ONLINE CONFORMAL PREDICTION WITH ADVERSARIAL SEMI-BANDIT FEEDBACK VIA REGRET MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncertainty quantification is crucial in safety-critical systems, where decisions must be made under uncertainty. In particular, we consider the problem of online uncertainty quantification, where data points arrive sequentially. Online conformal prediction is a principled online uncertainty quantification method that dynamically constructs a prediction set at each time step. While existing methods for online conformal prediction provide long-run coverage guarantees without any distributional assumptions, they typically assume a *full-feedback* setting in which the true label is always observed. In this paper, we propose a novel learning method for online conformal prediction with *partial feedback* from an adaptive adversary—a more challenging setup where the true label is revealed only when it lies inside the constructed prediction set. Specifically, we formulate online conformal prediction as an adversarial bandit problem by treating each candidate prediction set as an arm. Building on an existing algorithm for adversarial bandits, our method achieves a long-run coverage guarantee by explicitly establishing its connection to the regret of the learner. Finally, we empirically demonstrate the effectiveness of our method in both independent and identically distributed (i.i.d.) and non-i.i.d. settings, showing that it successfully controls the miscoverage rate while maintaining a reasonable size of the prediction set.

1 INTRODUCTION

Uncertainty quantification is essential in safety-critical domains such as autonomous driving (Lindemann et al., 2023), healthcare (Lin et al., 2022), and finance (Park & Cho, 2025), where uncertainty-aware decision making is required. Unlike point prediction methods that return the most likely outcome, conformal prediction (Vovk et al., 2005) is a promising uncertainty quantification method that constructs a *conformal set* for a given input, a set of outcomes that is guaranteed to contain the true label with a user-specified probability. We refer to the guarantee as a *coverage guarantee*. Here, the size of the conformal set quantifies the uncertainty in terms of making a prediction.

Moreover, the coverage guarantee is model-agnostic in the sense that the guarantee holds irrespective of the choice of the prediction model. Exchangeability assumption on the data generating process (Vovk et al., 2005) is the only requirement for the guarantee, where a typical independent and identically distributed (i.i.d.) scenario is the case that satisfies such assumption. Specifically, under the exchangeability assumption, the coverage guarantee of the conformal set constructed from training samples holds for an unseen test sample (Vovk, 2013). Therefore, since the coverage guarantee holds for arbitrary prediction models, conformal prediction has been applied to complex and large-scale models such as large language models (Mohri & Hashimoto, 2024; Cherian et al., 2024; Lee et al., 2024).

However, the exchangeability assumption is easily violated under scenarios such as distribution shift, where the training and test distributions differ. A number of conformal prediction methods have been proposed to provide coverage guarantees under such settings (Tibshirani et al., 2019; Podkopaev & Ramdas, 2021; Park et al., 2022; Gendler et al., 2022; Si et al., 2024). In contrast to the aforementioned batch conformal prediction methods, which require a batch of samples for training, methods for online conformal prediction are proposed to tackle online uncertainty quantification problems, where data points arrive sequentially (Gibbs & Candès, 2021; Bastani et al., 2022; Angelopoulos et al., 2023; 2024a). Even in adversarial settings, where no distributional assumptions are made on the data stream or on the functional form of the scoring functions, these methods provide a long-run coverage guarantee such that the empirical coverage reaches the target level after sufficiently many time steps.

Meanwhile, existing methods for online conformal prediction typically assume a *full feedback* scenario, where the true label is revealed every time step. Indeed, these methods are tailored to the full feedback setting, since they require a scoring function evaluated on the true label either for its quantile estimation or for the evaluation of the miscoverage loss over multiple conformal set candidates. Recently, Ge et al. (2025) proposed a method for online conformal prediction with *partial feedback*, specifically feedback referred to as *semi-bandit feedback*, where the true label is revealed only when it lies within the chosen conformal set. While it is a more challenging learning setting compared to the full feedback scenario, their coverage guarantee holds only under i.i.d. data streams.

In this paper, we present the first study of online conformal prediction with adversarial partial feedback. Specifically, by discretizing the continuous hypothesis space of thresholds that parameterize a conformal set, and then treating each candidate conformal set as an arm, we formulate the problem as an adversarial bandit problem. A bandit problem is a sequential game between a learner and an environment. In each round, the learner chooses an arm, and the environment provides feedback on the chosen arm. In particular, the adversarial bandit problem removes almost all assumptions about how the environment provides feedback, where the environment is often called the adversary accordingly. The performance of the learner is evaluated based on the regret, which typically quantifies how well the learner performs with respect to the best arm in hindsight (Bubeck et al., 2012; Lattimore & Szepesvári, 2020). There is a rich literature on adversarial bandit problems, devising algorithms with sublinear regret. `EXP3.P` algorithm Auer et al. (2002) is one of the algorithms that provides a high-probability sublinear regret even under the adaptive adversary, an adversary that can generate feedback based on the previous history.

Building on the `EXP3.P` algorithm, we propose a novel algorithm for online conformal prediction with adversarial partial feedback, in which we consider a semi-bandit feedback scenario, similar to Ge et al. (2025). Specifically, by devising a loss function tailored to conformal prediction, we explicitly establish a connection between the regret of the learner and a long-run coverage guarantee (Lemma 1), which in turn provides a long-run coverage guarantee of our algorithm. We further improve the performance in terms of the speed approaching the target coverage, by fully exploiting the monotonicity property of the miscoverage loss with respect to the threshold parameterizing a conformal set. Specifically, it enables partial inference of feedback from candidate conformal sets that are not chosen, even when the true label is unavailable. We empirically demonstrate the efficacy of our method on both classification and regression tasks, conducting experiments in i.i.d. and non-i.i.d. settings for each task. In particular, we show that our method approaches long-run coverage while maintaining a moderate average conformal set size, achieving performance comparable to Bastani et al. (2022), an online conformal prediction method with adversarial full feedback.

2 Related Work

2.1 Online Conformal Prediction

Gibbs & Candès (2021) first proposed an online conformal prediction method for arbitrary data streams. Based on online gradient descent, their method provides a long-run coverage guarantee over arbitrary sequences. While the method relies only on a single step size parameter, the optimal parameter requires knowledge of the degree of distribution shift, which is an unrealistic assumption. The same authors have resolved the issue by aggregating results from multiple experts, running in parallel with different step sizes, making the method adaptive to the type of distribution shift in a data-driven manner Gibbs & Candès (2024). While providing a biased result in terms of the long-run coverage, they provide a local coverage guarantee over all time intervals of a given width, under mild assumptions on the smoothness of the distribution of the scoring function and its quantile estimates. Building on the strongly adaptive online learning method, Bhatnagar et al. (2023) further improved the method by providing a simultaneous coverage guarantee over all local intervals of arbitrary window size. Unlike Gibbs & Candès (2024), they considered a dynamic set of experts, where each expert is active only for a specific period of time. Inspired by control theory, Angelopoulos et al. (2023) extended existing online gradient descent-based methods by incorporating online gradient descent steps, which they refer to as quantile tracking, as one of the components for the online quantile update.

Besides, there have been works to provide stronger theoretical guarantees. Bastani et al. (2022) proposed a method with a threshold-calibrated multivalid coverage guarantee, a group- and threshold-conditional coverage guarantee where a set of groups can be arbitrarily defined. Angelopoulos et al. (2024b) proposed a simple online gradient-descent method that has simultaneous guarantees both on the adversarial and i.i.d. settings. Recently, Zhang et al. (2025) devised an online conformal prediction algorithm, providing both privacy and coverage guarantees under arbitrary data streams.

In this paper, we also consider an online conformal prediction problem under an arbitrary data stream. Specifically, we consider an adaptive adversary that can generate data based on the learner's past actions.

2.2 Online Conformal Prediction with Partial Feedback

Existing methods for online conformal prediction with adversarial feedback typically assume the full feedback setting, where the true label is revealed at every time step. One exceptional case is Angelopoulos et al. (2024b), where the algorithm itself only requires the feedback on whether a chosen conformal set contains the true label. However, the authors are basically considering a full feedback scenario, and some of their theoretical results assume a problem setup where the scoring function is trained online, using the labeled data pairs from previous time steps.

On the other hand, there have been few papers addressing online conformal prediction with partial feedback, a scenario where access to the true label is limited. Wang & Qiao (2024) considered a bandit feedback scenario, where the true label is observed only when the predicted label corresponds to the true label. Recently, Ge et al. (2025) proposed a method under a semi-bandit feedback scheme, a less rigid partial feedback scenario where the true label is revealed as the true label lies within a chosen conformal set. Although partial feedback is inherently more challenging than full feedback, prior works still rely on the i.i.d. data-generating assumption, which restricts their applicability to real-world, non-i.i.d. data streams.

As such, we consider an online conformal prediction with adversarial partial feedback, where data streams deviate from the i.i.d. process and at the same time the true label is difficult to obtain.

3 Online Conformal Prediction with Adversarial Feedback

We consider online conformal prediction with adversarial partial feedback. Let $\mathcal{X}$ be a set of examples and $\mathcal{Y}$ be a set of labels. At each time step $t \in [T]$, a learner chooses a conformal set $\hat{C}_{\pi_t} : \mathcal{X} \to 2^{\mathcal{Y}}$, which is parameterized by the threshold parameter $\pi_t \in [0, 1]$ as follows:

$$\hat{C}_{\pi_t}(x) \coloneqq \{\tilde{y} \in \mathcal{Y} \mid f_t(x, \tilde{y}) \geq \pi_t\}.$$

Here, $f_t : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ is a scoring function that measures the conformity of a label for a given input. Note that the functional form of $f_t(\cdot, \cdot)$ may evolve over time.

In conformal set learning, we consider the following standard learning protocol from adversarial bandits. Specifically, an example $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ is chosen by an adversary, where only the input x_t is revealed to a learner. We assume that the adversary can be adaptive, in the sense that it may select a sample based on the learner's previous decisions. Once x_t is given, the learner outputs a conformal set $\hat{C}_{\pi_t}(x_t)$, where a threshold parameter π_t is chosen by the learner's current strategy. The learner's strategy can either be stochastic or deterministic, which can be updated online based on either the learner's previous interactions with the adversary or x_t. Then, the learner receives feedback, chosen before the learner's choice of a conformal set, from the adversary on whether $\hat{C}_{\pi_t}(x_t)$ contains the true label y_t, which we denote as $m_t(\pi_t) \coloneqq \mathbb{1}(y_t \notin \hat{C}_{\pi_t}(x_t))$, called *miscoverage* henceforth.

Here, we consider a *semi-bandit feedback* scenario (Ge et al., 2025), one of the partial feedback settings where a true label y_t is additionally revealed when $m_t(\pi_t) = 0$. Having access to y_t enables the learner to evaluate the miscoverage $m_t(\pi)$ for all $\pi \in [0, 1]$, since the scoring function $f_t(x_t, y_t)$ of the true label—a quantity sufficient to evaluate the miscoverage of a threshold-parameterized conformal set—can be computed. Although we have coined the term partial feedback, in contrast to full feedback, to encompass both bandit and semi-bandit feedback settings, we will use it to refer exclusively to the semi-bandit feedback scenario in the following sections for simplicity.

Under such online conformal prediction problems with adversarial partial feedback, our goal is to design a learner that provides a long-run coverage guarantee by controlling the miscoverage rate defined as follows:

$$\mathbf{MC}(T) \coloneqq \frac{1}{T}\sum_{t=1}^{T} m_t(\pi_t), \tag{1}$$

where T is the time horizon. Specifically, given a target miscoverage level $\alpha \in (0, 0.5)$, we aim to upper bound the miscoverage rate as $\mathbf{MC}(T) \leq \alpha + \varepsilon(T)$ such that $\varepsilon(T) \to 0$ as $T \to \infty$. As a trivial conformal set achieves this goal, our secondary goal is to minimize inefficiency, also called conformal set size, $\mathbf{Ineff}(T) \coloneqq \frac{1}{T}\sum_{t=1}^{T} S(\hat{C}_{\pi_t}(x_t))$ for some size metric S.

