Learning to Decide with AI Assistance under Human-Alignment

Nina Corvelo Benz
Max Planck Institute of Biochemistry
Planegg, Germany
corvelo@biochem.mpg.de

Eleni Straitouri
IMDEA Software Institute
Madrid, Spain
eleni.straitouri@software.imdea.org

Manuel Gomez-Rodriguez
Max Planck Institute for Software Systems
Kaiserslautern, Germany
manuelgr@mpi-sws.org

Abstract

It is widely agreed that when AI models assist decision-makers in high-stakes domains by predicting an outcome of interest, they should communicate the confidence of their predictions. However, empirical evidence suggests that decision-makers often struggle to determine when to trust a prediction based solely on this communicated confidence. In this context, recent theoretical and empirical work suggests a positive correlation between the utility of AI-assisted decision-making and the degree of alignment between the AI confidence and the decision-makers' confidence in their own predictions. Crucially, these findings do not yet elucidate the extent to which this alignment influences the complexity of learning to make optimal decisions through repeated interactions. In this paper, we address this question in the canonical case of binary predictions and binary decisions. We first show that this problem is equivalent to a binary online contextual learning problem with full feedback, and establish a lower bound of $\Omega(\sqrt{|\mathcal{H}| \cdot |\mathcal{B}| \cdot T})$ on the expected regret any learner can attain, where $\mathcal{H}$ and $\mathcal{B}$ denote the sets of human and AI confidence values. We then demonstrate that, under perfect alignment between AI and human confidence, a learner can attain an expected regret of $\mathcal{O}(\sqrt{|\mathcal{H}| \cdot T \log T})$ and, when $\sqrt{|\mathcal{H}|} = \mathcal{O}(\log T)$ and $\mathcal{B}$ is countable, a non-trivial generalization of the Dvoretzky-Kiefer-Wolfowitz inequality improves the regret bound to $\mathcal{O}(\sqrt{T \log T})$. Taken together, these results reveal that alignment can reduce the complexity of learning to make decisions with AI assistance. Experiments on real data from two different human-subject studies where participants solve simple decision-making tasks assisted by AI models show that our theoretical results are robust to violations of perfect alignment.

1 Introduction

The promise of AI-based decision support systems is that human experts using these systems will make *better* decisions than human experts not using them. However, multiple lines of empirical evidence across various domains suggest that this promise has not yet been reliably realized [Yin et al., 2019, Zhang et al., 2020, Suresh et al., 2020, Lai et al., 2023].

In the canonical case of binary predictions and binary decisions, Corvelo Benz and Gomez-Rodriguez [2023, 2025] have recently argued both theoretically and empirically that the way in which AI models quantify and communicate confidence in their predictions is one of the reasons AI-assisted decision

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

making falls short. First, they have shown that, if an AI model uses calibrated estimates of the probability that the prediction is correct as AI confidence values—as commonly done [Gneiting et al., 2007, Zadrozny and Elkan, 2001, Guo et al., 2017, Gupta and Ramdas, 2021, Huang et al., 2020, Wang et al., 2023]—a (rational) human expert who places more (less) trust on predictions with higher (lower) AI confidence may never make provably optimal decisions. Second, they have provided empirical evidence that closer alignment between AI and human confidence correlates with higher utility in AI-assisted decision making.

Crucially, the aforementioned results do not yet elucidate the extent to which the alignment between AI and human confidence influences the complexity of learning to make (optimal) decisions through repeated interactions [Noti and Chen, 2023, Buçinca et al., 2024, Noti et al., 2025, He et al., 2025]. In this paper, we start closing this gap.

Our contributions. We first show that, in the canonical case of binary predictions and binary decisions, the problem of learning to make decisions with AI assistance through repeated interactions is equivalent to a binary online contextual learning problem with full feedback [Slivkins, 2019]. Using this equivalence, we establish a lower bound of $\Omega(\sqrt{|\mathcal{H}| \cdot |\mathcal{B}| \cdot T})$ on the expected regret *any* learner can attain in solving the problem, where $\mathcal{H}$ and $\mathcal{B}$ denote the sets of human and AI confidence values. Further, we derive a simple online learning algorithm that, under perfect alignment between AI and human confidence, is guaranteed to attain an expected regret of $\mathcal{O}(\sqrt{|\mathcal{H}| \cdot T \log T})$. Additionally, when $\sqrt{|\mathcal{H}|} = \mathcal{O}(\log T)$ and $\mathcal{B}$ is countable, we show that a non-trivial generalization of the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [Dvoretzky et al., 1956, Massart, 1990] improves the regret bound to $\mathcal{O}(\sqrt{T \log T})$. Taken together, our theoretical results reveal that alignment between AI and human confidence can reduce the complexity of learning to make decisions with AI assistance. Finally, we complement our theoretical results with experiments on real data from two different human-subject studies where participants solve simple decision-making tasks assisted by AI models [Vodrahalli et al., 2022a, Corvelo Benz and Gomez-Rodriguez, 2025]. The experiments demonstrate that our theoretical results are robust to violations of perfect alignment—a learner who assumes perfect alignment can achieve lower expected regret than one making no such assumption, provided that a weaker notion of alignment holds.

Further related work. Our work builds upon further related work on AI-assisted decision making, multi-armed bandits, selective prediction and learning under algorithmic triage.

Our work is motivated by several empirical studies within the rapidly increasing literature on AI-assisted decision making (see Lai et al. [2023] for a recent review), which show that decision-makers have difficulties in modulating trust using confidence values [Vodrahalli et al., 2022b, Yona et al., 2022, Straitouri et al., 2023, 2025, Straitouri and Gomez-Rodriguez, 2024, De Toni et al., 2024, Göndöcs et al., 2025, Devic et al., 2025]. In this context, we should also note that other studies have analyzed how additional context, such as model explanations and accuracy, can help modulate trust [Papenmeier et al., 2019, Wang and Yin, 2021, Yin et al., 2019, Nourani et al., 2020, Zhang et al., 2020, Guo et al., 2025, Hullman et al., 2025, Mei et al., 2025].

Within the vast literature of multi-armed bandits (see Slivkins [2019] for a recent review) the DKW inequality has been used to derive regret bounds in settings in which the objective requires precise estimation of the entire reward distribution [Chen et al., 2016, Kearns et al., 2017, Keramati et al., 2020, Baudry et al., 2021, Howard and Ramdas, 2022], such as the conditional value at risk (CVaR). However, to the best of our knowledge, the DKW inequality has not been used to derive improved average regret bounds in contextual settings in which the optimal decision policy satisfies a monotonicity property with respect to the context as in our work.

There is a rich line of work on selective prediction, also referred to as predicting with the reject option [Cordella et al., 1995, De Stefano et al., 2000, Fumera and Roli, 2002, El-Yaniv et al., 2010, Cortes et al., 2016, Geifman and El-Yaniv, 2017, 2019, Hendrickx et al., 2024]. In selective prediction, the AI model has the option to abstain from (reject) a decision task whenever the confidence of the AI prediction is low. Algorithmic triage extends this paradigm by assigning the rejected decision tasks to a human while learning under algorithmic triage focuses on routing decision tasks to either a human or an AI model [Raghu et al., 2019, Mozannar and Sontag, 2020, De et al., 2020, 2021, Okati et al., 2021]. Our work focuses on a fundamentally different problem setting, where a human makes every decision, using the AI model strictly as an assistive tool.

2 Learning to Decide with AI Assistance

In this section, we first revisit the AI-assisted decision making task we focus in our work, which has also been studied elsewhere [Corvelo Benz and Gomez-Rodriguez, 2023, 2025]. Then, in the context of this task, we introduce the problem of learning an optimal decision policy through repeated interactions, and characterize its complexity.

2.1 AI-Assisted Decision Making with Binary Predictions and Binary Decisions

We consider an AI-assisted decision making task where, for each realization of the task, a decision maker first observes a set of features $(x, v) \in \mathcal{X} \times \mathcal{V}$, then takes a binary decision $a \in \{0, 1\}$ informed by a classifier's prediction $\hat{y} = \operatorname{argmax}_y f_y(x)$, as well as confidence $f_{\hat{y}}(x) \in \mathcal{B} \subseteq [0, 1]$, of a binary label of interest $y \in \{0, 1\}$, and finally receives a utility $u(a, y) \in \mathbb{R}$. Such an AI-assisted decision making process fits a variety of real-world applications. For example, in medical treatment, the features (x, v) may comprise multiple sources of information regarding a patient's health[1], the label y may indicate whether a patient would benefit from a specific treatment, the decision a may indicate whether the doctor applies the specific treatment to the patient, and the utility $u(a, y)$ may quantify the trade-off between health benefit to the patient and economic cost to the decision maker.

In what follows, rather than working with both $\hat{y}$ and $f_{\hat{y}}(x)$, we will work with just $b = f_1(x)$, which we will refer to as classifier's confidence, without loss of generality[2]. Moreover, we will assume that the gain from deciding $a = 1$ when $y = 1$ and the gain from deciding $a = 0$ when $y = 0$ is positive, *i.e.*,

$$u(1, 1) - u(0, 1) > 0 \quad \text{and} \quad u(0, 0) - u(1, 0) > 0, \tag{1}$$

a condition that it is natural under an appropriate choice of label and decision values. For example, in medical diagnosis, if $a = 1$ means the patient is tested early for a disease and $y = 1$ means the patient suffers the disease, the above condition implies that, for a patient who suffers the disease, the utility of testing them is greater than the utility of not testing them and, for a patient who does not suffer the disease, the utility of not testing them is greater than the utility of testing them.

Further, for each task instance, we formally characterize the decision maker's decision a through a decision policy $\pi(h, b) \in \{0, 1\}$, where $h \in \mathcal{H}$ and $b \in \mathcal{B} \subseteq [0, 1]$ denote the decision maker's confidence and the classifier's confidence that the value of the label of interest is $y = 1$, respectively.[3] Here, following previous behavioral studies showing that human's confidence h is discretized in a few distinct levels [Lisi et al., 2021, Zhang et al., 2015], we assume $\mathcal{H}$ is a totally ordered finite discrete set. Then, under this characterization, the problem of learning to make optimal decisions reduces to finding the optimal decision policy π^* that maximizes the expected utility, *i.e.*,

$$\pi^*(h, b) = \arg\max_a \mu(a \mid h, b) \text{ where } \mu(a \mid h, b) = \mathrm{E}_Y[u(a, Y) \mid H = h, B = b], \tag{2}$$

where the expectation is over the randomness of the label of interest Y.[4]

Next, we examine the extent to which a decision maker can efficiently learn the optimal policy π^* through repeated interactions—employing a sequence of policies π_t over subsequent task realizations.

2.2 Complexity of Learning to Decide with AI Assistance

Our starting point is the realization that, for binary predictions and binary decisions, the problem of learning to decide with AI assistance through repeated interactions is equivalent to a binary online contextual learning problem with full feedback [Slivkins, 2019]. In particular, at each time step t, the confidence values (h_t, b_t) correspond to the context, the decisions a corresponds to the arms, the utility $u(a_t, y_t)$ corresponds to the reward, and we have full feedback because, given the label of interest y_t, the decision maker can compute both $u(a_t, y_t)$ and $u(1 - a_t, y_t)$.[5] Such full feedback

[1] Our formulation allows for a subset of the features v to be available only to the decision maker but not to the classifier.

[2] We can recover $\hat{y}$ and $f_{\hat{y}}(x)$ from b, *i.e.*, if $b > 0.5$, we have that $\hat{y} = 1$ and $f_{\hat{y}}(x) = b$; if $b < 0.5$, $\hat{y} = 0$ and $f_{\hat{y}}(x) = 1 - b$.

[3] Similar to Corvelo Benz and Gomez-Rodriguez [2023], the decision maker's confidence h refers to the confidence the decision maker has that the label $Y = 1$ *before* observing the classifier's confidence b.

[4] We use upper-case for random variables and lower case for their realizations.

[5] In our setting, the decision maker is the learner, hence, there is no elicitation of confidence.

setting fits many high-stakes decision-making tasks well. For example, doctors order cultures for patients with infections to test for antimicrobial resistance, but they frequently initiate antibiotic treatment immediately before the results are finalized. Further, traders may decide to trade a few specific stocks, but they can observe the entire market's evolution over time. That said, there also exist decision-making tasks that do not fit the full-feedback setting. Refer to Section 5 for a more in-depth discussion.

Consequently, similarly as elsewhere in the online learning literature [Slivkins, 2019], we can measure to which extent a decision maker succeeds at efficiently finding the optimal decision policy by analyzing the expected cumulative regret. More formally, the expected cumulative regret $\mathbb{E}[R(T)]$ is defined as

$$\mathbb{E}[R(T)] = \mathbb{E}_{h_t,b_t\sim P(H,B)}\left[\sum_{t=1}^{T}\left(\mu(\pi^*(h_t,b_t)\,|\,h_t,b_t) - \mathbb{E}_{a_t\sim P(A_t\,|\,h_t,b_t)}[\mu(a_t\,|\,h_t,b_t)]\right)\right], \quad (3)$$

where $P(H,B)$ denotes the distribution of the decision maker's and the classifier's confidence, which implicitly depends on the feature distribution $P(X,V)$, and $P(A_t\,|\,h_t,b_t)$ denotes the distribution of decisions induced by the decision policy π_t used at time step t.

In what follows, we establish a lower bound on the expected regret using a non-trivial extension of a classical result for contextual bandits with partial feedback from Agarwal et al. [2012] to our full feedback setting.[6]

Theorem 1. *For any choice of decision policies π_t based on historical observations, there exists a problem instance such that $\mathbb{E}[R(T)] = \Omega(\sqrt{|\mathcal{H}|\cdot|\mathcal{B}|\cdot T})$.*

The above result reveals that, in general, the complexity of learning the optimal decision policy through repeated interactions depends on the size of the sets of possible human and AI confidence values, $\mathcal{H}$ and $\mathcal{B}$.[7] Importantly, in the next section, we will show that, under perfect alignment between AI and human confidence, this dependence vanishes, reducing the overall complexity.

3 Learning to Decide with AI Assistance under Perfect Alignment

In this section, we first show that, under perfect alignment between AI and human confidence, the optimal decision policy π^* is a threshold function defined over the AI confidence, with a threshold that depends on the human confidence. Next, we present a simple online algorithm that, by utilizing this structure, is guaranteed to attain an expected regret of $\mathcal{O}(\sqrt{|\mathcal{H}|\cdot T\log T})$ in learning the optimal decision policy π^*. Finally, when $\sqrt{|\mathcal{H}|} = \mathcal{O}(\log T)$ and $\mathcal{B}$ is countable, we improve the regret bound to $\mathcal{O}(\sqrt{T\log T})$. In Appendix B, we further show that, whenever perfect alignment does not hold, the degree of alignment bounds the gap between the expected utility achieved by the decision policy found by our algorithm and the optimal policy.[8]

3.1 Optimal Decision Policy Under Perfect Alignment

There is perfect alignment between AI and human confidence if, for any $h, h' \in \mathcal{H}$ and $b, b' \in \mathcal{B}$ such that $h \leq h'$ and $b \leq b'$, we have that [Corvelo Benz and Gomez-Rodriguez, 2023]

$$P(Y=1\,|\,H=h, B=b) - P(Y=1\,|\,H=h', B=b') \leq 0. \quad (4)$$

Intuitively, perfect alignment between the classifier's and human confidence means the odds of a positive outcome (Y=1) increase (monotonically) with both the AI and human confidence. Hence, under perfect alignment, if both the AI and human are more confident for instances in a group A than in another group B, it can never happen that the odds of a positive outcome in group B are higher than group A.

For utility functions u satisfying Eq. 1, the following theorem shows that, for each $h \in \mathcal{H}$, there is a threshold value $b^*(h)$ such that $\pi^*(h,b) = 1$ if $b > b^*(h)$ and $\pi^*(h,b) = 0$ otherwise.

[6] All proofs are provided in Appendix A.

[7] Our lower bound is only meaningful whenever both $\mathcal{H}$ and $\mathcal{B}$ are finite discrete sets. Note that, without further assumptions on the decision policy, learning is impossible under infinite context [Shalev-Shwartz and Ben-David, 2014].

[8] The algorithms presented are not meant to replace the human decision maker, but to help us understand the complexity of learning to make decisions under AI-assistance.

Algorithm 1 Online Learning under Perfect Alignment

Require: # interactions T, set of human confidence values $\mathcal{H}$, set of AI confidence values $\mathcal{B}$
  Initialize $\bar{\mu}_1(b \mid h)$, $\forall (h, b) \in \mathcal{H} \times \mathcal{B}$
  for t from 1 to T **do**
    Observe confidence values (h_t, b_t)
    Compute unbiased estimates $\bar{\mu}_t(b \mid h_t)$, $\forall b \in \mathcal{B}$ using Eq. 6
    $\bar{b}_t(h_t) \leftarrow \arg\max_{b \in \mathcal{B}} \bar{\mu}_t(b \mid h_t)$
    if $b_t > \bar{b}_t(h_t)$ **then**
      $a_t \leftarrow 1$
    else
      $a_t \leftarrow 0$
    end if
    Take decision a_t and observe label value y_t
  end for

Theorem 2. *Under perfect alignment, for any $h \in \mathcal{H}$, the optimal decision policy $\pi^*(h, b)$ is given by a threshold function $\pi^*(h, b) = \mathbb{I}[b > b^*(h)]$ where*

$$b^*(h) = \sup \left\{ b \in \mathcal{B} \;\middle|\; P(Y = 0 \mid H = h, B = b) \geq \frac{u(1,1) - u(0,1)}{u(1,1) - u(1,0) + u(0,0) - u(0,1)} \right\}.$$

Importantly, the above result reveals that learning the optimal policy essentially reduces to learning $b^*(h)$, $\forall h \in \mathcal{H}$. In what follows, we present a simple online algorithm to learn $b^*(h)$, $\forall h \in \mathcal{H}$, and show that it attains an expected regret that no longer depends on $\mathcal{B}$ and, when $\mathcal{B}$ is countable and $\sqrt{|\mathcal{H}|} = \mathcal{O}(\log T)$, no longer depends on $\mathcal{H}$.

