Near-Optimal Learning in Parametric Bandits with Action-Dependent Coarsened Feedback

Zhuohua Li[1] **Maoli Liu**[2*] **Yuwen Huang**[3] **Cheng Wen**[1]
Jie Su[1] **Cong Tian**[1] **Shengchao Qin**[1] **John C.S. Lui**[2]

[1]Xidian University [2]The Chinese University of Hong Kong
[3]The Hong Kong University of Science and Technology (Guangzhou)

Abstract

We study a structured stochastic multi-armed bandit problem in which each decision simultaneously determines the learner's reward and the feedback available for learning. At each round, the learner selects a *task* and an *option*. Each task is associated with an unknown parameter that governs its reward, while each option has a known reward function of this parameter and determines how information about that reward is observed. Some options may provide fine-grained feedback at the cost of lower immediate reward, whereas others may yield higher reward but return only coarse feedback. This creates an intrinsic trade-off between reward maximization and information acquisition. We formalize this setting as *parametric bandits with action-dependent coarsened feedback*. For stationary environments, we propose `ACF-UCB`, an optimistic algorithm that combines feedback from different options to construct a confidence set for each task. For piecewise-stationary tasks, we develop `SW-ACF-UCB`, a sliding-window variant that adapts to temporal changes. We prove high-probability regret bounds for stationary environments and dynamic regret guarantees for piecewise-stationary environments. We also establish lower bounds showing that the dependence on the number of tasks and the horizon is unavoidable, while methods that ignore the shared parametric structure suffer substantially larger regret.

1 Introduction

In many sequential decision-making problems, an action affects not only the reward obtained by the learner, but also *how information about the environment is revealed*. For example, in recommender systems, a platform may choose not only which item to recommend, but also how to elicit or observe user preference feedback: lightweight interaction signals, such as clicks or skips, are easy to collect but provide only coarse preference information, whereas more costly elicitation protocols can be more informative but may reduce short-term engagement. This creates a trade-off between exploiting currently preferred items and exploring through feedback that differs in cost and informativeness.

Motivated by this phenomenon, we study a new structured multi-armed bandit (MAB) problem in which each pull consists of choosing a *task* to serve and an *option* to use. Each task i is associated with an unknown scalar parameter θ_i that governs the rewards of all options for that task. Each option j, in turn, determines both the reward function applied to θ_i and the form of feedback returned to the learner. Unlike in the standard MAB model, the feedback is not necessarily a noisy sample of the reward. Rather, it is a coarsened observation whose informativeness and cost depend on the selected option. Usually, an option with a high immediate reward may provide weak feedback, while

*Maoli Liu is the corresponding author.

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

a lower-reward option may provide more informative observations. Therefore, the learner must balance immediate reward maximization with information acquisition.

This reward-information coupling appears in several real-world applications. For example, in adaptive inspection, an inspection protocol affects both throughput and how defect counts are reported. In adaptive sensing, a sensing mode affects both utility and measurement resolution. A particularly concrete example is online Forward-Error-Correcting (FEC) code selection. A sender repeatedly chooses a channel and a code, while the unknown channel error rate determines both which code maximizes throughput and which feedback is observed. We formalize this example in Section 3.2.

Bandit formulation. We model the above setting as a parametric bandit with action-dependent coarsened feedback. The learner is given K tasks and M options. At each round $t = 1, \dots, T$, the learner selects a task-option pair (i, j): the task i specifies the unknown environment to be served, and the option j specifies both how service is performed and how feedback is observed. Each task i has an unknown scalar parameter θ_i. When option j is applied to task i, the system receives a reward with mean $\mu(\theta_i, j)$, which need not be directly observed by the learner, and instead, the learner only observes a categorical signal $Y \sim g_j(\cdot; \theta_i)$. The goal is to minimize regret against an oracle that knows all task parameters $\{\theta_i\}_{i=1}^K$ and always selects the best task-option pair.

Why existing bandit algorithms fall short. A direct reduction to a standard KM-armed bandit [Auer et al., 2002] treats each task-option pair as an independent arm. Such a reduction ignores the fact that, within each task, all options share the same latent parameter. As a result, it explores in KM unrelated directions rather than using feedback from one option to inform the value of the others. Generalized-linear bandits [Filippi et al., 2010] exploit shared parametric reward structure, while the structured bandit framework of Combes et al. [2017] also allows rewards to remain unobserved and observation laws to differ from reward functions. Threshold and censored-feedback bandits [Abernethy et al., 2016, Bengs and Hüllermeier, 2022] are closer in spirit, but their feedback mechanisms are usually tied to thresholds, resource limits, or censored rewards, rather than to a family of option-dependent categorical experiments sharing a common task parameter.

Our approach. We introduce *Parametric Bandits with Action-Dependent Coarsened Feedback*, a framework in which each action is both a reward-generating decision and a statistical experiment. The main algorithmic idea is to construct confidence sets for each task parameter using the observed categorical feedback laws and a task-level constraint, and then choose task-option pairs optimistically.

Our analysis distinguishes two regimes. The favorable regime admits bounded statistics whose means vary in the same direction with the parameter and control reward differences for every option (Assumption 2). In this case, the task-level constraint controls reward uncertainty using all observations of a task. When the favorable condition does not hold, our algorithms still admit an alternative analysis based on a categorical identifiability constant Γ, which converts uncertainty in the feedback law into uncertainty about the underlying parameter (Assumption 3).

We consider both the stationary setting and the piecewise-stationary setting. For stationary tasks, we propose `ACF-UCB` (Section 4), which maintains coordinate-wise confidence constraints together with the task-level constraint and then plays the most optimistic parameter in the resulting set. For piecewise-stationary tasks, `SW-ACF-UCB` (Section 4.2) applies the same construction on sliding windows. Our regret analysis shows that, under Assumption 2, the shared parametric structure yields a $\widetilde{O}(\sqrt{KT})$ stationary regret bound, while lower bounds show that this dependence is unavoidable.

Contributions. We summarize our contributions as follows.

- We introduce *Parametric Bandits with Action-Dependent Coarsened Feedback*, in which each action simultaneously specifies a reward function and a coarsened categorical feedback law.
- We design two optimistic algorithms for this model. For stationary tasks, `ACF-UCB` constructs task-level confidence sets using the action-dependent categorical feedback laws and statistics. For piecewise-stationary tasks, `SW-ACF-UCB` applies the same principle over sliding windows to adapt to parameter changes.
- We develop two complementary regret analyses that connect coarsened categorical feedback to reward learning. Under Assumption 2, we obtain reward-sensitive bounds by controlling reward uncertainty directly through constraints on statistic means. Under an alternative condition (Assumption 3), we provide a fallback analysis that converts categorical confidence into parameter uncertainty via the identifiability constant Γ.

- We prove lower bounds showing that the $\sqrt{KT}$ dependence is unavoidable, while unstructured KM-armed methods that ignore the shared task parameter face an $\Omega(\sqrt{KMT})$ regret.
- We evaluate the algorithms on general instances and on the FEC specialization, and the results show that our algorithm outperforms existing baselines.

2 Related Work

Contextual bandits and applications. In standard contextual bandits, the learner observes contextual information before choosing an action and learns from the selected action's reward [Chu et al., 2011, Abbasi-Yadkori et al., 2011]. Applications include personalized recommendation [Li et al., 2010], large language model (LLM) response selection [Dai et al., 2026], federated learning and online user clustering [Gentile et al., 2014, Li et al., 2024, 2025a], as well as information acquisition for conversational recommendation [Li et al., 2025b, Liu et al., 2025]. Combinatorial extensions with linear generalization also support network routing through path selection [Wen et al., 2015]

Structured bandits. Filippi et al. [2010] study generalized-linear parametric bandits, where arms share reward parameters but observations remain reward-linked. Global and regional bandits learn shared parameters from reward observations across all arms or within groups [Atan et al., 2015, Wang et al., 2018]. Combes et al. [2017] study structured stochastic bandits with minimal exploration and allow rewards to remain unobserved. Confidence-based and structure-adaptive algorithms further exploit known reward structure [Tirinzoni et al., 2020, Degenne et al., 2020]. Lattimore and Szepesvári [2020] give a broader treatment of bandit models with shared structure and classical information constraints. These works are closest in their use of a common structure across arms. However, they typically observe rewards or reward-linked samples rather than action-dependent coarsened categorical laws. In our model, an action is simultaneously a reward decision and an experiment whose observation law may differ from the reward functional. This distinction is what makes confidence sets on $g_j(\cdot;\theta)$ useful.

Nonstandard feedback. Abernethy et al. [2016] study threshold bandits with censored feedback, where feedback is tied to thresholding rather than to a learner-chosen categorical experiment. Bengs and Hüllermeier [2022] study bandits with censored resource consumption, where censoring concerns resource observations rather than a shared task parameter. Censored semi-bandits also study resource-dependent censoring [Verma et al., 2019]. Cesa-Bianchi et al. [2006] and Bartók et al. [2014] study finite partial-monitoring games, where feedback depends on both the selected action and the outcome. Linear and kernelized partial monitoring further separate action-dependent observations from rewards [Kirschner et al., 2023]. Alon et al. [2015] study feedback graphs, where observability is determined by a graph rather than by a task-specific latent parameter. These models allow feedback to differ from direct reward observations. Our task-option model specifies known categorical feedback laws and reward functions that share a scalar task parameter. This lets feedback from one option improve the value estimates of other options.

Non-stationarity and exploration principles. Sliding-window and discounted UCB policies handle switching reward distributions [Garivier and Moulines, 2011], while other methods use variation budgets [Besbes et al., 2014] or adapt to an unknown number of changes [Auer et al., 2019]. Weighted and forgetting-based estimators extend these ideas to linear and generalized-linear bandits [Russac et al., 2019, 2021, Wang et al., 2023]. Regional bandits also admit sliding-window policies for slowly varying group parameters [Wang et al., 2018], while adaptive restart reductions adapt to unknown degrees of non-stationarity [Wei and Luo, 2021]. Following the sliding-window principle, `SW-ACF-UCB` constructs task-level confidence sets from option-dependent categorical feedback. Bayesian UCB and Thompson sampling offer alternative exploration strategies [Kaufmann et al., 2012, Agrawal and Goyal, 2012], while information-directed sampling explicitly balances regret and information gain [Russo and Van Roy, 2014, Kirschner and Krause, 2018].

3 Problem Setting

In this section, we introduce the general model of Parametric Bandits with Action-Dependent Coarsened Feedback. We first state the favorable regime (Assumption 2), then give a fallback regime (Assumption 3) based on global identifiability. The FEC application satisfies the favorable condition, and the fallback condition provides a guarantee when the favorable condition is unavailable.

3.1 General Action-Dependent Coarsened Feedback Model

There are K tasks and M options. Each task $i \in [K]$ has an *unknown* scalar parameter $\theta_i \in \Theta$, where $\Theta = [\theta_{\min}, \theta_{\max}] \subset \mathbb{R}$ is compact. At each round t, the learner chooses a task $I_t \in [K]$ and an option $J_t \in [M]$. For each option j, let $\mathcal{Y}_j$ be its feedback alphabet. When option j is used on a task with parameter θ, the learner observes a categorical feedback signal Y_t taking values in $\mathcal{Y}_j$. We denote the corresponding feedback law by $g_j(\cdot;\theta)$, so that

$$\Pr(Y_t = y \mid I_t = i, J_t = j) = g_j(y;\theta_i), \qquad y \in \mathcal{Y}_j.$$

The performance of option j at parameter θ is measured by a mean reward function $\mu(\theta, j)$. The algorithm does not require observing reward samples. Instead, learning is driven by the categorical feedback Y_t. The best option value at parameter θ is $v(\theta) := \max_{j\in[M]} \mu(\theta, j)$. The learner knows the parameter space Θ, the feedback alphabets $\{\mathcal{Y}_j\}_{j=1}^M$, the feedback laws $\{g_j(\cdot;\theta)\}_{j=1}^M$, and the mean reward functions $\{\mu(\theta, j)\}_{j=1}^M$. The learner does not know the task parameters $\theta_1, \dots, \theta_K$. We write $D_{\mathcal{Y}} := \sum_{j=1}^M |\mathcal{Y}_j|$ for the total number of feedback outcomes across all options.

Assumption 1 (Basic regularity)**.** The parameter space $\Theta \subset \mathbb{R}$ is compact. For every option j, the feedback alphabet $\mathcal{Y}_j$ is finite. For every option j and feedback outcome $y \in \mathcal{Y}_j$, the map $\theta \mapsto g_j(y;\theta)$ is continuous on Θ. For every option j, the reward function satisfies $0 \le \mu(\theta, j) \le 1$ on Θ. Conditional on the pre-observation history and the selected pair $(I_t, J_t) = (i, j)$, the feedback law is $g_j(\cdot;\theta_i)$ in the stationary setting and $g_j(\cdot;\theta_i(t))$ in the non-stationary setting.

We consider two settings in terms of the stationarity of unknown parameters.

