Efficient Algorithms for Contextual Apple Tasting with Log-Loss

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper introduces two novel algorithmic approaches designed for the Contextual Apple Tasting problem, where the learner faces an asymmetric feedback structure by observing binary labels only upon an 'Accept' action. To address the inherent decision bias and exploration challenges, we propose two distinct but complementary strategies. First, we introduce LogCBPSide-AT, an algorithm leveraging Confidence Bounds for Partial monitoring (CBP) to explicitly quantify predictive uncertainty and effectively balance the exploration-exploitation trade-off. Second, we present LogCB-AT, an approach that reduces the apple tasting problem to an online regression oracle. This reduction-based strategy offers a computationally efficient and scalable alternative that fundamentally bypasses the complex, often intractable confidence bound constructions required by traditional methods. Theoretically, we prove that both algorithms achieve sublinear regret bounds for losses associated with the binary labels, guaranteeing robust performance even under fundamentally restricted feedback. The practical utility of our methods is empirically validated through adaptive Large Language Model (LLM) cascading, where they effectively optimize the trade-off between inference cost and response accuracy.

1 Introduction

In the field of sequential decision-making, the standard contextual bandit framework serves as a foundational model for learning under uncertainty. Within this setup, a learner observes a context and selects an action, receiving feedback only for the chosen arm. Achieving high performance in such tasks depends on effectively managing the balance between exploration and exploitation. The agent must decide whether to exploit its greedy action to maximize immediate rewards or explore alternative actions to gather information that enables more accurate future decision-making. While this standard framework has proven effective across a wide range of practical problems[Li et al., 2010, Tewari and Murphy, 2017, Durand et al., 2018], more challenging settings arise when the information structure is fundamentally restricted. This challenge is especially prominent in environments characterized by partial monitoring, where feedback may be entirely absent for certain choices. A particularly notable subclass of this setting is the Apple Tasting problem, originally introduced by Helmbold et al. [2000]. In this framework, the learner faces an asymmetric information structure where the true label of an instance (e.g., whether the apple is rotten or not) is revealed only if the learner decides to 'Accept' (or 'Taste') it. If the learner chooses to remain 'Reject' (or 'Pass'), no feedback is provided, leaving the true status of the instance unknown. The learner's goal is to minimize the accumulated loss induced by the interaction between its actions (Taste vs. Pass) and the binary labels (rotten vs non-rotten), e.g., reducing the number of untasted rotten apples and tasted non-rotten apples. This objective faces a fundamental challenge where errors in rejection (false negatives) are never explicitly corrected, potentially leading to a persistent bias where the model

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

remains stuck in a suboptimal state due to the continuous absence of corrective feedback.

In this paper, we study the contextual apple tasting problem, where action selection is informed by contextual information. The decision to Accept reveals the binary label at a specific cost, while the Reject arm does not provide any feedback, and potentially results in large loss according to the value of y_t. This setup is motivated by modern AI application scenarios, for which we present two representative examples below:

LLM cascading The system initially processes a query through a Small Language Model (SLM) and must determine whether to cascade to a Large Language Model (LLM) for a more refined result. While the SLM provides an efficient baseline response, complex queries may require the LLM's superior reasoning at the cost of significantly higher resources and latency. Since the true performance gain is only observable upon invoking the LLM, the system operates under partial feedback: staying with the SLM prevents learning whether the LLM would have produced a better answer. Optimizing this trade-off is critical to avoid either unnecessary computational cost or degraded user satisfaction from suboptimal SLM outputs. In this example, invoking LLM and staying with SLM correspond to 'Accept' and 'Reject' actions respectively, and whether a performance gain was obtained corresponds to binary label. The sequence of queries serves as contextual information.

Digital Healthcare Wearable devices monitor bio-signals (contextual information) to determine if a patient needs a clinical visit. This is a quintessential contextual apple tasting problem because the actual presence of a disease (binary label) can usually only be confirmed if the device has notified the patient to see a doctor ('Accept'). If the device remains silent ('Reject'), the true medical status remains unobserved. In this context, failing to notify a patient who is actually ill results in an exceptionally large loss. Conversely, notifying a healthy patient leads to significant patient inconvenience and unnecessary medical expenses.

The asymmetric feedback structure necessitates a sophisticated exploration strategy to prevent the model from becoming permanently biased due to missing data. In this paper, we propose two distinct but complementary algorithmic approaches that efficiently tackle this challenge: LogCBPSide-AT, an confidence bound algorithm based on a logistic model assumption on the label, and LogCB-AT, a reduction-to-regression-oracle approach that accommodates any model class that can be learned under the log-loss objective. The two algorihtms cater to different modeling assumptions and theoretical guarantees: the former employs a refined exploration-exploitations strategy within a more restrictive model class with possibly logarithmic regret, while the latter accommodates a broader class of models with $O(\sqrt{T})$ regret, where T is the time horizon.

First, we develop LogCBPSide-AT, an algorithm leveraging Confidence Bounds for Partial monitoring (CBP) to explicitly quantify predictive uncertainty under the logistic model. While this approach provides a principled and theoretically rigorous mechanism for balancing exploration and exploitation under partial feedback [Bartók and Szepesvári, 2012, Heuillet et al., 2024], constructing explicit confidence bounds can become computationally demanding when scaling to complex function classes. To provide a scalable alternative that accommodates a broader function class, including the logistic model, we subsequently introduce LogCB-AT, a reduction-based approach that adapts the SquareCB algorithm of Foster and Rakhlin [2020] to the binary, partial feedback setting. This approach fundamentally bypasses the challenge of explicit bound construction by reducing the bandit learning task to repeated calls to an online regression oracle with a log-loss objective. Consequently, it eliminates the need for complex exploration terms and allows for straightforward integration of black-box or complex predictive models. Importantly, both algorithms, LogCBPSide-AT and LogCB-AT, adapt their respective base algorithms–CBPSide and SquareCB– to the one-sided feedback structure, by exploiting the structural property of the Apple Tasting problem: the latent true label y_t is a shared ground truth independent of the learner's action.

2 Problem Setting and Preliminaries

We consider a contextual binary decision problem over T rounds. At each round $t \in [T] := \{1, \ldots, T\}$, the learner observes a context vector $x_t \in \mathcal{X} \subset \mathbb{R}^d$ and must choose an action $a_t \in \mathcal{A} := \{0, 1\}$, where 0 and 1 denote 'Reject' and 'Accept', respectively. After the action is chosen, a latent binary label $y_t \in \{0, 1\}$ is realized according to an unknown conditional distribution. The defining feature of apple tasting is its one-sided feedback structure. Specifically, the true label y_t is revealed only when the learner chooses the revealing action ($a_t = 1$); if the learner chooses the Reject action ($a_t = 0$), no feedback is observed.

Our basic assumption is realizability: the learner has access to a function class $\mathcal{F} \subseteq (\mathcal{X} \to [0,1])$ that contains the true conditional expectation of the binary label.

Assumption 1 (Realizability)**.** *There exists a function $f^\star \in \mathcal{F}$ such that, for all $x \in \mathcal{X}$,*

$$f^\star(x) = \mathbb{E}[y \mid x].$$

As for LogCBPSide-AT, we further assume (in Section 4) that $\mathcal{F}$ is the class of logistic functions, whereas for LogCB-AT, we impose no additional restriction beyond the existence of a regression oracle that can efficiently learn f^* under a log-loss objective (Assumption 2 in Section 5). We denote $f^\star(x_t) = \mathbb{E}[y_t \mid x_t]$ by $y_t^\star$, the conditional mean at round t. The learner maintains an estimate $\hat{f}_{t-1}$ of $f^\star$, and its prediction at round t is $\hat{y}_t := \hat{f}_{t-1}(x_t)$. The learner is evaluated through the asymmetric loss matrix

	$y = 0$	$y = 1$
$a = 0$	0	ℓ_{01}
$a = 1$	ℓ_{10}	ℓ_{11}

where $\ell_{01}, \ell_{10}, \ell_{11} \in [0,1]$. Here, ℓ_{01} is the loss of rejecting an item with label 1, ℓ_{10} is the loss of accepting an item with label 0, and ℓ_{11} is the cost incurred when accepting an item with label 1. Throughout the paper, we assume the following cost conditions

$$\ell_{01} > \ell_{11} \geq 0 \quad \text{and} \quad \ell_{10} > 0.$$

The strict inequality $\ell_{01} > \ell_{11}$ is necessary to prevent the Reject action ($a = 0$) from trivially always being the optimal choice. This aligns naturally with practical scenarios like LLM cascading: the performance penalty of missing a better answer (ℓ_{01}) inherently outweighs the cost of invoking the LLM (ℓ_{11}). Additionally, $\ell_{11} \geq 0$ simply reflects that any invocation incurs a basic cost, while $\ell_{10} > 0$ ensures that an unnecessary call strictly incurs a computational penalty.

For any scalar $z \in [0,1]$, we define the loss function as $\psi_a(z) := \ell_{01} z(1-a) + \{\ell_{10} + (\ell_{11} - \ell_{10})z\}a$. For label y, $\psi_0(y)$ and $\psi_1(y)$ represent the losses of rejection and acceptance, respectively. Also, since $\psi_a(\cdot)$ is linear in z for each $a \in \mathcal{A}$, we have,

$$\mathbb{E}[\psi_a(y_t) \mid x_t] = \psi_a\big(f^\star(x_t)\big) = \psi_a(y_t^\star).$$

The optimal action at round t is defined as

$$a_t^\star \in \arg\min_{a \in \mathcal{A}} \psi_a(y_t^\star),$$

i.e., the action with the smaller conditional expected loss under the true model. The instantaneous regret is then defined by

$$r_t := \psi_{a_t}(y_t) - \psi_{a_t^\star}(y_t),$$

and the cumulative regret is $R(T) := \sum_{t=1}^{T} r_t$. In Sections 4 and 5, we present two novel algorithms, LogCBPSide-AT and LogCB-AT, respectively, both of which achieve $O(\sqrt{T})$ upper bounds on $R(T)$ with high probability under their respective assumptions. As for LogCBPSide-AT, we also present instance-dependent regret bound with polylogarithmic order in T.

3 Related work

Partial monitoring (PM) studies sequential decision problems where the learner does not observe the loss directly, but only a feedback signal determined by the chosen action and the hidden label. In stochastic finite PM, games are classically categorized into trivial, easy, hard, and hopeless classes according to their minimax regret rates [Bartók and Szepesvári, 2012, Bartók et al., 2014]. Apple tasting is one of the canonical PM examples, and is known to be a two-action, two-outcome easy game, for which the minimax regret scales on the order of $\sqrt{T}$ [Heuillet et al., 2024].

Helmbold et al. [2000] was the first to formalize the apple tasting model, demonstrating that an $O(\sqrt{T})$ mistake bound is achievable through randomized algorithms in the realizable setting, where the target hypothesis is assumed to reside within the hypothesis class. This foundational work established the essential framework for transforming standard online learning models into the apple tasting paradigm. Subsequently, Raman et al. [2024] extended this research from a combinatorial perspective, providing a comprehensive analysis across both realizable and agnostic settings. By introducing a new combinatorial parameter known as 'Effective Width,' they established a trichotomy for mistake bounds in the realizable setting, showing that the expected number of mistakes must fall into one of three categories: $\Theta(1), \Theta(\sqrt{T})$, or $\Theta(T)$. Notably, they resolved a long-standing open question by characterizing agnostic learnability via the Littlestone dimension. In contrast to previous literature that primarily focused on randomized algorithms, Chase and Mehalel [2024] explored the feasibility of deterministic algorithms within the realizable setting. They proved that any hypothesis class learnable under Apple tasting feedback is necessarily learnable by a deterministic algorithm. By providing an optimal mistake bound of $O(\sqrt{L(\mathcal{H})T \log T})$, where $L(\mathcal{H})$ denotes the Littlestone dimension of the hypothesis class $\mathcal{H}$, this study completed the theoretical foundation for deterministic Apple tasting.

While the aforementioned literature established the theoretical foundation of apple tasting in non-contextual settings, subsequent research expanded this framework to contextual bandits, where side information is available for decision-making. Bartók and Szepesvári [2012] extended the confidence bound based method of Bartók et al. [2012] for partial monitoring to the contextual setting, introducing CBPSide. More recently, Heuillet et al. [2024] proposed RandCBP and RandCBPSide by incorporating randomization into these frameworks to improve exploration efficiency. Bayesian strategies have also been adapted to better balance the exploration-exploitation trade-off in partial monitoring. Tsuchiya et al. [2020] established theoretical regret bounds for Thompson Sampling (TS), while Kirschner et al. [2020] proposed Information-Directed Sampling (IDS) to minimize the information ratio. Specifically, Grant and Leslie [2021] revisited the apple tasting problem, demonstrating the empirical superiority of these Bayesian mechanisms in managing restricted feedback. In addition, Harris et al. [2023] proposed an EXP-based algorithm to address apple tasting in adversarial scenarios. While LogCBPSide-AT is adapted from CBPSide, it is simplified and specialized for the logistic contextual apple tasting setting. Beyond this, we establish an instance-dependent regret bound for LogCBPSide-AT, a guarantee that, to our knowledge, has not been previously derived for CBPSide. We note that the partial monitoring literature predominantly focuses on minimax bounds; among the few works that establish instance-dependent guarantees, the MED-based approach of Komiyama et al. does not prove a minimax bound, whereas LogCBPSide-AT achieves both.

Recent contextual bandit research has increasingly explored reductions to online regression oracles, beginning with SquareCB[Foster and Rakhlin, 2020]. Within this line of work, Zhang et al. [2023] proposed SquareCB.G, adapting SquareCB [Foster and Rakhlin, 2020] to the partial monitoring setting, including Apple Tasting. However, the SquareCB.G framework is restricted to the use of a square-loss oracle, which may not be well-suited for binary labels, as in the original Apple Tasting problem and the aforementioned modern applications in LLMs and Digital Healthcare. In this work, we adapt the SquareCB algorithm to the binary feedback setting using a log-loss oracle. The derivation of the action-selection policy fundamentally differs from that of SquareCB.G., resulting in a fully randomized algorithm, as opposed to the partially deterministic policy used in SquareCB.G.

4 Confidence bound-based algorithm: LogCBPSide-AT

In this section, we present LogCBPSide-AT (Algorithm 1), a deterministic algorithm for the logistic contextual apple tasting problem, adapted from CBPSide Bartók and Szepesvári [2012]. We analyze logistic function class, $f^*(x) = \mathbb{E}[y \mid x] = \sigma(x^T\theta^*)$. The algorithm maintains an estimate of θ^* (maximum quasi-likelihood estimator) alongside a confidence interval of width $\beta_{t-1}^{x_t}(\delta_t)$ (defined in Algorithm 1) for the unknown latent label y_t^* given context x_t and selects actions based on the relationship between this confidence interval and a decision threshold.