4 Online Conformal Prediction as Adversarial Bandit Problem

We formulate the online conformal prediction problem with adversarial partial feedback as a multi-armed adversarial bandit problem, by treating each candidate conformal set as an arm (Section 4.1). Defining a finely discretized subset Π of the continuous hypothesis space $[0, 1]$ as an action space, we leverage the `EXP3.P` algorithm (Auer et al., 2002), an algorithm that provides a sublinear regret under adversarial bandit environments. It is a modified version of `EXP3` to encompass both non-adaptive and adaptive adversary settings.

To this end, we first design a loss function tailored to conformal prediction (Section 4.2), which in turn provides an explicit learner-agnostic relationship between a regret from a learner and its miscoverage rate $\mathbf{MC}(T)$ (Section 4.3). This relationship ensures the long-run coverage to achieve the target level $1 - \alpha$, for any learner that achieves a sublinear regret. However, directly applying an existing learner, *e.g.,* `EXP3.P`, does not make full use of the available information under the semi-bandit feedback setting, since we can evaluate $m_t(\pi)$ for all $\pi \in \Pi$ when $m_t(\pi_t) = 0$. Therefore, we further improve the algorithm by fully exploiting such additional information and the monotonicity property of a threshold-parameterized conformal set with respect to the miscoverage (Section 4.4).

4.1 Problem Reformulation

We begin by reformulating the online conformal prediction problem with adversarial partial feedback as an adversarial multi-armed bandit problem, specifying both the interaction protocol and the performance metric. For each time t, (1) the learner chooses an arm $\pi_t \in \Pi$, where π_t is drawn from its current arm selection strategy p_t, (2) the adversary simultaneously chooses a loss function $\ell_t : \Pi \to [0, 1]$, and (3) the learner observes the feedback $\ell_t(\pi_t)$ on its chosen arm and uses it to update its current strategy p_t. Here, we consider an adaptive adversary who leverages the learner's previous interaction history, *i.e.,* $(\pi_1, \ell_1(\pi_1)), \dots, (\pi_{t-1}, \ell_{t-1}(\pi_{t-1}))$, to choose the loss function ℓ_t.

We reduce the online conformal prediction problem with partial feedback to this adversarial bandit formulation (see Table 1). In our setting, each arm corresponds to a conformal threshold that indexes a prediction set, and the adversary plays the role of an adaptive data-generating process that induces a loss vector over these thresholds.

Table 1: Mapping between online conformal prediction and adversarial bandits.

	Online conformal prediction with adversarial feedback	Bandit
Option	Conformal set parameter: π_t	Arm: π_t
Feedback	Miscoverage: $m_t(\pi_t)$ True label: y_t if $m_t(\pi_t) = 0$	Loss: $\ell_t(\pi_t)$
Metric	Miscoverage rate: $\mathbf{MC}(T)$	Regret: $\mathbf{Reg}(T)$

We restrict the arm set Π to a finite collection of K candidate thresholds obtained by uniformly discretizing the score range, that is, a uniform grid on $[0, 1]$. This choice reflects the adversarial setting: since the sequence of scores may be chosen adaptively and need not obey any fixed distribution, we avoid data–dependent thresholds and instead work with a fixed, distribution-free grid.

Within this formulation, the performance of the learner is measured by its regret against the best fixed arm in hindsight,

$$\textbf{Reg}(T) \coloneqq \sum_{t=1}^{T} \ell_t(\pi_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(\pi), \tag{2}$$

which quantifies the excess cumulative loss incurred by the learner relative to the best static threshold. In the following subsections, we design a loss function $\ell_t(\pi)$ tailored to online conformal prediction and show how sublinear regret bounds for bandit algorithms translate into coverage guarantees for the resulting conformal predictor.

4.2 Design of the Loss Function

To connect the feedback in online conformal prediction to the loss-based feedback in adversarial bandits, we design a bandit loss for each threshold that summarizes the observable miscoverage information $m_t(\pi)$ into a single scalar signal. Concretely, for each threshold $\pi \in \Pi$ we fix a constant $c \in (0, 0.5)$ and a trade-off parameter $\lambda' > 0$, and define the loss

$$\ell_t(\pi, c) \coloneqq d_t(\pi, c) + \lambda' a_t(\pi) \in [\ell_{\min}, \ell_{\max}], \tag{3}$$

Miscoverage loss. The term $d_t(\pi, c) \in [0, 1]$ is the *miscoverage loss*, which depends on the miscoverage $m_t(\pi)$. We define

$$d_t(\pi, c) \coloneqq \big|m_t(\pi) - c\big|.$$

This quantity measures how far the miscoverage $m_t(\pi)$ is from the scalar c. Because $m_t(\pi) \in \{0, 1\}$ and $c \in (0, 0.5)$, the loss $d_t(\pi, c)$ equals c on coverage rounds ($m_t(\pi) = 0$) and $1 - c$ on miscoverage rounds ($m_t(\pi) = 1$), with $c < 1 - c$. Thus, miscoverage always incurs a strictly larger penalty than coverage, providing a simple mechanism that distinguishes between the two cases.

Inefficiency loss. The term $a_t(\pi) \in [0, 1]$ is an *inefficiency loss* that regularizes the size of the conformal set $\hat{C}_\pi(x_t)$. It is designed so that, on coverage rounds, it penalizes unnecessarily large sets, while on miscoverage rounds, it encourages enlarging the set, thereby preferring thresholds that are more likely to correct miscoverage. For the regret–coverage conversion in Section 4.3, we only require $a_t(\pi)$ to be bounded and to satisfy this qualitative dependence on the set size; a specific functional form will be introduced in Section 4.4 when we instantiate our `EXP3.P`-style algorithms.

4.3 Miscoverage Guarantees from Regret Bounds

Here, we connect the miscoverage rate in online conformal prediction with the regret notion in adversarial bandits, inspired by the conversion idea in selective generation (Lee et al., 2025). Using the loss $\ell_t(\pi, c)$ from Section 4.2, we show that a bound on the regret with respect to $\{\ell_t(\cdot, c)\}_{t=1}^{T}$ yields an explicit upper bound on the empirical miscoverage rate in terms of the target level α. This connection between regret and coverage is formalized in the following lemma. See Appendix C for a proof.

Lemma 1. *For any $T \in \mathbb{N}$, $\alpha \in (0, 0.5)$, and $\lambda > 0$, let $c = \frac{\alpha}{\lambda+2}$ and $\lambda' = \frac{\lambda\alpha}{\lambda+2}$. For losses ℓ_t of the form* (3) *with $d_t(\pi, c) = \big|m_t(\pi) - c\big|$, and $a_t(\pi) \in [0, 1]$, any learner with bounded regret satisfies the following empirical miscoverage guarantee:*

$$\textbf{\textit{MC}}(T) \le \alpha + \frac{1}{T}\textbf{\textit{Reg}}\left(T, \frac{\alpha}{\lambda + 2}\right).$$

This implies that if the regret is bounded by a sublinear function of T, then the excess miscoverage rate $\textbf{MC}(T) - \alpha$ is upper bounded by a vanishing term of order $\textbf{Reg}\Big(T, \frac{\alpha}{\lambda+2}\Big)/T$. In particular, any bandit algorithm that achieves sublinear regret with respect to the loss (3) can be used as a conformal set learner under our framework, regardless of whether it operates with full or partial feedback.

Among such algorithms, we adopt `EXP3.P` (Auer et al., 2002) in the adversarial bandit setting, as it is known to achieve sublinear regret against an adaptive adversary and thus, by Lemma 1, yields conformal sets whose coverage shortfall relative to the target level α is asymptotically negligible.

4.4 EXP3.P-STYLE ALGORITHMS AND THEIR REGRET BOUNDS

We first apply the adversarial bandit algorithm `EXP3.P` (Auer et al., 2002) to our setting, yielding the baseline method `EXP3.P-CP` that runs on the threshold set Π with the loss $\ell_t(\pi, c)$ from Section 4.2. By Lemma 1, this already provides coverage guarantees, but it still treats online conformal prediction as a generic bandit problem and does not exploit the additional information available under semi-bandit feedback. To leverage this structure, we further develop two strengthened variants, `EXP3.P-CP-SEMI` and `EXP3.P-CP-UNLOCK`, which reuse unlocked feedback across thresholds by exploiting conformal monotonicity and pseudo-gain constructions, respectively. All three bandit-based conformal learners in this subsection optimize the same loss $\ell_t(\pi, c)$; their differences lie solely in how they construct gain estimates from the available (partial) feedback.

To make this loss concrete, we now specify the inefficiency term $a_t(\pi)$ used in all of our bandit-based conformal learners. We set

$$a_t(\pi) \coloneqq \mathbb{1}\big(m_t(\pi) = 0\big)\ \exp\left(-\frac{\pi}{o(T)}\right) + \mathbb{1}\big(m_t(\pi) = 1\big)\ \exp\left(-\frac{1-\pi}{o(T)}\right),$$

where $o(T)$ is a positive function of the horizon T. In our analysis we choose $o(T) = \sqrt{T}$, but more generally any $o(T) = T^k$ with $k \in [0.5, 1)$ and $o(T)/T \to 0$ as $T \to \infty$ suffices for our regret bounds.

This choice ensures $a_t(\pi) \in [0, 1]$ and has the following effect: on coverage rounds ($m_t(\pi) = 0$), $a_t(\pi)$ decreases in π, so larger thresholds—corresponding to smaller prediction sets $\hat{C}_\pi(x_t)$—are preferred, whereas on miscoverage rounds ($m_t(\pi) = 1$), $a_t(\pi)$ decreases in $1 - \pi$, so smaller thresholds—corresponding to larger sets—are favored to correct miscoverage.

The miscoverage term $d_t(\pi, c)$ in $\ell_t(\pi, c)$ creates a fixed penalty gap between coverage ($m_t(\pi) = 0$) and miscoverage ($m_t(\pi) = 1$) and ensures that miscoverage is penalized more heavily overall, while the inefficiency term $a_t(\pi)$ only adjusts the set size within each miscoverage level.

`EXP3.P-CP`. Using our bandit reformulation together with the loss (3) and its miscoverage and inefficiency losses $d_t(\pi, c)$ and $a_t(\pi)$, we first apply the classical `EXP3.P` algorithm (Auer et al., 2002) directly to online conformal set learning.

As established in Theorem 2, the `EXP3.P` learner (Algorithm 2) achieves a regret bound of $\mathbf{Reg}(T) \leq \mathcal{O}\Big(\sqrt{\frac{TK}{\ln K}}\ln(\delta^{-1}) + 5.15\sqrt{TK \ln K}\Big)$ with probability at least $1 - \delta$, where $\delta \in (0, 1)$ is the confidence parameter. We obtain `EXP3.P-CP` (Algorithm 3) by running `EXP3.P` on the threshold set Π with the loss function (3). By Lemma 1, the resulting learner enjoys a corresponding high-probability bound on the miscoverage rate.

However, `EXP3.P-CP` still treats conformal set learning as a generic adversarial bandit and does not exploit conformal-specific structure (*e.g.,* semi-bandit feedback or the characteristics of conformal prediction), so corrections to coverage rely solely on bandit feedback, and in practice we observe that the empirical coverage moves toward the target level $1 - \alpha$ noticeably more slowly.

`EXP3.P-CP-SEMI`. Unlike `EXP3.P-CP`, which only uses the bandit feedback on the chosen arm, `EXP3.P-CP-SEMI` explicitly exploits the semi-bandit feedback available in our setting: when the constructed conformal set $\hat{C}_{\pi_t}(x_t)$ covers the true label ($m_t(\pi_t) = 0$), we additionally observe y_t, whereas when $m_t(\pi_t) = 1$ we only observe the binary coverage indicator for the chosen arm π_t. To take advantage of this information, the algorithm introduces an *unlocking mechanism*, originally proposed in the context of selective generation (Lee et al., 2025), that leverages the monotonicity of conformal miscoverage in π and proceeds differently depending on whether $m_t(\pi_t) = 0$ or $m_t(\pi_t) = 1$.

