Learning to Defer in Content Moderation: The Human-AI Interplay

Thodoris Lykouris* Wentao Weng†

Abstract

Ensuring successful content moderation is vital for a healthy online social platform where it is necessary to responsively remove harmful posts without jeopardizing non-harmful content. Due to the high-volume nature of online posts, human-only moderation is operationally challenging, and platforms often employ a human-AI collaboration approach. A typical machine-learning heuristic estimates the expected harmfulness of incoming posts and uses fixed thresholds to decide whether to remove the post (classification decision) and whether to send it for human review (admission decision). This can be inefficient as it disregards the uncertainty in the machine-learning estimation, the time-varying element of human review capacity and post arrivals, and the selective sampling in the dataset (humans only review posts filtered by the admission algorithm).

In this paper, we introduce a model to capture the human-AI interplay in content moderation. The algorithm observes contextual information for incoming posts, makes classification and admission decisions, and schedules posts for human review. Non-admitted posts do not receive reviews (selective sampling) and admitted posts receive human reviews on their harmfulness. These reviews help educate the machine-learning algorithms but are delayed due to congestion in the human review system. The classical learning-theoretic way to capture this human-AI interplay is via the framework of *learning to defer*, where the algorithm has the option to defer a classification task to humans for a fixed cost and immediately receive feedback. Our model contributes to this literature by introducing congestion in the human review system. Moreover, unlike work on *online learning with delayed feedback* where the delay in the feedback is exogenous to the algorithm's decisions, the delay in our model is endogenous to both the admission and the scheduling decisions.

We propose a near-optimal learning algorithm that carefully balances the classification loss from a selectively sampled dataset, the idiosyncratic loss of non-reviewed posts, and the delay loss of having congestion in the human review system. To the best of our knowledge, this is the first result for online learning in contextual queueing systems and hence our analytical framework may be of independent interest.

*Massachusetts Institute of Technology, `lykouris@mit.edu`
†Massachusetts Institute of Technology, `wweng@mit.edu`

1 Introduction

Recent advances in Artificial Intelligence (AI) provide the promise of freeing humans from repetitive tasks by *responsive* automation, thus enabling the humankind to focus on more creative endeavors [Yeh23]. One example of automating traditionally human-centric tasks is content moderation targeting misinformation and explicitly harmful content in social platforms such as Facebook [Met23], Twitter [X C23], and Reddit [Red23]. Historically (e.g., forums in the 2000s), human reviewers would monitor all exchanges to detect any content that violated the community standards [Rob19]. That said, the high volume of posts in current platforms coupled with the advances in AI has led platforms to automate their content moderation, harnessing the responsiveness of AI.[1] This trend of automating traditionally human-centric tasks applies broadly beyond content moderation; for example, speedy insurance claim process [Pin23] and domain-specific generative AI copilots [GO23].

However, excessive use of automation in such human-centric applications significantly reduces the *reliability* of the systems. AI models are trained based on historical data and therefore their predictions reflect patterns observed in the past that are not always accurate for the current task. On the other hand, humans' cognitive abilities and expertise make humans more attune to correct decisions. In content moderation, particular posts may have language that is unclear, complex, and too context-dependent, obscuring automated predictions [Met22]. Similarly, AI models may wrongfully reject a valid insurance claim [Eub18] and Large Language Model copilots may hallucinate non-existing legal cases [Nov23]. These errors can have significant ethical and legal repercussions.

The *learning to defer* paradigm is a common way to combine the responsiveness of AI and the reliability of humans. When a new job arrives, the AI model classifies it as *accept* or *reject*, and determines *whether to defer* the job for human review by admitting it to a corresponding queue (in content moderation, incoming jobs correspond to new posts and the classification decision pertains to whether the post is kept on or removed from the platform). When a human reviewer becomes available, the AI model determines which job to schedule for human review. As a result, the AI model directly determines which posts will be reviewed by humans.

At the same time, humans also affect the AI model as their labels for the reviewed jobs form the dataset based on which the AI is trained, creating an *interplay* between humans and AI. Hence, the AI model's admission decisions are not only useful for correctly classifying the current jobs but are also crucial for its future prediction ability, a phenomenon known as *selective sampling.*

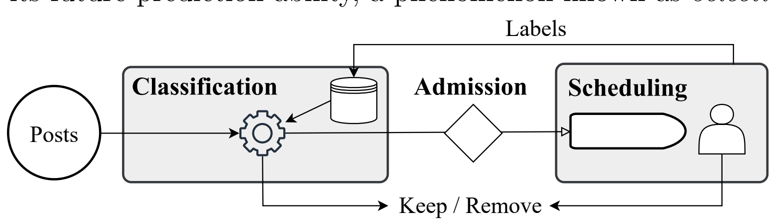

Figure 1: Pipeline of human-AI interplay in content moderation

In this paper, we model this *human-AI interplay* in the context of content moderation (though our insights can be transferable to other human-centric settings) and pose the following question:

How can we make classification, admission, and scheduling decisions that combine the responsiveness of AI and the reliability of humans?

[1]Facebook reports 60 billion posts per month with about 15,000 reviewers (Apr.-Jun. 2023 [Met23]).

1.1 Our contributions

Summary of operational insights. The theoretical results from this paper contribute to the following three operational insights on how to achieve a responsive and reliable content moderation system with efficient scheduling and contextual learning in a queueing system.

- First, existing content moderation practice typically uses static thresholds on the sample average of a post's harmfulness to decide which posts to admit for human reviews (see Appendix A). We demonstrate that such approaches can be inefficient by failing to consider the variance of post harmfulness and the fluctuation in post arrivals and human capacity prominently seen in practice (see [MSA$^+$21, Figure 2].) Instead, we propose a congestion-aware admission rule that balances posts' idiosyncrasy loss when we rely on AI prediction and the additional congestion in the human review system caused by admitting a new post.

- Second, a classical way to enable online learning on post harmfulness is to create optimistic estimates for the admission decision to balance exploration and exploitation. This optimistic approach ignores the downstream effect of human data on future AI classification decisions. Another approach to address learning is to separate a portion of human capacity to label posts offline. This offline labeling approach cannot adapt to the fluctuating post arrival patterns and can waste human capacity. To address the inefficiencies of both approaches, our work proposes an efficient learning algorithm by combining optimism with label-driven admission. A crucial challenge that we overcome is the delay in label collection process caused by congestion in the human review system, which is coupled with the admission decisions.

- Third, classical scheduling algorithms keep separate queues for jobs of different types. In applications like content moderation, jobs (posts) come with contextual information that defines a myriad number of types. Maintaining separate queues for all types is thus wasteful and can lead to an overly congested system. To the best of our knowledge, our work is the first to show that aggregating jobs with similar service requirement enables provably efficient scheduling even when job types are defined by contextual information.

We note that the above insights transcend the content moderation application and can inform responsive and reliable automation of other human-centric tasks with limited human capacity.

Learning to defer with capacity constraints. To formally tackle the problem, we introduce a model that combines *learning to defer* with soft capacity constraints via queueing delays. To the best of our knowledge, the impact of limited capacity for learning to defer has not been considered in the literature (see Section 1.2 for further discussion). In particular, in each period $t = 1, \dots, T$, a post j arrives with a type $k(j)$ drawn from a time-varying distribution over K types. A post j has an unknown cost $c_j = h_{k(j)} + \xi_j$ where $\xi_j \sim \mathcal{N}(0, \sigma^2_{k(j)})$ is a zero-mean idiosyncratic Gaussian noise while h_k and σ_k^2 are the expectation and variance of the cost of type-k posts. To capture the difference in reliability, we assume that the AI model only observes the type of the post; the true cost c_j is only observable via human reviews. To capture the impact of limited human capacity, we assume that posts that are admitted (deferred) by the AI wait in a review queue. At the end of a period, one post in the queue is scheduled for review; the review is completed in this period with a probability $\mu_k N(t)$ that depends on both the specific post type k and the fluctuating capacity of reviewers $N(t)$. We assume that a type-k post is viewed only within ℓ_k periods since its arrival and only if it is not removed.

We measure the *loss* of a policy by comparing it to an omniscient benchmark that keeps benign posts ($c_j \leq 0$) and removes harmful posts ($c_j > 0$). Note that this omniscient benchmark is not limited by capacity, which means that we need a capacity-constrained benchmark to argue about the efficiency of our policies. As a result, we define the *average regret* of a policy by comparing its time-averaged loss to a fluid benchmark that incorporates time-varying capacity constraints.

Balancing idiosyncrasy loss with delays. As typical in complex learning settings, we start by assuming knowledge of the latent parameters (i.e., the expected costs h_k). The classification decision is then to keep the post j if and only if the expected cost of its type is negative: $h_{k(j)} < 0$. Of course, deferring a post to human reviewers can improve upon this ex-ante classification decision as human reviewers can observe the true cost. The admission decision should thus identify the posts that would benefit the most from this more refined observation. If all types have equal variance then the human reviewing capacity is more efficiently allocated when it focuses on borderline posts, i.e., $h_{k(j)}$ close to 0, as borderline types have higher probability that the true cost of their posts has a different sign from their expected cost (which would lead to a different classification by the omniscient benchmark). This intuition drives the design of admission policies in existing content moderation practice [MSA+21, ABB+22] that operate based on two thresholds: a post j is rejected (resp. accepted) without admission if its expected cost $h_{k(j)}$ is above the higher (resp. below the lower) threshold and is admitted for human review if it lies between the two thresholds. A more detailed comparison to these works is provided in Appendix A.

That said, even when all expected costs h_k are known, there are two important shortcomings in these two-threshold heuristics. First, when the variance of different types is heterogeneous, the admission decision should not be restricted to the expected costs but rather also take variance into account. In particular, it may be more beneficial to defer to human reviewers post from a non-borderline type k with high idiosyncratic variance σ_k^2 instead of wasting human capacity on more deterministic borderline types. Second, even if we operate with variance-aware scores, using static thresholds does not allow the system to adapt to the time-varying arrival patterns of posts and the fluctuating human capacities that arise in practice (see [MSA+21, Figure 2]). In particular, in certain time periods, the number of borderline posts may be either too large resulting in high delays for admitted posts due to the overflow in the review system or too small resulting in misclassification for some non-borderline posts that could have been prevented if they were admitted.

Our approach (Section 3) quantifies and balances the two losses hinted above (due to idiosyncrasy and delay) and is henceforth called Balanced Admission Control for Idiosyncrasy and Delay or BACID as a shorthand. We first calculate the expected ex-ante per-period loss of keeping a post on the platform (resp. removing it) using the distributional information of the true cost c_j; this loss is denoted by r_k^O (resp. r_k^R). Hence, if a post is not admitted to human review, its per-period loss is $r_k = \min(r_k^O, r_k^R)$ and thus its *idiosyncrasy loss* that aggregates over its lifetime is $r_k \ell_k$. On the other hand, admitting a post incurs *delay loss*, which is the increase in congestion for future posts due to the limited human capacity. Motivated by the drift-plus-penalty literature [Nee22], we estimate the delay loss of a post by the number of posts of the same type currently waiting for reviews. We then admit a post if (a weighted version of) its idiosyncrasy loss (which one can avoid by admission) *exceeds* the estimated delay loss (which occurs due to the admission). For scheduling, we choose in each period the type with the most number of waiting posts (weighted by the difficulty to review this type) to review. We show that while heuristics like AI-Only or Human-Only have average regret at least $\Omega(\ell_{\max})$ (Propositions 1 and 2), where $\ell_{\max}$ is the maximum lifetime, BACID achieves a $O(\sqrt{K\ell_{\max}})$ average regret even with general time-varying arrivals and capacities (Theorem 1). This bound is optimal on the dependence on $\ell_{\max}$ as we also show a lower bound $\Omega(\sqrt{\ell_{\max}})$ for any deterministic policy (Theorem 2).

Efficient learning with selective sampling, optimism and forced scheduling. To deal with unknown expected costs h_k, a typical approach in the bandit literature is to use an *optimistic* estimator for the unknown quantities. The unknown quantity that affects our admission rule is r_k, which depends on the unknown expected cost h_k. Following the optimistic approach, we first create a confidence interval $[\underline{h}_k, \bar{h}_k]$ for the unknown expected cost h_k based on reviewed posts. We then compute an optimistic estimator $\bar{r}_k$ using the parameter $\widetilde{h}_k \in [\underline{h}_k, \bar{h}_k]$ that maximizes $r_k(\widetilde{h}_k)$.The optimistic admission rule then admits a new post if (a weighted version of) the optimistic idiosyncrasy loss, $\bar{r}_k \ell_k$, exceeds its estimated delay loss. The high-level idea behind optimism is that, if a post is admitted (due to the optimistic estimation), we collect additional samples for this post which result in shrinking the confidence interval and eventually leading the confidence interval to converge to the true parameter h_k. Without considering heterogeneity in variance, [ABB+22] adopts a similar optimism-only heuristic to address the online learning nature of h_k.

However, optimism-only learning heuristics disregard the selective sampling nature of the feedback. When there is uncertainty on h_k, the benefit of admitting a post is not restricted to avoiding the idiosyncrasy loss of the *current* post, but extends to having one more data points that can improve the classification decision of *future* posts. This benefit does not arise when h_k is known because the best classification decision is then clear (keep a post if and only if $h_k \le 0$). With unknown h_k, there is an additional *classification loss* due to incorrect estimation of the sign of h_k. Given that optimism-only heuristics emulate the admission rule with known h_k, they disregard this classification loss and more broadly the positive externality that labels for the current post have on future posts. This challenge highlights that optimism-only heuristics such as Upper Confidence Bound (UCB) do not employ *forward-looking exploration* but rather *optimistically myopic exploitation*. They are *myopic* in their nature: they create a confidence interval around the parameter of interest and are optimistic with respect to this confidence interval, but they make an *exploit* decision based on those optimistic estimates. Concretely, given the most optimistic estimates about the underlying parameters, these algorithms select the action that *myopically* maximizes the contribution of the current post without actively considering any positive externality of the labels on future posts. We formally illustrate this inefficiency of optimism-only learning heuristics in Proposition 3.

To circumvent the above issue, we design an online learning version of BACID (Section 4) , which we term OLBACID, that augments the optimistic admission with *label-driven admission* and *forced scheduling*. When a post arrives, we estimate its potential classification loss based on its current confidence interval; if this is higher than a threshold, we admit the post to a *label-driven queue* and prioritize posts in this queue via *forced scheduling* to ensure enough labels for future classification decisions. To avoid exhausting human capacity, we limit the number of posts in the label-driven queue to one at any time. We show that OLBACID enjoys an average regret of $\tilde{O}(\sqrt{K\ell_{\max}} + K\ell_{\max}/\sqrt{T})$ (Theorem 3), addressing the inefficiency of optimism-only heuristics.

Context-defined types with type aggregation and contextual learning. To remove the dependence on K, which can be large when types are defined based on contextual information (e.g., word embeddings of the contents), we provide a contextual bandit extension of our algorithm. In particular, we consider a linear contextual setting where $h_k = \boldsymbol{\phi}_k^\intercal \boldsymbol{\theta}^\star$ where $\boldsymbol{\phi}_k$ is a known $d-$dimensional feature vector for each type k and $\boldsymbol{\theta}^\star$ is an unknown $d-$dimensional vector; this linear cost structure is similar to the one used in [ABB+22].

Our contextual extension (Section 5), which we term COLBACID, addresses the two facets where the previous algorithms require dependence on the number of types K. First, recall that BACID estimates the delay loss of admitting a post by the number of same-type posts in the review system. This naturally creates a dependence on the number of types K. Instead, we aggregate types into G groups such that any two types k, k' in the same group satisfy $|\mu_k - \mu_{k'}| \leq \Delta$ for some parameter Δ. This aggregation enables us to estimate the delay loss by the number of same-group waiting posts, thus removing the aforementioned dependence on K. Second, recall that the learning in OLBACID creates a separate confidence interval for each type and refining those intervals introduces a dependence on K. By employing techniques from the linear contextual bandit literature [APS11], we replace this dependence with the dimension d. We note that our setting has the additional complexity that the feedback is received after a queueing delay which is endogenous to the admission and scheduling decisions of the algorithm (see Appendix D.1 for a comparison to contextual bandit with delays where delays are assumed to be exogenous). To handle this challenge, our scheduling employs a first-come-first-serve review order for posts in the same group and we only use data from the same group to estimate the expected cost of a type, which controls the delay by the number of same-group posts in the review system.

Our instance-dependent guarantee $\tilde{O}(\sqrt{G\ell_{\max}}+\ell_{\max}(\Delta+d\sqrt{G/T})$ (Theorem 4) scales optimally with the maximum lifetime $\ell_{\max}$ and avoids the dependence on K. We also provide a worst-case guarantee $\tilde{O}(\ell_{\max}^{2/3} + \ell_{\max}^{7/6} d/\sqrt{T})$ with no dependence on G and Δ at the cost of a worse dependence on $\ell_{\max}$ (Corollary 1). To the best of our knowledge, this is the first result for online learning in contextual queueing systems and hence our analytical framework may be of independent interest.

1.2 Related work

Human-AI collaboration. The nature of human-AI collaboration can be broadly classified into two types: *augmentation* and *automation* [BM17, RK21]. In particular, augmentation represents a "human-in-the-loop" type workforce where human experts combine machine learning with their own judgement to make better decisions. Such augmentation is often found in high-stake settings such as healthcare [LLL22], child maltreatment hotline screening [DFC20], refugee resettlement [BP22, AGP+23] and bail decisions [KLL+18]. A general concern for augmentation relates to humans' compliance patterns to [BBS21, LLL22], the impact of such patterns to decision accuracy [KLL+18, MS22] and the impact on fairness [MSG22, GBB23]. Automation, on the other hand, concerns the use of machine learning in place of humans and is widely applied in human and social services where colossal demands overwhelm limited human capacity, such as in data labelling [VLSV23], content moderation [GBK20, MSA+21] and insurance [Eub18]. Full automation is clearly undesirable in these applications as machine learning can err. A natural question is how one can better utilize the limited human capacity when machine is uncertain about the prediction. The literature on "learning to defer", which studies when machine learning algorithms should defer decisions to downstream experts, tries to answer this question and is where our work fits in.

Assuming that humans have perfect prediction ability but there is limited capacity, classical learning to defer has two streams of research, *learning with abstention* and *selective sampling*. Although these two streams have been studied separately (with a few exceptions, see discussion below), our work provides an endogenous approach to connect them. In particular, learning with abstention can be traced back to [Cho57, Cho70] and studies an offline classification problem with the option to not classify a data point for a fixed cost. For example, in content moderation, this corresponds to paying a fixed fee to an exogenous human reviewer to review a post. Since optimizing the original problem is computational infeasible, an extensive line of work investigates suitable surrogate loss functions and optimization methods; see [BW08, EW10, CDM16] and references therein. For online learning with expert advice, it is shown that even a limited amount of abstention allows better regret bound than without [SZB10, LLWS11, ZC16, NZ20]; [CDG+18] studies a similar problem but allows experts to also abstain. Selective sampling (or label efficient prediction) considers a different model where the algorithm makes predictions for every arrival; but the ground truth label is unavailable unless the algorithm queries for it; if queried, the label is available immediately [CLS05]. The goal is to obtain low regret while using as few queries possible. Typical solutions query only when a confidence interval exceeds a certain threshold [CLS05, CBGO09, OC11, DGS12]. There is recent work connecting the two directions by showing that allowing abstention of a small fixed cost can lead to better regret bound for selective sampling [ZN22, PZ22]. However, these works treat the abstention cost and selective sampling exogenously. Existing work neglects the impact of limited human capacity to learning to defer[LSFB22]. Our model serve as one step to capture it by endogenously connecting both costly abstention and selective sampling via delays. In particular, our admission component determines both abstention and selective sampling. In addition, the cost of abstention is dynamically affected by the delay in getting human reviews whereas the avoidance of frequent sampling (to get data) is captured by its impact on the delay.

Although the above line of work as well as ours assumes perfect labels from human predictions, a more recent stream of work on "learning to defer" considers imperfect human predictions and is focused on combining prediction ability of experts and learning algorithms; this moves towards the augmentation type of human-machine collaboration. In particular, [MCPZ18, MS20, WHK20, CMSS22] study a setting with an offline dataset and expert labels. The goal is to learn both a classifier, which predicts outcome, and a rejector that predicts when to defer to human experts. The loss is defined by the machine's classification loss over non-deferred data, humans' classification loss over deferred data, and the cost to query experts. [DKGG20, DOZR21] study a setting where given expert loss functions, the algorithm picks a size-limited subset of data to outsource to humans and solves a regression or a classification problem on remaining data, with a goal to minimize the total loss. [RBC+19] extends the model by allowing the expert classification loss to depend on human effort and further considering an allocation of human effort to different data points. [KLK21, VBN23, MMZ23] consider learning to defer with multiple experts.

Bandits with knapsacks or delays. Restricting our attention to admission decisions, our model bears similar challenges with the literature on bandits with knapsacks or broadly online learning with resource constraints, which finds applications in revenue management [BZ12, WDY14, FSW18]. In particular, for bandits with knapsacks, there are arrivals with rewards and required resources from an unknown distribution. The algorithm only observes the reward and required resources after admitting an arrival and the goal is to obtain as much reward as possible subject to resource constraints [BKS18, AD19]. A typical primal-dual approach learns the optimal dual variables of a fixed fluid model and explores with upper confidence bound [BLS14, LSY21]; these extend to contextual settings [WSLJ15, AD16, ADL16, SSF23], general linear constraints [PGBJ21, LLSY21] and constrained reinforcement learning [BDL+20, Che19]. These results cannot immediately apply to our setting for two reasons. First, the resource constraint in our model is dynamically captured by time-varying queue lengths instead of a single resource constraint over the entire horizon; thus learning fixed dual variables is insufficient for good performance. Second, our admission decisions must also consider the effect on classification; thus relying only on optimism-based exploration is insufficient (see Section 4.1).

The problem of bandits with delays is related to our setting where feedback of an admitted post gets delayed due to congestion. Motivated by conversion in online advertising, bandits with delays consider the problem where the reward of each pulled arm is only revealed after a random delay independently generated from a fixed distribution [DHK$^+$11, Cha14]. Assuming independence between rewards and delays as well as bounded delay expectation, [JGS13, MLBP15] propose a general reduction from non-delay settings to their delayed counterpart. Subsequent papers consider censored settings with unobservable rewards [VCP17], general (heavy-tail) delay distributions [MVCV20, WW22] and reward-dependent delay [LSKM21]. [VCL$^+$20, BXZ23] study (generalized) linear contextual bandit with delayed feedback. The key difference between bandits with delays and our model is that delays for posts in our model are not independent across posts due to queueing effect; we provide a more elaborate comparison to those works in Appendix D.1.

Learning in queueing systems. Learning in queueing systems can be classified into two types: 1) learning to schedule with unknown service rates to obtain low delay; 2) learning unknown utility of jobs / servers to obtain high reward in a congested system. For the first line of research, an intuitive approach to measure delay suboptimality is via the queueing regret, defined as the difference in queue lengths compared with a near-optimal algorithm [WX21]. The interest for queueing regret is in its asymptotic scaling in the time horizon, and it is studied for single-queue multi-server systems [Wal14, KSJS21, SSM21], multi-queue single-server systems [KAJS18], load balancing [CJWS21], queues with abandonment [ZBW22] and more general markov decision processes with countable infinite state space [AS23]. As an asymptotic metric may not capture the learning efficiency of the system, [FLW23b] considers an alternative metric (cost of learning in queueing) that measures transient performance by the maximum increase in time-averaged queue length. This metric is motivated by works that study stabilization of queueing systems without knowledge of parameters. In particular, [NRP12, YSY23, NM23] combine the celebrated MaxWeight scheduling algorithm [Tas92] with either discounted UCB or sliding-window UCB for scheduling with time-varying service rates. [FHL22, GT23] study decentralized learning with strategic queues and [SSM19, SBP21, FLW23a] consider efficient decentralized learning algorithms for cooperative queues. Although most work for learning in queues focus on an online stochastic setting, [HGH23, LM18] study online adversarial setting and [SGVM22] considers an offline feature-based setting.

Our work is closer to the second literature that learns job utility in a queueing system. In particular, [MX18, SGMV20] consider Bayesian learning in an expert system where jobs are routed to different experts for labels and the goal is to keep the expert system stable. [JKK21, HXLB22, FM22] study a matching system where incoming jobs have uncertain payoffs when served by different servers and the objective is to maximize the total utility of served jobs within a finite horizon. [JSS22b, JSS22a, CLH23] investigate regret-optimal learning algorithms and [LJWX23] studies randomized experimentation for online pricing in a queueing system.

We note that, although most performance guarantees for learning in queueing systems deteriorate as the number of job types K increases, our work allows admission, scheduling and learning in a many-type setting where the performance guarantee is independent of K. Although prior work obtains such a guarantee in a Bayesian setting, where the type of a job corresponds to a distribution over a finite set of labels [AZ09, MX18, SGMV20], their service rates are only server-dependent (thus finite). In contrast, our work allows for job-dependent service rates. In addition, a Bayesian setting does not immediately capture the contextual information between jobs that may be useful for learning. We note that [SGVM22] consider a multi-class queueing system where a job has an observed feature vector and an unobserved job type, and the task is to assign jobs to a fixed number of classes with the goal of minimizing mean holding cost, with known holding cost rate for any type. They find that directly optimizing a mapping from features to classes can greatly reduce the holding cost, compared with a predict-then-optimize approach. Different from their setting, we consider an online learning setting where reviewing a job type provides information for other types. This creates an explore-exploit trade-off complicated with the additional challenge that the feedback experiences queueing delay. To the best of our knowledge, our work is the first result for efficient online learning in a queueing system with contextual information.

Joint admission and scheduling. When there is no learning, our problem becomes a joint admission and scheduling problem that is widely studied in wireless networks and the general focus is on a decentralized system [KMT98, LSS06]. Our method is based on the drift-plus penalty algorithm [Nee22], which is a common approach for joint admission and scheduling, first noted as a greedy primal-dual algorithm in [Sto05]. The intuition is to view queue lengths as dual variables to guide admission; [HN11] formalizes this idea and exploits it to obtain better utility-delay tradeoffs.

2 Model

We consider a T-period discrete-time system to model content moderation on a platform. Each post has a type in a set $\mathcal{K}$ with $|\mathcal{K}| = K$. In period $t = 1, \ldots, T$, a new post $j(t) = j$ arrives with probability $\lambda(t)$; its type is $k(j) = k$ with probability $\frac{\lambda_k(t)}{\lambda(t)}$ where $\lambda_k(t)$ is the arrival rate of type k. We use $\Lambda_k(t) = \mathbb{1}(k(j(t)) = k)$ to denote whether a type-k post arrives in period t. If there is no new post, we denote $j(t) = \perp$.

The per-view harmfulness of a post is captured by its cost $c_j \in \mathbb{R}$; this quantity being negative means that the post is healthy. Condition on having type k, the cost of post j is given by $c_j = h_{k(j)} + \xi_j$ where h_k is the average cost of type-k posts and $\xi_j \sim \mathcal{N}(0, \sigma_k^2)$ is independent Gaussian noise. Although the type $k(j)$ is observable, we assume the cost c_j remains unknown until it is reviewed by a human. We next define the total harmfulness of a post. For every post of type k, we assume it has a lifetime ℓ_k and a per-period view v_k, i.e., if a post j arrives in period $t(j)$, then for periods $\{t(j), \ldots, \min(t(j) + \ell_k - 1, T)\}$, it will receive v_k views given that it is *on the platform*.[2] We define $O_j(t) \in \{0, 1\}$ such that it is equal to 1 when post j is on the platform for period t. The total harmfulness of a post j is given by $c_j v_k \sum_{t=t(j)}^{\min(t(j)+\ell_k-1,T)} O_j(t)$. If c_j is known, this quantity is minimized by setting $O_j(t) = 1$ if $c_j \leq 0$ or 0 if $c_j > 0$.

2.1 The human-AI pipeline

When the cost is unknown, the platform resorts to a human-AI pipeline (Figure 1) by making three decisions in any period t, *classification*, *admission*, and *scheduling*:

- *Classification.* Upon the arrival of a new post j of type k, the platform first makes a classification decision $Y_k(t)$ such that the post stays on the platform if $Y_k(t) = 1$ or is removed if $Y_k(t) = 0$. Then $O_j(\tau) = Y_k(t)$ for $\tau \geq t$ unless the decision is reversed by a human reviewer.
- *Admission.* For this new post, the platform may also decide to admit it into the human review

[2] Our analysis easily extends to the case where the view per period is a random variable with mean v_k.

system; $A_k(t) = 1$ if it is admitted and 0 if not. If $A_k(t) = 1$, then post j is included into an initially empty review queue $\mathcal{Q}$. We define $\mathcal{A}(t) = j(t)$ if $A_k(t) = 1$ and $\mathcal{A}(t) = \perp$ otherwise.

- *Scheduling.* At the end of period t, the platform selects a post from the review queue $\mathcal{Q}$ for humans to review; we denote this post by $M(t)$. To capture the service capacity, we assume that we have $N(t)$ reviewers in period t. Define $\psi_k(t) = \mathbb{1}(k(M(t)) = k)$ indicating whether humans review a type-k post. If $M(t)$ is of type k, it is reviewed with service probability equal to $N(t)\mu_k \le 1$ where μ_k is a known type-specific quantity;[3] we let $S_k(t) = 1$ and $\mathcal{S}(t) = M(t)$ if $M(t)$ is reviewed and $S_k(t) = 0, \mathcal{S}(t) = \perp$ otherwise. When $M(t)$ is reviewed, we assume human reviewers observe the exact cost $c_{M(t)}$ and reverse the previous classification decision if wrong, i.e., $O_{\mathcal{S}(t)}(\tau) = \mathbb{1}\left(c_{\mathcal{S}(t)} \le 0\right)$ for $\tau > t$. Let $\mathcal{Q}(t)$ be the set of posts in the review queue at the beginning of period t. Then $\mathcal{Q}(t+1) = \mathcal{Q}(t) \cup \{\mathcal{A}(t)\} \setminus \{\mathcal{S}(t)\}$. In addition, the data set of reviewed posts at the beginning of period t, $\mathcal{D}(t)$, is given by $\mathcal{D}(t) = \{(\mathcal{S}(\tau), c_{\mathcal{S}(\tau)})\}_{\tau < t}$.

We next discuss information a feasible policy can rely on. We assume that $\{\sigma_k, \ell_k, v_k\}_{k \in \mathcal{K}}$ are known; but h_k is *unknown* initially and must be learned via samples from human reviewers, i.e., the data set $\mathcal{D}(t)$. Since there are many types of contents, we assume that a type k comes with a known d-dimensional feature vector $\boldsymbol{\phi}_k \in \mathbb{R}^d$. The average cost satisfies a linear model $h_k = \boldsymbol{\phi}_k^\intercal \boldsymbol{\theta}^\star$ where $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is an unknown vector. We assume that $\|\boldsymbol{\theta}^\star\|_2, \|\boldsymbol{\phi}_k\|_2$ are bounded by a known value U, $|h_k|$ is bounded by a known value $H \ge 1$. We also define $\eta = \min(1, \min_{k \in \mathcal{K}} |h_k|)$ which is the margin of the average cost capped at 1 (and can be 0). The platform has no information of $\{\lambda_k(t), N(t)\}_{t \in [T]}$. A policy is *feasible* if its decisions for any period t are only based on the observed sample path $\{j(\tau), \mathcal{A}(\tau), \mathcal{S}(\tau)\}_{\tau < t} \cup \{j(t)\}$, the data set $\mathcal{D}(t)$ and the initial information $\{\sigma_k, \ell_k, v_k, \boldsymbol{\phi}_k\}_{k \in \mathcal{K}}$.

2.2 Objectives and Benchmark

Recall that the total harmfulness of a post is $c_j v_k \sum_{t=t(j)}^{\min(t(j)+\ell_k-1,T)} O_j(t)$ and the optimal clairvoyant that knows c_j will set $O_j^\star(t) = \mathbb{1}(c_j \le 0)$ for any j. Letting $\mathcal{J}(t)$ be the set of posts that arrive in the first t periods, the loss of a policy π with respect to this clairvoyant is thus:

$$\mathcal{L}^\pi(T) \coloneqq \sum_{j \in \mathcal{J}(t)} c_j v_{k(j)} \sum_{t=t(j)}^{\min(T, t(j)+\ell_{k(j)}-1)} (O_j(t) - \mathbb{1}(c_j \le 0)).$$

Due to the variance σ_k in posts of the same type k, any post that is not reviewed incurs a positive loss in expectation even if the average h_k is known. In addition, humans cannot review all posts due to limited capacity. As a result, aiming for vanishing loss is unattainable and we need to define a benchmark that captures both the effect of variance and capacity constraints. Our benchmark also needs to accommodate the non-stationarity in arrival rates $\{\lambda(t)\}$ and review capacity $\{N(t)\}$.

We consider a deterministic (fluid) benchmark where for every period t, there is a mass of $\lambda_k(t)$ posts from type k. The platform admits a mass of $a_k(t)$ posts to review and leaves a mass of $\lambda_k(t) - a_k(t)$ posts classified based on h_k and not reviewed at all. Admitted posts receive human reviews immediately and thus incur no costs. For a non-admitted post j of type k, the expected per-period loss of leaving this post on the platform is $r_k^O \coloneqq v_k \mathbb{E}\left[(c_j)^+ \mid k(j) = k\right]$ and the loss of rejecting this post is $r_k^R \coloneqq v_k \mathbb{E}\left[-(c_j)^- \mid k(j) = k\right]$, where we denote $x^+ = \max(x, 0)$ and $x^- = \min(x, 0)$. With the assumption of Gaussian noise, these quantities have explicit expressions (see Section 3).

[3] We aggregate the service power of reviewers for simplicity.

The per-period loss of an non-admitted post in the benchmark is thus $r_k := \min(r_k^O, r_k^R)$, by rejecting a post if $h_k > 0$ and keeping it if $h_k \le 0$. Across its lifetime ℓ_k , each non-admitted type-k post incurs loss $r_k \ell_k$. The expected total loss is then $\sum_{t=1}^T \sum_{k\in\mathcal{K}} r_k \ell_k(\lambda_k(t) - a_k(t))$.