We define the flip point $\alpha \in (0, 1)$ as the value at which the expected losses of actions Accept and Reject are equal:

$$\psi_0(\alpha) = \psi_1(\alpha) \implies \alpha = \ell_{10}/(\ell_{10} + \ell_{01} - \ell_{11})$$

Algorithm 1 LogCBPSide-AT

Input: λ, α, $\{\delta_t\}$
Observe the context x_1, Play action $a_1 = 1$ (Accept) and Observe y_1
$N_1 = 1$ (N_t is the number of times the action Accept ($a_t = 1$) is chosen up to time t)
$V_1 = \lambda I + x_1 x_1^T$
Compute $\hat{\theta}_1$ such that $\left(y_1 - \sigma(x_1^T \hat{\theta}_1)\right) x_1 = 0$
for $t > 1$ **do**
  Observe the context x_t, Calculate $\hat{y}_t = \sigma(x_t^T \hat{\theta}_{t-1})$
  Let $\beta_{t-1}^{x_t}(\delta_t) = \frac{1}{2c_\sigma} \|x_t\|_{V_{t-1}^{-1}} \sqrt{\left(3 + 2\log\left(1 + \frac{2}{\lambda}\right)\right) \cdot 2d \log(N_{t-1}) \cdot \log\left(\frac{d}{\delta_t}\right)}$
  if $|\hat{y}_t - \alpha| < \beta_{t-1}^{x_t}(\delta_t)$ **then**
    Exploration phase
    Play action $a_t = 1$ (Accept) and observe y_t
  else
    Exploitation phase
    if $\hat{y}_t \leq \alpha$ **then**
      Play action $a_t = 0$ (Reject)
    else
      Play action $a_t = 1$ (Accept) and observe y_t
    end if
  end if
  Update $V_t = V_{t-1} + x_t x_t^T \cdot \mathbb{1}\{a_t = 1\}$
  Compute $\hat{\theta}_t$ such that $\sum_{s=1}^{t} \mathbb{1}\{a_s = 1\} \left(y_s - \sigma(x_s^T \hat{\theta}_t)\right) x_s = 0$
  $N_t = N_{t-1} + \mathbb{1}\{a_t = 1\}$
end for

For any context x_t, action $a = 0$ (Reject) is the optimal action whenever $y_t^* < \alpha$, and action $a = 1$ (Accept) is optimal otherwise. Accordingly, when the confidence interval lies entirely to one side of α, LogCBPSide-AT acts greedily by exploiting its current estimate $\hat{y}_t$ to select the optimal action. When the confidence interval overlaps with α, the algorithm resolves this uncertainty by selecting Accept, thereby obtaining feedback to refine future estimates and confidence bounds. We present the full procedure for LogCBPSide-AT in Algorithm 1.

The theoretical analysis of LogCBPSide-AT is deferred to Appendix C. We state the main regret guarantee below. LogCBPSide-AT achieves the minimax-optimal regret bound of order $\sqrt{T}$, consistent with the known lower bounds for easy partial monitoring games with side information.

Theorem 1 (LogCBPSide-AT minimax regret bound)**.** *Assume* $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$ *and* $\sup_{x \in \mathcal{X}} |x^T \theta^*| \leq C_{\max}$. *For* $\delta_t = \frac{1}{t}$,*the expected cumulative regret of the LogCBPSide-AT algorithm is bounded as:*

$$\mathbb{E}\left[Reg_T\right] \leq O\left(\left(1 + e^{C_{\max}}\right)^2 e^{-C_{\max}} d\sqrt{T} \log^{\frac{3}{2}}(T)\right).$$

(Ignoring logarithmic factors in d)

Furthermore, an important advantage of confidence-based algorithms in online learning with limited feedback is that they often enjoy instance dependent regret bound of order $\mathrm{polylog}(T)$. We prove that LogCBPSide-AT also enjoys a $\log^2 T$ regret bound (see Theorem 2), a guarantee that, to our knowledge, has not been previously derived for CBPSide. Proof of Theorem 2 is deferred to Appendix C.2.

Theorem 2 (LogCBPSide-AT instance dependent regret bound)**.** *Assume* $\sup_{x \in \mathcal{X}} \|x\|_2 \leq 1$, $\sup_{x \in \mathcal{X}} |x^T \theta^*| \leq C_{\max}$, *and the existence of* $\Delta > 0$ *such that* $\Delta \leq \min_{x \in \mathcal{X}} |\alpha - \sigma(x^T \theta^*)|$.

For $\delta_t = \frac{1}{t}$, the expected cumulative regret of the LogCBPSide-AT algorithm is bounded as:

$$\mathbb{E}\left[Reg_T\right] \leq O(\frac{1}{\Delta} d^2 \log^2 T)$$

(ignoring doubly logarithmic factors and $\mathrm{polylog}(\frac{1}{\Delta})$)

5 SquareCB-based algorithm: LogCB-AT

We now introduce **LogCB-AT**, our adaptation of SquareCB to the contextual apple tasting setting. Our algorithm relies on an online regression oracle, denoted by $\mathrm{Alg}_{\mathrm{KL}}$, which produces a prediction $\hat{y}_t = \hat{f}_{t-1}(x_t)$ at round t. Since the true label is revealed only when the learner takes the Accept action ($a_t = 1$), the oracle is updated only on those rounds.

Assumption 2 (Logarithmic-loss oracle guarantee)**.** *For any (possibly adaptive) sequence of contexts, actions, and labels, the online log loss regression oracle* $\mathrm{Alg}_{\mathrm{KL}}$ *satisfies*

$$\sum_{t=1}^{T} \mathbf{1}\{a_t = 1\} \ell_{\log}(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\{a_t = 1\} \ell_{\log}\big(f(x_t), y_t\big) \leq \mathrm{Reg}_{\mathrm{KL}}(T_1) \leq \mathrm{Reg}_{\mathrm{KL}}(T),$$

where $\hat{y}_t = \hat{f}_{t-1}(x_t)$, $\ell_{\log}$ denotes the standard log-loss (binary cross-entropy),

$$\ell_{\log}(\hat{y}, y) = -\big[y \log(\hat{y}) + (1-y) \log(1-\hat{y})\big],$$

and $T_1 := \sum_{t=1}^{T} \mathbf{1}\{a_t = 1\}$ denotes the number of rounds on which the action $a_t = 1$ is taken.

Our proposed algorithm, LogCB-AT, is presented in Algorithm 2. It closely follows the structure of SquareCB. At each round t, the learner first computes the greedy action b_t and defines the action-selection policy which a learning rate $\gamma > 0$ and an exploration parameter μ. The learner then samples an action a_t, observes the label y_t only if $a_t = 1$, and updates $\mathrm{Alg}_{\mathrm{KL}}$ only on those rounds. A notable difference from SquareCB is that (i) the oracle uses a log-loss objective rather than a square-loss objective, and (ii) the oracle is updated only on rounds with $a_t = 1$. These differences call for non-trivial regret analysis and a new formula for the optimal value of γ which enables an $O(\sqrt{T})$ regret bound under mild guarantees of the oracle. We elaborate on details in the following subsections.

We analyze the regret of LogCB-AT algorithm, and propose a new formula for the learning rate γ which effectively minimizes the upper bound of the regret under the challenge of one-sided, binary feedback. Our analysis begines by partitioning the time steps $[T]$ into two disjoint sets, $\mathcal{I}_1$ and $\mathcal{I}_0$, according to the greedy action b_t: $\mathcal{I}_1 := \{t \in [T] : b_t = 1\}$ and $\mathcal{I}_0 := \{t \in [T] : b_t = 0\}$. The regret is then analyzed separately over the rounds in $\mathcal{I}_1$ and $\mathcal{I}_0$.

Regret for $t \in \mathcal{I}_1$ For $t \in \mathcal{I}_1$, the regret is decomposed according to the following corollary, which originates from Foster and Rakhlin [2020].

Corollary 1. *Suppose Assumptions 1 hold. With probability at least $1 - \delta$, we have*

$$\sum_{t \in \mathcal{I}_1} r_t \leq \sum_{t \in \mathcal{I}_1} \sum_{a \in \mathcal{A}} p_{t,a} \left[(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*)) - \frac{\gamma}{4} (\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2 \right]$$
$$+ \frac{\gamma}{4} \sum_{t \in \mathcal{I}_1} \sum_{a \in \mathcal{A}} p_{t,a} (\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2 + \max(\ell_{10}, \ell_{01} - \ell_{11}) \sqrt{2T \log(2\delta^{-1})}.$$

The first term of the right-hand side is upper bounded by $4T/\gamma$ under the action selection policy specified in Algorithm 2, with specific choice of $\mu = 2$. The challenge arises in the second term,

Algorithm 2 LogCB-AT

Require: $\gamma > 0$, $\mu > 0$, online regression oracle for log loss $\mathrm{Alg}_{\mathrm{KL}}$

for $t = 1, \dots, T$ **do**

  Observe context $x_t \in \mathcal{X}$

  Obtain prediction $\hat{y}_t = \hat{f}_{t-1}(x_t)$ from $\mathrm{Alg}_{\mathrm{KL}}$

  Set $b_t \in \arg\min_{a \in \{0,1\}} \psi_a(\hat{y}_t)$

  Set

$$p_{t,a} = \frac{1}{\mu + \gamma\big(\psi_a(\hat{y}_t) - \psi_{b_t}(\hat{y}_t)\big)} \text{ for } a \neq b_t, \qquad p_{t,b_t} = 1 - \sum_{a \neq b_t} p_{t,a}$$

  Sample $a_t \sim p_t$

  if $a_t = 1$ **then**

    Observe y_t and update $\mathrm{Alg}_{\mathrm{KL}}$ with (x_t, y_t)

  end if

end for

which represents the squared loss of the prediction $\hat{y}_t$ over all rounds in $\mathcal{I}_1$, whereas Assumption 2 guarantees an upper bound only on the cumulative log loss over rounds with $a_t = 1$. To circumvent this issue, we exploit the structural property of Apple Tasting that, at each time t, both actions share a common latent label y_t, and leverage a probability-shift argument between $p_{t,0}$ and $p_{t,1}$ for rounds in $\mathcal{I}_1$. Finally, we employ the conversion between squared loss and log-loss used in Foster and Krishnamurthy [2021] to bound the second term as follows,

$$\frac{\gamma}{4} \sum_{t \in \mathcal{I}_1} \sum_{a \in \mathcal{A}} p_{t,a}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2 \leq \frac{\gamma}{4} \Lambda_1 \left(8\mathrm{Reg}_{\mathrm{KL}}(T) + 4 \log(2\delta^{-1}) \right),$$

where $\Lambda_1 = (\ell_{10} - \ell_{11})^2 + \ell_{01}^2/(\mu - 1)$. The complete proofs are deferred to Appendix A. We note that choosing $\gamma = O(\sqrt{T/\mathrm{Reg}_{\mathrm{KL}}(T)})$ yields a bound of $O(\sqrt{T\mathrm{Reg}_{\mathrm{KL}}(T)})$ for the cumulative regret over rounds in $\mathcal{I}_1$.

Regret for $t \in \mathcal{I}_0$ A similar technique to that used for analyzing the regret over rounds in $\mathcal{I}_1$ could also be employed to bound the regret for $t \in \mathcal{I}_0$. However, when $b_t = 0$, the probability ratio $p_{t,0}/p_{t,1}$ introduces an additional dependence on γ in the second term. As a result, setting $\gamma = O(\sqrt{T/\mathrm{RegKL}(T)})$ leads to a regret bound of $O(T\mathrm{RegKL}(T))$. The optimal choice of γ for rounds in $\mathcal{I}_0$ instead turns out to be $O\left(\{T/\mathrm{RegKL}(T)\}^{1/3}\right)$, which yields a suboptimal regret bound of $O\left(\{T\mathrm{Reg}_{\mathrm{KL}}(T)\}^{2/3}\right)$. To achieve $O(\sqrt{T\mathrm{Reg}_{\mathrm{KL}}(T)})$ regret for rounds in $\mathcal{I}_0$ as well, we devise a novel relationship between the regret and prediction loss that fundamentally differs from Corollary 2. Specifically, we further decompose the rounds in $\mathcal{I}_0$ according to whether $a_t^* = 1$ or $a_t^* = 0$, and derive the following inequalities for each case:

For $t \in \mathcal{I}_0 \cap \{t \in [T] : a_t^* = 1\}$,

$$\mathbb{E}[r_t] \leq C_\Delta \mathbb{E}[I(a_t = 1)|y_t^* - \hat{y}_t|] + \gamma C_\Delta^2 \mathbb{E}[I(a_t = 1)(y_t^* - \hat{y}_t)^2],$$

and for $t \in \mathcal{I}_0 \cap \{t \in [T] : a_t^* = 0\}$,

$$\mathbb{E}[r_t] \leq \frac{1}{\gamma} + C_\Delta \mathbb{E}[I(a_t = 1)|\hat{y}_t - y_t^*|],$$

where $C_\Delta = |\ell_{01} - \ell_{11} + \ell_{10}|$. We note that neither inequality involves a quadratic dependence on γ, and moreover, the prediction loss in the right-hand side is accumulated only over rounds with $a_t = 1$. Consequently, the cumulated regret over rounds in $\mathcal{I}_0$ enjoys the same $O(\sqrt{T\mathrm{Reg}_{KL}(T)})$ bound with the same choice of γ. Detailed proof is provided in Appendix A.

Building upon the analysis above, we present the regret bound for the LogCB-AT algorithm.

Theorem 3 (LogCB-AT Regret Bound)**.** *Suppose Assumptions 1 and 2 hold. For the choice of* $\gamma = O(\sqrt{T/\mathrm{Reg}_{\mathrm{KL}}(T)})$*, the cumulative regret of the LogCB-AT algorithm is bounded by:*

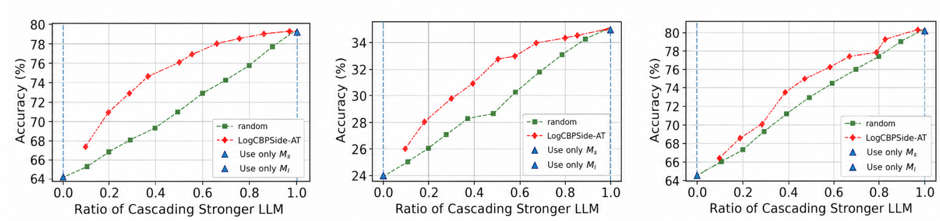
$$R(T) = O\left(\sqrt{T \cdot Reg_{KL}(T)}\right)$$

with probability at least $1 - \delta$*.*

Remark. LogCB-AT guarantees strictly positive, closed-form probabilities for all actions, satisfying $p_{t,a} \geq \frac{1}{2+\gamma \max(l_{01}-l_{11}, l_{10})}$ for all $a \in \{0, 1\}$. This fully randomized design contrasts with partially deterministic methods such as SquareCB.G and CBPSide, which may assign zero probability to non-greedy actions. This distinction is particularly important for Off-Policy Evaluation (OPE), where estimators such as IPW require non-zero action probabilities for valid counterfactual inference [Bian and Jun, 2022, Balagopalan and Jun, 2025]. By explicitly lower bounding the action probabilities, LogCB-AT avoids unstable IPW weights and enables robust offline evaluation.