We first define the *coverage-consistent* subset $\Pi_t^* \coloneqq \{\pi \in \Pi : \pi \leq f_t(x_t, y_t)\}$, which consists of all thresholds whose induced conformal sets include the true label y_t. In the semi-bandit setup, the label y_t is observed exactly when $m_t(\pi_t) = 0$, so Π_t^* is implementable on such rounds. The unlocking set $\Pi_t(\pi_t)$ is then defined by

$$\Pi_t(\pi_t) \coloneqq \begin{cases} \Pi_t^* & \text{if } m_t(\pi_t) = 0 \\ \{\pi \in \Pi : \pi \geq \pi_t\} & \text{if } m_t(\pi_t) = 1 \end{cases}.$$

This construction follows from the monotonicity of conformal prediction: larger thresholds produce smaller conformal sets, so $\hat{C}_{\pi_1}(x_t) \supseteq \hat{C}_{\pi_2}(x_t)$ whenever $\pi_1 \leq \pi_2$. Consequently, when $m_t(\pi_t) = 0$, all thresholds $\pi \leq f_t(x_t, y_t)$ also satisfy $m_t(\pi) = 0$, whereas when $m_t(\pi_t) = 1$, all larger thresholds $\pi \geq \pi_t$ incur the same miscoverage $m_t(\pi) = 1$.

Under semi-bandit feedback, we use the following biased gain estimator with unlocking:

$$\tilde{g}_t\left(\pi \mid \Pi_t(\pi_t)\right) \coloneqq \mathbb{1}(\pi \in \Pi_t(\pi_t)) \left\{ \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t(\pi_t)} p_t(\tilde{\pi})} + \frac{\beta}{p_t(\pi)} \right\} + \mathbb{1}(\pi \notin \Pi_t(\pi_t)) \frac{\beta}{p_t(\pi)}. \tag{4}$$

Here, p_t is a probability distribution over the K candidate thresholds, so that $p_t(\pi) \in [0, 1]$ and $\sum_{\pi \in \Pi} p_t(\pi) = 1$. If we disable unlocking by setting $\Pi_t(\pi_t) = \{\pi_t\}$, this estimator reduces to the standard `EXP3.P` gain estimator (6) applied to the loss $\ell_t(\pi, c)$. Hence `EXP3.P-CP-SEMI` (Algorithm 4) reduces to `EXP3.P-CP` when $\Pi_t(\pi_t) = \{\pi_t\}$.

`EXP3.P-CP-SEMI` (Algorithm 1) combines the loss (3) with the unlocking estimator (4) to reuse feedback across thresholds whenever monotonicity allows it. As established in Theorem 4, this algorithm achieves a high-probability regret bound of $\textbf{Reg}(T) \leq \mathcal{O}(5.15\sqrt{K \ln KT})$, so that the cumulative regret grows sublinearly with T and the learner's performance remains close to that of the best fixed threshold in hindsight. In practice, `EXP3.P-CP-SEMI` adjusts the empirical coverage toward the target level $1 - \alpha$ more quickly than `EXP3.P-CP`, but it still treats all thresholds within the unlocking set $\Pi_t(\pi_t)$ symmetrically and does not fully exploit the ordering induced by conformal monotonicity; this motivates the pseudo-gain variant described next.

`EXP3.P-CP-UNLOCK`. Like `EXP3.P-CP-SEMI`, `EXP3.P-CP-UNLOCK` operates under the same semi-bandit feedback, but it is more tightly aligned with the conformal structure. It sharpens the unlocking rule and modifies the gain estimator so that, inside the unlocked region, more desirable thresholds (in terms of set size and coverage correction) receive larger estimated gains.

First, we redefine the unlocking set $\Pi_t(\pi_t) \subset \Pi$ as

$$\Pi_t(\pi_t) \coloneqq \begin{cases} \Pi & \text{if } m_t(\pi_t) = 0 \\ \{\pi \in \Pi : \pi \geq \pi_t\} & \text{if } m_t(\pi_t) = 1 \end{cases}.$$

On coverage rounds ($m_t(\pi_t) = 0$), semi-bandit feedback reveals y_t, so $g_t(\pi)$ is evaluable for every $\pi \in \Pi$; `EXP3.P-CP-UNLOCK` therefore unlocks the entire threshold set $\Pi_t(\pi_t) = \Pi$ whenever $m_t(\pi_t) = 0$, enabling per-round updates on all arms whenever coverage occurs.

We then define the biased unlocking estimator $\tilde{g}_t(\pi \mid \Pi_t(\pi_t))$ under this semi-bandit feedback as

$$\tilde{g}_t\left(\pi \mid \Pi_t(\pi_t)\right) \coloneqq \underbrace{\mathbb{1}(m_t(\pi_t) = 0) \times (A)}_{\text{full unlocking}} + \underbrace{\mathbb{1}(m_t(\pi_t) = 1) \times (B)}_{\text{partial unlocking}}. \tag{5}$$

Recalling $\Pi_t^* \coloneqq \{\tilde{\pi} \in \Pi : \tilde{\pi} \leq f_t(x_t, y_t)\}$, the full-unlocking branch (A) is given by

$$(A) \coloneqq \mathbb{1}(\pi \in \Pi_t^*) \left\{ \frac{g_t(\pi) + \beta}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})} + \beta \right\} + \mathbb{1}(\pi \notin \Pi_t^*) \left\{ g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi} \leq \pi} p_t(\tilde{\pi})} \right\},$$

while the partial-unlocking branch (B) is

$$(B) \coloneqq \mathbb{1}(\pi \in \Pi_t(\pi_t)) \left\{ g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi} \leq \pi} p_t(\tilde{\pi})} \right\} + \underbrace{\mathbb{1}(\pi \notin \Pi_t(\pi_t)) \left\{ \tilde{g}_t(\pi) + \beta + \frac{\beta}{p_t(\pi)} \right\}}_{\text{(C)}}.$$

In the case of (C), the true gain cannot be unlocked. We therefore use a **pseudo-gain** $\tilde{g}_t(\pi)$, defined by rescaling a plug-in loss $\tilde{\ell}_t(\pi)$ into $[0, 1]$ using the known bounds $[\ell_{\min}, \ell_{\max}]$:

$$\tilde{g}_t(\pi) = \frac{\ell_{\max} - \tilde{\ell}_t(\pi, c)}{\ell_{\max} - \ell_{\min}}, \qquad \text{where} \qquad \tilde{\ell}_t(\pi, c) \coloneqq (1 - c) + \lambda' \exp\left(-\frac{1 - \pi}{o(T)} \right) \text{ from (3).}$$

where $\tilde{\ell}_t(\pi, c)$ is a plug-in surrogate for the loss $\ell_t(\pi, c)$ obtained by evaluating (3) under the miscoverage branch $m_t(\pi) = 1$. By construction $\tilde{g}_t(\pi) \in [0, 1]$, and since $\tilde{\ell}_t(\pi, c)$ is increasing in π, the pseudo-gain $\tilde{g}_t(\pi)$ is larger for smaller π, which prioritizes correcting miscoverage by favoring thresholds that expand $\hat{C}_\pi(x_t)$ on the locked side. In particular, we use the pseudo-gain only inside branch (C) of the estimator (5); the notation $\tilde{g}_t(\pi)$ there refers to this pseudo-gain, whereas $\tilde{g}_t(\pi \mid \Pi_t(\pi_t))$ denotes the overall biased estimator.

The biased unlocking estimator $\tilde{g}_t(\pi \mid \Pi_t(\pi_t))$ in (5) is designed to reflect this preference structure. When coverage information is available, thresholds whose sets cover y_t are assigned larger effective gains than those whose sets exclude y_t, and within each case (coverage or miscoverage), the estimator favors thresholds that adjust the set size in the desired direction (shrinking the set when coverage holds and expanding it when miscoverage occurs). Combining the loss (3) with this biased gain estimator yields our method `EXP3.P-CP-UNLOCK` (Algorithm 1), which fully exploits the additional feedback available under semi-bandit feedback.

In comparison to `EXP3.P-CP` and `EXP3.P-CP-SEMI`, our unlocking-based learner `EXP3.P-CP-UNLOCK` achieves a high-probability regret bound of the same order $\sqrt{K \ln KT}$, up to constant and logarithmic factors. Moreover, it admits an explicit, data-independent choice of the trade-off parameter λ by setting $o(T)^{-1} = \varepsilon(\lambda, \alpha)$ (Eq. 40), rather than treating λ as a user-tuned hyperparameter. The coverage guarantee is summarized in the following theorem.

Theorem 1. *For any given $\delta \in (0,1)$ and $\ell_t(\pi) \in [\ell_{min}, \ell_{max}]$, suppose* `EXP3.P-CP-UNLOCK` (Algorithm 1) *is run with $\lambda = \frac{2(1-\alpha)}{\sqrt{T}\alpha - 1}$, $\beta = \sqrt{\frac{\ln K}{CT}}$, $\gamma = 1.05\sqrt{\frac{K \ln K}{T}}$, and $\eta = 0.95\sqrt{\frac{\ln K}{KT}}$, then the empirical miscoverage rate satisfies*

$$\boldsymbol{MC}(T) \leq \alpha + \sqrt{\frac{C \ln K}{T}} + 4.15\sqrt{\frac{K \ln K}{T}}$$

with probability at least $1-\delta$, where C is the constant defined in Eq. 49.

This theorem follows by combining the high-probability regret bound in Theorem 5 with our conversion lemma (Lemma 1).

Algorithm 1 `EXP3.P` Learner for Conformal Prediction with Unlocking and Pseudo gain

```
procedure EXP3.P-CP-UNLOCK(Π, T, η, γ, β, λ, α, (f_t)_{t=1}^T)
    Initialize cumulative estimated gains G̃_0(π) ← 0 for all π ∈ Π
    for t ← 1, ..., T do
        Compute probabilities: p_t(π) ← (1 − γ) exp(η G̃_{t−1}(π)) / Σ_{π̃∈Π} exp(η G̃_{t−1}(π̃)) + γ 1/K, where K = |Π|
        Sample arm: π_t ∼ p_t
        Receive m_t(π_t) ← 𝟙(y_t ∉ Ĉ_{π_t}(x_t))
        if m_t(π_t) = 0 then                                  (▷) Exploit the structure of arms.
            Π_t(π_t) ← Π
        else                                                  (▷) Semi-bandit feedback: Observe the true label.
            Π_t(π_t) ← {π ∈ Π | π ≥ π_t}
        for π̃ ∈ Π_t(π_t) do                                   (▷) Reuse feedback m_t(π_t).
            ℓ_t(π̃, α) ← COMPUTELOSS(π̃, m_t(π̃), λ, α)
            Compute normalized gain: g_t(π̃) = (ℓ_max − ℓ_t(π̃, α)) / (ℓ_max − ℓ_min)
        Construct biased gain estimator g̃_t(π|Π_t(π_t)), defined in (5)
        Update cumulative gain: G̃_t(π) ← G̃_{t−1}(π) + g̃_t(π)
procedure COMPUTELOSS(π, m, λ, α)
    return |m − α/(λ+2)| + (λα)/(λ+2) {𝟙(m=0) exp(−π/√T) + 𝟙(m=1) exp(−(1−π)/√T)}
```

5 Experiment

In this section, we empirically evaluate the bandit-based conformal prediction methods `EXP3.P-CP`, `EXP3.P-CP-SEMI`, and our main algorithm `EXP3.P-CP-UNLOCK`. We study how the theoretical coverage guarantees derived from the regret bounds and the conversion lemma (Lemma 1) manifest in practice under different data-generating regimes.

Datasets. We evaluate one classification and one regression benchmark under three data-generating regimes: i.i.d. streams, score-shifted streams, and covariate-shifted streams. For classification, we use the *ImageNet* dataset with 1,000 classes, and for regression, the *UCI Airfoil Self-Noise* dataset (Dua & Graff, 2017). Further details, including the underlying scoring functions, base predictors, and the precise constructions of the shifted set-ups, are deferred to Appendix E.2 and Appendix E.3.