To motivate our fluid benchmark, we start from the stationary case where $\lambda_k(t) \equiv \lambda_k$ and $N(t) \equiv N$ across any period $t \le T$. The classical fluid benchmark corresponds to a linear program (LP) as in (1). In particular, the objective is the expected total loss and the first constraint captures the capacity constraint requiring all admitted posts are reviewed eventually in expectation. An alternative formulation is to satisfy the capacity constraint for *each period* as in (2). With the stationary property that $\lambda_k(t) \equiv \lambda_k$ and $N(t) \equiv N$, the two LPs are equivalent.

$$
\begin{aligned}
&\min_{\{a_k(t)\},\{\nu_k(t)\}} \sum_{t=1}^T \sum_{k\in\mathcal{K}} r_k \ell_k(\lambda_k - a_k(t)), \text{ s.t.} \\
&\boldsymbol{\sum_{t=1}^T a_k(t) \le \mu_k \sum_{t=1}^T N(t)\nu_k(t), \forall k \in \mathcal{K}} \\
&a_k(t) \le \lambda_k(t),\ \nu_k(t) \ge 0,\ \forall k \in \mathcal{K}, t \in [T] \\
&\sum_{k\in\mathcal{K}} \nu_k(t) \le 1,\ \forall t \in [T].
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
&\min_{\{a_k(t)\},\{\nu_k(t)\}} \sum_{t=1}^T \sum_{k\in\mathcal{K}} r_k \ell_k(\lambda_k - a_k(t)), \text{ s.t.} \\
&\boldsymbol{a_k(t) \le \mu_k N(t)\nu_k(t), \forall k \in \mathcal{K}, t \in [T]} \\
&a_k(t) \le \lambda_k(t),\ \nu_k(t) \ge 0,\ \forall k \in \mathcal{K}, t \in [T] \\
&\sum_{k\in\mathcal{K}} \nu_k(t) \le 1,\ \forall t \in [T].
\end{aligned} \tag{2}
$$

Extending the fluid benchmark to a non-stationary setting is non-trivial. In particular, the benchmark in (1) allows an early post to be reviewed by human capacity throughout the entire horizon. Its ignorance of delays in human reviews can greatly underestimate the loss any policy must incur, as illustrated in Example 1. Alternatively, using the benchmark in (2) can overestimate the loss of an optimal policy. It requires an admitted post to be reviewed in the same period it arrives, even though an admitted post can wait in the review queue and be reviewed by later human capacity.

Example 1. *Consider a setting with one type of posts and $r_1 = 1$. Arrival rates and review capacity are time-varying such that $\lambda(t) = 1, N(t) = 0$ for $t \le (T - \ell_1)/2$; $\lambda(t) = 0, N(t) = 0$ for $t \in ((T - \ell_1)/2, (T + \ell_1)/2)$; and $\lambda(t) = 0, N(t) = 1$ for $t \ge (T + \ell_1)/2$. Benchmark (1) will always give a loss of 0, indicating an "easy" setting to moderate. However, the uneven capacity indeed makes it a "difficult" setting where no post can be reviewed within its lifetime under any policy.*

Given that (1) is too strong and (2) is too weak as the benchmark, we consider a series of fluid benchmarks that interpolate between them, which we call the $w-$*fluid benchmarks*. In particular, we assume that the interval $\{1, \ldots, T\}$ can be partitioned into consecutive windows of sizes at most w, such that the admission mass of the benchmark is no larger than the service capacity in each window. In other words, a $w-$fluid benchmark limits the range of human capacity used to review an admitted post: a larger window size allows an admitted post to be reviewed by later human capacity. Benchmarks (1) and (2) are special cases where $w = T$ and $w = 1$ respectively.

Formally, consider the set of feasible window partitions

$$
\mathcal{P}(w) = \{\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{I+1})\colon 1 = \tau_1 < \cdots < \tau_{I+1} = T + 1;\ \tau_{i+1} - \tau_i < w,\ \forall i \le I;\ I \in \mathbb{N}\}.
$$

The w-fluid benchmark minimizes the expected total loss over admission vector $\{a_k(t)\}_{k\in\mathcal{K}, t\in[T]}$, partition $\boldsymbol{\tau}$ and probabilistic service vector $\{\nu_k(t)\}_{k\in\mathcal{K},t\in[T]}$ that satisfy w-capacity constraints:

$$\begin{aligned}
\mathcal{L}^\star(w,T) \coloneqq \; & \min_{\substack{\boldsymbol{\tau}\in\mathcal{P}(w)\\ \{a_k(t)\},\{\nu_k(t)\}}} \sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\lambda_k(t)-a_k(t)) \\
\text{s.t.} \; & \boldsymbol{\sum_{t=\tau_i}^{\tau_{i+1}-1} a_k(t) \le \mu_k \sum_{t=\tau_i}^{\tau_{i+1}-1} N(t)\nu_k(t),\ \forall i\le I, k\in\mathcal{K}} \\
& a_k(t)\le\lambda_k(t),\ \nu_k(t)\ge 0,\ \forall k\in\mathcal{K}, t\in[T] \\
& \sum_{k\in\mathcal{K}}\nu_k(t)\le 1,\ \forall t\in[T].
\end{aligned} \tag{w-fluid}$$

We note that the idea of enforcing capacity constraints in consecutive windows is also used in defining capacity regions for non-stationary queueing systems; see, e.g., [YSY23, NM23].

As discussed above, if arrival rates and review capacity are stationary, then $\mathcal{L}^\star(1,T)=\cdots=\mathcal{L}^\star(T,T)$ and thus the choice of w does not matter. For a general non-stationary system, we aim to design a feasible policy π that is robust to different choices of window size w. More formally, we define the average regret for a window size w as

$$\textsc{Reg}^\pi(w,T) \coloneqq \mathbb{E}\left[\mathcal{L}^\pi(T)\right]/T - \mathcal{L}^\star(w,T)/T.$$

We focus on the case where posts have long lifetimes ℓ_k. Our goal is to have average regret small with respect to the longest lifetime $\ell_{\max}=\max_{k\in\mathcal{K}}\ell_k$ even with initially unknown average cost $\{h_k\}$.

Notation. For ease of exposition when stating our results, we use $\lesssim$ to include super-logarithmic dependence only on the number of types K, feature dimension d, the maximum lifetime $\ell_{\max}=\max_{k\in\mathcal{K}}\ell_k$, the margin η, the window size w and the time horizon T with other parameters treated as constants. Note that we use $\perp$ to denote an empty element, so a set $\{\perp\}$ should be interpreted as an empty set $\emptyset$. We denote the density and distribution of a standard normal random variable by $\varphi(x)$ and $\Phi(x)$. We also follow the convention that $\frac{a}{0}=+\infty$ for any positive a. For a $d-$dimensional positive semi-definite (PSD) matrix $\boldsymbol{V}$, we define its corresponding vector norm by $\|\boldsymbol{\theta}\|_{\boldsymbol{V}}=\sqrt{\boldsymbol{\theta}^\intercal\boldsymbol{V}\boldsymbol{\theta}}$ for any $\boldsymbol{\theta}\in\mathbb{R}^d$. We denote $\boldsymbol{I}$ as the identity matrix with a suitable dimension, $\det(\boldsymbol{V})$ as the determinant of matrix $\boldsymbol{V}$ and $\lambda_{\min}(\boldsymbol{V})$ as the minimum eigenvalue of a PSD matrix $\boldsymbol{V}$.

3 Balancing Idiosyncrasy and Delay with Known Average Cost

Our starting point is the simpler setting where the average harm h_k is known for every type and the number of types K is small. In this setting, classification can be directly optimized by removing a post if and only if it has positive average harm, i.e., $Y_k(t)=\mathbb{1}(h_k\le 0)$. The two additional decisions (admission and scheduling) are not as straightforward and give rise to an interesting trade-off. To understand this trade-off, we decompose the loss of a policy $\mathcal{L}(\pi,T)$ into two components: idiosyncrasy loss and delay loss. In particular, for a period t, consider a new post j of type k.

- If the post is not admitted for review then it incurs an *idiosyncrasy loss* of $r_k\ell_k$ due to variance: it either stays on the platform for ℓ_k periods (if $h_k\le 0$) and incurs a loss $r_k^O\ell_k=r_k\ell_k$, or it is removed from the platform for ℓ_k periods (if $h_k>0$) and incurs a loss $r_k^R\ell_k=r_k\ell_k$.

- If the post is admitted and (successfully) reviewed in period $t+d-1$ (with a delay of d periods), then for periods $\{t,\ldots,\min(t+d-1,T)\}$, it incurs a *delay loss* $r_k \cdot \min(d,\ell_k) \leq r_k d$.

Formally, let $D(j)$ be a post's delay, i.e., a post arriving in period t gets its label in period $t+D(j)-1$. We set $D(j)=T+1-t$ if a post is never reviewed. In addition, let $\mathcal{Q}_k(t)$ be the set of type-k posts in the review queue $\mathcal{Q}(t)$ in period t and let $Q_k(t)=|\mathcal{Q}_k(t)|$. The above discussion thus shows that:

$$\mathbb{E}\left[\mathcal{L}^{\pi}(T)\right] = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t))\right]}_{\text{Idiosyncrasy Loss}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} A_{k(j(t))(t)}r_{k(j(t))}\cdot\min(D(j(t)),\ell_{k(j(t))})\right]}_{\text{Delay Loss}} \quad (3)$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t))\right] + \mathbb{E}\left[\sum_{t=1}^{T} A_{k(j(t))}r_{k(j(t))}D(j(t))\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t))\right] + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_kQ_k(t)\right]}_{\text{Relaxed Delay Loss}}, \quad (4)$$

where the last equality is by Little's Law [Lit61] (sum of post delays equal to sum of queue lengths).

3.1 Why AI-Only and Human-Only policies fail

We first consider two natural policies: AI-Only and Human-Only policies. In particular, for AI-Only, the platform purely relies on the classification and sends no post to human review. Although this policy has zero delay loss, we show that it can lead to $\Omega(\ell_{\max})$ average regret due to a high idiosyncrasy loss.

Proposition 1. *There is a one-type setting such that* $\textsc{Reg}^{\textsc{AI-Only}}(1,T) \geq r_1\ell_1/2$.

The second policy, Human-Only, admits every post to review by humans. Although this policy has zero idiosyncrasy loss, we show that its delay loss is high and the average regret is also $\Omega(\ell_{\max})$.

Proposition 2. *There is a one-type setting such that* $\textsc{Reg}^{\textsc{Human-Only}}(1,T) \geq r_1\ell_1/6$ *for* $T \geq 21\ell_1$.

We remark that any policy that has a static-threshold admission rule and does not admit every post (in which case it becomes Human-Only) is reduced to AI-Only in the worst case because we can always force every arriving post to have a type that is not admitted by the static threshold.

Proof of Propositions 1 and 2. We prove both propositions by the following example. There is only one type ($K=1$) with arrival rate $\lambda_1=1$ and lifetime $\ell_1\geq 25$. There is one human ($N(t)=1$) and the service rate is $\mu_1=0.5$. The fluid benchmark gives $\mathcal{L}^\star := \mathcal{L}^\star(1,T)=0.5r_1\ell_1T$.

The AI-only policy admits no posts ($A_1(t)=0$ for all t). Its idiosyncrasy loss is thus $\mathcal{L} := \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t))\right] = Tr_1\ell_1$ and $\textsc{Reg}^{\textsc{AI-only}}(1,T)=(\mathcal{L}^\star-\mathcal{L})/T \geq r_1\ell_1/2$.

The Human-Only policy admits a new post every period. For a post $j(t)$, there are $Q(t)=|\mathcal{Q}(t)|$ posts in front of it in the review queue and each takes at least one period to review. As a result, $D(j(t)) \geq \min(Q(t),T+1-t)$ where we take minimum because we define $D((j(t))=T+1-t$

if the post does not get review after period T. The delay loss of Human-Only is then at least $\mathbb{E}\left[\sum_{t=1}^{T-\ell_1} r_1 \min(Q(t), \ell_1)\right]$. We show (Lemma B.1 in Appendix B.1) that, for $T \geq 21\ell$, this is at least $\frac{2Tr_1\ell_1}{3}$ because the queue grows linearly, implying $\textsc{Reg}^{\textsc{Human-Only}}(1,T) \geq \frac{2r_1\ell_1}{3} - \mathcal{L}^\star \geq \frac{r_1\ell_1}{6}$. □

3.2 Balanced admission control for idiosyncrasy and delay (BACID)

Our analysis of AI-only and Human-Only reveals a trade-off between idiosyncrasy and delay loss: admitting more posts helps the first but harms the second. Our algorithm (Algorithm 1), adopts a simple admission rule to balance the two loss; we henceforth call it Balanced Admission for Classification, Idiosyncrasy and Delay, or BACID in short.

In period t, a new post j arrives with type k and the algorithm removes it if $h_k > 0$ and leaves it on the platform otherwise (Line 4). We admit this post to review if its idiosyncrasy loss $r_k\ell_k$, scaled by a hyper-parameter β (Line 1), is greater than the current number of type k posts in the review queue, $Q_k(t)$, i.e., $A_k(t) = \Lambda_k(t)\mathbb{1}\left(\beta r_k \ell_k \geq Q_k(t)\right)$ (Line 6). This rule requires an explicit calculation of r_k. We give exact formulae of r_k^O, r_k^R as a function when the average cost is h:

$$r_k^O(h) = v_k(h\Phi(h/\sigma_k) + \sigma_k\varphi(h/\sigma_k)) \qquad \text{and} \qquad r_k^R(h) = v_k(-h\Phi(-h/\sigma_k) + \sigma_k\varphi(-h/\sigma_k)). \quad (5)$$

These functional forms (proved in Appendix B.2) provide flexibility when h_k is unknown. We define $r_k(h) = \min(r_k^R(h), r_k^O(h))$ and use r_k^O, r_k^R, r_k as shorthands for $r_k^O(h_k), r_k^R(h_k), r_k(h_k)$, which are all known in this section because h_k is known.

For scheduling (Line 7), we follow the MaxWeight algorithm [Tas92] and select the earliest post in $\mathcal{Q}_{k'}(t)$ (first-come-first serve) from the type k' that maximizes the product of service rate and queue length, i.e., $\psi_{k'}(t) = 1$ for $k' \in \arg\max_{k' \in \mathcal{K}} \mu_{k'} Q_{k'}(t)$ (breaking ties arbitrarily). We then update the queues and the dataset by $\mathcal{Q}(t+1) = \mathcal{Q}(t) \cup \{\mathcal{A}(t)\} \setminus \{\mathcal{S}(t)\}$, $\mathcal{D}(t+1) = \mathcal{D}(t) \cup \{(\mathcal{S}(t), c_{\mathcal{S}(t)})\}$.

Algorithm 1: Balanced Admission for Classification, Idiosyncrasy & Delay

Data: h_k, ℓ_k, μ_k
1 $\beta \leftarrow 1/\sqrt{K\ell_{\max}}$ `// Admission Parameters`
2 **for** $t = 1$ **to** T **do**
3 Observe a new post $j(t) = j$ of type $k(j) = k$; if no new post, set $j(t) = \perp$
4 **if** $h_k > 0$ **then** $Y_k(t) \leftarrow 0$ **else** $Y_k(t) \leftarrow 1$ `// Classification`
5 $r_k \leftarrow \min(r_k^O(h_k), r_k^R(h_k))$
6 **if** $\beta \cdot r_k \cdot \ell_k \geq Q_k(t)$ **then** $A_k(t) = 1$ **else** $A_k(t) = 0$ `// Admission`
7 $k' \leftarrow \arg\max_{k' \in \mathcal{K}} \mu_{k'} \cdot Q_{k'}(t)$, $M(t) \leftarrow$ first post in $\mathcal{Q}_{k'}(t)$ if any `// Scheduling`
8 **if** the review finishes (with probability $N(t) \cdot \mu_k$) **then** $\mathcal{S}(t) = M(t)$ **else** $\mathcal{S}(t) = \perp$

Our main result is that BACID with $\beta = 1/\sqrt{K\ell_{\max}}$ achieves an average regret of $O(w\sqrt{K\ell_{\max}})$.

Theorem 1. *For a window size w, the average regret of* BACID *is upper bounded by*

$$\textsc{Reg}^{\text{BACID}}(w,T) \lesssim w\sqrt{K\ell_{\max}} + K.$$

Remark 1. *The above result is agnostic to the window size w; if w is known, $\beta = \sqrt{w/(K\ell_{\max})} + K$ leads to a bound of $O(\sqrt{wK\ell_{\max}})$. Note that our benchmark becomes stronger as w increases; the benchmark can schedule a post admitted in period t for review in period $t+w$ without delay loss.*

The next theorem (proof in Appendix B.3) shows that the dependence on $\sqrt{\ell_{\max}}$ is tight. To simplify the analysis, we consider a infinite-horizon stationary setting where $\frac{\mathcal{L}^{\pi}(T)}{T}$ converges in the stationary distribution and show the lower bound for any deterministic stationary policy.

Theorem 2. *There exists an infinite-horizon stationary setting where even with the knowledge of average cost h_k, any deterministic stationary policy must incur $\sqrt{\ell_{\max}}/6$ average regret.*

To prove Theorem 1, we rely on the loss decomposition in (4). We first upper bound the idiosyncrasy loss by showing that its difference to any w-window fluid benchmark is bounded by wT/β. As β increases (corresponding to more admissions), the idiosyncrasy loss thus decreases. The proof relies on a coupling with the benchmark using Lyapunov analysis and is given in Section 3.3.

Lemma 3.1. BACID*'s idiosyncrasy loss is* $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t))\right] \leq \mathcal{L}^{\star}(w,T)+\frac{wT}{\beta}$.

Moreover, the policy admits a new post when $\beta r_k\ell_k \geq Q_k(t)$, which upper bounds the queue length by $\beta r_k\ell_k+1$. This implies a delay loss of $r_{\max}K(\beta r_{\max}\ell_{\max}+1)T$ where $r_{\max}$ is an upper bound over r_k^O and r_k^R. Hence, a larger β leads to more delay loss, which matches our intuition on the trade-off between idiosyncrasy and delay loss. Setting $\beta = 1/\sqrt{K\ell_{\max}}$ balances this trade-off.

Lemma 3.2. BACID*'s relaxed delay loss is* $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_kQ_k(t)\right] \leq r_{\max}K(\beta r_{\max}\ell_{\max}+1)T$.

Proof. By induction on $t=1,2,\ldots$, we show that, for any type k, $Q_k(t)\leq \beta r_k\ell_k+1$. The basis of the induction ($Q_k(1)=0$) holds as the queue is initally empty. Our admission rule implies that $Q_k(t+1)\leq Q_k(t)+A_k(t)\leq Q_k(t)+\mathbb{1}\left(\beta r_k\ell_k\geq Q_k(t)\right)$. Combined with the induction hypothesis, $Q_k(t)\leq \beta r_k\ell_k+1$, this implies that $Q_k(t+1)\leq \beta r_k\ell_k+1$, proving the induction step. The lemma then follows as $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_kQ_k(t)\right] \leq T\sum_{k\in\mathcal{K}} r_k(\beta r_k\ell_k+1)\leq Tr_{\max}K(\beta r_{\max}\ell_{\max}+1)$. ☐

Proof of Theorem 1. For any window size w, applying Lemma 3.1 and Lemma 3.2 to (4) gives

$$\frac{\mathbb{E}\left[\mathcal{L}^{\mathrm{BACID}}(T)\right]}{T}\leq \frac{\mathcal{L}^{\star}(w,T)}{T}+\frac{w}{\beta}+r_{\max}K(\beta r_{\max}\ell_{\max}+1).\quad \text{Using that } \beta=1/\sqrt{K\ell_{\max}},$$
$$\textsc{Reg}^{\mathrm{BACID}}(w,T)\leq (w+r_{\max}^2)\sqrt{K\ell_{\max}}+Kr_{\max}\lesssim w\sqrt{K\ell_{\max}}+K.$$

☐

3.3 Coupling with $w-$window fluid benchmark (Lemma 3.1)

We fix a window size w and let $\boldsymbol{\tau}^{\star}=(\tau_1^{\star},\ldots,\tau_I^{\star}),\{a_k^{\star}(t),\nu_k^{\star}(t)\}_{k\in\mathcal{K},t\in[T]}$ be the optimal solution to the w-fluid benchmark (w-fluid). The problem (w-fluid) then becomes a linear program and multiplying the objective by β does not impact its optimal solution; the optimal value is simply multiplied by β. Taking the Lagrangian of the scaled program on capacity constraints and letting

$$f(\{\boldsymbol{a}(t)\}_t,\{\boldsymbol{\nu}(t)\}_t,\boldsymbol{u})=\beta\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\lambda_k(t)-a_k(t))-\sum_{i=1}^{I}\sum_{k\in\mathcal{K}} u_{i,k}\left(\mu_k\sum_{t=\tau_i^{\star}}^{\tau_{i+1}^{\star}-1}N(t)\nu_k(t)-\sum_{t=\tau_i^{\star}}^{\tau_{i+1}^{\star}-1}a_k(t)\right),$$

where $\boldsymbol{u} = (u_{i,k})_{i\in\mathcal{I},k\in\mathcal{K}} \geq 0$ are dual variables for the capacity constraints, the Lagrangian is

$$\begin{aligned} f(\boldsymbol{u}) \coloneqq & \min_{a_k(t),\nu_k(t)} f(\{\boldsymbol{a}(t)\}_t, \{\boldsymbol{\nu}(t)\}_t, \boldsymbol{u}) \\ & \text{s.t.} \quad a_k(t) \leq \lambda_k(t),\ \nu_k(t) \geq 0,\ \sum_{k'\in\mathcal{K}} \nu_{k'}(t) \leq 1,\ \forall k \in \mathcal{K}, t \in [T]. \end{aligned} \tag{6}$$

If the dual $\boldsymbol{u}$ is fixed, the optimal solution is given by $a_k(t) = \lambda_k(t)\mathbb{1}(\beta r_k \ell_k \geq y_{i,k})$ and $\nu_k(t) = \mathbb{1}(k \in \arg\max_{k'\in\mathcal{K}} \mu_k u_{i,k})$ for $t \in [\tau_i^\star, \tau_{i+1}^\star - 1]$. Comparing the induced optimal solution with BACID, BACID uses the queue length information $\boldsymbol{Q}(t) = (Q_k(t))_{k\in\mathcal{K}}$ as the dual to make decisions by setting $u_{i,k} = q_k(t)$. Under this setting of duals, the per-period Lagrangian is

$$f_t(\boldsymbol{a}, \boldsymbol{\nu}, \boldsymbol{q}) \coloneqq \beta \sum_{k\in\mathcal{K}} r_k \ell_k \lambda_k(t) - \left(\sum_{k\in\mathcal{K}} a_k(\beta r_k \ell_k - q_k) + \sum_{k\in\mathcal{K}} \nu_k q_k \mu_k N(t) \right). \tag{7}$$

Recalling that $\psi_k(t) = \mathbb{1}(k(M(t)) = k)$ indicates whether humans review a type-k post at time t, the expected Lagrangian of BACID is then $\sum_{t=1}^T \mathbb{E}[f_t(\boldsymbol{A}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t))]$. Our proof of Lemma 3.1 relies on a Lyapunov analysis of the function

$$L(t) = \beta \sum_{t'=1}^{t-1} \sum_{k\in\mathcal{K}} r_k \ell_k \left(\Lambda_k(t') - A_k(t')\right) + \frac{1}{2} \sum_{k\in\mathcal{K}} Q_k^2(t),$$

which connects the idiosyncrasy loss to the Lagrangian by the next lemma (proof in Appendix B.4).

Lemma 3.3. *The expected Lagrangian of* BACID *upper bounds its idiosyncrasy loss as following:*

$$\beta \mathbb{E}\left[\sum_{t=1}^T \sum_{k\in\mathcal{K}} r_k \ell_k (\Lambda_k(t) - A_k(t)) \right] \leq \mathbb{E}[L(T+1) - L(1)] \leq T + \sum_{t=1}^T \mathbb{E}[f_t(\boldsymbol{A}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t))].$$

Our second lemma shows that the expected Lagrangian of BACID is close to the optimal fluid.

Lemma 3.4. *For any window size w, the expected Lagrangian of* BACID *is upper bounded by:*

$$\sum_{t=1}^T \mathbb{E}[f_t(\boldsymbol{A}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t))] \leq \beta \mathcal{L}^\star(w, T) + (w-1)T.$$

Remark 2. *Lemma 3.4 holds even if the queue length sequence $\{\boldsymbol{Q}(t)\}$ is generated by another policy (instead of* BACID*); $\boldsymbol{A}(t)$ and $\boldsymbol{\psi}(t)$ are still the decisions made by* BACID *in period t given $\boldsymbol{Q}(t)$. This generalization is useful when we apply the lemma to a learning setting in Section 4.*

Proof of Lemma 3.1. Combining Lemmas 3.3 and 3.4 gives $\beta \mathbb{E}\left[\sum_{t=1}^T \sum_{k\in\mathcal{K}} r_k \ell_k (\Lambda_k(t) - A_k(t))\right] \leq wT + \beta \mathcal{L}^\star(w, T)$, which finishes the proof by dividing both sides by β. □

3.4 Connecting Lagrangian of BACID with Fluid Optimal (Lemma 3.4)

The scaled optimal primal objective $\beta \mathcal{L}^\star(w, T)$ is related to the Lagrangian $f(\{\boldsymbol{a}^\star(t)\}_t, \{\boldsymbol{\nu}^\star(t)\}_t, \boldsymbol{u})$ for some dual $\boldsymbol{u} = (u_{i,k})_{i\in[I],k\in\mathcal{K}}$. However, the left hand side of Lemma 3.4 has time-varying duals

$\boldsymbol{Q}(t)$. To connect it with the primal, we select a vector of dual variables $\boldsymbol{u}^\star$ consisting of the queue lengths in the first period of each window: $u^\star_{i,k} = Q_k(\tau^\star_i)$. Then we have

$$\sum_{t=1}^{T} \mathbb{E}\left[f_t(\boldsymbol{A}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t))\right] - \beta\mathcal{L}^\star(w,T) = \mathbb{E}\left[f(\{\boldsymbol{a}^\star(t)\}_t, \{\boldsymbol{\nu}^\star(t)\}_t, \boldsymbol{u}^\star)\right] - \beta\mathcal{L}^\star(w,T) \tag{8}$$

$$+ \sum_{i=1}^{I} \sum_{t=\tau^\star_i}^{\tau^\star_{i+1}-1} \left(\mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}(t))\right] - \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}(\tau^\star_i))\right]\right) \tag{9}$$

$$+ \sum_{t=1}^{T} \left(\mathbb{E}\left[f_t(\boldsymbol{A}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t))\right] - \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}(t))\right]\right). \tag{10}$$

Hence, BACID's suboptimality is captured by the sum of three terms: (8), the difference between the Lagrangian and the primal; (9), the suboptimality incurred by having different dual variables within a window; and (10), the difference in Lagrangian compared to the optimal fluid solution when the dual is given by per-period queue length. Our proof bounds these three terms independently.

The first step is to lower bound the first term by the optimal objective because of the definition of Lagrangian (proof in Appendix B.5), which shows that $\mathbb{E}\left[f(\{\boldsymbol{a}^\star(t)\}_t, \{\boldsymbol{\nu}^\star(t)\}_t, \boldsymbol{u}^\star)\right] \le \beta\mathcal{L}^\star(w,T)$.

Lemma 3.5. *For any dual variables $\boldsymbol{u} = (u_{i,k})_{i\in\mathcal{I},k\in\mathcal{K}} \ge 0$, $f(\{\boldsymbol{a}^\star(t)\}_t, \{\boldsymbol{\nu}^\star(t)\}_t, \boldsymbol{u}) \le \beta\mathcal{L}^\star(w,T)$.*

We next upper bound (9). The intuition is that the queue length changes at most linearly within a window and thus the difference in using per-period queue lengths or initial queue lengths of a window is not large. This is formalized in the next lemma (proof in Appendix B.6).

Lemma 3.6. *The difference between evaluating the expected Lagrangian by queue lengths and by window-based queue lengths scales at most linear with the window size, i.e.,*

$$\sum_{i=1}^{I} \sum_{t=\tau^\star_i}^{\tau^\star_{i+1}-1} \left(\mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}(t))\right] - \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}(\tau^\star_i))\right]\right) \le (w-1)T.$$

Our final lemma (proof in Appendix B.7) shows that (10) is nonpositive as BACID explicitly optimizes the per-period Lagrangian based on $\boldsymbol{Q}(t)$.

Lemma 3.7. *For every period t, we have $\mathbb{E}\left[f_t(\boldsymbol{A}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t)) \,|\, \boldsymbol{Q}(t)\right] \le \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}(t))\right]$.*

Proof of Lemma 3.4. The proof follows by applying Lemmas 3.5, 3.6 and 3.7 in (8), (9), (10). □

4 BACID with Learning: Optimism and Selective Sampling

In this section, we extend our approach to the setting where the expected costs h_k are initially unknown and the algorithm's classification, admission, and scheduling decisions should account for the need to learn these parameters online.

We first restrict our attention to BACID admission rule: defer a post of type k to human review if and only if $\beta r_k \ell_k \ge Q_k(t)$ where $r_k = \min(r^O_k(h_k), r^R_k(h_k))$ and, recalling Eq.(5):

$$r^O_k(h) = v_k(h\Phi(h/\sigma_k) + \sigma_k\varphi(h/\sigma_k)) \qquad \text{and} \qquad r^R_k(h) = v_k(-h\Phi(-h/\sigma_k) + \sigma_k\varphi(-h/\sigma_k)).$$

When the expected costs h_k are unknown, we cannot directly compute r_k and we need to instead use some estimate for r_k. A canonical way to resolve this problem in, e.g., bandits with knapsacks [AD19] is to use an optimistic estimate $\bar{r}_k(t)$. In particular, for each type k, we can compute the sample-average cost $\hat{h}_k(t) = \frac{\sum_{(j,c_j)\in\mathcal{D}(t):\, k(j)=k} c_j}{n_k(t)}$ where $n_k(t) = \sum_{(j,c_j)\in\mathcal{D}(t)} \mathbb{1}\left(k(j)=k\right)$ is the number of samples from type k; if $n_k(t) = 0$, we can set $\hat{h}_k(t) = 0$. We can then compute a confidence interval $[\underline{h}_k(t), \bar{h}_k(t)]$ which is, with high probability, *valid*, i.e., the true expected cost h_k lies within the confidence interval (see Lemma 4.5):

$$\underline{h}_k(t) = \max\left(-H, \hat{h}_k(t) - \sigma_k\sqrt{\frac{8\ln t}{n_k(t)}}\right), \bar{h}_k(t) = \min\left(H, \hat{h}_k(t) + \sigma_k\sqrt{\frac{8\ln t}{n_k(t)}}\right) \tag{11}$$

where we recall that $H \geq 1$ is a known upper bound on the absolute value of h_k. We can then admit a post if and only if $\beta\bar{r}_k(t)\ell_k \geq Q_k(t)$ where $\bar{r}_k(t) = \max_{h\in[\underline{h}_k(t),\bar{h}_k(t)]} r_k(h)$ is the most optimistic estimate for the per-period idiosyncrasy loss, i.e. it is the highest possible value that the per-period idiosyncrasy loss can have assuming that the sample-average cost is within the confidence interval.[4]

4.1 Why Optimism-Only is Insufficient for Classification

This optimistic admission rule gives a natural adaptation of BACID which we term BACID.UCB:

1. classify a type-k post as harmful if and only if $\hat{h}_k(t) \geq 0$ (similar to Line 4 in Algorithm 1);
2. admit a type-k post if $\beta\bar{r}_k(t)\ell_k \geq Q_k(t)$ (similar to Line 6 in Algorithm 1);
3. and schedule a type-k' post where k' maximizes $\mu_{k'}Q_{k'}(t)$ (same as Line 7 in Algorithm 1).

Such optimism-only heuristics generally work well in a constrained setting, such as bandits with knapsacks. The intuition is that assuming a valid confidence interval, we always admit a post that would have been admitted by BACID with known expected costs h_k. If we admit a post that would not have been admitted by BACID, we obtain one more sample; this shrinks the confidence interval which, in turn, limits the number of mistakes and leads to efficient learning.

Interestingly, this intuition does not carry over to our setting due to the additional error in the classification decisions of posts that are not admitted, for which the sign of $\hat{h}_k(t)$ may differ from that of h_k. To illustrate this point, consider the following instance with $K = 2$ types of posts: texts (type-1) and videos (type-2); see (Figure 2 for the exact instance parameters. Texts have a much larger lifetime than videos ($\ell_1 = 49\ell_2$). Proposition 3 (proved in Appendix C.2) shows that BACID.UCB with $\beta \in (48/\ell_1, 1)$ and initial knowledge of h_1 does not review video posts at all with high probability and thus there is no video post in the dataset for the entire horizon, although there are $\Omega(T)$ arrivals of them. As a result, when classifying type-2 posts, since there is no data, the algorithm estimates $\hat{h}_2(t) = 0$, and will remove all of videos, incurring $\Omega(\ell_2)$ average regret.[5]

Proposition 3. *With probability at least $1 - 2/T$, there is no type-2 post in the dataset $\mathcal{D}(T+1)$.*

BACID.UCB incurs linear regret in the above example as its decisions are inherently myopic to current rounds and disregard the importance of labels towards classification decisions in future rounds. Broadly speaking, optimism-only heuristics focus on the most optimistic estimate on the contribution of each action (in our setting, the admission decision) subject to a confidence interval

[4]The solution h giving the highest $r_k(h)$ has a simple form: if $0 \in [\underline{h}_k(t), \bar{h}_k(t)]$, the solution is zero; if $\bar{h}_k(t) < 0$, the solution is $\bar{h}_k(t)$; otherwise, $\underline{h}_k(t) > 0$, the solution is $\underline{h}_k(t)$. A formal proof is in Lemma C.1.

[5]We assume removal when $\hat{h}_k(t) = 0$. If the algorithm keep the post, the same issue exists by setting $h_2 = 1$.

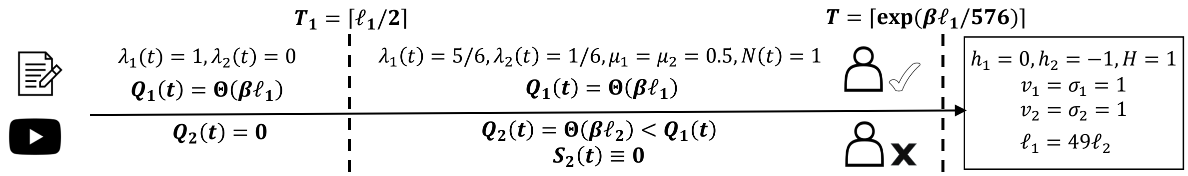

Figure 2: An example where BACID.UCB fails to correctly classify video (type-2) posts. Humans never review a video because the corresponding queue length is much smaller than that of texts.

and then select the action that maximizes this optimistic contribution in the current round. In our example, if we can only admit one type, the text posts have larger idiosyncrasy loss due to their lifetime and thus the benefit of admitting a text post outweighs even the most optimistic estimate on the idiosyncrasy loss of video posts in the current round. This mimics the admission rule of BACID which operates with known parameters and would never review a video post. Although always prioritizing text posts is myopically beneficial in the current round, this means that we collect no new video data, thus harming our classification performance in the long run.