Stationary setting: In the stationary setting, each θ_i is fixed over time. The benchmark is an oracle that knows all task parameters and always chooses the best task and option. The regret is defined as

$$R_T = \sum_{t=1}^{T} [\mu^* - \mu(\theta_{I_t}, J_t)], \qquad \mu^* := \max_{i,j} \mu(\theta_i, j). \tag{1}$$

Piecewise stationary setting: We also consider a non-stationary setting in which task parameters may change over time. Let $\theta_i(t) \in \Theta$ denote the parameter of task i at round t. Each sequence $\{\theta_i(t)\}_{t=1}^T$ is piecewise constant with S_i change points. The total number of change points is $S := \sum_{i=1}^K S_i$. The benchmark is an oracle that knows the current parameters and chooses the best task and option at each round. The dynamic regret is

$$R_T^{\mathrm{dyn}} = \sum_{t=1}^{T} \left[\max_{i\in[K],\, j\in[M]} \mu(\theta_i(t), j) - \mu(\theta_{I_t}(t), J_t) \right]. \tag{2}$$

Based on different applications, we use two structural conditions in the analysis. The first condition is the favorable case, which enables information sharing across options by comparing reward differences with differences in statistic means.

Assumption 2 (Monotonicity and reward transfer)**.** For each option a, there is a known statistic $h_a : \mathcal{Y}_a \to [0, 1]$, with mean $m_a(\theta) := \sum_{y\in\mathcal{Y}_a} h_a(y) g_a(y;\theta)$. The functions m_a are either all nondecreasing or all nonincreasing on Θ. There exists a constant $B > 0$ such that for all $\theta_1, \theta_2 \in \Theta$ and all source and reward options $i, j \in [M]$,

$$|\mu(\theta_1, j) - \mu(\theta_2, j)| \le B |m_i(\theta_1) - m_i(\theta_2)|.$$

Remark 1. Assumption 2 ensures that feedback from any option can be used to control reward uncertainty for every option of the same task. Common monotonicity prevents cancellation when differences in statistic means are combined across options. This assumption is natural for the FEC application. For codes with the same block length, the event that the error count does not exceed the smallest correction radius can be determined from the feedback of every code. Its indicator has the same monotone mean for all codes, and differences in this mean bound the reward differences up to a constant factor, as verified in Proposition 9.

When Assumption 2 does not hold, we use the following fallback condition. It converts confidence in feedback distributions into confidence in the unknown scalar parameter, then uses reward Lipschitzness to control regret.

Assumption 3 (Identifiability and Lipschitz rewards)**.** The following hold.

(i) *Uniform identifiability.* The categorical-feedback modulus

$$\Gamma := \min_{j\in[M]} \inf_{\substack{\theta,\theta'\in\Theta \\ \theta\neq\theta'}} \frac{\|g_j(\cdot;\theta) - g_j(\cdot;\theta')\|_2}{|\theta - \theta'|}$$

is strictly positive.

(ii) *Lipschitz rewards.* There exists $L < \infty$ such that

$$|\mu(\theta, j) - \mu(\theta', j)| \le L|\theta - \theta'| \qquad \forall \theta, \theta' \in \Theta,\ j \in [M].$$

Compared with Assumption 2, Assumption 3 provides an alternative guarantee, but it can lead to conservative constants when Γ is small.

3.2 FEC Code Selection as an Instance

Online FEC code selection is our main example. A sender transmits over multiple channels whose error rates are unknown. Each round chooses both a channel and a code, so the decision controls both the transmission reward and the feedback about the channel. A code with parameters (n, k, d) transmits k information bits in n physical bits and corrects up to $t = \lfloor (d-1)/2 \rfloor$ bit errors. This follows the classical error-correcting code premise that redundancy can detect and correct channel errors [Hamming, 1950]. High-rate codes are efficient when errors are rare, while high-distance codes are safer when errors are common. Table 1 illustrates this reward-information tradeoff for two simple codes. The $(10, 4, 3)$ code has higher rate and is better at $p = 0.05$, while the lower-rate $(10, 2, 5)$ code becomes better at $p = 0.25$ because it corrects more errors. The selected code also determines the feedback alphabet, because small error counts are reported exactly and large counts are censored.

Table 1: Two-code FEC example. The reward is the expected code rate, equal to the code rate multiplied by decoding success probability.

Code	Code Rate	Corrects	Reward at error rate $p = 0.05$	Reward at error rate $p = 0.25$
$(10, 4, 3)$	0.4	$t = 1$	0.366	0.098
$(10, 2, 5)$	0.2	$t = 2$	0.198	0.105

In the model, each task is a binary symmetric channel with unknown error rate $p_i \in [0, 1/2]$, and each action is a code with parameters (n_j, k_j, d_j) and correction radius $t_j = \lfloor (d_j - 1)/2 \rfloor$. When code j is used on a channel with error rate p, the latent error count is $E \sim \mathrm{Bin}(n_j, p)$, and the learner observes E if $E < d_j$ and a censorship symbol $\perp$ otherwise. The expected reward is

$$\mu(p, j) = \frac{k_j}{n_j} F_{\mathrm{Bin}}(t_j; n_j, p). \tag{3}$$

The FEC instance satisfies Assumptions 1, 2, and 3. For equal block lengths, a statistic shared by all codes is $h_j(Y) = \mathbb{1}\{E \le \min_a t_a\}$, an event visible under every code. Appendix A gives the model dictionary, the formal feedback law, and the verification of the assumptions.

4 Algorithms

Both algorithms are stated for the general model and apply to FEC after substituting the FEC feedback law and reward function. They build confidence sets for $g_j(\cdot;\theta)$, which control selected rewards under Assumption 2 and parameter error under Assumption 3. When only the fallback condition is available, set $h_j := 0$ and $m_j := 0$, making the task-level constraint vacuous.

For confidence level $\alpha \in (0, 1)$, define the anytime logarithmic factors

$$\beta_{\mathrm{st}}(t, \alpha) := \log\left(\frac{\pi^2 K(D_{\mathcal{Y}} + 1)t^2}{3\alpha}\right), \qquad \beta_{\mathrm{sw}}(t, \alpha) := \log\left(\frac{\pi^2 K(D_{\mathcal{Y}} + 1)t^3}{3\alpha}\right). \tag{4}$$

4.1 Stationary Algorithm: ACF-UCB

Write $N_i(t) := \sum_{s<t} \mathbb{1}\{I_s = i\}$. The stationary categorical confidence set is

$$\mathcal{C}_i(t) := \Bigg\{ \theta \in \Theta : \Bigg| \sum_{s<t: I_s=i, J_s=j} (\mathbb{1}\{Y_s = y\} - g_j(y;\theta)) \Bigg| \leq \sqrt{2\beta_{\mathrm{st}}(t,\alpha) \sum_{s<t} \mathbb{1}\{I_s = i, J_s = j\}} \quad \forall j \in [M],\ y \in \mathcal{Y}_j, \\ \Bigg| \sum_{s<t: I_s=i} (h_{J_s}(Y_s) - m_{J_s}(\theta)) \Bigg| \leq \sqrt{2N_i(t)\beta_{\mathrm{st}}(t,\alpha)} \Bigg\}. \tag{5}$$

Algorithm 1: ACF-UCB: Action-Dependent Coarsened-Feedback UCB

Input: Tasks $[K]$, actions $[M]$, horizon T, confidence α

1 **Initialize:** for $t = 1, \ldots, \min\{K, T\}$, play task t once using any fixed action j_0;
2 **for** $t = K+1$ **to** T **do**
3 **for** $i \in [K]$ **do**
4 Construct $\mathcal{C}_i(t)$ as in (5). If it is empty, set $\mathcal{C}_i(t) \leftarrow \Theta$;
5 Choose $\tilde{\theta}_i(t) \in \arg\max_{\theta \in \mathcal{C}_i(t)} v(\theta)$;
6 $U_i(t) \leftarrow v(\tilde{\theta}_i(t))$;
7 $I_t \in \arg\max_i U_i(t)$;
8 $J_t \in \arg\max_j \mu(\tilde{\theta}_{I_t}(t), j)$;
9 Play (I_t, J_t) and observe Y_t;

Algorithm 1 first initializes each task (Line 1), constructs $\mathcal{C}_i(t)$ (Line 4), selects the parameter with largest optimistic value (Line 5), forms the task upper confidence value (Line 6), selects the task with largest upper confidence value (Line 7), and plays the reward-optimal action at the selected optimistic parameter (Line 8). It then observes the coarsened feedback used in the next confidence update (Line 9). The confidence set directly bounds reward differences under Assumption 2. Under Assumption 3, it bounds parameter differences, which are then converted into reward bounds using Lipschitz continuity.

4.2 Sliding-Window Algorithm: SW-ACF-UCB

For the non-stationary setting, define the window $\mathcal{W}(t) := \{s : \max(1, t - W) \leq s < t\}$. Write $N_i^W(t) := \sum_{s \in \mathcal{W}(t)} \mathbb{1}\{I_s = i\}$. The sliding-window confidence set is

$$\mathcal{C}_i^W(t) := \Bigg\{ \theta \in \Theta : \Bigg| \sum_{s \in \mathcal{W}(t): I_s=i, J_s=j} (\mathbb{1}\{Y_s = y\} - g_j(y;\theta)) \Bigg| \leq \sqrt{2\beta_{\mathrm{sw}}(t,\alpha) \sum_{s \in \mathcal{W}(t)} \mathbb{1}\{I_s = i, J_s = j\}} \quad \forall j \in [M],\ y \in \mathcal{Y}_j, \\ \Bigg| \sum_{s \in \mathcal{W}(t): I_s=i} (h_{J_s}(Y_s) - m_{J_s}(\theta)) \Bigg| \leq \sqrt{2N_i^W(t)\beta_{\mathrm{sw}}(t,\alpha)} \Bigg\}. \tag{6}$$

Algorithm 2 assigns infinite priority to tasks with no recent observations (Line 4), constructs the sliding-window confidence set $\mathcal{C}_i^W(t)$ (Line 6), selects the optimistic parameter in that set (Line 7), and forms the task upper confidence value (Line 8). It then selects the task with largest upper confidence value (Line 9), refreshes an empty-window task with a fixed exploratory action (Line 11), otherwise chooses the reward-optimal action at the selected optimistic parameter (Line 13), and observes the resulting coarsened feedback (Line 14).

Algorithm 2: SW-ACF-UCB: Sliding-Window Action-Dependent Coarsened-Feedback UCB

Input: Tasks $[K]$, actions $[M]$, horizon T, window W, confidence α

for $t = 1$ **to** T **do**
  for $i \in [K]$ **do**
    if *no observation of task i lies in $\mathcal{W}(t)$* **then**
      $U_i(t) \leftarrow +\infty$;
    else
      Construct $\mathcal{C}_i^W(t)$. If it is empty, set $\mathcal{C}_i^W(t) \leftarrow \Theta$;
      Choose $\tilde{\theta}_i^W(t) \in \arg\max_{\theta \in \mathcal{C}_i^W(t)} v(\theta)$;
      $U_i(t) \leftarrow v(\tilde{\theta}_i^W(t))$;
  $I_t \in \arg\max_i U_i(t)$;
  if *no observation of task I_t lies in $\mathcal{W}(t)$* **then**
    Play (I_t, j_0) using any fixed action j_0 and observe Y_t;
  else
    $J_t \in \arg\max_j \mu(\tilde{\theta}_{I_t}^W(t), j)$;
    Play (I_t, J_t) and observe Y_t;

5 Theoretical Analysis

5.1 Main Reward-Sensitive Bounds

The main regret guarantees apply in the favorable regime of Assumption 2. They avoid converting categorical confidence into parameter confidence and therefore do not depend on $1/\Gamma$. Common monotonicity lets the task-level constraint bound the sum of absolute differences in statistic means, and reward transfer then gives a reward width of $B\sqrt{8\beta_{\mathrm{st}}(t,\alpha)/N_i(t)}$ when $N_i(t) > 0$. The denominator counts all observations of the task, regardless of the selected option. The dependence on the total feedback alphabet size $D_{\mathcal{Y}}$ is logarithmic.

Theorem 1 (Reward-sensitive stationary regret of ACF-UCB)**.** *Under Assumptions 1 and 2, for any stationary instance and for every $\alpha \in (0,1)$, Algorithm 1 satisfies, with probability at least $1-\alpha$,*

$$R_T \le K + 4\sqrt{2}\, B\sqrt{KT\beta_{\mathrm{st}}(T,\alpha)}.$$

For the piecewise-stationary setting, call a pair (i,t) *clean* if all observations of task i in $\mathcal{W}(t)$ were generated by the current parameter $\theta_i(t)$. The sliding-window confidence lemma in Appendix D applies only to clean windows. Contaminated rounds are charged separately. The resulting dynamic regret has three sources: contaminated windows after changes, forced refreshes of empty task windows, and ordinary confidence width on clean windows.

Theorem 2 (Reward-sensitive dynamic regret of SW-ACF-UCB)**.** *Under Assumptions 1 and 2, consider Algorithm 2 with window $W \in [T]$. For every $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$R_T^{\mathrm{dyn}} \le\ S(W+1) + K\left(1 + \frac{T}{W}\right) + B\sqrt{8\beta_{\mathrm{sw}}(T,\alpha)}\left\lceil \frac{T}{W} \right\rceil \left(K + 2\sqrt{KW}\right).$$

The proofs of Theorems 1 and 2 are in Appendices C.2 and D.2, respectively.