3.2 An Efficient Algorithm for Learning to Decide Under Perfect Alignment

Our starting point is the realization that the problem of learning $b^*(h)$, $\forall h \in \mathcal{H}$ through repeated interactions is equivalent to $|\mathcal{H}|$ independent multiclass online learning problems with full feedback—one per value of h. For each of these problems, at each time step t, the (candidate) thresholds $b \in \mathcal{B}$ correspond to the arms, the utility $u(\mathbb{I}[b_t > b], y_t)$ corresponds to the reward, and we have full feedback because, given the label of interest y_t, the decision maker can compute the utility $u(\mathbb{I}[b_t > b], y_t)$ for all thresholds $b \in \mathcal{B}$.

To solve the above multiclass online learning problems, our online algorithm leverages the following key theorem:

Theorem 3. *Let $\mu(b \mid h)$ be the conditional expected utility achieved by the decision policy π defined by a threshold function with threshold b given the human confidence $H = h$,* i.e.*,*

$$\mu(b \mid h) = \mathbb{E}\left[\mu(1 \mid H, B) \cdot \mathbb{I}[B > b] \mid H = h\right] + \mathbb{E}\left[\mu(0 \mid H, B) \cdot \mathbb{I}[B \leq b] \mid H = h\right], \tag{5}$$

where $\mu(a \mid H, B)$ is the conditional expected utility achieved by decision a given the human and AI confidence H and B, defined in Eq. 2. Then, it holds that $b^*(h) = \arg\max_{b \in \mathcal{B}} \mu(b \mid h)$.

In a nutshell, our online algorithm proceeds as follows. At each time step t, it first computes an unbiased estimate of the conditional expected utility $\mu(b \mid h_t)$ for each candidate threshold $b \in \mathcal{B}$, *i.e.*,

$$\begin{aligned} \bar{\mu}_t(b \mid h_t) = \frac{1}{n_t(h_t)} \sum_{t' < t : h_{t'} = h_t} & \left(\mathbb{I}[y_{t'} = 0, b_{t'} > b] \cdot (u(1,0) - u(1,1))\right. \\ & + \mathbb{I}[y_{t'} = 0, b_{t'} \leq b] \cdot (u(0,0) - u(0,1)) \\ & \left. + \mathbb{I}[b_{t'} \leq b] \cdot (u(0,1) - u(1,1)) + u(1,1)\right). \end{aligned} \tag{6}$$

where $n_t(h_t)$ is the number of observations with human confidence equal to h_t until time step t. Then, it picks the threshold $\bar{b}_t(h_t)$ with the highest estimated conditional expected utility $\bar{\mu}_t(b \mid h_t)$. Finally, it takes decision $a_t = 1$ if $b_t > \bar{b}_t(h_t)$ and $a_t = 0$ otherwise. Algorithm 1 summarizes the overall algorithm, where note that the $|\mathcal{H}|$ independent multiclass online learning problems are solved simultaneously.

The following theorem bounds the expected regret attained by our algorithm:

Theorem 4. *Algorithm 1 attains expected regret* $\mathbb{E}[R(T)] = \mathcal{O}\left(\sqrt{|\mathcal{H}| \cdot T \log T}\right)$ *under perfect alignment.*

The proof of the above theorem relies on two key steps. First, we show that, under perfect alignment, our algorithm's expected regret on the $|\mathcal{H}|$ independent instances of the multiclass online learning problem is equivalent to its expected regret on the binary online contextual learning problem with full feedback introduced in Section 2.2.[9] Second, we use the DKW inequality [Dvoretzky et al., 1956, Massart, 1990] to obtain a uniform bound on the estimation error of $\bar{\mu}(b \mid h_t)$ across all b values, which is the key to achieving a regret bound independent of the AI confidence b.

Next, when $\sqrt{|\mathcal{H}|} = \mathcal{O}(\log T)$ and $\mathcal{B}$ is countable, we derive a tighter regret bound, which is independent not only on the AI confidence b but also on the human confidence h. The key idea is to reformulate our algorithm in terms of a function $\ell : \mathcal{H} \to \mathcal{B}$ rather than in terms of $|\mathcal{H}|$ independent thresholds. This reformulation leverages the following key theorem, which is the counterpart of Theorem 3:

Theorem 5. *Let* $\mathcal{L} = \{\ell(\cdot) \mid \ell : \mathcal{H} \to \mathcal{B}\}$ *and, for each* $\ell \in \mathcal{L}$*, let* $\mu(\ell)$ *be the expected utility achieved by the decision policy* $\pi(h, b) = \mathbb{I}[b > \ell(h)]$*,* i.e.*,*

$$\mu(\ell) := \mathbb{E}\left[\mu(1 \mid H, B) \cdot \mathbb{I}[B > \ell(H)]\right] + \mathbb{E}\left[\mu(0 \mid H, B) \cdot \mathbb{I}[B \leq \ell(H)]\right], \tag{7}$$

where $\mu(a \mid H, B)$ *is the conditional expected utility achieved by decision* a *given the human and AI confidence* H *and* B*, and let* $\ell^* = \arg\max_{\ell \in \mathcal{L}} \mu(\ell)$*. Then, for all* $h \in \mathcal{H}$*, it holds that* $b^*(h) = \ell^*(h)$*.*

More specifically, under this reformulation, the algorithm proceeds as follows. At each time step t, it first computes an unbiased estimate of the expected utility $\mu(\ell)$, *i.e.*,

$$\begin{aligned}\bar{\mu}_t(\ell) = P_t(Y = 0, B > \ell(H)) \cdot (u(1,0) - u(1,1)) + P_t(Y = 0, B \leq \ell(H)) \cdot (u(0,0) - u(0,1)) \\ + P_t(B \leq \ell(H)) \cdot (u(0,1) - u(1,1)) + u(1,1),\end{aligned} \tag{8}$$

where $P_t(E) := \sum_{t' < t} \mathbb{I}[E_{t'}]$. Then, it picks the function $\bar{\ell}_t \in \mathcal{L}$ with the highest estimated expected utility $\bar{\mu}_t(\ell)$. Finally, it takes decision $a_t = 1$ if $b_t > \bar{\ell}_t(h_t)$ and $a_t = 0$ otherwise.

To see why the above reformulation of our algorithm is faithful to the original, note that

$$\max_{\ell \in \mathcal{L}} \bar{\mu}_t(\ell) = \frac{1}{n} \sum_{h \in \mathcal{H}} \max_{\ell(h) \in \mathcal{B}} n_t(h) \cdot \bar{\mu}_t(\ell(h) \mid h), \tag{9}$$

and, as an immediate consequence, it readily follows that $\bar{\ell}_t(h_t) = \bar{b}_t(h_t)\ \forall t$.

The following theorem bounds the expected regret attained by the above reformulation of our algorithm, which improves the bound in Theorem 4 by a factor of $\log T$ when $\sqrt{|\mathcal{H}|} = \mathcal{O}(\log T)$ and $\mathcal{B}$ is countable:

Theorem 6. *Under perfect alignment, there exists constant* $c > 0$ *such that, if* $\sqrt{|\mathcal{H}|} < c \cdot \log(T)$ *and* $\mathcal{B}$ *is countable, Algorithm 1 attains expected regret* $\mathbb{E}[R(T)] = \mathcal{O}\left(\sqrt{T \log T}\right)$*.*

The proof of the above theorem is analogous to the proof of Theorem 4 but relies on a non-trivial extension of the DKW inequality to obtain a uniform bound on the estimation error of $\bar{\mu}(\ell)$ across all $\ell \in \mathcal{L}$, which is the key to achieving a regret bound that is not only independent of the AI confidence b but also independent of the human confidence h and may be of independent interest (see Appendix A.7 and Theorem 16 for more details).

Remark. A naive implementation of Algorithm 1 may require computing $|\mathcal{B}|$ values of the estimated expected utility, one for each $b \in \mathcal{B}$. However, if $\mathcal{B}$ is large, such a naive implementation may be very inefficient and, if $\mathcal{B}$ is a continuous set, it may be infeasible. To overcome this issue, we can observe that the estimated expected utility can only take at most $T + 1$ different values at T time steps. This is because, for any observed AI confidence values b_t and $b_{t'}$ such that $b_t < b_{t'}$ and no other observed AI confidence value lies in between, all decision policies $\pi(b, h) = \mathbf{1}[b > \bar{b}_t(h)]$ using any threshold $\bar{b}_t(h)$ such that $b_{t'} \leq \bar{b}_t(h) < b_t$ will take the same decision for any observed $b \leq b_{t'}$ or $b_t \leq b$, and thus have the same estimated expected utility value.

Algorithm 2 Vanilla Contextual Online Learning

Require: # interactions T, set of human confidence values $\mathcal{H}$, set of AI confidence values $\mathcal{B}$
  Initialize $\bar{\mu}_1(a \mid h, b), \forall a \in \{0, 1\}, \forall (h, b) \in \mathcal{H} \times \mathcal{B}$
  for each time step t from 1 to T **do**
    Observe confidence values (h_t, b_t)
    for $a \in \{0, 1\}$ **do**
      $\bar{\mu}_t(a \mid h_t, b_t) \leftarrow \sum_{t' < t : (h_{t'}, b_{t'}) = (h_t, b_t)} (\mathbb{I}[y_{t'} = 1] \cdot u(a, 1) + \mathbb{I}[y_{t'} = 0] \cdot u(a, 0))$
    end for
    if $\bar{\mu}_t(0 \mid h_t, b_t) \geq \bar{\mu}_t(1 \mid h_t, b_t)$ **then**
      $a_t \leftarrow 0$
    else
      $a_t \leftarrow 1$
    end if
    Take action a_t and observe label value y_t
  end for

4 Experiments

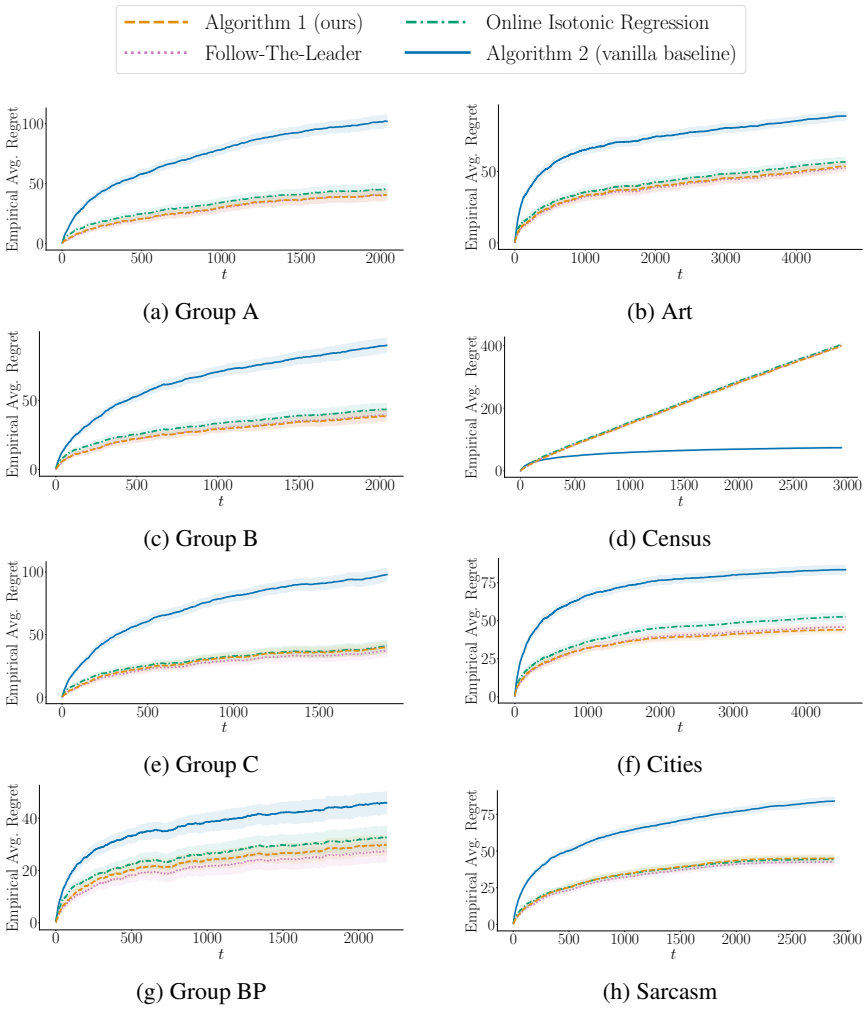
In this section, we estimate the expected regret attained by our algorithm (Algorithm 1), three classical online learning algorithms that can exploit the monotonicity of the outcome probabilities under perfect alignment, and a vanilla contextual online learning baseline (Algorithm 2) using real data from two human-subject studies. In these studies, participants solve simple decision-making tasks assisted by AI models with different degrees of alignment between AI and human confidence, as measured by the maximum alignment error (MAE) and the expected alignment error (EAE)[10].

Human-Alignment dataset by Corvelo Benz and Gomez-Rodriguez [2025]. The Human-Alignment dataset was gathered through a large-scale ($n = 703$) human-subject study where participants predicted the binary outcome of a card game and reported their confidence both before and after receiving AI assistance. Here, the participants' predictions and game outcomes correspond to the decisions a and the label of interest y, respectively. In the study, participants are assigned to one of four groups: A, B, BP, and C. In groups A, B, and C, participants are assisted by the same AI model; however, the distribution of cards shown to them differs, varying the degree of alignment between AI and human confidence across groups. In group BP, participants see the same card distribution as in group B, but they are assisted by a modified version of the AI model used in the other groups, which has been postprocessed to increase alignment using held-out data from group B. Throughout all groups, participants are asked to state their confidence in a positive prediction ($a = 1$) on a scale of $[0, 100]$. Afterwards the confidence of the AI assistant in a positive prediction is shown to the participants in the same scale and participants are asked to make a final prediction. Analogously to Corvelo Benz and Gomez-Rodriguez [2025], we bin participants' confidence into $|\mathcal{H}| = 4$ confidence value classes, {very low, low, high, very high}, and rescale the AI confidence to $[0, 1]$. The game outcomes are sampled from a distribution $P(Y \mid q)$, where the value q is a game-specific parameter unknown to participants (refer to Appendix C for details). Further, in groups A, B, and C, the AI confidence is independent of the human confidence and $|\mathcal{B}| = 13$. In contrast, in group BP, the AI confidence depends on the human confidence and, for each h, $|\mathcal{B}| = 5$. Table 1 in Appendix C shows the MAE and EAE, and the number of data samples per group.

Human-AI Interactions dataset by Vodrahalli et al. [2022a]. The Human-AI Interactions dataset was gathered through a large-scale human subject study where participants solved different binary classification tasks with AI assistance. Throughout our experiments, we used a subset of this dataset processed by Corvelo Benz and Gomez-Rodriguez [2023], which comprises $15{,}063$ unique binary predictions from 471 participants, together with their confidence, on four tasks: i) 'Art', ii) 'Sarcasm', iii) 'Cities', and iv) 'Census'. Participants had to make a binary prediction on i) the art period of a painting (given two options), ii) presence of sarcasm in text (yes/no option), iii) which US city is shown in an image (given two options), and iv) if an individual earns more than 50K a year given

[9] Note that our algorithm can be used to solve the binary online contextual learning problem with full feedback introduced in Section 2.2 because thresholds $b \in \mathcal{B}$ determine decisions $a \in \{0, 1\}$.

[10] The MAE and EAE are defined as $\text{MAE} = \max_{h \leq h', b \leq b'} P(Y = 1 \mid H = h, B = b) - P(Y = 1 \mid H = h', B = b')$ and $\text{EAE} = \frac{1}{|\mathcal{H}| \cdot |\mathcal{B}|} \sum_{h \leq h', b \leq b'} [P(Y = 1 \mid H = h, B = b) - P(Y = 1 \mid H = h', B = b')]_+$



(a) Group A (b) Art

(c) Group B (d) Census

(e) Group C (f) Cities

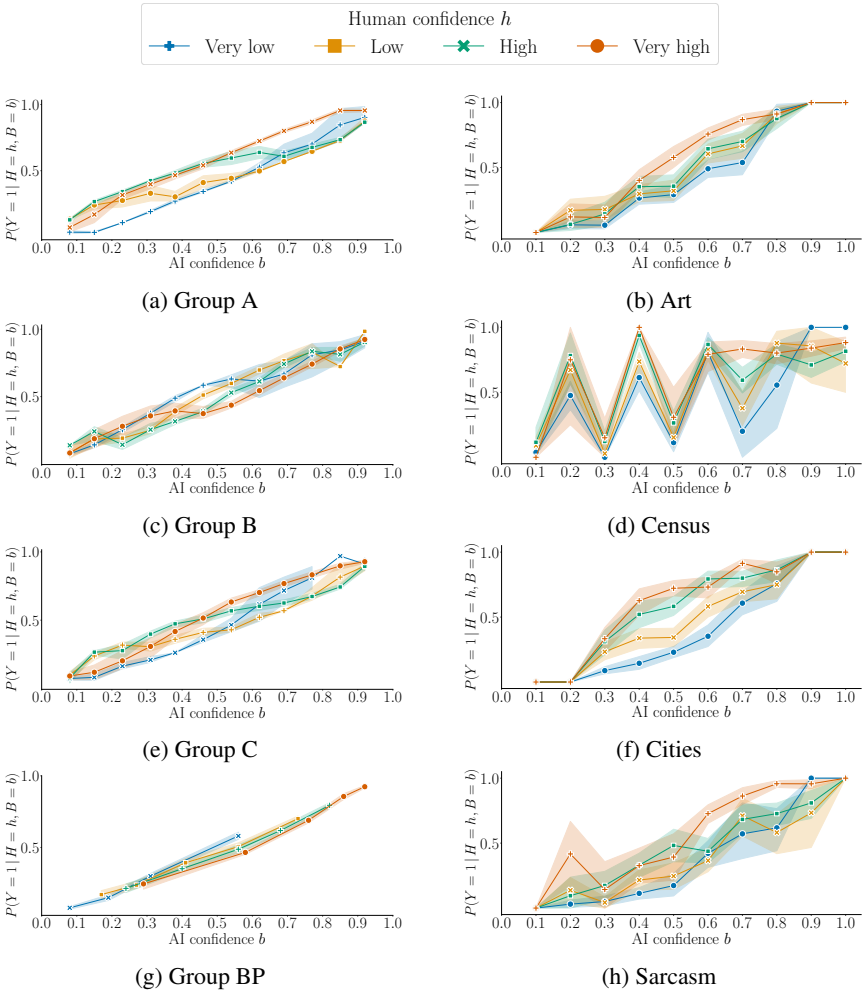
(g) Group BP (h) Sarcasm

Figure 1: Empirical average regret attained by our algorithm (Algorithm 1), online isotonic regression, follow-the-leader, and a vanilla contextual online learning baseline (Algorithm 2) per time step t for each group in the Human-Alignment dataset (left) and each task in the Human-AI Interactions dataset (right). The average is over 100 executions and shaded areas represent 95% confidence intervals.

their census data (yes/no option), respectively. Participants' confidence is recorded on a scale of $[-1, 1]$, where -1 and 1 represent full confidence in one of the options respectively, and 0 represents no confidence in either option. After the initial confidence elicitation[11], the AI confidence is shown to participants on the same scale as the one used to record the human confidence, and participants can adjust their confidence and final prediction. Throughout all tasks, the participants' predictions and the groundtruth correspond to the decisions a and the label of interest y, respectively, where we define the positive outcome ($y = 1$) and decision ($a = 1$) at random to one of the two given options for tasks i) and iii) and to the "yes" option for task ii) and iv). In each task, the AI model used to provide assistance is fitted on labels by human annotators, the AI confidence is a noisy average of the annotators' label confidence, and the degree of alignment varies across tasks. For better comparison in our experiments, we bin participants' confidence into $|\mathcal{H}| = 4$ confidence value classes, {very low, low, high, very high}, and the AI confidence to $|\mathcal{B}| = 10$ confidence value classes rescaled to $[0, 1]$. Table 2 in Appendix C shows the MAE and EAE, and the number of data samples per task.