4.2 Label-Driven Admission and Forced Scheduling for Classification

The inefficiency of BACID.UCB suggests the need to complement the myopic nature of optimism-only approaches by a forward-looking exploration that enhances classification decisions. Our algorithm, Optimistic and Label-driven Admission for Balanced Classification, Idiosyncrasy and Delay or OLBACID in short (Algorithm 2)) incorporates this forward-looking exploration and evades the shortcomings of optimism-only approaches. When a new post of type k arrives, we classify it as harmful/remove it ($Y_k(t) = 0$) if and only if the empirical average cost $\hat{h}_k(t)$ is positive. Unlike BACID which assumes knowledge of h_k, when $\underline{h}_k(t) < 0 < \bar{h}_k(t)$, we cannot confidently infer the sign of h_k from the sign of $\hat{h}_k(t)$.

To enhance future classification decisions on those posts, we complement the optimism-based admission (Line 7) with a *label-driven admission* (Line 6). Specifically, we maintain a new *label-driven queue* $\mathcal{Q}^{\text{LD}}(t)$ and add the post to that queue ($E_k(t) = 1$) if the queue is empty and there is high uncertainty on the sign of $\hat{h}_k(t)$, i.e., $\underline{h}_k(t) \leq -\gamma$ and $\gamma \leq \bar{h}_k(t)$ where γ is a parameter that avoids wasting reviewing capacity on posts that are already well classified. If the post is admitted into the review queue by the optimism-based admission ($E_k(t) = 0$, $\beta\bar{r}_k(t)\ell_k \geq Q_k(t)$), we denote $A_k(t) = 1$. We stress that $Q_k(t)$ includes only posts in the review queue $\mathcal{Q}(t)$ and not in $\mathcal{Q}^{\text{LD}}(t)$.

We use *forced scheduling* to prioritize reviews for the label-driven queue. If $\mathcal{Q}^{\text{LD}}(t)$ is not empty, we review a post from $\mathcal{Q}^{\text{LD}}(t)$. Otherwise, we follow the MaxWeight scheduling (as in BACID): select a type k' that maximizes $\mu_{k'}Q_{k'}(t)$ and review the earliest waiting post in the review queue $\mathcal{Q}(t)$ that has type k'. We let $\psi_{k'}(t) = 1$ if a type-k' post in $\mathcal{Q}(t)$ is scheduled to review in period t.

Algorithm 2: OPTIMISTIC AND LABEL-DRIVEN BACID (OLBACID)

Data: $(\ell_k, \mu_k, \sigma_k\}_{k\in\mathcal{K}}, H$

$\gamma = \beta \leftarrow 1/\sqrt{K\ell_{\max}}$ `// Label-driven & Optimistic Admission Parameters`

for $t = 1$ **to** T **do**

- Observe a new post $j(t) = j$ of type $k(j) = k$; if no new post, set $j(t) = \perp$
- **if** $\hat{h}_k(t) > 0$ **then** $Y_k(t) \leftarrow 0$ **else** $Y_k(t) \leftarrow 1$ `// Empirical Classification`
- `/* Label-Driven and Optimistic Admission */`
- $\bar{r}_k(t) = \max_{h\in[\underline{h}_k(t),\bar{h}_k(t)]} r_k(h)$ where $\underline{h}_k(t), \bar{h}_k(t)$ are given by (11)
- **if** $\underline{h}_k(t) < -\gamma < \gamma < \bar{h}_k(t)$ **and** $|\mathcal{Q}^{\text{LD}}(t)| = 0$ **then** $E_k(t) \leftarrow 1$ **else** $E_k(t) = 0$
- **if** $E_k(t) = 0$ **and** $\beta \cdot \bar{r}_k(t) \cdot \ell_k \geq Q_k(t)$ **then** $A_k(t) = 1$ **else** $A_k(t) = 0$
- `/* Forced Scheduling and MaxWeight Scheduling */`
- **if** $\mathcal{Q}^{\text{LD}}(t) \neq \emptyset$ **then** $M(t) \leftarrow$ the post in $\mathcal{Q}^{\text{LD}}(t)$
- **else** $k' \leftarrow \arg\max_{k'\in\mathcal{K}} \mu_{k'} \cdot Q_{k'}(t)$, $M(t) \leftarrow$ first post in $\mathcal{Q}_{k'}(t)$ if any
- **if** the review finishes (with probability $N(t) \cdot \mu_{k(M(t))}$) **then** $\mathcal{S}(t) = M(t)$ **else** $\mathcal{S}(t) = \perp$

Our main result (Theorem 3) is that, setting $\beta = \gamma = 1/\sqrt{K\ell_{\max}}$, OLBACID achieves an average regret of $\tilde{O}(\sqrt{\ell_{\max}} + \ell_{\max}/\sqrt{T} + \ell_{\max}^{1.5}/T)$ when there is a large margin η and small number of types K. In particular, the guarantee matches the lower bound $O(\sqrt{\ell_{\max}})$ when $T \geq \ell_{\max}$. Even if there is no margin (so classification is difficult), OLBACID still obtains an average regret of $\tilde{O}(\sqrt{\ell_{\max}} + \ell_{\max}/\sqrt{T} + \ell_{max}^2/T)$, which matches the lower bound when $T \geq \ell_{\max}^{1.5}$.

Theorem 3. *For a window size w, the average regret of* OLBACID *is upper bounded by*

$$\text{REG}^{\text{OLBACID}}(w, T) \lesssim w\sqrt{K\ell_{\max}} + K\ell_{\max}\sqrt{\frac{1}{T}} + \frac{\sqrt{K}\ell_{\max}^{1.5}}{T} + \min\left(\frac{K\ell_{\max}}{\eta^2 T}, \frac{K^2\ell_{\max}^2}{T}\right).$$

To prove Theorem 3, we need a loss decomposition of $\mathcal{L}^{\text{OLBACID}}(T)$ as in (3) that also captures the possible incorrect classification decisions. The incurred loss per period is thus given by $r_k^O \cdot (1 - Y_k(t)) + r_k^R \cdot Y_k(t)$ instead of $r_k = \min(r_k^O, r_k^R)$. Following the same analysis of (3), we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}^{\text{OLBACID}}(T)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}\left(r_k^O \cdot Y_k(t) + r_k^R \cdot (1 - Y_k(t))\right)\ell_k(\Lambda_k(t) - A_k(t) - E_k(t))\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T}(A_{k(j(t))}(t) + E_{k(j(t))}(t))\left(r_k^O \cdot Y_k(t) + r_k^R \cdot (1 - Y_k(t))\right)\min\left(D(j(t)), \ell_{k(j)}\right)\right].\end{aligned}$$

Using $r_k^O, r_k^R \leq r_{\max}$ and applying Little's Law, we upper bound $\mathbb{E}\left[\mathcal{L}^{\text{OLBACID}}(T)\right]$ by

$$\begin{aligned}&\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(Y_k(t)(r_k^O - r_k^R)^+ + (1 - Y_k(t))(r_k^R - r_k^O)^+)\ell_k(\Lambda_k(t) - A_k(t) - E_k(t))\right]}_{\text{Classification Loss}} \\ &+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t) - A_k(t) - E_k(t))\right]}_{\text{Idiosyncrasy Loss}} + \underbrace{r_{\max}\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{k\in\mathcal{K}} Q_k(t) + |\mathcal{Q}^{\text{LD}}(t)|\right)\right]}_{\text{Relaxed Delay Loss}}.\end{aligned} \tag{12}$$

Comparing (12) with (4), there is a new term on classification loss, which captures the loss when the classification $Y_k(t)$ is incorrect. We adjust the other terms to capture label-driven admission.

To bound those three losses, we define the maximum per-period views and cost standard deviation as $v_{\max} = \max(1, \max_k v_k)$, $\sigma_{\max} = \max(1, \max_{k\in\mathcal{K}} \sigma_k)$ respectively[6], and the minimum per-period review rate as $\hat{\mu}_{\min} = \min_{t\in[T],k\in\mathcal{K}} N(t)\mu_k$. To account for uncertainty, we take $r_{\max} = v_{\max}(H + \sigma_{\max}) \geq 1$; this is an upper bound on $r_k^O(\cdot), r_k^R(\cdot)$ by their definition Eq. (5). We focus on $T \geq 3$ and $K \leq \ell_{\max}$ (the bound in Theorem 3 becomes trivial if $K > \ell_{\max}$). Our bounds (Lemmas 4.1, 4.2 and 4.3) hold for general β, γ, and are proven in Sections 4.3, 4.4 and 4.5 respectively. The lemmas are defined for $T \geq 3$ and $K \leq \ell_{\max}$ (the bound in Theorem 3 becomes trivial otherwise). The proof of Theorem 3 (provided in Appendix C.4) directly combines the lemmas.

Lemma 4.1. *For any $\beta, \gamma \in (1/\ell_{\max}, 1]$, the Relaxed Delay Loss of* OLBACID *is at most*

$$2K\beta r_{\max}^2 \ell_{\max} T + \frac{33 r_{\max} K \sigma_{\max}^2 \ln T}{\hat{\mu}_{\min} \max(\eta, \gamma)^2}.$$

Lemma 4.2. *For any $\beta, \gamma \in (1/\ell_{\max}, 1]$, the Classification Loss of* OLBACID *is at most*

$$4\ell_{\max}\left(v_{\max}\gamma T + K r_{\max}\right) + \frac{33 r_{\max} \ell_{\max} K \sigma_{\max}^2 \ln T}{\hat{\mu}_{\min} \max(\eta, \gamma)^2}.$$

Lemma 4.3. *For any $\beta, \gamma \in (1/\ell_{\max}, 1]$, the Idiosyncrasy Loss of* OLBACID *is at most*

$$\mathcal{L}^\star(w, T) + \frac{wT}{\beta} + \frac{66 r_{\max} \ell_{\max} K \sigma_{\max}^2 \ln T}{\max(\eta, \gamma)^2 \hat{\mu}_{\min}} + 18 K r_{\max} \ell_{\max} \left(\sqrt{T \ln T} + \beta r_{\max} \ell_{\max}\right)$$

4.3 Bounding Forced Scheduling and Relaxed Delay Loss (Lemma 4.1)

A key ingredient in the proof of all of Lemmas 4.1, 4.2, and 4.3 is to provide an upper bound on the number of periods that the label-driven queue is non-empty. Given that $\mathcal{Q}^{\mathrm{LD}}(t)$ has at most one post at any period t, letting $Q^{\mathrm{LD}}(t) = |\mathcal{Q}^{\mathrm{LD}}(t)|$, this is equal to $\mathbb{E}\left[\sum_{t=1}^T Q^{\mathrm{LD}}(t)\right]$. This quantity allows to bound 1) the number of posts which we do not admit into the label-driven queue despite not being able to confidently estimate the sign of their expected cost; 2) the delay loss; 3) the number of periods that we do not follow MaxWeight scheduling.

Our first lemma connects $\mathbb{E}\left[\sum_{t=1}^T Q^{\mathrm{LD}}(t)\right]$ to the number of posts admitted to the label-driven queue $\mathcal{Q}^{\mathrm{LD}}(t)$. The proof (Appendix C.3) relies on only admitting posts when $\mathcal{Q}^{\mathrm{LD}}(t)$ is empty.

Lemma 4.4. *The label-driven queue admits at most $\mathbb{E}\left[\sum_{t=1}^T Q^{\mathrm{LD}}(t)\right] \leq \frac{\mathbb{E}\left[\sum_{t=1}^T \sum_{k\in\mathcal{K}} E_k(t)\right]}{\hat{\mu}_{\min}}$ posts.*

Our second lemma applies concentration bounds (Appendix C.5) to show that the event $\mathcal{E}_{k,t} = \{h_k \in [\underline{h}_k(t), \bar{h}_k(t)]\}$ (expected cost lies in confidence interval) holds with high probability.

Lemma 4.5. *For any k, t, the expected cost lies in $[\underline{h}_k(t), \bar{h}_k(t)]$ with probability $\mathbb{P}\{\mathcal{E}_{k,t}\} \geq 1 - 2t^{-3}$.*

Our next lemma (proof in Appendix C.6) bounds $\mathbb{E}\left[\sum_{t=1}^T \sum_{k\in\mathcal{K}} E_k(t)\right]$ via considering the type-k confidence interval when the last type-k post is admitted to the label-driven queue $\mathcal{Q}^{\mathrm{LD}}$.

Lemma 4.6. *The label-driven queue admits at most $\mathbb{E}\left[\sum_{t=1}^T \sum_{k\in\mathcal{K}} E_k(t)\right] \leq \frac{33K\sigma_{\max}^2 \ln T}{\max(\eta,\gamma)^2}$ posts.*

[6] We take maximum with 1 to simplify the exposition but this does not impose an assumption on the instance.

Our final lemma bounds the length of the review queue (similar to Lemma 3.2).

Lemma 4.7. *For any k, t, the optimistic queue has at most $Q_k(t) \le 2\beta r_{\max}\ell_{\max}$ posts.*

Proof. We admit a type-k post into $\mathcal{Q}$ only if $\beta \bar{r}_k(t)\ell_k \le Q_k(t)$ and $\bar{r}_k(t) \le r_{\max}$. As in the proof of Lemma 3.2, the queue length is bounded by $Q_k(t) \le \beta r_{\max}\ell_{\max} + 1 \le 2\beta r_{\max}\ell_{\max}$. □

Proof of Lemma 4.1. By definition, the Relaxed Delay Loss is upper bounded by

$$r_{\max}\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{k\in\mathcal{K}} Q_k(t) + |\mathcal{Q}^{\mathrm{LD}}(t)|\right)\right] \le 2r_{\max}K\beta r_{\max}\ell_{\max}T + \frac{33 r_{\max}K\sigma_{\max}^2 \ln T}{\max(\eta,\gamma)^2\hat{\mu}_{\min}},$$

where we use Lemma 4.7 for the first term and Lemmas 4.4 and 4.6 for the second term. □

4.4 Bounding the Classification Loss (Lemma 4.2)

Our first lemma (proof in Appendix C.7) shows a Lipschitz property for functions $r_k^O(\cdot), r_k^R(\cdot), r_k(\cdot)$.

Lemma 4.8. *For any type k, functions $r_k^O(h), r_k^R(h)$ and $r_k(h)$ are $1.5v_k$-Lipschitz continuous.*

Our next lemma (proof in Appendix C.8) uses this property to establish a bound on the type-k period-t classification loss $Z_k(t) = (Y_k(t)(r_k^O - r_k^R)^+ + (1-Y_k(t))(r_k^R - r_k^O)^+)\ell_k(\Lambda_k(t) - A_k(t) - E_k(t))$.

Lemma 4.9. *For any type k and period t, $Z_k(t)\mathbb{1}(\mathcal{E}_{k,t}) \le \left(3v_{\max}\gamma\ell_{\max} + r_{\max}\ell_{\max}Q^{\mathrm{LD}}(t)\right)\Lambda_k(t)$.*

Proof of Lemma 4.2. By definition, the total classification loss is $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} Z_k(t)\right]$. As a result,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} Z_k(t)\right] &\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} Z_k(t)\mathbb{1}(\mathcal{E}_{k,t})\right] + r_{\max}\ell_{\max}\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}\mathbb{P}\{\mathcal{E}_{k,t}^c\} && (Z_k(t) \le r_{\max}\ell_{\max}) \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} Z_k(t)\mathbb{1}(\mathcal{E}_{k,t})\right] + Kr_{\max}\ell_{\max}\sum_{t=1}^{T}\frac{2}{t^3} && \text{(By Lemma 4.5)} \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\left(3v_{\max}\gamma\ell_{\max} + r_{\max}\ell_{\max}Q^{\mathrm{LD}}(t)\right)\right] + 4Kr_{\max}\ell_{\max} && \text{(By Lemma 4.9)} \\
&\le 3v_{\max}\gamma\ell_{\max}T + \frac{33 r_{\max}\ell_{\max}K\sigma_{\max}^2\ln T}{\max(\eta,\gamma)^2\hat{\mu}_{\min}} + 4Kr_{\max}\ell_{\max}. && \text{(By Lemmas 4.4, 4.6)}
\end{aligned}$$

□

4.5 Bounding the Idiosyncrasy Loss (Lemma 4.3)

We follow a similar strategy as in the proof of Lemma 3.1, but we encounter three new challenges due to the algorithmic differences between BACID and OLBACID.

Our first challenge arises because of the additional label-driven admission. Without the label-driven admission, we would admit a type-k post to the review queue $\mathcal{Q}(t)$ in period t if its (optimistic) idiosyncrasy loss outweighs the delay loss, i.e., $\bar{A}_k(t) = \Lambda_k(t)\mathbb{1}(\beta\bar{r}_k(t)\ell_k \ge Q_k(t))$ is equal to one. However, we now only admit such a post to $\mathcal{Q}(t)$ if it is not admitted to the label-driven queue $\mathcal{Q}^{\mathrm{LD}}(t)$, i.e., the real admission decision is $A_k(t) = \bar{A}_k(t)(1 - E_k(t))$. The following Lyapunov function defined based on only the length of $\mathcal{Q}(t)$ accounts for this difference and offers an analogue of Lemma 3.3, which we prove in Appendix C.9.

$$L(t) = \beta \sum_{t'=1}^{t-1} \sum_{k \in \mathcal{K}} r_k \ell_k (\Lambda_k(t') - A_k(t') - E_k(t')) + \frac{1}{2} \sum_{k \in \mathcal{K}} Q_k^2(t),$$

Lemma 4.10. *The expected Lagrangian of* OLBACID *upper bounds the idiosyncrasy loss by:*

$$\beta \mathbb{E}\left[\sum_{t=1}^{T} \sum_{k \in \mathcal{K}} r_k \ell_k (\Lambda_k(t) - A_k(t) - E_k(t))\right] \leq \mathbb{E}\left[L(T+1) - L(1)\right] \leq T + \sum_{t=1}^{T} \mathbb{E}\left[f_t(\bar{\boldsymbol{A}}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t))\right].$$

We next bound the right hand side of Lemma 4.10. By Lemma 3.4, we know how to bound this quantity when admission and scheduling decisions are made according to BACID, i.e., $A_k^{\mathrm{BACID}}(t) = \mathbb{1}\left(\beta r_k \ell_k \geq Q_k(t)\right) \Lambda_k(t)$ and $\psi_k^{\mathrm{BACID}}(t) = \mathbb{1}\left(k = \arg\max_{k' \in \mathcal{K}} \mu_k Q_k(t)\right)$. To bound the Lagrangian under OLBACID, we connect it to its analogue under BACID via the *regret in Lagrangian*:

$$\begin{aligned}
\textsc{RegL}(T) &= \mathbb{E}\left[\sum_{t=1}^{T} f_t(\bar{\boldsymbol{A}}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t)) - f_t(\boldsymbol{A}^{\mathrm{BACID}}(t), \boldsymbol{\psi}^{\mathrm{BACID}}(t), \boldsymbol{Q}(t))\right] \\
&= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \sum_{k \in \mathcal{K}} (A_k^{\mathrm{BACID}}(t) - \bar{A}_k(t))(\beta r_k \ell_k - Q_k(t))\right]}_{\textsc{RegA}(T)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \sum_{k \in \mathcal{K}} (\psi_k^{\mathrm{BACID}}(t) - \psi_k(t)) Q_k(t) \mu_k N(t)\right]}_{\textsc{RegS}(T)}
\end{aligned}$$

Our second challenge is to upper bound the regret in admission $\textsc{RegA}(T)$. Unlike BACID which admits based on the ground-truth per-period idiosyncrasy loss r_k, OLBACID uses the optimistic estimation $\bar{r}_k(t)$. Unlike works in bandits with knapsacks, the following lemma also needs to account for the endogenous queueing delay in label acquisition (proof in Appendix C.10).

Lemma 4.11. *The regret in admission is* $\textsc{RegA}(T) \leq 12K\beta^2 r_{\max}^2 \ell_{\max}^2 + 6K\beta r_{\max} \ell_{\max} \sqrt{8T \ln T}$.

Our third challenge is to bound the regret in scheduling $\textsc{RegS}(T)$. Unlike BACID which schedules based on $\textsc{MaxWeight}$, OLBACID prioritizes posts in the $\mathcal{Q}^{\mathrm{LD}}(t)$. The effect of those deviations is bounded in the next lemma (proof in Appendix C.11), using Lemmas 4.4 and 4.6.

Lemma 4.12. *The regret in scheduling is* $\textsc{RegS}(T) \leq \frac{66\beta r_{\max} \ell_{\max} K \sigma_{\max}^2 \ln T}{\max(\eta, \gamma)^2 \hat{\mu}_{\min}}$.

Proof of Lemma 4.3. By Lemma 4.10, the idiosyncrasy loss is upper bounded by

$$
\begin{aligned}
\beta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t)-E_k(t))\right] &\le T+\sum_{t=1}^{T}\mathbb{E}\left[f_t(\bar{\boldsymbol{A}}(t),\boldsymbol{\psi}(t),\boldsymbol{Q}(t))\right] \\
&= T+\sum_{t=1}^{T}\mathbb{E}\left[f_t(\boldsymbol{A}^{\text{BACID}}(t),\boldsymbol{\psi}^{\text{BACID}}(t),\boldsymbol{Q}(t))\right]+\textsc{RegL}(T) \\
&\le wT+\beta\mathcal{L}^\star(w,T)+\textsc{RegA}(T)+\textsc{RegS}(T) \qquad \text{(By Lemma 3.4)} \\
\le wT+\beta\mathcal{L}^\star(w,T)+12K\beta^2 r_{\max}^2\ell_{\max}^2+6K\beta r_{\max}\ell_{\max}\sqrt{8T\ln T}&+\frac{66\beta r_{\max}\ell_{\max}K\sigma_{\max}^2\ln T}{\max(\eta,\gamma)^2\hat{\mu}_{\min}} \\
&\qquad \text{(By Lemmas 4.11 and 4.12)} \\
\le wT+\beta\mathcal{L}^\star(w,T)+\frac{66\beta r_{\max}\ell_{\max}K\sigma_{\max}^2\ln T}{\max(\eta,\gamma)^2\hat{\mu}_{\min}}&+18K\beta r_{\max}\ell_{\max}\left(\sqrt{T\ln T}+\beta r_{\max}\ell_{\max}\right).
\end{aligned}
$$

Dividing both sides of the inequality by β gives the desired result. $\square$

5 OLBACID with Type Aggregation and Contextual Learning

In this section, we design an algorithm whose performance does not deteriorate with the number of types K. We adopt a linear contextual structure assumption that is common in the bandit literature [LCLS10, CLRS11, APS11], and in online content moderation practice [ABB+22]. Each post comes with a d-dimensional feature vector $\boldsymbol{\phi}_k \in \mathbb{R}^d$ associated with its type k and its expected cost satisfies a linear model $h_k = \boldsymbol{\phi}_k^\intercal\boldsymbol{\theta}^\star$ for a fixed unknown vector $\boldsymbol{\theta}^\star$.

We incorporate the contextual information in our confidence intervals around h_k by classical bandit techniques [APS11]; we provide a more detailed comparison in Appendix D.1. We construct a confidence set $\mathcal{C}_t$ for $\boldsymbol{\theta}^\star$ based on the collected data set $\mathcal{D}(t)$. Recalling that $\mathcal{S}(\tau)$ is the reviewed post for period τ (or $\perp$ when no post is reviewed), the dataset in period t is $\{(\boldsymbol{X}_\tau, Z_\tau)\}_{\tau\le t}$ with $\boldsymbol{X}_\tau = \boldsymbol{\phi}_{k(\mathcal{S}(\tau))}$ and $Z_\tau = c_{\mathcal{S}(\tau)}$. We use the ridge estimator with regularization parameter κ, i.e,

$$\hat{\boldsymbol{\theta}}(t) = \bar{\boldsymbol{V}}_t^{-1}[\boldsymbol{X}_1,\dots,\boldsymbol{X}_t][Z_1,\dots,Z_t]^\intercal \quad \text{where} \quad \bar{\boldsymbol{V}}_t = \kappa\boldsymbol{I}+\sum_{\tau\le t}\boldsymbol{X}_\tau\boldsymbol{X}_\tau^\intercal \quad \text{and} \quad \hat{\boldsymbol{\theta}}(0)=\boldsymbol{0}. \tag{13}$$

Recalling that $\sigma_{\max} = \max_{k\in\mathcal{K}}\sigma_k$ is the maximum idiosyncrasy variance and U upper bounds the Euclidean norm of $\boldsymbol{\theta}^\star$ and any feature $\boldsymbol{\phi}_k$, we define the confidence set with a confidence level δ by

$$\mathcal{C}_t := \left\{\boldsymbol{\theta}\in\mathbb{R}^d\colon \|\hat{\boldsymbol{\theta}}(t)-\boldsymbol{\theta}\|_{\bar{\boldsymbol{V}}_t} \le B_\delta(t)\right\} \quad \text{where } B_\delta(t) := \sigma_{\max}\sqrt{d\ln\left(\frac{1+tU^2/\kappa}{\delta}\right)}+\sqrt{\kappa}U. \tag{14}$$

Using confidence set $\mathcal{C}_{t-1}$, we can modify our confidence intervals from (11) for period t as follows:

$$\underline{h}_k(t) = \max\left(-H, \min_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}}\boldsymbol{\phi}_k^\intercal\boldsymbol{\theta}\right), \quad \bar{h}_k(t) = \min\left(H, \max_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}}\boldsymbol{\phi}_k^\intercal\boldsymbol{\theta}\right), \quad \bar{r}_k(t) = \max_{h\in[\underline{h}_k(t),\bar{h}_k(t)]} r_k(h). \tag{15}$$

5.1 Type Aggregation for Better Idiosyncrasy and Delay Tradeoff

An additional challenge when the number of types K is large (that arises even without learning) is the difficulty to estimate delay loss; this is crucial in the design of BACID which admits a post if and only if its idiosyncrasy loss is above an estimated delay loss. BACID estimates the delay loss of a post based on the number of same-type waiting posts. With a large K, this approach underestimates the real delay loss, leading to overly admitting posts into the review system. An alternative delay estimator uses the total number of waiting posts $|\mathcal{Q}(t)|$. This ignores the heterogeneous delay loss of admitting different types. In particular, our scheduling algorithm prioritizes posts with less review workload (higher μ_k) to effectively manage the limited capacity. A post with a higher service rate thus has smaller delay loss and neglecting this heterogeneity results in overestimating its delay loss.

To address this challenge, we create an estimator for the delay loss that lies in the middle ground of the aforementioned estimators. In particular, we map each type k to a group $g(k)$ based on its service rates; we denote this partition by $\mathcal{K}_{\mathcal{G}} = \{\mathcal{K}_g\}_{g\in\mathcal{G}}$ where $\mathcal{G}$ is the set of groups, $G = |\mathcal{G}|$ is its cardinality, and $\mathcal{K}_g$ is the set of types in group $g \in \mathcal{G}$. For a group g, we define its proxy service rate $\tilde{\mu}_g = \min_{k\in\mathcal{K}_g} \mu_k$ as the minimum service rate across types in this group. For a new post of type k, we estimate its delay loss by the number of posts of types in $\mathcal{K}_{g(k)}$ waiting in the review queue. This estimator is efficient if the number of groups is small, and service rates of types in a group are close to each other. Specifically, letting $N_{\max} = \max_t N(t)$ be the maximum number of reviewers, we define the *aggregation gap* $\Delta(\mathcal{K}_{\mathcal{G}}) = \max_{g\in\mathcal{G}} \max_{k\in\mathcal{K}_g} N_{\max}(\mu_g - \tilde{\mu}_g)$ of a group partition $\mathcal{K}_{\mathcal{G}}$ as the maximum within-group service rate difference (scaled by reviewing capacity).

5.2 Algorithm, Theorem and Proof Sketch

Our algorithm, Contextual OLBACID (Algorithm 3), or COLBACID in short, works as follows. In period t, we first compute a ridge estimator $\hat{\boldsymbol{\theta}}(t)$ of $\boldsymbol{\theta}^\star$ by (13). The algorithm then classifies a new type-k post as harmful ($Y_k(t) = 0$) if its empirical average cost, $\boldsymbol{\phi}_k^\intercal\hat{\boldsymbol{\theta}}(t)$, is positive. We follow the same label-driven admission with OLBACID in Line 7 but we set the confidence interval on h_k by (15) using the confidence set $\mathcal{C}_t$ from (14). The optimistic admission rule (Line 8) similarly finds an optimistic per-period idiosyncrasy loss $\bar{r}_k(t)$ but estimates the delay loss by type-aggregated queue lengths. Specifically, letting $\tilde{\mathcal{Q}}_g(t) = \{j \in \mathcal{Q}(t) \colon g(k(j)) = g\}$ be the set of waiting posts whose types belong to group g and $\tilde{Q}_g(t) = |\tilde{\mathcal{Q}}_g(t)|$, we admit a type-$k$ post if and only if its (optimistic) idiosyncrasy loss is higher than the estimated delay loss, i.e., $\beta\bar{r}_{k(t)}(t)\ell_k \geq \tilde{Q}_{g(k(t))}(t)$. We still prioritize the label-driven queue for scheduling. If there is no post in the label-driven queue, we use a type-aggregated MaxWeight scheduling: we first pick a group g maximizing $\tilde{\mu}_g\tilde{Q}_g(t)$, and then pick the earliest admitted post in the review queue $\mathcal{Q}(t)$ whose type is of group g to review.

Algorithm 3: CONTEXTUAL OLBACID (COLBACID)

Data: $(\ell_k, \mu_k, \sigma_k\}_{k\in\mathcal{K}}$, upper bounds U, H, and group partition $\mathcal{K}_{\mathcal{G}}$

$\beta, \gamma \leftarrow 1/\sqrt{G\ell_{\max}},\ \delta \leftarrow \min(\gamma, 1/T),\ \kappa \leftarrow \max(1, U^2)$

for $t = 1$ **to** T **do**

  Observe a new post $j(t) = j$ of type $k(j) = k$; if no new post, set $j(t) = \perp$

  Compute ridge estimator $\hat{\boldsymbol{\theta}}(t)$ by (13)

  if $\boldsymbol{\phi}_k^\intercal \hat{\boldsymbol{\theta}}(t) > 0$ **then** $Y_k(t) \leftarrow 0$ **else** $Y_k(t) \leftarrow 1$ `// Empirical Classification`

  `/* Label-Driven and Optimistic Admission */`

  $\bar{r}_k(t) = \max_{h\in[\underline{h}_k(t), \bar{h}_k(t)]} r_k(h)$ where $\underline{h}_k(t), \bar{h}_k(t)$ are given by (15)

  if $\underline{h}_k(t) < -\gamma < \gamma < \bar{h}_k(t)$ **and** $|\mathcal{Q}^{\mathrm{LD}}(t)| = 0$ **then** $E_k(t) \leftarrow 1$ **else** $E_k(t) = 0$

  if $E_k(t) = 0$ **and** $\beta \cdot \bar{r}_k(t) \cdot \ell_k \geq \tilde{Q}_{g(k)}(t)$ **then** $A_k(t) = 1$ **else** $A_k(t) = 0$ `// Admission`

  `/* Forced Scheduling and Type-Aggregated MaxWeight Scheduling */`

  if $\mathcal{Q}^{\mathrm{LD}}(t) \neq \emptyset$ **then** $M(t) \leftarrow$ the post in $\mathcal{Q}^{\mathrm{LD}}(t)$

  else $g \leftarrow \arg\max_{g\in\mathcal{G}} \tilde{\mu}_g \cdot \tilde{Q}_g(t)$, $M(t) \leftarrow$ first post in $\tilde{\mathcal{Q}}_g(t)$ if any

  if the review finishes (with probability $N(t) \cdot \mu_{k(M(t))}$) **then** $\mathcal{S}(t) = M(t)$ **else** $\mathcal{S}(t) = \perp$

For ease of exposition, we follow the $\lesssim$ notation to include only super-logarithmic dependence on the number of groups G, feature dimension d, maximum lifetime $\ell_{\max}$, the margin η, window size w, the time horizon T, and the aggregation gap $\Delta(\mathcal{K}_{\mathcal{G}})$. Our main result is as follows.

Theorem 4. *For a window size w, the average regret of* COLBACID *is upper bounded by*

$$\mathrm{REG}^{\mathrm{COLBACID}}(w, T) \lesssim w\sqrt{G\ell_{\max}} + \ell_{\max}\left(\Delta(\mathcal{K}_{\mathcal{G}}) + d\sqrt{\frac{G}{T}} + \frac{\sqrt{G}\ell_{\max}^{0.5}d^{1.5}}{T} + \frac{d^{2.5}}{T}\min\left(\frac{1}{\eta^2}, G\ell_{\max}\right)\right).$$

Note that, if, for all groups $g \in \mathcal{G}$, all types $k \in \mathcal{K}_g$ have the same service rate, then $\Delta(\mathcal{K}_{\mathcal{G}}) = 0$, but the regret still depends on G which is large if there are many types with unequal service rates.

To provide a worst-case guarantee, we select a fixed aggregation gap $0 < \zeta \leq 1$ and create a partition $\mathcal{K}_{\mathcal{G}}^{\zeta}$ that segments types based on the their maximum scaled service rate $N_{\max}\mu_k$ into intervals $(0, \zeta], (\zeta, 2\zeta], \ldots, (\lfloor\frac{1}{\zeta}\rfloor\zeta, 1]$. The number of groups is at most $G \leq \frac{1}{\zeta} + 1 \leq \frac{2}{\zeta}$ and the aggregation gap $\Delta(\mathcal{K}_{\mathcal{G}}^{\zeta})$ is at most ζ. Optimizing the bound in Theorem 4 for the term $w\sqrt{G\ell_{\max}} + \Delta(\mathcal{K}_{\mathcal{G}})\ell_{\max}$, and setting $\zeta = \ell_{\max}^{-1/3}$ we obtained the following result.