5.2 Fallback Γ-Based Bounds

When Assumption 2 is unavailable, the same algorithms still admit regret guarantees under Assumption 3. The constants in these bounds depend on the parameter-identification cost L/Γ. The fallback proof first converts confidence bounds for the feedback probabilities into a radius for the latent scalar parameter. Reward error is then controlled by Lipschitzness of $\mu(\cdot, j)$. This route is robust to non-linear reward functions, but it pays for the least informative option through the minimum in Γ.

Theorem 3 (Stationary regret of ACF-UCB)**.** *Under Assumptions 1 and 3, fix any stationary instance with $\theta_i \in \Theta$. For every $\alpha \in (0,1)$, Algorithm 1 satisfies, with probability at least $1-\alpha$,*

$$R_T \le K + 4\sqrt{2D_{\mathcal{Y}}}\,\frac{L}{\Gamma}\sqrt{KT\beta_{\mathrm{st}}(T,\alpha)}.$$

In particular, setting $\alpha = 1/T$ gives

$$\mathbb{E}[R_T] = O\left(K + \frac{L\sqrt{D_{\mathcal{Y}}}}{\Gamma}\sqrt{KT\log(KD_{\mathcal{Y}}T)}\right)$$

up to universal logarithmic constants.

Theorem 4 (Dynamic regret of `SW-ACF-UCB`)**.** *Under Assumptions 1 and 3, consider Algorithm 2 with window $W \in [T]$. For every $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$R_T^{\mathrm{dyn}} \le S(W+1) + K\left(1+\frac{T}{W}\right) + 2\sqrt{2D_{\mathcal{Y}}}\,\frac{L}{\Gamma}\sqrt{\beta_{\mathrm{sw}}(T,\alpha)}\left\lceil\frac{T}{W}\right\rceil\left(K+2\sqrt{KW}\right).$$

The proofs of Theorems 3 and 4 are in Appendices C.4 and D.4, respectively.

6 Lower Bounds

We show that the $\sqrt{KT}$ dependence is unavoidable, even in a two-action instance with an explicit reward-information tradeoff. The first lower bound embeds the ordinary K-armed Bernoulli bandit as a one-option subclass. The second lower bound shows that the same scaling persists when the learner has a more informative but lower-reward action available. The unstructured lower bound then isolates the additional price paid by methods that regard each task-option pair as an unrelated arm.

Theorem 5 (General minimax lower bound)**.** *There exists an instance of the model satisfying Assumptions 1, 2, and 3 such that, for every algorithm and every $T \ge K \ge 2$,*

$$\sup_{\theta_1,\dots,\theta_K\in\Theta}\mathbb{E}[R_T] \ge c\sqrt{KT},$$

where $c > 0$ is a universal numerical constant.

Theorem 6 (Two-action lower bound)**.** *There exists a model with two actions and binary coarsened feedback, satisfying Assumptions 1, 2, and 3, such that the following holds. For every algorithm, every $K \ge 2$, and every $T \ge K$,*

$$\sup_{\theta_1,\dots,\theta_K\in\Theta}\mathbb{E}[R_T] \ge c_{\mathrm{tr}}\sqrt{KT},$$

where $c_{\mathrm{tr}} > 0$ is a universal numerical constant. In this model, one action has larger reward but weaker feedback sensitivity, while the other action has lower reward but more informative feedback.

Theorems 5 and 6 show that the K, T dependence of `ACF-UCB` is optimal up to logarithmic factors.

Theorem 7 (Unstructured baseline lower bound)**.** *Consider algorithms that discard the shared-parameter structure and treat the KM task-option pairs as unrelated arms. In the resulting unstructured model class, for every such algorithm and every $T \ge KM \ge 2$,*

$$\sup_{\nu}\mathbb{E}_{\nu}[R_T] \ge c_{\mathrm{un}}\sqrt{KMT},$$

where the supremum is over ordinary Bernoulli bandit instances on KM arms and $c_{\mathrm{un}} > 0$ is a universal constant.

The proofs of Theorems 5, 6, and 7 are in Appendices E.1, E.2, and E.3, respectively.

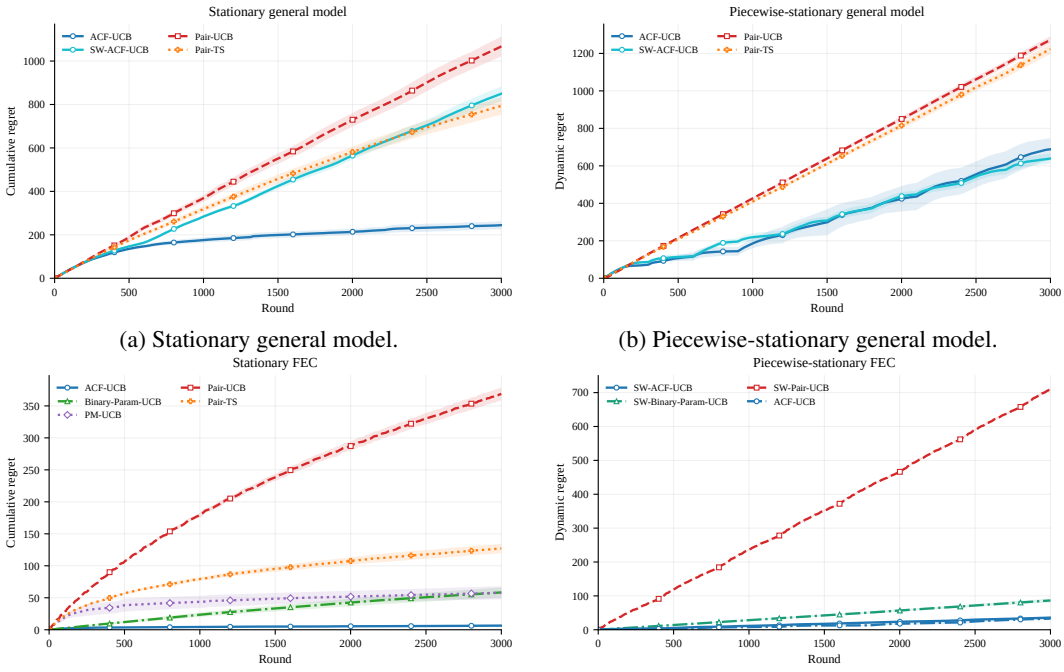
7 Experiments

We evaluate whether the shared-parameter coarsened-feedback structure is useful beyond the proofs.

Setup. For the general-model experiments, each action has binary feedback $g_j(1;\theta) = s_j\theta$ and reward $\mu(\theta, j) = r_j(1 - s_j\theta)$. The rates r_j and sensitivities s_j are chosen so that high-rate actions are attractive at small θ, high-sensitivity actions lose value faster as θ increases, and no action is trivially dominated. General-model task parameters lie in $[0.04, 0.66]$, and the piecewise-stationary instances change the identity of the best task at segment boundaries. For FEC, we use length-15 code

profiles $(15, 11, 3)$, $(15, 7, 5)$, $(15, 5, 7)$, and $(15, 2, 9)$ with channel error rates in $[0.002, 0.035]$. The statistics are $h_j(y) = y$ in the general model and $h_j(Y) = \mathbb{1}\{E \leq 1\}$ in FEC.

Protocol and baselines. All curves show pseudo-regret averaged over independent seeds, with shaded 95% confidence intervals. The main general-model panels use $T = 3000$, $K = 30$, $M = 10$, 10 random seeds. The FEC panels use $T = 3000$, $K = 6$, $M = 4$, and 8 random seeds. All experiments were run on a Linux machine with a 5.58 GHz AMD Ryzen 7 9700X (16) CPU and 32 GB RAM. For the general model, we compare against reward-only PAIR-UCB [Auer et al., 2002] and PAIR-TS [Agrawal and Goyal, 2012]. These baselines treat each task-option pair as an independent reward arm, so they test the cost of ignoring the shared task parameter. For FEC, we also include BINARY-PARAM-UCB, which knows the scalar channel model but only observes decoding success, and PM-UCB, which uses finite partial-monitoring matrices [Bartók et al., 2014] without the binomial one-parameter structure. BINARY-PARAM-UCB isolates the value of observing censored error counts beyond binary success, while PM-UCB tests whether a generic finite-feedback method can exploit the same structure. Appendix B gives the full configuration, scaling experiments, and practical radius factors used in the simulator.



(a) Stationary general model.

(b) Piecewise-stationary general model.

(c) Stationary FEC.

(d) Piecewise-stationary FEC.

Figure 1: Main experimental results.

Figure 1 shows that `ACF-UCB` has lower stationary regret than baselines that treat each task-option pair as an independent arm and use only reward observations. In the piecewise-stationary general model, `SW-ACF-UCB` adapts better to changes than `ACF-UCB`, which uses all past observations. The FEC panels show the benefit of observing error counts subject to censoring when channel error rates are small, as these counts provide information about the channel error rate beyond whether decoding succeeds. The scaling experiments in Appendix B further show that the structured method is nearly insensitive to M, while reward-only baselines degrade as the number of task-option pairs grows.

8 Limitations

Our analysis focuses on scalar parametric families with known action-dependent feedback laws. The main regret guarantees require Assumption 2, which ensures that observations from every option control the reward uncertainty of every other option. Extending the framework to vector-valued parameters, misspecified feedback laws, or partially unknown model classes would require new confidence constructions and different identifiability conditions. Our non-stationary results assume

piecewise-constant task parameters and use a fixed sliding-window length. Finally, our experiments are synthetic and primarily centered on the FEC application. Evaluations on larger-scale real-channel data and non-FEC domains would further clarify the empirical scope of the proposed framework.

9 Conclusion

In this paper, we introduced Parametric Bandits with Action-Dependent Coarsened Feedback, where each action determines both a reward function and a coarsened categorical feedback law. `ACF-UCB` achieves near-optimal stationary regret, and the analysis under Assumption 2 avoids the potentially pessimistic $1/\Gamma$ factor for applications such as FEC. `SW-ACF-UCB` extends the same optimism principle to piecewise-stationary tasks through per-task sliding windows. The lower bounds show that the stationary $\sqrt{KT}$ rate is tight up to logarithmic factors, while the unstructured KM-armed model pays a larger minimax rate.

Acknowledgments and Disclosure of Funding

The authors thank the anonymous reviewers for their helpful comments and suggestions. The work of Zhuohua Li was supported in part by the Xiaomi Young Talent Program and the Fundamental Research Funds for the Central Universities under Grant QTZX26133. The work of Yuwen Huang was supported in part by the Department of Education of Guangdong Province under Grant 2026KQNCX066 and in part by the Hong Kong University of Science and Technology (Guangzhou) under Grant G0101000351. The work of Jie Su and Cong Tian was supported by the National Natural Science Foundation of China under Grants 62402371 and 62192734. The work of Shengchao Qin was supported by the National Natural Science Foundation of China under Grant 62472339. The work of John C.S. Lui was supported in part by the RGC GRF-14202923.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2312–2320, 2011.

Jacob Abernethy, Kareem Amin, and Ruihao Zhu. Threshold bandit, with and without censored feedback. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4896–4904, 2016.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26. PMLR, 25–27 Jun 2012.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35, Paris, France, 03–06 Jul 2015. PMLR.

Onur Atan, Cem Tekin, and Mihaela van der Schaar. Global multi-armed bandits with H"older continuity. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 28–36, San Diego, California, USA, 09–12 May 2015. PMLR.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, May 2002.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 138–158. PMLR, 25–28 Jun 2019.

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.

Viktor Bengs and Eyke Hüllermeier. Multi-armed bandits with censored consumption of resources. *Machine Learning*, 112(1):217–240, November 2022.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 199–207, 2014.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

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

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 1761–1769, 2017.

Xiangxiang Dai, Yuejin Xie, Maoli Liu, Xuchuang Wang, Zhuohua Li, Huanyu Wang, and John C.S. Lui. A multi-agent conversational bandit approach to online evaluation and selection of user-aligned llm responses. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44): 37323–37331, Mar. 2026.

Rémy Degenne, Han Shao, and Wouter Koolen. Structure adaptive algorithms for stochastic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2443–2452. PMLR, 13–18 Jul 2020.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: the generalized linear case. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pages 586–594, 2010.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory*, pages 174–188. Springer Berlin Heidelberg, 2011.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 757–765. PMLR, 22–24 Jun 2014.

R. W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29 (2):147–160, 1950.

Emilie Kaufmann, Olivier Cappe, and Aurelien Garivier. On bayesian upper confidence bounds for bandit problems. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600. PMLR, 21–23 Apr 2012.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 358–384. PMLR, 06–09 Jul 2018.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Linear partial monitoring for sequential decision making: Algorithms, regret bounds and applications. *Journal of Machine Learning Research*, 24(346):1–45, 2023.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

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*, WWW '10, page 661–670, 2010.