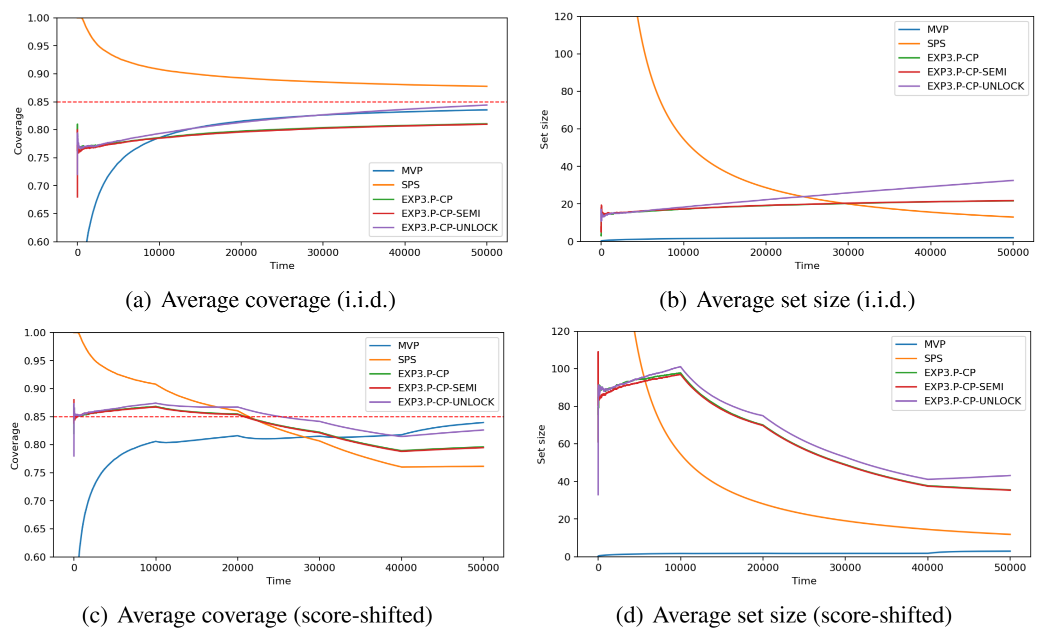
Baselines. *MultiValid Prediction* (MVP) (Bastani et al., 2022) is an online conformal set learning method that provides coverage guarantees under an adversarial set-up with full feedback, so that the loss of all conformal set parameters can be evaluated each round. We also consider the *Semi-bandit Prediction Set* (SPS) method (Ge et al., 2024), which leverages semi-bandit feedback and the nested structure of conformal sets to estimate losses for all parameters from a single labeled example per round. On top of these baselines, `EXP3.P-CP` (Algorithm 3) is a modification of `EXP3.P` tailored to conformal set learning, and `EXP3.P-CP-SEMI` (Algorithm 4) and `EXP3.P-CP-UNLOCK` (Algorithm 1) are semi-bandit variants incorporating an unlocking mechanism, with MVP serving as an oracle baseline for our partial-feedback setting.

5.1 Classification: I.I.D. and Adversarial Score Shifts

Figure 5.1 shows the ImageNet results under both the i.i.d. and adversarially score-shifted streams.

Under the i.i.d. set-up (target coverage $1-\alpha = 0.85$), all methods attain empirical coverage close to the nominal target but with different efficiency profiles. MVP achieves slightly sub-nominal coverage with the smallest prediction sets, whereas SPS attains higher coverage at the cost of substantially larger sets, illustrating a standard coverage–efficiency trade-off. Our bandit-based methods (`EXP3.P-CP`, `EXP3.P-CP-SEMI`, `EXP3.P-CP-UNLOCK`) also reach coverage near the target, with `EXP3.P-CP-UNLOCK` closest to the target while using wider sets than MVP due to the partial-feedback constraint.

Under the adversarial score-shifted set-up, all methods exhibit some undercoverage relative to the target. MVP remains competitive, preserving relatively high coverage with compact sets, whereas SPS suffers a marked drop in coverage despite moderately large sets. The bandit-based methods respond by enlarging their prediction sets to maintain reasonably high coverage under partial feedback, with `EXP3.P-CP-UNLOCK` achieving the highest coverage among them at the expense of the widest sets.



(a) Average coverage (i.i.d.)

(b) Average set size (i.i.d.)

(c) Average coverage (score-shifted)

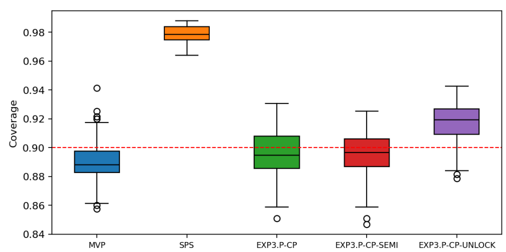
(d) Average set size (score-shifted)

Figure 1: Average coverage and prediction set size on the *ImageNet* dataset under the i.i.d. (top row) and adversarially score-shifted (bottom row) set-ups, averaged over 50 independent runs.

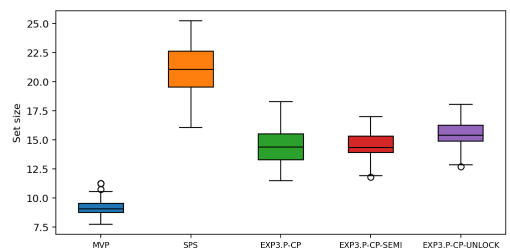
5.2 Regression: I.I.D. and Covariate Shift

Under the i.i.d. *Airfoil* set-up (target coverage $1 - \alpha = 0.9$), MVP attains slightly sub-nominal coverage with the most compact prediction sets, whereas SPS achieves the highest coverage but with substantially wider sets, again illustrating a clear coverage–efficiency trade-off. Our bandit-based baselines `EXP3.P-CP` and `EXP3.P-CP-SEMI` reach coverage very close to the nominal level with intermediate set sizes between MVP and SPS, and `EXP3.P-CP-UNLOCK` further increases coverage to slightly above the target while incurring only a modest additional increase in width.

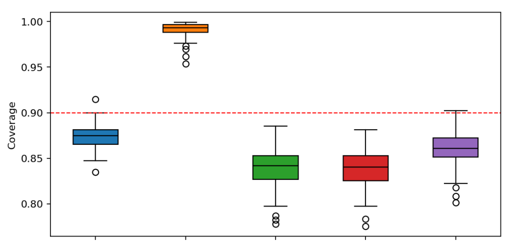
Under the covariate-shift set-up, all methods experience some degradation in coverage relative to the i.i.d. case. MVP remains reasonably well-calibrated with a mild increase in set size, while SPS continues to prioritize coverage, attaining near-perfect coverage at the cost of the largest widths. Among the bandit-based methods, `EXP3.P-CP` and `EXP3.P-CP-SEMI` maintain moderately wide sets with somewhat reduced coverage, and `EXP3.P-CP-UNLOCK` again improves coverage relative to these baselines with only a small increase in width. Taken together, the Airfoil experiments mirror the classification results: semi-bandit variants and unlocking enhance coverage under both i.i.d. and shifted regimes at the price of a moderate increase in prediction set size.



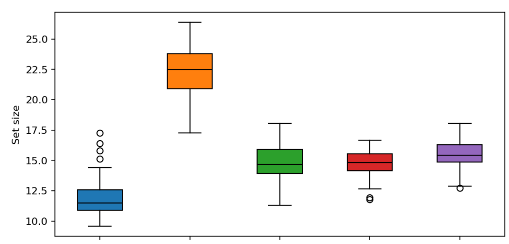
(a) Average coverage (i.i.d.)



(b) Average set size (i.i.d.)



(c) Average coverage (covariate shift)



(d) Average set size (covariate shift)

Figure 2: Average coverage and prediction set size on the *Airfoil Self-Noise* dataset under the i.i.d. (top row) and covariate-shifted (bottom row) set-ups, averaged over 100 independent runs.

6 Conclusion

We introduce an online conformal prediction algorithm that operates with semi-bandit feedback in both stochastic (i.i.d.) and adversarial settings. The method can be applied to several tasks, like classification and regression, constructing prediction sets while observing only partial information each round. We establish that, under semi feedback coupled with adversarial bandit updates, minimizing an appropriate regret objective implies coverage at least $1 - \alpha$, and that the coverage shortfall decays at rate $\mathcal{O}\left(\sqrt{\frac{K \ln K}{T}}\right)$. This contrasts with prior online CP approaches—which typically assume full feedback to update thresholds—and supports more realistic human-in-the-loop workflows where the ground-truth label may be unobservable unless it lies in the prediction set. Our method is currently context-free—it does not leverage the context, x_t; extending both the algorithm and its analysis to contextual semi-bandit settings is a promising direction for future work.

REFERENCES

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074, 2023.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes, 2024a. URL https://arxiv.org/abs/2402.01139.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *International Conference on Machine Learning*, pp. 1616–1630. PMLR, 2024b.

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

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *Advances in neural information processing systems*, 35:29362–29373, 2022.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2337–2363. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bhatnagar23a.html.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods, June 2024. URL http://arxiv.org/abs/2406.09714. arXiv:2406.09714 [cs, stat].

Dheeru Dua and Casey Graff. Uci machine learning repository. https://archive.ics.uci.edu/ml, 2017. Accessed: 2025-09-22.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Stochastic online conformal prediction with semi-bandit feedback. *arXiv preprint arXiv:2405.13268*, 2024.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Stochastic online conformal prediction with semi-bandit feedback. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=IdRrKDsTZ8.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9L1BsI4wP1H.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift, 2021. URL https://arxiv.org/abs/2106.00170.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

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

Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective generation for controllable language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=glfYOAzh2f.

Minjae Lee, Yoonjae Jung, and Sangdon Park. A regret perspective on online selective generation, 2025. URL https://arxiv.org/abs/2506.14067.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Conformal prediction with temporal quantile adjustments. *Advances in Neural Information Processing Systems*, 35:31017–31030, 2022.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8(8):5116–5123, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

Ji Won Park and Kyunghyun Cho. Semiparametric conformal prediction. In *International Conference on Artificial Intelligence and Statistics*, pp. 3880–3888. PMLR, 2025.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DhP9L8vIyLc.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. *arXiv preprint arXiv:2103.03323*, 2021.

Wenwen Si, Sangdon Park, Insup Lee, Edgar Dobriban, and Osbert Bastani. PAC prediction sets under label shift. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4vPVBh3fhz.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019. ISSN 1935-8237. doi: 10.1561/2200000068. URL http://dx.doi.org/10.1561/2200000068.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32:2530–2540, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine learning*, 92(2-3): 349–376, 2013.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Zhou Wang and Xingye Qiao. Efficient online set-valued classification with bandit feedback. *arXiv preprint arXiv:2405.04393*, 2024.

Qiangqiang Zhang, Ting Li, Xinwei Feng, Xiaodong Yan, and Jinhan Xie. Online differentially private conformal prediction for uncertainty quantification. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dmZQrojdVU.

A PRELIMINARY

A.1 REGRET MINIMIZATION WITH PARTIAL FEEDBACK

Sequential prediction, represented by multi-armed bandits (Slivkins, 2019), is modeled by a game between a learner and an adversary (also called an environment). For multi-armed bandits with K arms $\Pi = \{\pi_1, \ldots, \pi_K\}$ over T rounds, at each round $t \in [1, T]$, a learner selects an arm $\pi_t \in \Pi$ and an adversary provides a loss $\ell_t(\pi_t) \in [0, 1]$ on the selected arm as feedback, where the learner leverages the feedback to update its arm-selection strategy. Here, we call the above feedback type *partial feedback* as the adversary provides feedback on the selected arm, while *full feedback* represents a setup where the adversary provides feedback on all arms. Note that the adversary may provide feedback regardless of the learner's arm selection. This adversary is called *oblivious*. However, if the feedback at round t depends on the learner's previous selections $\pi_1, \ldots, \pi_{t-1}$, we say that the adversary is *adaptive*. In this paper, we consider the adaptive adversary, which is stronger than the oblivious one. For both oblivious and adaptive adversaries, we denote our bandit problem by *adversarial bandits*.