Corollary 1. *For a window size w,* COLBACID *with group partition $\mathcal{K}_{\mathcal{G}}^{\zeta}$ satisfies:*

$$\mathrm{REG}^{\mathrm{COLBACID}}(w, T) \lesssim w\ell_{\max}^{2/3} + \ell_{\max}\left(\frac{\ell_{\max}^{1/6}d}{\sqrt{T}} + \frac{\ell_{\max}^{2/3}d^{1.5}}{T} + \frac{d^{2.5}}{T}\min\left(\frac{1}{\eta^2}, \ell_{\max}^{4/3}\right)\right).$$

The proof of Theorem 4 (Appendix D.2) bounds the losses in the same decomposition as in (12),[7]

$$\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(Y_k(t)(r_k^O - r_k^R)^+ + (1-Y_k(t))(r_k^R - r_k^O)^+)\ell_k(\Lambda_k(t) - A_k(t) - E_k(t))\right]}_{\text{Classification Loss}}$$

$$+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t) - A_k(t) - E_k(t))\right]}_{\text{Idiosyncrasy Loss}} + \underbrace{r_{\max}\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{g\in\mathcal{G}}\tilde{Q}_g(t) + Q^{\mathrm{LD}}(t)\right)\right]}_{\text{Relaxed Delay Loss}}. \tag{16}$$

Similar to the proof of Theorem 3, we set $v_{\max} = \max_k v_k, \sigma_{\max} = \max(1, \max_k \sigma_k)$, $\hat{\mu}_{\min} = \min_{t\le T, k\in\mathcal{K}} N(t)\mu_k$, $r_{\max} = v_{\max}(H + \sigma_{\max}) \ge 1$ and focus on $T \ge 3$ and $U, H \ge 1$. The following lemmas also assume $G \le \ell_{\max}$; Theorem 4 holds directly otherwise.

Lemma 5.1. *For any $\beta, \gamma \in (1/\ell_{\max}, 1), \kappa \ge U^2$ and $\delta \in (0, 1/T)$,*

$$\textit{Relaxed Delay Loss} \le 2G\beta r_{\max}^2 \ell_{\max} T + \frac{9 r_{\max} B_\delta^2(T) d\ln(1 + T/d)}{\max(\eta,\gamma)^2 \hat{\mu}_{\min}}.$$

Lemma 5.2. *For any $\beta, \gamma \in (1/\ell_{\max}, 1]$, $\kappa \ge U^2$ and $\delta \in (0, \min(1/T, \gamma))$,*

$$\textit{Classification Loss} \le 4\gamma r_{\max}\ell_{\max} T + \frac{9 r_{\max}\ell_{\max} B_\delta^2(T) d\ln(1 + T/d)}{\max(\eta,\gamma)^2 \hat{\mu}_{\min}}$$

Lemma 5.3. *For any $\beta, \gamma \in (1/\ell_{\max}, 1]$, $\kappa \ge U^2$ and $\delta \in (0, 1/T)$,*

$$\textit{Idiosyncrasy Loss} \lesssim \mathcal{L}^\star(w, T) + \frac{wT}{\beta} + \ell_{\max}\left(\Delta(\mathcal{K}_{\mathcal{G}})T + \beta\ell_{\max} G d^{1.5} + d\sqrt{GT} + \frac{d^{2.5}}{\max(\eta,\gamma)^2}\right).$$

5.3 Bounding Forced Scheduling and Relaxed Delay Loss (Lemma 5.1)

Similar to Section 4.3, we bound the relaxed delay loss by the sum of the label-driven queue length $\mathbb{E}\left[\sum_{t=1}^{T} Q^{\mathrm{LD}}(t)\right]$ and the review queue length $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{g\in\mathcal{G}}\tilde{Q}_g(t)\right]$. By Lemma 4.4, bounding the first term requires bounding the the expected number of posts that we admit into the label-driven queue, i.e., $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} E_k(t)\right]$. We first define the "good" event $\mathcal{E}$, where the confidence set $\mathcal{C}_t$ is valid for any period t: $\mathcal{E} := \{\forall t, \boldsymbol{\theta}^\star \in \mathcal{C}_t\}$. Our first lemma shows that this event happens with probability at least $1 - \delta$ using [APS11, Theorem 2] (proof in Appendix D.3).

Lemma 5.4. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for any $t \ge 0$, the confidence set $\mathcal{C}_t$ is valid, i.e., $\boldsymbol{\theta}^\star \in \mathcal{C}_t$.*

Our second lemma bounds the expected number of posts admitted into the label-driven queue. The proof is based on classical linear contextual bandit analysis and is provided in Appendix D.4.

Lemma 5.5. *If $\kappa \ge U^2, \gamma \le 1$, $\delta \le 1/T$, the number of posts admitted to the label-driven queue is*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} E_k(t)\right] \le \frac{9B_\delta^2(T)d\ln(1 + T/d)}{\max(\eta,\gamma)^2}.$$

[7] Note that $\sum_{g\in\mathcal{G}}\tilde{Q}_g(t) = \sum_{k\in\mathcal{K}} Q_k(t)$ but summing across groups facilitates our per-group analysis.

The next lemma bounds the review queue length $\tilde{Q}_g(t)$ in a similar way with Lemma 4.7.

Lemma 5.6. *For any g, t, the number of group-g posts in the review queue is $\tilde{Q}_g(t) \le 2\beta r_{\max}\ell_{\max}$.*

Proof. For any group g and period t, COLBACID admits a post in this group in period t only if $\beta \bar{r}_{k(j(t))}(t)\ell_k \ge \tilde{Q}_g(t)$, which happens only if $\beta r_{\max}\ell_{\max} \ge \tilde{Q}_g(t)$ since $\bar{r}_k(t) \le r_{\max}$ for any k, t. By induction, we have $\tilde{Q}_g(t) \le \beta r_{\max}\ell_{\max} + 1 \le 2\beta r_{\max}\ell_{\max}$. □

Proof of Lemma 5.1. Applying Lemmas 4.4, 5.5 and 5.6 gives

$$
\begin{aligned}
\text{Relaxed Delay Loss} &= r_{\max}\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{g\in\mathcal{G}}\tilde{Q}_g(t) + Q^{\mathrm{LD}}(t)\right)\right] \\
&\le r_{\max}\left(2GT\beta r_{\max}\ell_{\max} + \frac{9B_\delta^2(T)d\ln(1+T/d)}{\max(\eta,\gamma)^2\hat{\mu}_{\min}}\right) = 2GT\beta r_{\max}^2\ell_{\max} + \frac{9r_{\max}B_\delta^2(T)d\ln(1+T/d)}{\max(\eta,\gamma)^2\hat{\mu}_{\min}}.
\end{aligned}
$$

□

5.4 Bounding Classification Loss (Lemma 5.2)

Similar to Lemma 4.2, we first bound the per-period classification loss (proved in Appendix D.5);

$$
Z_k(t) = (Y_k(t)(r_k^O - r_k^R)^+ + (1-Y_k(t))(r_k^R - r_k^O)^+)\ell_k(\Lambda_k(t) - A_k(t) - E_k(t)).
$$

Lemma 5.7. *For any type k and period t, $Z_k(t)\mathbb{1}(\mathcal{E}) \le \left(3v_{\max}\gamma\ell_{\max} + r_{\max}\ell_{\max}Q^{\mathrm{LD}}(t)\right)\Lambda_k(t)$.*

Proof of Lemma 5.2. The classification loss is given by $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}Z_k(t)\right]$, which we bound by

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}Z_k(t)\right] &\le r_{\max}\ell_{\max}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}(\mathcal{E}^c)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}Z_k(t)\mathbb{1}(\mathcal{E})\right] && \left(\textstyle\sum_{k\in\mathcal{K}}Z_k(t) \le r_{\max}\ell_{\max}\right) \\
&\le r_{\max}\ell_{\max}\delta T + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}Z_k(t)\,\middle|\,\mathcal{E}\right] && \text{(By Lemma 5.4)} \\
&\le r_{\max}\ell_{\max}\gamma T + \mathbb{E}\left[\sum_{t=1}^{T}\left(3v_{\max}\gamma\ell_{\max} + r_{\max}\ell_{\max}Q^{\mathrm{LD}}(t)\right)\right] && \text{(By Lemma 5.7 and } \delta \le \gamma) \\
&\le 4\gamma r_{\max}\ell_{\max}T + \frac{9r_{\max}\ell_{\max}B_\delta^2(T)d\ln(1+T/d)}{\max(\eta,\gamma)^2\hat{\mu}_{\min}}. && \text{(By Lemmas 4.4, 5.5)}
\end{aligned}
$$

□

5.5 Bounding the Idiosyncrasy Loss (Lemma 5.3)

To bound the idiosyncrasy loss, we connect it with the fluid benchmark via Lagrangians. In the corresponding lemmas of previous section (Lemmas 3.1 and 4.3), we evaluate the Lagrangians by the queue length vector across types $\boldsymbol{Q}(t)$ because we estimate the delay loss of admitting a new type-k post by the number of type-k posts in the review queue. To avoid greatly underestimating the delay loss (due to the larger number of types), a key innovation of COLBACID is to use the number of waiting posts in the same group $\tilde{Q}_{g(k)}(t)$ as an estimator. This new estimator motivates us to use $\tilde{Q}_{g(k)}(t)$ as the dual for Lagrangian analysis and to define a new Lyapunov function

$$\tilde{L}(t) = \beta \sum_{t'=1}^{t-1} \sum_{k \in \mathcal{K}} r_k \ell_k (\Lambda_k(t') - A_k(t') - E_k(t')) + \frac{1}{2} \sum_{g \in \mathcal{G}} \tilde{Q}_g^2(t).$$

We also define the *type-aggregated* queue length vector, $\boldsymbol{Q}^{\mathrm{TA}}(t) = (Q_k^{\mathrm{TA}}(t))_{k \in \mathcal{K}}$, such that $Q_k^{\mathrm{TA}}(t) = \tilde{Q}_{g(k)}(t)$ for any $k \in \mathcal{K}$. We denote $\bar{A}_k(t) = \Lambda_k(t) \mathbb{1}\left(\beta \bar{r}_k(t) \geq \tilde{Q}_{g(k)}(t)\right)$, which captures whether the post would have been admitted in the absence of the label-driven admission. For scheduling, we select a group g that maximizes $\tilde{\mu}_g \tilde{Q}_g(t)$ and choose the first waiting post of that group to review if the label-driven queue is empty. We denote $\psi_{k'}(t) = 1$ where k' is the type of that reviewed post and let $\boldsymbol{\psi}(t) = (\psi_k(t))_{k \in \mathcal{K}}$. The following lemma (proved in Appendix D.6) connects the idiosyncrasy loss and the per-period Lagrangian (7) with the type-aggregated queue lengths $\boldsymbol{Q}^{\mathrm{TA}}(t)$ as the dual.

Lemma 5.8. *The expected Lagrangian of* COLBACID *upper bounds the idiosyncrasy loss by:*

$$\mathbb{E}\left[\beta \sum_{t=1}^{T} \sum_{k \in \mathcal{K}} r_k \ell_k (\Lambda_k(t) - A_k(t) - E_k(t))\right] \leq \mathbb{E}\left[\tilde{L}(T+1) - \tilde{L}(1)\right] \leq T + \sum_{t=1}^{T} \mathbb{E}\left[f_t(\bar{\boldsymbol{A}}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right].$$

Our next step is to bound the Lagrangian with $\boldsymbol{Q}^{\mathrm{TA}}(t)$ as the dual. In Lemma 3.4 (Section 3.4), the Lagrangian of BACID for dual $\boldsymbol{Q}(t)$ is connected to the fluid benchmark via a Lagrangian optimality result of BACID (Lemma 3.7). We also use this result to bound the Lagrangian of OLBACID (Section 4.5). However, with the type-aggregated queue length $\boldsymbol{Q}^{\mathrm{TA}}(t)$ as the dual, the Lagrangian optimality of BACID no longer holds. To deal with this challenge, we thus introduce another benchmark policy which we call TYPE AGGREGATED BACID or TABACID.

Letting $\{\boldsymbol{A}^{\mathrm{TABACID}}(t)\}, \{\boldsymbol{\psi}^{\mathrm{TABACID}}(t)\}$ be the admission and scheduling decisions, TABACID admits a type-k post if $A_k^{\mathrm{TABACID}}(t) = \Lambda_k(t) \mathbb{1}\left(\beta r_k \ell_k \geq \tilde{Q}_{g(k)}(t)\right) = 1$. For scheduling, we first pick a group g maximizing $\tilde{\mu}_g \tilde{Q}_g(t)$ and then select the earliest post j in the review queue that belongs to this group (unless there is no waiting post). We set $\psi_{k(j)}^{\mathrm{TABACID}}(t) = 1$ for the corresponding type. Note that the admission decisions of TABACID are the same as COLBACID except that TABACID uses the ground-truth r_k for admission (not the optimistic estimation) and does not consider the impact of label-driven admission. The scheduling differs from MAXWEIGHT and is sub-optimal due to grouping types with different service rates which is captured by $\Delta(\mathcal{K}_{\mathcal{G}}) \max_g \tilde{Q}_g(t)$. We now upper bounds the expected Lagrangian of TABACID (proved in Appendix D.7).

Lemma 5.9. *For any window size w, the expected Lagrangian of* TABACID *is upper bounded by:*

$$\sum_{t=1}^{T} \mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t), \boldsymbol{\psi}^{\mathrm{TABACID}}(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right] \leq \beta \mathcal{L}^\star(w, T) + (w-1)T + 2\beta r_{\max} \ell_{\max} \Delta(\mathcal{K}_{\mathcal{G}}) T.$$

As in the proof of Lemma 4.3 (Section 4.5), we relate the right hand side in Lemma 5.8 (Lagrangian of COLBACID) to the left hand side in Lemma 5.9 (Lagrangian of TABACID) by

$$\begin{aligned}\textsc{RegL}(T) &= \mathbb{E}\left[\sum_{t=1}^{T} f_t(\bar{\boldsymbol{A}}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}^{\mathrm{TA}}(t)) - f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t), \boldsymbol{\psi}^{\mathrm{TABACID}}(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} (A_k^{\mathrm{TABACID}}(t) - \bar{A}_k(t))(\beta r_k \ell_k - \tilde{Q}_{g(k)}(t))\right]}_{\textsc{RegA}(T)} \\ &\quad + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} (\psi_k^{\mathrm{TABACID}}(t) - \psi_k(t))\tilde{Q}_{g(k)}(t)\mu_k N(t)\right]}_{\textsc{RegS}(T)}. \end{aligned} \tag{17}$$

Letting $Q_{\max} = 2\beta r_{\max}\ell_{\max}$, we bound $\textsc{RegS}(T)$ as for Lemma 4.12 (proof in Appendix D.8).

Lemma 5.10. *If $\kappa \geq U^2, \gamma \leq 1, \delta \leq 1/T$, the regret in scheduling is* $\textsc{RegS}(T) \leq \frac{9Q_{\max}B_\delta^2(T)d\ln(1+T/d)}{\max(\eta,\gamma)^2\hat{\mu}_{\min}}$.

Our novel contribution is the following result bounding the regret in admission $\textsc{RegA}(T)$. The proof handles contextual learning with queueing delayed feedback, which we discuss in Section 5.6.

Lemma 5.11. *If $\kappa \geq U^2, \gamma \leq 1, \delta \leq 1/T$, the regret in admission is*

$$\textsc{RegA}(T) \leq 3\beta r_{\max}\ell_{\max}B_\delta(T)\left(d\ln(1+T/d)\left(4GQ_{\max} + \frac{9B_\delta^2(T)}{\max(\eta,\gamma)^2}\right) + \sqrt{2GTd\ln(1+T/d)}\right).$$

Proof of Lemma 5.3. By Lemma 5.8, we have

$$\begin{aligned}\mathbb{E}\left[\beta\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t) - A_k(t) - E_k(t))\right] &\leq T + \sum_{t=1}^{T}\mathbb{E}\left[f_t(\bar{\boldsymbol{A}}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right] \\ &= T + \sum_{t=1}^{T}\mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t), \boldsymbol{\psi}^{\mathrm{TABACID}}(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right] + \textsc{RegS}(T) + \textsc{RegA}(T) \\ &\leq \beta\mathcal{L}^\star(w,T) + wT + 2\beta r_{\max}\ell_{\max}\Delta(\mathcal{K}_{\mathcal{G}})T + \textsc{RegS}(T) + \textsc{RegA}(T). \qquad \text{(Lemma 5.9)}\end{aligned}$$

The result follows by bounding $\textsc{RegS}(T)$ and $\textsc{RegA}(T)$ respectively by Lemmas 5.10 and 5.11, and by dividing both sides by β and noting that $B_\delta(T) \lesssim \sqrt{d}, Q_{\max} \lesssim \beta\ell_{\max}$. □

5.6 Contextual Learning with Queueing-Delayed Feedback (Lemma 5.11)

To bound the regret in admission in a way that avoids dependence on K (that Lemma 4.11 exhibits), we rely on the contextual structure to more effectively bound the total estimation error of admitted posts. This has the additional complexity that feedback is observed after a queueing delay (that is endogenous on the algorithmic decisions) and is handled via Lemma 5.13 below. The proof of Lemma 5.11 (Appendix D.9) then follows from classical linear contextual bandit analysis.

The estimation error for one post with feature $\boldsymbol{\phi}_k$ given observed data points that form matrix $\bar{V}_{t-1}$ (defined in (13)) corresponds to $\|\boldsymbol{\phi}_k\|_{\bar{V}_{t-1}^{-1}}$. To see the correspondence, in a non-contextual setting, $\boldsymbol{\phi}_k$ is a unit vector for the $k-$th dimension, and $\bar{\boldsymbol{V}}_{t-1}$ is a diagonal matrix where the $k-$th element is the number of type-k reviewed posts $n_k(t-1)$. As a result, $\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}} = 1/\sqrt{n_k(t-1)}$, which is the estimation error we expect from a concentration inequality.

Our first result (proof in Appendix D.10) bounds the estimation error when feedback of all admitted posts is delayed by a fixed duration, which we utilize to accommodate random delays.

Lemma 5.12. *Given a sequence of M vectors $\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_M$ in $\mathbb{R}^d$, let $\hat{\boldsymbol{V}}_j = \kappa \boldsymbol{I} + \sum_{j'=1}^{j} \hat{\boldsymbol{\phi}}_{j'} \hat{\boldsymbol{\phi}}_{j'}^\intercal$. If $\|\hat{\boldsymbol{\phi}}_i\|_2 \leq U$ for any $i \leq M$ and $\kappa \geq U^2$, the estimation error for a fixed delay $q \geq 1$ is*

$$\sum_{i=q}^{M} \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} \leq \sqrt{2Md\ln(1+M/d)} + 2qd\ln(1+M/d).$$

The challenge in our setting is that the feedback delay is not fixed, but is indeed affected by both the admission and scheduling decisions. The following lemma upper bounds the estimation error under this queueing delayed feedback, enabling our contextual online learning result.

Lemma 5.13. *If $\kappa \geq U^2$ and $T \geq 3$, the estimation error of admitted posts is bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{k} \|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}} \bar{A}_k(t)\right] \leq \sqrt{2GTd\ln(1+T/d)} + d\ln(1+T/d)\left(3GQ_{\max} + \frac{9B_\delta^2(T)}{\max(\eta,\gamma)^2}\right).$$

Proof sketch. The proof contains three steps. The first step is to connect the error of a post in our setting to a fixed-delay setting by the first-come-first-serve (FCFS) property of our scheduling algorithm (Lemma D.5). In particular, consider the sequence of admitted group-g posts. For a post j on this sequence, the set of posts before j whose feedback is still not available can include at most the $Q_{\max}$ posts right before j on the sequence (where $Q_{\max}$ is controlled by our admission rule). Therefore, the error of group-g posts accumulates as in a setting with a fixed delay $Q_{\max}$.

Based on this result, the second step (Lemma D.6) bounds $\mathbb{E}\left[\sum_{t=1}^{T} \sum_k \|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}} A_k(t)\right]$, which is the estimation error for posts admitted *by the optimistic admission rule.* Enabled by the connection to the fixed-delay setting, we bound the estimation error for each group separately by Lemma 5.12 and aggregate them to get the total error, i.e, the first two terms in Lemma 5.13.

For the third step, corresponding to the last term of our bound, we bound the difference between the error of all admitted posts and the error of posts admitted by the optimistic admission (which we upper bounded in the second step). We show that this difference is at most the number of label-driven admissions $\mathbb{E}\left[\sum_{t=1}^{T} \sum_k E_k(t)\right]$ (bounded by Lemma 5.5) because $\bar{A}_k(t) - A_k(t) \leq E_k(t)$ and $\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}} \leq 1$. The full proof is provided in Appendix D.11. □

6 Conclusion

Motivated by the human-AI interplay in online content moderation, we propose a learning to defer model with limited and time-varying human capacity. In particular, for each period, the AI makes three decisions: (i) whether to keep or remove a new post (classification); (ii) whether to admit this new post for human review (admission); and (iii) which post to send to the next available reviewer (scheduling). The cost of a post is unknown until a human reviews it. The objective is to minimize the total loss with respect to an omniscient benchmark that knows the cost of posts.

Since achieving vanishing loss is unattainable, we aim to minimize the average regret of a policy, i.e., the difference between the loss of a policy and a fluid benchmark. When the average cost of posts is known, we propose BACID which balances the idiosyncrasy loss avoided by admitting a post and the delay loss the admission could incur to other posts. We show that BACID achieves a near-optimal $O(\sqrt{\ell_{\max}})$ regret, with $\ell_{\max}$ being the maximum lifetime of a post. When the average cost of posts is unknown, we show that an optimism-only extension of BACID fails to learn because of the *selective sampling* nature of the system. That is, humans only see posts that are admitted by the AI, while labels from humans affect AI's classification and admission accuracy. To address this issue, we carefully balance label-driven admissions and admissions aiming to reduce idiosyncrasy loss. Finally, we extend our algorithm to a contextual setting an derive a performance guarantee without dependence on the number of types. Our type aggregation technique for a many-class queueing system and analysis for queueing-delayed feedback enable us to provide (to the best of our knowledge) the first online learning result in contextual queueing systems, which may be of independent interest.

Our work also opens up a list of interesting questions. First, we assume reviewers produce perfect labels; how can we schedule posts when reviewers have skill-dependent non-perfect review quality? Second, our algorithm requires the knowledge of post lifetime, views per-period, and assumes constant views per-period. Can we relax these assumptions? Third, we focus on the w-fluid benchmark to handle non-stationary arrivals and human capacities, which is not necessary the tightest benchmark. Is there an alternative benchmark to consider? Fourth, we assume stationary expected cost of posts for the learning problem. It is interesting to extend our work to a non-stationary learning setting. Finally, our work focuses on the human-AI interplay in online content moderation and it is interesting to study this interplay in settings beyond content moderation.

References

[ABB+22] Vashist Avadhanula, Omar Abdul Baki, Hamsa Bastani, Osbert Bastani, Caner Gocmen, Daniel Haimovich, Darren Hwang, Dima Karamshuk, Thomas J. Leeper, Jiayuan Ma, Gregory Macnamara, Jake Mullett, Christopher Palow, Sung Park, Varun S. Rajagopal, Kevin Schaeffer, Parikshit Shah, Deeksha Sinha, Nicolás Stier Moses, and Peng Xu. Bandits for online calibration: An application to content moderation on social media platforms. 2022. 3, 4, 23, 40, 41

[AD16] Shipra Agrawal and Nikhil R. Devanur. Linear contextual bandits with knapsacks. In *Annual Conference on Neural Information Processing Systems 2016*, pages 3450–3458, 2016. 6, 51

[AD19] Shipra Agrawal and Nikhil R Devanur. Bandits with global convex constraints and objective. *Operations Research*, 67(5):1486–1502, 2019. 6, 17, 51

[ADL16] Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *the 29th Conference on Learning Theory, COLT*, 2016. 6

[AGP+23] Narges Ahani, Paul Gölz, Ariel D Procaccia, Alexander Teytelboym, and Andrew C Trapp. Dynamic placement in refugee resettlement. *Operations Research*, 2023. 5

[APS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *25th Annual Conference on Neural Information Processing Systems 2011*, pages 2312–2320, 2011. 5, 23, 26, 54, 55, 62, 66

[AS23] Saghar Adler and Vijay Subramanian. Bayesian learning of optimal policies in markov decision processes with countably infinite state-space. *arXiv preprint arXiv:2306.02574*, 2023. 7

[AZ09] Nilay Tanık Argon and Serhan Ziya. Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management*, 11(4):674–693, 2009. 7

[BBS21] Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454*, 2021. 5

[BDL+20] Kianté Brantley, Miroslav Dudík, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Annual Conference on Neural Information Processing Systems, NeurIPS 2020*, 2020. 6

[BKS18] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018. 6

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. 65, 66

[BLS14] Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR, 2014. 6

[BM17] Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534, 2017. 5

[BP22] Kirk Bansak and Elisabeth Paulson. Outcome-driven dynamic refugee assignment with allocation balancing. In *The 23rd ACM Conference on Economics and Computation*, pages 1182–1183. ACM, 2022. 5

[BW08] Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 9:1823–1840, 2008. 6

[BXZ23] Jose Blanchet, Renyuan Xu, and Zhengyuan Zhou. Delay-adaptive learning in generalized linear contextual bandits. *Mathematics of Operations Research*, 2023. 7, 54

[BZ12] Omar Besbes and Assaf Zeevi. Blind network revenue management. *Operations research*, 60(6):1537–1550, 2012. 6

[CBGO09] Nicolo Cesa-Bianchi, Claudio Gentile, and Francesco Orabona. Robust bounds for classification via selective sampling. In *Proceedings of the 26th annual international conference on machine learning*, pages 121–128, 2009. 6

[CDG+18] Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1067–1075. PMLR, 2018. 6

[CDM16] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *Algorithmic Learning Theory - 27th International Conference, ALT*, volume 9925 of *Lecture Notes in Computer Science*, pages 67–82, 2016. 6

[Cha14] Olivier Chapelle. Modeling delayed feedback in display advertising. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 1097–1105. ACM, 2014. 7

[Che19] Wang Chi Cheung. Regret minimization for reinforcement learning with vectorial feedback and complex objectives. In *Annual Conference on Neural Information Processing Systems, NeurIPS 2019*, pages 724–734, 2019. 6

[Cho57] C. K. Chow. An optimum character recognition system using decision functions. *IRE Trans. Electron. Comput.*, 6(4):247–254, 1957. 5

[Cho70] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16(1):41–46, 1970. 5

[CJWS21] Tuhinangshu Choudhury, Gauri Joshi, Weina Wang, and Sanjay Shakkottai. Job dispatching policies for queueing systems with unknown service rates. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 181–190, 2021. 7

[CLH23] Xinyun Chen, Yunan Liu, and Guiyu Hong. Online learning and optimization for queues with unknown demand curve and service distribution. *arXiv preprint arXiv:2303.03399*, 2023. 7

[CLRS11] Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011*, 2011. 23

[CLS05] Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Trans. Inf. Theory*, 51(6):2152–2162, 2005. 6

[CMSS22] Mohammad-Amin Charusaie, Hussein Mozannar, David A. Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning, ICML 2022*, 2022. 6

[DFC20] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM, 2020. 5

[DGS12] Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *J. Mach. Learn. Res.*, 13:2655–2697, 2012. 6

[DHK+11] Miroslav Dudík, Daniel J. Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178. AUAI Press, 2011. 7

[DKGG20] Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression under human assistance. In *The Thirty-Fourth Conference on Artificial Intelligence, AAAI 2020*, pages 2611–2620. AAAI Press, 2020. 6

[DOZR21] Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. Classification under human assistance. In *Thirty-Fifth Conference on Artificial Intelligence, AAAI 2021*, pages 5905–5913. AAAI Press, 2021. 6

[Eub18] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, 2018. 1, 5

[EW10] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 2010. 6

[FHL22] Hu Fu, Qun Hu, and Jia'nan Lin. Stability of decentralized queueing networks beyond complete bipartite cases. In *Web and Internet Economics - 18th International Conference, WINE*, volume 13778 of *Lecture Notes in Computer Science*, pages 96–114. Springer, 2022. 7

[FLW23a] Daniel Freund, Thodoris Lykouris, and Wentao Weng. Efficient decentralized multi-agent learning in asymmetric bipartite queueing systems. *Operations Research*, 2023. 7

[FLW23b] Daniel Freund, Thodoris Lykouris, and Wentao Weng. Quantifying the cost of learning in queueing systems. *arXiv preprint arXiv:2308.07817*, 2023. 7

[FM22] Xinzhe Fu and Eytan Modiano. Optimal routing to parallel servers with unknown utilities—multi-armed bandit with queues. *IEEE/ACM Transactions on Networking*, 2022. 7

[FSW18] Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018. 6

[GBB23] Haosen Ge, Hamsa Bastani, and Osbert Bastani. Rethinking fairness for human-ai collaboration. *arXiv:2310.03647*, 2023. 5

[GBK20] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, 2020. 5

[GO23] Github and OpenAI. Github copilot, 2023. `https://github.com/features/copilot`. Accessed on December 19, 2023. 1

[GT23] Jason Gaitonde and Éva Tardos. The price of anarchy of strategic queuing systems. *Journal of the ACM*, 2023. 7

[HGH23] Jiatai Huang, Leana Golubchik, and Longbo Huang. Queue scheduling with adversarial bandit learning. *arXiv preprint arXiv:2303.01745*, 2023. 7

[HN11] Longbo Huang and Michael J. Neely. Delay reduction via lagrange multipliers in stochastic network optimization. *IEEE Trans. Autom. Control.*, 56(4):842–857, 2011. 8

[HXLB22] Wei-Kang Hsu, Jiaming Xu, Xiaojun Lin, and Mark R Bell. Integrated online learning and adaptive control in queueing systems with uncertain payoffs. *Operations Research*, 70(2):1166–1181, 2022. 7

[JGS13] Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, volume 28, pages 1453–1461. JMLR.org, 2013. 7

[JKK21] Ramesh Johari, Vijay Kamble, and Yash Kanoria. Matching while learning. *Operations Research*, 69(2):655–681, 2021. 7

[JSS22a] Huiwen Jia, Cong Shi, and Siqian Shen. Online learning and pricing for network revenue management with reusable resources. In *NeurIPS*, 2022. 7

[JSS22b] Huiwen Jia, Cong Shi, and Siqian Shen. Online learning and pricing for service systems with reusable resources. *Operations Research*, 2022. 7

[KAJS18] Subhashini Krishnasamy, Ari Arapostathis, Ramesh Johari, and Sanjay Shakkottai. On learning the cμ rule in single and parallel server networks. *arXiv preprint arXiv:1802.06723*, 2018. 7

[KLK21] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165. ACM, 2021. 6

[KLL$^+$18] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018. 5

[KMT98] Frank P Kelly, Aman K Maulloo, and David Kim Hong Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49:237–252, 1998. 8

[KSJS21] Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. Learning unknown service rates in queues: A multiarmed bandit approach. *Oper. Res.*, 69(1):315–330, 2021. 7

[LCLS10] 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 2010*, pages 661–670. ACM, 2010. 23

[Lit61] John DC Little. A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387, 1961. 12

[LJWX23] Shuangning Li, Ramesh Johari, Stefan Wager, and Kuang Xu. Experimenting under stochastic congestion. *arXiv preprint arXiv:2302.12093*, 2023. 7

[LLL22] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. To engage or not to engage with ai for critical judgments: How professionals deal with opacity when using ai for medical diagnosis. *Organization Science*, 33(1):126–148, 2022. 5

[LLSY21] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. In *Annual Conference on Neural Information Processing Systems*, pages 24075–24086, 2021. 6

[LLWS11] Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: a framework for self-aware learning. *Mach. Learn.*, 82(3):399–443, 2011. 6

[LM18] Qingkai Liang and Eytan Modiano. Minimizing queue length regret under adversarial network models. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):1–32, 2018. 7

[LSFB22] Diogo Leitão, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. Human-ai collaboration in decision-making: Beyond learning to defer. *CoRR*, abs/2206.13202, 2022. 6

[LSKM21] Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5969–5978. PMLR, 2021. 7

[LSS06] Xiaojun Lin, Ness B Shroff, and Rayadurgam Srikant. A tutorial on cross-layer optimization in wireless networks. *IEEE Journal on Selected areas in Communications*, 24(8):1452–1463, 2006. 8

[LSY21] Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 6483–6492. PMLR, 2021. 6

[MCPZ18] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3381–3390. PMLR, 2018. 6

[Met22] Meta Platforms, Inc. How meta prioritizes content for review, 2022. `https://transparency.fb.com/policies/improving/prioritizing-content-review/`. Accessed on December 19, 2023. 1

[Met23] Meta Platforms, Inc. Facebook community standards, 2023. `https://transparency.fb.com/policies/community-standards/`. Accessed on December 19, 2023. 1

[MLBP15] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popovic. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2849–2856, 2015. 7

[MMZ23] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. *arXiv preprint arXiv:2310.14774*, 2023. 6

[MS20] Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR, 2020. 6

[MS22] Bryce McLaughlin and Jann Spiess. Algorithmic assistance with recommendation-dependent preferences. *arXiv preprint arXiv:2208.07626*, 2022. 5

[MSA+21] Rahul Makhijani, Parikshit Shah, Vashist Avadhanula, Caner Gocmen, Nicolás E Stier-Moses, and Julián Mestre. Quest: Queue simulation for content moderation at scale. *arXiv preprint arXiv:2103.16816*, 2021. 2, 3, 5, 40, 41

[MSG22] Bryce McLaughlin, Jann Spiess, and Talia Gillis. On the fairness of machine-assisted human decisions. In *ACM Conference on Fairness, Accountability, and Transparency*, page 890. ACM, 2022. 5

[MVCV20] Anne Gael Manegueu, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *the 37th International Conference on Machine Learning, ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3348–3356. PMLR, 2020. 7

[MX18] Laurent Massoulié and Kuang Xu. On the capacity of information processing systems. *Operations Research*, 66(2):568–586, 2018. 7

[Nee22] Michael Neely. *Stochastic network optimization with application to communication and queueing systems*. 2022. 3, 8

[NM23] Quang Minh Nguyen and Eytan Modiano. Learning to schedule in non-stationary wireless networks with unknown statistics. *arXiv preprint arXiv:2308.02734*, 2023. 7, 11

[Nov23] Matt Novak. Lawyer uses chatgpt in federal court and it goes horribly wrong, 2023. `https://www.forbes.com/sites/mattnovak/2023/05/27/lawyer-uses-chatgpt-in-federal-court-and-it-goes-horribly-wrong/?sh=735755923494`. Accessed on December 19, 2023. 1

[NRP12] Michael J. Neely, Scott Rager, and Thomas F. La Porta. Max weight learning algorithms for scheduling in unknown environments. *IEEE Trans. Autom. Control.*, 57(5):1179–1191, 2012. 7

[NZ20] Gergely Neu and Nikita Zhivotovskiy. Fast rates for online prediction with abstention. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 3030–3048. PMLR, 2020. 6

[OC11] Francesco Orabona and Nicolò Cesa-Bianchi. Better algorithms for selective sampling. In *Proceedings of the 28th International Conference on Machine Learning, ICML*, pages 433–440, 2011. 6

[PGBJ21] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter L. Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 2827–2835. PMLR, 2021. 6

[Pin23] Ping An Insurance. Insurance business, 2023. `https://group.pingan.com/about_us/our_businesses/insurance.html`. Accessed on December 19, 2023. 1

[PZ22] Nikita Puchkin and Nikita Zhivotovskiy. Exponential savings in agnostic active learning through abstention. *IEEE Trans. Inf. Theory*, 68(7):4651–4665, 2022. 6

[RBC+19] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019. 6

[Red23] Reddit Inc. Automoderator, 2023. `https://www.reddit.com/wiki/automoderator/`. Accessed on December 19, 2023. 1

[RK21] Sebastian Raisch and Sebastian Krakowski. Artificial intelligence and management: The automation–augmentation paradox. *Academy of management review*, 46(1):192–210, 2021. 5

[Rob19] Sarah T Roberts. *Behind the screen*. Yale University Press, 2019. 1

[SBP21] Flore Sentenac, Etienne Boursier, and Vianney Perchet. Decentralized learning in online queuing systems. pages 18501–18512, 2021. 7

[SGMV20] Virag Shah, Lennart Gulikers, Laurent Massoulié, and Milan Vojnovic. Adaptive matching for expert systems with uncertain task types. *Oper. Res.*, 68(5):1403–1424, 2020. 7

[SGVM22] Simrita Singh, Itai Gurvich, and Jan A Van Mieghem. Feature-based priority queuing. *Available at SSRN 3731865*, 2022. 7

[SSF23] Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J. Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *The Thirty Sixth Annual Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4633–4656. PMLR, 2023. 6

[SSM19] Thomas Stahlbuhk, Brooke Shrader, and Eytan H. Modiano. Learning algorithms for scheduling in wireless networks with unknown channel statistics. *Ad Hoc Networks*, 85:131–144, 2019. 7

[SSM21] Thomas Stahlbuhk, Brooke Shrader, and Eytan H. Modiano. Learning algorithms for minimizing queue length regret. *IEEE Trans. Inf. Theory*, 67(3):1759–1781, 2021. 7

[Sto05] Alexander L Stolyar. Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm. *Queueing Systems*, 50:401–457, 2005. 8

[SZB10] Amin Sayedi, Morteza Zadimoghaddam, and Avrim Blum. Trading off mistakes and don't-know predictions. In *24th Annual Conference on Neural Information Processing Systems*, pages 2092–2100, 2010. 6

[Tas92] Leandros Tassiulas. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, 31(12), 1992. 7, 13

[VBN23] Rajeev Verma, Daniel Barrejón, and Eric T. Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11415–11434. PMLR, 2023. 6

[VCL+20] Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brückner. Linear bandits with stochastic delayed feedback. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9712–9721. PMLR, 2020. 7

[VCP17] Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI*. AUAI Press, 2017. 7, 54

[VLSV23] Harit Vishwakarma, Heguang Lin, Frederic Sala, and Ramya Korlakai Vinayak. Promises and pitfalls of threshold-based auto-labeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5

[Wal14] Neil S. Walton. Two queues with non-stochastic arrivals. *Oper. Res. Lett.*, 42(1):53–57, 2014. 7

[WDY14] Zizhuo Wang, Shiming Deng, and Yinyu Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014. 6

[WHK20] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020. 6

[WSLJ15] Huasen Wu, R. Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Annual Conference on Neural Information Processing Systems*, pages 433–441, 2015. 6

[WW22] Han Wu and Stefan Wager. Thompson sampling with unrestricted delays. In *The 23rd ACM Conference on Economics and Computation*, pages 937–955. ACM, 2022. 7

[WX21] Neil Walton and Kuang Xu. Learning and information in stochastic networks and queues. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 161–198. INFORMS, 2021. 7

[X C23] X Corp. The X rules, 2023. `https://help.twitter.com/en/rules-and-policies/x-rules`. Accessed on December 19, 2023. 1

[Yeh23] Guy Yehiav. Will ai augment or replace workers?, 2023. `https://www.forbes.com/sites/forbestechcouncil/2023/08/08/will-ai-augment-or-replace-workers/`. 1

[YSY23] Zixian Yang, R Srikant, and Lei Ying. Learning while scheduling in multi-server systems with unknown statistics: Maxweight with discounted ucb. In *International Conference on Artificial Intelligence and Statistics*, pages 4275–4312. PMLR, 2023. 7, 11

[ZBW22] Yueyang Zhong, John R Birge, and Amy Ward. Learning the scheduling policy in time-varying multiclass many server queues with abandonment. *SSRN*, 2022. 7

[ZC16] Chicheng Zhang and Kamalika Chaudhuri. The extended littlestone's dimension for learning with mistakes and abstentions. In *Proceedings of the 29th Conference on Learning Theory, COLT*, volume 49, pages 1584–1616. JMLR.org, 2016. 6

[ZN22] Yinglun Zhu and Robert Nowak. Efficient active learning with abstention. In *NeurIPS*, 2022. 6

A Further discussion on content moderation practice (Section 1)

In this section we summarize the key algorithmic approaches in existing content moderation practice [MSA+21, ABB+22]. [MSA+21] presents a queueing simulation framework QUEST and claims that it is "used at Facebook" (see [ABB+22, Section 1, Page 2]). In addition, [ABB+22] claims that "Our approach has been deployed at scale at Meta." (see [ABB+22, Section 5, Page 6]).