Zhuohua Li, Maoli Liu, and John C. S. Lui. Fedconpe: Efficient federated conversational bandits with heterogeneous clients. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 4533–4541. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

Zhuohua Li, Maoli Liu, Xiangxiang Dai, and John C.S. Lui. Demystifying online clustering of bandits: Enhanced exploration under stochastic and smoothed adversarial contexts. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zhuohua Li, Maoli Liu, Xiangxiang Dai, and John C.S. Lui. Towards efficient conversational recommendations: Expected value of information meets bandit learning. In *THE WEB CONFERENCE 2025*, 2025b.

Maoli Liu, Zhuohua Li, Xiangxiang Dai, and John C.S. Lui. Leveraging the power of conversations: Optimal key term selection in conversational contextual bandits. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 1833–1844, 2025.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-concordant analysis of generalized linear bandits with forgetting. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 658–666. PMLR, 13–15 Apr 2021.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Andrea Tirinzoni, Alessandro Lazaric, and Marcello Restelli. A novel confidence-based algorithm for structured bandits. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3175–3185. PMLR, 26–28 Aug 2020.

Arun Verma, Manjesh Hanawal, Arun Rajkumar, and Raman Sankaran. Censored semi-bandits: A framework for resource allocation with censored feedback. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jing Wang, Peng Zhao, and Zhi-Hua Zhou. Revisiting weighted strategy for non-stationary parametric bandits. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7913–7942. PMLR, 25–27 Apr 2023.

Zhiyang Wang, Ruida Zhou, and Cong Shen. Regional multi-armed bandits. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 510–518. PMLR, 09–11 Apr 2018.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4300–4354. PMLR, 15–19 Aug 2021.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1113–1122, Lille, France, 07–09 Jul 2015. PMLR.

A Application Details

A.1 FEC Code Selection

We first spell out the FEC specialization used in the experiments. There are K independent binary symmetric channels with unknown error rates $p_i \in [0, 1/2]$. Here $\theta_i = p_i$ and $\Theta = [0, 1/2]$.

Table 2: Dictionary between the general model and the FEC running example.

General object	General role	FEC specialization
Task i	Unknown environment to operate on	Binary symmetric channel i
Parameter θ_i	Scalar condition shared by all actions	Error rate $p_i \in [0, 1/2]$
Action j	Decision/action choice	Code with parameters (n_j, k_j, d_j)
Feedback Y	Observed outcome	Reveal E if $E < d_j$, else $\perp$
Feedback alphabet $\mathcal{Y}_j$	Possible observed outcomes	$\{0, \dots, d_j - 1, \perp\}$
Feedback law g_j	Categorical law used by `ACF-UCB`	Censored binomial law on $\{0, \dots, d_j - 1, \perp\}$
Reward μ	Value of action at the task parameter	$(k_j/n_j) F_{\mathrm{Bin}}(t_j; n_j, p_i)$

Definition 1 (Code parameters)**.** Each action $j \in [M]$ is an error-correcting code with parameters (n_j, k_j, d_j), where $1 \le k_j < n_j$, $2 \le d_j \le n_j$, and $t_j = \lfloor (d_j - 1)/2 \rfloor$. Let $n_{\max} := \max_{j \in [M]} n_j$.

When code j is used on a channel with error rate p, the latent statistic is the error count

$$E \sim \mathrm{Bin}(n_j, p).$$

The decoder reports the exact count if it is below the code distance and reports a censorship symbol otherwise:

$$\psi_j(E) = \begin{cases} E, & E < d_j, \\ \perp, & E \ge d_j. \end{cases}$$

Thus $\mathcal{Y}_j = \{0, 1, \dots, d_j - 1, \perp\}$, and

$$g_j(y; p) = \begin{cases} f_{\mathrm{Bin}}(y; n_j, p), & y = 0, \dots, d_j - 1, \\ 1 - F_{\mathrm{Bin}}(d_j - 1; n_j, p), & y = \perp. \end{cases}$$

Consequently $D_{\mathcal{Y}} = \sum_j (d_j + 1)$. The realized reward is

$$r_t = \frac{k_{J_t}}{n_{J_t}} \mathbb{1}\{E_t \le t_{J_t}\},$$

and the expected reward is the quantity in Equation (3).

Proposition 8 (FEC specialization)**.** Under Definition 1, the online FEC code-selection problem under binary symmetric channels satisfies Assumptions 1 and 3. Moreover,

$$\Gamma \ge \min_j n_j 2^{-(n_j - 1)} > 0.$$

Proof. The feedback probabilities are polynomials in p, so they are continuous on $[0, 1/2]$. The zero-error atom satisfies $g_j(0; p) = (1 - p)^{n_j}$, and hence for $p, q \in [0, 1/2]$,

$$\|g_j(\cdot; p) - g_j(\cdot; q)\|_2 \ge |(1 - p)^{n_j} - (1 - q)^{n_j}| \ge n_j 2^{-(n_j - 1)} |p - q|.$$

This proves the stated lower bound on Γ. The reward $\mu(p, j) = (k_j/n_j) F_{\mathrm{Bin}}(t_j; n_j, p)$ is a polynomial in p on $[0, 1/2]$, so it is Lipschitz. □

Proposition 9 (Reward transfer for FEC)**.** Under Definition 1, Assumption 2 holds with the zero-error statistic $h_a(y) = \mathbb{1}\{y = 0\}$, mean $m_a(p) = (1 - p)^{n_a}$, and

$$B_0 := \max_{p \in [0, 1/2]} \frac{\max_{j \in [M]} k_j f_{\mathrm{Bin}}(t_j; n_j - 1, p)}{\min_{a \in [M]} n_a (1 - p)^{n_a - 1}} \le 2^{n_{\max} - 1}.$$

If all codes have the same block length n, one can instead set $r := \min_j t_j$ and use the statistic

$$h_a(y) := \mathbb{1}\{y \in \{0, \dots, r\}\}, \qquad m_a(p) = F_{\mathrm{Bin}}(r; n, p),$$

with

$$B := \max_{j \in [M]} \frac{k_j}{n} \frac{\binom{n-1}{t_j}}{\binom{n-1}{r}} \le 2^{n_{\max}-1}. \tag{7}$$

Proof. The binomial derivative identity gives

$$\left| \frac{\mathrm{d}\mu(p, j)}{\mathrm{d}p} \right| = k_j f_{\mathrm{Bin}}(t_j; n_j - 1, p).$$

Each zero-error mean is decreasing, with $|m_a'(p)| = n_a(1-p)^{n_a-1} > 0$ on $[0, 1/2]$. The denominator defining B_0 is therefore positive and continuous on this compact interval. For every target option j and observed option a, $|\mu'(p, j)| \le B_0 |m_a'(p)|$. Integrating this inequality between any two parameters proves the reward comparison in Assumption 2, because each m_a is monotone. The fixed block-length bound yields $B_0 \le 2^{n_{\max}-1}$, because $k_j < n_j \le n_{\max}$ and $n_a 2^{-(n_a-1)} \ge n_{\max} 2^{-(n_{\max}-1)}$.

For a common block length, $r \le t_a < d_a$ for every option a, so the event $E \le r$ is determined by every coarsened observation. All statistic means equal the decreasing function $m(p) = F_{\mathrm{Bin}}(r; n, p)$. For $p \in (0, 1/2]$,

$$\frac{|\mu'(p, j)|}{|m'(p)|} = \frac{k_j}{n} \frac{\binom{n-1}{t_j}}{\binom{n-1}{r}} \left(\frac{p}{1-p} \right)^{t_j - r} \le B.$$

The derivative inequality extends to $p = 0$ by continuity. Integration again gives the required comparison for all pairs of parameters, without requiring a positive lower bound on $|m'(p)|$. The stated bound follows from $k_j/n < 1$, $\binom{n-1}{r} \ge 1$, and $\binom{n-1}{t_j} \le 2^{n-1}$. □

Corollary 1 (FEC regret bounds)**.** *For the FEC application,*

$$\Theta = [0, 1/2], \qquad D_{\mathcal{Y}} = \sum_{j=1}^{M} (d_j + 1).$$

The bounds in Theorems 1 and 2 apply to `ACF-UCB` *and* `SW-ACF-UCB` *with either statistic and its corresponding constant from Proposition 9. The bounds in Theorems 3 and 4 also apply, with* $\Gamma \ge \min_j n_j 2^{-(n_j-1)}$.

B Additional Experiment Details

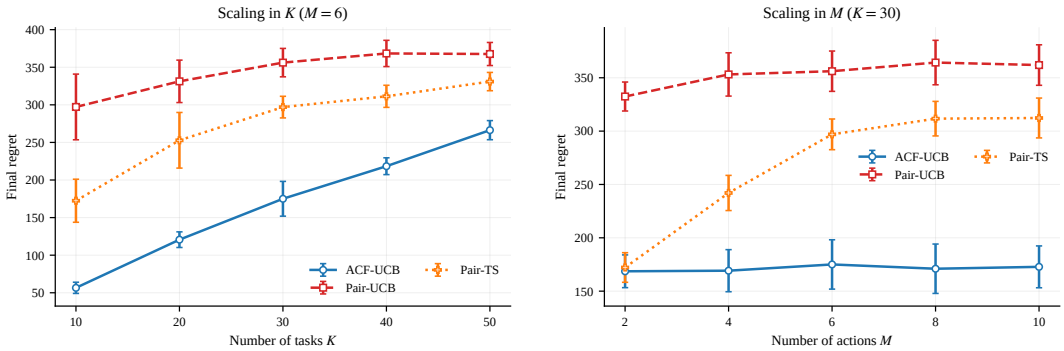
In the general-model experiments, each action has binary feedback alphabet $\mathcal{Y}_j = \{0, 1\}$. Action j has nominal rate $r_j \in [0.55, 0.95]$ and sensitivity $s_j \in [0.45, 1.35]$, evenly spaced over their respective intervals. The largest sensitivity satisfies $1.35 \cdot 0.72 < 1$, so the binary feedback law is valid on the full numerical grid $[0, 0.72]$. The crossover point of adjacent reward lines is

$$\theta_{j,j+1} = \frac{r_j - r_{j+1}}{r_j s_j - r_{j+1} s_{j+1}}.$$

With $M = 10$, these crossovers lie between approximately 0.30 and 0.56, inside the sampling range $[0.04, 0.66]$. Thus no action is trivially dominated in the M-scaling experiment. The statistic is $h_j(y) = y$, with $m_j(\theta) = s_j \theta$. Assumption 2 holds with $B = \max_j r_j s_j / \min_a s_a \le 2.85$.

Both general-model panels use $T = 3000$, $K = 30$, $M = 10$, 10 seeds, and a grid of 151 parameter values. The general non-stationary experiment uses segment length 300 and sliding window $W = 600$. The same window $W = 600$ is used for `SW-ACF-UCB` in the stationary general panel. The scaling experiment uses $T = 1000$, 5 seeds, grid size 101, $K \in \{10, 20, 30, 40, 50\}$ at fixed $M = 6$, and $M \in \{2, 4, 6, 8, 10\}$ at fixed $K = 30$. We set $\alpha = 0.05$.

The FEC non-stationary experiment uses the same piecewise construction as the general non-stationary experiment, with segment length 600. Both FEC experiments use $T = 3000$, $K = 6$, $M = 4$, 8 seeds, and grid size 151. The non-stationary FEC experiment uses window $W = 400$.



(a) Scaling in K with $M = 6$.

(b) Scaling in M with $K = 30$.

Figure 2: Scaling of final regret in the general model. The structured algorithm is nearly insensitive to M, while unstructured reward-only baselines degrade as the number of task-option pairs grows.

Figure 1 reports the main general-model and FEC panels. The fixed-window variant is included in the stationary general panel for comparability with the non-stationary panel. As expected, it is less efficient in stationary environments because it discards old observations.

Figure 2 shows two one-dimensional scaling slices. As K increases, all methods pay for more tasks, but the structured method remains below the unstructured baselines. As M increases at fixed K, the structured method is nearly flat, while reward-only baselines degrade because they must explore more task-option pairs. This is the empirical counterpart of the theoretical comparison between learning one scalar per task and learning KM unrelated reward arms.

In the stationary FEC panel of Figure 1, the full-feedback parametric method is below the unstructured reward-only baselines and the finite partial-monitoring baseline. It also improves over the binary-threshold ablation in the low-error regime. For small p, a single error is much more likely than two or more errors. The censored error count distinguishes $E = 0$ from $E = 1$, whereas the event $E \le 1$ combines these outcomes. The counts therefore provide additional information about the channel error rate, even when the reward depends only on whether decoding succeeds.

C Proofs for the Stationary Setting

Throughout the appendix, let $(U_t)_{t\ge1}$ be the learner's internal randomization, independent of the observation noise, and expose the round-t random seed before the action is chosen. For deterministic policies, take U_t to be constant. Define the pre-observation filtration

$$\mathcal{F}_1 := \sigma(U_1), \qquad \mathcal{F}_t := \sigma(U_1, I_1, J_1, Y_1, \dots, U_{t-1}, I_{t-1}, J_{t-1}, Y_{t-1}, U_t), \quad t \ge 2.$$

The selected pair (I_t, J_t) is $\mathcal{F}_t$-measurable, and conditional on $\mathcal{F}_t$ and $(I_t, J_t) = (i, j)$, the observation has law $g_j(\cdot; \theta_i)$ in the stationary setting and $g_j(\cdot; \theta_i(t))$ in the non-stationary setting. For any integer-valued stopping time $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$, define the stopped sigma-field

$$\mathcal{F}_\tau := \{A \in \mathcal{F} : A \cap \{\tau \le t\} \in \mathcal{F}_t \text{ for all } t \ge 1\}.$$

Thus $\mathcal{F}_\tau$ is the information before the observation at time τ, while $\mathcal{F}_{\tau+1}$ is the information after that observation has been incorporated.