In adversarial bandit problems, the objective of learning is modeled by regret, which is the gap between the learner's cumulative loss and the best arm's cumulative loss in hindsight, which is formally quantified as follows:

$$\textbf{Reg}(T) \coloneqq \sum_{t=1}^{T} \ell_t(\pi_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(\pi).$$

Here, we have two sources of randomness: (1) a learner's randomized strategy, *i.e.,* an arm π_t is drawn from an arm distribution updated by the learners and (2) an adversary's randomized strategy, *i.e.,* the adversary's feedback vector ℓ_t is drawn from a feedback distribution which depends on the learner's previously chosen arms $\pi_1, \ldots, \pi_{t-1}$ without looking at the current learner's arm choice π_t.

The goal of the adversarial bandit is to find a learner's strategy such that the corresponding regret is sub-linear in T with high probability. Note that we do not consider a deterministic learner, as it is known that it cannot achieve the sub-linear regret bound Bubeck et al. (2012). In the following, we introduce a known regret minimization method, called the Exponential-weight algorithm for Exploration and Exploitation to control the regret variance (`EXP3.P`), for the adversarial bandit under the adaptive adversary.

A.2 `EXP3.P` FOR ADVERARIAL BANDITS

`EXP3.P` (Algorithm 2) maintains an estimate of the cumulative biased gain for each arm $\pi \in \Pi$, where $|\Pi| = K$. Following the conventional descriptions on `EXP3.P`, we illustrate the algorithm in terms of gain instead of loss for clarity. In particular, at each round t, the learner updates a probability distribution over arms as

$$p_t(\pi) = (1-\gamma)\frac{\exp\big(\eta \tilde{G}_{t-1}(\pi)\big)}{\sum_{\tilde{\pi} \in \Pi} \exp\big(\eta \tilde{G}_{t-1}(\tilde{\pi})\big)} + \gamma \frac{1}{K},$$

where $\gamma \in (0, 1]$ is a mixing weight, $\eta > 0$ is a learning rate, and $\tilde{G}_{t-1}(\pi)$ denotes the cumulative estimated gain of arm π up to round $t-1$. The learner samples an arm $\pi_t \sim p_t$, observes the loss $\ell_t(\pi_t) \in [\ell_{\min}, \ell_{\max}]$ or equivalently the gain $g_t(\pi) = \frac{\ell_{\max} - \ell_t(\pi, \frac{\alpha}{\lambda+2})}{\ell_{\max} - \ell_{\min}}$, , and forms the biased gain estimator:

$$\tilde{g}_t(\pi) = \frac{g_t(\pi)\mathbb{1}(\pi_t = \pi) + \beta}{p_t(\pi)}, \tag{6}$$

where $\beta > 0$ is a bias parameter. Then, the cumulative gain estimate for each arm is updated as $\tilde{G}_t(\pi) = \tilde{G}_{t-1}(\pi) + \tilde{g}_t(\pi)$, which is then used to update the arm-selection strategy at round $t+1$. Theorem 2 shows that `EXP3.P` achieves the high probability regret bound under the properly chosen hyperparameters. See Appendix B for the proof.

Theorem 2 (High Probability Bound (Auer et al., 2002)). *For any given $\delta \in (0, 1)$ and $\ell_t(\pi) \in [\ell_{min}, \ell_{max}]$, if Algorithm 2 is run with $\beta = \sqrt{\frac{\ln K}{KT}}$, $\eta = 0.95\sqrt{\frac{\ln K}{KT}}$, $\gamma = 1.05\sqrt{\frac{K \ln K}{T}}$, then the*

Algorithm 2 EXP3.P for Adversarial Bandits

1: **procedure** EXP3.P($\Pi, T, \eta, \gamma, \beta$)
2: Initialize cumulative gains $\tilde{G}_0(\pi) \leftarrow 0$ for all $\pi \in \Pi$
3: **for** $t \leftarrow 1, \ldots, T$ **do**
4: Compute probabilities: $p_t(\pi) \leftarrow (1-\gamma)\frac{\exp(\eta \tilde{G}_{t-1}(\pi))}{\sum_{\tilde{\pi}\in\Pi}\exp(\eta\tilde{G}_{t-1}(\tilde{\pi}))} + \gamma\frac{1}{K}$ (▷) $K = |\Pi|$
5: Sample an arm: $\pi_t \sim p_t$
6: Observe a loss: $\ell_t(\pi_t) \in [\ell_{\min}, \ell_{\max}]$ (▷) $\ell_t(\pi) \in [\ell_{\min}, \ell_{\max}]$ $\forall \pi \in \Pi$ and $\forall t \in \mathbb{N}$
7: Compute a normalized gain: $g_t(\pi_t) = \frac{\ell_{\max} - \ell_t(\pi_t)}{\ell_{\max} - \ell_{\min}}$
8: Construct a biased gain estimator: $\tilde{g}_t(\pi) \leftarrow \frac{g_t(\pi)\mathbb{1}(\pi_t = \pi) + \beta}{p_t(\pi)}$ for all $\pi \in \Pi$
9: Update cumulative gains: $\tilde{G}_t(\pi) \leftarrow \tilde{G}_{t-1}(\pi) + \tilde{g}_t(\pi)$ for all $\pi \in \Pi$

following holds with probability at least $1-\delta$*:*

$$\boldsymbol{Reg}(T) \leq (\ell_{max} - \ell_{min})\left(\sqrt{\frac{TK}{\ln K}}\ \ln(\delta^{-1}) + 5.15\sqrt{TK\ln K}\right).$$

Note that the original proof on the regret bound of EXP3.P algorithm requires $\ell_t(\pi) \in [0,1]$ for any $t \in \mathbb{N}$ and $\pi \in \Pi$ (Auer et al., 2002). But, here we consider a simple loss normalization in the algorithm and bound, providing the equivalent result for any bounded loss functions.

A.3 EXP3.P FOR CONFORMAL PREDICTION

Algorithm 3 EXP3.P Learner for Conformal Prediction

1: **procedure** EXP3.P-CP($\Pi, T, \eta, \gamma, \beta, \lambda, \alpha$)
2: Initialize cumulative estimated gains $\tilde{G}_0(\pi) \leftarrow 0$ for all $\pi \in \Pi$
3: **for** $t \leftarrow 1, \ldots, T$ **do**
4: Compute probabilities: $p_t(\pi) \leftarrow (1-\gamma)\frac{\exp(\eta \tilde{G}_{t-1}(\pi))}{\sum_{\tilde{\pi}\in\Pi}\exp(\eta\tilde{G}_{t-1}(\tilde{\pi}))} + \gamma\frac{1}{K}$, where $K = |\Pi|$
5: Sample arm: $\pi_t \sim p_t$
6: Receive $m_t(\pi_t) \leftarrow \mathbb{1}\left(y_t \notin \hat{C}_{\pi_t}(x_t)\right)$
7: Observe loss: $\ell_t(\pi_t, \alpha) \leftarrow$ COMPUTELOSS($\pi_t, m_t(\pi_t), \lambda, \alpha$)
8: Compute normalized gain: $g_t(\pi_t) = \frac{\ell_{\max} - \ell_t(\pi_t, \alpha)}{\ell_{\max} - \ell_{\min}}$
9: Construct biased gain estimator: $\tilde{g}_t(\pi) \leftarrow \frac{g_t(\pi)\mathbb{1}(\pi_t = \pi) + \beta}{p_t(\pi)}$
10: Update cumulative gain: $\tilde{G}_t(\pi) \leftarrow \tilde{G}_{t-1}(\pi) + \tilde{g}_t(\pi)$
11: **procedure** COMPUTELOSS(π, m, λ, α)
12: **return** $|m - \alpha| + \frac{\lambda\alpha}{\lambda+2}\{\mathbb{1}(m=0)\exp(-\pi) + \mathbb{1}(m=1)\}$

A.4 EXP3.P FOR CONFORMAL PREDICTION WITH UNLOCKING

Algorithm 4 EXP3.P Learner for Conformal Prediction with Unlocking

procedure EXP3.P-CP-SEMI$(\Pi, T, \eta, \gamma, \beta, \lambda, \alpha, (f_t)_{t=1}^T)$
  Initialize cumulative estimated gains $\tilde{G}_0(\pi) \leftarrow 0$ for all $\pi \in \Pi$
  for $t \leftarrow 1, \ldots, T$ **do**
    Compute probabilities: $p_t(\pi) \leftarrow (1-\gamma)\frac{\exp(\eta \tilde{G}_{t-1}(\pi))}{\sum_{\tilde{\pi}\in\Pi} \exp(\eta \tilde{G}_{t-1}(\tilde{\pi}))} + \gamma \frac{1}{K}$, where $K = |\Pi|$
    Sample arm: $\pi_t \sim p_t$
    Receive $m_t(\pi_t) \leftarrow \mathbb{1}\left(y_t \notin \hat{C}_{\pi_t}(x_t)\right)$
    if $m_t(\pi_t) = 0$ **then** (▷) Exploit the structure of arms.
      $\Pi_t(\pi_t) \leftarrow \{\pi \in \Pi \mid \pi \leq f_t(x_t, y_t)\}$
    else (▷) *Semi-bandit feedback*: Observe the true label.
      $\Pi_t(\pi_t) \leftarrow \{\pi \in \Pi \mid \pi \geq \pi_t\}$
    for $\tilde{\pi} \in \Pi_t(\pi_t)$ **do** (▷) Reuse feedback $m_t(\pi_t)$.
      $\ell_t(\tilde{\pi}, \alpha) \leftarrow$ COMPUTELOSS$(\tilde{\pi}, m_t(\pi_t), \lambda, \alpha)$
      Compute normalized gain: $g_t(\tilde{\pi}) = \frac{\ell_{\max} - \ell_t(\tilde{\pi}, \alpha)}{\ell_{\max} - \ell_{\min}}$
    Construct biased gain estimator $\tilde{g}_t(\pi | \Pi_t(\pi_t))$, defined in (18)
    Update cumulative gain: $\tilde{G}_t(\pi) \leftarrow \tilde{G}_{t-1}(\pi) + \tilde{g}_t(\pi)$
procedure COMPUTELOSS$(\pi, m, \lambda, \alpha)$
  return $\left|m - \frac{\alpha}{\lambda+2}\right| + \frac{\lambda\alpha}{\lambda+2}\left\{\mathbb{1}(m=0)\exp\left(-\frac{\pi}{\sqrt{T}}\right) + \mathbb{1}(m=1)\exp\left(-\frac{1-\pi}{\sqrt{T}}\right)\right\}$

B PROOF OF THEOREM 2

This theorem is due to Auer et al. (Auer et al., 2002). For completeness, we reproduce their regret analysis for EXP3.P, adapting it to our normalized loss setting. We begin by recalling the following key lemma.