Experimental Setup. We estimate the expected regret attained by our algorithm (Algorithm 1), two classical online learning algorithms that can exploit the monotonicity of the outcome probabilities

[11] In the experiments, we take the reported confidence as ground truth for the decision maker's confidence. However, noisy reporting of the decision maker's confidence can have an effect on the alignment of the AI which we discuss in section 5.



(a) Group A (b) Art

(c) Group B (d) Census

(e) Group C (f) Cities

(g) Group BP (h) Sarcasm

Figure 2: Empirical probability $P(Y = 1 \mid H = h, B = b)$ for each value of the human confidence h and AI confidence b for each group in the the Human-Alignment dataset (left) in the Human-AI Interactions dataset (right). Each line corresponds to the empirical probability value for fixed h against values of b. Shaded areas represent 95% confidence intervals.

under perfect alignment, and a vanilla contextual online learning baseline (Algorithm 2) on each of the groups (A, B, BP and C) and tasks ('Art', 'Sarcasm', 'Cities' and 'Census'), using all the data samples. The two classical algorithms are: i) online isotonic regression [Kotłowski et al., 2016], ii) follow-the-leader algorithm (FTL) [Kotłowski, 2018]. In both algorithms, one instance of the algorithm is used for each $h \in \mathcal{H}$ observed and, for FTL, we define the experts as threshold classifiers over $\mathcal{B}$ and $\mathcal{B}$ needs to be known a priori. See Appendix C for detailed descriptions of each algorithm. In all our experiments, we use the utility function $u(a, y) = \mathbb{I}[a = y] - \mathbb{I}[a \neq y], \forall a, y \in \{0, 1\}$ and, to estimate the expected regret, we repeat each experiment 100 times using different random seeds and, in each repetition, we randomly shuffle all data samples.

Results. Figure 1 shows the (estimated) expected regret attained by our algorithm, online isotonic regression, FTL, and the baseline on both datasets. We find that, with the exception of the 'Census' task, our algorithm, the online isotonic regression algorithm and the FTL algorithm learn the optimal policy for all groups and tasks and achieve comparable regret, lower than the vanilla baseline, even though perfect alignment does not hold. This is because the regret guarantees of the implementation of our algorithm, online isotonic regression and FTL only require the probability $P(Y = 1 \mid H, B)$ to be monotone in B for any given H—a condition representing a weaker notion of perfect alignment. In our experiments, this probability is approximately monotone in B for most H values across nearly all groups and tasks, with the exception of the 'Census' task as seen in Figure 2. These findings

strongly support our main claim—a high degree of alignment speeds up learning under AI-assistance, regardless of the specific learning algorithm.

5 Discussion and Limitations

In this section, we highlight several limitations of our work, discuss its broader impact, and point out avenues for future research.

Monotonicity of the optimal decision policy. In our work, we have shown that, under perfect alignment, a decision-maker can learn the optimal decision policy more efficiently by focusing only on the set of decision policies monotone in the AI confidence. However, Corvelo Benz and Gomez-Rodriguez [2023] have shown that, under perfect alignment, the optimal decision policy lies within the set of policies monotone in both the human and the AI confidence, which is a subset of the set of decision policies monotone in the AI confidence. Therefore, it would be very interesting to investigate whether, under perfect alignment, a decision-maker can learn the optimal decision policy even more efficiently by focusing on this latter set.

Decision making setting. We have considered AI-assisted decision making processes where the label y is always observable independently of the decision taken—there is full-feedback. However, there exist decision making tasks in which the label y is only realized whenever the decision $a = 1$. For example, we can only observe if a treatment is eventually beneficial or not if the patient is treated, or we can only observe if a customer does or does not default on a loan if the loan is granted. Our lower bound on the expected regret can be easily extended to such partial-feedback setting using a classical result for contextual bandits by Agarwal et al. [2012]. However, investigating the extent to which a decision-maker can leverage the alignment between AI and human confidence to learn more efficiently under partial-feedback is significantly more challenging, and it is left for future work. Further, in our setting, we have assumed that the AI model and the degree of alignment are fixed. It would be interesting to extend our analysis to a setting where both the AI model and the degree of alignment can change over time. To this end, recent work on online calibration [Gupta and Ramdas, 2023] and online multicalibration [Garg et al., 2024, Ghuge et al., 2025] may be proven useful.

Broader impact. AI promises to enhance decision making in a variety of high-stakes domains; however, empirical evidence suggests that decision-makers often struggle to make optimal decisions under AI assistance. Through theoretical and empirical analysis, our work suggests that increasing the alignment between the AI confidence and the decision maker's confidence on their own predictions may help decision-makers learn to make optimal decisions through repeated interactions. In practice, we can test for alignment by gathering confidence values from a decision maker and the corresponding outcomes from their decisions. Using this data, one can i) measure the degree of alignment of different AI models, and ii) improve alignment of any model through post-processing using multicalibration algorithms [Corvelo Benz and Gomez-Rodriguez, 2023]. Models with high degree of alignment and good performance can then be used to assist the human decision maker. Note that problems with noisy reporting of confidence values by human decision makers can affect the eventual alignment checks or post-processing of the AI-assistant for alignment. However, we have empirically shown that our results are robust to violations of perfect alignment.

6 Conclusions

We have theoretically shown that, when AI models assist decision-makers by predicting a binary outcome of interest, perfect alignment between the AI confidence and the decision-makers' confidence in their own predictions can be used to learn the optimal decision policy more efficiently. Moreover, using real data from two human-subject studies, we have empirically demonstrated that our theoretical results are robust to certain violations of perfect alignment—a learner who assumes perfect alignment can achieve lower expected regret than one making no such assumption, provided that a weaker notion of alignment holds.

Acknowledgements. Gomez-Rodriguez acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101169607). Straitouri acknowledges support from a Google PhD Fellowship.

References

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

Harini Suresh, Natalie Lao, and Ilaria Liccardi. Misplaced trust: Measuring the interference of machine learning in human decision-making. In *Proceedings of the ACM Conference on Web Science*, pages 315–324. ACM, 2020.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2023.

Nina Corvelo Benz and Manuel Gomez-Rodriguez. Human-aligned calibration for ai-assisted decision making. *Advances in Neural Information Processing Systems*, 36:14609–14636, 2023.

Nina Corvelo Benz and Manuel Gomez-Rodriguez. Human-alignment influences the utility of ai-assisted decision making. *Scientific Reports*, 15(1):29154, 2025.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, 2017.

Chirag Gupta and Aaditya K. Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 2020.

Lequn Wang, Thorsten Joachims, and Manuel Gomez-Rodriguez. Improving screening processes via calibrated subset selection. In *Proceedings of the 39th International Conference on Machine Learning*, 2023.

Gali Noti and Yiling Chen. Learning when to advise human decision makers. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3038–3048, 2023.

Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Susan A Murphy, and Krzysztof Z Gajos. Towards optimizing human-centric objectives in ai-assisted decision-making with offline reinforcement learning. *arXiv preprint arXiv:2403.05911*, 2024.

Gali Noti, Kate Donahue, Jon Kleinberg, and Sigal Oren. Ai-assisted decision making with human learning. In *Proceedings of the 26th ACM Conference on Economics and Computation*, pages 38–39, 2025.

Kevin He, Ran Shorrer, and Mengjia Xia. Human misperception of generative-ai alignment: A laboratory experiment. In *Proceedings of the 26th ACM Conference on Economics and Computation*, pages 739–739, 2025.

A. Slivkins. *Introduction to Multi-Armed Bandits*. Foundations and Trends in Machine Learning Series. Now Publishers, 2019. ISBN 9781680836219. URL `https://books.google.com/books?id=6ViCzQEACAAJ`.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 763–777, 2022a.

Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. Uncalibrated models can improve human-ai collaboration. In *Advances in Neural Information Processing Systems*, 2022b.

Gal Yona, Amir Feder, and Itay Laish. Useful confidence measures: Beyond the max score. *arXiv preprint arXiv:2210.14070*, 2022.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez-Rodriguez. Improving expert predictions with conformal prediction. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Eleni Straitouri, Stratis Tsirtsis, Ander Artola Velasco, and Manuel Gomez-Rodriguez. Narrowing action choices with ai improves human sequential decisions. In *AISTATS*, 2025.

Eleni Straitouri and Manuel Gomez-Rodriguez. Designing decision support systems using counterfactual prediction sets. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46722–46744, 2024.

Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. Towards human-ai complementarity with prediction sets. *Advances in Neural Information Processing Systems*, 37: 31380–31409, 2024.

Dóra Göndöcs, Szabolcs Horváth, and Viktor Dörfler. Uncovering the dynamics of human-ai hybrid performance: A qualitative meta-analysis of empirical studies. *International journal of human-computer studies*, page 103622, 2025.

Siddartha Devic, Tejas Srinivasan, Jesse Thomason, Willie Neiswanger, and Vatsal Sharan. From calibration to collaboration: Llm uncertainty quantification should be more human-centered. *arXiv preprint arXiv:2506.07461*, 2025.

Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*, 2019.

Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

Mahsan Nourani, Joanie King, and Eric Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 112–121, 2020.

Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. The value of information in human-ai decision-making. In *The Fourteenth International Conference on Learning Representations*, 2025.

Jessica Hullman, Ziyang Guo, and Berk Ustun. Explanations are a means to an end. *arXiv preprint arXiv:2506.22740*, 2025.

Haiyang Mei, Difei Gao, Xiaopeng Wei, Xin Yang, and Mike Zheng Shou. Can i trust you? advancing gui task automation with action trust score. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8692–8700, 2025.

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29, 2016.

Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *International conference on machine learning*, pages 1828–1836. PMLR, 2017.

Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4436–4443, 2020.

Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In *International Conference on Machine Learning*, pages 716–726. PMLR, 2021.

Steven R Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to a/b testing and best-arm identification. *Bernoulli*, 28(3):1704–1728, 2022.

Luigi Pietro Cordella, Claudio De Stefano, Francesco Tortorella, and Mario Vento. A method for improving classification reliability of multilayer perceptrons. *IEEE transactions on neural networks*, 6(5):1140–1147, 1995.

Claudio De Stefano, Carlo Sansone, and Mario Vento. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(1):84–94, 2000.

Giorgio Fumera and Fabio Roli. Support vector machines with embedded reject option. In *International workshop on support vector machines*, pages 68–82. Springer, 2002.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. *Advances in neural information processing systems*, 29, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR, 2019.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *Machine Learning*, 113(5):3073–3110, 2024.

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.

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

Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez-Rodriguez. Classification under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. In *Advances in Neural Information Processing Systems*, 2021.

Matteo Lisi, Gianluigi Mongillo, Georgia Milne, Tessa Dekker, and Andrei Gorea. Discrete confidence levels revealed by sequential decisions. *Nature Human Behaviour*, 5(2):273–280, 2021.

Hang Zhang, Nathaniel D Daw, and Laurence T Maloney. Human representation of visuo-motor uncertainty as mixtures of orthogonal basis distributions. *Nature neuroscience*, 18(8):1152–1158, 2015.

Alekh Agarwal, Miroslav Dudik, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. PMLR, 21–23 Apr 2012.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Wojciech Kotłowski, Wouter M. Koolen, and Alan Malek. Online isotonic regression. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1165–1189, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v49/kotlowski16.html`.

Wojciech Kotłowski. On minimaxity of follow the leader strategy in the stochastic setting. *Theoretical Computer Science*, 742:50–65, 2018.

Chirag Gupta and Aaditya Ramdas. Online platt scaling with calibeating. In *International Conference on Machine Learning*, pages 12182–12204. PMLR, 2023.

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2725–2792. SIAM, 2024.

Rohan Ghuge, Vidya Muthukumar, and Sahil Singla. Improved and oracle-efficient online ℓ_1-multicalibration. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 19437–19457. PMLR, 13–19 Jul 2025.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

Olivier Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, pages 213–247. Springer, 2003.

Bodhisattva Sen. A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 11:28–29, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

A Proofs

A.1 Proof of Theorem 1

Our proof closely follows the proof of Theorem 5.1 in Agarwal et al. [2012], which shows that learning the optimal policy π^* for the contextual multi-arm bandit problem with a policy class Π and K arms (refer to Slivkins [2019] Chapter 8.4 for the exact definition) has expected regret $\mathbb{E}(R(T)) = \Omega\left(\sqrt{K \cdot T \cdot \log(|\Pi|)/\log(K)}\right)$, as long as $\pi^* \in \Pi$. This is because learning the best-response policy $\pi^*(h, b)$ given by Eq. 2 is equivalent to an instance of the contextual bandit problem with a policy class with context $z = (h, b)$, $K = 2$ arms with each decision a being an arm, and a policy class $\Pi = \{0, 1\}^{|\mathcal{H}|\cdot|\mathcal{B}|}$. However, the key difference between our setup and the one considered in Agarwal et al. [2012] is that at each time step, rather than a single utility, we observe the utility for all arms, *i.e.*, we assume *full feedback*.

Under this assumption we show that we achieve a tighter lower bound for the expected regret, which is essentially due to achieving a tighter upper bound in Lemma A.1 by Auer et al. [2002]. This Lemma considers a K-arm bandit problem with a horizon of T time steps and the following utility distributions: a) the distribution P_i, where the utility $R \sim Bernoulli(1/2 + \epsilon)$, $\epsilon \in (0, 1/2)$ for an arm i chosen uniformly at random before the first time step and for any arm $i' \neq i$ the utility $R \sim Bernoulli(1/2)$; b) the distribution P_{unif}, where the utility $R \sim Bernoulli(1/2)$ for all arms. In what follows, we will first derive the equivalent of Lemma A.1 by Auer et al. [2002] under full feedback, on which we will base the rest of our proof.

Lemma 1. *Let f be any function $\{0, 1\}^{K\times T} \to [0, M]$ defined on a utility matrix $\mathbf{R}$ comprising the utilities under full feedback for any bandit algorithm over T time steps. Then for any decision i,*

$$\mathbb{E}_i[f(\mathbf{R})] - \mathbb{E}_{unif}[f(\mathbf{R})] \leq \frac{M}{2}\sqrt{T \cdot \frac{-\log(1 - 4\epsilon^2)}{4}},$$

where $\mathbb{E}_i$, $\mathbb{E}_{unif}$ denote the expectation with respect to the distribution P_i and P_{unif} respectively.

Proof. Similar to Lemma A.1 Auer et al. [2002] we have

$$\begin{aligned}\mathbb{E}_i[f(\mathbf{R})] - \mathbb{E}_{unif}[f(\mathbf{R})] &= \sum_{\mathbf{R}} f(\mathbf{R})(P_i[\mathbf{R}] - P_{unif}[\mathbf{R}]) \leq M \cdot \sum_{\mathbf{R}}(P_i[\mathbf{R}] - P_{unif}[\mathbf{R}]) \\ &\leq M \cdot \sum_{\mathbf{R}:P_i[\mathbf{R}]\geq P_{unif}[\mathbf{R}]} (P_i[\mathbf{R}] - P_{unif}[\mathbf{R}]) = \frac{M}{2}||P_i - P_{unif}||_1,\end{aligned} \tag{10}$$

where $||P_i - P_{unif}||_1$ is the variational distance between the distributions P_i and P_{unif}, defined as

$$||P_i - P_{unif}||_1 = \sum_{\mathbf{R}\in\{0,1\}^{K\times T}} |P_i(\mathbf{R}) - P_{unif}(\mathbf{R})|$$

the last equality in Eq. 10 holds because $P_i[\mathbf{R}] \geq P_{unif}[\mathbf{R}]$ whenever the realized utility values in $\mathbf{R} \in \{0, 1\}^{T\times K}$ are 1 for the optimal arm i —since in this case these values are sampled with probability $1/2 + \epsilon$ for arm i and with $1/2$ for all other arms.