We use the following count notation:

$$N_i(t) := \sum_{s<t} \mathbb{1}\{I_s = i\}, \qquad N_{i,j}(t) := \sum_{s<t} \mathbb{1}\{I_s = i, J_s = j\}.$$

C.1 Stationary Confidence Lemmas

Lemma 1 (Predictable subsampling Hoeffding bound)**.** *Let $\{\mathcal{F}_t\}_{t\ge1}$ be a filtration, $A_t \in \{0, 1\}$ and p_t be $\mathcal{F}_t$-measurable, and Z_t be observed after time t, i.e., Z_t is $\mathcal{F}_{t+1}$-measurable. Suppose that, on $\{A_t = 1\}$,*

$$\mathbb{E}[Z_t \mid \mathcal{F}_t] = p_t, \qquad |Z_t - p_t| \le 1 \quad \text{a.s.}$$

Define

$$S_t := \sum_{s<t} A_s(Z_s - p_s), \qquad N_t := \sum_{s<t} A_s.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $t \geq 1$,

$$|S_t| \leq \sqrt{2N_t \log\left(\frac{\pi^2 N_t^2}{3\delta}\right)},$$

with the convention that the right-hand side is 0 when $N_t = 0$.

Proof. Define the active times

$$\tau_n := \inf\left\{ r \geq 1 : \sum_{s=1}^{r} A_s \geq n \right\}, \qquad n \geq 1,$$

with the convention $\inf \emptyset = \infty$. Since A_s is $\mathcal{F}_s$-measurable, each τ_n is a stopping time.

Define

$$\xi_n := \begin{cases} Z_{\tau_n} - p_{\tau_n}, & \tau_n < \infty, \\ 0, & \tau_n = \infty. \end{cases}$$

Let

$$\mathcal{G}_0 := \mathcal{F}_1, \qquad \mathcal{G}_n := \mathcal{F}_{\tau_n+1}, \qquad n \geq 1,$$

where we use the convention $\mathcal{F}_{\infty+1} = \mathcal{F}_\infty := \sigma(\bigcup_{t\geq 1} \mathcal{F}_t)$. Since Z_t is $\mathcal{F}_{t+1}$-measurable, each ξ_n is $\mathcal{G}_n$-measurable.

Let $B \in \mathcal{F}_{\tau_n}$. Since $\xi_n = 0$ on $\{\tau_n = \infty\}$, we have

$$\mathbb{E}[\mathbb{1}_B \xi_n] = \sum_{r\geq 1} \mathbb{E}\big[\mathbb{1}_{B\cap\{\tau_n = r\}} \xi_n\big] = \sum_{r\geq 1} \mathbb{E}\big[\mathbb{1}_{B\cap\{\tau_n = r\}} (Z_r - p_r)\big].$$

For each $r \geq 1$, because $B \in \mathcal{F}_{\tau_n}$, we have

$$B \cap \{\tau_n = r\} \in \mathcal{F}_r.$$

Moreover, on $\{\tau_n = r\}$, we have $A_r = 1$. Hence

$$B \cap \{\tau_n = r\} \subseteq \{A_r = 1\}.$$

Using the assumption $\mathbb{E}[Z_r \mid \mathcal{F}_r] = p_r$ on $\{A_r = 1\}$, we get

$$\mathbb{E}\left[\mathbb{1}_{B\cap\{\tau_n=r\}}(Z_r - p_r)\right] = \mathbb{E}\left[\mathbb{1}_{B\cap\{\tau_n=r\}}\, \mathbb{E}[Z_r - p_r \mid \mathcal{F}_r]\right] = 0.$$

Therefore, $\mathbb{E}[\mathbb{1}_B \xi_n] = 0$ for all $B \in \mathcal{F}_{\tau_n}$. By definition, this implies $\mathbb{E}[\xi_n \mid \mathcal{F}_{\tau_n}] = 0$.

Also, on $\{\tau_n < \infty\}$, we have $A_{\tau_n} = 1$, and hence $|\xi_n| = |Z_{\tau_n} - p_{\tau_n}| \leq 1$. On $\{\tau_n = \infty\}$, $\xi_n = 0$. Thus $|\xi_n| \leq 1$ almost surely.

Next, we claim that $\mathcal{G}_{n-1} \subseteq \mathcal{F}_{\tau_n}$. For $n = 1$, this follows from $\mathcal{G}_0 = \mathcal{F}_1$ and $\tau_1 \geq 1$. For $n \geq 2$, the n-th active time is strictly after the $(n-1)$-st active time whenever both are finite, and hence $\tau_{n-1} + 1 \leq \tau_n$. Therefore

$$\mathcal{G}_{n-1} = \mathcal{F}_{\tau_{n-1}+1} \subseteq \mathcal{F}_{\tau_n}.$$

By the tower property,

$$\mathbb{E}[\xi_n \mid \mathcal{G}_{n-1}] = \mathbb{E}[\mathbb{E}[\xi_n \mid \mathcal{F}_{\tau_n}] \mid \mathcal{G}_{n-1}] = 0.$$

Thus $\{\xi_n\}_{n\geq 1}$ is a bounded martingale difference sequence with respect to $\{\mathcal{G}_n\}_{n\geq 0}$. By the Hoeffding–Azuma inequality, for every deterministic $n \geq 1$ and every $x > 0$,

$$\Pr\left(\left|\sum_{\ell=1}^{n} \xi_\ell\right| > x\right) \leq 2\exp\left(-\frac{x^2}{2n}\right).$$

Set

$$x_n := \sqrt{2n \log\left(\frac{\pi^2 n^2}{3\delta}\right)}.$$

Then

$$\Pr\left(\left|\sum_{\ell=1}^{n}\xi_\ell\right| > x_n\right) \le \frac{6\delta}{\pi^2 n^2}.$$

Taking a union bound over all $n \ge 1$, the total failure probability is at most $\sum_{n=1}^{\infty}\frac{6\delta}{\pi^2 n^2} = \delta$. Therefore, with probability at least $1-\delta$, for all $n \ge 1$,

$$\left|\sum_{\ell=1}^{n}\xi_\ell\right| \le \sqrt{2n\log\left(\frac{\pi^2 n^2}{3\delta}\right)}.$$

It remains to translate this active-prefix bound back to calendar time. Fix $t \ge 1$. If $N_t = 0$, then no active observation occurs before time t, so $S_t = 0$. If $N_t = n \ge 1$, then the active times before t are exactly $\tau_1, \tau_2, \ldots, \tau_n$. Therefore

$$S_t = \sum_{s<t} A_s(Z_s - p_s) = \sum_{\ell=1}^{n}(Z_{\tau_\ell} - p_{\tau_\ell}) = \sum_{\ell=1}^{n}\xi_\ell.$$

Substituting $n = N_t$ gives

$$|S_t| \le \sqrt{2N_t\log\left(\frac{\pi^2 N_t^2}{3\delta}\right)}.$$

Since the high-probability event above holds simultaneously for all $n \ge 1$, the bound holds simultaneously for all $t \ge 1$. The case $N_t = 0$ is covered by the stated convention. □

Lemma 2 (Stationary confidence)**.** *Under Assumption 1, fix $\alpha \in (0,1)$. There is an event $\mathcal{E}_{\mathrm{st}}$ with probability at least $1-\alpha$ such that, for all tasks i and all rounds t,*

$$\theta_i \in \mathcal{C}_i(t),$$

and every $q \in \mathcal{C}_i(t)$ satisfies, for every $j \in [M]$ and $y \in \mathcal{Y}_j$,

$$N_{i,j}(t)|g_j(y;q) - g_j(y;\theta_i)| \le 2\sqrt{2N_{i,j}(t)\beta_{\mathrm{st}}(t,\alpha)}.$$

The task-level confidence constraint also gives

$$\left|\sum_{a=1}^{M} N_{i,a}(t)(m_a(q) - m_a(\theta_i))\right| \le 2\sqrt{2N_i(t)\beta_{\mathrm{st}}(t,\alpha)}. \tag{8}$$

Proof. Fix a task i, an action j, and an outcome $y \in \mathcal{Y}_j$. For each round s, define the predictable selection indicator

$$A_s^{i,j} := \mathbb{1}\{I_s = i, J_s = j\}.$$

Since the selected pair (I_s, J_s) is $\mathcal{F}_s$-measurable, $A_s^{i,j}$ is also $\mathcal{F}_s$-measurable. On the event $\{A_s^{i,j} = 1\}$, the feedback is generated from $g_j(\cdot;\theta_i)$. Hence, for

$$Z_s^y := \mathbb{1}\{Y_s = y\}, \qquad p := g_j(y;\theta_i),$$

we have, on $\{A_s^{i,j} = 1\}$,

$$\mathbb{E}[Z_s^y \mid \mathcal{F}_s] = g_j(y;\theta_i) = p, \qquad |Z_s^y - p| \le 1.$$

Therefore Lemma 1, applied with confidence level $\delta := \frac{\alpha}{K(D_{\mathcal{Y}}+1)}$, implies that, with probability at least $1-\delta$, simultaneously for all $t \ge 1$,

$$\left|\sum_{s<t} A_s^{i,j}(\mathbb{1}\{Y_s = y\} - g_j(y;\theta_i))\right| \le \sqrt{2N_{i,j}(t)\log\left(\frac{\pi^2 K(D_{\mathcal{Y}}+1)N_{i,j}(t)^2}{3\alpha}\right)},$$

where $N_{i,j}(t) = \sum_{s<t} A_s^{i,j}$. If $N_{i,j}(t) = 0$, both sides are interpreted as zero. If $N_{i,j}(t) \ge 1$, then $N_{i,j}(t) \le t$, and hence

$$\log\left(\frac{\pi^2 K(D_{\mathcal{Y}}+1)N_{i,j}(t)^2}{3\alpha}\right) \le \log\left(\frac{\pi^2 K(D_{\mathcal{Y}}+1)t^2}{3\alpha}\right) = \beta_{\mathrm{st}}(t,\alpha).$$

Thus, for the fixed triple (i, j, y), with probability at least $1 - \delta$, for all $t \geq 1$,

$$\left| \sum_{s<t: I_s=i, J_s=j} (\mathbb{1}\{Y_s = y\} - g_j(y; \theta_i)) \right| \leq \sqrt{2N_{i,j}(t)\beta_{\mathrm{st}}(t, \alpha)}. \tag{9}$$

For the task-level confidence constraint of task i, apply Lemma 1 with

$$A_s^i := \mathbb{1}\{I_s = i\}, \qquad Z_s := h_{J_s}(Y_s), \qquad p_s := m_{J_s}(\theta_i).$$

The center p_s is predictable because the option J_s is selected before observing Y_s. On $\{A_s^i = 1\}$, the conditional mean of Z_s is p_s, and $|Z_s - p_s| \leq 1$ because $h_j \in [0, 1]$. With the same failure probability $\delta = \alpha/[K(D_{\mathcal{Y}} + 1)]$, this gives, simultaneously for all t,

$$\left| \sum_{s<t: I_s=i} (h_{J_s}(Y_s) - m_{J_s}(\theta_i)) \right| \leq \sqrt{2N_i(t)\beta_{\mathrm{st}}(t, \alpha)}. \tag{10}$$

A union bound over the $KD_{\mathcal{Y}}$ coordinate streams and K task streams has total failure probability at most $KD_{\mathcal{Y}}\delta + K\delta = \alpha$. Call this event $\mathcal{E}_{\mathrm{st}}$. We work on $\mathcal{E}_{\mathrm{st}}$ for the remainder of the proof.

First, fix any task i and round t. Equations (9) and (10) are the defining constraints of $\mathcal{C}_i(t)$ evaluated at the true parameter θ_i, for every $j \in [M]$ and $y \in \mathcal{Y}_j$. Therefore $\theta_i \in \mathcal{C}_i(t)$ for all i,t.