[MSA+21] focuses on scheduling in the human review system. In particular, the system consists of jobs, reviewers and queues of jobs. A job represents a piece of content and is created a) when a user reports this piece, or b) from an existing job, or c) when a classifier considers it to potentially violate the policy. In addition, a borderline post that is removed proactively by the classifier may also join the review system based on criteria depending on its type, realized content views, and whether its classifier violation score is below a manually set threshold. When a job is created, it joins a specific queue based on its attributes. Each queue is a priority queue where jobs are ranked based on their severity-aware review value that depends on predicted content views, the predicted severity, and the review time. When a reviewer becomes available, the system first selects a queue for them (constrained by the reviewer's skills such as language and system configuration), and schedules the job with the highest review value in the queue to this reviewer. The prioritization of queues may be percentage-based to avoid starvation. The reviewer then decides whether to keep or remove the post, transfer the job to another queue, or reinsert it into the original queue.

[ABB+22] describes in more details how Meta leverages various machine learning models to flag potentially violating content that enters the human review system by using contextual bandit to calibrate an aggregated score that dynamically adapts to new trends. In particular, a job arrives in period t with features $x_{t,i}$ given by outputs from different machine learning models. If the corresponding post is unambiguously harmful based on these models, it is automatically removed from the platform. Otherwise, the system makes an admission decision a_t on whether this new job needs to have human reviews like in ours. When $a_t = 1$, it is assumed that a human reviewer observes the ground-truth severity y_t. Different from c_j in our model, which can be negative, y_t is non-negative such that $y_t = 0$ if it is non-violating, and $y_t > 0$ if it is harmful. If $y_t > 0$, the system removes the corresponding content. This paper regards y_t as a reward such that the algorithm aims to maximize the total violation score from admitted posts.

Their algorithm works as follows. When a post arrives, the system calculates $\hat{y}_t = \max_i f_{\beta_i}(x_{t,i})$, where β_i is a vector representing rescaling parameters for model i and $f_{\beta_i}(z) = \sum_{j=1}^{k} \mathbb{1}(z \in B_j)\, \beta_{i,j} z$ is a piecewise linear function where B_j defines a bin. The algorithm assumes a linear model such that $y_t = \beta_{i,j} x_{t,j} + \varepsilon_{t,i,j}$ with $\varepsilon_{t,j}$ being sub-Gaussian noise, which enables the use of linear regression to obtain an estimator $\{\hat{\beta}_i\}$ of the unknown calibration parameters $\{\beta_i\}$. For each arrival, the algorithm uses upper confidence bound of $\{\beta_i\}$ to promote exploration. In particular, it generates a confidence interval $\{u_i\}$ such that $\hat{\beta}_{i,j} + u_{i,j} \geq \beta_{i,j}$ with high probability, and then estimates y_t by $\hat{y}_t = \max_i f_{\hat{\beta}_i + u_i}(x_{t,i})$. It admits a post if $\hat{y}_t > 0$, and updates parameters when true severity y_t is observed. To account for content lifetime, [ABB+22] consider a metric called *integrity value* (IVs) for each post, defined as (predicted future views + constant) $\times$ severity, and aim to minimize the total integrity value. The goal is to maximize the sum of IVs of posts reviewed by humans, which is close to our objective of minimizing cost. Instead of making binary admission decision, they consider a new algorithm that maintains a pool of all currently available posts C_t, where a time step is defined as the time a new reviewer is available. The decision is to pick a post from C_t for every t. They estimate the IVs for each post in C_t by replacing the ground-truth severity (which is unknown) by an upper confidence bound estimation $\max_i f_{\hat{\beta}_i + u_i}(\phi_i(c))$ like $\hat{y}_t$ mentioned above, where $\phi_i(c)$ is the output of model i for post c. They then select a post that maximizes the *estimated rate of change in IVs* for review.

Although [MSA+21, ABB+22] do not provide a formal model, their views of the system are largely captured by our model. In particular, the severity of a content is modelled by the cost c_j of a post in our model. Their admission policy, which removes posts that ML views as unambiguously harmful, and admits posts with predicted positive severity, can be viewed as a static two-threshold admission policy. Delaying reviews of harmful contents is undesirable for the system, as emphasized by our objective and the IV metric in [ABB+22]. We note that our paper does not aim to capture the full picture of content moderation. We assume that average cost is unknown but stationary, that per-period views of posts do not change over time, that we know the per-period views, the lifetime, and the idiosyncratic variance and that human reviewers are perfect (see Section 6 for further discussion). Approaches in [MSA+21, ABB+22] do not rely on these assumptions. However, the goal of our paper is to provide a stylized model that illustrates challenges arising from idiosyncratic variance in posts, time-varying capacity and selective sampling, and that relying on sample average and optimism-only learning policies may be inefficient in this setting. These justify the contribution of our paper which presents a formal model to address the two challenges.

B Supplementary materials on BACID (Section 3)

B.1 Delay Loss of Human-Only (Section 3.1)

Recall that for the Human-Only example, the review queue has an arrival and a post is reviewed (and leaves the queue) with probability 0.5 every period. Also recall that $\ell_1 \geq 25$. The following lemma shows that when $T \geq 21\ell_1$, we must have $\mathbb{E}\left[\sum_{t=1}^{T-\ell_1} \min(Q(t), \ell_1)\right] \geq \frac{2T\ell_1}{3}$.

Lemma B.1. *Under the above setting, we have* $\mathbb{E}\left[\sum_{t=1}^{T-\ell_1} \min(Q(t), \ell_1)\right] \geq \frac{2T\ell_1}{3}$ *when* $T \geq 21\ell_1$*.*

Proof. Recall the queue dynamic is $Q(t+1) = Q(t) + A(t) - S(t)$ where $S(t)$ is a Bernoulli random variable with mean 0.5 and $A(t) = 1$. As a result, $Q(t) = t - \sum_{\tau=1}^{t-1} S(\tau) \geq t - \sum_{\tau=1}^{t} S(\tau)$. For every period t, define a good event $\mathcal{C}_t = \{\sum_{\tau=1}^{t} S(\tau) \leq \frac{t}{2} + \sqrt{t \ln t}\}$. By Hoeffding's Inequality (Fact 1), we have $\mathbb{P}\{\mathcal{C}_t\} \geq 1 - 1/t^2$. In addition, by Fact 8, we have $t/4 \geq \sqrt{t \ln t}$ for $t \geq 100$. As a result, for $t \geq \max(100, 4\ell_1) = 4\ell_1$, condition on $\mathcal{C}_t$, we have

$$Q(t) \geq t - (\frac{t}{2} + \sqrt{t \ln t}) \geq \frac{t}{4} \geq \ell_1.$$

The delay loss of Human-Only is thus lower bounded by

$$\mathbb{E}\left[\sum_{t=1}^{T-\ell_1} \min(Q(t), \ell_1)\right] \geq \sum_{t=4\ell_1}^{T-\ell_1} \ell_1 \mathbb{P}\{\mathcal{C}_t\} \geq \sum_{t=4\ell_1}^{T-\ell_1} \ell_1(1 - 1/t^2) \geq (T - 5\ell_1)\ell_1 - 2\ell_1 \geq (T - 7\ell_1)\ell_1$$

where the third inequality is because $\sum_{t=1}^{\infty} 1/t^2 \leq 2$. We finish the proof by noting that $T \geq 21\ell_1$ and thus $(T - 7\ell_1)\ell_1 \geq \frac{2T\ell_1}{3}$. □

B.2 Formulae to calculate per-period loss (Eq. (5))

Proposition 4. *For any normal random variable X with mean h and variance σ^2, we have*

$$\mathbb{E}\left[X^+\right] = h\Phi(h/\sigma) + \sigma\varphi(h/\sigma) \qquad \textit{and} \qquad \mathbb{E}\left[X^-\right] = -h\Phi(-h/\sigma) + \sigma\varphi(-h/\sigma).$$

Proof. We prove the result for $\mathbb{E}[X^-]$. The result for $\mathbb{E}[X^+]$ follows by considering $\mathbb{E}[(-X)^-]$. We have $\mathbb{E}[X^-] = -\mathbb{P}\{X \le 0\}\mathbb{E}[X \mid X \le 0]$. Conditioning on $X \le 0$, X has a truncated normal distribution; the mean $\mathbb{E}[X \mid X \le 0]$ is thus $h + \frac{-\varphi(-h/\sigma)}{\Phi(-h/\sigma)}\sigma$. Since $\mathbb{P}\{X \le 0\} = \Phi(-h/\sigma)$, we have

$$\mathbb{E}\left[X^-\right] = -\Phi(-h/\sigma)\left(h + \frac{-\varphi(-h/\sigma)}{\Phi(-h/\sigma)}\sigma\right) = -h\Phi(-h/\sigma) + \sigma\varphi(-h/\sigma).$$

□

B.3 Proof of the Lower Bound (Theorem 2)

Proof. Consider a setting where there is only one type of posts ($K = 1$). Since there is only one post, we will omit notational dependence on the type. The arrival rate is $\lambda = 1/2$ and the service rate is $\mu = 1/2$. We assume $N(t) = 1$. For this type, $h = 0$ and $\sigma = 1$; so the inaccuracy loss per post is $r = 1$ by Lemma 4. The lifetime is ℓ. Recall $D(j)$ is the delay for a post to review. Let $Q(t)$ be the length of the review queue at the beginning of period t. Then for a policy π,

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}^\pi(T)\right] &= \mathbb{E}\left[\ell\sum_{t=1}^{T}(\Lambda(t) - A(t)) + \sum_{t=1}^{T} A(t)\min(D(j(t)), \ell)\right] \\ &\ge \mathbb{E}\left[\ell\sum_{t=1}^{T}(\Lambda(t) - A(t)) + \sum_{t=1}^{T} A(t)\min(Q(t), \ell)\right]\end{aligned}$$

where the first term corresponds to the inaccuracy loss, and the second term is the loss when a post is waiting for review (note that it is capped by ℓ since a post has no view after ℓ periods), and the inequality is because $D(j(t)) \ge Q(t)$ due to FCFS.

We consider the case that $T \to \infty$ and thus the system is effectively a discrete-time Markov Decision Process, where the state is defined by the length of the review queue $Q(t)$. Let π be a deterministic stationary policy such that the policy is a function $\pi(q) \in \{0, 1\}$ from the observed queue length to a an admit or non-admit decision. By our model assumption, $Q(1) = 0$. If $\pi(q') = 0$ for some q', no state after q' is reachable. Therefore, let Θ be the minimum q with $\pi(q) = 0$. We only need to focus on states $\{1, \ldots, \Theta\}$. If $\Theta = \infty$, i.e., the policy admits all posts, we have $Q(t) \to \infty$ as t increases because $\lambda = \mu$ and thus $\lim_{T\to\infty} \frac{\mathbb{E}[\mathcal{L}^\pi(T)]}{T} = \ell$. Therefore, it suffices to consider finite Θ. Let $p_{q,q'}$ be the transition probability from state q to state q'. We have

$$\begin{aligned}&p_{0,0} = 1 - \lambda + \lambda\mu,\ p_{0,1} = \lambda(1-\mu) \\ &p_{q,q-1} = (1-\lambda)\mu,\ p_{q,q} = (1-\lambda)(1-\mu),\ p_{q,q+1} = \lambda(1-\mu),\ \forall\, 0 < q < \Theta \\ &\quad p_{\Theta,\Theta-1} = \mu,\ p_{\Theta,\Theta} = 1-\mu.\end{aligned}$$

Since the Markov chain has a finite state space and is irreducible, there is a stationary distribution.

The $w-$ fluid benchmark has $\ell^\star(w, T) = 0$ as in a fluid sense, all posts can be admitted without violating the capacity. It thus remains to lower bound $\lim_{T\to\infty} \frac{\mathbb{E}[\mathcal{L}^\pi(T)]}{T}$ which is lower bounded by

$\frac{\ell(1-p^{\text{admit}})}{2} + \mathbb{E}\left[A(\infty)\min(Q(\infty),\ell)\right]$ where p^{admit} is the long-run average of admission probability, $A(\infty)$ is the stationary random variable of whether a post is admitted, and $Q(\infty)$ is the stationary random variable for queue length. Let ν_q be the stationary distribution of states. Since our system is a Geo/Geo/1 queue with balking at $q = \Theta$ and $\lambda = \mu = 1/2$, solving the local balance quation gives $\nu_0 = \ldots = \nu_{\Theta-1}, \nu_{\Theta-1} = 2\nu_\Theta$, which gives $\nu_q = \frac{2}{2\Theta+1}$ for $q < \Theta$ and $\nu_\Theta = \frac{1}{2\Theta+1}$. As a result,

$$\begin{aligned}\lim_{T\to\infty}\frac{\mathbb{E}\left[\mathcal{L}^\pi(T)\right]}{T} \geq \frac{\ell(1-p^{\text{admit}})}{2} + \mathbb{E}\left[A(\infty)\min(Q(\infty),\ell)\right] &= \frac{\ell\nu_\Theta}{2} + \sum_{q=0}^{\Theta-1}\lambda\nu_q\min(q,\ell) \\ &= \frac{\ell}{4\Theta+2} + \frac{1}{2\Theta+1}\sum_{q=0}^{\Theta-1}\min(q,\ell)\end{aligned}$$

where the first equality is because the policy admits a new post if and only if $q < \Theta$. We then lower bound $f(\Theta) \coloneqq \frac{\ell}{4\Theta+2} + \frac{1}{2\Theta+1}\sum_{q=0}^{\Theta-1}\min(q,\ell)$ by two cases:

- $\Theta \geq \ell$. Then $f(\Theta) \geq \frac{1}{2\Theta+1}\left(\frac{\ell(\ell+1)}{2} + \ell(\Theta-\ell)\right) = \frac{\ell}{4\Theta+2}(2\Theta-\ell+1) \geq \frac{\ell}{6}$.
- $\Theta < \ell$. Then if $\Theta = 0$, we simply have $f(\Theta) \geq \frac{\ell}{2}$. If $\Theta \geq 1$, we have $f(\Theta) = \frac{\ell}{4\Theta+2} + \frac{(\Theta-1)\Theta}{4\Theta+2} \geq \frac{\ell}{6\Theta} + \frac{\Theta-1}{6} \geq \sqrt{\ell}/6$ since the optimal Θ is $\Theta = \sqrt{\ell}$.

Therefore, for any choice of Θ, we must have $f(\Theta) \geq \sqrt{\ell}/6$, which concludes the proof by

$$\lim_{T\to\infty}\frac{\mathbb{E}\left[\mathcal{L}^\pi(T)\right]}{T} - \frac{\mathcal{L}^\star(w,T)}{T} = \lim_{T\to\infty}\frac{\mathbb{E}\left[\mathcal{L}^\pi(T)\right]}{T} \geq \min_{\Theta\in\mathbb{N}} f(\Theta) \geq \sqrt{\ell}/6.$$

□

B.4 Bounding idiosyncrasy loss by Lagrangian via drift analysis (Lemma 3.3)

Proof of Lemma 3.3. The first inequality is because

$$\begin{aligned}\mathbb{E}\left[L(T+1) - L(1)\right] &= \beta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t) - A_k(t))\right] + \frac{1}{2}\mathbb{E}\left[\sum_{k\in\mathcal{K}} Q_k^2(T+1)\right] \\ &\geq \beta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t) - A_k(t))\right]\end{aligned}$$

since $Q_k(T+1) \geq 0$ and $Q_k(1) = 0$.

For the second inequality, we bound the drift for a period t by

$$
\begin{aligned}
\mathbb{E}\left[L(t+1)-L(t)\right] &= \beta\mathbb{E}\left[\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t))\right] + \mathbb{E}\left[\sum_{k\in\mathcal{K}} \frac{1}{2}(Q_k(t+1)^2 - Q_k(t)^2)\right] \\
&= \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) + \mathbb{E}\left[-\beta\sum_{k\in\mathcal{K}} r_k\ell_k A_k(t) + \sum_{k\in\mathcal{K}}\frac{1}{2}((Q_k(t)+A_k(t)-S_k(t))^2 - Q_k^2(t))\right] \\
&= \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) + \mathbb{E}\left[-\beta\sum_{k\in\mathcal{K}} r_k\ell_k A_k(t) + \frac{1}{2}\sum_{k\in\mathcal{K}}(A_k(t)-S_k(t))^2 + \sum_{k\in\mathcal{K}} Q_k(t)(A_k(t)-S_k(t))\right] \\
&\le 1 + \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) - \mathbb{E}\left[\sum_{k\in\mathcal{K}} A_k(t)(\beta r_k\ell_k - Q_k(t)) + \sum_{k\in\mathcal{K}} Q_k(t)S_k(t)\right] \\
&= 1 + \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) - \mathbb{E}\left[\sum_{k\in\mathcal{K}} A_k(t)(\beta r_k\ell_k - Q_k(t)) + \sum_{k\in\mathcal{K}} Q_k(t)\psi_k(t)\mu_k N(t)\right] \\
&= 1 + \mathbb{E}\left[f_t(\boldsymbol{A}(t),\boldsymbol{\psi}(t),\boldsymbol{Q}(t))\right]
\end{aligned}
$$

where the second equality uses $\mathbb{E}\left[\Lambda_k(t)\right] = \lambda_k(t)$ and the queueing dynamic ($S_k(t) = 1$ if a type k post is reviewed); the first inequality uses the fact that $A_k(t) = 1$ for at most one type and $S_k(t) = 1$ for at most one type; the last equality uses that condition on $\psi_k(t) = 1$ (i.e., type k is chosen to review), $S_k(t) = 1$ with probability $\mu_k N(t)$. The result follows by telescoping over periods. □

B.5 Connecting Lagrangian with Primal Objective (Lemma 3.5)

Proof of Lemma 3.5. Expanding over the I windows of the benchmark, $f(\{\boldsymbol{a}^\star(t)\}_t, \{\boldsymbol{\nu}^\star(t)\}_t, \boldsymbol{u})$ is given by

$$
\beta\sum_{i=1}^{I}\sum_{t=\tau_i^\star}^{\tau_{i+1}^\star-1}\sum_{k\in\mathcal{K}} r_k\ell_k(\lambda_k(t)-a_k^\star(t)) - \sum_{i=1}^{I}\sum_{k\in\mathcal{K}} u_{i,k}\left(\mu_k\sum_{t=\tau_i^\star}^{\tau_{i+1}^\star-1}\nu_k^\star(t)N(t) - \sum_{t=\tau_i^\star}^{\tau_{i+1}^\star-1} a_k^\star(t)\right).
$$

As $\{\{\boldsymbol{a}^\star(t)\}_t, \{\boldsymbol{\nu}^\star(t)\}_t\}$ is a feasible solution to (w-fluid), for any interval $i = 1,\ldots,I$, it holds that $\mu_k\sum_{t=\tau_i^\star}^{\tau_{i+1}^\star-1}\nu_k^\star(t)N(t) - \sum_{t=\tau_i^\star}^{\tau_{i+1}^\star-1} a_k^\star(t) \ge 0$. Therefore, since $u_{i,k}\ge 0$, we have

$$
f(\{\boldsymbol{a}^\star(t)\}_t, \{\boldsymbol{\nu}^\star(t)\}_t, \boldsymbol{u}) \le \beta\sum_{i=1}^{I}\sum_{t=\tau_i^\star}^{\tau_{i+1}^\star-1}\sum_{k\in\mathcal{K}} r_k\ell_k(\lambda_k(t)-a_k^\star(t)) = \beta\mathcal{L}^\star(w,T).
$$

□

B.6 Difference in Lagrangian by Having Window-Based Duals (Lemma 3.6)

Proof of Lemma 3.6. First note that for any $t_1 \le t_2, k\in\mathcal{K}$, we have

$$
|Q_k(t_2)-Q_k(t_1)| \le \sum_{t=t_1}^{t_2-1}|Q_k(t+1)-Q_k(t)| = \sum_{t=t_1}^{t_2-1}|A_k(t)-S_k(t)| \le t_2 - t_1.
$$

As a result,

$$
\begin{aligned}
&|f_{t_2}(\boldsymbol{a}^\star(t_2), \boldsymbol{\nu}^\star(t_2), \boldsymbol{Q}(t_2)) - f_{t_2}(\boldsymbol{a}^\star(t_2), \boldsymbol{\nu}^\star(t_2), \boldsymbol{Q}(t_1))| \le \sum_{k\in\mathcal{K}} (a_k^\star(t_2) + \nu_k^\star(t_2)\mu_k N(t_2))|Q_k(t_2) - Q_k(t_1)| \\
&\quad \le (t_2 - t_1)\sum_{k\in\mathcal{K}} (a_k^\star(t_2) + \nu_k^\star(t_2)\mu_k N(t_2)) \overset{(a)}{\le} (t_2 - t_1)\left(\sum_{k\in\mathcal{K}} \lambda_k(t_2) + \sum_{k\in\mathcal{K}} \nu_k\right) \le 2(t_2 - t_1)
\end{aligned}
$$

where (a) is by $a_k^\star(t_2) \le \lambda_k(t_2)$ and the assumption $\mu_k N(t_2) \le 1$; the next inequality is by $\sum_k \nu_k^\star(t_2) \le 1$. Therefore,

$$
\begin{aligned}
\sum_{i=1}^{I}\sum_{t=\tau_i^\star}^{\tau_{i+1}^\star - 1} \left(\mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}(t))\right] - \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}(\tau_i^\star))\right]\right) &\le 2\sum_{i=1}^{I}\sum_{t=\tau_i^\star}^{\tau_{i+1}^\star - 1} (t - \tau_i^\star) \\
&= \sum_{i=1}^{I} (\tau_{i+1}^\star - \tau_i^\star)(\tau_{i+1}^\star - \tau_i^\star - 1) \le (w-1)T
\end{aligned}
$$

where the last inequality is because $\tau_{i+1}^\star - \tau_i^\star \le w$ and $\sum_{i=1}^{I} \tau_{i+1}^\star - \tau_i^\star = T$. $\square$

B.7 BACID optimizes per-period Lagrangian (Lemma 3.7)

Proof of Lemma 3.7. For a period t, the expectation only involves the uncertainty in the arrival and the review so

$$
\begin{aligned}
&\mathbb{E}\left[f_t(\boldsymbol{A}(t), \boldsymbol{\psi}(t), \boldsymbol{Q}(t)) \,|\, \boldsymbol{Q}(t) = \boldsymbol{q}\right] \\
&\quad = \beta\sum_{k\in\mathcal{K}} r_k \ell_k \lambda_k(t) - \left(\sum_{k\in\mathcal{K}} \lambda_k(t)\mathbb{1}\left(\beta r_k \ell_k \ge q_k\right)\left(\beta r_k \ell_k - q_k\right) + \sum_{k\in\mathcal{K}} \psi_k(t) q_k \mu_k N(t)\right) \\
&\quad \le \beta\sum_{k\in\mathcal{K}} r_k \ell_k \lambda_k(t) - \left(\sum_{k\in\mathcal{K}} a_k^\star(t)(\beta r_k \ell_k - q_k) + N(t)\sum_{k\in\mathcal{K}} \nu_k^\star(t) q_k \mu_k\right) = f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{q})
\end{aligned}
$$

where the inequality is because $0 \le a_k^\star(t) \le \lambda_k(t)$ and $\sum_{k\in\mathcal{K}} \nu_k \le 1$ and the definition that $\psi_k(t) = 1$ for the type k with the largest $q_k \mu_k$. $\square$

C Supplementary materials on OLBACID (Section 4)

C.1 Solution to the optimistic estimator (Footnote 4)

Lemma C.1. *For any $h_1 < h_2$ and $k \in \mathcal{K}$, the solution $h^\star$ of $\arg\max_{h\in[h_1,h_2]} r_k(h)$ is as follows: if $0 \in [h_1, h_2]$, then $h^\star = 0$; if $0 < h_1$, then $h^\star = h_1$; if $h_2 < 0$, then $h^\star = h_2$.*

Proof. By the definition of r_k^O and r_k^R in (5), we know for $h > 0$ that

$$
\begin{aligned}
r_k^R(h) = v_k(-h\Phi(-h/\sigma_k) + \sigma_k\varphi(-h/\sigma_k)) &= v_k(-h\Phi(-h/\sigma_k) + \sigma_k\varphi(h/\sigma_k)) \\
&\le v_k(h\Phi(h/\sigma_k) + \sigma_k\varphi(h/\sigma_k)) = r_k^O(h).
\end{aligned}
$$

Similarly, we have $r_k^O(h) \le r_k^R(h)$ for $h < 0$. Therefore, $r_k(h) = r_k^O(h)$ for $h < 0$, $r_k(h) = r_k^R(h)$ for $h > 0$, and $r_k(0) = r_k^O(0) = r_k^R(0)$. In addition, $r_k^R(h)$ is decreasing for $h \ge 0$ and $r_k^O(h)$ is increasing for $h \le 0$ by Fact 7. As a result, $r_k(0) \ge r_k(h)$ for any h, so $h^\star = 0$ if $0 \in [h_1, h_2]$. The remaining result follows from the monotonicity properties of r_k^R and r_k^O. $\square$

C.2 Optimism-only fails to learn classification decisions (Proposition 3)

We recall the setting: $K = 2$, $h_1 = 0, h_2 = -1, v_1 = \sigma_1 = v_2 = \sigma_2 = 1$ and $\ell_1 = 49\ell_2$. The algorithm is run with $H = 1$, with exact knowledge that $h_1 = 0$, and with no knowledge of h_2. The hyper-parameter β is in $(48/\ell_1, 1)$. The time horizon T is $\lceil \exp(\beta\ell_1/576) \rceil$. Arrival rates are such that $\lambda_1(t) = 1, \lambda_2(t) = 0$ for $t \le T_1 := \lfloor \frac{\beta\ell_1}{2} \rfloor$, and $\lambda_1(t) = 5/6, \lambda_2(t) = 1/6$ for $T_1 < t \le T$. The service rate is $\mu_1 = \mu_2 = 1/2$ with $N(t) = 1$ for any t. We know that $r_2^R - r_2^O = \Phi(1) + \varphi(1) - (\Phi(-1) + \varphi(-1)) = \Phi(1) - \Phi(-1) > 0$, meaning that wrongly removal of a type-2 post incurs an additional loss of $(\Phi(1) - \Phi(-1))\ell_2$. Let us define $Q_{\max} = \beta\ell_1$.

Lemma C.2. *For any period t, the number of type-2 posts admitted for review is $Q_2(t) < \frac{Q_{\max}}{24}$.*

Proof. We have $r_1 = \min(r_1^O, r_1^R) = 1, r_2 = \min(r_2^O, r_2^R) \le 1$ by (5). In addition, the admission rule is that $A_k(t) = 1$ if $\beta\bar{r}_k(t)\ell_k > Q_k(t)$. Since $\bar{r}_k(t) \le \sigma_k = 1$ by Fact 7, it holds that $A_2(t) \le \mathbb{1}(Q_2(t) \le \beta\ell_2)$, which, by the same analysis in the proof of Lemma 3.2, implies that $Q_2(t) \le \beta\ell_2 + 1 < \frac{\beta\ell_1}{24} = \frac{Q_{\max}}{24}$ because $\ell_1 \ge 49\ell_2$ and $\beta \ge \frac{48}{\ell_1}$. $\square$

We define the event $\mathcal{E}$ where for all period t in $[T_1, T]$, we have $Q_1(t) \ge \frac{Q_{\max}}{24}$, i.e., $\mathcal{E} = \{\forall t \in [T_1 + 1, T], Q_1(t) \ge Q_{\max}/24\}$. The following lemma shows event $\mathcal{E}$ happens with high probability.

Lemma C.3. *With high probability, there are at least $\frac{Q_{\max}}{24}$ type-1 admitted posts: $\mathbb{P}\{\mathcal{E}\} \ge 1-2/T$.*

Proof. Our proof strategy is as follows. We first show that by a concentration bound, the queue length of type-1 posts must increase linearly in the interval, given that the queue length is less than $Q_{\max}$ in the entire interval. In this case, for any period t, either there is a period τ with $Q_1(\tau) \ge Q_{\max}$ in the last $Q_{\max}/4$ periods, which implies $Q_1(t) \ge Q_{\max} - Q_{\max}/4$; or $Q_1(\tau)$ grows linearly in the last $Q_{\max}/4$ periods, which also implies $Q_1(t) = \Omega(Q_{\max})$.

Formally, we define $\hat{S}_k(t)$ as the Bernoulli random variable indicating whether the review of a type-k post will finish in this period, given that we schedule such a post. Then $\hat{S}_k(t)$ has mean $N(t)\mu_k = 1/2$, and the true review outcome is $S_k(t) = \hat{S}_k(t)\psi_k(t)$. We denote $E = \lceil Q_{\max}/4 \rceil$. For $t > T_1 > E$, we define

$$\tilde{\mathcal{E}}_t = \left\{ \sum_{\tau=t-E}^{t-1} \Lambda_1(\tau) \ge \frac{5E}{6} - \sqrt{E \ln T} \right\} \cap \left\{ \sum_{\tau=t-E}^{t-1} \hat{S}_1(t) \le \frac{E}{2} + \sqrt{E \ln T} \right\}.$$

By Hoeffding Inequality and union bound, we have $\mathbb{P}\{\tilde{\mathcal{E}}_t\} \ge 1 - 2/T^2$ since $\mathbb{E}[\Lambda_1(t)] \ge \frac{5}{6}$. Let us set $\tilde{\mathcal{E}} = \cap_{t>T_1}\{\tilde{\mathcal{E}}_t\}$. By union bound, we have $\mathbb{P}\{\tilde{\mathcal{E}}\} \ge 1 - 2/T$. It remains to show that under $\tilde{\mathcal{E}}$, we have $Q_1(t) \ge \frac{Q_{\max}}{24}$ for any $t > T_1$, and thus $\mathbb{P}\{\mathcal{E}\} \ge \mathbb{P}\{\tilde{\mathcal{E}}\} \ge 1 - 2/T$.