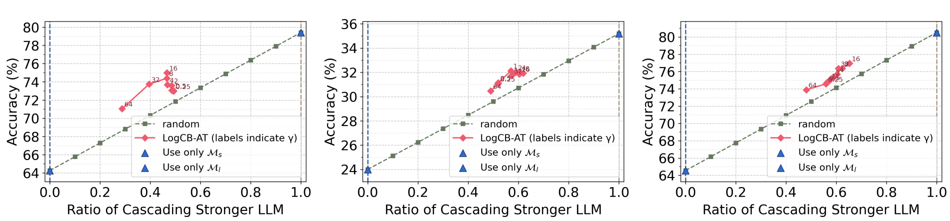
Comparison of LogCBPSide-AT and LogCB-AT In this work, we proposed two complementary algorithms for the contextual Apple Tasting problem with provable guarantees under respective assumptions. In particular, LogCBPSide-AT exhibits strong theoretical guarantees, with the possibility of achieving logarithmic regret under suitable assumptions in the logistic-linear setting. On the other hand, LogCB-AT does not attain logarithmic regret, but instead offers greater flexibility by accommodating richer hypothesis classes beyond the logistic-linear setting. Therefore, when a good and rich representation is available and the logistic-linear assumption is appropriate, LogCBPSide-AT may be preferred due to its stronger guarnatees. In contrast, when the logistic-linear assumption is hard to validate, LogCB-AT provides a more flexible alternative. In this sense, the two approaches serve as complementary alternatives, each being advantageous under different modeling assumptions and application scenarios.

6 Experiments



(a) ARC-Easy (b) OpenBookQA (c) BoolQ

Figure 1: Accuracy vs. LLM Invocation Ratio: LogCBPSide-AT



(a) ARC-Easy (b) OpenBookQA (c) BoolQ

Figure 2: Accuracy vs. LLM Invocation Ratio: LogCB-AT

Task Formulation In this section, we empirically evaluate the performance of the proposed LogCB-AT and LogCBPSide-AT algorithms on the practical real-world application of test-time adaptive Large Language Model (LLM) cascading, which we frame as a contextual Apple Tasting problem, namely AppleTeA: Test-Time Adaptive LLM Cascading via Logistic Apple Tasting. The goal is to dynamically route streaming queries to either a computationally efficient Small Language Model (SLM) or a highly capable Large Language Model (LLM), thereby optimizing the trade-off between inference cost and response accuracy. At each time step t, the system receives a query q_t and first generates a response using the SLM ($\mathcal{M}_s$). The algorithm then makes a binary decision $a_t \in 1, 0$: choosing $a_t = 0$ (Reject) finalizes the SLM response without additional cost, while choosing $a_t = 1$ (Accept) invokes the stronger LLM ($\mathcal{M}_l$) at a higher computational cost. The ground-truth label is defined as $y_t = 1$ if the response generated by $\mathcal{M}_s$ differs from that of $\mathcal{M}_l$, and $y_t = 0$ otherwise. Importantly, this label is observed only when $a_t = 1$, naturally inducing a partial-feedback apple tasting structure. We carefully construct the loss matrix to encode two competing cost components 1) the invocation cost of the strong LLM (i.e., token-level inference cost) and 2) the price of error incurred when the weak SLM is invoked in contexts where prediction failure is likely. Further details on the experiments, including the construction of contexts, datasets, and the choice of SLM/LLM models, are provided in Appendix D.

Baselines and Evaluation Metrics. We compare our algorithms against three baselines: (i) $\mathcal{M}_s$-only, (ii) $\mathcal{M}_l$-only, and (iii) random cascading with a fixed probability $p_{random} \in [0, 1]$. In the figure, the x-axis represents the proportion of total queries routed to the stronger LLM. Moving to the right indicates an increase in invocations, which directly translates to higher operational costs. The y-axis reflects the final accuracy resulting from these routing decisions.

Experimental Results Figure 1 shows that LogCBPSide-AT forms a smooth, concave curve, demonstrating a highly efficient trade-off between accuracy and LLM invocations. Figure 2 illustrates that LogCB-AT, as a probabilistic approach, does not deterministically greedily invoke the LLM solely to maximize accuracy. Instead, it inherently preserves exploration, effectively managing the exploration-exploitation trade-off to maintain robust cost-efficiency.

7 Conclusion

In this paper, we introduced two novel algorithms, LogCBPSide-AT and LogCB-AT, designed to tackle the contextual Apple Tasting problem. Specifically, LogCBPSide-AT specializes the partial monitoring-based CBPSide framework for the Apple Tasting scenario by introducing a logistic setting. Concurrently, LogCB-AT adapts the contextual bandit-based SquareCB algorithm to operate within a log-loss setting. By leveraging these frameworks, we provided rigorous theoretical guarantees, demonstrating that our algorithms achieve sublinear regret bounds. Beyond the theoretical formulations, we empirically validated the practical utility of our approaches in a real-world application of test-time adaptive Large Language Model (LLM) cascading. Our experiments showed that dynamically routing queries between a computationally efficient Small Language Model (SLM) and a highly capable LLM via these algorithms effectively optimizes the trade-off between inference cost and response accuracy. Building upon these foundational results, our future work aims to extend this framework to achieve first-order regret bounds. By adapting the FastCB algorithm [Foster and Krishnamurthy, 2021], we plan to develop a more advanced variant that scales with the loss of the optimal policy, yielding tighter regret guarantees and further accelerating convergence in practical environments.

Appendix

A LogCB-AT Regret Analysis

In this section, we derive a high-probability regret bound for the LogCB-AT algorithm. We aim to show that with probability at least $1-\delta$, the cumulative regret is bounded by $O(\sqrt{T \cdot \mathrm{Reg}_{\mathrm{KL}}(T)})$.

A.1 Definition of Regret

The instantaneous regret at round t is defined as the difference between the loss of the action taken a_t and the loss of the optimal action a_t^*.

$$r_t = \psi_{a_t}(y_t) - \psi_{a_t^*}(y_t). \tag{1}$$

The cumulative regret $R(T)$ over T rounds is the sum of these instantaneous regrets.

$$R(T) = \sum_{t=1}^{T} r_t. \tag{2}$$

Lemma 1. *For any $\hat{y}, y^* \in [0,1]$, the KL-divergence between two Bernoulli distributions with parameters $\hat{y}$ and y^* satisfies:*

$$d_{\mathrm{KL}}(\hat{y}, y^*) \geq \frac{1}{2}\frac{(\hat{y}-y^*)^2}{\hat{y}+y^*}. \tag{3}$$

Consequently, the following pointwise inequality holds for all $\hat{y}, y^* \in [0,1]$:

$$(\hat{y}-y^*)^2 \leq 4d_{\mathrm{KL}}(\hat{y}, y^*). \tag{4}$$

Proof. The first inequality is a known lower bound for the KL-divergence of Bernoulli distributions (see, e.g., Proposition 5 of Foster and Krishnamurthy [2021]).

Since $\hat{y}, y^* \in [0,1]$, their sum is at most 2 ($\hat{y}+y^* \leq 2$). Rearranging the first inequality, we obtain

$$(\hat{y}-y^*)^2 \leq 2(\hat{y}+y^*)d_{\mathrm{KL}}(\hat{y}, y^*) \leq 4d_{\mathrm{KL}}(\hat{y}, y^*).$$

□

Lemma 2. *Let $C_1 = |\ell_{11} - \ell_{10}|$ be the constant for the Accept action ($a=1$). From the definition of ψ_1, it follows that:*

$$(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 = C_1^2(\hat{y}_t - y_t^*)^2. \tag{5}$$

Then, under the Assumption 2 and Freedman's inequality, with probability at least $1-\delta$, the following bound holds:

$$\sum_{t=1}^{T} p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 \leq 8C_1^2\mathrm{Reg}_{\mathrm{KL}}(T) + 4C_1^2\log(2\delta^{-1}). \tag{6}$$

Proof. Let $\mathcal{F}_{t-1} = \sigma((x_1, a_1, y_1), \ldots, (x_{t-1}, a_{t-1}, y_{t-1}), x_t)$ be the filtration. Define

$$M_t := \mathbb{1}\{a_t = 1\}(\hat{y}_t - y_t^*)^2.$$

Set $Z_t := M_t - \mathbb{E}[M_t \mid \mathcal{F}_{t-1}]$. Since $(\hat{y}_t - y_t^*)^2 \leq 1$ and $0 \leq \mathbb{1}\{a_t = 1\} \leq 1$, the range is $0 \leq M_t \leq 1$. Then,

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = p_{t,1}(\hat{y}_t - y_t^*)^2, \quad \mathbb{E}[Z_t^2 \mid \mathcal{F}_{t-1}] \leq \mathbb{E}[M_t \mid \mathcal{F}_{t-1}].$$

Applying Freedman's inequality with range bound 1 and $\eta = \frac{1}{2}$, with probability at least $1 - \delta$

$$\sum_{t=1}^{T} \mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \leq 2 \sum_{t=1}^{T} M_t + 4 \log(2\delta^{-1}). \tag{7}$$

By Lemma 1, we know $d_{\mathrm{KL}}(\hat{y}_t, y_t^*) \geq \frac{1}{2} \frac{(\hat{y}_t - y_t^*)^2}{\hat{y}_t + y_t^*}$. Since $\hat{y}_t, y_t^* \in [0,1]$, we have $\hat{y}_t + y_t^* \leq 2$, which implies $(\hat{y}_t - y_t^*)^2 \leq 4 d_{\mathrm{KL}}(\hat{y}_t, y_t^*)$. Therefore, based on the Log-loss oracle assumption, we can bound the sum of M_t:

$$\sum_{t=1}^{T} M_t = \sum_{t=1}^{T} \mathbb{1}\{a_t = 1\}(\hat{y}_t - y_t^*)^2 \leq 4 \sum_{t=1}^{T} \mathbb{1}\{a_t = 1\} d_{\mathrm{KL}}(\hat{y}_t, y_t^*) \leq 4 \mathrm{Reg}_{\mathrm{KL}}(T).$$

Plugging this into (7), we get

$$\sum_{t=1}^{T} p_{t,1}(\hat{y}_t - y_t^*)^2 \leq 8 \mathrm{Reg}_{\mathrm{KL}}(T) + 4 \log(2\delta^{-1}).$$

Finally, multiplying both sides by C_1^2 and using $(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 = C_1^2 (\hat{y}_t - y_t^*)^2$, we obtain

$$\sum_{t=1}^{T} p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 \leq 8 C_1^2 \mathrm{Reg}_{\mathrm{KL}}(T) + 4 C_1^2 \log(2\delta^{-1}).$$

□

Lemma 3. *Suppose Assumptions 1 and 2 hold. Let $\mathcal{A} = \{1, 0\}$ be the action set, where $a = 1$ denotes the Accept action and $a = 0$ denotes the Reject action. Then, with probability at least $1 - \delta$, the cumulative regret $R(T)$ is bounded as*

$$R(T) \leq \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} p_{t,a} \left(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*) \right) + R' \sqrt{2T \log(2\delta^{-1})} \tag{8}$$

where $R' = \max(\ell_{10}, \ell_{01} - \ell_{11})$.

Proof. Let $\mathcal{F}_{t-1} = \sigma((x_1, a_1, y_1), \ldots, (x_{t-1}, a_{t-1}, y_{t-1}), x_t)$ be the filtration representing the history up to round t. The conditional expectation of the instantaneous regret $r_t = \psi_{a_t}(y_t) - \psi_{a_t^*}(y_t)$ with respect to $\mathcal{F}_{t-1}$ is given by

$$\begin{aligned} \mathbb{E}_{t-1}[r_t] &= \mathbb{E}\left[\psi_{a_t}(y_t) - \psi_{a_t^*}(y_t) \mid \mathcal{F}_{t-1} \right] \\ &= \sum_{a \in \mathcal{A}} p_{t,a} \left(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*) \right). \end{aligned}$$

We define a martingale difference sequence Z_t as the deviation of the realized regret from its conditional expectation

$$Z_t = r_t - \mathbb{E}_{t-1}[r_t].$$

By construction, $\mathbb{E}_{t-1}[Z_t] = 0$. To apply Azuma-Hoeffding's inequality, we determine the range of r_t by considering the possible values of the true label $y_t \in \{0, 1\}$:

- **Case** $y_t = 0$: The losses are $\psi_0(0) = 0$ (Reject) and $\psi_1(0) = \ell_{10}$ (Accept). Since $\ell_{10} \geq 0$ by Assumption 1, the optimal action is $a_t^* = 0$. Thus, $r_t = \psi_{a_t}(0) - \psi_0(0) = \psi_{a_t}(0)$. Since $a_t \in \{0, 1\}$, we have $r_t \in \{0, \ell_{10}\}$.

- **Case** $y_t = 1$: The losses are $\psi_0(1) = \ell_{01}$ (Reject) and $\psi_1(1) = \ell_{11}$ (Accept). Since $\ell_{01} \geq \ell_{11}$ by Assumption 2, the optimal action is $a_t^* = 1$. Thus, $r_t = \psi_{a_t}(1) - \psi_1(1)$. If $a_t = 1$, $r_t = 0$; if $a_t = 0$, $r_t = \ell_{01} - \ell_{11}$. This implies $r_t \in \{0, \ell_{01} - \ell_{11}\}$.

Combining both cases, the instantaneous regret is bounded within the interval $r_t \in [0, \max(\ell_{10}, \ell_{01} - \ell_{11})]$. Let $R' = \max(\ell_{10}, \ell_{01} - \ell_{11})$ be this upper bound. Since $0 \leq \mathbb{E}_{t-1}[r_t] \leq R'$, the range of the martingale difference Z_t is also bounded by R'.

Applying Azuma-Hoeffding's inequality to the sum $\sum_{t=1}^{T} Z_t$, we have that with probability at least $1 - \delta$

$$\sum_{t=1}^{T} (r_t - \mathbb{E}_{t-1}[r_t]) \leq R' \sqrt{2T \log(\delta^{-1})}.$$

we obtain

$$\begin{aligned} R(T) = \sum_{t=1}^{T} r_t &\leq \sum_{t=1}^{T} \mathbb{E}_{t-1}[r_t] + R' \sqrt{2T \log(2\delta^{-1})} \\ &= \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} p_{t,a} \left(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*) \right) + R' \sqrt{2T \log(2\delta^{-1})}. \end{aligned}$$

This completes the proof. □

A.2 Regret Decomposition

We partition the total rounds $[T]$ into two disjoint sets based on the learner's greedy decision $b_t = \arg\min_{a \in \{1,0\}} \psi_a(\hat{y}_t)$:

- B_1 **Rounds ($\mathcal{I}_1$):** $\{t \in [T] : b_t = 1\}$.
- B_0 **Rounds ($\mathcal{I}_0$):** $\{t \in [T] : b_t = 0\}$.

The regret in $\mathcal{I}_0$ requires an additional refined decomposition to address the lack of direct feedback. Specifically, we further categorize the rounds in $\mathcal{I}_0$ by comparing the greedy choice with the true optimal action a_t^*:

- $a_t^* = 1$ (Optimal is Accept): The optimal action is to Accept, but the greedy choice is to Reject. This represents a misclassification where the learner misses an opportunity to observe.
- $a_t^* = 0$ (Optimal is Reject): The optimal action is to Reject, and the greedy choice is correct. In this case, regret is only incurred due to the learner's stochastic exploration, specifically when the Accept action is sampled with probability $p_{t,1}$.