Second, fix $q \in \mathcal{C}_i(t)$, an action j, and an outcome $y \in \mathcal{Y}_j$. Since q and θ_i belong to $\mathcal{C}_i(t)$, we have

$$\left| \sum_{s<t: I_s=i, J_s=j} (\mathbb{1}\{Y_s = y\} - g_j(y; q)) \right| \leq \sqrt{2N_{i,j}(t)\beta_{\mathrm{st}}(t, \alpha)},$$

and

$$\left| \sum_{s<t: I_s=i, J_s=j} (\mathbb{1}\{Y_s = y\} - g_j(y; \theta_i)) \right| \leq \sqrt{2N_{i,j}(t)\beta_{\mathrm{st}}(t, \alpha)}.$$

Subtracting the two sums cancels the empirical indicator terms and gives

$$\sum_{s<t: I_s=i, J_s=j} (g_j(y; \theta_i) - g_j(y; q)) = N_{i,j}(t)(g_j(y; \theta_i) - g_j(y; q)).$$

Hence, by the triangle inequality,

$$\begin{aligned} N_{i,j}(t)|g_j(y; q) - g_j(y; \theta_i)| &\leq \left| \sum_{s<t: I_s=i, J_s=j} (\mathbb{1}\{Y_s = y\} - g_j(y; q)) \right| \\ &\quad + \left| \sum_{s<t: I_s=i, J_s=j} (\mathbb{1}\{Y_s = y\} - g_j(y; \theta_i)) \right| \\ &\leq 2\sqrt{2N_{i,j}(t)\beta_{\mathrm{st}}(t, \alpha)}. \end{aligned}$$

This proves the claimed coordinate confidence bound. Finally, apply the same subtraction and triangle inequality to the task-level constraints at q and θ_i. The empirical $h_{J_s}(Y_s)$ terms cancel, and the remaining sum is $\sum_a N_{i,a}(t)(m_a(q) - m_a(\theta_i))$, proving Equation (8). □

Lemma 3 (Stationary reward-sensitive confidence)**.** *Under Assumption 2, on the event $\mathcal{E}_{\mathrm{st}}$, for every task i, round t, action j, and $q \in \mathcal{C}_i(t)$, if $N_i(t) \geq 1$, then*

$$|\mu(q, j) - \mu(\theta_i, j)| \leq B\sqrt{\frac{8\beta_{\mathrm{st}}(t, \alpha)}{N_i(t)}}.$$

Proof. For fixed q and θ_i, the differences $m_a(q) - m_a(\theta_i)$ have the same sign for every observed option a, because their means have a common monotonicity direction. The reward comparison in Assumption 2 therefore gives

$$\begin{aligned} N_i(t)|\mu(q,j) - \mu(\theta_i,j)| &\le B\sum_{a=1}^{M} N_{i,a}(t)|m_a(q) - m_a(\theta_i)| \\ &= B\left|\sum_{a=1}^{M} N_{i,a}(t)(m_a(q) - m_a(\theta_i))\right| \\ &\le 2B\sqrt{2N_i(t)\beta_{\mathrm{st}}(t,\alpha)}, \end{aligned}$$

where the last step is Equation (8). Division by $N_i(t)$ proves the claim. □

C.2 Proof of Theorem 1

Theorem 1 (Reward-sensitive stationary regret of `ACF-UCB`)**.** *Under Assumptions 1 and 2, for any stationary instance and for every $\alpha \in (0,1)$, Algorithm 1 satisfies, with probability at least $1-\alpha$,*

$$R_T \le K + 4\sqrt{2}\,B\sqrt{KT\beta_{\mathrm{st}}(T,\alpha)}.$$

Proof. Let $\mathcal{E}_{\mathrm{st}}$ be the event from Lemma 2. Then $\Pr(\mathcal{E}_{\mathrm{st}}^c) \le \alpha$. Work on $\mathcal{E}_{\mathrm{st}}$. The initialization phase incurs regret at most K and ensures $N_i(t) \ge 1$ for each subsequent selection.

At such a remaining round, let $i = I_t$, $j = J_t$, and $\tilde{\theta} = \tilde{\theta}_i(t)$. Let $i^\star \in \arg\max_i v(\theta_i)$. Since $\theta_{i^\star} \in \mathcal{C}_{i^\star}(t)$ on $\mathcal{E}_{\mathrm{st}}$, the optimistic value of the oracle-optimal task satisfies $U_{i^\star}(t) \ge v(\theta_{i^\star}) = \mu^*$. The algorithm selects a task with maximal UCB, so $U_i(t) \ge \mu^*$. By the definition of the selected action, $U_i(t) = v(\tilde{\theta}) = \mu(\tilde{\theta}, j)$. Therefore

$$\mu^* - \mu(\theta_i, j) \le \mu(\tilde{\theta}, j) - \mu(\theta_i, j).$$

Lemma 3 yields

$$\mu^* - \mu(\theta_{I_t}, J_t) \le B\sqrt{\frac{8\beta_{\mathrm{st}}(T,\alpha)}{N_{I_t}(t)}}.$$

Grouping the remaining rounds by task, and using $\sum_{n=1}^{m} n^{-1/2} \le 2\sqrt{m}$, gives

$$\begin{aligned} R_T &\le K + B\sqrt{8\beta_{\mathrm{st}}(T,\alpha)}\sum_{i=1}^{K}\sum_{n=1}^{N_i(T+1)}\frac{1}{\sqrt{n}} \\ &\le K + 4\sqrt{2}B\sqrt{\beta_{\mathrm{st}}(T,\alpha)}\sum_{i=1}^{K}\sqrt{N_i(T+1)} \\ &\le K + 4\sqrt{2}B\sqrt{KT\beta_{\mathrm{st}}(T,\alpha)}, \end{aligned}$$

where the last inequality follows from Cauchy-Schwarz and $\sum_i N_i(T+1) = T$. This proves the high-probability bound. On $\mathcal{E}_{\mathrm{st}}^c$, regret is at most T, which gives the expectation bound. □

C.3 Fallback Γ-Based Stationary Proof

The parameter radius used in the stationary Γ-based analysis is

$$\rho_{\mathrm{st}}(n,t,\alpha) := \frac{\sqrt{8D_{\mathcal{Y}}\beta_{\mathrm{st}}(t,\alpha)}}{\Gamma\sqrt{n}}.$$

Lemma 4 (Stationary parameter width)**.** *Under Assumption 3, on the event $\mathcal{E}_{\mathrm{st}}$, every task i, round t, and $q \in \mathcal{C}_i(t)$ satisfy*

$$|q - \theta_i| \le \rho_{\mathrm{st}}(N_i(t), t, \alpha) \qquad \textit{whenever } N_i(t) \ge 1.$$

Proof. Fix i, t, and $q \in \mathcal{C}_i(t)$. On the event $\mathcal{E}_{\mathrm{st}}$, both θ_i and q belong to $\mathcal{C}_i(t)$, and the coordinate bound of Lemma 2 holds for every action-outcome pair. Terms with $N_{i,j}(t) = 0$ contribute zero. Squaring the coordinate bound and summing over all outcomes gives

$$\sum_{j=1}^{M} N_{i,j}(t) \|g_j(\cdot; q) - g_j(\cdot; \theta_i)\|_2^2 \le 8 D_{\mathcal{Y}} \beta_{\mathrm{st}}(t, \alpha).$$

By the definition of Γ, the left-hand side is at least

$$\Gamma^2 N_i(t)(q - \theta_i)^2.$$

Combining the last two displays proves the claimed width bound. □

C.4 Proof of Theorem 3

Theorem 3 (Stationary regret of `ACF-UCB`)**.** *Under Assumptions 1 and 3, fix any stationary instance with $\theta_i \in \Theta$. For every $\alpha \in (0, 1)$, Algorithm 1 satisfies, with probability at least $1 - \alpha$,*

$$R_T \le K + 4\sqrt{2 D_{\mathcal{Y}}} \frac{L}{\Gamma} \sqrt{KT \beta_{\mathrm{st}}(T, \alpha)}.$$

In particular, setting $\alpha = 1/T$ gives

$$\mathbb{E}[R_T] = O\left(K + \frac{L\sqrt{D_{\mathcal{Y}}}}{\Gamma} \sqrt{KT \log(K D_{\mathcal{Y}} T)} \right)$$

up to universal logarithmic constants.

Proof. If $T < K$, the initialization phase stops after T rounds and the bound is trivial because $R_T \le T \le K$. We therefore assume $T \ge K$ below. Let $\mathcal{E}_{\mathrm{st}}$ be the event from Lemma 2. Then $\Pr(\mathcal{E}_{\mathrm{st}}^c) \le \alpha$. Work on $\mathcal{E}_{\mathrm{st}}$. Consider a non-initialization round t, let $i = I_t$, set $n = N_i(t)$, and write $\tilde{\theta} = \tilde{\theta}_i(t)$. Let $i^\star \in \arg\max_i v(\theta_i)$. Since $\theta_{i^\star} \in \mathcal{C}_{i^\star}(t)$,

$$U_{i^\star}(t) \ge v(\theta_{i^\star}) = \mu^*.$$

Since the algorithm chooses a task with maximal UCB, $U_i(t) \ge \mu^*$. The chosen action satisfies

$$U_i(t) = v(\tilde{\theta}) = \mu(\tilde{\theta}, J_t).$$

By Lemma 4, $|\tilde{\theta} - \theta_i| \le \rho_{\mathrm{st}}(n, t, \alpha)$. Hence

$$\mu^* - \mu(\theta_{I_t}, J_t) \le \mu(\tilde{\theta}, J_t) - \mu(\theta_i, J_t) \le L \rho_{\mathrm{st}}(N_{I_t}(t), t, \alpha).$$

Since $t \le T$ and $\beta_{\mathrm{st}}(\cdot, \alpha)$ is nondecreasing, $\rho_{\mathrm{st}}(n, t, \alpha) \le \rho_{\mathrm{st}}(n, T, \alpha)$. Summing and grouping by task,

$$\begin{aligned}
R_T &\le K + L \sum_{i=1}^{K} \sum_{n=1}^{N_i(T+1)} \rho_{\mathrm{st}}(n, T, \alpha) \\
&\le K + \sqrt{8 D_{\mathcal{Y}}} \frac{L}{\Gamma} \sqrt{\beta_{\mathrm{st}}(T, \alpha)} \sum_{i=1}^{K} \sum_{n=1}^{N_i(T+1)} \frac{1}{\sqrt{n}} \\
&\le K + 4\sqrt{2 D_{\mathcal{Y}}} \frac{L}{\Gamma} \sqrt{\beta_{\mathrm{st}}(T, \alpha)} \sum_{i=1}^{K} \sqrt{N_i(T+1)} \\
&\le K + 4\sqrt{2 D_{\mathcal{Y}}} \frac{L}{\Gamma} \sqrt{KT \beta_{\mathrm{st}}(T, \alpha)}.
\end{aligned}$$

On $\mathcal{E}_{\mathrm{st}}^c$, regret is at most T, which gives the expectation bound. □

D Proofs for the Non-Stationary Setting

For a window length W, let

$$\mathcal{W}(t) := \{s : \max(1, t - W) \le s < t\},$$

and define the sliding-window counts

$$N_i^W(t) := \sum_{s \in \mathcal{W}(t)} \mathbb{1}\{I_s = i\}, \qquad N_{i,j}^W(t) := \sum_{s \in \mathcal{W}(t)} \mathbb{1}\{I_s = i, J_s = j\}.$$

D.1 Sliding-Window Confidence Lemmas

Lemma 5 (Sliding-window confidence)**.** *Under Assumption 1, fix $\alpha \in (0,1)$. There is an event $\mathcal{E}_{\mathrm{sw}}$ with probability at least $1-\alpha$ such that, for every clean window (i,t), meaning that all observations of task i in $\mathcal{W}(t)$ were generated by the current parameter $\theta_i(t)$,*

$$\theta_i(t) \in \mathcal{C}_i^W(t),$$

and every $q \in \mathcal{C}_i^W(t)$ satisfies, for every $j \in [M]$ and $y \in \mathcal{Y}_j$,

$$N_{i,j}^W(t)|g_j(y;q) - g_j(y;\theta_i(t))| \le 2\sqrt{2N_{i,j}^W(t)\beta_{\mathrm{sw}}(t,\alpha)}.$$

The task-level confidence constraint also gives

$$\left|\sum_{a=1}^{M} N_{i,a}^W(t)(m_a(q) - m_a(\theta_i(t)))\right| \le 2\sqrt{2N_i^W(t)\beta_{\mathrm{sw}}(t,\alpha)}. \tag{11}$$

Proof. Fix a task i, an action j, and an outcome $y \in \mathcal{Y}_j$. Define the active times

$$\tau_m^{i,j} := \inf\left\{ r \ge 1 : \sum_{s=1}^{r} \mathbb{1}\{I_s = i, J_s = j\} \ge m \right\}, \qquad m \ge 1,$$

with $\inf \emptyset = \infty$, and the active filtration

$$\mathcal{G}_0^{i,j} := \mathcal{F}_1, \qquad \mathcal{G}_m^{i,j} := \mathcal{F}_{\tau_m^{i,j}+1}, \quad m \ge 1.$$

We use the convention $\mathcal{F}_{\infty+1} = \mathcal{F}_\infty := \mathcal{F}$. When $\tau_m^{i,j} < \infty$, set

$$\xi_m^{i,j,y} := \mathbb{1}\{Y_{\tau_m^{i,j}} = y\} - g_j(y;\theta_i(\tau_m^{i,j})).$$

For finite-prefix martingale arguments, set $\xi_m^{i,j,y} = 0$ on $\{\tau_m^{i,j} = \infty\}$. The centering uses the parameter at the calendar time when the active observation is generated. Since (I_r, J_r) is $\mathcal{F}_r$-measurable, each $\tau_m^{i,j}$ is a stopping time. On $\{\tau_m^{i,j} < \infty\}$, the selected pair is (i,j), and conditional on the pre-observation sigma-field $\mathcal{F}_{\tau_m^{i,j}}$, the feedback law is $g_j(\cdot;\theta_i(\tau_m^{i,j}))$. Hence

$$\mathbb{E}\left[\xi_m^{i,j,y} \,\middle|\, \mathcal{F}_{\tau_m^{i,j}}\right] = 0.$$

As in the stationary proof, $\mathcal{G}_{m-1}^{i,j} \subseteq \mathcal{F}_{\tau_m^{i,j}}$ on $\{\tau_m^{i,j} < \infty\}$, so the tower property gives

$$\mathbb{E}\left[\xi_m^{i,j,y} \,\middle|\, \mathcal{G}_{m-1}^{i,j}\right] = 0.$$

Thus the time-varying active increments are bounded martingale differences.