By the Pinsker's inequality it holds

$$||P_{unif} - P_i||_1 \leq \sqrt{\frac{1}{2}KL(P_{unif}||P_i)}, \tag{11}$$

where $KL(P_{unif}||P_i)$ is KL divergence between P_i and P_{unif}. By the chain rule of total entropy we have

$$KL(P_{unif}||P_i) = \sum_{t=1}^{T} KL(P_{unif}(\mathbf{r_t})||P_i(\mathbf{r_t})), \tag{12}$$

where $\mathbf{r_t} \in \{0, 1\}^K$ is a K column vector of the utility matrix $\mathbf{R}$ at column t, and we use that $\mathbf{r_t}$ is independent of $\mathbf{r_{t-1}}$ for $t \geq 2$. We will now use the expression of KL divergence for multivariate

Bernoulli distributions $\mathbf{p}, \mathbf{q}$ with K variables which is as follows

$$KL(\mathbf{p}||\mathbf{q}) = \sum_{j=1}^{K} p_j \log \frac{p_j}{q_j} + (1-p_j) \log \frac{1-p_j}{1-q_j},$$

where p_j, q_j is the j-th coefficient of $\mathbf{p}, \mathbf{q}$ respectively. Since $P_i(\mathbf{r_t}), P_{unif}(\mathbf{r_t})$ are multivariate Bernoulli distributions for any utility vector $\mathbf{r}_t = \mathbf{r}$ it holds

$$\begin{aligned}
KL(P_{unif}(\mathbf{r})||P_i(\mathbf{r})) &= \sum_{j=1}^{K} P_{unif}(r_j) \log \frac{P_{unif}(r_j)}{P_i(r_j)} + (1 - P_{unif}(r_j)) \log \frac{1 - P_{unif}(r_j)}{1 - P_i(r_j)} \\
&= P_{unif}(r_i) \log \frac{P_{unif}(r_i)}{P_i(r_i)} + (1 - P_{unif}(r_i) \log \frac{1 - P_{unif}(r_i)}{1 - P_i(r_i)} \\
&= \frac{1}{2} \log \frac{1/2}{1/2 + \epsilon} + \frac{1}{2} \log \frac{1/2}{1/2 - \epsilon} \\
&= \frac{1}{2} \log (\frac{1}{1+2\epsilon} \cdot \frac{1}{1-2\epsilon}) \\
&= \frac{1}{2} \log (\frac{1}{1+2\epsilon} \cdot \frac{1}{1-2\epsilon}) \\
&= \frac{1}{2} \log \frac{1}{1-4\epsilon^2} \\
&= -\frac{1}{2} \log (1 - 4\epsilon^2),
\end{aligned}$$

where we use that for any arm $j \neq i, P_{unif}(r_j) = P_i(r_j) = 1/2$. Therefore Eq. 12 becomes

$$KL(P_{unif}||P_i) = \sum_{t=1}^{T} KL(P_{unif}(\mathbf{r_t})||P_i(\mathbf{r_t})) = -\frac{T}{2} \log(1 - 4\epsilon^2) \tag{13}$$

By Eq. 10, 11, and 13 we have

$$\mathbb{E}_i[f(\mathbf{R})] - \mathbb{E}_{unif}[f(\mathbf{R})] \leq \frac{M}{2} \sqrt{T \cdot \frac{-\log(1 - 4\epsilon^2)}{4}}$$

□

Note that the above bound no longer depends on the expected number of times that the algorithm chose to play the arm i as in Auer et al. [2002].

We will use the above result to prove Corollary 7 below, in the context of the construction used by Agarwal et al. [2012], which we restate for completeness: We assume M different contexts, a policy class Π comprising all possible K^M policies over the contexts, *i.e.*, $M = \log(|\Pi|)/\log(K)$, and the utility distributions as in the above lemma for each context z. Fix a context z, let T_z be the set of time steps such that $z_t = z$, and let i be the optimal arm for this context under the utility distribution P_i chosen uniformly at random before the first time step. Then by Lemma 1, for the number of time steps in T_z such that any algorithm chooses $a_t = a$ we can show:

Corollary 7. *Given context $z_t = z$ for $t = 1, \ldots, T$*

$$\mathbb{E}_a[S_a] \leq \mathbb{E}_{unif}[S_a] + |T_z|^{\frac{3}{2}} \cdot \sqrt{2}\epsilon.$$

Proof. We use Lemma 1 for $f(\mathbf{R}) = S_a \in [0, |T_z|]$ (because given the context z, $S_a \leq |T_z|$), horizon $T = |T_z|$, therefore we have

$$\mathbb{E}_a[S_a] - \mathbb{E}_{unif}[S_a] \leq \frac{|T_z|}{2} \sqrt{|T_z|} \cdot \frac{-\log(1 - 4\epsilon^2)}{4} \leq |T_z|^{\frac{3}{2}} \sqrt{\frac{-\log(1 - 4\epsilon^2)}{16}}.$$

We will now use that $(-\log(1-4\epsilon^2))/16 \leq 2\epsilon^2$, which we can rewrite as $0 \leq 8w - \log(1-w)$, for $w = 4\epsilon^2$, $w \in (0,1)$. Let $g(w) = 8w - \log(1-w)$. Since, $g'(w) = 8 + \frac{1}{1-w} > 0$ for $w \in (0,1)$, g is increasing for $w \in (0,1)$, and thus $g(w) \geq \lim_{w \to 0} g(w) = \lim_{w \to 0} 8w - \log(1-w) = 0$. □

By using Corollary 7 instead of Corollary 5.1 in the rest of the proof by Agarwal et al. [2012] we have that the expected regret

$$\mathbb{E}[R(T)] = \Omega\left(M \cdot \epsilon - \frac{\epsilon^2}{\sqrt{M}} \cdot T^{3/2}\right),$$

which for $\epsilon = \Theta(\sqrt{M/T})$ becomes

$$\mathbb{E}[R(T)] = \Omega\left(\sqrt{T \cdot M}\right) = \Omega\left(\sqrt{T \cdot \frac{\log(|\Pi|)}{\log K}}\right).$$

Since $|\Pi| = K^{|\mathcal{H}|\cdot|\mathcal{B}|}$ we have

$$\mathbb{E}[R(T)] = \Omega\left(\sqrt{T \cdot |\mathcal{H}| \cdot |\mathcal{B}|}\right)$$

A.2 Proof of Theorem 2

Let $h \in \mathcal{H}$. Due to perfect alignment, for any b, b' it holds that

$$b \leq b' \iff P(Y = 0 \,|\, H = h, B = b) \geq P(Y = 0 \,|\, H = h, B = b')$$

Using this and by definition of $b^*(h)$, we have that

$$b \leq b^*(h) \iff P(Y = 0 \,|\, H = h, B = b) \geq \frac{u(1,1) - u(0,1)}{u(1,1) - u(1,0) + u(0,0) - u(0,1)}.$$

By Lemma 2, stated and proven below, it must hold that

$$b \leq b^*(h) \iff \pi^*(h, b) = 0.$$

Lemma 2. *For any utility function satisfying Eq. 1 and any fixed confidence values h, b, let $\pi^*(h, b)$ denote the optimal decision in this context,* i.e.*, $\pi^*(h, b) := \arg\max_{a \in \{0,1\}} \mu(a \,|\, h, b)$. The following are equivalent statements*

a) $\pi^*(h, b) = 0$

b) $P(Y = 0 \,|\, H = h, B = b) \geq \frac{u(1,1) - u(0,1)}{u(1,1) - u(1,0) + u(0,0) - u(0,1)}$

Proof. It holds that

$$\begin{aligned}
\mu(1 \,|\, h, b) &= E_Y\left[u(1, Y) | H = h, B = b\right] \\
&= P(Y = 1 | H = h, B = b) \cdot u(1,1) + P(Y = 0 | H = h, B = b) \cdot u(1,0) \\
&= (1 - P(Y = 0 | H = h, B = b)) \cdot u(1,1) + P(Y = 0 | H = h, B = b) \cdot u(1,0) \\
&= P(Y = 0 | H = h, B = b) \cdot (u(1,0) - u(1,1)) + u(1,1), \qquad (14)
\end{aligned}$$

since $Y \in \{0, 1\}$ and thus $P(Y = 1 | H = h, B = b) = 1 - P(Y = 0 | H = h, B = b)$. Similarly, we show:

$$\begin{aligned}
\mu(0 \,|\, h, b) &= E_Y\left[u(0, Y) | H = h, B = b\right] \\
&= P(Y = 1 | H = h, B = b) \cdot u(0,1) + P(Y = 0 | H = h, B = b) \cdot u(0,0) \\
&= (1 - P(Y = 0 | H = h, B = b)) \cdot u(0,1) + P(Y = 0 | H = h, B = b) \cdot u(0,0) \\
&= P(Y = 0 | H = h, B = b) \cdot (u(0,0) - u(0,1)) + u(0,1). \qquad (15)
\end{aligned}$$

By definition a decision a' is optimal iff $a' = \arg\max_{a \in \{0,1\}} \mu(a \,|\, h, b)$. Thus, $\pi^*(h, b) = 0$ iff $\mu(1 \,|\, h, b) \leq \mu(0 \,|\, h, b)$. We can rewrite this inequality using Eq. 14 and Eq. 15 as

$$\begin{aligned}
&P(Y = 0 | H = h, B = b) \cdot (u(1,0) - u(1,1)) + u(1,1) \\
&\qquad \leq P(Y = 0 | H = h, B = b) \cdot (u(0,0) - u(0,1)) + u(0,1).
\end{aligned}$$

Since $u(0,0) - u(1,0) > 0$ and $u(1,1) - u(0,1) > 0$, by rearranging the terms we get

$$P(Y = 0 \,|\, H = h, B = b) \geq \frac{u(1,1) - u(0,1)}{u(1,1) - u(1,0) + u(0,0) - u(0,1)}.$$

Thus,

$$\begin{aligned}
\pi^*(h, b) = 0 &\iff \mu(1 \,|\, h, b) \leq \mu(0 \,|\, h, b) \\
&\iff P(Y = 0 \,|\, H = h, B = b) \geq \frac{u(1,1) - u(0,1)}{u(1,1) - u(1,0) + u(0,0) - u(0,1)}
\end{aligned}$$

□

A.3 Proof of Theorem 3

For any $h \in \mathcal{H}$, consider any $b \in [0,1], b \neq b^*(h)$, we distinguish two cases. If $b = b^*(h) + \epsilon$ for some $\epsilon > 0$, then

$$\begin{aligned}
\mu(b \mid h) &= \mathbb{E}[\mu(1 \mid H, B) \cdot \mathbb{I}[B > b] \mid H = h] \\
&\quad + \mathbb{E}[\mu(0 \mid H, B) \cdot \mathbb{I}[b \geq B > b^*(H)] \mid H = h] + \mathbb{E}[\mu(0 \mid H, B) \cdot \mathbb{I}[B \leq b^*(H)] \mid H = h] \\
&< \mathbb{E}[\mu(1 \mid H, B) \cdot \mathbb{I}[B > b] \mid H = h] + \mathbb{E}[\mu(1 \mid H, B) \cdot \mathbb{I}[b \geq B > b^*(H)] \mid H = h] \\
&\quad + \mathbb{E}[\mu(0 \mid H, B) \cdot \mathbb{I}[B \leq b^*(H)] \mid H = h] \qquad (16) \\
&= \mu(b^*(h) \mid h)
\end{aligned}$$

where Eq. 16 follows from Theorem 2 since $\mu(0 \mid h, b') < \mu(1 \mid h, b')$ for all $b' > b^*(h)$. On the other hand, if $b = b^* - \epsilon$ for some $\epsilon > 0$, then

$$\begin{aligned}
\mu(b \mid h) &= \mathbb{E}[\mu(1 \mid H, B) \cdot \mathbb{I}[B > b^*(H)] \mid H = h] \\
&\quad + \mathbb{E}[\mu(1 \mid H, B) \cdot \mathbb{I}[b^*(H) \geq B > b] \mid H = h] + \mathbb{E}[\mu(0 \mid H, B) \cdot \mathbb{I}[B \leq b] \mid H = h] \\
&\leq \mathbb{E}[\mu(1 \mid H, B) \cdot \mathbb{I}[B > b^*(H)] \mid H = h] + \mathbb{E}[\mu(0 \mid H, B) \cdot \mathbb{I}[b^*(H) \geq B > b] \mid H = h] \\
&\quad + \mathbb{E}[\mu(0 \mid H, B) \cdot \mathbb{I}[B \leq b] \mid H = h] \qquad (17) \\
&= \mu(b^*(h) \mid h)
\end{aligned}$$

where Eq. 17 follows from Theorem 2 since $\mu(1 \mid h, b') \leq \mu(0 \mid h, b')$ for all $b' < b^*(h)$. In both cases, $\mu(b \mid h) \leq \mu(b^*(h) \mid h)$, thus $b^*(h) = \arg\max_{b \in \mathcal{B}} \mu(b \mid h)$.

A.4 Proof of Theorem 4

We first state the well-known DKW inequality (or DKWM inequality) and its complimentary corollary, which we use to bound the error of the conditional probability estimators in Lemma 3 used to estimate the conditional expected utility $\mu(b \mid h)$ in Algorithm 1. Then, we show that under perfect alignment, our algorithm's expected regret on the $|\mathcal{H}|$ independent instances of the multiclass online learning problem is equivalent to its expected regret on the binary online contextual learning problem with full feedback

Theorem 8. *(DKW inequality Dvoretzky et al. [1956], Massart [1990]) Given a natural number n, let $Z_1, Z_2, \ldots, Z_n$ be real-valued independent and identically distributed random variables with cumulative distribution function $F(\cdot)$. Let F_n denote the associated empirical distribution function defined by*

$$F_n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[Z_i \leq z] \quad \textit{for} \quad z \in \mathbb{R}$$

Then,

$$P\left(\sup_{z \in \mathcal{Z}} \; |F_n(z) - F(z)| > \epsilon \right) \leq 2 \exp^{-2n\epsilon^2}$$

Corollary 9. *Given a natural number n, let $Z_1, Z_2, \ldots, Z_n$ be real-valued independent and identically distributed random variables with cumulative distribution function $F(\cdot)$. Let $F^+(\cdot) = P(Z > z)$ and F_n^+ denote the associated empirical distribution function defined by*

$$F_n^+(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[Z_i > z] \quad \textit{for} \quad z \in \mathbb{R}$$

Then,

$$P\left(\sup_{z \in \mathcal{Z}} \; |F_n^+(z) - F^+(z)| > \epsilon \right) \leq 2 \exp^{-2n\epsilon^2}$$

Proof. First, note that,

$$\begin{aligned}
|F_n^+(z) - F^+(z)| = |F_n^+(z) + 1 - 1 - F^+(z)| &= |-(1 - F_n^+(z)) + 1 - F^+(z)| \\
&= |-F_n(z) + F(z)| = |F_n(z) - F(z)| \qquad (18)
\end{aligned}$$

Substituting the above in Theorem 8 (DKW inequality), we get that

$$P\left(\sup_{z\in\mathcal{Z}} \ |F_n^+(z) - F^+(z)| > \epsilon\right) = P\left(\sup_{z\in\mathcal{Z}} \ |F_n(z) - F(z)| > \epsilon\right) \leq 2\exp^{-2n\epsilon^2}$$

□

Lemma 3. *For any fixed h, let $(B_i, Y_i)_{i\in[n]}$ be independent and identically distributed random variables with $B_i, Y_i \sim P(B, Y \mid H = h)$. Consider the following estimators for any fixed $b \in \mathcal{B}$*

$$P_n(B \leq b \mid H = h) = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[B_i \leq b\right]$$

$$P_n(Y = 0, B \leq b \mid H = h) = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[B_i \leq b\right](1 - Y_i)$$

$$P_n(Y = 0, B > b \mid H = h) = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[B_i > b\right](1 - Y_i).$$

Then, for any $\alpha \in (0, 1)$, it holds that

$$P\left(\sup_{b\in\mathcal{B}} |P(B \leq b \mid H = h) - P_n(B \leq b \mid H = h)| \leq \sqrt{\log(2/\alpha)/(2n)}\right) \geq 1 - \alpha, \quad (19)$$

$$P\left(\sup_{b\in\mathcal{B}} |P(Y = 0, B \leq b \mid H = h) - P_n(Y = 0, B \leq b \mid H = h)| \leq \sqrt{\log(2/\alpha)/(2n)}\right) \geq 1 - \alpha, \quad (20)$$

and

$$P\left(\sup_{b\in\mathcal{B}} |P(Y = 0, B > b \mid H = h) - P_n(Y = 0, B > b \mid H = h)| \leq \sqrt{\log(2/\alpha)/(2n)}\right) \geq 1 - \alpha \quad (21)$$

Proof. Consider any fixed $h \in \mathcal{H}$. Eq. 19 follows directly from Theorem 8 by setting $Z_i = B_i, \forall i \leq [n]$, and $\epsilon = \sqrt{\log(2/\alpha)/(2n)}$.

To show Eq. 20, consider the following random variables

$$Z_i = B_i + Y_i \qquad \forall i \leq [n].$$

Let $F(\cdot)$ denote the cumulative distribution function $P(Z \leq z) = P(B + Y \leq z \mid H = h)$ and

$$F_n(z) = \frac{1}{n}\sum_{i:h_i=h} \mathbb{I}[B_i + Y_i \leq z] \quad \forall z \in \mathbb{R}.$$

Applying Theorem 8, we have that, for any $\alpha \in (0, 1)$, with probability $1 - \alpha$

$$\sup_{z\in\mathbb{R}} |F(z) - F_n(z)| \leq \sqrt{\log(2/\alpha)/(2n_h)} \quad (22)$$

Further, note that for any $z \in [0, 1]$ the following are equivalent statements

$$B + Y \leq z \iff Y = 0 \text{ and } B \leq z \iff \mathbb{I}[B \leq z](1 - Y) = 1$$

because if $Y = 0$ and $B \leq z$ then it implies that all statements are true (direction $\Longleftarrow \Longrightarrow$) and if either $Y = 1$ or $B > z$ then it implies that all statements are false (direction $\Longrightarrow \Longleftarrow$). Hence, we have that, for $z \in [0, 1]$, it holds that

$$F(z) = P(B + Y \leq z \mid H = h) = P(Y = 0, B \leq z \mid H = h)$$

and similarly

$$F_n(z) = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}[B_i + Y_i \leq z] = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[B_i \leq z\right](1 - Y_i) = P_n(Y = 0, B \leq z \mid H = h).$$

Thus, because of Eq. 22, Eq. 20 must hold for all $b \in [0, 1]$.