Lemma 2. *For $\beta \le 1$ and $g_t(\cdot) \in [0,1]$, let $\tilde{g}_t(\pi) = \frac{g_t(\pi)\mathbb{1}(\pi_t=\pi)+\beta}{p_t(\pi)} \in (0,\infty)$. Then, for each $\pi \in \Pi$, the following holds with probability at least $1-\delta$:*

$$\sum_{t=1}^{T} g_t(\pi) \le \sum_{t=1}^{T} \tilde{g}_t(\pi) + \frac{\ln(\delta^{-1})}{\beta}.$$

Proof. Let $\mathbb{E}_t$ be the expectation conditioned on $\pi_1, ...\pi_{t-1}$. Since $\exp(x) \le 1 + x + x^2$ for $x \le 1$, for $\beta \le 1$, by letting $\Delta_t(\pi_t) \coloneqq \beta g_t(\pi) - \frac{\beta g_t(\pi)\mathbb{1}(\pi_t=\pi)}{p_t(\pi)} \le 1$, we have

$$\begin{aligned}
\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t) - \frac{\beta^2}{p_t(\pi)}\right)\right] &\le \left(1 + \mathbb{E}_t[\Delta_t(\pi_t)] + \mathbb{E}_t[\Delta_t(\pi_t)^2]\right)\exp\left(-\frac{\beta^2}{p_t(\pi)}\right) \\
&\le \left(1 + \frac{\beta^2 g_t(\pi)^2}{p_t(\pi)}\right)\exp\left(-\frac{\beta^2}{p_t(\pi)}\right) \\
&\le \left(1 + \frac{\beta^2}{p_t(\pi)}\right)\exp\left(-\frac{\beta^2}{p_t(\pi)}\right) \ (\because\ g_t(\cdot) \in [0,1]) \\
&\le 1 \ (\because 1 + u \le \exp(u)).
\end{aligned}$$

By sequentially applying the double expectation rule for $t = T, \ldots, 1$,

$$\mathbb{E}\exp\left[\sum_{t=1}^{T}\left(\Delta_t(\pi_t) - \frac{\beta^2}{p_t(\pi)}\right)\right] \le 1. \tag{7}$$

Moreover, from the Markov's inequality, we have $\mathbb{P}\left(X > \ln(1/\delta)\right) = \mathbb{P}\left(\exp(X) > 1/\delta\right) \le \delta\mathbb{E}\exp(X)$. Combined with Eq. 7, we have

$$\beta\sum_{t=1}^{T} g_t(\pi) \le \beta\sum_{t=1}^{T} \tilde{g}_t(\pi) + \ln(\delta^{-1})$$

with probability at least $1-\delta$. This completes the proof. □

Now, we show the proof on the regret bound of EXP3.P for any bounded loss functions, which consists of five steps.

First, our goal is to show that, if $\gamma \le \frac{1}{2}$ and $(1+\beta)K\eta \le \gamma$,

$$\mathbf{Reg}(T) \le (\ell_{\text{max}} - \ell_{\text{min}})\left(\beta TK + \gamma T + (1+\beta)\eta Kn + \frac{\ln(K/\delta)}{\beta} + \frac{\ln K}{\eta}\right). \tag{8}$$

Irrespective of the hyperparameter setup, note that Eq. 8 always holds if $T \ge 5.15\sqrt{TK\ln(K\delta^{-1})}$. If $T < 5.15\sqrt{TK\ln(K\delta^{-1})}$, this implies that $\gamma \le \frac{1}{2}$ and $(1+\beta)K\eta \le \gamma$, which makes it suffice to show that Eq. 8 holds for $\gamma \le \frac{1}{2}$ and $(1+\beta)K\eta \le \gamma$.

Step 1: Simple equalities. By the definition of $\tilde{g}_t(\pi)$, the following holds:

$$\mathbb{E}_{\pi\sim p_t}\tilde{g}_t(\pi) = g_t(\pi_t) + \beta K. \tag{9}$$

Here, the gain $g_t(\pi) \in [0,1]$ is defined with respect to the loss $\ell_t(\pi) \in [\ell_{\text{min}}, \ell_{\text{max}}]$ as the following:

$$g_t(\pi) = \frac{\ell_{\text{max}} - \ell_t(\pi)}{\ell_{\text{max}} - \ell_{\text{min}}}.$$

Then, for all $\pi \in \Pi$, the following equality holds:

$$\begin{aligned} R_\pi(T) &= \sum_{t=1}^T \ell_t(\pi_t) - \sum_{t=1}^T \ell_t(\pi) \\ &= (\ell_{\max} - \ell_{\min}) \left(\sum_{t=1}^T g_t(\pi) - \sum_{t=1}^T g_t(\pi_t) \right) \; (\because \text{ Definition of } g_t(\cdot)) \quad (10) \\ &= (\ell_{\max} - \ell_{\min}) \left(\beta K T + \sum_{t=1}^T g_t(\pi) - \sum_{t=1}^T \mathbb{E}_{\tilde{\pi}\sim p_t} \tilde{g}_t(\tilde{\pi}) \right) \; (\because \text{ Eq. 9}). \end{aligned}$$

Using the definition of cumulant generating function and the relationship that $p_t = (1-\gamma)\omega_t + \gamma u$ where $\omega_t(\pi) = \frac{\exp\left(\eta \tilde{G}_{t-1}(\pi)\right)}{\sum_{\tilde{\pi}\in\Pi} \exp\left(\eta \tilde{G}_{t-1}(\tilde{\pi})\right)}$ and u is the uniform distribution over K arms, the following holds:

$$\begin{aligned} -\mathbb{E}_{\tilde{\pi}\sim p_t} \tilde{g}_t(\tilde{\pi}) &= -(1-\gamma)\mathbb{E}_{\tilde{\pi}\sim \omega_t} \tilde{g}_t(\pi_i) - \gamma \mathbb{E}_{\tilde{\pi}\sim u} \tilde{g}_t(\tilde{\pi}) \; (\because p_t = (1-\gamma)\omega_t + \gamma u) \\ &= (1-\gamma) \left[\frac{1}{\eta} \ln \mathbb{E}_{\tilde{\pi}\sim\omega_t} \exp\left(\tilde{g}_t(\tilde{\pi}) - \mathbb{E}_{\tilde{\pi}\sim\omega_t} \tilde{g}_t(\tilde{\pi}) \right) - \right. \quad (11) \\ &\qquad \left. \frac{1}{\eta} \ln \mathbb{E}_{\tilde{\pi}\sim\omega_t} \exp(\eta \tilde{g}_t(\tilde{\pi})) \right] - \gamma \mathbb{E}_{\tilde{\pi}\sim u} \tilde{g}_t(\tilde{\pi}) \end{aligned}$$

Step 2: Bounding the first term of Eq. 11. Since $\ln x \le x - 1$, $\exp(x) \le 1 + x + x^2$ for all $x \le 1$, and $\eta \tilde{g}_t(\tilde{\pi}) = \eta \frac{g_t(\tilde{\pi})+\beta}{p_t(\tilde{\pi})} \le \eta \frac{1+\beta}{(1-\gamma)w_t(\tilde{\pi}) + \gamma\frac{1}{K}} \le \frac{\gamma\frac{1}{K}}{(1-\gamma)w_t(\tilde{\pi}) + \gamma\frac{1}{K}} \le 1$ ($\because$ $(1+\beta)\eta K \le \gamma$),

$$\begin{aligned} \ln \mathbb{E}_{\tilde{\pi}\sim\omega_t} \exp\left(\eta(\tilde{g}_t(\tilde{\pi}) - \mathbb{E}_{\tilde{\pi}\sim\omega_t} \tilde{g}_t(\tilde{\pi})) \right) &= \ln \mathbb{E}_{\tilde{\pi}\sim\omega_t} \exp(\eta \tilde{g}_t(\tilde{\pi})) - \eta \mathbb{E}_{\tilde{\pi}\sim\omega_t} \tilde{g}_t(\tilde{\pi}) \\ &\le \mathbb{E}_{\tilde{\pi}\sim\omega_t} \left\{ \exp(\eta \tilde{g}_t(\tilde{\pi})) - 1 - \eta \tilde{g}_t(\tilde{\pi}) \right\} \; (\because \ln x \le x-1) \\ &\le \mathbb{E}_{\tilde{\pi}\sim\omega_t} \eta^2 \tilde{g}_t(\tilde{\pi})^2 \; (\because \exp(x) \le 1 + x + x^2) \\ &\le \eta^2 \frac{1+\beta}{1-\gamma} \sum_{\tilde{\pi}\in\Pi} \tilde{g}_t(\tilde{\pi}) \; \left(\because \; \frac{w_t(\tilde{\pi})}{p_t(\tilde{\pi})} \le \frac{1}{1-\gamma} \right). \end{aligned} \quad (12)$$

Step 3: Summing. Let $\tilde{G}_0(\tilde{\pi}) = 0$. Then, combining Eq. 11-Eq. 12 and summing over t yield

$$\begin{aligned} -\sum_{t=1}^T \mathbb{E}_{\tilde{\pi}\sim p_t} \tilde{g}_t(\tilde{\pi}) &\le (1+\beta)\eta \sum_{t=1}^T \sum_{\tilde{\pi}\in\Pi} \tilde{g}_t(\tilde{\pi}) - \frac{1-\gamma}{\eta} \sum_{t=1}^T \ln \left(\sum_{\tilde{\pi}\in\Pi} w_t(\tilde{\pi}) \exp(\eta \tilde{g}_t(\tilde{\pi})) \right) \\ &= (1+\beta)\eta \sum_{t=1}^T \sum_{\tilde{\pi}\in\Pi} \tilde{g}_t(\tilde{\pi}) - \frac{1-\gamma}{\eta} \ln \left(\frac{\sum_{\tilde{\pi}\in\Pi} \exp(\eta \tilde{G}_T(\tilde{\pi}))}{\sum_{\tilde{\pi}\in\Pi} \exp(\eta \tilde{G}_0(\tilde{\pi}))} \right) \; (\because \text{ Definition of } \omega_t(\tilde{\pi}), \tilde{G}_t(\tilde{\pi})) \\ &\le (1+\beta)\eta K \max_{\tilde{\pi}\in\Pi} \tilde{G}_T(\tilde{\pi}) + \frac{\ln K}{\eta} - \frac{1-\gamma}{\eta} \ln \left(\sum_{\tilde{\pi}\in\Pi} \exp(\eta \tilde{G}_T(\tilde{\pi})) \right) \; (\because 1-\gamma \le 1 \text{ and } \tilde{G}_0(\tilde{\pi}) = 0) \\ &\le -(1-\gamma-(1+\beta)\eta K) \max_{\tilde{\pi}\in\Pi} \tilde{G}_T(\tilde{\pi}) + \frac{\ln K}{\eta} \; (\because \text{ Property of log-sum-exponential}) \\ &\le -(1-\gamma-(1+\beta)\eta K) \max_{\tilde{\pi}\in\Pi} \sum_{t=1}^T g_t(\tilde{\pi}) + \frac{\ln(K\delta^{-1})}{\beta} + \frac{\ln K}{\eta}, \end{aligned} \quad (13)$$

where the last inequality holds due to the Lemma 2, union bound (the reason for using the confidence term of $\frac{\delta}{K}$), and the initial assumption that $\gamma \le \frac{1}{2}$ and $(1+\beta)K\eta \le \gamma$. Plugging Eq. 13 into Eq. 10, the following holds with probability $1 - \frac{\delta}{K}$ for all $\pi \in \Pi$:

$$R_\pi(T) \le (\ell_{\max} - \ell_{\min}) \left(\beta T K + \gamma T + (1+\beta)\eta K n + \frac{\ln(K/\delta)}{\beta} + \frac{\ln K}{\eta} \right).$$

Since $\textbf{Reg}(T) := \max R_\pi(T)$, this completes the proof by taking the union bound.