For any deterministic active interval $0 \le a < b$, the event $\{\tau_b^{i,j} < \infty\}$ makes all increments in the block $a+1, \dots, b$ well-defined. Hoeffding–Azuma applied conditionally after the a-th active observation gives

$$\Pr\left(\tau_b^{i,j} < \infty, \left|\sum_{m=a+1}^{b} \xi_m^{i,j,y}\right| > \sqrt{2(b-a)\beta_{\mathrm{sw}}(b,\alpha)}\right) \le 2e^{-\beta_{\mathrm{sw}}(b,\alpha)}.$$

For each fixed active endpoint b, there are b possible starting indices $a = 0, \dots, b-1$. For the fixed triple (i,j,y), the total failure probability over all intervals is

$$\sum_{b=1}^{\infty}\sum_{a=0}^{b-1} 2e^{-\beta_{\mathrm{sw}}(b,\alpha)} = \frac{6\alpha}{\pi^2 K(D_{\mathcal{Y}}+1)}\sum_{b=1}^{\infty}\frac{1}{b^2} = \frac{\alpha}{K(D_{\mathcal{Y}}+1)}.$$

For each task i, also consider the active times τ_m^i of $\mathbb{1}\{I_s = i\}$ and the task-level statistic increments

$$\zeta_m^i := h_{J_{\tau_m^i}}(Y_{\tau_m^i}) - m_{J_{\tau_m^i}}(\theta_i(\tau_m^i))$$

when $\tau_m^i < \infty$, with value zero otherwise. Their calendar-time centers are predictable, their conditional means are zero, and their absolute values are at most one. The active-time construction in

Lemma 1 therefore applies to this task stream, including when the selected options vary. The preceding argument gives the same interval bound and the same total failure probability $\alpha/[K(D_{\mathcal{Y}}+1)]$ for each task stream. A union bound over the $KD_{\mathcal{Y}}$ coordinate streams and K task streams gives an event $\mathcal{E}_{\mathrm{sw}}$ of probability at least $1-\alpha$ on which all these active-interval inequalities hold simultaneously.

Consider a clean window (i,t). For fixed j, the observations in $\mathcal{W}(t)$ with $I_s = i, J_s = j$ form a contiguous active interval for the pair (i,j). If the interval is empty, the corresponding coordinate sum is zero. Otherwise, write it as $a+1,\dots,b$. Because the window is clean, every observation in this active interval was generated with parameter $\theta_i(t)$. Thus the sum of $\xi_m^{i,j,y}$'s over this interval is exactly the coordinate sum appearing in $\mathcal{C}_i^W(t)$ evaluated at $\theta_i(t)$. The observations of task i in the same window also form a contiguous active interval in its task stream. Cleanliness makes the center of each task-level increment equal to $m_{J_s}(\theta_i(t))$, although the option J_s may vary within the interval. Thus both the coordinatewise constraints and the task-level constraint hold at $\theta_i(t)$, using $\beta_{\mathrm{sw}}(b,\alpha) \le \beta_{\mathrm{sw}}(t,\alpha)$ for each stream's active endpoint $b \le t$. Consequently $\theta_i(t) \in \mathcal{C}_i^W(t)$.

Now fix $q \in \mathcal{C}_i^W(t)$. Since both q and $\theta_i(t)$ are in $\mathcal{C}_i^W(t)$, for every (j,y),

$$N_{i,j}^W(t)|g_j(y;q) - g_j(y;\theta_i(t))| \le 2\sqrt{2N_{i,j}^W(t)\beta_{\mathrm{sw}}(t,\alpha)}.$$

This proves the coordinate confidence bound. Subtracting the task-level constraints at q and $\theta_i(t)$ cancels the observed statistics and yields Equation (11) by the same triangle inequality. □

Lemma 6 (Sliding-window reward-sensitive confidence)**.** *Under Assumption 2, on the event of Lemma 5, for every clean window (i,t), action j, and $q \in \mathcal{C}_i^W(t)$, if $N_i^W(t) \ge 1$, then*

$$|\mu(q,j) - \mu(\theta_i(t),j)| \le B\sqrt{\frac{8\beta_{\mathrm{sw}}(t,\alpha)}{N_i^W(t)}}.$$

Proof. The common monotonicity direction in Assumption 2 prevents cancellation between the differences $m_a(q) - m_a(\theta_i(t))$. Applying the reward comparison for every observed option a, then Equation (11), gives

$$\begin{aligned} N_i^W(t)|\mu(q,j) - \mu(\theta_i(t),j)| &\le B\sum_{a=1}^{M} N_{i,a}^W(t)|m_a(q) - m_a(\theta_i(t))| \\ &= B\left|\sum_{a=1}^{M} N_{i,a}^W(t)(m_a(q) - m_a(\theta_i(t)))\right| \\ &\le 2B\sqrt{2N_i^W(t)\beta_{\mathrm{sw}}(t,\alpha)}. \end{aligned}$$

Dividing by $N_i^W(t)$ proves the claim. □

D.2 Proof of Theorem 2

Theorem 2 (Reward-sensitive dynamic regret of `SW-ACF-UCB`)**.** *Under Assumptions 1 and 2, consider Algorithm 2 with window $W \in [T]$. For every $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$R_T^{\mathrm{dyn}} \le\ S(W+1) + K\left(1+\frac{T}{W}\right) + B\sqrt{8\beta_{\mathrm{sw}}(T,\alpha)}\left\lceil\frac{T}{W}\right\rceil\left(K+2\sqrt{KW}\right).$$

Proof. Call a round t contaminated if there exists a task change at a time τ with $t-W \le \tau \le t$. As in the proof of Theorem 4, there are at most $S(W+1)$ contaminated rounds, and we charge regret at most one to each.

Let $\mathcal{E}_{\mathrm{sw}}$ be the event from Lemma 5. Then $\Pr(\mathcal{E}_{\mathrm{sw}}^c) \le \alpha$. Work on $\mathcal{E}_{\mathrm{sw}}$ and consider a non-contaminated round t. Let $i = I_t$ and $j = J_t$. By Lemma 8, there are at most $K(1+T/W)$ rounds with $N_{I_t}^W(t) = 0$, and each incurs regret at most one. It remains to consider non-contaminated rounds with $N_{I_t}^W(t) \ge 1$. If some task had an empty window, its UCB would be $+\infty$, so the algorithm would select an empty-window task. Therefore every task, including an oracle-optimal task, has a nonempty clean window at this round.

Let $\tilde{\theta} = \tilde{\theta}_i^W(t)$. Let $i^\star \in \arg\max_i v(\theta_i(t))$. Since $\theta_{i^\star}(t) \in \mathcal{C}_{i^\star}^W(t)$, optimism gives

$$U_{i^\star}(t) \geq v(\theta_{i^\star}(t)) = \max_{i,j} \mu(\theta_i(t), j).$$

The selected task has maximal UCB and $U_i(t) = v(\tilde{\theta}) = \mu(\tilde{\theta}, j)$. Therefore

$$\max_{i,j} \mu(\theta_i(t), j) - \mu(\theta_i(t), j) \leq \mu(\tilde{\theta}, j) - \mu(\theta_i(t), j).$$

By Lemma 6,

$$\max_{i,j} \mu(\theta_i(t), j) - \mu(\theta_i(t), j) \leq B\sqrt{\frac{8\beta_{\mathrm{sw}}(T, \alpha)}{N_i^W(t)}}.$$

Summing over the remaining rounds and applying Lemma 9 gives

$$R_T^{\mathrm{dyn}} \leq S(W+1) + K\left(1 + \frac{T}{W}\right) + B\sqrt{8\beta_{\mathrm{sw}}(T,\alpha)}\lceil\frac{T}{W}\rceil\left(K + 2\sqrt{KW}\right).$$

On $\mathcal{E}_{\mathrm{sw}}^c$, regret is at most T, which gives the expectation bound. □

D.3 Fallback Γ-Based Non-Stationary Proof

The parameter radius used in the sliding-window Γ-based analysis is

$$\rho_{\mathrm{sw}}(n, t, \alpha) := \frac{\sqrt{8D_{\mathcal{Y}}\beta_{\mathrm{sw}}(t, \alpha)}}{\Gamma\sqrt{n}}.$$

Lemma 7 (Sliding-window parameter width)**.** *Under Assumption 3, on the event of Lemma 5, for every clean window (i, t), every $q \in \mathcal{C}_i^W(t)$ satisfies*

$$|q - \theta_i(t)| \leq \rho_{\mathrm{sw}}(N_i^W(t), t, \alpha) \qquad \textit{whenever } N_i^W(t) \geq 1.$$

Proof. Fix a clean window (i, t) and $q \in \mathcal{C}_i^W(t)$. On the event of Lemma 5, both q and $\theta_i(t)$ belong to $\mathcal{C}_i^W(t)$, and the coordinate bound of Lemma 5 holds for every action-outcome pair. Terms with $N_{i,j}^W(t) = 0$ contribute zero. Squaring the coordinate bound and summing over all outcomes gives

$$\sum_{j=1}^{M} N_{i,j}^W(t)\|g_j(\cdot; q) - g_j(\cdot; \theta_i(t))\|_2^2 \leq 8D_{\mathcal{Y}}\beta_{\mathrm{sw}}(t, \alpha).$$

By the definition of Γ, the left-hand side is at least

$$\Gamma^2 N_i^W(t)(q - \theta_i(t))^2.$$

Combining the last two displays proves the claimed width bound. □

Lemma 8 (Empty-window selections)**.** *For any action sequence, the number of rounds $t \leq T$ such that $N_{I_t}^W(t) = 0$ is at most $K(1 + T/W)$.*

Proof. Fix a task i. If i is selected at round t with $N_i^W(t) = 0$, then task i was not selected in the previous W rounds. Hence two such rounds for the same task are separated by more than W time steps. There are at most $1 + T/W$ of them for each task, and the result follows by summing over i. □

Lemma 9 (Sliding-window summation)**.** *For any action sequence,*

$$\sum_{\substack{t \leq T \\ N_{I_t}^W(t) \geq 1}} \frac{1}{\sqrt{N_{I_t}^W(t)}} \leq \lceil\frac{T}{W}\rceil\left(K + 2\sqrt{KW}\right).$$

Proof. Partition $\{1, \dots, T\}$ into consecutive blocks of length at most W. Fix one block and one task i. Suppose task i is selected m_i times in this block. For the first such selection, the contribution is at most 1 if the window count is nonzero, and it is absent from the sum if the window count is zero. For the r-th selection of task i within the block, $r \geq 2$, the previous $r-1$ selections in the same block all lie inside the length-W window, so $N_i^W(t) \geq r-1$. Therefore the contribution of task i inside the block is at most

$$1 + \sum_{r=2}^{m_i} \frac{1}{\sqrt{r-1}} \leq 1 + 2\sqrt{m_i}.$$

Summing over tasks and using $\sum_i m_i \leq W$,

$$\sum_i (1 + 2\sqrt{m_i}) \leq K + 2\sqrt{K \sum_i m_i} \leq K + 2\sqrt{KW}.$$

Multiplying by the number of blocks proves the lemma. □

D.4 Proof of Theorem 4

Theorem 4 (Dynamic regret of `SW-ACF-UCB`)**.** *Under Assumptions 1 and 3, consider Algorithm 2 with window $W \in [T]$. For every $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$R_T^{\mathrm{dyn}} \leq S(W+1) + K\left(1 + \frac{T}{W}\right) + 2\sqrt{2D_{\mathcal{Y}}}\,\frac{L}{\Gamma}\sqrt{\beta_{\mathrm{sw}}(T,\alpha)}\left\lceil \frac{T}{W} \right\rceil \left(K + 2\sqrt{KW}\right).$$

Proof. Call a round t contaminated if there exists a task change at a time τ with $t - W \leq \tau \leq t$. Each change contaminates at most $W+1$ rounds, so there are at most $S(W+1)$ contaminated rounds. We charge regret at most one to each of them.

On the complement of contaminated rounds, every task window contains observations generated by the current parameter of that task. Let $\mathcal{E}_{\mathrm{sw}}$ be the event from Lemma 5. Then $\Pr(\mathcal{E}_{\mathrm{sw}}^c) \leq \alpha$.