Analogously, to show Eq. 21, let

$$Z_i = B_i(1 - Y_i) \qquad \forall i \leq [n]$$

let $F^+(\cdot) = P(Z > z) = P(B(1-Y) > z \,|\, H = h)$ and

$$F_n^+(z) = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}[B_i(1 - Y_i) > z] \quad \forall z \in \mathbb{R}.$$

Applying Corollary 9, we have that for any $\alpha \in (0, 1)$, with probability $1 - \alpha$

$$\sup_{z \in \mathbb{R}} |F^+(z) - F_n^+(z)| \leq \sqrt{\log(2/\alpha)/(2n_h)} \tag{23}$$

Similarly, note that, for any $z \in [0, 1]$ the following are equivalent statements

$$B(1-Y) > z \iff Y = 0 \text{ and } B > z \iff \mathbb{I}[B > z](1 - Y) = 1.$$

Hence, we have that

$$F^+(z) = P(B(1-Y) > z \,|\, H = h) = P(Y = 0, B > z \,|\, H = h)$$

and

$$F_n^+(z) = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}[B_i(1 - Y_i) > z] = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\left[B_i > z\right](1 - Y_i) = P_n(Y = 0, B > z \,|\, H = h).$$

Thus, because of Eq. 23, Eq. 21 must hold for all $b \in [0, 1]$. □

Lemma 4. *Under perfect alignment, consider a selection algorithm that for confidence values H_t, B_t picks threshold $\bar{B}_t(H_t) \in \mathcal{B}$ and decides $A_t \in \{0, 1\}$ according to this threshold,* i.e.*, $A_t = \mathbb{I}[B_t > \bar{B}_t(H_t)]$. Let $P(\bar{B}_t(H_t) \,|\, H_t)$ be the distribution over thresholds and $P(A_t \,|\, H_t, B_t)$ be the distribution over decisions induced by this selection algorithm in time step t and distributions $P(Y \,|\, H, B)$ and $P(B \,|\, H)$ (these distributions are relevant as they influence the history $(H_i, B_i, Y_i)_{i<t}$ that the algorithm observes). Then, we have that*

$$\begin{aligned}
\mathbb{E}[R(T)] &= \mathbb{E}_{h_t, b_t \sim P(H,B), a_t \sim P(A_t \,|\, H_t, B_t)} \left[\sum_{t=1}^{T} \mu(\pi^*(h_t, b_t) \,|\, h_t, b_t) - \mu(a_t \,|\, h_t, b_t) \right] \\
&= \mathbb{E}_{h_t \sim P(H), \bar{b}_t(h_t) \sim P(\bar{B}_t(H_t) \,|\, H_t = h_t)} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \,|\, h_t) - \mu(\bar{b}_t(h_t) \,|\, h_t) \right]. \qquad (24)
\end{aligned}$$

Proof. By the law of total expectation

$$\begin{aligned}
\mathbb{E}[R(T)] &= \mathbb{E}_{h_t, b_t \sim P(H,B), a_t \sim P(A_t \,|\, H_t, B_t)} \left[\sum_{t=1}^{T} \mu(\pi^*(h_t, b_t) \,|\, h_t, b_t) - \mu(a_t \,|\, h_t, b_t) \right] \\
&= \mathbb{E}_{h_t \sim P(H)} \left[\sum_{t=1}^{T} \mathbb{E}_{b_t \sim P(B \,|\, H = h_t), a_t \sim P(A \,|\, H = h_t, B = b_t)} \left[\mu(\pi^*(h_t, b_t) \,|\, h_t, b_t) - \mu(a_t \,|\, h_t, b_t) \right] \right]
\end{aligned}$$

For a fixed $H = h_t$ and any fixed threshold $\bar{b}$, the choice of decision A_t is defined deterministically by the fixed threshold $\bar{b}$. Thus, we have that

$$\begin{aligned}
&\mathbb{E}_{b_t \sim P(B \,|\, H = h_t)} \left[\mu(A_t \,|\, H_t = h_t, B_t = b_t) \,|\, \bar{B}_t = \bar{b} \right] \\
&\qquad = \mathbb{E}_{b_t \sim P(B \,|\, H = h_t)} \left[\mathbb{E}[\mu(1 \,|\, h_t, b_t) \cdot \mathbb{I}[b_t > \bar{b}]] + \mathbb{E}[\mu(0 \,|\, h_t, b_t) \cdot \mathbb{I}[b_t \leq \bar{b}]] \right] \\
&\qquad = \mu(\bar{b} \,|\, h_t)
\end{aligned}$$

Since, at each timestep t, $\bar{b}_t(h_t)$ is chosen by the selection algorithm given h_t under distribution $P(\bar{B}_t(H_t) \,|\, H_t = h_t)$, we can substitute the induced distribution $P(A_t \,|\, H_t = h_t, B_t = b_t)$ for decision A_t by this distribution for threshold $\bar{B}_t(H_t)$ and obtain

$$\begin{aligned}
&\mathbb{E}_{b_t \sim P(B \,|\, H = h_t), a_t \sim P(A_t \,|\, H_t = h_t, B_t = b_t)} \left[\mu(a_t \,|\, h_t, b_t) \right] \\
&\quad = \mathbb{E}_{\bar{b}_t \sim P(\bar{B}_t(H_t) \,|\, H_t = h_t)} \left[\mathbb{E}_{b_t \sim P(B \,|\, H = h_t)} \left[\mu(A_t \,|\, h_t, b_t) \,|\, \bar{B}_t(H_t) = \bar{b}_t(h_t) \right] \right] \\
&\quad = \mathbb{E}_{\bar{b}_t \sim P(\bar{B}_t(H_t) \,|\, H_t = h_t)} \left[\mu(\bar{b}_t(h_t) \,|\, h_t) \right]
\end{aligned}$$

Since the optimal policy π^* correspond to a threshold function with fixed optimal thresholds $b^*(h)$ for each $h \in \mathcal{H}$, we get

$$\mathbb{E}_{b_t \sim P(B \mid H=h_t)} \left[\mu(\pi(h_t, b_t) \mid h_t, b_t)\right] \\ = \mathbb{E}_{b_t \sim P(B \mid H=h_t)} \left[\mu(A^* \mid h_t, b_t) \mid B^*(H_t) = b^*(h_t)\right] = \mu(b^*(h_t) \mid h_t)$$

where optimal decision A^* given confidence H_t depends on optimal threshold $B^*(H_t)$ with $A^* = \mathbb{I}[b_t > B^*(H_t)]$. Thus, we have that

$$\mathbb{E}[R(T)] = \mathbb{E}_{h_t \sim P(H)} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \mid h_t) - \mathbb{E}_{\bar{b}_t \sim P(\bar{B}_t \mid H_t)}[\mu(\bar{b}_t(h_t) \mid h_t)]\right] \\ = \mathbb{E}_{h_t \sim P(H), \bar{b}_t \sim P(\bar{B}_t(H_t) \mid H_t = h_t)} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \mid h_t) - \mu(\bar{b}_t(h_t) \mid h_t)\right]$$

□

With the above auxiliary Lemmas, we are ready to prove Theorem 4.

By Lemma 4 we have that

$$\mathbb{E}[R(T)] = \mathbb{E}_{h_t \sim P(H), \bar{b}_t \sim P(\bar{B}_t(H_t) \mid H_t = h_t)} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \mid h_t) - \mu(\bar{b}_t(h_t) \mid h_t)\right],$$

where $\bar{b}_t(h_t)$ is the value of the threshold selected by Algorithm 1 at time step t and $P(\bar{B}_t(H_t) \mid H_t = h_t)$ is the corresponding distribution. In what follows, we will show that selecting $\bar{b}_t(h_t)$ using Algorithm 1 achieves expected regret $\mathbb{E}(R(T)) = \mathcal{O}\left(\sqrt{|\mathcal{H}| T \log T}\right)$.

Let us first define an estimator for the conditional expected utility $\mu(b \mid h)$ achieved by the decision policy π defined by a threshold function with threshold b given the human confidence $H = h$, where as a reminder

$$\mu(b \mid h) = \mathbb{E}[\mu(1 \mid h, B) \cdot \mathbb{I}[B > b] \mid H = h] + \mathbb{E}[\mu(0 \mid h, B) \cdot \mathbb{I}[B \leq b] \mid H = h].$$

By Eq. 14 and Eq. 15 the above becomes

$$\begin{aligned} \mu(b \mid h) &= \mathbb{E}\left[(P(Y = 0 \mid H = h, B)(u(1,0) - u(1,1)) + u(1,1)) \cdot \mathbb{I}[B > b] \mid H = h\right] \\ &+ \mathbb{E}\left[(P(Y = 0 \mid H = h, B)(u(0,0) - u(0,1)) + u(0,1)) \cdot \mathbb{I}[B \leq b] \mid H = h\right] \\ &= P(Y = 0, B > b \mid H = h)(u(1,0) - u(1,1)) \\ &+ P(Y = 0, B \leq b \mid H = h)(u(0,0) - u(0,1)) \\ &+ P(B \leq b \mid H = h)(u(0,1) - u(1,1)) + u(1,1). \end{aligned}$$

We can now define the estimator $\bar{\mu}_t(b \mid h)$ at time step t as

$$\begin{aligned} \bar{\mu}_t(b \mid h) &= P_{n_t}(Y = 0, B > b \mid H = h) \cdot (u(1,0) - u(1,1)) \\ &+ P_{n_t}(Y = 0, B \leq b \mid H = h) \cdot (u(0,0) - u(0,1)) \\ &+ P_{n_t}(B \leq b \mid H = h) \cdot (u(0,1) - u(1,1)) + u(1,1), \end{aligned} \tag{25}$$

where, as in Lemma 3,

$$P_{n_t}(Y = 0, B > b \mid H = h) = \frac{1}{n_t(h)} \sum_{i \leq t : h_i = h} \mathbb{I}\left[b_i > b\right](1 - y_i),$$

$$P_{n_t}(Y = 0, B \leq b \mid H = h) = \frac{1}{n_t(h)} \sum_{i \leq t : h_i = h} \mathbb{I}\left[b_i \leq b\right](1 - y_i),$$

$$P_{n_t}(B \leq b \mid H = h) = \frac{1}{n_t(h)} \sum_{i \leq t : h_i = h} \mathbb{I}\left[b_i \leq b\right],$$

with $n_t(h) = \sum_{i=1}^{t} \mathbb{I}\left[h_i = h\right]$.

In what follows, we will first express the regret bound in terms of the regret under the clean and the bad event for the above estimator. We define the clean event as

$$\mathcal{E} = \{|\bar{\mu}_t(b \,|\, h) - \mu(b \,|\, h)| \leq \texttt{rad}_t(h),\ \forall b \in \mathcal{B}, \forall h \in \mathcal{H}, 1 \leq t \leq T\}\,,$$

where $\texttt{rad}_t(h) = 3\sqrt{\frac{\log(6|\mathcal{H}|T^3)}{2n_t(h)}}$. By Lemma 4 it holds

$$\begin{aligned}
\mathbb{E}(R(T)) &= P(\mathcal{E}) \cdot \mathbb{E}_{h_t, \bar{b}_t(h_t) \sim P(H, \bar{B}_t(H_t))} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \,|\, h_t) - \mu(\bar{b}_t(h_t) \,|\, h_t) \,|\, \mathcal{E} \right] \\
&+ P(\mathcal{E}^c) \cdot \mathbb{E}_{h_t, \bar{b}_t(h_t) \sim P(H, \bar{B}_t(H_t))} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \,|\, h_t) - \mu(\bar{b}_t(h_t) \,|\, h_t) \,|\, \mathcal{E}^c \right], \\
&\leq \mathbb{E}_{h_t, \bar{b}_t(h_t) \sim P(H, \bar{B}_t(H_t))} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \,|\, h_t) - \mu(\bar{b}_t(h_t) \,|\, h_t) \,|\, \mathcal{E} \right] + P(\mathcal{E}^c) \cdot T \qquad (26)
\end{aligned}$$

where $\mathcal{E}^c$ is the complementary of the clean event $\mathcal{E}$ (and we have written the joint distribution $P(H, \bar{B}_t(H_t))$ for brevity). By Lemma 3 for $\alpha = \frac{1}{3|\mathcal{H}|T^3}$, the union bound over the three estimators, all $h \in \mathcal{H}$ and all $t \in \{1, .., T\}$, and given $u(\cdot) \in [0, 1]$, we have that $P(\mathcal{E}) \geq 1 - \frac{3|\mathcal{H}|T}{3|\mathcal{H}|T^3} = 1 - \frac{1}{T^2}$ and consequently $P(\mathcal{E}^c) \leq \frac{1}{T^2}$. We can now rewrite Eq. 26 as

$$\mathbb{E}(R(T)) \leq \mathbb{E}_{h_t, \bar{b}_t(h_t) \sim P(H, \bar{B}_t(H_t))} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \,|\, h_t) - \mu(\bar{b}_t(h_t) \,|\, h_t) \,|\, \mathcal{E} \right] + \frac{1}{T}. \qquad (27)$$

We proceed by deriving an upper bound for

$$\mathbb{E}_{h_t, \bar{b}_t(h_t) \sim P(H, \bar{B}_t(H_t))} \left[\sum_{t=1}^{T} \mu(b^*(h_t) \,|\, h_t) - \mu(\bar{b}_t(h_t) \,|\, h_t) \,|\, \mathcal{E} \right].$$

Let $\bar{b}_t(h_t) = \text{argmax}_{b \in \mathcal{B}} \bar{\mu}_t(b \,|\, h_t)$, *i.e.*, the arm pulled by Algorithm 1 in time step t. For $\bar{b}_t(h_t)$ at time step t, assuming that the clean event holds, we have:

$$\bar{\mu}_t(\bar{b}_t(h_t) \,|\, h_t) \geq \bar{\mu}_t(b^*(h_t) \,|\, h_t). \qquad (28)$$

Under the clean event it holds that $\mu(b \,|\, h_t) + \texttt{rad}_t(h_t) \geq \bar{\mu}_t(b \,|\, h_t)$ and that $\bar{\mu}_t(b \,|\, h_t) \geq \mu(b \,|\, h_t) - \texttt{rad}_t(h_t)$ for any arm b. Therefore, given Eq. 28 we have

$$\mu(\bar{b}_t(h_t) \,|\, h_t) + \texttt{rad}_t(h_t) \geq \bar{\mu}_t(\bar{b}_t(h_t) \,|\, h_t) \geq \bar{\mu}_t(b^*(h_t) \,|\, h_t) \geq \mu(b^*(h_t) \,|\, h_t) - \texttt{rad}_t(h_t),$$

and thus

$$\mu(b^*(h_t) \,|\, h_t) - \mu(\bar{b}_t(h_t) \,|\, h_t) \leq 2 \cdot \texttt{rad}(h_t),$$

which is the regret due to pulling a suboptimal arm $\bar{b}_t(h_t)$ at a single time step t given confidence h_t. For the cumulative regret given confidence h up to time step t given the above, we have

$$\begin{aligned}
\sum_{i=1}^{t} \mathbb{I}[h_i = h] \cdot \big(\mu(b^*(h_t) \,|\, h) - \mu(\bar{b}_i(h_t) \,|\, h)\big) &\leq \sum_{i=1}^{t} \mathbb{I}[h_i = h] \cdot 6\sqrt{\frac{\log(6|\mathcal{H}|T^3)}{2n_i(h)}} \qquad (29) \\
&= \sqrt{18\log(6|\mathcal{H}|T^3)} \sum_{i=1}^{n_t(h)} \frac{1}{\sqrt{i}} \\
&= \sqrt{18\log(6|\mathcal{H}|T^3)} \sum_{i=1}^{n_t(h)} \int_{i-1}^{i} \frac{dx}{\sqrt{i}} \qquad (30) \\
&\leq \sqrt{18\log(6|\mathcal{H}|T^3)} \sum_{i=1}^{n_t(h)} \int_{i-1}^{i} \frac{dx}{\sqrt{x}} \\
&= \sqrt{18\log(6|\mathcal{H}|T^3)} \int_{0}^{n_t(h)} \frac{dx}{\sqrt{x}} \\
&= 2\sqrt{18 n_t(h) \log(6|\mathcal{H}|T^3)}. \qquad (31)
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(H_t))}\left[\sum_{t=1}^{T}\mu(b^*(h_t)\,|\,h_t)-\mu(\bar{b}_t(h_t)\,|\,h_t)\,|\,\mathcal{E}\right]\\
&\quad=\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(H_t))}\left[\sum_{t=1}^{T}\sum_{h\in\mathcal{H}}\mathbb{I}[h_t=h]\cdot(\mu(b^*(h_t)\,|\,h)-\mu(\bar{b}_t(h_t)\,|\,h))\,|\,\mathcal{E}\right]\\
&\quad=\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(H_t))}\left[\sum_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{I}[h_t=h]\cdot(\mu(b^*(h_t)\,|\,h)-\mu(\bar{b}_t(h_t)\,|\,h))\,|\,\mathcal{E}\right] \qquad (32)
\end{aligned}$$

By Eq. 29, it follows that $\sum_{t=1}^{T}\mathbb{I}[h_t = h]\cdot\big(\mu(b^*(h)\,|\,h)-\mu(\bar{b}_t(h)\,|\,h)\big) \le 2\sqrt{18n_T(h)\log(6|\mathcal{H}|T^3)}$. As a result, we can bound Eq. 32 as follows

$$\begin{aligned}
&\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(h_t))}\left[\sum_{t=1}^{T}\mu(b^*(h_t)\,|\,h_t)-\mu(\bar{b}_t(h_t)\,|\,h_t)\,|\,\mathcal{E}\right] \qquad (33)\\
&\qquad\le\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(H_t))}\left[\sum_{h\in\mathcal{H}}2\sqrt{18n_T(h)\log(6|\mathcal{H}|T^3)}\,|\,\mathcal{E}\right]\\
&\qquad=\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(H_t))}\left[2\sqrt{18\log(6|\mathcal{H}|T^3)}\sum_{h\in\mathcal{H}}\sqrt{n_T(h)}\,|\,\mathcal{E}\right]. \qquad (34)
\end{aligned}$$

By Jensen's inequality it holds

$$\frac{\sum_{h\in\mathcal{H}}\sqrt{n_T(h)}}{|\mathcal{H}|}\le\sqrt{\frac{\sum_{h\in\mathcal{H}}n_T(h)}{|\mathcal{H}|}}$$

and since $\sum_{h\in\mathcal{H}}n_T(h)=T$, we have

$$\sum_{h\in\mathcal{H}}\sqrt{n_T(h)}\le|\mathcal{H}|\sqrt{\frac{T}{|\mathcal{H}|}}=\sqrt{|\mathcal{H}|T}.$$