The resulting comprehensive decomposition of the cumulative regret $R(T)$ is as follows

$$\begin{aligned} R(T) &= \sum_{t \in \mathcal{I}_1} r_t + \sum_{t \in \mathcal{I}_0} r_t \\ &= \sum_{t \in \mathcal{I}_1} r_t + \sum_{t \in \mathcal{I}_0, a_t^* = 1} r_t + \sum_{t \in \mathcal{I}_0, a_t^* = 0} r_t \\ &= R_1(T) + R_{0,1}(T) + R_{0,0}(T). \end{aligned} \tag{9}$$

By applying Lemma 3, the realized cumulative regret $R(T)$ can be bounded by the sum of conditional expected regrets (denoted as $\bar{R}$) and the martingale concentration term. Specifically, with probability at least $1 - \delta$:

$$R(T) \leq \bar{R}_1(T) + \bar{R}_{0,1}(T) + \bar{R}_{0,0}(T) + R' \sqrt{2T \log(2\delta^{-1})}, \tag{10}$$

where $\bar{R}_1(T) = \sum_{t \in \mathcal{I}_1} \mathbb{E}_{t-1}[r_t]$, $\bar{R}_{0,1}(T) = \sum_{t \in \mathcal{I}_0, a_t^*=1} \mathbb{E}_{t-1}[r_t]$, and $\bar{R}_{0,0}(T) = \sum_{t \in \mathcal{I}_0, a_t^*=0} \mathbb{E}_{t-1}[r_t]$.

Corollary 2. *Suppose Assumptions 1 and 2 hold. Following Lemma 3, with probability at least $1-\delta$, the cumulative regret $R(T)$ can be decomposed and bounded as*

$$R(T) \leq \sum_{t\in\mathcal{I}_1}\sum_{a\in\mathcal{A}} p_{t,a}\left[(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*)) - \frac{\gamma}{4}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2\right] \tag{11}$$
$$+ \frac{\gamma}{4}\sum_{t\in\mathcal{I}_1}\sum_{a\in\mathcal{A}} p_{t,a}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2 \tag{12}$$
$$+ \bar{R}_0(T) + R'\sqrt{2T\log(2\delta^{-1})}. \tag{13}$$

Proof. From Lemma 3, with probability at least $1-\delta$, the realized cumulative regret $R(T)$ is bounded by the sum of conditional expected regrets and the martingale concentration term.

$$R(T) \leq \bar{R}_1(T) + \bar{R}_0(T) + R'\sqrt{2T\log(2\delta^{-1})},$$

where the conditional expected regrets are defined as $\bar{R}_1(T) = \sum_{t\in\mathcal{I}_1}\sum_{a\in\mathcal{A}} p_{t,a}(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*))$ and $\bar{R}_0(T) = \sum_{t\in\mathcal{I}_0}\sum_{a\in\mathcal{A}} p_{t,a}(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*))$.

To connect the regret in B_1 rounds ($\bar{R}_1(T)$) to the regression oracle, we employ an additive-subtractive trick. We add and subtract the term $\frac{\gamma}{4}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2$ for each action in the $\bar{R}_1(T)$ summation.

$$\begin{aligned}
\bar{R}_1(T) &= \sum_{t\in\mathcal{I}_1}\sum_{a\in\mathcal{A}} p_{t,a}\left[\Big(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*)\Big) - \frac{\gamma}{4}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2 + \frac{\gamma}{4}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2\right] \\
&= \sum_{t\in\mathcal{I}_1}\sum_{a\in\mathcal{A}} p_{t,a}\left[(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*)) - \frac{\gamma}{4}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2\right] \\
&\quad + \frac{\gamma}{4}\sum_{t\in\mathcal{I}_1}\sum_{a\in\mathcal{A}} p_{t,a}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2.
\end{aligned}$$

□

Lemma 4. *Suppose Assumptions 1 and 2 hold. Let $C_1 = |\ell_{11} - \ell_{10}|$. By the definition of the loss function ψ, we have $(\psi_0(\hat{y}_t) - \psi_0(y_t^*))^2 = \ell_{01}^2(\hat{y}_t - y_t^*)^2$. Then, with probability at least $1-\delta$, the expected squared error of the loss estimates for rounds in $\mathcal{I}_1$ is bounded as*

$$\sum_{t\in\mathcal{I}_1}\sum_{a\in\{1,0\}} p_{t,a}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2 \leq \Lambda_1\left(8\mathit{Reg}_{KL}(T) + 4\log(2\delta^{-1})\right), \tag{14}$$

where the constant Λ_1 is defined as $\Lambda_1 := C_1^2 + \frac{\ell_{01}^2}{\mu-1}$.

Proof. For rounds $t \in \mathcal{I}_1$, the greedy action is to Accept ($b_t = 1$). We decompose the total squared error sum into components for each action $a \in \{1, 0\}$

$$\sum_{t\in\mathcal{I}_1}\sum_{a\in\{1,0\}} p_{t,a}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2 = \sum_{t\in\mathcal{I}_1} p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 + \sum_{t\in\mathcal{I}_1} p_{t,0}(\psi_0(\hat{y}_t) - \psi_0(y_t^*))^2. \tag{15}$$

For the first term (Accept action $a = 1$), we can apply the bound derived from Freedman's inequality in Lemma 2, summing only over $t \in \mathcal{I}_1 \subseteq [T]$

$$\sum_{t\in\mathcal{I}_1} p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 \leq \sum_{t=1}^{T} p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 \leq C_1^2\left(8\mathrm{Reg}_{\mathrm{KL}}(T) + 4\log(2\delta^{-1})\right). \tag{16}$$

For the second term (Reject action $a = 0$), we relate its error to the Accept action by bounding the probability ratio. Since $b_t = 1$ for $t \in \mathcal{I}_1$, we know that $\psi_0(\hat{y}_t) > \psi_1(\hat{y}_t)$. Based on the SquareCB sampling rule, the probabilities are defined as

$$p_{t,0} = \frac{1}{\mu + \gamma(\psi_0(\hat{y}_t) - \psi_1(\hat{y}_t))}, \quad p_{t,1} = 1 - p_{t,0}.$$

The probability ratio is then:

$$\frac{p_{t,0}}{p_{t,1}} = \frac{p_{t,0}}{1 - p_{t,0}} = \frac{1}{\mu + \gamma(\psi_0(\hat{y}_t) - \psi_1(\hat{y}_t)) - 1}.$$

Since this ratio is decreasing with respect to the difference $\psi_0(\hat{y}_t) - \psi_1(\hat{y}_t)$ (which is strictly positive), the maximum occurs as the difference approaches 0. Thus, we have the upper bound

$$\frac{p_{t,0}}{p_{t,1}} < \frac{1}{\mu - 1}.$$

Using the identity $(\psi_0(\hat{y}_t) - \psi_0(y_t^*))^2 = \frac{\ell_{01}^2}{C_1^2}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2$, we relate the squared error terms

$$\begin{aligned} p_{t,0}(\psi_0(\hat{y}_t) - \psi_0(y_t^*))^2 &= \frac{p_{t,0}}{p_{t,1}} \cdot \frac{\ell_{01}^2}{C_1^2} \cdot p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 \\ &< \frac{1}{\mu - 1} \cdot \frac{\ell_{01}^2}{C_1^2} \cdot p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2. \end{aligned}$$

Summing this over all $t \in \mathcal{I}_1$ and applying Lemma 2 yields

$$\begin{aligned} \sum_{t \in \mathcal{I}_1} p_{t,0}(\psi_0(\hat{y}_t) - \psi_0(y_t^*))^2 &\leq \frac{1}{\mu - 1} \frac{\ell_{01}^2}{C_1^2} \sum_{t=1}^{T} p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 \\ &\leq \frac{\ell_{01}^2}{\mu - 1} \left(8\text{Reg}_{\text{KL}}(T) + 4\log(2\delta^{-1})\right). \end{aligned} \tag{17}$$

Substituting (16) and (17) back into (15), we arrive at the final bound

$$\begin{aligned} \sum_{t \in \mathcal{I}_1} \sum_{a \in \{1,0\}} p_{t,a}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2 &\leq \left(C_1^2 + \frac{\ell_{01}^2}{\mu - 1}\right) \left(8\text{Reg}_{\text{KL}}(T) + 4\log(2\delta^{-1})\right) \\ &= \Lambda_1 \left(8\text{Reg}_{\text{KL}}(T) + 4\log(2\delta^{-1})\right). \end{aligned}$$

This completes the proof. □

Lemma 5. *Suppose Assumptions 1 and 2 hold. Let $\mathcal{I}_0 = \{t \in [T] : b_t = 0\}$ be the set of rounds where the greedy choice is the Reject action. Define the loss gap between the Reject and Accept actions as $\Delta(y) := \psi_0(y) - \psi_1(y)$, and let $C_\Delta := |\ell_{01} - \ell_{11} + \ell_{10}|$. By the definition of the linear loss functions, we have $|\Delta(y) - \Delta(y')| = C_\Delta |y - y'|$ for all $y, y' \in [0, 1]$. For the SquareCB exploration parameter $\mu = 2$, with probability at least $1 - \delta$, the sum of conditional expected regrets in B_0 rounds, $\bar{R}_0(T) = \bar{R}_{0,1}(T) + \bar{R}_{0,0}(T)$, is bounded by*

$$\begin{aligned} \bar{R}_0(T) \leq \frac{T}{\gamma} &+ 2C_\Delta \sqrt{T\left(8\textit{Reg}_{\textit{KL}}(T) + 4\log(2\delta^{-1})\right)} \\ &+ \gamma C_\Delta^2 \left(8\textit{Reg}_{\textit{KL}}(T) + 4\log(2\delta^{-1})\right). \end{aligned} \tag{18}$$

Proof. For any round $t \in \mathcal{I}_0$, the greedy choice is $b_t = 0$, which implies that the estimated gap is non-positive $\hat{\Delta}_t := \Delta(\hat{y}_t) \leq 0$. The true gap is denoted as $\Delta_t^* := \Delta(y_t^*)$. According to the LogCB sampling rule, the exploration probability for the Accept action ($a = 1$) is given by $p_{t,1} = \frac{1}{2+\gamma(\psi_1(\hat{y}_t) - \psi_0(\hat{y}_t))} = \frac{1}{2+\gamma(-\hat{\Delta}_t)}$.

We analyze the conditional expected regret $\mathbb{E}_{t-1}[r_t]$ for $t \in \mathcal{I}_0$. By definition, the conditional expected regret is the sum of instantaneous regrets over all possible sampled actions $a \in \{1, 0\}$, weighted by their exploration probabilities $p_{t,a}$:

$$\mathbb{E}_{t-1}[r_t] = \sum_{a \in \{1,0\}} p_{t,a}(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*)). \tag{19}$$

Notice that when the sampled action matches the optimal action ($a = a_t^*$), the term becomes zero (e.g., $\psi_1(y_t^*) - \psi_1(y_t^*) = 0$). Thus, regret is strictly incurred only when the algorithm explores a suboptimal action ($a \neq a_t^*$). We consider two exhaustive cases based on the true optimal action $a_t^* \in \{1, 0\}$.

Case 1: $a_t^* = 1$ **(**$\bar{R}_{0,1}$**)** The optimal action is Accept, meaning the true gap is positive ($\Delta_t^* > 0$). Expanding the expected regret definition from (19), the regret for sampling the correct action ($a = 1$) vanishes:

$$\begin{aligned}\mathbb{E}_{t-1}[r_t] &= p_{t,1}(\psi_1(y_t^*) - \psi_1(y_t^*)) + p_{t,0}(\psi_0(y_t^*) - \psi_1(y_t^*)) \\ &= 0 + p_{t,0}\Delta_t^* = p_{t,0}\Delta_t^*.\end{aligned}$$

Using the fact that $p_{t,0} = 1 - p_{t,1}$, the probability ratio is

$$\frac{p_{t,0}}{p_{t,1}} = \frac{1 - p_{t,1}}{p_{t,1}} = \Big(2 + \gamma(-\hat{\Delta}_t)\Big) - 1 = 1 + \gamma(-\hat{\Delta}_t).$$

We can rewrite the expected regret as

$$\mathbb{E}_{t-1}[r_t] = p_{t,1}\frac{p_{t,0}}{p_{t,1}}\Delta_t^* = p_{t,1}(1 + \gamma(-\hat{\Delta}_t))\Delta_t^*.$$

Since $\Delta_t^* > 0$ and $\hat{\Delta}_t \leq 0$, we bound the terms using the definition of Δ:

- $-\hat{\Delta}_t \leq \Delta_t^* - \hat{\Delta}_t \leq |\Delta_t^* - \hat{\Delta}_t| = C_\Delta |y_t^* - \hat{y}_t|$.
- $\Delta_t^* = \hat{\Delta}_t + (\Delta_t^* - \hat{\Delta}_t) \leq \Delta_t^* - \hat{\Delta}_t \leq |\Delta_t^* - \hat{\Delta}_t| = C_\Delta |y_t^* - \hat{y}_t|$ (because $\hat{\Delta}_t \leq 0$).

Substituting these upper bounds yields:

$$\mathbb{E}_{t-1}[r_t] \leq p_{t,1}(1 + \gamma C_\Delta |y_t^* - \hat{y}_t|)C_\Delta |y_t^* - \hat{y}_t| = C_\Delta p_{t,1}|y_t^* - \hat{y}_t| + \gamma C_\Delta^2 p_{t,1}(y_t^* - \hat{y}_t)^2.$$

Case 2: $a_t^* = 0$ **(**$\bar{R}_{0,0}$**)** The optimal action is Reject, meaning $\Delta_t^* \leq 0$. Even though the greedy action is correct ($b_t = a_t^* = 0$), regret is incurred due to stochastic exploration when $a = 1$ is sampled. Expanding the expected regret, the term for sampling the correct action ($a = 0$) vanishes:

$$\begin{aligned}\mathbb{E}_{t-1}[r_t] &= p_{t,1}(\psi_1(y_t^*) - \psi_0(y_t^*)) + p_{t,0}(\psi_0(y_t^*) - \psi_0(y_t^*)) \\ &= p_{t,1}(-\Delta_t^*) + 0 = p_{t,1}(-\Delta_t^*).\end{aligned}$$

We decompose $-\Delta_t^*$ by adding and subtracting $\hat{\Delta}_t$:

$$-\Delta_t^* = -\hat{\Delta}_t + (\hat{\Delta}_t - \Delta_t^*) \leq -\hat{\Delta}_t + |\hat{\Delta}_t - \Delta_t^*| = -\hat{\Delta}_t + C_\Delta |\hat{y}_t - y_t^*|.$$

Multiplying by $p_{t,1}$, we obtain:

$$\mathbb{E}_{t-1}[r_t] \leq p_{t,1}(-\hat{\Delta}_t) + C_\Delta p_{t,1}|\hat{y}_t - y_t^*|.$$

Notice that $p_{t,1}(-\hat{\Delta}_t) = \frac{-\hat{\Delta}_t}{2+\gamma(-\hat{\Delta}_t)}$. Since the function $f(x) = \frac{x}{2+\gamma x}$ is monotonically increasing for $x \geq 0$ and bounded by $\frac{1}{\gamma}$, and given $-\hat{\Delta}_t \geq 0$, we have $p_{t,1}(-\hat{\Delta}_t) \leq \frac{1}{\gamma}$. Therefore,

$$\mathbb{E}_{t-1}[r_t] \leq \frac{1}{\gamma} + C_\Delta p_{t,1}|\hat{y}_t - y_t^*|.$$

Combining the Bounds Summing the expected regrets from both cases over all rounds $t \in \mathcal{I}_0$, and relaxing the summation to all T rounds since $\mathcal{I}_0 \subseteq [T]$ and all terms are non-negative, we get:

$$\bar{R}_0(T) = \sum_{t\in\mathcal{I}_0} \mathbb{E}_{t-1}[r_t] \le \sum_{t\in\mathcal{I}_0} \left[\frac{1}{\gamma} + 2C_\Delta p_{t,1}|y_t^* - \hat{y}_t| + \gamma C_\Delta^2 p_{t,1}(y_t^* - \hat{y}_t)^2\right]$$
$$\le \frac{T}{\gamma} + 2C_\Delta \sum_{t=1}^{T} p_{t,1}|y_t^* - \hat{y}_t| + \gamma C_\Delta^2 \sum_{t=1}^{T} p_{t,1}(y_t^* - \hat{y}_t)^2.$$