Work on $\mathcal{E}_{\mathrm{sw}}$. Non-contaminated rounds with $N_{I_t}^W(t) = 0$ are bounded by Lemma 8, and each contributes at most one regret. It remains to consider a non-contaminated round with $N_{I_t}^W(t) \geq 1$. If some task had an empty window, its UCB would be $+\infty$, and the algorithm would select an empty-window task. Hence every task, including an oracle-optimal task, has a nonempty window at this round.

Let $i = I_t$, let $\tilde{\theta} = \tilde{\theta}_i^W(t)$, and let $i^\star \in \arg\max_i v(\theta_i(t))$. Since every task has a nonempty clean window at this round, Lemma 5 gives $\theta_{i^\star}(t) \in \mathcal{C}_{i^\star}^W(t)$. Hence

$$U_{i^\star}(t) \geq v(\theta_{i^\star}(t)) = \max_{i,j} \mu(\theta_i(t), j).$$

The algorithm chooses a task with maximal UCB, so $U_i(t) \geq U_{i^\star}(t)$. Moreover J_t maximizes the reward at $\tilde{\theta}$, and therefore

$$U_i(t) = v(\tilde{\theta}) = \mu(\tilde{\theta}, J_t).$$

Applying Lemma 7 once more to the selected task gives

$$|\tilde{\theta} - \theta_i(t)| \leq \rho_{\mathrm{sw}}(N_{I_t}^W(t), t, \alpha).$$

Therefore

$$\max_{i,j} \mu(\theta_i(t), j) - \mu(\theta_{I_t}(t), J_t) \leq L\rho_{\mathrm{sw}}(N_{I_t}^W(t), t, \alpha).$$

Since $t \leq T$ and $\beta_{\mathrm{sw}}(\cdot, \alpha)$ is nondecreasing, $\rho_{\mathrm{sw}}(n, t, \alpha) \leq \rho_{\mathrm{sw}}(n, T, \alpha)$. Using Lemma 9,

$$\begin{aligned} R_T^{\mathrm{dyn}} &\leq S(W+1) + K\left(1 + \frac{T}{W}\right) + L \sum_{\substack{t \leq T \\ N_{I_t}^W(t) \geq 1}} \rho_{\mathrm{sw}}(N_{I_t}^W(t), T, \alpha) \\ &\leq S(W+1) + K\left(1 + \frac{T}{W}\right) \\ &\quad + 2\sqrt{2D_{\mathcal{Y}}}\,\frac{L}{\Gamma}\sqrt{\beta_{\mathrm{sw}}(T,\alpha)}\left\lceil \frac{T}{W} \right\rceil \left(K + 2\sqrt{KW}\right). \end{aligned}$$

On $\mathcal{E}_{\mathrm{sw}}^c$, regret is at most T, which gives the expectation bound. □

E Lower Bound Proofs

E.1 Proof of Theorem 5 (General Minimax Lower Bound)

Theorem 5 (General minimax lower bound)**.** *There exists an instance of the model satisfying Assumptions 1, 2, and 3 such that, for every algorithm and every $T \geq K \geq 2$,*

$$\sup_{\theta_1,\dots,\theta_K \in \Theta} \mathbb{E}[R_T] \geq c\sqrt{KT},$$

where $c > 0$ is a universal numerical constant.

Proof. Embedding ordinary stochastic bandits. Consider the following instance of the model:

$$\Theta = [0,1], \qquad M = 1, \qquad \mathcal{Y}_1 = \{0,1\},$$
$$g_1(1;\theta) = \theta, \qquad g_1(0;\theta) = 1-\theta, \qquad \mu(\theta,1) = \theta.$$

There is only one action, so the learner only chooses a task I_t. The feedback $Y_t \sim \mathrm{Bernoulli}(\theta_{I_t})$ can be viewed as the reward sample itself. Therefore this instance is exactly the classical K-armed Bernoulli bandit with arm means $(\theta_1,\dots,\theta_K)$.

Checking the assumptions. Assumption 1 is immediate because the alphabet is finite and $\theta \mapsto g_1(y;\theta)$ is continuous. Assumption 2 holds with statistic $h_1(y) = y$, mean $m_1(\theta) = \theta$, and $B = 1$. Assumption 3 also holds: for any $\theta, \theta' \in [0,1]$,

$$\|g_1(\cdot;\theta) - g_1(\cdot;\theta')\|_2 = \sqrt{(\theta-\theta')^2 + (\theta'-\theta)^2} = \sqrt{2}|\theta-\theta'|,$$

so $\Gamma = \sqrt{2}$, and $\mu(\theta,1) = \theta$ is 1-Lipschitz.

Applying the standard minimax lower bound. For this subclass, the regret in (1) is precisely the regret of a K-armed Bernoulli bandit against the best arm. The standard minimax lower bound for stochastic bandits [Lattimore and Szepesvári, 2020] states that there exists a universal constant $c > 0$ such that, for every $K \geq 2$, every algorithm, and every $T \geq K$,

$$\sup_{\theta_1,\dots,\theta_K \in [0,1]} \mathbb{E}[R_T] \geq c\sqrt{KT}.$$

Since every algorithm for the structured model induces a valid bandit algorithm on the embedded subclass above, the same lower bound holds for the structured model. □

E.2 Proof of Theorem 6 (Two-Action Lower Bound)

Theorem 6 (Two-action lower bound)**.** *There exists a model with two actions and binary coarsened feedback, satisfying Assumptions 1, 2, and 3, such that the following holds. For every algorithm, every $K \geq 2$, and every $T \geq K$,*

$$\sup_{\theta_1,\dots,\theta_K \in \Theta} \mathbb{E}[R_T] \geq c_{\mathrm{tr}}\sqrt{KT},$$

where $c_{\mathrm{tr}} > 0$ is a universal numerical constant. In this model, one action has larger reward but weaker feedback sensitivity, while the other action has lower reward but more informative feedback.

Proof. The two-action model. Let

$$\Theta = [-1/4, 1/4], \qquad M = 2, \qquad \mathcal{Y}_1 = \mathcal{Y}_2 = \{0,1\}.$$

Define the feedback probabilities by

$$g_1(1;\theta) = \frac{1}{2} + \frac{\theta}{2}, \qquad g_2(1;\theta) = \frac{1}{2} + \theta.$$

Define the rewards

$$\mu(\theta,1) = \frac{1}{2} + \theta, \qquad \mu(\theta,2) = \frac{1}{2} + \theta - \frac{1}{8}.$$

Action 1 is always better by $1/8$, but action 2 has twice the feedback sensitivity. This is the reward–information tradeoff used in the lower-bound construction.

Checking the assumptions. Assumption 1 is immediate. For Assumption 2, take $h_1(y) = h_2(y) = y$. Then $m_1(\theta) = 1/2 + \theta/2$ and $m_2(\theta) = 1/2 + \theta$ are increasing, and both reward functions have slope one. The cross-option reward comparison therefore holds with $B = 2$. For identifiability, for any $\theta, \theta' \in \Theta$,

$$\|g_1(\cdot;\theta) - g_1(\cdot;\theta')\|_2 = \frac{|\theta - \theta'|}{\sqrt{2}}, \qquad \|g_2(\cdot;\theta) - g_2(\cdot;\theta')\|_2 = \sqrt{2}|\theta - \theta'|.$$

Hence Assumption 3 holds with $\Gamma = 1/\sqrt{2}$, and both reward functions are 1-Lipschitz.

Hard instances. Fix $K \geq 2$ and $T \geq K$, and set

$$\varepsilon := \frac{1}{8}\sqrt{\frac{K}{T}}.$$

Then $\varepsilon \leq 1/8$, so all parameters below lie in Θ. Let $\mathbb{P}_0$ denote the instance with $\theta_i = 0$ for every task. For each $u \in [K]$, let $\mathbb{P}_u$ denote the instance with

$$\theta_u = \varepsilon, \qquad \theta_i = 0 \quad (i \neq u).$$

Under $\mathbb{P}_u$, the unique optimal task is u, and the optimal action on that task is action 1. Let

$$N_u(T+1) := \sum_{t=1}^{T} \mathbb{1}\{I_t = u\}.$$

At every round with $I_t \neq u$, the regret is at least ε, regardless of which action is chosen. Therefore

$$\mathop{\mathbb{E}}_{u}[R_T] \geq \varepsilon \mathop{\mathbb{E}}_{u}[T - N_u(T+1)]. \tag{12}$$

This lower bound ignores the additional $1/8$ regret paid when the learner uses the more informative action 2, so it remains valid for arbitrary adaptive use of both actions.

Change of measure. The distributions $\mathbb{P}_0$ and $\mathbb{P}_u$ differ only on observations collected from task u. For either action $j \in \{1, 2\}$, the Bernoulli parameter changes from $1/2$ to $1/2 + a_j\varepsilon$, where $a_1 = 1/2$ and $a_2 = 1$. Since $\varepsilon \leq 1/8$, the elementary Bernoulli KL bound

$$\mathrm{KL}(\mathrm{Bernoulli}(1/2) \,\|\, \mathrm{Bernoulli}(1/2 + x)) \leq 8x^2, \qquad |x| \leq 1/8,$$

implies, by the chain rule for KL divergence under adaptive sampling, that

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_u) \leq 8\varepsilon^2 \mathop{\mathbb{E}}_{0}[N_u(T+1)]. \tag{13}$$

Pinsker's inequality gives

$$\mathop{\mathbb{E}}_{u}[N_u(T+1)] \leq \mathop{\mathbb{E}}_{0}[N_u(T+1)] + T\sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_u)}.$$

Using (13) and averaging over $u \in [K]$,

$$\begin{aligned}\frac{1}{K}\sum_{u=1}^{K} \mathop{\mathbb{E}}_{u}[N_u(T+1)] &\leq \frac{1}{K}\sum_{u=1}^{K} \mathop{\mathbb{E}}_{0}[N_u(T+1)] + \frac{2\varepsilon T}{K}\sum_{u=1}^{K}\sqrt{\mathop{\mathbb{E}}_{0}[N_u(T+1)]} \\ &\leq \frac{T}{K} + 2\varepsilon T\sqrt{\frac{T}{K}},\end{aligned}$$

where the last step uses Cauchy–Schwarz and $\sum_u \mathbb{E}_0[N_u(T+1)] = T$. With $\varepsilon = (1/8)\sqrt{K/T}$, the last display is at most

$$\frac{T}{K} + \frac{T}{4} \leq \frac{3T}{4},$$

because $K \geq 2$. Combining this with (12),

$$\frac{1}{K}\sum_{u=1}^{K} \mathop{\mathbb{E}}_{u}[R_T] \geq \varepsilon\left(T - \frac{3T}{4}\right) = \frac{1}{32}\sqrt{KT}.$$

Hence at least one instance u has expected regret at least $(1/32)\sqrt{KT}$, which proves the theorem. □

E.3 Proof of Theorem 7 (Unstructured Baseline)

Theorem 7 (Unstructured baseline lower bound)**.** *Consider algorithms that discard the shared-parameter structure and treat the KM task-option pairs as unrelated arms. In the resulting unstructured model class, for every such algorithm and every $T \geq KM \geq 2$,*

$$\sup_{\nu} \mathbb{E}_{\nu}[R_T] \geq c_{\mathrm{un}} \sqrt{KMT},$$

where the supremum is over ordinary Bernoulli bandit instances on KM arms and $c_{\mathrm{un}} > 0$ is a universal constant.

Proof. If the learner discards the shared-parameter structure, then each task–action pair (i, j) is treated as a separate arm. There are KM such pairs. The unstructured model class therefore contains the ordinary stochastic bandit class with KM independent Bernoulli arms.

Applying the standard minimax lower bound for stochastic bandits [Lattimore and Szepesvári, 2020] with $A = KM \geq 2$ arms gives a universal constant $c_{\mathrm{un}} > 0$ such that, for every unstructured algorithm and every $T \geq KM$,

$$\sup_{\nu} \mathbb{E}_{\nu}[R_T] \geq c_{\mathrm{un}} \sqrt{KMT},$$

where ν ranges over the KM-armed Bernoulli bandit instances. This establishes the claimed baseline lower bound. □

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 state the new bandit model, algorithms, regret guarantees, and experiment results.

   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: Section 8 discusses the limitations of this work.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Assumptions are stated in Section 3, the theoretical results are stated in Sections 5 and 6, and their complete proofs are given 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: Section 7 and Appendix B give the simulation models, baselines, horizons, random seeds, grids, confidence settings, etc.

   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 simulation code is provided in the supplemental material. No external datasets are required.

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: Section 7 and Appendix B specify the details used in the reported experiments.

   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: All figures show confidence intervals over independent seeds.

   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: Section 7 and Appendix B specify the computer resources used in the experiments.

   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 work is theoretical and simulation-based, uses no human-subject data, and does not train or release datasets.

   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: The paper is foundational theory with synthetic simulations and does not introduce a deployed system or a dataset involving people.

    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 is theoretical and thus does not pose such risks.

    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: [N/A]

    Justification: The experiments use synthetic data and do not rely on external assets.

    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 simulation code is well-documented and a README.md is provided.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with 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: The work does not involve crowdsourcing or research with human subjects.

    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: LLM is used only for writing, editing, or formatting purposes.

    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.