We can now bound Eq. 33 using the above

$$\begin{aligned}
&\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(H_t))}\left[\sum_{t=1}^{T}\mu(b^*(h_t)\,|\,h_t)-\mu(\bar{b}_t(h_t)\,|\,h_t)\,|\,\mathcal{E}\right]\\
&\qquad\le\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(H_t))}\left[2\sqrt{18|\mathcal{H}|T\log(6|\mathcal{H}|T^3)}\,|\,\mathcal{E}\right]=2\sqrt{18|\mathcal{H}|T\log(6|\mathcal{H}|T^3)}.
\end{aligned}$$

As long as $|\mathcal{H}|<T$, it holds that under the clean event $\mathcal{E}$

$$\mathbb{E}_{h_t,\bar{b}_t\sim P(H,\bar{B})}\left[\sum_{t=1}^{T}\mu(b^*\,|\,h_t)-\mu(\bar{b}_t\,|\,h_t)\,|\,\mathcal{E}\right]=\mathcal{O}\left(\sqrt{|\mathcal{H}|T\log T}\right)$$

Finally, using the above and Eq. 27 we have

$$\mathbb{E}_{h_t,\bar{b}_t(h_t)\sim P(H,\bar{B}_t(H_t))}\left[\sum_{t=1}^{T}\mu(b^*(h_t)\,|\,h_t)-\mu(\bar{b}_t(h_t)\,|\,h_t)\right]=\mathcal{O}\left(\sqrt{|\mathcal{H}|T\log T}\right)$$

A.5 Proof of Theorem 5

For any function $\ell\in\mathcal{L}$, we have that by the law of total expectation

$$\begin{aligned}
\mu(\ell)&=\mathbb{E}[\mu(1\,|\,H,B)\cdot\mathbb{I}[B>\ell(H)]]+\mathbb{E}[\mu(0\,|\,H,B)\cdot\mathbb{I}[B\le\ell(H)]]\\
&=\mathbb{E}_H[\mathbb{E}[\mu(1\,|\,H,B)\cdot\mathbb{I}[B>\ell(H)]\,|\,H]+\mathbb{E}[\mu(0\,|\,H,B)\cdot\mathbb{I}[B\le\ell(H)]\,|\,H]]\\
&=\mathbb{E}_H[\mu(\ell(H)\,|\,H)]
\end{aligned}$$

From Theorem 3, we know that for any $h\in\mathcal{H}$, $b^*(h)=\arg\max_{b\in\mathcal{B}}\mu(b\,|\,h)$. Thus, $\ell(h)=b^*(h)$ for all $h\in\mathcal{H}$ maximizes the value of $\mathbb{E}_H[\mu(\ell(H)\,|\,H)]$.

A.6 Proof of Theorem 6

We proceed the proof analogously to the proof of Theorem 4. Here, we use a non-trivial extension of the DKW inequality DKW-like inequality and its complimentary corollary, stated and proven in Section A.7, to bound the error of the conditional probability estimators in Lemma 5—these estimators estimate the expected utility $\mu(\ell)$ as defined in Eq. 7. We use this to show that in each time step t with high probability the utility of the selected function (arm) $\bar{\ell}_t$ is within certain bounds of the utility of the optimal function ℓ^*. Then, we show that, under perfect alignment, the expected regret of the binary online contextual learning problem with full feedback is equivalent to the expected regret of the multiclass online contextual learning problem over functions (arms) $\ell \in \mathcal{L} := \{\ell(\cdot) | \ell : \mathcal{H} \to \mathcal{B}\}$ where decisions $A = \{0, 1\}$ are defined by a threshold function $A = \mathbb{I}[B_t > \ell(H_t)]$ with $H_t, B_t \sim P(H, B)$.

More formally, we define the class of policies induced by functions in $\mathcal{L}$ as follows

Definition 10. *Given a fixed $\vec{b}$ in vector form, $\vec{b} = (\vec{b}_h)_{h \in \mathcal{H}} \in \mathcal{B}^{\mathcal{H}}$, let $\Delta_{\vec{b}}$ denote the following policy for $h, b \in \mathcal{H} \times \mathcal{B}$*

$$\Delta_{\vec{b}}(h, b) = \mathbb{I}[b \leq \vec{b}_h].$$

Let $\mathcal{L} := \{\ell(\cdot) | \ell : \mathcal{H} \to \mathcal{B}\}$. We define the class $\mathcal{D}_{\mathcal{L}}$ of policies induced by $\mathcal{L}$ as

$$\mathcal{D}_{\mathcal{L}} := \{\Delta_{\vec{b}} \,|\, \ell \in \mathcal{L} : \vec{b}_h = \ell(h) \text{ for all } h \in \mathcal{H}\} = \{\Delta_{\vec{b}} \,|\, \vec{b} \in \mathcal{B}^{\mathcal{H}}\}$$

Lemma 5. *Let $(H_i, B_i, Y_i)_{i \in [n]}$ be independent and identically distributed random variables with $H_i, B_i, Y_i \sim P(H, B, Y)$. Let $\mathcal{B}$ be countable (and recall $\mathcal{H}$ is discrete) and let $\mathcal{D}_{\mathcal{L}}$ be the class of (countable) policies as in Definition 10. Consider the following estimators for any fixed $\Delta_{\vec{b}} \in \mathcal{D}$*

$$P_n(\Delta_{\vec{b}}(H, B) = 1) = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\left[B_i \leq \vec{b}_{H_i}\right]$$

$$P_n(Y = 0, \Delta_{\vec{b}}(H, B) = 1) = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\left[B_i \leq \vec{b}_{H_i}\right] (1 - Y_i)$$

$$P_n(Y = 0, \Delta_{\vec{b}}(H, B) = 0) = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\left[B_i > \vec{b}_{H_i}\right] (1 - Y_i).$$

Then, there exist constant $c_3 > 0$ such that, for any $\alpha \in (0, 1)$ with $\log(1/\alpha) > c_3 \sqrt{|\mathcal{H}|}$, it holds that

$$P\left(\sup_{\Delta_{\vec{b}} \in \mathcal{D}_{\mathcal{L}}} |P(\Delta_{\vec{b}}(H, B) = 1) - P_n(\Delta_{\vec{b}}(H, B) = 1)| \leq \sqrt{4c_2 \sqrt{|\mathcal{H}|} \log(1/\alpha)/n}\right) \geq 1 - \alpha, \tag{35}$$

$$P\left(\sup_{\Delta_{\vec{b}} \in \mathcal{D}_{\mathcal{L}}} |P(Y = 0, \Delta_{\vec{b}}(H, B) = 1) - P_n(Y = 0, \Delta_{\vec{b}}(H, B) = 1)| \leq \sqrt{4c_2 \sqrt{|\mathcal{H}|} \log(1/\alpha)/n}\right) \geq 1 - \alpha, \text{ and} \tag{36}$$

$$P\left(\sup_{\Delta_{\vec{b}} \in \mathcal{D}_{\mathcal{L}}} |P(Y = 0, \Delta_{\vec{b}}(H, B) = 0) - P_n(Y = 0, \Delta_{\vec{b}}(H, B)) = 0| \leq \sqrt{4c_2 \sqrt{|\mathcal{H}|} \log(1/\alpha)/n}\right) \geq 1 - \alpha \tag{37}$$

Proof. The proof follows the proof of Theorem 3 closely using Theorem 16 and Corollary 17 instead of the DKW inequality. Accordingly, to use the previous results to prove Eq. 35 to 37, we use that the probability of an indicator random variable obtaining value of 1 is equal the expectation of the random variable, that is, for example

$$P(\Delta_{\vec{b}}(H, B) = 1) = \mathbb{E}_{H,B \sim P(H,B)}[\Delta_{\vec{b}}(H, B)]$$

Eq. 35 follows directly from Theorem 16 by setting $K_i = H_i$ and $X_i = B_i, \forall i \leq [n]$, and $\epsilon = \sqrt{4c_2 \sqrt{|\mathcal{H}|} \log(1/\alpha)/n}$. More formally, because $\Delta_{\vec{b}}(H_i, B_i) = \mathbb{I}\left[B_i \leq b_{H_i}\right]$, Eq. 35 is equivalent

to

$$P\left(\sup_{\Delta_{\vec{b}}\in\mathcal{D}_{\mathcal{L}}}\left|\mathbb{E}_{H,B\sim P(H,B)}[\Delta_{\vec{b}}(H,B)]-\frac{1}{n}\sum_{i\in[n]}\Delta_{\vec{b}}(H_i,B_i)\right|\leq\sqrt{4c_2\sqrt{|\mathcal{H}|}\log(1/\alpha)/n}\right)\geq 1-\alpha.$$

which holds by Theorem 16 when $\epsilon=\sqrt{4c_2\sqrt{|\mathcal{H}|}\log(1/\alpha)/n}>c_1\sqrt{|\mathcal{H}|/n}$, that is, for $\log(1/\alpha)>c_3\sqrt{|\mathcal{H}|}$ where $c_3:=\frac{c_1^2}{4c_2}$.

To show Eq. 36, consider the following random variables $K_i=H_i$ and

$$X_i=B_i+Y_i\qquad\forall i\leq[n].$$

where $\mathcal{X}=\mathcal{B}+\mathcal{Y}$ is countable as $\mathcal{B}$ is countable and $\mathcal{Y}=\{0,1\}$. Consider any policy $\Delta_{\vec{x}}$ in countable class $\mathcal{D}$ as in Definition 15 where $\vec{x}\in\mathcal{X}^{\mathcal{H}}$ and the following estimator

$$\frac{1}{n}\sum_{i\in[n]}\Delta_{\vec{x}}(H_i,B_i+Y_i)=\frac{1}{n}\sum_{i\in[n]}\mathbb{I}[B_i+Y_i\leq\vec{x}_{H_i}]$$

Applying Theorem 16, we have that, for any $\alpha\in(0,1)$ with $\log(1/\alpha)>c_3\sqrt{|\mathcal{H}|}$, with probability $1-\alpha$

$$\sup_{\Delta_{\vec{x}}\in\mathcal{D}}\left|\mathbb{E}_{H,B,Y\sim P(H,B,Y)}[\Delta_{\vec{x}}(H,B+Y)]-\frac{1}{n}\sum_{i\in[n]}\Delta_{\vec{x}}(H_i,B_i+Y_i)\right|\leq\sqrt{4c_2\sqrt{|K|}\log(1/\alpha)/n}\tag{38}$$

Further, note that $\mathcal{B}\subset(\mathcal{Y}+\mathcal{B})=\mathcal{X}$ and, for any $\vec{b}\in\mathcal{B}^{\mathcal{H}}$, the following are equivalent statements

$$B+Y\leq\vec{b}_H\iff Y=0\text{ and }B\leq\vec{b}_H\iff\mathbb{I}[B\leq\vec{b}_H](1-Y)=1$$

because if $Y=0$ and $B\leq\vec{b}_H$ then it implies that all statements are true (direction $\Longleftarrow\Longrightarrow$) and if either $Y=1$ or $B>\vec{b}_H$ then it implies that all statements are false (direction $\Longrightarrow\Longleftarrow$). Hence, we have that, for any $\Delta_{\vec{b}}\in\mathcal{D}_{\mathcal{L}}$ and $\Delta_{\vec{x}}\in\mathcal{D}$ with $\vec{x}=\vec{b}$, it holds that

$$\mathbb{E}_{H,B,Y\sim P(H,B,Y)}[\Delta_{\vec{x}}(H,B+Y)]=P(B+Y\leq\vec{b}_H)=P(Y=0,\Delta_{\vec{b}}(H,B)=1)$$

and similarly

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}\Delta_{\vec{x}}(H_i,B_i+Y_i)&=\frac{1}{n}\sum_{i\in[n]}\mathbb{I}[B_i+Y_i\leq\vec{x}_{H_i}]\\&=\frac{1}{n}\sum_{i\in[n]}\mathbb{I}\left[B_i\leq\vec{x}_{H_i}\right](1-Y_i)=P_n(Y=0,\Delta_{\vec{b}}(H,B)=1).\end{aligned}$$

Thus, because of Eq. 38 and $\mathcal{B}\subset\mathcal{X}$, Eq. 36 must hold for all $\Delta_{\vec{b}}\in\mathcal{D}_{\mathcal{L}}$.

Analogously, to show Eq. 37, let $K_i=H_i$ and

$$X_i=B_i(1-Y_i)\qquad\forall i\leq[n].$$

where $\mathcal{X}=\mathcal{B}\cup\{0\}$ is countable. Consider any policy $\Delta_{\vec{x}}$ in countable class $\mathcal{D}$ where $\vec{x}\in\mathcal{X}^{\mathcal{H}}$ and the following estimator

$$\frac{1}{n}\sum_{i\in[n]}(1-\Delta_{\vec{x}}(H_i,B_i+Y_i))=\frac{1}{n}\sum_{i\in[n]}\mathbb{I}[B_i(1-Y_i)>\vec{x}_{H_i}]$$

Applying Corollary 17, we have that, for any $\alpha\in(0,1)$ with $\log(1/\alpha)>c_3\sqrt{|\mathcal{H}|}$, with probability $1-\alpha$

$$\sup_{\Delta_{\vec{x}}\in\mathcal{D}}\left|\mathbb{E}_{H,B,Y\sim P(H,B,Y)}[1-\Delta_{\vec{x}}(H,B(1-Y))]-\frac{1}{n}\sum_{i\in[n]}(1-\Delta_{\vec{x}}(H_i,B_i(1-Y_i)))\right|\leq\sqrt{4c_2\sqrt{|K|}\log(1/\alpha)/n}\tag{39}$$

Similarly, note that, for $\vec{b} \in \mathcal{B}^{\mathcal{H}}$, the following are equivalent statements

$$B(1-Y) > \vec{b}_H \iff Y = 0 \text{ and } B > \vec{b}_H \iff \mathbb{I}[B > \vec{b}_H](1-Y) = 1$$

Hence, we have that, for any $\Delta_{\vec{b}} \in \mathcal{D}_{\mathcal{L}}$ and $\Delta_{\vec{x}} \in \mathcal{D}$ with $\vec{x} = \vec{b}$, it holds that

$$\mathbb{E}_{H,B,Y\sim P(H,B,Y)}[1 - \Delta_{\vec{x}}(H, B(1-Y)] = P(B(1-Y) \leq \vec{b}_H) = P(Y = 0, \Delta_{\vec{b}}(H, B) = 0)$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}(1 - \Delta_{\vec{x}}(H_i, B_i(1-Y_i))) &= \frac{1}{n}\sum_{i\in[n]}\mathbb{I}[B_i(1-Y_i) > \vec{x}_{H_i}] \\ &= \frac{1}{n}\sum_{i\in[n]}\mathbb{I}\left[B_i > \vec{x}_{H_i}\right](1-Y_i) = P_n(Y = 0, \Delta_{\vec{b}}(H, B) = 0). \quad (40)\end{aligned}$$

Thus, because of Eq. 39 and $\mathcal{B} \subseteq \mathcal{X}$, Eq. 37 must hold for all $\Delta_{\vec{b}} \in \mathcal{D}_{\mathcal{L}}$. □

Lemma 6. *Let $\mathcal{B}$ be countable (and recall $\mathcal{H}$ is discrete and $u : \mathcal{A} \times \mathcal{Y} \to [0, 1]$). At time step t, let $\bar{\ell}_t \in \mathcal{L}$ be the function with highest average utility,* i.e.*, $\bar{\ell}_t = \arg\max_{\ell\in\mathcal{L}} \bar{\mu}_t(\ell)$ with*

$$\begin{aligned}\bar{\mu}_t(\ell) = &P_t(Y = 0, B > \ell(H)) \cdot (u(1,0) - u(1,1)) \\ &+ P_t(Y = 0, B \leq \ell(H)) \cdot (u(0,0) - u(0,1)) \\ &+ P_t(B \leq \ell(H)) \cdot (u(0,1) - u(1,1)) + u(1,1), \quad (41)\end{aligned}$$

where $P_t(\cdot)$ is the probability estimated with observations $(H_{t'}, B_{t'}, Y_{t'})_{t'<t}$ as defined in Lemma 5. Then, there exists constant $c_3 > 0$ such that, for any $\alpha \in (0, 1)$ with $\log(3/\alpha) > c_3\sqrt{|\mathcal{H}|}$, it holds that with probability $1 - \alpha$

$$\mu(b^*) - \mu(\bar{\ell}_t) \leq 6 \cdot \sqrt{4c_2\sqrt{|\mathcal{H}|}\log(3/\alpha)/t}.$$

Proof. Let $\texttt{rad}_t := \sqrt{4c_2\sqrt{|\mathcal{H}|}\log(3/\alpha)/t}$. Because each function in $\mathcal{L}$ induces a policy in $\mathcal{D}_{\mathcal{L}}$, the supremum over $\mathcal{L}$ can be substituted by the supremum over $\mathcal{D}_{\mathcal{L}}$. Using Lemma 5 for $\alpha \in (0, 1)$ with $\log(3/\alpha) > c_3\sqrt{|\mathcal{H}|}$, we have that

$$\begin{aligned}&P\left(\sup_{\ell\in\mathcal{L}} |P(B > \ell(H)) - P_t(B > \ell(H))| \leq \texttt{rad}_t\right) \\ &= P\left(\sup_{\Delta_{\vec{b}}\in\mathcal{D}_{\mathcal{L}}} |P(\Delta_{\vec{b}}(H, B) = 1) - P_n(\Delta_{\vec{b}}(H, B) = 1)| \leq \sqrt{4c_2\sqrt{|\mathcal{H}|}\log(3/\alpha)/t}\right) \\ &\geq 1 - \alpha/3.\end{aligned}$$