Applying the Cauchy-Schwarz inequality to the middle term:

$$\sum_{t=1}^{T} p_{t,1}|y_t^* - \hat{y}_t| = \sum_{t=1}^{T} \sqrt{p_{t,1}} \cdot \sqrt{p_{t,1}}|y_t^* - \hat{y}_t| \le \sqrt{\sum_{t=1}^{T} p_{t,1}} \sqrt{\sum_{t=1}^{T} p_{t,1}(y_t^* - \hat{y}_t)^2} \le \sqrt{T \sum_{t=1}^{T} p_{t,1}(y_t^* - \hat{y}_t)^2}.$$

Substituting this back into the expected regret bound yields:

$$\bar{R}_0(T) \le \frac{T}{\gamma} + 2C_\Delta \sqrt{T \sum_{t=1}^{T} p_{t,1}(y_t^* - \hat{y}_t)^2} + \gamma C_\Delta^2 \sum_{t=1}^{T} p_{t,1}(y_t^* - \hat{y}_t)^2.$$

From Lemma 2, we know that with probability at least $1-\delta$, the sum of squared errors is bounded by $\sum_{t=1}^{T} p_{t,1}(\hat{y}_t - y_t^*)^2 \le 8\mathrm{Reg}_{\mathrm{KL}}(T) + 4\log(2\delta^{-1})$. Applying this upper bound yields the final result. □

> **Lemma 6** (Lemma 3 of SquareCB[Foster and Rakhlin, 2020])**.** *For any round t, the probability distribution p_t over the action set $\mathcal{A} = \{1, 0\}$ (where $K = 2$), chosen by the SquareCB algorithm, ensures that for any true loss vector defined by y_t^*:*
>
> $$\sum_{a\in\mathcal{A}} p_{t,a} \left[(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*)) - \frac{\gamma}{4}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2\right] \le \frac{4}{\gamma}. \tag{20}$$
>
> *Summing this inequality over all rounds $t \in \mathcal{I}_1$, we obtain:*
>
> $$\sum_{t\in\mathcal{I}_1}\sum_{a\in\mathcal{A}} p_{t,a} \left[(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*)) - \frac{\gamma}{4}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2\right] \le \frac{4|\mathcal{I}_1|}{\gamma} \le \frac{4T}{\gamma}. \tag{21}$$

Proof. The first inequality is a direct application of Lemma 3 in SquareCB[Foster and Rakhlin, 2020], adapted to our binary action setting with $K = 2$. By substituting the arbitrary loss f_a^* with our specific loss function $\psi_a(y_t^*)$ and the estimator $\widehat{y}_a$ with $\psi_a(\hat{y}_t)$, the bound holds for each individual round t. Summing this bound over all rounds $t \in \mathcal{I}_1$ and using the trivial upper bound $|\mathcal{I}_1| \le T$ yields the final cumulative result. □

> **Theorem 3** (LogCB-AT Regret Bound)**.** *Suppose Assumptions 1 and 2 hold. For the choice of $\gamma = O(\sqrt{T/\mathrm{Reg}_{\mathrm{KL}}(T)})$, the cumulative regret of the LogCB-AT algorithm is bounded by:*
>
> $$R(T) = O\left(\sqrt{T \cdot Reg_{KL}(T)}\right)$$
>
> *with probability at least $1-\delta$.*

Proof. From Corollary 2,

$$R(T) \le \sum_{t\in\mathcal{I}_1}\sum_{a\in\{1,0\}} p_{t,a} \left[(\psi_a(y_t^*) - \psi_{a_t^*}(y_t^*)) - \frac{\gamma}{4}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2\right]$$
$$+ \frac{\gamma}{4}\sum_{t\in\mathcal{I}_1}\sum_{a\in\{1,0\}} p_{t,a}(\psi_a(\hat{y}_t) - \psi_a(y_t^*))^2$$
$$+ \bar{R}_0(T) + R'\sqrt{2T\log(2\delta^{-1})}.$$

By substituting the upper bounds derived in Lemma 4, Lemma 5, and Lemma 6 into the inequality, we obtain

$$
\begin{aligned}
R(T) \leq &\frac{4T}{\gamma} + \frac{\gamma}{4}\Lambda_1 \left(8\mathrm{Reg}_{\mathrm{KL}}(T) + 4\log(2\delta^{-1})\right) \\
&+ \frac{T}{\gamma} + 2C_\Delta\sqrt{T\cdot\left(8\mathrm{Reg}_{\mathrm{KL}}(T) + 4\log(2\delta^{-1})\right)} + \gamma C_\Delta^2 \left(8\mathrm{Reg}_{\mathrm{KL}}(T) + 4\log(2\delta^{-1})\right) \\
&+ R'\sqrt{2T\log(2\delta^{-1})}.
\end{aligned}
$$

Grouping the terms by $\frac{1}{\gamma}$ and γ, we have:

$$
\begin{aligned}
R(T) \leq \frac{5T}{\gamma} + \gamma &\left[\left(2\Lambda_1 + 8C_\Delta^2\right)\mathrm{Reg}_{\mathrm{KL}}(T) + \left(\Lambda_1 + 4C_\Delta^2\right)\log(2\delta^{-1})\right] \\
&+ 2C_\Delta\sqrt{T\cdot\left(8\mathrm{Reg}_{\mathrm{KL}}(T) + 4\log(2\delta^{-1})\right)} + R'\sqrt{2T\log(2\delta^{-1})}.
\end{aligned}
$$

By setting the learning rate γ to balance the first two terms:

$$
\gamma = \sqrt{\frac{5T}{\left(2\Lambda_1 + 8C_\Delta^2\right)\mathrm{Reg}_{\mathrm{KL}}(T) + \left(\Lambda_1 + 4C_\Delta^2\right)\log(2\delta^{-1})}}, \tag{22}
$$

the sum of the $\frac{1}{\gamma}$ and γ terms simplifies to $2\sqrt{5T\left[\left(2\Lambda_1 + 8C_\Delta^2\right)\mathrm{Reg}_{\mathrm{KL}}(T) + \left(\Lambda_1 + 4C_\Delta^2\right)\log(2\delta^{-1})\right]}$.

This yields the final bounded expression:

$$
\begin{aligned}
R(T) \leq 2&\sqrt{5T\left[\left(2\Lambda_1 + 8C_\Delta^2\right)\mathrm{Reg}_{\mathrm{KL}}(T) + \left(\Lambda_1 + 4C_\Delta^2\right)\log(2\delta^{-1})\right]} \\
&+ 2C_\Delta\sqrt{T\cdot\left(8\mathrm{Reg}_{\mathrm{KL}}(T) + 4\log(2\delta^{-1})\right)} + R'\sqrt{2T\log(2\delta^{-1})}.
\end{aligned}
$$

Extracting the dominant terms with respect to T, we conclude

$$
R(T) = O\left(\sqrt{T\cdot\mathrm{Reg}_{\mathrm{KL}}(T)}\right). \tag{23}
$$

□

B SquareCB-AT Regret Analysis

In this section, we present the regret analysis for the Contextual Apple Tasting problem under the square loss setting, serving as an alternative to the log-loss framework used in LogCB-AT. To ensure compatibility with the standard SquareCB architecture, which inherently relies on square loss regression oracles, we adopt a **Linear Probability Model (LPM)** as our generative foundation. Under this setting, the conditional expectation is modeled linearly with a fixed offset:

$$
y_t \mid x_t \sim \text{Bernoulli}\left(x_t^\top\theta^\star + \frac{1}{2}\right). \tag{24}
$$

Accordingly, we introduce our new algorithm (Algorithm 3) and assumption (Assumption 3.

Assumption 3 (Squared-loss oracle guarantee)**.** *For any (possibly adaptive) sequence of contexts, actions, and labels, the regression oracle satisfies*

$$
\sum_{t=1}^{T}\mathbf{1}\{a_t = 1\}\left(\hat{y}_t - f^\star(x_t)\right)^2 \leq \mathrm{Reg}_{\mathrm{Sq}}(T_1) \leq \mathrm{Reg}_{\mathrm{Sq}}(T),
$$

Algorithm 3 SquareCB-AT

Require: Learning rate $\gamma > 0$, exploration parameter $\mu > 1$, online regression oracle $\mathrm{Alg}_{\mathrm{Sq}}$

```
for t = 1, ..., T do
    Observe context x_t ∈ 𝒳
    Obtain prediction ŷ_t = f̂_{t-1}(x_t) from SqAlg
    Compute ψ_0(ŷ_t) and ψ_1(ŷ_t)
    Set
```

$$b_t \in \arg\min_{a \in \{0,1\}} \psi_a(\hat{y}_t)$$

```
    for each a ≠ b_t do
        Set
```

$$p_{t,a} = \frac{1}{\mu + \gamma\big(\psi_a(\hat{y}_t) - \psi_{b_t}(\hat{y}_t)\big)}$$

```
    end for
    Set
```

$$p_{t,b_t} = 1 - \sum_{a \neq b_t} p_{t,a}$$

```
    Sample a_t ~ p_t
    if a_t = 1 then
        Observe y_t and update Alg_Sq with (x_t, y_t)
    end if
end for
```

where

$$T_1 := \sum_{t=1}^{T} \mathbf{1}\{a_t = 1\}$$

denotes the number of rounds on which the action A is taken.

Next, we introduce Lemma 7. This lemma serves a similar role to Lemma 2 in LogCB-AT.

Lemma 7. *Let $C_1 = |\ell_{11} - \ell_{10}|$ be the constant for the Accept action ($a = 1$). From the definition of ψ_1, it follows that:*

$$(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 = C_1^2(\hat{y}_t - y_t^*)^2. \tag{25}$$

Then, under Assumption 2 and Freedman's inequality, with probability at least $1 - \delta$, the following bound holds:

$$\sum_{t=1}^{T} p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 \leq 2C_1^2 \mathit{Reg}_{Sq}(T) + 4C_1^2 \log(2\delta^{-1}). \tag{26}$$

Proof. Let $\mathcal{F}_{t-1} = \sigma((x_1, a_1, y_1), \ldots, (x_{t-1}, a_{t-1}, y_{t-1}), x_t)$ be the filtration. Define

$$M_t := \mathbb{1}\{a_t = 1\}(\hat{y}_t - y_t^*)^2.$$

Set $Z_t := M_t - \mathbb{E}[M_t \mid \mathcal{F}_{t-1}]$. Since $(\hat{y}_t - y_t^*)^2 \leq 1$ and $0 \leq \mathbb{1}\{a_t = 1\} \leq 1$, the range is $0 \leq M_t \leq 1$. Then,

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = p_{t,1}(\hat{y}_t - y_t^*)^2, \quad \mathbb{E}[Z_t^2 \mid \mathcal{F}_{t-1}] \leq \mathbb{E}[M_t \mid \mathcal{F}_{t-1}].$$

Applying Freedman's inequality with range bound 1 and $\eta = \frac{1}{2}$, with probability at least $1 - \delta$:

$$\sum_{t=1}^{T} \mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \leq 2\sum_{t=1}^{T} M_t + 4\log(2\delta^{-1}). \tag{27}$$

By Assumption 2, $\sum_{t=1}^{T} M_t = \sum_{t=1}^{T} \mathbb{1}\{a_t = 1\}(\hat{y}_t - y_t^*)^2 \leq \text{Reg}_{\text{Sq}}(T)$. Plugging this into (27)

$$\sum_{t=1}^{T} p_{t,1}(\hat{y}_t - y_t^*)^2 \leq 2\text{Reg}_{\text{Sq}}(T) + 4\log(2\delta^{-1}).$$

Finally, multiplying both sides by C_1^2 and using $(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 = C_1^2(\hat{y}_t - y_t^*)^2$, we obtain

$$\sum_{t=1}^{T} p_{t,1}(\psi_1(\hat{y}_t) - \psi_1(y_t^*))^2 \leq 2C_1^2\text{Reg}_{\text{Sq}}(T) + 4C_1^2\log(2\delta^{-1}).$$

□

Theorem 4 (SquareCB-AT Regret Bound)**.** *Suppose Assumptions 1 and 3 hold. Then, with probability at least* $1 - \delta$*, the cumulative regret of the SquareCB-AT algorithm is bounded by:*

$$R(T) = O\left(\sqrt{T \cdot Reg_{Sq}(T)}\right). \tag{28}$$

Proof. The proof structure is essentially identical to the regret analysis of LogCB-AT presented in Theorem 3. The core steps and the regret decomposition remain exactly the same, with the only modification being the substitution of Lemma 2 (used for the log-loss setting) with Lemma 7 (derived for the square-loss setting).

we configure the optimal learning rate γ as

$$\gamma = \sqrt{\frac{5T}{\left(\frac{\Lambda_1}{2} + 2C_\Delta^2\right)\text{Reg}_{\text{Sq}}(T) + (\Lambda_1 + 4C_\Delta^2)\log(2\delta^{-1})}}. \tag{29}$$

Substituting this γ simplifies the sum of the $\frac{1}{\gamma}$ and γ terms, yielding the explicit bounded expression:

$$\begin{aligned} R(T) \leq 2\sqrt{5T\left[\left(\frac{\Lambda_1}{2} + 2C_\Delta^2\right)\text{Reg}_{\text{Sq}}(T) + (\Lambda_1 + 4C_\Delta^2)\log(2\delta^{-1})\right]} \\ + 2C_\Delta\sqrt{T \cdot \left(2\text{Reg}_{\text{Sq}}(T) + 4\log(2\delta^{-1})\right)} + R'\sqrt{2T\log(2\delta^{-1})}. \end{aligned} \tag{30}$$

Finally, we conclude

$$R(T) = O\left(\sqrt{T \cdot \text{Reg}_{\text{Sq}}(T)}\right). \tag{31}$$

□

C LogCBPSide-AT Regret Analysis

In this section we provide the detailed proof for Theorem 1.

Our reward model is,

$$\mathbb{E}\left[y_t \mid x_t\right] = \sigma\left(x_t^T \cdot \theta^*\right)$$

Following the work of Filippi et al. [2010], we use *maximum quasi-likelihood estimator* for $\hat{\theta}_t$. That is,

$$\sum_{s=1}^{t} \mathbb{1}\{a_s = 1\}\left(y_s - \sigma(x_s^T \hat{\theta}_t)\right) x_s = 0.$$

Since, obtained rewards y_s are conditionally independent of the context (x_s), we can modify the Proposition 1 of Filippi et al. [2010] from Appendix A.2 by replacing t with N_{t-1} (number of times action Reject was chosen).