We fix a period $t > T_1$ and look at the interval $[t - E, t - 1]$. Since $Q_{\max} = \beta\ell_1 \ge 24$, we have $E < Q_{\max}/2$. We consider two cases:

- there exists $\tau \in [t - E, t - 1]$ such that $Q_1(\tau) \ge Q_{\max}$. In this case, we have

$$Q_1(t) \ge Q_1(\tau) - (t - \tau) \ge Q_1(\tau) - E \ge Q_{\max} - E \ge Q_{\max}/2$$

  where the first inequality is because at most one type-1 post is reviewed per period.

- for every $\tau \in [t-E, t-1]$, we have $Q_1(\tau) < Q_{\max}$. By the admission rule and the assumption that h_1 is known perfectly, we have $A_1(\tau) = \Lambda_1(\tau)\mathbb{1}\left(r_1\ell_1 > Q_1(\tau)\right) = \Lambda_1(\tau)$ for every $\tau \in [t-E, t-1]$. As a result, condition on $\tilde{\mathcal{E}}$,

$$\begin{aligned} Q_1(t) &\geq Q_1(t-E) + \sum_{\tau=t-E}^{t-1} A_1(\tau) - \sum_{\tau=t-E}^{t-1} \hat{S}_1(\tau) = \sum_{\tau=t-E}^{t-1} \Lambda_1(\tau) - \sum_{\tau=t-E}^{t-1} \hat{S}_1(\tau) \\ &\geq \frac{5E}{6} - \sqrt{E\ln T} - \frac{E}{2} - \sqrt{E\ln T} = \frac{E}{3} - 2\sqrt{E\ln T} \overset{(a)}{\geq} \frac{E}{6} \geq \frac{Q_{\max}}{24}, \end{aligned}$$

  where (a) is because $T \leq \exp(Q_{\max}/576) \leq \exp(E/144)$.

Combining the above discussions then shows conditioning on $\tilde{\mathcal{E}}$, we have $Q_1(t) \geq Q_{\max}/24$, which finishes the proof. □

Proof of Proposition 3. Let us condition on $\mathcal{E}$, which happens with probability at least $1 - 2/T$ by Lemma C.3. Since the algorithm follows the scheduling rule that the type with the larger $\mu_k Q_k(t)$ get reviewed in every period, it is guaranteed that only type-1 posts get reviewed after T_1 because $\mu_1 = \mu_2$ and $Q_1(t) \geq \frac{Q_{\max}}{24} > Q_2(t)$ where the last inequality is by Lemma C.2. But type-2 posts only arrive after T_1. As a result, there is no review for type-2 posts. □

C.3 Number of posts admitted to label-driven queue (Lemma 4.4)

Proof of Lemma 4.4. We define $S_k^{\mathrm{LD}}(t) \in \{0,1\}$ to be equal to one if and only if $S_k(t) = 1$ and $Q^{\mathrm{LD}}(t) = 1$, i.e., humans review a type-k post from the label-driven queue. This implies that:

$$Q^{\mathrm{LD}}(t+1) - Q^{\mathrm{LD}}(t) = \sum_{k\in\mathcal{K}} (E_k(t) - S_k^{\mathrm{LD}}(t)) = \sum_{k\in\mathcal{K}} (E_k(t) - S_k^{\mathrm{LD}}(t) Q^{\mathrm{LD}}(t)).$$

We then define an indicator $\mathcal{I}_{k,t}$ which is equal to 1 if $\mathcal{Q}^{\mathrm{LD}}(t)$ contains a type-k post. Then conditioning on $Q^{\mathrm{LD}}(t) = 1$, the expectation of $\sum_{k\in\mathcal{K}} S_k^{\mathrm{LD}}(t)$ is lower bounded by

$$\begin{aligned} \mathbb{E}\left[\sum_{k\in\mathcal{K}} S_k^{\mathrm{LD}}(t) \,\middle|\, Q^{\mathrm{LD}}(t) = 1\right] &= \sum_{k\in\mathcal{K}} \mathbb{P}\{\mathcal{I}_{k,t} = 1 | Q^{\mathrm{LD}}(t) = 1\} \mathbb{E}\left[S_k^{\mathrm{LD}}(t) \,\middle|\, Q^{\mathrm{LD}}(t) = 1, \mathcal{I}_{k,t} = 1\right] \\ &= \sum_{k\in\mathcal{K}} \mathbb{P}\{\mathcal{I}_{k,t} = 1 | Q^{\mathrm{LD}}(t) = 1\} N_k(t)\mu_k \geq \hat{\mu}_{\min}. \end{aligned} \tag{18}$$

The second equality is because OLBACID prioritizes posts in the label-driven queue (Line 8 in Algorithm 2); the last inequality uses $N_k(t)\mu_k \geq \hat{\mu}_{\min}$ and $\sum_{k\in\mathcal{K}} \mathbb{P}\{\mathcal{I}_{k,t} = 1 | Q^{\mathrm{LD}}(t) = 1\} = 1$. Therefore,

$$\mathbb{E}\left[Q^{\mathrm{LD}}(t+1) - Q^{\mathrm{LD}}(t)\right] = \mathbb{E}\left[\sum_{k\in\mathcal{K}} E_k(t)\right] - \mathbb{E}\left[Q^{\mathrm{LD}}(t) \sum_{k\in\mathcal{K}} S_k^{\mathrm{LD}}(t)\right] \geq \mathbb{E}\left[\sum_{k\in\mathcal{K}} E_k(t)\right] - \hat{\mu}_{\min}\mathbb{E}\left[Q^{\mathrm{LD}}(t)\right]$$

where we use (18) for the inequality. Telescoping for $t = 1, \ldots, T$, we obtain $\hat{\mu}_{\min} \sum_{t=1}^{T} \mathbb{E}\left[Q^{\mathrm{LD}}(t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{k\in\mathcal{K} E_k(t)}\right]$, which finishes the proof. □

C.4 Regret guarantee for OLBACID (Theorem 3)

Proof of Theorem 3. The average regret $\textsc{Reg}^{\text{OLBACID}}(w,T)$ is upper bounded by $r_{\max}\ell_{\max}$, which is less than the first term when $K \geq \ell_{\max}$. Moreover, if $T \leq 2$,

$$\textsc{Reg}^{\text{OLBACID}}(w,T) \leq \mathcal{L}^{\text{OLBACID}}(T) \leq 2r_{\max}\ell_{\max} \leq 2r_{\max}\frac{K\ell_{\max}\sqrt{\ln(T+2)}}{\sqrt{T}} \lesssim \frac{K\ell_{\max}\sqrt{\ln(T)}}{\sqrt{T}}.$$

We thus focus on $T \geq 3$, $K \leq \ell_{\max}$ and $\beta = \gamma = 1/\sqrt{K\ell_{\max}} \geq 1/\ell_{\max}$, which is the setting Lemmas 4.1, 4.2 and 4.3 are applicable. Omitting dependence on $v_{\max}, r_{\max}, \sigma_{\max}, \hat{\mu}_{\min}, \ln T$, constants and applying Lemmas 4.1, 4.2 and 4.3 to the corresponding terms in (12), we obtain:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}^{\text{OLBACID}}(T)\right] - \mathcal{L}^{\star}(w,T) &\lesssim \left(\sqrt{\ell_{\max}}T + K\ell_{\max} + \frac{K\ell_{\max}}{\max(\eta,\gamma)^2}\right) + \left(\sqrt{K\ell_{\max}}T + \frac{K}{\max(\eta,\gamma)^2}\right) \\ &\quad + \left(w\sqrt{K\ell_{\max}}T + \frac{K\ell_{\max}}{\max(\eta,\gamma)^2} + K\ell_{\max}\sqrt{T} + K\ell_{\max}^{1.5}\right) \\ &\lesssim w\sqrt{K\ell_{\max}}T + K\ell_{\max}\sqrt{T} + \sqrt{K}\ell_{\max}^{1.5} + K\ell_{\max}\min\left(\frac{1}{\eta^2}, K\ell_{\max}\right).\end{aligned}$$

Dividing by T, $\frac{\mathbb{E}\left[\mathcal{L}^{\text{OLBACID}} - \mathcal{L}^{\star}(w,T)\right]}{T} \lesssim w\sqrt{K\ell_{\max}} + K\ell_{\max}\sqrt{\frac{1}{T}} + \frac{\sqrt{K}\ell_{\max}^{1.5}}{T} + \min\left(\frac{K\ell_{\max}}{\eta^2 T}, \frac{K^2\ell_{\max}^2}{T}\right)$. □

C.5 Confidence intervals are valid with high probability (Lemma 4.5)

Proof of Lemma 4.5. Fix $k \in \mathcal{K}, t \in [T]$. The number $n_k(t)$ of type-k posts in the dataset $\mathcal{D}_t$ is upper bounded by t. Conditioning on $n_k(t) = n \leq t$, there are n posts whose costs are i.i.d. normal with mean h_k and variance σ_k^2. By Chernoff Bound (Fact 2), the empirical average $\hat{h}_k(t)$ satisfies

$$\mathbb{P}\left\{|\hat{h}_k(t) - h_k| > \sigma_k\sqrt{\frac{8\ln t}{n}}\right\} \leq 2\exp\left(\frac{-8\sigma_k^2\ln t \cdot n}{2\sigma_k^2 \cdot n}\right) \leq \frac{2}{t^4}.$$

By union bound over $n = 0, \ldots, t-1$, this bound is simultaneously valid for all values of $n_k(t)$, and thus $\mathbb{P}\left\{|\hat{h}_k(t) - h_k| > \sigma_k\sqrt{\frac{8\ln t}{n_k(t)}}\right\} \leq 2t^{-3}$. We finish the proof by using the definition of the confidence interval in (11) and the assumption that $h_k \in [-H, H]$. □

C.6 Bounding number of admitted posts to label-driven queue (Lemma 4.6)

Proof of Lemma 4.6. We decompose the number of admitted posts $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} E_k(t)\right]$ by

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} E_k(t)(\mathbb{1}\left(\mathcal{E}_{k,t} + \mathcal{E}_{k,t}^c\right))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} E_k(t)\mathbb{1}\left(\mathcal{E}_{k,t}\right)\right] + \sum_{t=1}^{T}\frac{2}{t^3} \leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} E_k(t)\mathbb{1}\left(\mathcal{E}_{k,t}\right)\right] + 4$$

where the first inequality uses Lemma 4.5 and the fact that $\sum_{k\in\mathcal{K}} E_k(t) \leq 1$, and the second inequality uses $\sum_{t=1}^{\infty} 1/t^3 \leq 2$. It remains to bound the first sum.

Fixing a type k, we consider the last time period T_k such that $E_k(T_k)\mathbb{1}\left(\mathcal{E}_{k,T_k}\right) = 1$. By Line 6 in Algorithm 2, $E_k(T_k) = 1$ if and only if $\underline{h}_k(T_k) < -\gamma < \gamma < \bar{h}_k(T_k)$ and $Q^{\text{LD}}(T_k) = 0$. In addition, the setting of the confidence interval, (11), shows that $\bar{h}_k(T_k) - \underline{h}_k(T_k) \le 2\sigma_k\sqrt{\frac{8\ln T_k}{n_k(T_k)}}$ where $n_k(T_k)$ is the number of type-k posts in the dataset $\mathcal{D}(T_k)$. We next bound $n_k(T_k)$ for the case of $h_k \ge 0$. Since $\mathbb{1}(\mathcal{E}_{k,T_k}) = 1$, we have $h_k \in [\underline{h}_k(T_k), \bar{h}_k(T_k)]$. by definition,

$$\underline{h}_k(T_k) = \bar{h}_k(T_k) - (\bar{h}_k(T_k) - \underline{h}_k(T_k)) \ge h_k - 2\sigma_k\sqrt{\frac{8\ln T_k}{n_k(T_k)}} \ge \eta - 2\sigma_k\sqrt{\frac{8\ln T_k}{n_k(T_k)}}$$

where the last inequality is by the margin definition that $\eta \le |h_k| = h_k$. Since $\underline{h}_k(T_k) < -\gamma$, we have $-\gamma \ge \eta - 2\sigma_k\sqrt{\frac{8\ln T_k}{n_k(T_k)}}$. As a result, $n_k(T_k) \le \frac{32\sigma_{\max}^2 \ln T}{\max(\eta,\gamma)^2}$. The same analysis applies for $h_k < 0$.

Recall that we also have $Q^{\mathrm{LD}}(t) = 0$, i.e., all previous type-k posts added to the label-driven queue have received human reviews. Therefore, we have $n_k(T_k) \ge \sum_{t=1}^{T_k-1} E_k(t)$ and, since $\eta, \gamma \le 1$,

$$\sum_{t=1}^{T} E_k(t) = \sum_{t=1}^{T_k-1} E_k(t) + 1 \le n_k(T_k) + 1 \le \frac{32\sigma_{\max}^2 \ln T}{\max(\eta,\gamma)^2} + 1 \le \frac{33\sigma_{\max}^2 \ln T}{\max(\eta,\gamma)^2}.$$

Summing across all types gives $\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} E_k(t) \le \frac{33K\sigma_{\max}^2 \ln T}{\max(\eta,\gamma)^2}$. □

C.7 Lipschitz properties of functions r^O, r^R, r (Lemma 4.8)

Proof of Lemma 4.8. We fix the type k and recall the definition of r_k^O and r_k^R in (5). We first show the Lipschitz property for r_k^O. By definition (5), we have

$$\left|r_k^O(h_1) - r_k^O(h_2)\right| \le v_k|h_1\Phi(h_1/\sigma_k) - h_2\Phi(h_2/\sigma_k)| + v_k\sigma_k|\varphi(h_1/\sigma_k) - \varphi(h_2/\sigma_k)|. \tag{19}$$

Let $f(x) = x\Phi(x/\sigma_k)$. Then $f'(x) = \Phi(x/\sigma_k) + (x/\sigma_k)\varphi(x/\sigma_k) = \Phi(x/\sigma_k) - \varphi'(x/\sigma_k)$ where we use the fact that the derivative of $\varphi(y)$, $\varphi'(y)$, is given by $-y\varphi(y)$. By Fact 6, we have $|\varphi'(y)| \le 1/4$ for any y. As a result, $|f'(x)| \le |\Phi(x/\sigma_k)| + |\varphi'(x/\sigma_k)| \le 1 + 1/4 = 5/4$ and thus $f(x)$ is 5/4-Lipschitz continuous. This shows that $v_k|h_1\Phi(h_1/\sigma_k) - h_2\Phi(h_2/\sigma_k)| \le \frac{5}{4}v_k|h_2 - h_1|$.

In addition, since $|\varphi'(y)| \le 1/4$, we know $\varphi(y)$ is 1/4-Lipschitz continuous, which shows $v_k\sigma_k|\varphi(h_1/\sigma_k) - \varphi(h_2/\sigma_k)| \le v_k\sigma_k/4|h_1 - h_2|/\sigma_k = \frac{1}{4}v_k|h_1 - h_2|$. As a result of (19), we have $|r_k^O(h_1) - r_k^O(h_2)| \le \frac{5}{4}v_k|h_1 - h_2|$ and thus $r_k^O(\cdot)$ is $\frac{3}{2}v_k$-Lipschitz continuous. Since $r_k^R(h) = r_k^O(-h)$, we also show the Lipschitz property for r_k^R.

We next show the property for $r_k(\cdot)$. Let $h_1 \le h_2 \in [-H, H]$. We need to bound $|r_k(h_1) - r_k(h_2)|$. There are three cases:

- $h_1 \le h_2 \le 0$. Then $r_k(h_1) = r_k^O(h_1), r_k(h_2) = r_k^O(h_2)$. Since r_k^O is $(5/4)v_k$-Lipschitz, we have $|r_k(h_1) - r_k(h_2)| \le (3/2)v_k|h_1 - h_2|$;
- $0 \le h_1 \le h_2$, same as in the last case but now we have $r_k(h_1) = r_k^R(h_1), r_k(h_2) = r_k^R(h_2)$;
- $h_1 < 0 < h_2$. We have $r_k(h_1) = r_k^O(h_1)$ and $r_k(h_2) = r_k^R(h_2)$, so $|r_k(h_1) - r_k(h_2)| = |r_k^O(h_1) - r_k^R(h_2)|$. In addition,

$$\begin{aligned}|r_k^O(h_1) - r_k^R(h_2)| &= v_k\left|h_1\Phi(h_1/\sigma_k) + \sigma_k\varphi(h_1/\sigma_k) + h_2\Phi(-h_2/\sigma_k) - \sigma_k\varphi(-h_2/\sigma_k)\right| \\ &\le v_k\left(\sigma_k|\varphi(h_1/\sigma_k) - \varphi(-h_2/\sigma_k)| + |h_1\Phi(h_1/\sigma_k) + h_2\Phi(-h_2/\sigma_k)|\right).\end{aligned}$$

We next bound the two terms in the parenthesis. For the first term, since $\varphi(y)$ is $1/4-$Lipschitz, we have $\sigma_k|\varphi(h_1/\sigma_k) - \varphi(-h_2/\sigma_k)| \leq \frac{1}{4}|h_1 + h_2| \leq (h_2 - h_1)/4$ since $h_1 < 0 < h_2$. For the second term, recall that $f(x) = x\Phi(x/\sigma_k)$ and $f(x)$ is 5/4-Lipschitz. We have $|h_1\Phi(h_1/\sigma_k) + h_2\Phi(-h_2/\sigma_k)| = f(h_1) - f(-h_2) \leq (5/4)|h_1 + h_2| \leq (5/4)(h_2 - h_1)$. Summing the two terms give $|r_k(h_1) - r_k(h_2)| = |r_k^O(h_1) - r_k^R(h_2)| \leq (3/2)v_k(h_2 - h_1) = (3/2)v_k|h_1 - h_2|$.

Summing the three cases shows that $r_k(\cdot)$ is $\frac{3}{2}v_k$-Lipschitz continuous, completing the proof. □

C.8 Bounding per-period classification loss (Lemma 4.9)

Proof of Lemma 4.9). We fix a period t and a type k with $\Lambda_k(t) = 1$. If $\Lambda_k(t) = 0$, this directly implies that $Z_k(t) = 0$. We focus on bounding $Z_k(t)$ condition on $\mathcal{E}_{k,t}$, which implies $h_k \in [\underline{h}_k(t), \bar{h}_k(t)]$. We consider three cases:

- $\underline{h}_k(t) \geq -\gamma$: We first suppose $\hat{h}_k(t) > 0$; we set $Y_k(t) = 0$ and $Z_k(t) \leq (r_k^R - r_k^O)^+\ell_k$. If $h_k \geq 0$, we have $Z_k(t) = 0$ since $r_k^R \leq r_k^O$. If not, we know $-\gamma \leq \underline{h}_k(t) \leq h_k \leq 0$ and hence

$$\begin{aligned}\frac{Z_k(t)}{\ell_k} \leq |r_k^R - r_k^O| &= |r_k^R(h_k) - r_k^O(h_k)| \\ &\leq |r_k^R(0) - r_k^O(0)| + |r_k^R(h_k) - r_k^R(0)| + |r_k^O(h_k) - r_k^O(0)| \\ &\leq |r_k^R(h_k) - r_k^R(0)| + |r_k^O(h_k) - r_k^O(0)| \leq 3v_k|h_k| \leq 3v_k\gamma, \end{aligned} \tag{20}$$

  where the second inequality uses $r_k^R(0) = r_k^O(0)$; the third inequality uses the Lipschitz property in Lemma 4.8; and the last inequality uses $h_k \in [-\gamma, 0]$.

  Alternatively, if $\hat{h}_k(t) \leq 0$, we set $Y_k(t) = 1$. By definition of the confidence interval (11), we have $\bar{h}_k(t) \leq \hat{h}_k(t) + (\hat{h}_k(t) - \underline{h}_k(t)) \leq \gamma$ because $\underline{h}_k(t) \geq -\gamma \geq -H$ (recall that we assume $H \geq 1 \geq \gamma$). Since the confidence bound is valid conditioning on $\mathcal{E}_{k,t}$, we have $h_k \in [-\gamma, \gamma]$. As a result, $Z_k(t)/\ell_k \leq |r_k^R - r_k^O| \leq 3v_k|h_k| \leq 3v_k\gamma$ as in (20).

- $\bar{h}_k(t) \leq \gamma$; This case is symmetric to the above case. Following the same analysis gives $Z_k(t) \leq 3v_k\gamma\ell_k$ as well.

- $\underline{h}_k(t) \leq -\gamma \leq \gamma \leq \bar{h}_k(t)$. In this case, our label-driven admission wishes to send the post to the review post. If the label-driven queue is empty, it does so by setting $E_k(t) = 1$ and thus $Z_k(t) = 0$. Otherwise, the algorithm sets $Y_k(t) = 0$ and $E_k(t) = 0$. Therefore, we have for this case $Z_k(t) \leq r_{\max}\ell_k Q^{\mathrm{LD}}(t)$.

Summarizing the above three cases finishes the proof. □

C.9 Bounding idiosyncrasy loss by Lagrangian via drift analysis (Lemma 4.10)

The proof is similar to that of Lemma 3.3, and we include it for completeness.

Proof of Lemma 4.10. By definition, $L(1) = 0$ and $L(T+1) \geq \beta \sum_{t=1}^{T} \sum_{k \in \mathcal{K}} r_k\ell_k(\Lambda_k(t) - A_k(t) - E_k(t))$. That is, the Idiosyncrasy Loss of OLBACID (as given in (12)), is upper bounded by $\frac{1}{\beta}\mathbb{E}\left[L(T+1) - L(1)\right]$. We thus prove the first inequality in the lemma.

To show the second inequality, we have

$$\begin{aligned}
\mathbb{E}\left[L(t+1)-L(t)\right] &= \beta\mathbb{E}\left[\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t)-E_k(t))\right]+\mathbb{E}\left[\sum_{k\in\mathcal{K}}\frac{1}{2}(Q_k(t+1)^2-Q_k(t)^2)\right]\\
&= \mathbb{E}\left[\beta\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t)-E_k(t))+\frac{1}{2}\sum_{k\in\mathcal{K}}(A_k(t)-S_k(t))^2+\sum_{k\in\mathcal{K}}Q_k(t)(A_k(t)-S_k(t))\right]\\
&\le 1+\beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t)+\mathbb{E}\left[-\beta\sum_{k\in\mathcal{K}} r_k\ell_k(A_k(t)+E_k(t))+\sum_{k\in\mathcal{K}}Q_k(t)(A_k(t)-S_k(t))\right]\\
&\le 1+\beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t)+\mathbb{E}\left[-\beta\sum_{k\in\mathcal{K}} r_k\ell_k\bar{A}_k(t)+\sum_{k\in\mathcal{K}}Q_k(t)(\bar{A}_k(t)-S_k(t))\right],
\end{aligned}$$

where the last inequality is because $A_k(t)\le\bar{A}_k(t)\le A_k(t)+E_k(t)$. We know that, conditioned on $\psi_k(t)=1$, the expectation of $S_k(t)$ is $\mu_k N(t)$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[L(t+1)-L(t)\right] &\le 1+\beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t)-\mathbb{E}\left[\sum_{k\in\mathcal{K}}\bar{A}_k(t)\left(\beta r_k\ell_k-Q_k(t)\right)+\sum_{k\in\mathcal{K}}Q_k(t)\psi_k(t)\mu_k N(t)\right]\\
&= 1+\mathbb{E}\left[f_t(\bar{\boldsymbol{A}}\boldsymbol{(t)},\boldsymbol{\psi(t)},\boldsymbol{Q}(t))\right].
\end{aligned}$$

Telescoping across $t=1,\dots,T$, we get

$$\mathbb{E}\left[L(T+1)-L(1)\right]\le T+\sum_{t=1}^{T}\mathbb{E}\left[f_t(\bar{\boldsymbol{A}}\boldsymbol{(t)},\boldsymbol{\psi(t)},\boldsymbol{Q}(t))\right].$$

□

C.10 Bounding the Regret in Admission for OLBACID (Lemma 4.11)

Our proof structure is similar to that of bandits with knapsacks when restricting to the admission component [AD19]. A difference in the settings is that works on bandits with knapsacks also need to learn the uncertain resource consumption whereas the consumption is known in our setting. However, our setting presents an additional challenge: post labels are delayed by an endogenous queueing effect. In particular, a key step in the proof is Lemma C.4, which bounds the estimation error for all admitted posts where the estimation error is proxied by $1/\sqrt{n_k(t)}$ and $n_k(t)$ is the number of reviewed type-k posts, and is a finite-type version of [AD16, Lemma 3] (we later prove a similar result in the contextual setting; see Lemma 5.13). To prove Lemma C.4, we need to address the feedback delay, which adds the first term in our bound. The key insight is that, whenever we admit a new post of one type, the number of labels from that type we have not received is exactly the number of waiting posts of that type. By the optimistic admission rule (Line 7 of Algorithm 2), this number is upper bounded by $\beta r_{\max}\ell_{\max}$, which enables us to bound the effect of the delay.

Lemma C.4. *We have* $\sum_{t=1}^{T}\bar{A}_k(t)\min\left(r_{\max},3v_k\sigma_k\sqrt{\frac{8\ln t}{n_k(t)}}\right)\le 4\beta r_{\max}^2\ell_{\max}+6v_k\sigma_k\sqrt{8T\ln T}$.

Proof. Recall that $A_k(t)$ captures whether the algorithm admits a type-k post into the review queue and $E_k(t)$ captures whether it admits a type-k post into the label-driven queue. By Line 7 of Algorithm 2, we know that $A_k(t)=\bar{A}_k(t)(1-E_k(t))$ and thus $\bar{A}_k(t)\le A_k(t)+E_k(t)$. We thus aim to upper bound $\sum_{t=1}^{T}(A_k(t)+E_k(t))\min\left(r_{\max},3v_k\sigma_k\sqrt{\frac{8\ln t}{n_k(t)}}\right)$. Denoting the sequence of periods with $A_k(t)+E_k(t)=1$ as $t_1,\dots,t_Y$ where $t_Y=\sum_{t=1}^{T}A_k(t)+E_k(t)$ and setting $n'(y)=n_k(t_y)$, we then need to upper bound $\sum_{y=1}^{Y}\min\left(r_{\max},3v_k\sigma_k\sqrt{\frac{8\ln T}{n'(y)}}\right)$.

We next connect y and $n'(y)$. We fix a period t. Since $n_k(t)$ is the number of type-k posts in the dataset $\mathcal{D}(t)$, we have $n_k(t)$ equal to the number of periods that a type-k post is successfully served. In addition, the number of type-k posts in the label-driven $\mathcal{Q}^{\mathrm{LD}}(t)$ queue and the review queue $\mathcal{Q}(t)$ is given by the number of admitted type-k posts deducted by the number of reviewed type-k posts. As a result, $Q^{\mathrm{LD}}(t)+Q_k(t)\geq\sum_{\tau=1}^{t-1}(A_k(\tau)+E_k(\tau))-n_k(t)$. Since $Q^{\mathrm{LD}}(t)\leq 1$ and $Q_k(t)\leq 2\beta r_{\max}\ell_{\max}$ by Lemma 4.7, we have $Q^{\mathrm{LD}}(t)+Q_k(t)\leq 3\beta r_{\max}\ell_{\max}$ and thus $n_k(t)\geq\sum_{\tau=1}^{t-1}(A_k(\tau)+E_k(\tau))-3\beta r_{\max}\ell_{\max}$. For any y, we consider $t=t_y$ in the above analysis. Then we know that $n'(y)\geq y-1-3\beta r_{\max}\ell_{\max}\geq y-4\beta r_{\max}\ell_{\max}$. Therefore,

$$
\begin{aligned}
\sum_{t=1}^{T}\bar{A}_k(t)\min\left(r_{\max},3v_k\sigma_k\sqrt{\frac{8\ln t}{n_k(t)}}\right) &\leq\sum_{y=1}^{Y}\min\left(r_{\max},3v_k\sigma_k\sqrt{\frac{8\ln T}{n'(y)}}\right)\\
&\leq\sum_{y=1}^{Y}\min\left(r_{\max},3v_k\sigma_k\sqrt{\frac{8\ln T}{\max(0,y-4\beta r_{\max}\ell_{\max})}}\right)\\
\overset{u=y-4\beta r_{\max}\ell_{\max}}{\leq} 4\beta r_{\max}^2\ell_{\max}+\sum_{u=1}^{T}3v_k\sigma_k\sqrt{\frac{8\ln T}{u}} &\leq 4\beta r_{\max}^2\ell_{\max}+6v_k\sigma_k\sqrt{8T\ln T}.
\end{aligned}
$$

where we use the fact that $\sum_{u=1}^{T}\sqrt{1/u}\leq 1+\int_1^T x^{-1/2}\mathrm{d}x\leq 2\sqrt{T}$. $\square$

We now upper bound the regret in admission.

Proof of Lemma 4.11. Recall the good event $\mathcal{E}_{k,t}$ where $h_k\in[\underline{h}_k(t),\bar{h}_k(t)]$. We first decompose RegA by

$$
\begin{aligned}
\mathrm{RegA}(T)&=\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(A_k^{\mathrm{BACID}}(t)-\bar{A}_k(t))(\beta r_k\ell_k-Q_k(t))\left(\mathbb{1}\left(\mathcal{E}_{k,t}\right)+\mathbb{1}\left(\mathcal{E}_{k,t}^c\right)\right)\right]\\
&\leq\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(A_k^{\mathrm{BACID}}(t)-\bar{A}_k(t))(\beta r_k\ell_k-Q_k(t))\mathbb{1}\left(\mathcal{E}_{k,t}\right)\right]+\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(\beta r_k\ell_k+t)\mathbb{P}\{\mathcal{E}_{k,t}^c\}\\
&\leq\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(A_k^{\mathrm{BACID}}(t)-\bar{A}_k(t))(\beta r_k\ell_k-Q_k(t))\mathbb{1}\left(\mathcal{E}_{k,t}\right)\right]+4\beta K r_{\max}\ell_{\max}+4K \qquad (21)
\end{aligned}
$$

where the first inequality uses $Q_k(t)\leq t$ and the second inequality uses Lemma 4.5 that $\mathbb{P}\{\mathcal{E}_{k,t}\}\geq 1-2t^{-3}$ and that $\sum_{t=1}^{T}t^{-2}\leq 2$.

We fix a type k and upper bound $\sum_{t=1}^{T}(A_k^{\mathrm{BACID}}(t)-\bar{A}_k(t))(\beta r_k\ell_k-Q_k(t))\mathbb{1}\left(\mathcal{E}_{k,t}\right)$. Fixing a period t and assuming $\mathcal{E}_{k,t}$ holds, we have $h_k\in[\underline{h}_k(t),\bar{h}_k(t)]$ and thus $\bar{r}_k(t)=\max_{h\in[\underline{h}_k(t),\bar{h}_k(t)]}r_k(h)\geq r_k(h_k)=r_k$. As a result,

$$
\bar{r}_k(t)-r_k=\max_{h\in[\underline{h}_k(t),\bar{h}_k(t)]}r_k(h)-r_k(h_k)\leq\frac{3}{2}v_k\left(\bar{h}_k(t)-\underline{h}_k(t)\right)\leq 3v_k\sigma_k\sqrt{\frac{8\ln t}{n_k(t)}} \qquad (22)
$$

where the first inequality uses the Lipschitz property of $r_k(\cdot)$ in Lemma 4.8 and the fact that $h_k \in [\underline{h}_k(t), \bar{h}_k(t)]$; the second inequality uses (11). In addition, if $\Lambda_k(t) = 0$, we have $\bar{A}_k(t) = A_k^{\text{BACID}}(t) = 0$; otherwise,

$$\bar{A}_k(t) = \mathbb{1}\left(\beta \bar{r}_k(t)\ell_k \geq Q_k(t)\right) \geq \mathbb{1}\left(\beta r_k \ell_k \geq Q_k(t)\right) \geq A_k^{\text{BACID}}(t).$$

Therefore, $(A_k^{\text{BACID}}(t) - \bar{A}_k(t))(\beta r_k \ell_k - Q_k(t))$ is positive if and only if $A_k^{\text{BACID}}(t) = 0$ and $\bar{A}_k(t) = 1$. In this case,

$$\begin{aligned}(A_k^{\text{BACID}}(t) - \bar{A}_k(t))(\beta r_k \ell_k - Q_k(t)) &= Q_k(t) - \beta r_k \ell_k = Q_k(t) - \beta \bar{r}_k(t)\ell_k + \beta(\bar{r}_k(t) - r_k)\ell_k \\ &\leq \beta(\bar{r}_k(t) - r_k)\ell_k \leq \beta \min\left(r_{\max}, 3v_k\sigma_k\sqrt{\frac{8\ln t}{n_k(t)}}\right)\ell_k\end{aligned}$$

where the last inequality is because $\bar{r}_k(t) \leq r_{\max}$ by $\underline{h}_k(t), \bar{h}_k(t) \in [-H, H]$ and (22). Therefore,

$$\begin{aligned}\sum_{t=1}^{T}(A_k^{\text{BACID}}(t) - \bar{A}_k(t))(\beta r_k \ell_k - Q_k(t))\mathbb{1}\left(\mathcal{E}_{k,t}\right) &\leq \beta\ell_k \sum_{t=1}^{T} \bar{A}_k(t) \min\left(r_{\max}, 3v_k\sigma_k\sqrt{\frac{8\ln t}{n_k(t)}}\right) \\ &\leq 4\beta^2 r_{\max}^2 \ell_{\max}^2 + 6\beta v_k \sigma_{\max}\ell_{\max}\sqrt{8T\ln T} \qquad \text{by Lemma C.4.}\end{aligned}$$

Using (21) gives

$$\begin{aligned}\text{RegA}(T) &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(A_k^{\text{BACID}}(t) - \bar{A}_k(t))(\beta r_k \ell_k - Q_k(t))\mathbb{1}\left(\mathcal{E}_{k,t}\right)\right] + 4\beta K r_{\max}\ell_{\max} + 4K \\ &\leq 4K\beta^2 r_{\max}^2 \ell_{\max}^2 + 6K\beta v_{\max}\sigma_{\max}\ell_{\max}\sqrt{8T\ln T} + 4\beta K r_{\max}\ell_{\max} + 4K \\ &\leq 12K\beta^2 r_{\max}^2 \ell_{\max}^2 + 6K\beta v_{\max}\sigma_{\max}\ell_{\max}\sqrt{8T\ln T} \\ &\leq 12K\beta^2 r_{\max}^2 \ell_{\max}^2 + 6K\beta r_{\max}\ell_{\max}\sqrt{8T\ln T}\end{aligned}$$

where we use the assumption that $K \leq \ell_{\max}$, $\beta \geq 1/\ell_{\max}$ and $r_{\max} = v_{\max}(H + \sigma_{\max})$. □

C.11 Bounding the Regret in Scheduling for OLBACID (Lemma 4.12)

Proof of Lemma 4.12. By the scheduling rule in OLBACID, if $Q^{\text{LD}}(t) = 0$ for a period t, we have $\psi_k(t) = \mathbb{1}\left(k = \arg\max Q_k(t)\mu_k\right) = \psi_k^{\text{BACID}}(t)$. If $Q^{\text{LD}}(t) = 1$, then $\psi_k(t) = 0$ for any k. Therefore,

$$\begin{aligned}\text{RegS}(T) &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}\psi_k^{\text{BACID}}(T)Q_k(t)\mu_k N(t)Q^{\text{LD}}(t)\right] \\ &\leq 2\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}\psi_k^{\text{BACID}}(T)\beta r_{\max}\ell_{\max}Q^{\text{LD}}(t)\right] \\ &\leq 2\beta r_{\max}\ell_{\max}\mathbb{E}\left[\sum_{t=1}^{T}Q^{\text{LD}}(t)\right] \leq \frac{66\beta r_{\max}\ell_{\max}K\sigma_{\max}^2\ln T}{\max(\eta,\gamma)^2\hat{\mu}_{\min}}\end{aligned}$$

where the first inequality is by Lemma 4.7 and the assumption that $N(t)\mu_k \leq 1$; the second inequality is by $\sum_{k\in\mathcal{K}}\psi_k^{\text{BACID}}(t) \leq 1$; the last inequality is by Lemma 4.6. □

D Supplementary materials on COLBACID (Section 5)

D.1 Comparison with Literature of Linear Contextual Bandits

Contrasting to linear contextual bandit, e.g. [APS11], and linear contextual bandit with delays, such as [VCP17, BXZ23], there are three main distinctions.