Analogously, we can show the same result for the other estimators in $\bar{\mu}_t(\ell)$. Now let $\bar{\ell}_t = \text{argmax}_{\ell\in\mathcal{L}} \bar{\mu}_t(\ell)$ and let $\ell^* = \text{argmax}_{\ell\in\mathcal{L}} \mu(\ell)$, then we have that

$$\bar{\mu}_t(\bar{\ell}_t) \geq \bar{\mu}_t(\ell^*). \quad (42)$$

Because $u(\cdot) \in [0, 1]$, the error of each of the three estimators cannot be amplified when estimating $\bar{\mu}_t(\ell)$ but only aggregated. Hence, using a union bound over the three estimators in $\bar{\mu}_t(\ell)$, with probability greater that $1 - \alpha$, it holds that $\mu(\ell) + 3\texttt{rad}_t \geq \bar{\mu}_t(\ell)$ and that $\bar{\mu}_t(\ell) + 3\texttt{rad}_t \geq \mu(\ell)$ for any arm $\ell \in \mathcal{L}$. Therefore, given Eq. 42, we have

$$\mu(\bar{\ell}_t) + 6 \cdot \texttt{rad}_t \geq \bar{\mu}_t(\bar{\ell}_t) + 3\texttt{rad}_t \geq \bar{\mu}_t(\ell^*) + 3\texttt{rad}_t \geq \mu(\ell^*),$$

and thus

$$\mu(\ell^*) - \mu(\bar{\ell}_t) \leq 6 \cdot \texttt{rad}_t.$$

□

Lemma 7. *Under perfect alignment, consider a selection algorithm that for context H_t, B_t picks a function $\bar{\ell}_t \in \mathcal{L}$ and selects decision $A_t = \mathbb{I}[B_t > \bar{\ell}_t(H_t)]$. Let $P(\bar{\ell}_t)$ be the distribution over thresholds and $P(A_t \mid H_t, B_t)$ be the distribution over decisions induced by this selection algorithm and distributions $P(Y \mid H, B)$ and $P(H, B)$ at time step t (these distributions are relevant as they influence the history $(H_i, B_i, Y_i)_{i<t}$ that the algorithm observes). Then, we have that*

$$\mathbb{E}[R(T)] = \mathbb{E}_{h_t, b_t \sim P(H,B), a_t \sim P(A_t \mid H_t, B_t)} \left[\sum_{t=1}^{T} \mu(\pi^*(h_t, b_t) \mid h_t, b_t) - \mu(a_t \mid h_t, b_t) \right]$$
$$= \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} \left[\sum_{t=1}^{T} \mu(\ell^*) - \mu(\bar{\ell}_t) \right]. \quad (43)$$

Proof. For any fixed function $\bar{\ell}$, the decision is defined deterministically by $\mathbb{I}[B_t > \bar{\ell}(H_t)]$ given the fixed threshold $\bar{\ell}(H_t)$. Thus, we have that

$$\begin{aligned}
&\mathbb{E}_{h_t, b_t \sim P(H,B)} \left[\mu(\mathbb{I}[b_t > \bar{\ell}(h_t)] \mid h_t, b_t) \right] \\
&\quad = \mathbb{E}_{h_t, b_t \sim P(H,B)} \left[\mathbb{E}[\mu(1 \mid h_t, b_t) \cdot \mathbb{I}[b_t > \bar{\ell}(h_t)]] + \mathbb{E}[\mu(0 \mid h_t, b_t) \cdot \mathbb{I}[b_t \leq \bar{\ell}(h_t)]] \right] \\
&\quad = \mu(\bar{\ell})
\end{aligned}$$

Since, at each time step t, $\bar{\ell}_t$ is chosen by the selection algorithm under distribution $P(\bar{\ell}_t)$, and the induced decision is $A_t = \mathbb{I}[B_t > \bar{\ell}(H_t)]$, we can substitute the induced decision $A_t \sim P(A \mid H_t = h_t, B_t = b_t)$ for the function $\bar{\ell}_t \sim P(\bar{\ell}_t)$ and obtain

$$\begin{aligned}
&\mathbb{E}_{h_t, b_t \sim P(H,B), a_t \sim P(A_t \mid H_t = h_t, B_t = b_t)} \left[\mu(a_t \mid h_t, b_t) \right] \\
&\quad = \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} [\mathbb{E}_{h_t, b_t \sim P(H,B)} \left[\mu(\mathbb{I}[b_t > \bar{\ell}(h_t)] \mid h_t, b_t) \right] \\
&\quad = \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} [\mu(\bar{\ell}_t)]
\end{aligned}$$

Since by Theorem 5 the optimal policy π^* correspond to a fixed optimal function $\ell^* \in \mathcal{L}$, we get

$$\mathbb{E}_{h_t, b_t \sim P(H,B)} \left[\mu(\pi^*(h_t, b_t) \mid h_t, b_t) \right] = \mathbb{E}_{h_t, b_t \sim P(H,B)} \left[\mu(A^* \mid h_t, b_t) \right] = \mu(\ell^*)$$

where optimal decision A^* depends on optimal function ℓ^* with $A^* = \mathbb{I}[b_t > \ell^*(h_t)]$. Thus, by linearity of expectation, we have that

$$\mathbb{E}[R(T)] = \sum_{t=1}^{T} \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} [\mu(\ell^*) - \mu(\bar{\ell}_t)] = \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} \left[\sum_{t=1}^{T} \mu(\ell^*) - \mu(\bar{\ell}_t) \right]$$

□

With the above auxiliary Lemmas, we are ready to prove Theorem 6.

From Lemma 7, we have that

$$\mathbb{E}[R(T)] = \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} \left[\sum_{t=1}^{T} \mu(\ell^*) - \mu(\bar{\ell}_t) \right].$$

We show in the following that the expected regret of the multiclass online contextual learning problem (right hand side of the equation) is in $\mathcal{O}\left(\sqrt{T \log T}\right)$. Let $\bar{\ell}_t = \operatorname{argmax}_{\ell \in \mathcal{L}} \bar{\mu}(\ell)$, and consider the following event

$$\mathcal{E} = \{\mu(\ell^*) - \mu(\bar{\ell}_t) \leq 6 \cdot \sqrt{4 c_2 \sqrt{|\mathcal{H}|} \log(3T^3)/t}, \ \forall 1 \leq t \leq T\} \quad (44)$$

Then, using Lemma 6 and an union bound, we can bound the probability of the complimentary event $\mathcal{E}^c$ in terms of T, that is

$$P(\mathcal{E}^c) \leq \sum_{t \leq T} P\left(\mu(\ell^*) - \mu(\bar{\ell}_t) > 6 \cdot \sqrt{4 c_2 \sqrt{|\mathcal{H}|} \log(3T^3)/t} \right) = \frac{1}{T^2} \quad (45)$$

when $\log(3T^3) > c_3\sqrt{|\mathcal{H}|}$, that is, when $\sqrt{|\mathcal{H}|} < c_4 \log(T)$ for some $c_4 > 0$ (c_4 equals constant c in the statement of the theorem).

We can now rewrite and bound the expected utility conditional on this event as

$$\begin{aligned}
\mathbb{E}[R(T)] &= P(\mathcal{E}) \cdot \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} \left[\sum_{t=1}^{T} \mu(\ell^*) - \mu(\bar{\ell}_t) \,|\, \mathcal{E} \right] + P(\mathcal{E}^c) \cdot \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} \left[\sum_{t=1}^{T} \mu(\ell^*) - \mu(\bar{\ell}_t) \,|\, \mathcal{E}^c \right] \\
&\leq \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} \left[\sum_{t=1}^{T} \mu(\ell^*) - \mu(\bar{\ell}_t) \,|\, \mathcal{E} \right] + P(\mathcal{E}^c) \cdot T \\
&\leq \mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} \left[\sum_{t=1}^{T} \mu(\ell^*) - \mu(\bar{\ell}_t) \,|\, \mathcal{E} \right] + \frac{1}{T}.
\end{aligned}$$

The first term of the bound is bounded given the definition of event $\mathcal{E}$, *i.e.*,

$$\mathbb{E}_{\bar{\ell}_t \sim P(\bar{\ell}_t)} \left[\sum_{t=1}^{T} \mu(\ell^*) - \mu(\bar{\ell}_t) \,|\, \mathcal{E} \right] \leq \sum_{t=1}^{T} 6 \cdot \sqrt{4c_2\sqrt{|\mathcal{H}|}\log(3T^3)/t} \leq 12 \cdot \sqrt{12c_2 T\sqrt{|\mathcal{H}|}\log(3T)}$$

where the last inequality holds as $\sum_{t=1}^{T} 1/\sqrt{t} \leq 2\sqrt{T}$. For $\sqrt{|\mathcal{H}|} < c_4 \log(T)$, we can conclude that

$$\mathbb{E}[R(T)] \leq 12 \cdot \sqrt{12c_2 T\sqrt{|\mathcal{H}|}\log(3T)} + \frac{1}{T} = \mathcal{O}(\sqrt{T\log(T)}).$$

A.7 Proof of a DKW-like Inequality for Threshold Policies

We first introduce two theorems that will form the basis to prove the extension of the DKW inequality for functions of the form $\Delta_{\vec{x}}(k, x) := \mathbb{I}[x \leq \vec{x}_k]$ for $\vec{x} \in \mathcal{X}^{\mathcal{K}}$.

Theorem 11 (Talagrand's inequality Talagrand [1996], Bousquet [2003], Sen [2018])**.** *Let X_i, $i \in [n]$ be independent $\mathcal{X}$-valued random variables. Let $\mathcal{F}$ be a (countable) class of measurable real-valued functions on $\mathcal{X}$ such that $||f||_\infty \leq U < \infty$ and $\mathbb{E}[f(X_1)] = ... = \mathbb{E}[f(X_n)] = 0$, for all $f \in \mathcal{F}$. Let*

$$Z = \sup_{f \in \mathcal{F}} \sum_{i \in [n]} f(X_i) \qquad \textit{or} \qquad Z = \sup_{f \in \mathcal{F}} \left| \sum_{i \in [n]} f(X_i) \right|$$

and let parameters σ^2 and ν_n be defined as

$$U^2 \geq \sigma^2 \geq \frac{1}{n} \sum_{i \in [n]} \sup_{f \in \mathcal{F}} \mathbb{E}[f^2(X_i)] \qquad \textit{and} \qquad \nu_n = 2U\mathbb{E}[Z] + n\sigma^2.$$

Then, for all $\epsilon \geq 0$

$$P(Z \geq \mathbb{E}[Z] + \epsilon) \leq \exp\left(-\frac{\nu_n}{U^2} h_1\left(\frac{\epsilon U}{\nu_n} \right) \right) \leq \exp\left(\frac{-\epsilon^2}{2\nu_n + 2\epsilon U/3} \right) \tag{46}$$

where $h_1(w) = (1+w)\log(1+w) - w$ for $w \in \mathbb{R}$ and

$$P(Z \geq \mathbb{E}[Z] + \sqrt{2\nu_n\delta} + U\delta/3) \leq e^{-\delta}, \qquad \delta \geq 0.$$

Definition 12 (VC Dimension Sen [2018])**.** *Given a function class $\mathcal{F}$ of binary-valued functions, we say that the set $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{F}$ if*

$$|\mathcal{F}(x_1, \ldots, x_n)| = 2^n \quad \textit{where} \quad \mathcal{F}(x_1, \ldots, x_n) = \{(f(x_1), \ldots, f(x_n)) \,|\, f \in \mathcal{F}\},$$

i.e., *for any binary vector $\vec{u}$ of length n there exists a function $f \in \mathcal{F}$, so that $(f(x_1), \ldots, f(x_n)) = \vec{u}$. The VC dimension $V(\mathcal{F})$ of $\mathcal{F}$ is defined as the largest integer n for which there is some collection $x_1, \ldots, x_n$ of n points that can be shattered by $\mathcal{F}$.*

Definition 13 (Envelope function Sen [2018])**.** *An envelope function of a class $\mathcal{F}$ of functions is any function $F(\cdot)$ such that $|f(x)| \leq F(x)$, for every $x \in \mathcal{X}$ and $f \in \mathcal{F}$.*

Theorem 14 (Sen [2018]). *Let X_i, $i \in [n]$ be independent $\mathcal{X}$-valued random variables with distribution $P(X)$. Let $\mathcal{F}$ be a class of measurable real-valued functions on $\mathcal{X}$ with envelope F and VC dimension $V(\mathcal{F})$. Then, for some constant $C > 0$,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i\in[n]} f(X_i) - \mathbb{E}_{X\sim P(X)}(f(X))\right|\right] \leq C||F||_{P,2}\sqrt{\frac{V(\mathcal{F})}{n}} \tag{47}$$

where $||F||_{P,2} = \left(\int F^2(x)\cdot P(X=x)dx\right)^{1/2}$.

Next, we formally describe the class of functions we are interested in.

Definition 15. *Let $\mathcal{X}$ and $\mathcal{K}$ be countable, totally ordered sample spaces. Given a fixed vector $\vec{x} \in \mathcal{X}^{\mathcal{K}}$, $\vec{x} = (\vec{x}_k)_{k\in\mathcal{K}}$, let $\Delta_{\vec{x}}$ be the following function for points $(k, x) \in \mathcal{K}\times\mathcal{X}$*

$$\Delta_{\vec{x}}(k, x) := \mathbb{I}[x \leq \vec{x}_k].$$

We define $\mathcal{D}$ as the class of all such functions over $\mathcal{X}$, i.e.*,*

$$\mathcal{D} := \{\Delta_{\vec{x}} \,|\, \vec{x} \in \mathcal{X}^{\mathcal{K}}\}$$

We also define the complementary class of functions denoted by $\mathcal{D}^+$ as

$$\mathcal{D}^+ := \{1 - \Delta \,|\, \Delta \in \mathcal{D}\}$$

This class is simply the class of functions of the type $\Delta^+_{\vec{x}}(k, x) := \mathbb{I}[x > \vec{x}_k]$ for $\vec{x}_k \in \mathcal{X}$, i.e.*, $\mathcal{D}^+ = \{\Delta^+_{\vec{x}} \,|\, \vec{x} \in \mathcal{X}^{\mathcal{K}}\}$.*

Lemma 8. *Let (K_i, X_i), $i \in [n]$ be independent and identically distributed random variables with distribution $P(K, X)$ over sample space $\mathcal{K}\times\mathcal{X}$ and consider the following (countable) class of functions*

$$\mathcal{F} = \{\Delta(\cdot) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta(K, X)] \,|\, \Delta \in \mathcal{D}\}$$

where $\mathcal{D}$ is the class of functions as in Definition 15 and let

$$Z := n\sup_{\Delta\in\mathcal{D}}\left|\frac{1}{n}\sum_{i\in[n]}\Delta(K_i, X_i) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta(K, X)]\right|.$$

Then,

a) $Z = \sup_{f\in\mathcal{F}}\left|\sum_{i\in[n]} f(X_i)\right|$

b) $||f||_\infty = \sup_{k,x\in\mathcal{K}\times\mathcal{H}}|f(k, x)| \leq 1$ *for all* $f \in \mathcal{F}$

c) $\mathbb{E}[f(K_i, X_i)] = 0$ *for all* $f \in \mathcal{F}$, $i \in [n]$

d) $\mathbb{E}[f^2(K_i, X_i)] \leq 1/4$ *for all* $f \in \mathcal{F}$, $i \in [n]$

e) $||\mathbf{\Delta}||_{P,2} \leq 1$ *where $\mathbf{\Delta}$ is the envelope of $\mathcal{D}$*

f) $V(\mathcal{D}) \leq |\mathcal{K}|$

g) $\mathbb{E}[Z] \leq C\sqrt{|\mathcal{K}|n}$ *for some constant C*

Proof. We will show each statement in order. Statement a) follows from simple arithmetic manipulations and by substituting the supremum over class $\mathcal{D}$ by the supremum over class $\mathcal{F}$, that

is

$$
\begin{aligned}
Z &= n \sup_{\Delta \in \mathcal{D}} \left| \frac{1}{n} \sum_{i \in [n]} \Delta(K_i, X_i) - \mathbb{E}_{K,X \sim P(K,X)}(\Delta(K,X)) \right| \\
&= n \sup_{\Delta \in \mathcal{D}} \left| \frac{1}{n} \sum_{i \in [n]} \Delta(K_i, X_i) - \frac{1}{n} n \cdot \mathbb{E}_{K,X \sim P(K,X)}(\Delta(K,X)) \right| \\
&= \sup_{\Delta \in \mathcal{D}} \left| \sum_{i \in [n]} \left[\Delta(K_i, X_i) - \mathbb{E}_{K,X \sim P(K,X)}(\Delta(K,X)) \right] \right| \\
&= \sup_{f \in \mathcal{F}} \left| \sum_{i \in [n]} f(X_i) \right|
\end{aligned}
$$

To show statement b), note that all $\Delta \in \mathcal{D}$ are binary-valued functions. Thus, $f(k,x) \in [-1,1]$ for all $f \in \mathcal{F}$ and $k, x \in \mathcal{K} \times \mathcal{X}$ which implies $||f||_\infty \leq 1$ for all $f \in \mathcal{F}$.

Statement c) can be derived as follows. Let f be any function in $\mathcal{F}$, then using the linearity of expectation we have that

$$
\begin{aligned}
\mathbb{E}[f(K_i, X_i)] &= \mathbb{E}_{K_i, X_i \sim P(K,X)}[\Delta(K_i, X_i) - \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)]] \\
&= \mathbb{E}_{K_i, X_i \sim P(K,X)}[\Delta(K_i, X_i)] - \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)] = 0
\end{aligned}
$$

Statement d) can be derived as follows. Let f be any function in $\mathcal{F}$, then we have that

$$
\begin{aligned}
\mathbb{E}[f^2(K_i, X_i)] &= \mathbb{E}_{K_i, X_i \sim P(K,X)}[(\Delta(K_i, X_i) - \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)])^2] \\
&= \mathbb{E}_{K_i, X_i \sim P(K,X)}[\Delta(K_i, X_i)^2] + \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)]^2 \\
&\qquad - 2 \cdot \mathbb{E}_{K_i, X_i \sim P(K,X)}[\Delta(K_i, X_i)] \cdot \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)] \\
&= \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)^2] - \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)]^2 \\
&= \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)] - \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)]^2
\end{aligned}
$$

where the last equation follows because $\Delta(K,X)^2 = \Delta(K,X)$ as $\Delta(K,X) \in \{0,1\}$. Note that, since $\mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)] \in [0,1]$, it holds that

$$
\mathbb{E}[f^2(K_i, X_i)] = \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)](1 - \mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)]) \leq 1/4
$$

for all $f \in \mathcal{F}$ (with equality satisfied for Δ such that $\mathbb{E}_{K,X \sim P(K,X)}[\Delta(K,X)] = 1/2$).