Proposition 1. *Take any* δ_t, N_{t-1} *such that* $0 < \delta_t < \min\{1, \frac{d}{e}\}, 1 \leq N_{t-1} \leq T$. *Let* x_t *be any random variable. Let*

$$\beta_{t-1}^{x_t}(\delta_t) = \frac{2k_\sigma R_{\max}}{c_\sigma} \|x_t\|_{V_{t-1}^{-1}} \sqrt{\left(3 + 2\log\left(1 + \frac{2}{\lambda}\right)\right) \cdot 2d\log(N_{t-1}) \cdot \log\left(\frac{d}{\delta_t}\right)}.$$

Then, with probability at least $1 - \delta_t$*, it holds that*

$$|\sigma(x_t^T\theta^*) - \sigma(x_t^T\hat{\theta}_{t-1})| \leq \beta_{t-1}^{x_t}(\delta_t).$$

where,

- $R_{\max} := \max_t y_t = 1.$
- k_σ *:= Lipschitz constant of* σ.

$$\begin{aligned} \sigma'(z) &= \sigma(z)(1 - \sigma(z)) \\ &\leq \frac{1}{4} \\ k_\sigma &= \frac{1}{4}. \end{aligned}$$

- $C_{\max} = \sup_{x_t \in \mathcal{X}, \theta^* \in \Theta} x_t^T \cdot \theta^*$
- $c_\sigma := \inf \sigma'(z).$

$$\begin{aligned} c_\sigma &= \inf \sigma(z)(1 - \sigma(z)) \\ &= \sup(|\sigma(z)|)(1 - \sup(|\sigma(z)|)) \\ &= \frac{e^{C_{\max}}}{\left(1 + e^{C_{\max}}\right)^2}. \end{aligned}$$

Furthermore, we introduce 2 more Lemmas necessary for the regret analysis.

The Elliptical Potential Lemma (EPL),

Lemma 8 (Elliptical potential lemma adapted from Proposition 2 of Abeille and Lazaric [2017])**.** *Let* $x_1, x_2, ..., x_t \in \mathbb{R}^d$ *be a sequence of vectors with* $\|x_s\|_2 \leq 1, \forall s \in [t]$. *Let*

$V_{t-1} = \lambda I + \sum_{s=1}^{t-1} x_s x_s^T$ *for some* $\lambda > 0$. *Then,*

$$\sum_{s=1}^{t} \|x_s\|^2_{V_{s-1}^{-1}} \leq 2d \log(1 + \frac{t}{d\lambda}).$$

Adapting this to our case:
Let $x_1 \cdot \mathbb{1}\{a_1 = 1\}, x_2 \cdot \mathbb{1}\{a_2 = 1\}, ..., x_t \cdot \mathbb{1}\{a_t = 1\} \in \mathbb{R}^d$ *for some* $\{a_s\}_{s=1}^t \in \{0,1\}$. *Let* $V_{t-1} = \lambda I + \sum_{s=1}^{t-1} x_s x_s^T \mathbb{1}\{a_s = 1\}$ *for some* $\lambda > 0$. *Then,*

$$\sum_{s=1}^{t} \|x_s\|^2_{V_{s-1}^{-1}} \mathbb{1}\{a_s = 1\} \leq 2d \log(1 + \frac{t}{d\lambda}).$$

and Elliptical Potential Count Lemma (EPC)

Lemma 9 (Elliptical potential count lemma adapted from Lemma C.2 of Jun and Kim [2024])**.** *Let* $x_1, x_2, ..., x_t \in \mathbb{R}^d$ *be a sequence of vectors with* $\|x_s\|_2 \leq 1, \forall s \in [t]$. *Let* $V_{t-1} = \lambda I + \sum_{s=1}^{t-1} x_s x_s^T$ *for some* $\lambda > 0$. *Let* $J = \{s \in [t] : \|x_s\|^2_{V_{s-1}^{-1}} \geq L^2\}$ *for some* $L^2 \leq 1$. *Then,*

$$|J| \leq 3\frac{d}{L^2} \ln\left(1 + \frac{2}{L^2\lambda}\right).$$

Adapting this to our case:
Let $x_1 \cdot \mathbb{1}\{a_1 = 1\}, x_2 \cdot \mathbb{1}\{a_2 = 1\}, ..., x_t \cdot \mathbb{1}\{a_t = 1\} \in \mathbb{R}^d$ *for some* $\{a_s\}_{s=1}^t \in \{0,1\}$. *Let* $V_{t-1} = \lambda I + \sum_{s=1}^{t-1} x_s x_s^T \mathbb{1}\{a_s = 1\}$ *for some* $\lambda > 0$. *Let* $J = \{s \in [t] : \|x_s \mathbb{1}\{a_s = 1\}\|^2_{V_{s-1}^{-1}} \geq L^2\}$ *for some* $L^2 \leq 1$. *Then,*

$$|J| \leq 3\frac{d}{L^2} \ln\left(1 + \frac{2}{L^2\lambda}\right).$$

C.1 Minimax regret analysis

Theorem 1 (LogCBPSide-AT minimax regret bound)**.** *For* $\delta_t = \frac{1}{t}$, *with* $\forall t \quad \|x_t\|_2 \leq 1$ *and* $\sup_{x_t \in \mathcal{X}} x_t^T \theta^* \leq C_{\max}$, *the expected cumulative regret of the LogCBPSide-AT algorithm is bounded as:*

$$\mathbb{E}\left[Reg_T\right] \leq O\left(\frac{\left(1 + e^{C_{\max}}\right)^2}{e^{C_{\max}}} d\sqrt{T} \log^{\frac{3}{2}}(T)\right).$$

(Ignoring logarithmic terms in d)

Throughout the analysis, we use $\hat{y}_t = \sigma(x_t^\top \hat{\theta}_{t-1})$ and $y_t^* = \sigma(x_t^\top \theta^*)$ interchangeably to denote the estimated and true latent labels, respectively.

Let us define the good event $\mathcal{G}_t$ as,

$$\mathcal{G}_t = \left\{|\sigma(x_t^T \theta^*) - \sigma(x_t^T \hat{\theta}_{t-1})| \leq \beta_{t-1}^{x_t}(\delta_t).\right\}$$

Instantaneous regret,

$$r_t = \psi_{a_t}(y_t) - \psi_{a_t^*}(y_t)$$

Let $\mathcal{F}_{t-1} = \sigma((x_1, a_1, y_1), \ldots, (x_{t-1}, a_{t-1}, y_{t-1}), x_t)$ be the filtration representing the history up to round t. Furthermore, use the shorthand $\mathbb{E}_{t-1}[\cdot]$ to denote $\mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$.

$$\mathbb{E}_{t-1}[r_t] = \psi_{a_t}(y_t^*) - \psi_{a_t^*}(y_t^*)$$

Cumulative regret,

$$\begin{aligned}
\mathrm{Reg}_T &= \sum_{t=1}^{T} r_t \\
\mathbb{E}[\mathrm{Reg}_T] &= \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \\
\mathbb{E}[\mathrm{Reg}_T] &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} r_t \mathbb{1}\{\mathcal{G}_t\}\right]}_{F_1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} r_t \mathbb{1}\{\bar{\mathcal{G}}_t\}\right]}_{F_2}
\end{aligned}$$

First, let us bound F_2

$$\begin{aligned}
F_2 &= \mathbb{E}\left[\sum_{t=1}^{T} r_t \mathbb{1}\{\bar{\mathcal{G}}_t\}\right] \\
&= \max(\ell_{10}, \ell_{01} - \ell_{11}) \sum_{t=1}^{T} \mathbb{P}\left(\bar{\mathcal{G}}_t\right) \\
&= \max(\ell_{10}, \ell_{01} - \ell_{11}) \sum_{t=1}^{T} \delta_t \\
&= \max(\ell_{10}, \ell_{01} - \ell_{11}) \sum_{t=1}^{T} \frac{1}{t} && \text{(Proposition 1, set } \delta_t = \tfrac{1}{t}) \\
&\lessapprox \max(\ell_{10}, \ell_{01} - \ell_{11}) \log T \\
&= O(\log T)
\end{aligned}$$

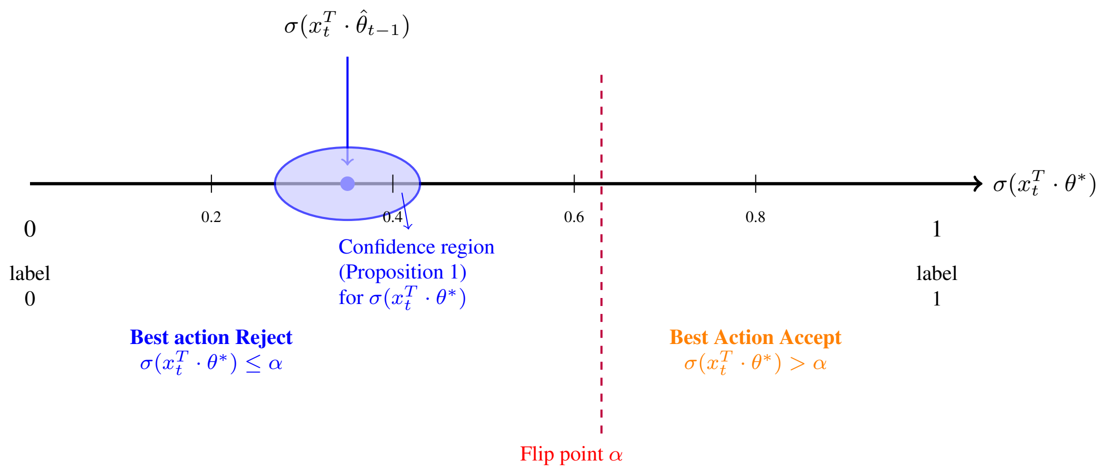
Now let us bound F_1,

$$\begin{aligned}
F_1 &= \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\{\mathcal{G}_t\}\right] \\
&= \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\{\mathcal{G}_t\} \, \mathbb{1}\left\{\sigma(x_t^T \hat{\theta}_{t-1}) \le \alpha - \beta_{t-1}^{x_t}(\delta_t)\right\}\right]}_{F_{11}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\{\mathcal{G}_t\} \, \mathbb{1}\left\{\sigma(x_t^T \hat{\theta}_{t-1}) \ge \alpha + \beta_{t-1}^{x_t}(\delta_t)\right\}\right]}_{F_{12}} \\
&\quad + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\{\mathcal{G}_t\} \, \mathbb{1}\left\{\alpha - \beta_{t-1}^{x_t}(\delta_t) < \sigma(x_t^T \hat{\theta}_{t-1}) < \alpha + \beta_{t-1}^{x_t}(\delta_t)\right\}\right]}_{F_{13}}.
\end{aligned}$$

Case F_{11}: $\hat{y}_t = \boldsymbol{\sigma(x_t^T \hat{\theta}_{t-1}) \le \alpha - \beta_{t-1}^{x_t}(\delta_t)}$

First of all $|\hat{y}_t - \alpha| \ge \beta_{t-1}^{x_t}(\delta_t)$ and $\hat{y}_t \le \alpha$, hence, $a_t = 0(\text{Reject})$ by the behavior of Algorithm 1

This is the case where, estimated distribution $\hat{y}_t$ is comfortably to the left of flip point, such that, based on Proposition 1, true mean of the distribution $\sigma(x_t^T \theta^*)$ itself will be on the left side of the flip point.



Since $\mathcal{G}_t$ happens,

$$
\begin{aligned}
|\sigma(x_t^T\theta^*) - \sigma(x_t^T\hat{\theta}_{t-1})| &\leq \beta_{t-1}^{x_t}(\delta_t) \\
\sigma(x_t^T\hat{\theta}_{t-1}) - \beta_{t-1}^{x_t}(\delta_t) \leq \sigma(x_t^T\theta^*) &\leq \sigma(x_t^T\hat{\theta}_{t-1}) + \beta_{t-1}^{x_t}(\delta_t) \\
\sigma(x_t^T\theta^*) &\leq \alpha
\end{aligned}
$$

Hence, $a_t^* = 0$.

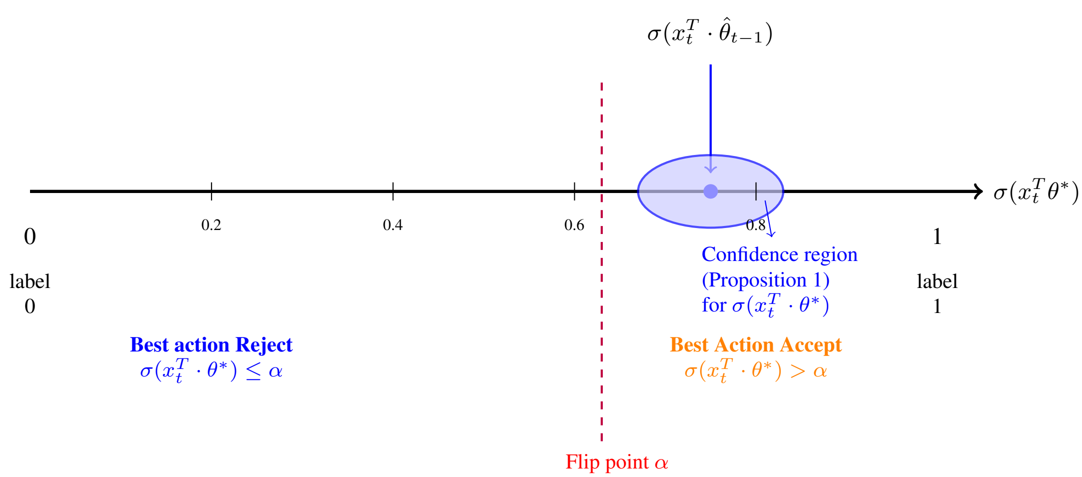
$$a_t = a_t^* = 0.$$

$$F_{11} = 0$$

Case F_{12}: $\hat{y}_t = \sigma(x_t^T\hat{\theta}_{t-1}) \geq \alpha + \beta_{t-1}^{x_t}(\delta_t)$

First of all $|\hat{y}_t - \alpha| \geq \beta_{t-1}^{x_t}(\delta_t)$ and $\hat{y}_t \geq \alpha$, hence, $a_t = 1(\text{Accept})$

This is the case where, estimated distribution $\hat{y}_t$ is comfortably to the right of flip point, such that, based on Proposition 1, true mean of the distribution $\sigma(x_t^T\theta^*)$ itself will be on the right side of the flip point.



Since $\mathcal{G}_t$ happens,

$$|\sigma(x_t^T\theta^*) - \sigma(x_t^T\hat{\theta}_{t-1})| \le \beta_{t-1}^{x_t}(\delta_t)$$
$$\sigma(x_t^T\hat{\theta}_{t-1}) - \beta_{t-1}^{x_t}(\delta_t) \le \sigma(x_t^T\theta^*) \le \sigma(x_t^T\hat{\theta}_{t-1}) + \beta_{t-1}^{x_t}(\delta_t)$$
$$\sigma(x_t^T\theta^*) \ge \alpha$$

Hence $a_t^* = 1(\text{Accept})$.