C A Proof on Lemma 1

We have

$$\begin{aligned}
\textbf{Reg}\left(T, \frac{\alpha}{\lambda+2}\right) &\coloneqq \sum_{t=1}^{T} \ell_t\left(\pi_t, \frac{\alpha}{\lambda+2}\right) - \min_{\pi\in\Pi}\sum_{t=1}^{T}\ell_t\left(\pi, \frac{\alpha}{\lambda+2}\right) \\
&= \sum_{t=1}^{T}\left\{\lambda\frac{\alpha}{\lambda+2}a_t(\pi_t) + d_t\left(\pi_t, \frac{\alpha}{\lambda+2}\right)\right\} - \min_{\pi\in\Pi}\sum_{t=1}^{T}\left\{\lambda\frac{\alpha}{\lambda+2}a_t(\pi) + d_t\left(\pi, \frac{\alpha}{\lambda+2}\right)\right\} \\
&\geq \sum_{t=1}^{T}\left\{\lambda\frac{\alpha}{\lambda+2}a_t(\pi_t) + d_t\left(\pi_t, \frac{\alpha}{\lambda+2}\right)\right\} - \lambda\frac{\alpha}{\lambda+2}\sum_{t=1}^{T}a_t(\bar{\pi}) - T\frac{\alpha}{\lambda+2} \qquad (14) \\
&= \lambda\frac{\alpha}{\lambda+2}\sum_{t=1}^{T}\{a_t(\pi_t) - a_t(\bar{\pi})\} + \sum_{t=1}^{T}d_t\left(\pi_t, \frac{\alpha}{\lambda+2}\right) - T\frac{\alpha}{\lambda+2} \\
&\geq -\lambda\frac{\alpha}{\lambda+2}T + \sum_{t=1}^{T}d_t\left(\pi_t, \frac{\alpha}{\lambda+2}\right) - T\frac{\alpha}{\lambda+2}, \qquad (15)
\end{aligned}$$

where $\bar{\pi} = \text{argmin}_{\pi\in\Pi}\sum_{t=1}^{T}d_t\left(\pi, \frac{\alpha}{\lambda+2}\right)$ and thus $\sum_{t=1}^{T}d_t\left(\bar{\pi}, \frac{\alpha}{\lambda+2}\right) = T\frac{\alpha}{\lambda+2}$, so (14) holds as

$$\begin{aligned}
\min_{\pi\in\Pi}\sum_{t=1}^{T}\left\{\lambda\frac{\alpha}{\lambda+2}a_t(\pi) + d_t\left(\pi, \frac{\alpha}{\lambda+2}\right)\right\} &\leq \sum_{t=1}^{T}\left\{\lambda\frac{\alpha}{\lambda+2}a_t(\bar{\pi}) + d_t\left(\bar{\pi}, \frac{\alpha}{\lambda+2}\right)\right\} \\
&= \lambda\frac{\alpha}{\lambda+2}\sum_{t=1}^{T}a_t(\bar{\pi}) + T\frac{\alpha}{\lambda+2}
\end{aligned}$$

and (15) holds as $\pi_t, \bar{\pi} \in \mathbb{R}_{\geq 0}$.

Thus, this implies

$$\sum_{t=1}^{T}d_t\left(\pi_t, \frac{\alpha}{\lambda+2}\right) - \frac{\lambda+1}{\lambda+2}\alpha T \leq \textbf{Reg}\left(T, \frac{\alpha}{\lambda+2}\right) \qquad (16)$$

Thus, considering that

$$\begin{aligned}
\sum_{t=1}^{T}d_t\left(\pi_t, \frac{\alpha}{\lambda+2}\right) - \frac{\lambda+1}{\lambda+2}\alpha T &= \sum_{t=1}^{T}\left|\mathbb{1}(\mathbf{y}_t \notin \hat{C}_{\pi_t}(\mathbf{x}_t)) - \frac{\alpha}{\lambda+2}\right| - \frac{\lambda+1}{\lambda+2}\alpha T \\
&\geq \sum_{t=1}^{T}\mathbb{1}(\mathbf{y}_t \notin \hat{C}_{\pi_t}(\mathbf{x}_t)) - \alpha T.
\end{aligned}$$

Dividing each side by T, we have

$$\textbf{MC}(T) - \alpha \leq \frac{\textbf{Reg}\left(T, \frac{\alpha}{\lambda+2}\right)}{T}.$$

D EXP3.P, EXP3.P-CP-SEMI, AND EXP3.P-CP-UNLOCK

D.1 BIASED UNLOCKING ESTIMATOR AND ITS PROPERTIES

Definition. First of all, we consider the following three different biased estimators $\tilde{g}_t(\pi|\Pi_t(\pi_t))$ under the semi-bandit feedback scenario as the following:

$$\tilde{g}_t\left(\pi \mid \Pi_t(\pi_t)\right) \coloneqq \underbrace{\mathbb{1}(m_t(\pi_t) = 0)}_{\text{full unlocking}} \times (A) + \underbrace{\mathbb{1}(m_t(\pi_t) = 1)}_{\text{partial unlocking}} \times (B), \tag{17}$$

where

- EXP3.P
$$(A) \coloneqq \mathbb{1}(\pi_t = \pi)\left\{\frac{g_t(\pi)}{p_t(\pi)} + \frac{\beta}{p_t(\pi)}\right\} + \mathbb{1}(\pi_t \neq \pi)\frac{\beta}{p_t(\pi)}$$
- EXP3.P-CP-SEMI
$$(A) \coloneqq \mathbb{1}(\pi \in \Pi_t^*)\left\{\frac{g_t(\pi)}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})} + \frac{\beta}{p_t(\pi)}\right\} + \mathbb{1}(\pi \in (\Pi_t^*)^c)\frac{\beta}{p_t(\pi)}$$
- EXP3.P-CP-UNLOCK
$$(A) \coloneqq \mathbb{1}(\pi \in \Pi_t^*)\left\{\frac{g_t(\pi)}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})} + \left(1 + \frac{1}{\sum_{\tilde{\pi}\leq\pi} p_t(\tilde{\pi})}\right)\beta\right\} + \mathbb{1}(\pi \in (\Pi_t^*)^c)\left\{g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi}\leq\pi} p_t(\tilde{\pi})}\right\}$$

and

- EXP3.P
$$(B) \coloneqq \mathbb{1}(\pi_t = \pi)\left\{\frac{g_t(\pi)}{p_t(\pi)} + \frac{\beta}{p_t(\pi)}\right\} + \mathbb{1}(\pi_t \neq \pi)\frac{\beta}{p_t(\pi)}$$
- EXP3.P-CP-SEMI
$$(B) \coloneqq \mathbb{1}(\pi \in \Pi_t(\pi_t))\left\{\frac{g_t(\pi)}{\sum_{\tilde{\pi}\in\Pi_t(\pi_t)} p_t(\tilde{\pi})} + \frac{\beta}{p_t(\pi)}\right\} + \mathbb{1}(\pi \in \Pi_t(\pi_t)^c)\frac{\beta}{p_t(\pi)}$$
- EXP3.P-CP-UNLOCK
$$(B) \coloneqq \mathbb{1}(\pi \in \Pi_t(\pi_t))\left\{g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi}\leq\pi} p_t(\tilde{\pi})}\right\} + \mathbb{1}(\pi \in \Pi_t(\pi_t)^c)\left\{\tilde{g}_t(\pi) + \left(1 + \frac{1}{p_t(\pi)}\right)\beta\right\}.$$

D.2 THEORETICAL ANALYSIS

Theorem 3 (EXP3.P)**.** *For any given $\delta \in (0,1)$ and $\ell_t(\pi) \in [\ell_{min}, \ell_{max}]$, if* EXP3.P *is run with $\beta = \sqrt{\frac{\ln K}{KT}}, \gamma = 1.05\sqrt{\frac{K\ln K}{T}}, \eta = 0.95\sqrt{\frac{\ln K}{KT}}$, then the following holds with probability at least $1-\delta$:*

$$\textbf{\textit{Reg}}(T) \leq 5.15\sqrt{K\ln KT}.$$

Theorem 4 (EXP3.P-CP-SEMI)**.** *For any given $\delta \in (0,1)$ and $\ell_t(\pi) \in [\ell_{min}, \ell_{max}]$, if* EXP3.P-CP-SEMI *is run with $\beta = \sqrt{\frac{\ln K}{KT}}, \gamma = 1.05\sqrt{\frac{K\ln K}{T}}, \eta = 0.95\sqrt{\frac{\ln K}{KT}}$, then the following holds with probability at least $1-\delta$:*

$$\textbf{\textit{Reg}}(T) \leq 5.15\sqrt{K\ln KT}.$$

Theorem 5 (EXP3.P-CP-UNLOCK)**.** *For any given $\delta \in (0,1)$ and $\ell_t(\pi) \in [\ell_{min}, \ell_{max}]$, if* EXP3.P-CP-UNLOCK *is run with $\varepsilon(\lambda,\alpha) = \frac{1}{\sqrt{T}}, \beta = \sqrt{\frac{\ln K}{CT}}, \gamma = 1.05\sqrt{\frac{K\ln K}{T}}$, and $\eta = 0.95\sqrt{\frac{\ln K}{KT}}$, then the following holds with probability at least $1-\delta$:*

$$\textbf{\textit{Reg}}(T) \leq \sqrt{C\ln KT} + 4.15\sqrt{K\ln KT},$$

where C and $\varepsilon(\lambda,\alpha)$ are the terms defined in Eq. 49 and Eq. 40, respectively.

E Experiment Setup

E.1 Parameter Choices

Our bandit-based methods depend on two main design parameters: the number K of candidate thresholds and the trade-off parameter λ in the loss function for `EXP3.P-CP` and `EXP3.P-CP-SEMI`. We briefly summarize the theoretical and practical considerations that guide our choices, and specify the values used in our experiments.

Number of thresholds K. For the EXP3.P-based algorithms, we adopt the standard exploration parameter

$$\gamma = 1.05\sqrt{\frac{K \ln K}{T}},$$

and the theoretical guarantees require $\gamma \le \frac{1}{2}$. This condition implicitly upper-bounds the feasible number of thresholds K for a given horizon T; if K is chosen too large, then γ would exceed $\frac{1}{2}$ and the original EXP3.P regret guarantees would no longer apply.

Even within this feasible range, there is a trade-off between coverage and efficiency. On the one hand, taking K very small yields a coarse grid of thresholds, which makes it relatively easy for the learned conformal sets to attain coverage at or above the target level, but typically at the cost of larger prediction sets. On the other hand, taking K very large yields a much finer grid and can in principle improve efficiency, but the regret bounds scale as $\mathcal{O}\big(\sqrt{K \ln K/T}\big)$, so larger K slows the convergence of the empirical coverage toward the target level. Consequently, for a fixed time horizon T, K cannot be increased arbitrarily without degrading the finite-sample coverage behavior, and if T is too small, even a moderate value of K may not be sufficient to bring the empirical coverage close to the target level.

In our main experiments, we balance these considerations by choosing $K = 1{,}000$ for the ImageNet classification experiments and $K = 20$ for the Airfoil regression experiments. These choices satisfy the EXP3.P constraint on γ and provide a practically useful compromise between coverage and prediction set size at the respective horizons.

Trade-off parameter λ. The loss function for `EXP3.P-CP` and `EXP3.P-CP-SEMI` combines a miscoverage term and an inefficiency term, weighted by a trade-off parameter $\lambda > 0$. Smaller values of λ reduce the influence of the inefficiency loss, encouraging the algorithm to prioritize eliminating miscoverage and thus reach the target coverage more quickly, typically at the expense of larger prediction sets. Larger values of λ place more weight on inefficiency, promoting smaller sets once coverage has been largely stabilized. Empirically, we find that the algorithms are reasonably robust to the precise choice of λ; moderate changes in λ tend not to qualitatively alter the coverage trajectories.

In all of our main experiments (excluding the λ-ablation in Section F), we fix $\lambda = 1$ for `EXP3.P-CP` and `EXP3.P-CP-SEMI`. The `EXP3.P-CP-UNLOCK` algorithm does not introduce an additional free trade-off parameter: its learning-rate and exploration parameters are fully determined by the horizon T and the target coverage level $1 - \alpha$ through our theoretical construction, so no separate tuning of λ is required.

E.2 Classification: I.I.D. and Adversarial Score Shifts

Setup. We consider ImageNet classification with a pre-trained ResNet-18 (He et al., 2015), and let $p_\theta(y \mid x)$ denote the softmax probability of label y given input x.