The first distinction is that those works use a different estimator $\hat{\boldsymbol{\theta}}$ by including all arrived data pointss into the dataset (unobserved labels are set as 0). In contract, our confidence set is constructed based on all *reviewed* posts in (13), which is a subset of all *arrived* posts *endogenously* determined by our algorithm.

The second distinction, as a result of the first, is that conditioned on the set of arrived posts, the matrix $\bar{V}_t$ in (13) is a fixed matrix in their settings, but is a random matrix in our setting because it is constructed from the set of reviewed posts. We thus must deal with intricacy of the randomness in the regression.

The third distinction lies in the analysis. The analysis of prior work crucially relies on an independence assumption, i.e., the event of observing feedback for a post within a particular delay period is independent from other posts, which enables a concentration bound on the number of observed feedback. In our setting, this is no longer the case as the delay in one post implies that other posts wait in the queue and thus delays could correlate with each other. We thus resort to the more intuitive estimator and analyze it via properties of our queueing systems.

We note that [BXZ23] consider a setting where delays are correlated because they are generated from a Markov chain. That said, the proof requires concentration and stationary properties of the Markov chain, which are not available in our non-Markovian setting.

D.2 Regret guarantee of COLBACID (Theorem 4)

Proof of Theorem 4. By definition of $B_\delta(T)$ in (14), we have $B_\delta(T) \lesssim \sqrt{d}$. Applying Lemmas 5.1, 5.2 and 5.3 to (16), we have the loss of COLBACID upper bounded by

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{L}^{\text{COLBACID}}(T)\right] = \textsc{Classification Loss} + \textsc{Idiosyncrasy Loss} + \textsc{Relaxed Delay Loss} \\
&\lesssim \mathcal{L}^\star(w,T) + (G\beta+\gamma)\ell_{\max}T + \frac{\ell_{\max}d^2}{\max(\eta,\gamma)^2} + \frac{wT}{\beta} \\
&\qquad + \ell_{\max}\left(\Delta(\mathcal{K}_\mathcal{G})T + \beta\ell_{\max}Gd^{1.5} + d\sqrt{GT} + \frac{d^{2.5}}{\max(\eta,\gamma)^2}\right) \\
&\lesssim \mathcal{L}^\star(w,T) + w\sqrt{G\ell_{\max}}T + \ell_{\max}\Delta(\mathcal{K}_\mathcal{G})T + \ell_{\max}d\sqrt{GT} + \sqrt{G}\ell_{\max}^{1.5}d^{1.5} + d^{2.5}\min\left(\frac{\ell_{\max}}{\eta^2}, G\ell_{\max}^2\right)
\end{aligned}
$$

where we use the setting that $\beta = \gamma = 1/\sqrt{G\ell_{\max}}$. We finish the proof by

$$
\begin{aligned}
\textsc{Reg}^{\text{COLBACID}}(w,T) &= \frac{\mathcal{L}^{\text{COLBACID}}(T) - \mathcal{L}^\star(w,T)}{T} \\
&\lesssim w\sqrt{G\ell_{\max}} + \ell_{\max}\left(\Delta(\mathcal{K}_\mathcal{G}) + d\sqrt{\frac{G}{T}} + \frac{\sqrt{G}\ell_{\max}^{0.5}d^{1.5}}{T} + \frac{d^{2.5}}{T}\min\left(\frac{1}{\eta^2}, G\ell_{\max}\right)\right).
\end{aligned}
$$

$\square$

D.3 Confidence Sets are Valid with High Probability (Lemma 5.4)

Proof of Lemma 5.4. The proof uses [APS11, Theorem 2], which we restate in Fact 3. In particular, we recall that $Z_t = c_{\mathcal{S}(t)} = \boldsymbol{\phi}_{k(\mathcal{S}(t))}^\intercal \boldsymbol{\theta}^\star + \xi_{\mathcal{S}(t)} = \boldsymbol{X}_t \boldsymbol{\theta}^\star + \xi_{\mathcal{S}(t)}$. To apply Fact 3, it suffices to show that the sequence $\{\xi_{\mathcal{S}(t)}\}$ is conditionally $\sigma_{\max}-$sub-Gaussian for the filtration $\mathcal{F}_{t-1} = \sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t, \xi_{\mathcal{S}(1)}, \ldots, \xi_{\mathcal{S}(t-1)})$, i.e., $\mathbb{E}\left[e^{u\xi_{\mathcal{S}(t)}} \,\middle|\, \mathcal{F}_{t-1}\right] \le \exp(u^2\sigma_{\max}^2/2)$ for any $u \in \mathbb{R}$.

Given that we assume the idiosyncratic noises $\{\xi_j\}$ of posts are independent samples, it holds that $\mathbb{E}\left[e^{u\xi_{\mathcal{S}(t)}} \,\middle|\, \mathcal{F}_{t-1}\right] = \mathbb{E}\left[e^{u\xi_{\mathcal{S}(\tau)}}\right]$. Since $\xi_{\mathcal{S}(t)} \sim \mathcal{N}(0, \sigma_k^2)$ conditioned on the type of the post $k(\mathcal{S}(t)) = k$, we have $\mathbb{E}\left[e^{u\xi_{\mathcal{S}(t)}}\right] = \exp(u^2\sigma_{k(\mathcal{S}(t))}^2/2) \le \exp(u^2\sigma_{\max}^2/2)$, and thus $\{\xi_{\mathcal{S}(t)}\}$ is conditionally $\sigma_{\max}-$sub-Gaussian. Hence the conditions of Fact 3 are met and we finish the proof by applying Fact 3. $\square$

D.4 Bounding the Number of Label-Driven Admissions (Lemma 5.5)

We define $\mathcal{T}_E$ as the set of periods where a new post is admitted into the label-driven queue, i.e., $\mathcal{T}_E = \{t \le T \colon \sum_{k\in\mathcal{K}} E_k(t) = 1\}$, and $\mathbb{E}\left[\sum_{t=1}^T \sum_{k\in\mathcal{K}} E_k(t)\right] = \mathbb{E}\left[|\mathcal{T}_E|\right]$. The following lemma connects $\mathbb{E}\left[|\mathcal{T}_E|\right]$ with the estimation error $\|\boldsymbol{\phi}_{k(j(t))}\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}^2$ for posts admitted into the label-driven queue.

Lemma D.1. *For $\delta \in (0,1)$, it holds that* $\mathbb{E}\left[|\mathcal{T}_E|\right] \le \frac{4B_\delta^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_E} \|\boldsymbol{\phi}_{k(j(t))}\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}^2\right]}{\max(\eta,\gamma)^2} + T\delta.$

Proof of Lemma D.1. By the law of total expectation, the fact that $|\mathcal{T}_E| \le T$ and Lemma 5.4 that $\mathbb{P}\{\mathcal{E}^c\} \le \delta$,

$$\mathbb{E}\left[|\mathcal{T}_E|\right] = \mathbb{E}\left[|\mathcal{T}_E| \,|\, \mathcal{E}\right] \mathbb{P}\{\mathcal{E}\} + \mathbb{E}\left[|\mathcal{T}_E \,|\, \mathcal{E}^c|\right] \mathbb{P}\{\mathcal{E}^c\} \le \mathbb{E}\left[|\mathcal{T}_E| \,|\, \mathcal{E}\right] + T\delta. \tag{23}$$

It remains to upper bound $\mathbb{E}\left[|\mathcal{T}_E| \,|\, \mathcal{E}\right]$. We condition on $\mathcal{E}$, which implies that $\boldsymbol{\theta}^\star \in \mathcal{C}_t$ for any $t \ge 0$. Focusing on posts that are admitted in the label-driven queue $t \in \mathcal{T}_E$ and letting $k = k(j(t))$ be the type of the post arrived in period t, it holds that $E_k(t) = 1$. This also implies that $\underline{h}_k(t) < -\gamma$ and $\gamma < \bar{h}_k(t)$ by the label-driven admission rule (Line 7 of Algorithm 3). In addition, since $\boldsymbol{\theta}^\star \in \mathcal{C}_t$, it holds that $\underline{h}_k(t) \le h_k(t) \le \bar{h}_k(t)$. By the margin assumption $|h_k(t)| \ge \eta$, we must have $\bar{h}_k(t) - \underline{h}_k(t) \ge \eta + \gamma \ge \max(\eta, \gamma)$. To see this, if $h_k(t) \ge 0$, then $\bar{h}_k(t) \ge \eta$ and $\bar{h}_k(t) - \underline{h}_k(t) \ge \eta - (-\gamma) = \eta + \gamma$; same analysis holds for $h_k(t) < 0$. In addition, for any $\boldsymbol{\theta} \in \mathcal{C}_{t-1}$,

$$\left|\boldsymbol{\phi}_k^\intercal \boldsymbol{\theta} - \boldsymbol{\phi}_k^\intercal \hat{\boldsymbol{\theta}}_{t-1}\right| = \left|\boldsymbol{\phi}_k^\intercal (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{t-1})\right| \le \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{t-1}\|_{\bar{\boldsymbol{V}}_{t-1}} \|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}} \le B_\delta(t-1)\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}$$

where the first inequality is by Cauchy Inequality and the second inequality is by the definition of $\mathcal{C}_{t-1}$ in (14). Combining this result with the definition of confidence intervals in (15), we have

$$\max(\eta,\gamma) \le \bar{h}_k(t) - \underline{h}_k(t) \le \max_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}} \boldsymbol{\phi}_k^\intercal \boldsymbol{\theta} - \min_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}} \boldsymbol{\phi}_k^\intercal \boldsymbol{\theta} \le 2B_\delta(t-1)\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}},$$

so $4B_\delta^2(T)\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}^2 \ge \max(\eta,\gamma)^2$ and $4B_\delta^2(T)\sum_{t\in|\mathcal{T}_E|} \|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}^2 \ge |\mathcal{T}_E| \max(\eta,\gamma)^2$ conditioned on $\mathcal{E}$. This shows that $\mathbb{E}\left[|\mathcal{T}_E| \,|\, \mathcal{E}\right] \le \frac{4B_\delta^2(T)\sum_{t\in|\mathcal{T}_E|} \|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}^2}{\max(\eta,\gamma)^2}$, which finishes the proof by (23). $\square$

To upper bound the estimation error, the following result shows that the estimation error for a feature vector is larger when using a subset of data points.

Lemma D.2. *Given $\kappa > 0$ and two subsets $\mathcal{T}_1, \mathcal{T}_2$ of time periods such that $\mathcal{T}_1 \subseteq \mathcal{T}_2$, we define $\boldsymbol{V}_1 = \kappa \boldsymbol{I} + \sum_{t \in \mathcal{T}_1} \boldsymbol{\phi}_{k(\mathcal{S}(t))} \boldsymbol{\phi}_{k(\mathcal{S}(t))}^\intercal$ and $\boldsymbol{V}_2 = \kappa \boldsymbol{I} + \sum_{t \in \mathcal{T}_2} \boldsymbol{\phi}_{k(\mathcal{S}(t))} \boldsymbol{\phi}_{k(\mathcal{S}(t))}^\intercal$ coming from the data points in periods $\mathcal{T}_1, \mathcal{T}_2$ respectively. Then for any vector $\boldsymbol{u} \in \mathbb{R}^d$, it holds that $\|\boldsymbol{u}\|_{\boldsymbol{V}_1^{-1}} \geq \|\boldsymbol{u}\|_{\boldsymbol{V}_2^{-1}}$.*

Proof. Let $\boldsymbol{V}' = \boldsymbol{V}_2 - \boldsymbol{V}_1$. We have $\boldsymbol{V}' = \sum_{t \in \mathcal{T}_2 \setminus \mathcal{T}_1} \|\boldsymbol{\phi}_{k(\mathcal{S}(t))}\|^2_{\bar{\boldsymbol{V}}_{t-1}^{-1}}$, which is positive-semi definite (PSD). Also, $\boldsymbol{V}_1^{-1/2}$ exists because $\boldsymbol{V}_1$ is PSD. As a result, for any $\boldsymbol{u} \in \mathbb{R}^d$, denoting the minimum eigenvalue of a matrix $\boldsymbol{A}$ as $\lambda_{\min}(\boldsymbol{A})$, we have

$$\begin{aligned}
\|\boldsymbol{u}\|^2_{\boldsymbol{V}_2^{-1}} = \boldsymbol{u}^\intercal \boldsymbol{V}_2^{-1} \boldsymbol{u} = \boldsymbol{u}^\intercal (\boldsymbol{V}_1 + \boldsymbol{V}')^{-1} \boldsymbol{u} &= \boldsymbol{u}^\intercal \boldsymbol{V}_1^{-1/2} \left(\boldsymbol{I} + \boldsymbol{V}_1^{1/2} \boldsymbol{V}' \boldsymbol{V}_1^{1/2} \right)^{-1} \boldsymbol{V}_1^{-1/2} \boldsymbol{u} \\
&\leq \frac{\|\boldsymbol{V}_1^{-1/2} \boldsymbol{u}\|_2^2}{\lambda_{\min}\left(\boldsymbol{I} + \boldsymbol{V}_1^{1/2} \boldsymbol{V}' \boldsymbol{V}_1^{1/2} \right)} \leq \|\boldsymbol{V}_1^{-1/2} \boldsymbol{u}\|_2^2 = \|\boldsymbol{u}\|^2_{\boldsymbol{V}_1^{-1}}
\end{aligned}$$

where the first inequality is by the Courant-Fischer theorem and the fact that eigenvalues of a matrix inverse are inverses of the matrix; the second inequality is because $\boldsymbol{V}_1, \boldsymbol{V}'$ are PSD and thus $\lambda_{\min}\left(\boldsymbol{I} + \boldsymbol{V}_1^{1/2} \boldsymbol{V}' \boldsymbol{V}_1^{1/2} \right) \geq 1$. □

We next bound $\mathbb{E}\left[|\mathcal{T}_E| \right]$ by bounding $\sum_{t \in \mathcal{T}_E} \|\boldsymbol{\phi}_{k(j(t))}\|^2_{\bar{\boldsymbol{V}}_{t-1}^{-1}}$.

Proof of Lemma 5.5. We denote elements in the set $\mathcal{T}_E$ by $1 \leq t_1 < \ldots < t_M$ with $M = |\mathcal{T}_E|$, and define $\tilde{\boldsymbol{V}}_m = \kappa \boldsymbol{I} + \sum_{i \leq m} \boldsymbol{\phi}_{k(j(t_i))} \boldsymbol{\phi}_{k(j(t_i))}^\intercal$ for any $m \leq M$, which resembles the definition of $\bar{\boldsymbol{V}}_t$ in (13) but is restricted to data collected from posts sent to the label-driven queue. Let us fix $m \leq M$ and consider the norm $\|\boldsymbol{\phi}_{k(j(t_m))}\|_{\bar{\boldsymbol{V}}_{t_m-1}^{-1}}$. Since the label-driven admission rule (Line 7 in Algorithm 3) only admits a post to the label-driven queue when it is empty, we have that in period t_m, posts that arrived in periods $\{t_1, \ldots, t_{m-1}\}$ are already reviewed as they were admitted into the label-driven queue and the queue is now empty. As a result, posts that arrived in periods $\{t_1, \ldots, t_{m-1}\}$ are in the dataset $\mathcal{D}_{t_m-1}$. By Lemma D.2, we have $\|\boldsymbol{\phi}_{k(j(t_m))}\|^2_{\bar{\boldsymbol{V}}_{t_m-1}^{-1}} \leq \|\boldsymbol{\phi}_{k(j(t_m))}\|^2_{\tilde{\boldsymbol{V}}_{m-1}^{-1}}$ and thus

$$\sum_{t \in \mathcal{T}_E} \|\boldsymbol{\phi}_{k(j(t))}\|^2_{\bar{\boldsymbol{V}}_{t-1}^{-1}} \leq \sum_{m=1}^{M} \|\boldsymbol{\phi}_{k(j(t_m))}\|^2_{\tilde{\boldsymbol{V}}_{m-1}^{-1}}.$$

We thus bound $\sum_{t \in \mathcal{T}_E} \|\boldsymbol{\phi}_{k(j(t))}\|^2_{\bar{\boldsymbol{V}}_{t-1}^{-1}}$ by bounding

$$\sum_{m=1}^{M} \|\boldsymbol{\phi}_{k(j(t_m))}\|^2_{\tilde{\boldsymbol{V}}_{m-1}^{-1}} \leq 2 \ln \frac{\det(\tilde{\boldsymbol{V}}_M)}{\det(\kappa \boldsymbol{I})} \leq 2 \ln \frac{(\kappa + MU^2/d)^d}{\kappa^d} = 2d \ln(1 + MU^2/(d\kappa)) \leq 2d \ln(1 + T/d),$$

where the first inequality is by Fact 4 (with $\kappa \geq \max(1, U^2)$); the second inequality is by Fact 5; the last inequality is by $\kappa \geq U^2, M \leq T$. Applying Lemma D.1 gives $\mathbb{E}\left[|\mathcal{T}_E| \right] \leq \frac{8B_\delta^2(T) d \ln(1+T/d)}{\max(\eta,\gamma)^2} + T\delta \leq \frac{9B_\delta^2(T) d \ln(1+T/d)}{\max(\eta,\gamma)^2}$ where the last inequality is because $\max(\eta, \gamma) \leq 1$, $T\delta \leq 1 \leq B_\delta^2(T) d \ln(1 + T/d)$ as $B_\delta^2(T) \geq \kappa U^2 \geq 1$ (we assume $U \geq 1$) and $d \ln(1 + T/d) \geq d \ln(1 + 3/d) \geq 1$ by $T \geq 3$ and Fact 9. □

D.5 Bounding Per-Period Classification Loss (Lemma 5.7)

Proof of Lemma 5.7. Let us condition on $\mathcal{E}$ (the event that the confidence set is correct) and bound

$$Z_k(t) = (Y_k(t)(r_k^O - r_k^R)^+ + (1 - Y_k(t))(r_k^R - r_k^O)^+)\ell_k(\Lambda_k(t) - A_k(t) - E_k(t)).$$

We fix a period t and a type k such that a type-k post arrives in period t and it is not admitted into the label-driven queue, i.e., $\Lambda_k(t) = 1$ and $E_k(t) = 0$; otherwise $Z_k(t) = 0$. Since $\mathcal{E}$ holds, we have $\boldsymbol{\theta}^\star \in \mathcal{C}_{t-1}$ and thus $h_k \in [\underline{h}_k(t), \bar{h}_k(t)]$ for any period t. We consider the three possible cases when the algorithm makes the classification decision $Y_k(t)$:

- $\underline{h}_k(t) \geq -\gamma$: we first suppose the empirical estimate of expected cost $\boldsymbol{\phi}_k^\intercal \hat{\boldsymbol{\theta}}(t) > 0$. In this case we classify the post by $Y_k(t) = 0$ (Line 5 in Algorithm 3). Then, if $h_k > 0$, we have $Z_k(t) = 0$ since $r_k^R \leq r_k^O$ by (5). If $h_k \leq 0$, we have $-\gamma \leq \underline{h}_k(t) \leq h_k \leq 0$ and $Z_k(t) \leq 3\gamma v_k \ell_k$ as in (20).

  We next consider the case where the empirical estimate of expected cost $\boldsymbol{\phi}_k^\intercal \hat{\boldsymbol{\theta}}(t) \leq 0$. We argue that we must have $\bar{h}_k(t) \leq \gamma$. In this case, since $h_k \in [\underline{h}_k(t), \bar{h}_k(t)] \subseteq [-\gamma, \gamma]$, we also have $Z_k(t) \leq 3\gamma v_k \ell_k$ by the same argument in (20). It remains to show $\bar{h}_k(t) \leq \gamma$. Suppose not, then there exists $\boldsymbol{\theta} \in \mathcal{C}_{t-1}$ such that $\boldsymbol{\phi}_k^\intercal \boldsymbol{\theta} > \gamma$. In addition, we know $\boldsymbol{\theta}' = 2\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}$ is also in $\mathcal{C}_{t-1}$ since $\boldsymbol{\theta}' - \hat{\boldsymbol{\theta}}(t) = -(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t))$ and thus the matrix norms $\|\boldsymbol{\theta}' - \hat{\boldsymbol{\theta}}(t)\|_{\bar{V}_{t-1}^{-1}}, \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t)\|_{\bar{V}_{t-1}^{-1}}$ are the same. However, we have $\boldsymbol{\phi}_k^\intercal \boldsymbol{\theta}' = \boldsymbol{\phi}_k^\intercal (2\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}) < 0 - \gamma = -\gamma$, which contradicts to the assumption that $\underline{h}_k(t) = \min_{\boldsymbol{\theta}_1 \in \mathcal{C}_{t-1}} \boldsymbol{\phi}_k^\intercal \boldsymbol{\theta}_1 \geq -\gamma$. Therefore, we must have $\bar{h}_k(t) \leq \gamma$. Summarizing the two cases shows that if $\underline{h}_k(t) \geq -\gamma$ then $Z_k(t) \leq 3\gamma v_k \ell_k$.

- $\bar{h}_k(t) \leq \gamma$: we also have $Z_k(t) \leq 3\gamma v_k \ell_k$ by a symmetric argument of the above case;

- $\underline{h}_k(t) \leq -\gamma < \gamma \leq \bar{h}_k(t)$: in this case the label-driven admission will admit the post unless $Q^{\mathrm{LD}}(t) = 1$. Since $E_k(t) = 0$, we must have $Q^{\mathrm{LD}}(t) = 1$, and thus $Z_k(t) \leq r_{\max} \ell_{\max} Q^{\mathrm{LD}}(t)$.

Combining the above cases, $Z_k(t) \leq (3\gamma v_{\max} \ell_{\max} + r_{\max} \ell_{\max} Q^{\mathrm{LD}}(t))\Lambda_k(t)$ conditioned on $\mathcal{E}$. $\square$

D.6 Idiosyncrasy Loss and Lagrangians with Type-Aggregated Duals (Lemma 5.8)

The proof is close to that of Lemmas 3.3 and 4.10, but relies on the analysis of a different Lyapunov function.

Proof. We have $\tilde{L}(1) = 0$ and that $\mathbb{E}\left[\tilde{L}(T+1)/\beta\right]$ upper bounds the idiosyncrasy loss, which proves the first inequality. For the second inequality, for any period t:

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{L}(t+1)-\tilde{L}(t)\right] \\
&=\beta\mathbb{E}\left[\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t)-E_k(t))\right]+\mathbb{E}\left[\sum_{g\in\mathcal{G}}\frac{1}{2}\left((\tilde{Q}_g(t+1))^2-(\tilde{Q}_g(t))^2\right)\right] \\
&=\mathbb{E}\left[\beta\sum_{k\in\mathcal{K}} r_k\ell_k(\Lambda_k(t)-A_k(t)-E_k(t))\right]+\frac{1}{2}\mathbb{E}\left[\sum_{g\in\mathcal{G}}\left(\sum_{k:g(k)=g}(A_k(t)-S_k(t))\right)^2\right] \\
&\quad+\mathbb{E}\left[\sum_{g\in\mathcal{G}}\tilde{Q}_g(t)\left(\sum_{k:g(k)=g}(A_k(t)-S_k(t))\right)\right] \\
&\overset{(a)}{\leq} 1+\beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t)+\mathbb{E}\left[-\beta\sum_{k\in\mathcal{K}} r_k\ell_k(A_k(t)+E_k(t))+\sum_{k\in\mathcal{K}}\tilde{Q}_g(t)\left(A_k(t)-S_k(t)\right)\right] \\
&\leq 1+\beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t)+\mathbb{E}\left[-\beta\sum_{k\in\mathcal{K}} r_k\ell_k\bar{A}_k(t)+\sum_{k\in\mathcal{K}}\tilde{Q}_{g(k)}(t)\left(\bar{A}_k(t)-S_k(t)\right)\right]
\end{aligned}
$$

where inequality (a) is because there is at most one type with $A_k(t)=1$ and at most one type with $S_k(t)=1$; the last inequality is because $A_k(t)\leq\bar{A}_k(t)\leq A_k(t)+E_k(t)$. In addition, $\mathbb{E}\left[S_k(t)\,|\,\psi_k(t)\right]=\mu_k N(t)\psi_k(t)$. As a result, recalling the definition of f_t from (7),

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{L}(t+1)-\tilde{L}(t)\right] \\
&\leq 1+\beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t)-\mathbb{E}\left[\sum_{k\in\mathcal{K}}\bar{A}_k(t)\left(\beta r_k\ell_k-\tilde{Q}_{g(k)}(t)\right)+\sum_{k\in\mathcal{K}}\tilde{Q}_{g(k)}(t)\psi_k(t)\mu_k N(t)\right] \\
&=1+\mathbb{E}\left[f_t(\bar{\boldsymbol{A}}(t),\boldsymbol{\psi}(t),\boldsymbol{Q}^{\mathrm{TA}}(t))\right]
\end{aligned}
$$

where $Q_k^{\mathrm{TA}}(t)=\tilde{Q}_{g(k)}(t)$. We obtain the desired result by telescoping from $t=1$ to $t=T$. $\square$

D.7 Lagrangians with Type-Aggregated Duals and Fluid Benchmark (Lemma 5.9)

We fix a window size w and denote the corresponding optimal partition, admission and service vectors to (w-fluid) by $\{\tau_i^\star\}_{i\in[I]},\{\boldsymbol{a}^\star(t)\}_{t\in[T]},\{\boldsymbol{\nu}^\star(t)\}_{t\in[T]}$ where $\tau_i^\star$ is the beginning period of each window for (w-fluid) which gives $\mathcal{L}^\star(w,T)$. The proof follows a similar structure with Section 3.4 but now uses a different dual $\boldsymbol{Q}^{\mathrm{TA}}$. To address the time-varying nature of $\boldsymbol{Q}^{\mathrm{TA}}$, we define a new vector of dual variables $\boldsymbol{u}^\star$ such that at the beginning of the $i-$th window, the dual $u_{i,k}^\star$ takes the corresponding queue length in $\boldsymbol{Q}^{\mathrm{TA}}$, i.e., $u^\star_{i,k} = Q^{\mathrm{TA}}_k(\tau^\star_i) = \tilde{Q}_{g(k)}(\tau^\star_i)$ for $i \in [I], k \in \mathcal{K}$. Then

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t), \boldsymbol{\psi}^{\mathrm{TABACID}}(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right] - \beta \mathcal{L}^\star(w,T) \\
&= \mathbb{E}\left[f(\{\boldsymbol{a}^\star(t)\}_t, \{\boldsymbol{\nu}^\star(t)\}_t, \boldsymbol{u}^\star)\right] - \beta \mathcal{L}^\star(w,T) \qquad (24)\\
&+ \sum_{i=1}^{I} \sum_{t=\tau^\star_i}^{\tau^\star_{i+1}-1} \left(\mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right] - \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}^{\mathrm{TA}}(\tau^\star_i))\right]\right) \qquad (25)\\
&+ \sum_{t=1}^{T} \left(\mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t), \boldsymbol{\psi}^{\mathrm{TABACID}}(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right] - \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right]\right). \qquad (26)
\end{aligned}$$

The first term (difference between a Lagrangian and the primal) is non-positive by Lemma 3.5.

In a similar way with Lemma 3.6, the next lemma bounds the second term (suboptimality due to time-varying queue lengths as dual variables within a window).

Lemma D.3. *(25) $\le (w-1)T$.*

Proof. The proof is similar to that of Lemma 3.6 and we include it here for completeness. We first have that, for any $t_1 \le t_2, k \in \mathcal{K}$,

$$\begin{aligned}
|Q^{\mathrm{TA}}_k(t_2) - Q^{\mathrm{TA}}_k(t_1)| = |\tilde{Q}_{g(k)}(t_2) - \tilde{Q}_{g(k)}(t_1)| &\le \sum_{t=t_1}^{t_2-1} |\tilde{Q}_{g(k)}(t+1) - \tilde{Q}_{g(k)}(t)| \\
&= \sum_{t=t_2}^{t_2-1} \left| \sum_{k' :\, g(k')=g(k)} A_{k'}(t) - \sum_{k' :\, g(k')=g(k)} S_{k'}(t)) \right| \le t_2 - t_1,
\end{aligned}$$

where the last inequality is because for each period at most one post arrives and at most one post gets reviewed. Therefore, the difference in Lagrangian under the optimal decisions to (w-fluid) is

$$\begin{aligned}
&|f_{t_2}(\boldsymbol{a}^\star(t_2), \boldsymbol{\nu}^\star(t_2), \boldsymbol{Q}^{\mathrm{TA}}(t_2)) - f_{t_2}(\boldsymbol{a}^\star(t_2), \boldsymbol{\nu}^\star(t_2), \boldsymbol{Q}^{\mathrm{TA}}(t_1))| \\
&\le \sum_{k \in \mathcal{K}} (a^\star_k(t_2) + \nu^\star_k(t_2)\mu_k N(t_2)) |Q^{\mathrm{TA}}_k(t_2) - Q^{\mathrm{TA}}_k(t_1)| \\
&\le (t_2 - t_1) \sum_{k \in \mathcal{K}} (a^\star_k(t_2) + \nu^\star_k(t_2)\mu_k N(t_2)) \le (t_2 - t_1) \left(\sum_{k \in \mathcal{K}} \lambda_k(t_2) + \sum_{k \in \mathcal{K}} \nu^\star_k \right) \le 2(t_2 - t_1)
\end{aligned}$$

where the second-to-last inequality holds because we can only admit when there is an arrival, and $\mu_k N(t_2) \le 1$ by our model assumption.

We finish the proof by taking expectation and summing across windows

$$\begin{aligned}
\sum_{i=1}^{I} \sum_{t=\tau^\star_i}^{\tau^\star_{i+1}-1} \left(\mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}^{\mathrm{TA}}(t))\right] - \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t), \boldsymbol{\nu}^\star(t), \boldsymbol{Q}^{\mathrm{TA}}(\tau^\star_i))\right]\right) &\le 2 \sum_{i=1}^{I} \sum_{t=\tau^\star_i}^{\tau^\star_{i+1}-1} (t - \tau^\star_i) \\
&= \sum_{i=1}^{I} (\tau^\star_{i+1} - \tau^\star_i)(\tau^\star_{i+1} - \tau^\star_i - 1) \le (w-1)T.
\end{aligned}$$

□

The next lemma is a per-period bound of (26).

Lemma D.4. *For any t,*

$$\mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t),\boldsymbol{\psi}^{\mathrm{TABACID}}(t),\boldsymbol{Q}^{\mathrm{TA}}(t))\right] \le \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t),\boldsymbol{\nu}^\star(t),\boldsymbol{Q}^{\mathrm{TA}}(t))\right] + 2\beta r_{\max}\ell_{\max}\Delta(\mathcal{K}_{\mathcal{G}}).$$

Proof. For any $\tilde{\boldsymbol{q}}$ such that $\tilde{q}_g \le 2\beta r_{\max}\ell_{\max}$, we expand the expectation on the left hand side, condition on $\tilde{\boldsymbol{Q}}(t) = \tilde{\boldsymbol{q}}$, by

$$\begin{aligned}
&\mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t),\boldsymbol{\psi}^{\mathrm{TABACID}}(t),\boldsymbol{Q}^{\mathrm{TA}}(t)) \,\middle|\, \tilde{\boldsymbol{Q}}(t) = \tilde{\boldsymbol{q}}\right]\\
&= \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) - \left(\sum_{k\in\mathcal{K}}\lambda_k(t)\mathbb{1}\left(\beta r_k\ell_k \ge \tilde{q}_{g(k)}\right)(\beta r_k\ell_k - \tilde{q}_{g(k)}) + \sum_{k\in\mathcal{K}}\psi_k^{\mathrm{TABACID}}(t)\tilde{q}_{g(k)}\mu_k N(t)\right)\\
&\overset{(a)}{\le} \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) - \sum_{k\in\mathcal{K}} a_k^\star(t)(\beta r_k\ell_k - \tilde{q}_{g(k)}) - \sum_{k\in\mathcal{K}}\psi_k^{\mathrm{TABACID}}(t)\tilde{q}_{g(k)}\mu_k N(t)\\
&\overset{(b)}{\le} \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) - \sum_{k\in\mathcal{K}} a_k^\star(t)(\beta r_k\ell_k - \tilde{q}_{g(k)}) - \sum_{g\in\mathcal{G}}\tilde{\mu}_g\tilde{q}_g N(t)\sum_{k\in\mathcal{K}_g}\psi_k^{\mathrm{TABACID}}(t)\\
&\overset{(c)}{\le} \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) - \sum_{k\in\mathcal{K}} a_k^\star(t)(\beta r_k\ell_k - \tilde{q}_{g(k)}) - \sum_{g\in\mathcal{G}}\tilde{\mu}_g\tilde{q}_g N(t)\sum_{k\in\mathcal{K}_g}\nu_k^\star(t)\\
&\overset{(d)}{\le} \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) - \sum_{k\in\mathcal{K}} a_k^\star(t)(\beta r_k\ell_k - \tilde{q}_{g(k)}) - \sum_{g\in\mathcal{G}}\tilde{q}_g \sum_{k\in\mathcal{K}_g}\left(N(t)\mu_k - \Delta(\mathcal{K}_{\mathcal{G}})\right)\nu_k^\star(t)\\
&\overset{(e)}{\le} \beta\sum_{k\in\mathcal{K}} r_k\ell_k\lambda_k(t) - \sum_{k\in\mathcal{K}} a_k^\star(t)(\beta r_k\ell_k - \tilde{q}_{g(k)}) - \sum_{g\in\mathcal{G}}\tilde{q}_g N(t)\sum_{k\in\mathcal{K}_g}\mu_k\nu_k^\star(t) + 2\beta r_{\max}\ell_{\max}\Delta(\mathcal{K}_{\mathcal{G}})\\
&= f_t(\boldsymbol{a}^\star,\boldsymbol{\nu}^\star,\boldsymbol{Q}^{\mathrm{TA}}(t)) + 2\beta r_{\max}\ell_{\max}\Delta(\mathcal{K}_{\mathcal{G}}),
\end{aligned} \tag{27}$$

where the first equality is by the admission rule of TABACID; Inequality (a) is because $a_k^\star(t) \le \lambda_k(t)$ and $(\beta r_k\ell_k - \tilde{q}_{g(k)}) \le \mathbb{1}\left(\beta r_k\ell_k \ge \tilde{q}_{g(k)}\right)(\beta r_k\ell_k - \tilde{q}_{g(k)})$; Inequality (b) is because $\tilde{\mu}_g = \min_{k\in\mathcal{K}_g}\mu_k$; Inequality (c) is by the scheduling rule of TABACID that picks a group maximizing $\tilde{\mu}_g\tilde{Q}_g(t)$ for review; Inequality (d) is by the definition of aggregation gap in Section 5.1; and inequality (e) is because $\tilde{q}_g \le 2\beta r_{\max}\ell_{\max}$ for any group g and $\sum_{k\in\mathcal{K}}\nu_k^\star(t) \le 1$.