To show statement e), first note that we can take the envelope of function class $\mathcal{D}$ to be the constant function $\mathbf{1}$ (as $\mathcal{D}$ is a class of binary-valued functions). It is straightforward to see that

$$
||\mathbf{1}||_{P,2} = \left(\int 1^2 \cdot P(K = k, X = x) dk dx \right)^{1/2} = 1.
$$

Statement f) can be derived by contradiction. First, note that for a class of simple threshold functions on $\mathcal{X}$ the VC-dimension is 1. More formally, for class

$$
\mathcal{G} = \{ g_\theta(x) := \mathbb{I}[x \leq \theta] \,|\, \theta \in \mathcal{X} \},
$$

and any two points $(x_1, x_2) \in \mathcal{X}^2$, if the larger point is labeled 1 by a function in $\mathcal{G}$, then the smaller point must also be labeled 1, so not all labelings are possible (*i.e.*, only $(0,1)$, if $x_1 > x_2$, or $(1,0)$, if $x_1 \leq x_2$, is possible but not both). Thus, we have that

$$
|\{ (g(x_1), g(x_2)) \,|\, g \in \mathcal{G} \}| < 2^2,
$$

that is no two points $(x_1, x_2) \in \mathcal{X}^2$ can be shattered by $\mathcal{G}$. By the same argument no larger sets can be shattered by $\mathcal{G}$ and we have that $V(\mathcal{G}) \leq 1$ (see Definition 12). Next, assume that $V(\mathcal{D}) = m > |\mathcal{K}|$. Then, there must exist a set of points $((k_1, x_1), \ldots, (k_m, x_m))$ such that

$$
|\{ (\Delta(k_1, x_1), \ldots, \Delta(k_m, x_m)) \,|\, \Delta \in \mathcal{D} \}| = 2^m \tag{48}
$$

By the pigeon hole principle, there must exist $k \in \mathcal{K}$ such that $|\mathcal{I}| \geq 2$, where $\mathcal{I} := \{i \in [m] \,|\, k_i = k\}$, and

$$|\{(\Delta(k, x_i))_{i\in\mathcal{I}} \,|\, \Delta \in \mathcal{D}\}| = 2^{|\mathcal{I}|} \geq 2^2$$

as otherwise Eq. 48 would not hold. This implies $V(\{\Delta(k,\cdot) \,|\, \Delta \in \mathcal{D}\}) \geq 2$. However, since $\{\Delta(k,\cdot) \,|\, \Delta \in \mathcal{D}\} \subseteq \mathcal{G}$, it follows that $V(\mathcal{G}) \geq 2$ which is a contradiction. Hence, $V(\mathcal{D}) \leq |\mathcal{K}|$.

Statement g) can be derived by using Theorem 14 together with statement e) and f),

$$\mathbb{E}[Z] = n \cdot \mathbb{E}\left[\sup_{\Delta\in\mathcal{D}} \left| \frac{1}{n} \sum_{i\in[n]} \Delta(K_i, X_i) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta(K, X)] \right|\right]$$
$$\leq n \cdot C||\boldsymbol{\Delta}||_{P,2} \sqrt{\frac{V(\mathcal{D})}{n}} \leq C\sqrt{|\mathcal{K}|n}$$

□

Theorem 16. *Let $(K_1, X_1), \ldots, (K_n, X_n)$ be n independent and identically distributed random variables with distribution $P(K, X)$ over sample space $\mathcal{K} \times \mathcal{X}$ and let $\mathcal{D}$ be the (countable) class of functions as in Definition 15. Then, there exist constant $c_1 > 0$ such that for any $\epsilon > c_1\sqrt{|\mathcal{K}|/n}$*

$$P\left(\sup_{\Delta\in\mathcal{D}} \left| \frac{1}{n} \sum_{i\in[n]} \Delta(K_i, X_i) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta(K, X)] \right| > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{4c_2\sqrt{|\mathcal{K}|}}\right)$$

for some constant $c_2 > c_1$.

Proof. This proof closely follows the proof of the DKW inequality for outcome space $\mathbb{R}^d$ outlined in Sen [2018]. Let Z be as in Lemma 8. We have to upper-bound the probability

$$P\left(\frac{Z}{n} > \epsilon\right) = P(Z > n\epsilon) = P(Z - \mathbb{E}[Z] > n\epsilon - \mathbb{E}[Z])$$

Note that for $\epsilon > 2$, this probability is zero (as $Z \leq 2n$). For $\epsilon > c_1\sqrt{|\mathcal{K}|/n}$ with $c_1 := 2C$ as in Lemma 8 g), then $n\epsilon - \mathbb{E}[Z] \geq (n\epsilon - c_1\sqrt{|\mathcal{K}|n}) > 0$ and we can apply Theorem 11 (other prerequisites shown in Lemma 8 a)- d)). Hence, for $2 \geq \epsilon > c_1\sqrt{|\mathcal{K}|/n}$,

$$P(Z > \mathbb{E}[Z] + (n\epsilon - \mathbb{E}[Z])) \leq \exp\left(\frac{-(n\epsilon - \mathbb{E}[Z])^2}{2\nu_n + 2(n\epsilon - \mathbb{E}[Z])U/3}\right). \tag{49}$$

We can use Lemma 8 b) and d) to obtain bounds $U \leq 1$ and $\nu_n \leq 2c_1\sqrt{|\mathcal{K}|n} + n/4$, inserting it in Eq. 49 we get

$$P(Z > \mathbb{E}[Z] + (n\epsilon - \mathbb{E}[Z])) \leq \exp\left(-\frac{(n\epsilon - \mathbb{E}[Z])^2}{4C\sqrt{|\mathcal{K}|n} + n/2 + \frac{2}{3}(n\epsilon - \mathbb{E}[Z])}\right) \tag{50}$$
$$\leq \exp\left(-\frac{(n\epsilon - C\sqrt{|\mathcal{K}|n})^2}{c_2 n\sqrt{|\mathcal{K}|}}\right) = \exp\left(-\frac{n(\epsilon - C\sqrt{|\mathcal{K}|/n})^2}{c_2\sqrt{|\mathcal{K}|}}\right)$$
$$\leq \exp\left(-\frac{n\epsilon^2}{4c_2\sqrt{|\mathcal{K}|}}\right) \tag{51}$$

where for $\epsilon \leq 2$ the denominator in Eq. 50 is upper bounded by $4C\sqrt{|\mathcal{K}|n} + n/4 + 4n/3$ which is in turn upper bounded by $c_2\sqrt{|\mathcal{K}|}n$ (for some $c_2 > c_1$) and Eq. 51 holds as for $\epsilon > 2C\sqrt{|\mathcal{K}|/n}$, $(\epsilon - C\sqrt{|\mathcal{K}|/n})^2 \geq (\epsilon - \epsilon/2)^2 = \epsilon^2/4$.

Thus, there exists constant $c_1 > 0$ such that for any $\epsilon > c_1\sqrt{|\mathcal{K}|/n}$

$$P\left(\sup_{\Delta\in\mathcal{D}} \left| \frac{1}{n} \sum_{i\in[n]} \Delta(K_i, X_i) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta(K, X)] \right| > \epsilon\right)$$
$$= P(Z > n\epsilon) \leq \exp\left(-\frac{n\epsilon^2}{4c_2\sqrt{|\mathcal{K}|}}\right)$$

for some constant $c_2 > c_1$. □

Corollary 17. *Given a natural number n, let $(K_1, X_1), \ldots, (K_n, X_n)$ be independent and identically distributed random variables with distribution $P(K, X)$ over sample space $\mathcal{K} \times \mathcal{X}$ and let $\mathcal{D}^+$ be the (countable) class of functions as in Definition 15. Then, there exist constant $c_1 > 0$ such that for any $\epsilon > c_1\sqrt{|\mathcal{K}|/n}$*

$$P\left(\sup_{\Delta^+ \in \mathcal{D}^+} \left|\frac{1}{n}\sum_{i\in[n]} \Delta^+(K_i, X_i) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta^+(K, X)]\right| > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{4c_2\sqrt{|\mathcal{K}|}}\right)$$

for some constant $c_2 > c_1$.

Proof. First, note that, for any $\Delta \in \mathcal{D}$, we have that

$$\left|\frac{1}{n}\sum_{i\in[n]} \Delta(K_i, X_i) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta(K, X)]\right| \tag{52}$$

$$= \left|-1 + \frac{1}{n}\sum_{i\in[n]} \Delta(K_i, X_i)) + 1 - \mathbb{E}_{K,X\sim P(K,X)}[\Delta(K, X)]\right| \tag{53}$$

$$= \left|-\left(\frac{1}{n}\sum_{i\in[n]} 1 - \Delta(K_i, X_i))\right) + \mathbb{E}_{K,X\sim P(K,X)}[1 - \Delta(K, X)]\right|. \tag{54}$$

By definition of $\mathcal{D}$ and $\mathcal{D}^+$, we also have that $\Delta \in \mathcal{D} \iff 1 - \Delta \in \mathcal{D}^+$. Hence, the supremum of the left expression in Eq. 52 over $\mathcal{D}$ and the right expression over $\mathcal{D}^+$ are the same. Using this fact and Theorem 8, we get that

$$P\left(\sup_{\Delta^+ \in \mathcal{D}^+} \left|\frac{1}{n}\sum_{i\in[n]} \Delta^+(K_i, X_i) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta^+(K, X)]\right| > \epsilon\right)$$

$$= P\left(\sup_{\Delta \in \mathcal{D}} \left|\frac{1}{n}\sum_{i\in[n]} \Delta(K_i, X_i) - \mathbb{E}_{K,X\sim P(K,X)}[\Delta(K, X)]\right| > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{4c_2\sqrt{|\mathcal{K}|}}\right).$$

□

B Learning to Decide with AI Assistance under Imperfect Alignment

In this section, we show that, under imperfect alignment, the decision policy found by Algorithm 1 is near-optimal, and the degree of alignment, as measured by the maximum alignment error (MAE), bounds the near-optimality gap.

To this end, we begin with the observation that the expected utility achieved by the decision policy $\pi(h, b) = \mathbb{I}[b > b^*(h)]$, defined in Theorem 2, matches that achieved by the optimal decision policy $\pi^*_{\texttt{mono}}$ within the set of decision policies monotone in the AI confidence.[12] This is because any policy $\pi_{\texttt{mono}}$ monotone in the AI confidence can be expressed as a threshold function $\pi_{\texttt{mono}}(h, b) = \mathbb{I}[b > b(h)]$, as shown elsewhere [Corbett-Davies et al., 2017].

Based on the above observation and Theorem 5 in Corvelo Benz and Gomez-Rodriguez [2023], which characterizes the maximum gap between the expected utility of the optimal decision policy π^* and the optimal decision policy within the set of decision policies monotone in both the human and AI confidence[13], it readily follows that

$$\mathbb{E}_{H,B\sim P(H,B)}[\mu(\pi^*(H,B) \mid H,B)] - \mathbb{E}_{H,B\sim P(H,B)}[\mu(\pi^*_{\texttt{mono}}(H,B) \mid H,B)] \\ \leq \text{MAE} \cdot \left[u(1,1) - u(0,1) + \frac{3}{2}(u(0,0) - u(0,1))\right], \quad (55)$$

because the set of decision policies that are monotone with respect to the AI confidence includes the set of decision policies that are monotone with respect to the human and the AI confidence.

Finally, since Algorithm 1 is guaranteed to learn $\pi^*_{\texttt{mono}}$ as T grows, as shown in Theorems 4 and 6, we can conclude that Algorithm 1 is near-optimal under imperfect alignment.

[12] A decision policy π is monotone in the AI confidence if, for any $h \in \mathcal{H}$ and any $b, b' \in \mathcal{H}$ such that $b \leq b'$, it holds that $\pi(h, b) \leq \pi(h, b')$.

[13] A decision policy π is monotone in the human and the AI confidence if, for any $h, h' \in \mathcal{H}$ and $b, b' \in \mathcal{H}$ such that $h \leq h'$ and $b \leq b'$, it holds that $\pi(h, b) \leq \pi(h', b')$.

C Additional Experimental Details

We provide the MAE and EAE for each group in the Human-Alignment dataset in Table 1 and each task in the Human-AI Interactions dataset in Table 2. The Human-Alignment dataset is publicly available under GNU General Public License v3.0 and Human-AI Interactions dataset is publicly available under MIT License. Further, we clarify that in the Human-Alignment dataset the game-specific parameter q is noted as r in Corvelo Benz and Gomez-Rodriguez [2025] and represents the fraction of red cards in the card pile of the game instance. For more details and the exact card game procedures, refer to Corvelo Benz and Gomez-Rodriguez [2025].

We have adapted two classical algorithms to exploit the monotonicity of the outcome probabilities under perfect alignment, namely, online isotonic regression (Algorithm 3) and follow-the-leader (Algorithm 4). Algorithm 3 uses isotonic regression to fit an estimator $\texttt{isotonic-regression}(\mathcal{D}_{h_t}, b_t)$ for the probability of $Y = 1$ given b_t using the observed AI confidences and outcomes for h_t, $\mathcal{D}_{h_t} = \{(b_i, y_i) \,|\, 0 < i < t \text{ and } h_i = h_t\}$. This estimate is used to make decision a_t according to Lemma 2. Algorithm 4 keeps track of the cumulative rewards $\text{reward}_t(h, b)$ of each threshold $b \in \mathcal{B}$ for each fixed $h \in \mathcal{H}$. At round t, the threshold $\bar{b}$ with highest cumulative reward for h_t is chosen setting decision $a_t = \mathbb{I}[b_t > \bar{b}]$. Finally, below we present further details on the implementation of our experiment.

Implementation details. Our experiments ran with `Python 3.13.5`, mainly using the open-source libraries `numpy` (BSD-3-Clause License) and `pandas` (BSD-3-Clause License). We provide the exact versions of all used libraries with our code as supplementary material. Our experiments ran on one node of an internal cluster of Debian machines with 2 AMD EPYC 7702 64-core CPU processors at 2.00GHz and 2TB of memory with total execution time below five minutes.

Group	EAE	MAE	# data samples
A	$6.5 \cdot 10^{-4}$	0.10	2,040
B	$69.3 \cdot 10^{-4}$	0.20	2,014
BP	$23.6 \cdot 10^{-4}$	0.12	2,184
C	$3 \cdot 10^{-4}$	0.06	1,896

Table 1: Expected Alignment Error (EAE) and Maximum Alignment Error (MAE) as in Corvelo Benz and Gomez-Rodriguez [2023] of the AI model within each participant group in the Human-Alignment dataset together with total number of data samples collected from participants in each group.

Task	EAE	MAE	# data samples
Art	$9.2 \cdot 10^{-4}$	0.108	4,701
Census	$298.4 \cdot 10^{-4}$	0.673	2,941
Cities	$2.7 \cdot 10^{-4}$	0.065	4,543
Sarcasm	$9.7 \cdot 10^{-4}$	0.130	2,878

Table 2: Expected Alignment Error (EAE) and Maximum Alignment Error (MAE) as in Corvelo Benz and Gomez-Rodriguez [2023] of the AI model used in each task in the Human-AI Interactions dataset together with total number of data samples in each task.

Algorithm 3 Online isotonic regression for online learning under perfect alignment

Require: # interactions T, set of human confidence values $\mathcal{H}$, set of AI confidence values $\mathcal{B}$.
  Initialize $\mathcal{D}_h \leftarrow \emptyset, \forall h \in \mathcal{H}$
  for t from 1 to T **do**
    Observe confidence values (h_t, b_t)
    if $\mathcal{D}_{h_t} = \emptyset$ **then**
      $\bar{p}_t(Y = 1|H = h_t, B = b_t) \leftarrow 0.5$
    else
      $\bar{p}_t(Y = 1|H = h_t, B = b_t) \leftarrow \texttt{isotonic-regression}(\mathcal{D}_{h_t}, b_t)$
    end if
    if $1 - \bar{p}_t(Y = 1|H = h_t, B = b_t) < \frac{u(1,1)-u(0,1)}{u(1,1)-u(1,0)+u(0,0)-u(0,1)}$ **then**
      $a_t \leftarrow 1$
    else
      $a_t \leftarrow 0$
    end if
    Take decision a_t and observe label value y_t
    Add (b_t, y_t) to set $\mathcal{D}_{h_t}$
  end for

Algorithm 4 Follow-the-leader algorithm for online learning under perfect alignment

Require: # interactions T, set of human confidence values $\mathcal{H}$, set of AI confidence values $\mathcal{B}$.
  Initialize $\text{reward}_1(h, b) \leftarrow 0, \forall (h, b) \in \mathcal{H} \times \mathcal{B}$
  for t from 1 to T **do**
    Observe confidence values (h_t, b_t)
    Compute leader $\bar{b} \leftarrow \arg\max_{b \in \mathcal{B}} \text{reward}_t(h_t, b)$
    if $b_t > \bar{b}$ **then**
      $a_t \leftarrow 1$
    else
      $a_t \leftarrow 0$
    end if
    Take decision a_t and observe label value y_t
    for each $b \in \mathcal{B}$ **do**
      $\text{reward}_{t+1}(h_{t+1}, b) \leftarrow \text{reward}_t(h_t, b) + u(\mathbb{I}[b_t > b], y_t)$
    end for
  end for

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction contain a summary of the contributions of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included a "Discussion and Limitations" section (Section 5) discussing limitations of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We state the assumptions clearly in Section 2 and 3. Proofs for every theoretical result are provided in the Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all necessary information in Section 4 and in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code and data as supplementary material which we will publicly release upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are specified in Section 4 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In Figure 1 and 2, we report 95% confidence intervals as measures of statistical significance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: We provide information on compute resources in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper does not violate the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss how our work may help to better understand human decision making under AI-assistance (see, for example, Section 5).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release no data or models. The online algorithms we propose pose no risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit the datasets' owners in Section 4, and provide licensing information on both code and data in Appendix C.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide documentation and an (anonymized) License with our code as supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper uses publicly available data from human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper uses publicly available data from human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This research does not involve LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.