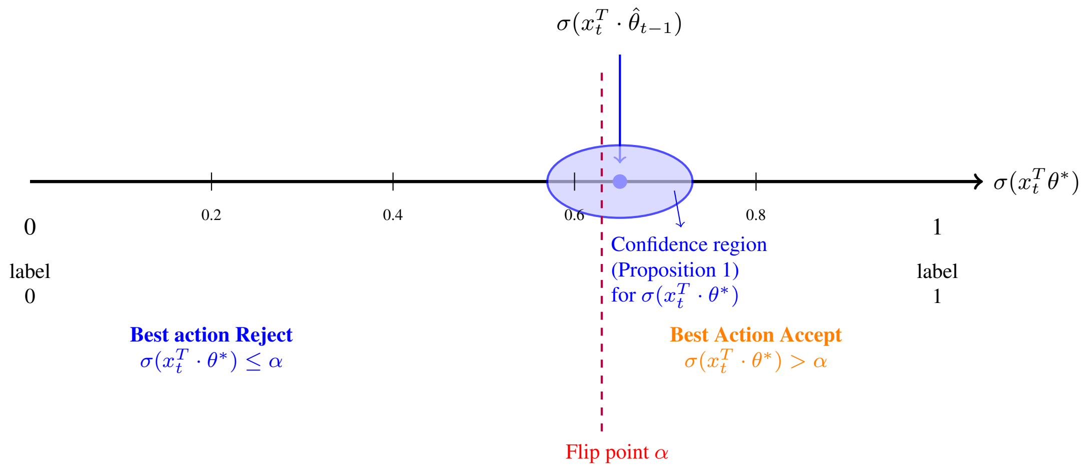
$$a_t = a_t^* = 1.$$

$$F_{12} = 0$$

Case F_{13}: $\alpha - \beta_{t-1}^{x_t}(\delta_t) < \sigma(x_t^T\hat{\theta}_{t-1}) < \alpha + \beta_{t-1}^{x_t}(\delta_t)$

First of all $|\hat{y}_t - \alpha| < \beta_{t-1}^{x_t}(\delta_t)$, hence, $a_t = 1(\text{Accept})$

This is an uncertain region which requires more exploration. However the nature of the confidence region $\beta_{t-1}^{x_t}(\delta_t)$ is such that it shrinks with more exploration (because its dependence on $\|x_t\|^2_{V_{t-1}^{-1}}$ which is a monotonically decreasing function). Hence the regret could be controlled using Elliptical potential lemma.



Since $\mathcal{G}_t$ happens,

$$|\sigma(x_t^T\theta^*) - \sigma(x_t^T\hat{\theta}_{t-1})| \le \beta_{t-1}^{x_t}(\delta_t)$$
$$\sigma(x_t^T\hat{\theta}_{t-1}) - \beta_{t-1}^{x_t}(\delta_t) \le \sigma(x_t^T\theta^*) \le \sigma(x_t^T\hat{\theta}_{t-1}) + \beta_{t-1}^{x_t}(\delta_t)$$
$$\alpha - 2\beta_{t-1}^{x_t}(\delta_t) \le \sigma(x_t^T\theta^*) \le \alpha + 2\beta_{t-1}^{x_t}(\delta_t)$$
$$-2\beta_{t-1}^{x_t}(\delta_t) \le \sigma(x_t^T\theta^*) - \alpha \le 2\beta_{t-1}^{x_t}(\delta_t)$$

$$
\begin{aligned}
\mathbb{E}_{t-1}\left[r_t\right] &= \mathbb{E}_{t-1}\left[\psi_{a_t}(y_t) - \psi_{a_t^*}(y_t)\right] \\
&= \mathbb{E}_{t-1}\left[\psi_1(y_t) - \psi_{a_t^*}(y_t)\right] \\
&\leq \left| \mathbb{E}_{t-1}\left[\psi_1(y_t) - \psi_0(y_t)\right] \right| \\
&= \left| \mathbb{E}_{t-1}\left[\ell_{10} + y_t(\ell_{11} - \ell_{10}) - \ell_{01} y\right] \right| \\
&= \left| \left[\ell_{10} + \mathbb{E}_{t-1}\left[y_t\right](\ell_{11} - \ell_{10}) - \ell_{01}\,\mathbb{E}_{t-1}\left[y_t\right]\right] \right| \\
&= \left| \left[\ell_{10} + \sigma(x_t^T\theta^*)(\ell_{11} - \ell_{10}) - \ell_{01}\sigma(x_t^T\theta^*)\right] \right| \\
&= \left| \left[\left(\sigma(x_t^T\theta^*) - \alpha\right)(\ell_{11} - \ell_{10}) - \ell_{01}\left(\sigma(x_t^T\theta^*) - \alpha\right)\right] \right| && \text{(definition of } \alpha\text{)} \\
&\leq 2\beta_{t-1}^{x_t}(\delta_t)\max(\ell_{10}, \ell_{01} - \ell_{11})
\end{aligned}
$$

$$
\begin{aligned}
F_{13} &= \sum_{t=1}^{T} \mathbb{E}\left[r_t\,\mathbb{1}\left\{\mathcal{G}_t\right\}\mathbb{1}\left\{\alpha - \beta_{t-1}^{x_t}(\delta_t) < \sigma(x_t^T\hat{\theta}_{t-1}) < \alpha + \beta_{t-1}^{x_t}(\delta_t)\right\}\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[r_t\,\mathbb{1}\left\{a_t = 1\right\}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_{t-1}\left[r_t\right]\mathbb{1}\left\{a_t = 1\right\}\right] \\
&\leq 2\max(\ell_{10}, \ell_{01} - \ell_{11})\,\mathbb{E}\left[\sum_{t=1}^{T} \beta_{t-1}^{x_t}(\delta_t)\,\mathbb{1}\left\{a_t = 1\right\}\right] \\
&= 2\max(\ell_{10}, \ell_{01} - \ell_{11})\,\mathbb{E}\left[\sum_{t=1}^{T} \frac{2k_\sigma R_{\max}}{c_\sigma}\|x_t\|_{V_{t-1}^{-1}}\sqrt{\left(3 + 2\log\left(1 + \frac{2}{\lambda}\right)\right)\cdot 2d\log(N_{t-1})\cdot\log\left(\frac{d}{\delta_t}\right)}(\delta_t)\,\mathbb{1}\left\{a_t = 1\right\}\right] \\
&= 2\max(\ell_{10}, \ell_{01} - \ell_{11})\,\mathbb{E}\left[\sum_{t=1}^{T} \frac{2k_\sigma R_{\max}}{c_\sigma}\|x_t\|_{V_{t-1}^{-1}}\sqrt{\left(3 + 2\log\left(1 + \frac{2}{\lambda}\right)\right)\cdot 2d\log(T)\cdot\log\left(dT\right)}(\delta_t)\,\mathbb{1}\left\{a_t = 1\right\}\right] \\
&= \frac{4\max(\ell_{10}, \ell_{01} - \ell_{11})k_\sigma R_{\max}}{c_\sigma}\sqrt{\left(3 + 2\log\left(1 + \frac{2}{\lambda}\right)\right)\cdot 2d\log(T)\cdot\log\left(Td\right)}\,\mathbb{E}\left[\sum_{t=1}^{T}\|x_t\|_{V_{t-1}^{-1}}\,\mathbb{1}\left\{a_t = 1\right\}\right] \\
&\leq \frac{4\sqrt{T}\max(\ell_{10}, \ell_{01} - \ell_{11})k_\sigma R_{\max}}{c_\sigma}\sqrt{\left(3 + 2\log\left(1 + \frac{2}{\lambda}\right)\right)\cdot 2d\log(T)\cdot\log\left(Td\right)}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\|x_t\|_{V_{t-1}^{-1}}^2\,\mathbb{1}\left\{a_t = 1\right\}\right]} \\
& && \text{(Cauchy-Schwartz)}
\end{aligned}
$$

528 By applying Lemma 8,

$$F_{13} \leq \left(\sqrt{2d\log(1+\frac{T}{d\lambda}})\right) \frac{4\sqrt{T}\max(\ell_{10},\ell_{01}-\ell_{11})k_\sigma R_{\max}}{c_\sigma} \sqrt{\left(3+2\log\left(1+\frac{2}{\lambda}\right)\right)\cdot 2d\log(T)\cdot\log(Td)}$$
$$= O\left(\frac{\left(1+e^{C_{\max}}\right)^2}{e^{C_{\max}}} d\sqrt{T}\log^{\frac{3}{2}}(T)\right) \qquad \text{(Ignoring logarithmic terms in } d\text{)}$$

Hence,

$$\mathbb{E}\left[\text{Reg}_T\right] \leq \left(\sqrt{2d\log(1+\frac{T}{d\lambda}})\right) \frac{4\sqrt{T}\max(\ell_{10},\ell_{01}-\ell_{11})k_\sigma R_{\max}}{c_\sigma} \sqrt{\left(3+2\log\left(1+\frac{2}{\lambda}\right)\right)\cdot 2d\log(T)\cdot\log(Td)}$$
$$+\log T$$
$$= O\left(\frac{\left(1+e^{C_{\max}}\right)^2}{e^{C_{\max}}} d\sqrt{T}\log^{\frac{3}{2}}(T)\right)$$
$$\text{(Ignoring logarithmic terms in } d\text{)}$$

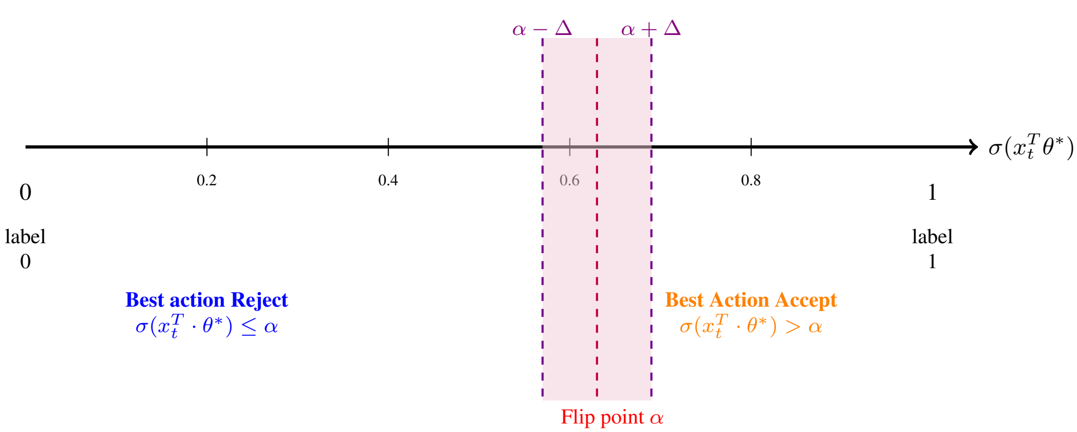
C.2 Instance dependent regret analysis

Theorem 2 (LogCBPSide-AT instance dependent regret bound)**.** *For $\delta_t = \frac{1}{t}$, with $\forall x \in \mathcal{X}$ $\|x\|_2 \leq 1$ and $\sup_{x\in\mathcal{X}} x^T\theta^* \leq C_{\max}$, and $0 < \Delta \leq \min_{x\in\mathcal{X}}|\alpha - \sigma(x^T\cdot\theta^*)|$ the expected cumulative regret of the LogCBPSide-AT algorithm is bounded as:*

$$\mathbb{E}\left[Reg_T\right] \leq O(\frac{1}{\Delta}d^2\log^2 T)$$
$$\text{(ignoring doubly logarithmic factors and polylog}(\tfrac{1}{\Delta})\text{)}$$

Let define the problem dependent parameter $\Delta > 0$ as,

$$\Delta \leq \min_{x\in\mathcal{X}}|\alpha - \sigma(x^T\cdot\theta^*)|$$



This defines how close $y_t^* = \sigma(x_t^T\cdot\theta^*)$ is to the decision boundary α. The closer, the problem instance become more difficult owing to the fact that we need more exploration to distinguish it. In the figure above, there are no instances of x_t falls into the shaded region.

With this additional assumption, we can refine our previous analysis.

F_2, F_{11}, F_{12} all go though for this case as well. We skip them and move to the analysis of F_{13}.

Case $\boldsymbol{F_{13}}$: $\boldsymbol{\alpha - \beta_{t-1}^{x_t}(\delta_t) < \sigma(x_t^T \hat{\theta}_{t-1}) < \alpha + \beta_{t-1}^{x_t}(\delta_t)}$

First of all $|\hat{y}_t - \alpha| < \beta_{t-1}^{x_t}(\delta_t)$, hence, $a_t = 1(\text{Accept})$

Since $\mathcal{G}_t$ happens,

$$
\begin{aligned}
|\sigma(x_t^T\theta^*) - \sigma(x_t^T\hat{\theta}_{t-1})| &\leq \beta_{t-1}^{x_t}(\delta_t) \\
\sigma(x_t^T\hat{\theta}_{t-1}) - \beta_{t-1}^{x_t}(\delta_t) \leq \sigma(x_t^T\theta^*) &\leq \sigma(x_t^T\hat{\theta}_{t-1}) + \beta_{t-1}^{x_t}(\delta_t) \\
\alpha - 2\beta_{t-1}^{x_t}(\delta_t) \leq \sigma(x_t^T\theta^*) &\leq \alpha + 2\beta_{t-1}^{x_t}(\delta_t) \\
-2\beta_{t-1}^{x_t}(\delta_t) \leq \sigma(x_t^T\theta^*) - \alpha &\leq 2\beta_{t-1}^{x_t}(\delta_t).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{t-1}[r_t] &= \mathbb{E}_{t-1}\left[\psi_{a_t}(y_t) - \psi_{a_t^*}(y_t)\right] \\
&= \mathbb{E}_{t-1}\left[\psi_1(y_t) - \psi_{a_t^*}(y_t)\right] \\
&\leq \left|\mathbb{E}_{t-1}\left[\psi_1(y_t) - \psi_0(y_t)\right]\right| \\
&= \left|\mathbb{E}_{t-1}\left[\ell_{10} + y_t(\ell_{11} - \ell_{10}) - \ell_{01}y\right]\right| \\
&= \left|\left[\ell_{10} + \mathbb{E}_{t-1}[y_t](\ell_{11} - \ell_{10}) - \ell_{01}\,\mathbb{E}_{t-1}[y_t]\right]\right| \\
&= \left|\left[\ell_{10} + \sigma(x_t^T\theta^*)(\ell_{11} - \ell_{10}) - \ell_{01}\sigma(x_t^T\theta^*)\right]\right| \\
&= \left|\left[\left(\sigma(x_t^T\theta^*) - \alpha\right)(\ell_{11} - \ell_{10}) - \ell_{01}\left(\sigma(x_t^T\theta^*) - \alpha\right)\right]\right| && \text{(definition of } \alpha) \\
&\leq 2\beta_{t-1}^{x_t}(\delta_t)\max(\ell_{10}, \ell_{01} - \ell_{11})
\end{aligned}
$$

$$
\begin{aligned}
F_{13} &= \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{\mathcal{G}_t\right\} \mathbb{1}\left\{\alpha - \beta_{t-1}^{x_t}(\delta_t) < \sigma(x_t^T \hat{\theta}_{t-1}) < \alpha + \beta_{t-1}^{x_t}(\delta_t)\right\}\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{a_t = 1\right\} \mathbb{1}\left\{-2\beta_{t-1}^{x_t}(\delta_t) \leq \sigma(x_t^T \theta^*) - \alpha \leq 2\beta_{t-1}^{x_t}(\delta_t)\right\}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{a_t = 1\right\} \mathbb{1}\left\{|\sigma(x_t^T \theta^*) - \alpha| \leq 2\beta_{t-1}^{x_t}(\delta_t)\right\}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{a_t = 1\right\} \mathbb{1}\left\{|\sigma(x_t^T \theta^*) - \alpha| \leq 2\beta_{t-1}^{x_t}(\delta_t)\right\} \mathbb{1}\left\{\beta_{t-1}^{x_t}(\delta_t) < \frac{\Delta}{2}\right\}\right] \\
&\quad + \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{a_t = 1\right\} \mathbb{1}\left\{|\sigma(x_t^T \theta^*) - \alpha| \leq 2\beta_{t-1}^{x_t}(\delta_t)\right\} \mathbb{1}\left\{\beta_{t-1}^{x_t}(\delta_t) \geq \frac{\Delta}{2}\right\}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{a_t = 1\right\} \mathbb{1}\left\{|\sigma(x_t^T \theta^*) - \alpha| < \Delta\right\}\right] \\
&\quad + \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{a_t = 1\right\} \mathbb{1}\left\{|\sigma(x_t^T \theta^*) - \alpha| \leq 2\beta_{t-1}^{x_t}(\delta_t)\right\} \mathbb{1}\left\{\beta_{t-1}^{x_t}(\delta_t) \geq \frac{\Delta}{2}\right\}\right] \\
&= 0 \qquad \text{(By definition of } \Delta\text{)} \\
&\quad + \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{a_t = 1\right\} \mathbb{1}\left\{|\sigma(x_t^T \theta^*) - \alpha| \leq 2\beta_{t-1}^{x_t}(\delta_t)\right\} \mathbb{1}\left\{\beta_{t-1}^{x_t}(\delta_t) \geq \frac{\Delta}{2}\right\}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[r_t \, \mathbb{1}\left\{a_t = 1\right\} \mathbb{1}\left\{|\sigma(x_t^T \theta^*) - \alpha| \leq 2\beta_{t-1}^{x_t}(\delta_t)\right\} \mathbb{1}\left\{\beta_{t-1}^{x_t}(\delta_t) \geq \frac{\Delta}{2}\right\}\right] \\
&\leq 2\max(\ell_{10}, \ell_{01} - \ell_{11}) \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left\{a_t = 1\right\} \beta_{t-1}^{x_t}(\delta_t) \, \mathbb{1}\left\{\beta_{t-1}^{x_t}(\delta_t) \geq \frac{\Delta}{2}\right\}\right]
\end{aligned}
$$