By Lemma 5.6, we always have $\tilde{Q}_g(t) \le 2\beta r_{\max}\ell_{\max}$. We then finish the proof by

$$\begin{aligned}
&\mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t),\boldsymbol{\psi}^{\mathrm{TABACID}}(t),\boldsymbol{Q}^{\mathrm{TA}}(t))\right]\\
&= \mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t),\boldsymbol{\psi}^{\mathrm{TABACID}}(t),\boldsymbol{Q}^{\mathrm{TA}}(t))\mathbb{1}\left(\forall g, \tilde{Q}_g(t) \le 2\beta r_{\max}\ell_{\max}\right)\right]\\
&\overset{(27)}{\le} \mathbb{E}\left[(f_t(\boldsymbol{a}^\star,\boldsymbol{\nu}^\star,\boldsymbol{Q}^{\mathrm{TA}}(t)) + 2\beta r_{\max}\ell_{\max}\Delta(\mathcal{K}_{\mathcal{G}}))\mathbb{1}\left(\forall g, \tilde{Q}_g(t) \le 2\beta r_{\max}\ell_{\max}\right)\right]\\
&= \mathbb{E}\left[f_t(\boldsymbol{a}^\star(t),\boldsymbol{\nu}^\star(t),\boldsymbol{Q}^{\mathrm{TA}}(t))\right] + 2\beta r_{\max}\ell_{\max}\Delta(\mathcal{K}_{\mathcal{G}}).
\end{aligned}$$

□

Proof of Lemma 5.9. We have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[f_t(\boldsymbol{A}^{\mathrm{TABACID}}(t),\boldsymbol{\psi}^{\mathrm{TABACID}}(t),\boldsymbol{Q}^{\mathrm{TA}}(t))\right] - \beta\mathcal{L}^\star(w,T) &= (24) + (25) + (26)\\
&\le 0 + (w-1)T + 2\beta r_{\max}\ell_{\max}\Delta(\mathcal{K}_{\mathcal{G}})T,
\end{aligned}$$

where we use Lemmas 3.5, D.3 and D.4 to bound the three terms respectively. □

D.8 Bounding the regret in scheduling (Lemma 5.10)

The proof is similar to that of Lemma 4.12 and we include it for completeness.

Proof. For a period t, by the scheduling decision of COLBACID and TABACID, we have $\psi_k(t) = \psi_k^{\text{TABACID}}(t)$ for any k if $Q^{\text{LD}}(t) = 0$. As a result,

$$\begin{aligned}
\textsc{RegS}(T) &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} \psi_k^{\text{TABACID}}(t)\tilde{Q}_{g(k)}(t)\mu_k N(t) Q^{\text{LD}}(t)\right] \\
&\leq 2\beta r_{\max}\ell_{\max}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} \psi_k^{\text{TABACID}}(t) Q^{\text{LD}}(t)\right] \\
&\leq 2\beta r_{\max}\ell_{\max}\mathbb{E}\left[\sum_{t=1}^{T} Q^{\text{LD}}(t)\right] \leq \frac{18\beta r_{\max}\ell_{\max}B_\delta^2(T)d\ln(1+T/d)}{\max(\eta,\gamma)^2\hat{\mu}_{\min}},
\end{aligned}$$

where the first inequality is because $\tilde{Q}_g(t) \leq 2\beta r_{\max}\ell_{\max}$ by Lemma 5.6; the second inequality is because $\sum_{k\in\mathcal{K}} \psi_k^{\text{TABACID}}(t) \leq 1$; the last inequality is by Lemmas 4.4 and 5.5 that together bound $\mathbb{E}\left[\sum_{t=1}^{T} Q^{\text{LD}}(t)\right]$. □

D.9 Bounding the Regret in Admission (Lemma 5.11)

Proof of Lemma 5.11. Recall that the TABACID and optimistic admission decisions are

$$A_k^{\text{TABACID}}(t) = \Lambda_k(t)\mathbb{1}\left(\beta r_k \ell_k \geq \tilde{Q}_{g(k)}(t)\right) \quad \text{and} \quad \bar{A}_k(t) = \Lambda_k(t)\mathbb{1}\left(\beta \bar{r}_k(t)\ell_k \geq \tilde{Q}_{g(k)}(t)\right).$$

Recall also the good event $\mathcal{E}$ for which $\boldsymbol{\theta}^\star \in \mathcal{C}_t$ for any $t \geq 0$. The regret in admission is

$$\begin{aligned}
\textsc{RegA}(T) &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(A_k^{\text{TABACID}}(t) - \bar{A}_k(t))(\beta r_k\ell_k - \tilde{Q}_{g(k)}(t))\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(A_k^{\text{TABACID}}(t) - \bar{A}_k(t))(\beta r_k\ell_k - \tilde{Q}_{g(k)}(t))\left(\mathbb{1}(\mathcal{E}) + \mathbb{1}(\mathcal{E}^c)\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(A_k^{\text{TABACID}}(t) - \bar{A}_k(t))(\beta r_k\ell_k - \tilde{Q}_{g(k)}(t))\mathbb{1}(\mathcal{E})\right] + \mathbb{P}\{\mathcal{E}^c\}\sum_{t=1}^{T} 3\beta r_{\max}\ell_{\max} \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}(A_k^{\text{TABACID}}(t) - \bar{A}_k(t))(\beta r_k\ell_k - \tilde{Q}_{g(k)}(t))\mathbb{1}(\mathcal{E})\right] + 3\beta r_{\max}\ell_{\max} \qquad (28)
\end{aligned}$$

where the first inequality is because $\sum_k A_k^{\text{TABACID}}(t) \leq 1, \sum_k \bar{A}_k(t) \leq 1$ and $\tilde{Q}_{g(k)}(t) \leq 2\beta r_{\max}\ell_{\max}$ by Lemma 5.6; the last inequality is by Lemma 5.4 that $\mathbb{P}\{\mathcal{E}^c\} \leq \delta$ and the assumption that $\delta \leq 1/T$.

We next upper bound the first term in (28). Conditioning on $\mathcal{E}$, we bound $(A_k^{\text{TABACID}}(t) - \bar{A}_k(t))(\beta r_k\ell_k - \tilde{Q}_{g(k)}(t))$ for a fixed period t and a fixed type k. Since $\boldsymbol{\theta}^\star \in \mathcal{C}_{t-1}$ by $\mathcal{E}$, we have that

h_k is inside the confidence interval $[\underline{h}_k(t), \bar{h}_k(t)]$ and thus $\bar{r}_k(t) \geq r_k$. If there is no type-k arrival, i.e., $\Lambda_k(t) = 0$, we have that $A_k^{\text{TABACID}}(t) = \bar{A}_k(t) = 0$ and thus the term is zero. Otherwise, since $\bar{r}_k(t) \geq r_k$, $\bar{A}_k(t) = \mathbb{1}\left(\beta\bar{r}_k(t)\ell_k \geq \tilde{Q}_{g(k)}(t)\right) \geq \mathbb{1}\left(\beta r_k\ell_k \geq \tilde{Q}_{g(k)}(t)\right) = A_k^{\text{TABACID}}(t)$. Therefore, $(A_k^{\text{TABACID}}(t) - \bar{A}_k(t))(\beta r_k\ell_k - \tilde{Q}_{g(k)}(t))$ is positive only if $A_k^{\text{TABACID}}(t) = 0$ and $\bar{A}_k(t) = 1$, under which we have

$$\begin{aligned}(A_k^{\text{TABACID}}(t) - \bar{A}_k(t))(\beta r_k\ell_k - \tilde{Q}_{g(k)}(t)) &= (\tilde{Q}_{g(k)}(t) - \beta r_k\ell_k)\bar{A}_k(t) \\ &= (\tilde{Q}_{g(k)}(t) - \beta\bar{r}_k(t)\ell_k + \beta\bar{r}_k(t)\ell_k - \beta r_k\ell_k)\bar{A}_k(t) \\ &\leq \beta\ell_{\max}\left(\bar{r}_k(t) - r_k\right)\bar{A}_k(t), \end{aligned} \tag{29}$$

where the last inequality is because the optimistic admission rule only admits a post ($\bar{A}_k(t) = 1$) when $\tilde{Q}_{g(k)}(t) \leq \beta\bar{r}_k(t)\ell_k$. To bound the right hand side of (29), we first bound $\bar{r}_k(t) - r_k$ by

$$\begin{aligned}\bar{r}_k(t) - r_k &= \max_{h\in[\underline{h}_k(t),\bar{h}_k(t)]} r_k(h) - r_k(h_k) = \max_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}}\left(r_k(\boldsymbol{\phi}_k^\intercal\boldsymbol{\theta}) - r_k(\boldsymbol{\phi}_k^\intercal\boldsymbol{\theta}^\star)\right) \\ &\leq 1.5v_k \max_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}} |\boldsymbol{\phi}_k^\intercal\boldsymbol{\theta} - \boldsymbol{\phi}_k^\intercal\boldsymbol{\theta}^\star| && \text{(By Lemma 4.8)} \\ &\leq 1.5v_k \max_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}}\left(|\boldsymbol{\phi}_k^\intercal\boldsymbol{\theta} - \boldsymbol{\phi}_k^\intercal\hat{\boldsymbol{\theta}}_{t-1}| + |\boldsymbol{\phi}_k^\intercal\hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\phi}_k^\intercal\boldsymbol{\theta}^\star|\right) \\ &\leq 1.5v_k \max_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}}\left(\|\boldsymbol{\phi}_k\|_{\bar{V}_{t-1}^{-1}}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{t-1}\|_{\bar{V}_{t-1}} + \|\boldsymbol{\phi}_k\|_{\bar{V}_{t-1}^{-1}}\|\hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^\star\|_{\bar{V}_{t-1}}\right) \\ & && \text{(Cauchy-Schwarz Inequality)} \\ &\leq 3r_{\max}B_\delta(T)\|\boldsymbol{\phi}_k\|_{\bar{V}_{t-1}^{-1}} && (30)\end{aligned}$$

where the last inequality is because we condition on $\mathcal{E}$, which requires $\boldsymbol{\theta}^\star \in \mathcal{C}_{t-1}$, the definition of $\mathcal{C}_{t-1}$ in (14), and our definition of $r_{\max} = v_{\max}(H + \sigma_{\max}) > v_{\max}$. We complete the proof by

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^T\sum_{k\in\mathcal{K}}(A_k^{\text{TABACID}}(t) - \bar{A}_k(t))(\beta r_k\ell_k - \tilde{Q}_{g(k)}(t))\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^T\sum_{k\in\mathcal{K}}\beta\ell_{\max}(\bar{r}_k(t) - r_k)\bar{A}_k(t)\mathbb{1}\left(\mathcal{E}\right)\right] + 3\beta r_{\max}\ell_{\max} && \text{(By (29))} \\ &\leq 3\beta r_{\max}\ell_{\max}B_\delta(T)\left(\mathbb{E}\left[\sum_{t=1}^T\sum_{k\in\mathcal{K}}\|\boldsymbol{\phi}_k\|_{\bar{V}_{t-1}^{-1}}\bar{A}_k(t)\right] + 1\right) && \text{(By (30))} \\ &\leq 3\beta r_{\max}\ell_{\max}B_\delta(T)\left(d\ln(1+T/d)\left(4GQ_{\max} + \frac{9B_\delta^2(T)}{\max(\eta,\gamma)^2}\right) + \sqrt{2GTd\ln(1+T/d)}\right)\end{aligned}$$

where we use Lemma 5.13 for the last inequality. □

D.10 Bounding estimation error with fixed delay (Lemma 5.12)

Proof. We first bound the case when there is no delay, i.e., $q = 1$, which is the case in [APS11]. In particular, assuming that we have M data points, by Cauchy-Schwarz Inequality

$$\begin{aligned}\sum_{i=1}^M\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \leq \sqrt{M\sum_{i=1}^M\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}}^2} \leq \sqrt{2M\ln(\det(\hat{\boldsymbol{V}}_n)/\kappa^d)} &\leq \sqrt{2Md\ln(1+MU^2/(d\kappa))} \\ &\leq \sqrt{2Md\ln(1+M/d)} \end{aligned} \tag{31}$$

where we use Fact 4 for the second inequality, Fact 5 for the third inequality, and $\kappa \geq U^2$ (assumption in the lemma) for the last inequality. To bound for arbitrary q, we observe that

$$\sum_{i=q}^{M} \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} \leq \sum_{i=1}^{M} \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} + \sum_{i=q}^{M} \left(\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} - \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \right). \tag{32}$$

Note that the first sum is already bounded by (31). The second sum measures the difference between the estimation errors when there is no delay and when there is a delay of q. We next upper bound the second term for a fixed data point i by

$$\begin{aligned}
\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} - \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} &= \left(\sqrt{\hat{\boldsymbol{\phi}}_i^{\intercal} \hat{\boldsymbol{V}}_{i-q}^{-1} \hat{\boldsymbol{\phi}}_i} - \sqrt{\hat{\boldsymbol{\phi}}_i^{\intercal} \hat{\boldsymbol{V}}_{i-1}^{-1} \hat{\boldsymbol{\phi}}_i} \right) \\
&\leq \frac{\hat{\boldsymbol{\phi}}_i^{\intercal} \left(\hat{\boldsymbol{V}}_{i-q}^{-1} - \hat{\boldsymbol{V}}_{i-1}^{-1} \right) \hat{\boldsymbol{\phi}}_i}{\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}}} \qquad \left(\forall a, b > 0,\ \sqrt{a} - \sqrt{b} = \tfrac{a-b}{\sqrt{a}+\sqrt{b}} \leq \tfrac{a-b}{\sqrt{a}}\right) \\
= \frac{\hat{\boldsymbol{\phi}}_i^{\intercal} \hat{\boldsymbol{V}}_{i-1}^{-1} \hat{\boldsymbol{V}}_{i-1} \left(\hat{\boldsymbol{V}}_{i-q}^{-1} - \hat{\boldsymbol{V}}_{i-1}^{-1} \right) \hat{\boldsymbol{V}}_{i-q} \hat{\boldsymbol{V}}_{i-q}^{-1} \hat{\boldsymbol{\phi}}_i}{\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}}} &= \frac{\hat{\boldsymbol{\phi}}_i^{\intercal} \hat{\boldsymbol{V}}_{i-1}^{-1} \left(\sum_{\tau=i-q+1}^{i-1} \hat{\boldsymbol{\phi}}_\tau \hat{\boldsymbol{\phi}}_\tau^{\intercal} \right) \hat{\boldsymbol{V}}_{i-q}^{-1} \hat{\boldsymbol{\phi}}_i}{\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}}}.
\end{aligned} \tag{33}$$

Further expanding (33), we have

$$\begin{aligned}
\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} - \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} &\leq \frac{\sum_{\tau=i-q+1}^{i-1} \hat{\boldsymbol{\phi}}_i^{\intercal} \hat{\boldsymbol{V}}_{i-1}^{-1} \hat{\boldsymbol{\phi}}_\tau \hat{\boldsymbol{\phi}}_\tau^{\intercal} \hat{\boldsymbol{V}}_{i-q}^{-1} \hat{\boldsymbol{\phi}}_i}{\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}}} \\
&\leq \frac{\sum_{\tau=i-q+1}^{i-1} \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \|\hat{\boldsymbol{\phi}}_\tau\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \|\hat{\boldsymbol{\phi}}_\tau\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} \cancel{\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}}}}{\cancel{\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}}}} \\
&\leq \sum_{\tau=i-q+1}^{i-1} \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \|\hat{\boldsymbol{\phi}}_\tau\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}}
\end{aligned}$$

where the second inequality is because

$$\hat{\boldsymbol{\phi}}_i^{\intercal} \hat{\boldsymbol{V}}_{i-1}^{-1} \hat{\boldsymbol{\phi}}_\tau = \left(\hat{\boldsymbol{V}}_{i-1}^{-1/2} \hat{\boldsymbol{\phi}}_i \right)^{\intercal} \left(\hat{\boldsymbol{V}}_{i-1}^{-1/2} \hat{\boldsymbol{\phi}}_\tau \right) \leq \|\hat{\boldsymbol{V}}_{i-1}^{-1/2} \hat{\boldsymbol{\phi}}_i\|_2 \|\hat{\boldsymbol{V}}_{i-1}^{-1/2} \hat{\boldsymbol{\phi}}_\tau\|_2 = \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \|\hat{\boldsymbol{\phi}}_\tau\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}}$$

by Cauchy-Schwarz Inequality (same argument for $\hat{\boldsymbol{\phi}}_\tau^{\intercal} \hat{\boldsymbol{V}}_{i-q}^{-1} \hat{\boldsymbol{\phi}}_i$); the last inequality is because $\|\hat{\boldsymbol{\phi}}_\tau\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} \leq 1$ under the assumption that $\kappa \geq U^2$. As a result,

$$\begin{aligned}
\sum_{i=q}^{M} \left(\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} - \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \right) &\leq \sum_{i=q}^{M} \left(\sum_{\tau=i-q+1}^{i-1} \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \|\hat{\boldsymbol{\phi}}_\tau\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}} \right) \\
&\leq \frac{1}{2} \sum_{i=q}^{M} \sum_{\tau=i-q+1}^{i-1} \left(\|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}}^2 + \|\hat{\boldsymbol{\phi}}_\tau\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}}^2 \right) \qquad (ab \leq (a^2+b^2)/2) \\
&\leq q \sum_{i=1}^{M} \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-1}^{-1}}^2 \leq 2qd \ln(1 + M/d),
\end{aligned} \tag{34}$$

where the third inequality is because every term is counted for q times; and the last two inequalities again follow again from Facts 4 and 5. Combining (31), (32), and (34), we complete the proof by

$$\sum_{i=q}^{M} \|\hat{\boldsymbol{\phi}}_i\|_{\hat{\boldsymbol{V}}_{i-q}^{-1}} \leq \sqrt{2Md\ln(1+M/d)} + 2qd\ln(1+M/d).$$

□

D.11 Bounding Estimation Error with Queueing Delays (Lemma 5.13)

Our proof starts by bounding $\sum_{t=1}^{T}\sum_k \|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}} A_k(t)$ where recall that $A_k(t) = \bar{A}_k(t)(1-E_k(t))$ denotes whether a post is admitted by the optimistic admission rule (Line 8). We bound the estimation error separately for different groups. In particular, fixing a group g, suppose that in the first T periods there are M_g group-g posts such that $A_k(t) = 1$ where k is the type of this post and t is the period it arrives. We denote the sequence of these posts by $\mathcal{J}_g$, their arrival periods by $t_{g,1}^A < \ldots < t_{g,M_g}^A$, and the time these posts are reviewed by $t_{g,1}^R, \ldots, t_{g,M_g}^R$, where we denote $t_{g,i}^R = T + i$ if the $i-$th post is not reviewed by period T.

Denoting $\hat{\boldsymbol{\phi}}_{g,i} = \boldsymbol{\phi}_{k(t_{g,i}^A)}$, which is the feature vector of the $i-$ arrival in group g ($k(t_{g,i}^A)$ is its type), the estimation error of the $i-$th post on sequence $\mathcal{J}_g$ is $\|\hat{\boldsymbol{\phi}}_{g,i}\|_{\bar{\boldsymbol{V}}_{t_{g,i}^A-1}^{-1}}$. Akin to the matrix $\bar{\boldsymbol{V}}_t$ in (13) but only with the first i posts in $\mathcal{J}_g$, we define $\hat{\boldsymbol{V}}_{g,i} = \kappa \boldsymbol{I} + \sum_{i'=1}^{i} \hat{\boldsymbol{\phi}}_{g,i'}\hat{\boldsymbol{\phi}}_{g,i'}^{\intercal}$. Our first result connects the estimation error of a post in a setting with queueing delayed feedback, to a setting where the feedback delay is fixed to $Q_{\max} = 2\beta r_{\max}\ell_{\max}$, the maximum number of group-g posts in the review queue (Lemma 5.6).

Lemma D.5. *For any group g and $i \geq Q_{\max}$, the estimation error is $\|\hat{\boldsymbol{\phi}}_{g,i}\|_{\bar{\boldsymbol{V}}_{t_{g,i}^A-1}^{-1}} \leq \|\hat{\boldsymbol{\phi}}_{g,i}\|_{\hat{\boldsymbol{V}}_{g,i-Q_{\max}}}$.*

Proof. Fix a group g. Since COLBACID follows a first-come-first-serve scheduling rule for posts of the same group in the review queue (Line 10 in Algorithm 3), we have $t_{g,1}^R < \ldots < t_{g,M_g}^R$. Since $Q_{\max}$ upper bounds the number of group-g posts in the review queue by Lemma 5.6, for the $i-$th post with $i \geq Q_{\max}$, the first $(i-Q_{\max})$ posts must have already been reviewed before the admission of post i. Otherwise there would be $Q_{\max}+1$ group-g posts in the review system when we admit post i, which contradicts Lemma 5.6. As a result, when we admit the $i-$th post, the first $i - Q_{\max}$ posts in $\mathcal{J}_g$ are already reviewed by the first-come-first-serve scheduling and are all accounted in $\bar{\boldsymbol{V}}_{t_{g,i}^A-1}$. By Lemma D.2, we have $\|\hat{\boldsymbol{\phi}}_{g,i}\|_{\bar{\boldsymbol{V}}_{t_{g,i}^A-1}^{-1}} \leq \|\hat{\boldsymbol{\phi}}_{g,i}\|_{\hat{\boldsymbol{V}}_{g,i-Q_{\max}}}$. □

As a result of Lemma D.5, the estimation error of posts in sequence $\mathcal{J}_g$ behaves as if having a fixed feedback delay. This allows the use of Lemma 5.12 to bound their total estimation error. Aggregating the error across groups gives the following result.

Lemma D.6. *The total estimation error of posts admited by the optimistic admission rule is bounded by $3GQ_{\max}d\ln(1+T/d) + \sqrt{2GTd\ln(1+T/d)}$.*

Proof. The estimation error can be decomposed into groups by

$$\sum_{t=1}^{T}\sum_{k}\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}A_k(t)=\sum_{g\in\mathcal{G}}\sum_{i=1}^{M_g}\|\hat{\boldsymbol{\phi}}_{g,i}\|_{\bar{\boldsymbol{V}}_{t^A_{g,i}-1}^{-1}}\le\sum_{g\in\mathcal{G}}\sum_{i=1}^{Q_{\max}}\|\hat{\boldsymbol{\phi}}_{g,i}\|_{(\kappa\boldsymbol{I})^{-1}}+\sum_{g\in\mathcal{G}}\sum_{i=Q_{\max}}^{M_g}\|\hat{\boldsymbol{\phi}}_{g,i}\|_{\bar{\boldsymbol{V}}_{t^A_{g,i}-1}^{-1}}$$

$$\le GQ_{\max}+\sum_{g\in\mathcal{G}}\sum_{i=Q_{\max}}^{M_g}\|\hat{\boldsymbol{\phi}}_{g,i}\|_{\hat{\boldsymbol{V}}_{g,i-Q_{\max}}}$$

(By $\kappa \ge U^2$ and Lemma D.5)

$$\le GQ_{\max}+\sum_{g\in\mathcal{G}}\left(\sqrt{2M_g d\ln(1+T/d)}+2Q_{\max}d\ln(1+T/d)\right)\quad\text{(By Lemma 5.12)}$$

$$\le 3GQ_{\max}d\ln(1+T/d)+\sqrt{2GTd\ln(1+T/d)}$$

where the last inequality is by $\sum_{g\in\mathcal{G}} M_g \le T$ and Cauchy-Schwarz Inequality. □

Proof of Lemma 5.13. Recall that $\bar{A}_k(t)=\Lambda_k(t)\mathbb{1}\left(\beta\bar{r}_k(t)\ell_k\ge\tilde{Q}_{g(k)}(t)\right)$ captures whether the optimistic admission would have admitted a type-k post, $E_k(t)$ captures whether the label-driven admission admits a type-k post, and $A_k(t)=\bar{A}_k(t)(1-E_k(t))$ captures whether the optimistic admission ends up admitting a type-k post into the review queue. We finish the proof by combining Lemma D.6 with the impact of label-driven admissions:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k}\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}\bar{A}_k(t)\right]\le\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k}\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}(A_k(t)+E_k(t))\right]\quad(A_k(t)=\bar{A}_k(t)(1-E_k(t)))$$

$$\le\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k}\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}A_k(t)\right]+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}}E_k(t)\right]\quad(\|\boldsymbol{\phi}_k\|_{\bar{\boldsymbol{V}}_{t-1}^{-1}}\le 1)$$

$$\le 3GQ_{\max}d\ln(1+T/d)+\sqrt{2GTd\ln(1+T/d)}+\frac{9B_\delta^2(T)d\ln(1+T/d)}{\max(\eta,\gamma)^2}$$

(By Lemmas 5.5 and D.6)

$$\le d\ln(1+T/d)\left(3GQ_{\max}+\frac{9B_\delta^2(T)}{\max(\eta,\gamma)^2}\right)+\sqrt{2GTd\ln(1+T/d)}.$$

□

E Some Useful Known Results

E.1 Concentration Inequality

We frequently use the following Hoeffding's Inequality [BLM13].

Fact 1 (Hoeffding's Inequality)**.** *Given N independent random variables X_n taking values in $[0,1]$ almost surely. Let $X=\sum_{n=1}^{N}X_n$. Then for any $x>0$,*

$$\mathbb{P}\{X-\mathbb{E}[X]>x\}\le e^{-2x^2/N},\ \mathbb{P}\{X-\mathbb{E}[X]<-x\}\le e^{-2x^2/N}.$$

We also use the Chernoff bound for normal random variables.

Fact 2 (Chernoff Bound). *Given N i.i.d. random variables X_i that are normally distributed with variance σ^2. Let $X = \sum_{n=1}^{N} X_n$. Then for any $x > 0$,*

$$\mathbb{P}\{X - \mathbb{E}[X] > x\} \le e^{-x^2/(2N\sigma^2)},\ \mathbb{P}\{X - \mathbb{E}[X] < -x\} \le e^{-x^2/(2N\sigma^2)}.$$

Proof. We know that X is normally distributed with mean $\mathbb{E}[X]$ and variance $N\sigma^2$. The result then follows from a Chernoff bound on X; see [BLM13, page 22]. □

E.2 Facts on Matrix Norm

Our results rely on a self-normalized tail inequality derived in [APS11] which we restate here. Let us consider a unknown $\boldsymbol{\theta}^\star \in \mathbb{R}^d$, an arbitrary sequence $\{\boldsymbol{X}_t\}_{t=1}^\infty$ with $\boldsymbol{X}_t \in \mathbb{R}^d$, a real-valued sequence $\{\eta_t\}_{t=1}^\infty$, and a sequence $\{Z_t\}_{t=1}^\infty$ with $Z_t = \boldsymbol{X}_t^\intercal \boldsymbol{\theta}^\star + \eta_t$. We define the $\sigma-$algebra $\mathcal{F}_t = \sigma(\boldsymbol{X}_1, \dots, \boldsymbol{X}_{t+1}, \eta_1, \dots, \eta_t)$. In addition, recall the definition of $\hat{\boldsymbol{\theta}}_t$ as the solution to the ridge regression with regularizer κ and $\bar{V}_t$ in (13). Then we have the following result implied by the second result in [APS11, Theorem 2].

Fact 3. *Assume that $\|\boldsymbol{\theta}^\star\|_2,\ \|\boldsymbol{X}_t\|_2 \le U$ and that η_t is conditionally $R-$sub-Gaussian for some $R \ge 0$ such that $\forall u \in \mathbb{R}$, $\mathbb{E}[e^{u\eta_t} \mid \mathcal{F}_{t-1}] \le \exp(u^2R^2/2)$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $t \ge 0$, $\boldsymbol{\theta}_\star$ lies in the set*

$$\mathcal{C}_t = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \colon \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|_{\bar{V}_t} \le R\sqrt{d \ln\left(\frac{1 + tU^2/\kappa}{\delta}\right)} + \sqrt{\kappa}U \right\}.$$

The following result is a restatement of a result in [APS11, Lemma 11].

Fact 4. *Let $\{\boldsymbol{X}_t\}$ be a sequence in $\mathbb{R}^d$, $\boldsymbol{V}$ a $d \times d$ positive definite matrix and define $\boldsymbol{V}_t = \boldsymbol{V} + \sum_{\tau=1}^{t} \boldsymbol{X}_t \boldsymbol{X}_t^\intercal$. If $\|\boldsymbol{X}_t\|_2 \le U$ for all t, and $\lambda_{\min}(\boldsymbol{V}) \ge \max(1, U^2)$, then $\sum_{t=1}^{n} \|\boldsymbol{X}_t\|^2_{\boldsymbol{V}_{t-1}^{-1}} \le 2\ln\frac{\det(\boldsymbol{V}_n)}{\det(\boldsymbol{V})}$ for any n.*

We also use a determinant-trace inequality from [APS11, Lemma 10].

Fact 5. *Suppose $\boldsymbol{X}_1, \dots, \boldsymbol{X}_t \in \mathbb{R}^d$ and for any $\tau \le t$, $\|\boldsymbol{X}_\tau\|_2 \le U$. Let $\boldsymbol{V}_t = \kappa \boldsymbol{I} + \sum_{\tau=1}^{t} \boldsymbol{X}_\tau \boldsymbol{X}_\tau^\intercal$ for some $\kappa > 0$. Then $\det(\boldsymbol{V}_t) \le (\kappa + tU^2/d)^d$.*

E.3 Additional Analytical Facts

Fact 6. *For any x, we have $\varphi'(x) - x\varphi(x)$ and $|\varphi'(x)| \le 1/4$.*

Proof. By definition, $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and thus $\varphi'(x) = \frac{-x}{\sqrt{2\pi}} e^{-x^2/2} = -x\varphi(x)$. To bound $|\varphi'(x)|$, it suffices to bound it for $x > 0$ due to symmetry. Let $f(x) = xe^{-x^2/2}$. We have $f'(x) = (1-x^2)e^{-x^2/2}$. Then the only $x \ge 0$ with $f'(x) = 0$ is $x = 1$. We know $f(0) = 0$, $f(1) = e^{-1/2}$ and $f(+\infty) = 0$. As a result, the maximum value of $f(x)$ is $e^{-1/2}$, which shows $|\varphi'(x)| \le \frac{1}{\sqrt{2\pi}} e^{-1/2} \le \frac{1}{4}$. □

Fact 7. *We have that $r_k^O(h)$ is increasing for $h \le 0$ and $r_k^R(h)$ is decreasing for $h \ge 0$. Further,for any t and k, we have $\bar{r}_k(t) \le v_k \sigma_k$.*

Proof. Since $r_k^O(h) = r_k^R(-h)$ for $h < 0$ by (5), we only need to show $r_k^R(h)$ is decreasing for $h \geq 0$. We know that $r_k^R(h) = v_k(-h\Phi(-h/\sigma_k) + \sigma_k\varphi(-h/\sigma_k))$. Then its derivative is given by

$$v_k\left(-\Phi(-h/\sigma_k) + \frac{h}{\sigma_k}\varphi(-h/\sigma_k) - \frac{h}{\sigma_k}\varphi(-h/\sigma_k)\right) = -v_k\Phi(-h/\sigma_k) < 0.$$

As a result, $r_k^R(h)$ is decreasing for $h \geq 0$, and is maximized at $h = 0$, which also shows $r_k^O(h)$ is increasing for $h < 0$.

By definition, $\bar{r}_k(t) = \max_{h\in[\underline{h}_k(t),\bar{h}_k(t)]} r_k(h) \leq \max_h r_k(h) = r_k(0) = r_k^R(0)$. We finish the proof by noting that $r_k^R(0) = v_k\sigma_k$. □

Fact 8. *For $t \geq 100$, we have $t/4 \geq \sqrt{t\ln t}$.*

Proof. It suffices to show $t/16 \geq \ln t$ for $t \geq 100$. Let $f(t) = t/16 - \ln(t)$. Then $f'(t) = 1/16 - 1/t \geq 0$ for $t \geq 16$. Thus $f(t)$ increases for $t \geq 16$. We prove the desired result by noting $f(100) \geq 0$. □

Fact 9. *We have $x\ln(1+3/x) \geq 1$ for any $x \geq 1$.*

Proof. Leg $f(x) = x\ln(1+3/x)$. Then $f'(x) = \ln(1+3/x) - \frac{3}{x+3}$ and $f''(x) = \frac{-3}{x(x+3)} + \frac{3}{(x+3)^2)} \leq 0$. Therefore, $f'(x)$ is a decreasing function with $f'(+\infty) = 0$, and $f'(x) \geq 0$. This shows $f(x)$ is an increasing function. Since $f(1) = \ln(4) \geq 1$, we have $f(x) \geq 1$ for any $x \geq 1$. □