From Proposition 1,

$$
\begin{aligned}
\beta_{t-1}^{x_t}(\delta_t) &= \frac{2k_\sigma R_{\max}}{c_\sigma} \|x_t\|_{V_{t-1}^{-1}} \sqrt{\left(3 + 2\log\left(1 + \frac{2}{\lambda}\right)\right) \cdot 2d\log(N_{t-1}) \cdot \log\left(\frac{d}{\delta_t}\right)} \\
&\leq \frac{2k_\sigma R_{\max}}{c_\sigma} \|x_t\|_{V_{t-1}^{-1}} \sqrt{\left(3 + 2\log\left(1 + \frac{2}{\lambda}\right)\right) \cdot 2d\log T \cdot \log(dT)} \\
&:= h(T) \cdot \|x_t\|_{V_{t-1}^{-1}} \\
\beta_{t-1}^{x_t}(\delta_t) \geq \frac{\Delta}{2} &\implies \|x_t\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)}
\end{aligned}
$$

$$
\text{(Let } h(T) := \tfrac{2k_\sigma R_{\max}}{c_\sigma} \sqrt{\left(3 + 2\log\left(1 + \tfrac{2}{\lambda}\right)\right) \cdot 2d\log T \cdot \log(dT)}\text{)}
$$

$$
\begin{aligned}
F_{13} &\leq 2\max(\ell_{10}, \ell_{01}-\ell_{11}) \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left\{a_t=1\right\} \beta_{t-1}^{x_t}(\delta_t)\, \mathbb{1}\left\{\|x_t\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)}\right\}\right] \\
&\leq 2\max(\ell_{10}, \ell_{01}-\ell_{11}) h(T) \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left\{a_t=1\right\} \|x_t\|_{V_{t-1}^{-1}}\, \mathbb{1}\left\{\|x_t\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)}\right\}\right]
\end{aligned}
$$

By applying peeling method,

$$
\begin{aligned}
F_{13} &\leq 2\max(\ell_{10}, \ell_{01}-\ell_{11}) h(T) \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left\{a_t=1\right\} \|x_t\|_{V_{t-1}^{-1}}\, \mathbb{1}\left\{\|x_t\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)}\right\}\right] \\
&= 2\max(\ell_{10}, \ell_{01}-\ell_{11}) h(T) \sum_{t=1}^{T} \sum_{\ell=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left\{a_t=1\right\} \|x_t\|_{V_{t-1}^{-1}}\, \mathbb{1}\left\{2^{\ell+1}\frac{\Delta}{2h(T)} \geq \|x_t\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)} \cdot 2^{\ell}\right\}\right] \\
&\leq 2\max(\ell_{10}, \ell_{01}-\ell_{11}) h(T) \sum_{t=1}^{T} \sum_{\ell=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left\{a_t=1\right\} 2^{\ell+1}\frac{\Delta}{2h(T)}\, \mathbb{1}\left\{2^{\ell+1}\frac{\Delta}{2h(T)} \geq \|x_t\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)} \cdot 2^{\ell}\right\}\right] \\
&\leq \max(\ell_{10}, \ell_{01}-\ell_{11}) \sum_{t=1}^{T} \sum_{\ell=1}^{\infty} \mathbb{E}\left[2^{\ell+1}\Delta\, \mathbb{1}\left\{2^{\ell+1}\frac{\Delta}{2h(T)} \geq \|x_t\, \mathbb{1}\left\{a_t=1\right\}\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)} \cdot 2^{\ell}\right\}\right] \\
&= \max(\ell_{10}, \ell_{01}-\ell_{11}) \sum_{\ell=1}^{\infty} 2^{\ell+1}\Delta \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left\{2^{\ell+1}\frac{\Delta}{2h(T)} \geq \|x_t\, \mathbb{1}\left\{a_t=1\right\}\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)} \cdot 2^{\ell}\right\}\right] \\
&\leq \max(\ell_{10}, \ell_{01}-\ell_{11}) \sum_{\ell=1}^{\infty} 2^{\ell+1}\Delta \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left\{\|x_t\, \mathbb{1}\left\{a_t=1\right\}\|_{V_{t-1}^{-1}} \geq \frac{\Delta}{2h(T)} \cdot 2^{\ell}\right\}\right] \\
&= \max(\ell_{10}, \ell_{01}-\ell_{11}) \sum_{\ell=1}^{\infty} 2^{\ell+1}\Delta\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\|x_t\, \mathbb{1}\left\{a_t=1\right\}\|^2_{V_{t-1}^{-1}} \geq \frac{\Delta^2}{4h^2(T)} \cdot 2^{2\ell}\right\}\right] \\
&\leq \max(\ell_{10}, \ell_{01}-\ell_{11}) \sum_{\ell=1}^{\infty} 2^{\ell+1}\Delta\, \mathbb{E}\left[\frac{12dh^2(T)}{2^{2\ell}\Delta^2} \log\left(1+\frac{8h^2(T)}{\lambda 2^{2\ell}\Delta^2}\right)\right] \qquad \text{(Lemma 9 (EPC))} \\
&\leq \frac{1}{\Delta} 24dh^2(T) \max(\ell_{10}, \ell_{01}-\ell_{11}) d \log\left(1+\frac{8h^2(T)}{\lambda 4\Delta^2}\right) \sum_{\ell=1}^{\infty} 2^{-\ell} \\
&\leq \frac{1}{\Delta} 24dh^2(T) \max(\ell_{10}, \ell_{01}-\ell_{11}) d \log\left(1+\frac{8h^2(T)}{\lambda 4\Delta^2}\right) \\
&\qquad \text{(here } h(T) = \tfrac{2k_\sigma R_{\max}}{c_\sigma}\sqrt{\left(3+2\log\left(1+\tfrac{2}{\lambda}\right)\right) \cdot 2d\log T \cdot \log(dT)}\text{)}
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\mathbb{E}\left[Reg_T\right] &\leq \frac{1}{\Delta} 24dh^2(T) \max(\ell_{10}, \ell_{01}-\ell_{11}) d \log\left(1+\frac{8h^2(T)}{\lambda 4\Delta^2}\right) + O(\log(T)) \\
&\qquad \text{(here } h(T) = \tfrac{2k_\sigma R_{\max}}{c_\sigma}\sqrt{\left(3+2\log\left(1+\tfrac{2}{\lambda}\right)\right) \cdot 2d\log T \cdot \log(dT)}\text{)} \\
&= O\left(\frac{1}{\Delta} d^2 \log^2 T\right) \qquad \text{(ignoring doubly logarithmic factors and } \mathrm{polylog}(\tfrac{1}{\Delta})\text{)}
\end{aligned}
$$

D Experiment Details

We call our framework AppleTeA: Test-Time Adaptive LLM Cascading via Logistic Apple Tasting, which is presented in Algorithm D

Algorithm 4 AppleTeA

  Input: $\mathcal{M}_s$, $\mathcal{M}_\ell$, n, ϕ, α, γ, λ, μ, algorithm Alg
  if Alg = "LogCB-AT" **then**
      Instantiate $\mathcal{A}$ = LogCB-AT (γ, μ)
  else
      Instantiate $\mathcal{A}$ = LogCBPSide-AT (λ, α)
  end if
  for $t = 1, \cdots, n$ **do**
      Observe the query q_t
      Run SLM $r_t = \mathcal{M}_s(q_t)$
      Extract the answer: $\hat{y}_{w,t} = \text{answer}(r_t)$
      Compute the context: $x_t = \phi(q_t, r_t)$
      Run single step of algorithm $\mathcal{A}$ and receive the action recommendation a_t
      if a_t = 0 (Reject) **then**
          output $\hat{y}_{w,t}$
      else
          Run the strong LLM: $r_t = \mathcal{M}_\ell(q_t)$
          Extract the answer: $\hat{y}_{s,t} = \text{answer}(r_t)$
          Compute the context: $x_t = \phi(q_t, r_t)$
          Update $\mathcal{A}$ with context x_t and reward = $\mathbb{1}\left\{\hat{y}_{s,t} \neq \hat{y}_{w,t}\right\}$
          output $\hat{y}_{s,t}$
      end if
  end for

We have established that both LogCB-AT and LogCBPSide-AT attain sub-linear regret with respect to the optimal policy. This regret bound directly translates into a cumulative cost guarantee for AppleTeA. By carefully constructing the loss matrix to encode two competing cost components 1) the invocation cost of the strong LLM (i.e., token-level inference cost) and 2) the price of error incurred when the weak SLM is invoked in contexts where prediction failure is likely; AppleTeA is equipped to navigate this cost-quality tradeoff in a principled manner. Consequently, AppleTeA achieves optimal performance in expectation, a guarantee that follows directly from the formulation of the model-selection problem as an Apple Tasting problem and the sub-linear regret guarantees of the underlying core algorithms.

Context Construction. To enable effective decision-making, we construct the context vector $x_t \in \mathbb{R}^d$ by capturing the internal state and the uncertainty score of the SLM model. Specifically, x_t is formed by concatenating the dimensionally-reduced hidden state vector of the final token from $\mathcal{M}_s$ and its generative uncertainty score. The uncertainty score can be quantified using either the maximum probability of the generated tokens or the logit margin between the top two candidate answers. This context provides a rich signal for the the proposed methods to estimate the probability of the LLM model altering the final answer.

Datasets and Models. We adopt Pythia-2.8B [Biderman et al., 2023] and LLaMA2-13B [Touvron et al., 2023] as $\mathcal{M}_s$ and $\mathcal{M}_l$, respectively. We evaluate the effectiveness of our algorithms on the commonsense reasoning benchmark: ARC-Easy [Clark et al., 2018], OpenBookQA [Mihaylov et al., 2018] and BoolQ [Sakaguchi et al., 2021]

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 clearly state the contributions, scope, and assumptions 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 address the limitations of our approach in the Conclusion.

   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: All assumptions are explicitly stated in the main text and complete proofs are provided in the Appendix.

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: All experimental details are in Experiments section.

   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: The source code and datasets are provided in the supplementary material.

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: The core experimental settings are described in Experiments section.

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: They are included in the experimental results.

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: These details are provided in the Experiments section.

   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 this paper strictly conforms to the NeurIPS 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: [N/A]

    Justification: This work focuses on foundational theoretical algorithms and has no direct societal impact.

    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: This paper focuses on theoretical algorithms and does not release models or datasets with a high risk of misuse.

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 appropriately cite all existing datasets and code used in our experiments and adhere to their respective licenses.

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: The code for the our proposed algorithms are provided in the 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: This research does not involve crowdsourcing or human subjects.

    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: This study does not involve human subjects or IRB-related research.

    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: LLMs were not used as a core component of the methodology in this research.

    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.

References

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), volume 54, pages 176–184, 2017.

Kapilan Balagopalan and Kwang-Sung Jun. Minimum empirical divergence for sub-gaussian linear bandits. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, Proceedings of The 28th International Conference on Artificial Intelligence and Statistics, volume 258 of Proceedings of Machine Learning Research, pages 1585–1593. PMLR, 03–05 May 2025.

Gábor Bartók and Csaba Szepesvári. Partial monitoring with side information. In International Conference on Algorithmic Learning Theory, pages 305–319. Springer, 2012.

Gábor Bartók, Navid Zolghadr, and Csaba Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. ICML'12, page 1779–1786, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. Mathematics of Operations Research, 39(4): 967–997, 2014.

Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, volume 151 of Proceedings of Machine Learning Research, pages 54–72. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/bian22a.html.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In International conference on machine learning, pages 2397–2430. PMLR, 2023.

Zachary Chase and Idan Mehalel. Deterministic apple tasting. arXiv preprint arXiv:2410.10404, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In Machine learning for healthcare conference, pages 67–82. PMLR, 2018.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric Bandits: The Generalized Linear Case. In Advances in Neural Information Processing Systems (NeurIPS), pages 586–594, 2010.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In International conference on machine learning, pages 3199–3210. PMLR, 2020.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. Advances in Neural Information Processing Systems, 34:18907–18919, 2021.

James A Grant and David S Leslie. Apple tasting revisited: Bayesian approaches to partially monitored online binary classification. arXiv preprint arXiv:2109.14412, 2021.

Keegan Harris, Chara Podimata, and Steven Z Wu. Strategic apple tasting. Advances in Neural Information Processing Systems, 36:79918–79945, 2023.

David P Helmbold, Nicholas Littlestone, and Philip M Long. Apple tasting. Information and Computation, 161(2):85–139, 2000.

Maxime Heuillet, Ola Ahmad, and Audrey Durand. Randomized confidence bounds for stochastic partial monitoring. arXiv preprint arXiv:2402.05002, 2024.

Kwang-Sung Jun and Jungtaek Kim. Noise-adaptive confidence sets for linear bandits and application to bayesian optimization. Proceedings of the International Conference on Machine Learning (ICML), 2024.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In Conference on Learning Theory, pages 2328–2369. PMLR, 2020.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring.

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

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Proceedings of the 2018 conference on empirical methods in natural language processing, pages 2381–2391, 2018.

Vinod Raman, Unique Subedi, Ananth Raman, and Ambuj Tewari. Apple tasting: Combinatorial dimensions and minimax rates. In The Thirty Seventh Annual Conference on Learning Theory, pages 4358–4380. PMLR, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM, 64(9):99–106, 2021.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In Mobile health: sensors, analytic methods, and applications, pages 495–517. Springer, 2017.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

Taira Tsuchiya, Junya Honda, and Masashi Sugiyama. Analysis and design of thompson sampling for stochastic partial monitoring. Advances in Neural Information Processing Systems, 33:8861–8871, 2020.

Mengxiao Zhang, Yuheng Zhang, Olga Vrousgou, Haipeng Luo, and Paul Mineiro. Practical contextual bandits with feedback graphs. Advances in Neural Information Processing Systems, 36:30592–30617, 2023.