For the *i.i.d.* case, we use a time-homogeneous scoring function

$$f_t^{\text{iid}}(x, y) := p_\theta(y \mid x)^{1/3}, \qquad t = 1, \dots, T.$$

For the *adversarial score-shifted* case, we keep the underlying data stream fixed and i.i.d., but make the scoring function time-varying by rescaling the probabilities with a piecewise-constant exponent

γ_t:

$$\gamma_t = \begin{cases} 1/6 & 1 \le t \le 10{,}000, \\ 1/4 & 10{,}001 \le t \le 20{,}000, \\ 1/2 & 20{,}001 \le t \le 30{,}000, \\ 1/1.2 & 30{,}001 \le t \le 40{,}000, \\ 1/3 & 40{,}001 \le t \le 50{,}000, \end{cases}$$

and define

$$f_t^{\text{adv}}(x, y) := p_\theta(y \mid x)^{\gamma_t}.$$

Smaller values of γ_t make the transformed scores $p_\theta(y \mid x)^{\gamma_t}$ more concentrated near 1, making the true label harder to distinguish from competing labels and thus creating a challenging adversarial score-shift scenario.

Protocol. We run experiments for $T = 50{,}000$ rounds with $K = 1{,}000$ candidate thresholds and target coverage $1-\alpha = 0.85$. The order of the data stream is fixed and shared across all baselines. For all online baselines (MVP, SPS, `EXP3.P-CP`, `EXP3.P-CP-SEMI`, and `EXP3.P-CP-UNLOCK`), we use the full stream of $T = 50{,}000$ examples and update the conformal prediction sets online at every round, under either f_t^{iid} or f_t^{adv}.

Metrics. We track the empirical marginal coverage and average prediction set size at each round $t = 1, \dots, T$. For both the i.i.d. and adversarial score-shifted cases, we plot the trajectories of these two metrics over time $T = 50{,}000$ across 50 independent runs.

E.3 Regression: I.I.D. and Covariate Shift Cases

The *UCI Airfoil Self-Noise* dataset consists of five-dimensional features—frequency, angle of attack, chord length, free-stream velocity, and suction-side displacement thickness—used to predict the scaled sound pressure level (Dua & Graff, 2017).

Setup. We use a linear regression predictor $\hat{y}(x)$, trained using the recursive least squares algorithm. As a scoring function, we use the residual-based score

$$f(x, y) = \frac{u - |y - \hat{y}(x)|}{u - l},$$

where l and u denote lower and upper bounds, respectively, on the residuals $|y - \hat{y}(x)|$.

Protocol. For the i.i.d. set-up, we run experiments for $T = 1{,}127$ rounds with $K = 20$ candidate thresholds and target coverage $1 - \alpha = 0.9$. The order of the data stream is fixed and shared across all baselines. All considered methods (MVP, SPS, and our bandit-based algorithms) use the entire stream of T samples and update the conformal prediction sets in an online manner. For the covariate shift set-up, we use the same total horizon and thresholds, but the first 33% of the samples are drawn from a different input distribution than the remaining 67%, while the update protocol for all methods remains identical.

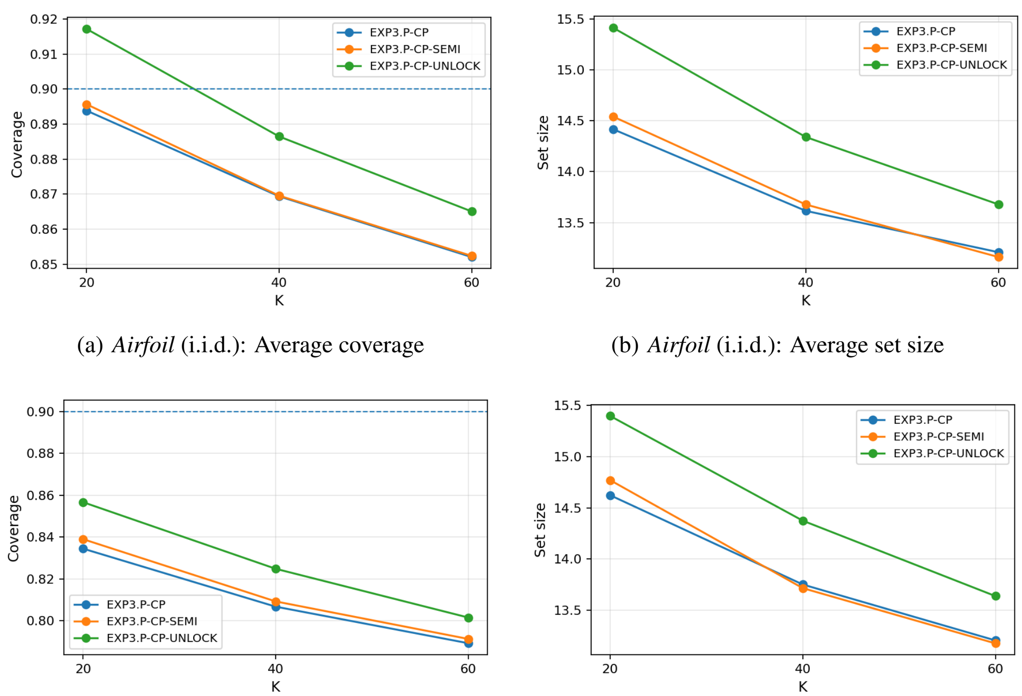
Metrics. For the i.i.d. set-up, we evaluate empirical marginal coverage and average prediction set size on the last two-thirds of the stream, discarding the first third. For the covariate shift set-up, we compute these quantities over all T rounds. In both set-ups, we perform 100 independent trials and summarize the results using box plots of the trial-wise empirical mean coverage and average prediction set size for each method.

F Ablation Study

F.1 Ablation on the Number of Thresholds

We first study the effect of the number of candidate thresholds K, complementing the discussion in Section E.1, while fixing $\alpha = 0.1$ and setting $\lambda = 1$ for the bandit-based baselines `EXP3.P-CP` and `EXP3.P-CP-SEMI`. For the regression task on *Airfoil*, we consider $K \in \{20, 40, 60\}$ and aggregate results over 100 runs; for the classification task on *ImageNet*, we consider $K \in \{200, 500, 1000\}$ and aggregate over 50 runs.

Across both the i.i.d. and shifted set-ups, all bandit-based methods exhibit a mild degradation in coverage as K increases, while the average prediction set size consistently decreases with larger K, reflecting the expected coverage–efficiency trade-off from the regret bounds. Among the proposed methods, `EXP3.P-CP-UNLOCK` consistently achieves the highest coverage for all choices of K in both tasks and under both i.i.d. and shifted regimes, with moderately larger prediction sets compared to `EXP3.P-CP` and `EXP3.P-CP-SEMI`. Overall, the qualitative conclusions from the main experiments are stable across this range of K. In particular, for the *Airfoil* regression task we keep $K = 20$ in the main results, and for the *ImageNet* classification task we retain the finer grid with $K = 1{,}000$; the additional curves for $K = 200$ and $K = 500$ in this section confirm that our conclusions are robust to the specific choice of K.



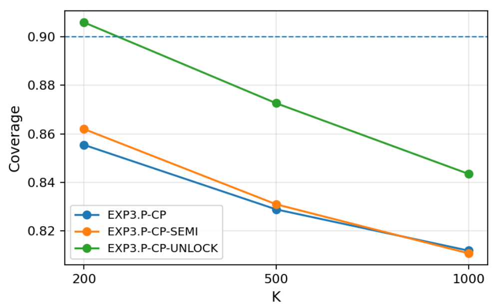
(a) *Airfoil* (i.i.d.): Average coverage

(b) *Airfoil* (i.i.d.): Average set size

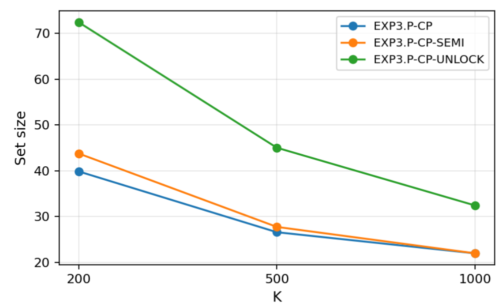
(c) *Airfoil* (shifted): Average coverage

(d) *Airfoil* (shifted): Average set size

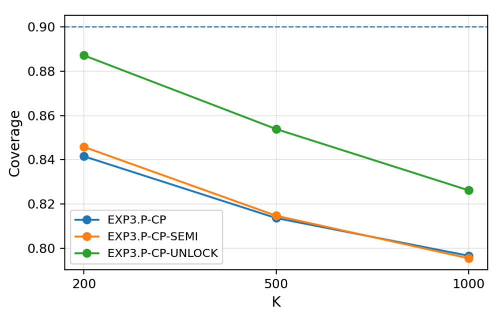
Figure 3: Ablation on the number of thresholds K for the *Airfoil* regression task under the i.i.d. (top) and covariate-shift (bottom) set-ups, averaged over 100 independent runs.



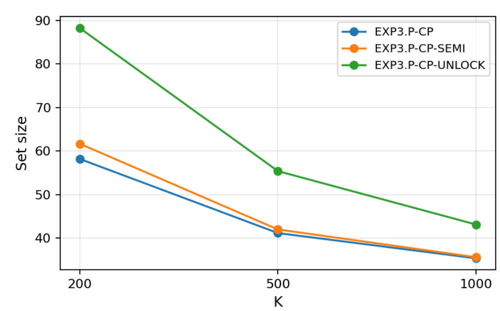
(a) *ImageNet* (i.i.d.): Average coverage



(b) *ImageNet* (i.i.d.): Average set size



(c) *ImageNet* (shifted): Average coverage



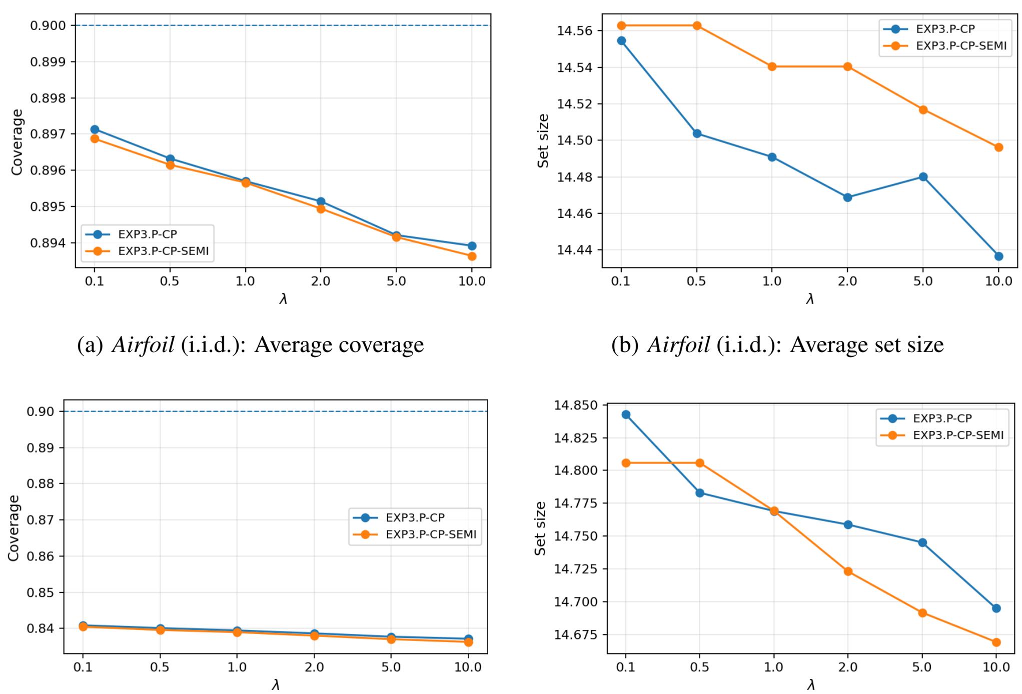
(d) *ImageNet* (shifted): Average set size

Figure 4: Ablation on the number of thresholds K for the *ImageNet* classification task under the i.i.d. (top row) and distribution-shifted (bottom row) set-ups, averaged over 50 independent runs.

F.2 ABLATION ON THE TRADE-OFF PARAMETER

Next, we investigate the sensitivity with respect to the trade-off parameter λ in the loss, again complementing the qualitative discussion in Section E.1, while fixing $\alpha = 0.1$ and focusing on the bandit-based baselines `EXP3.P-CP` and `EXP3.P-CP-SEMI`. For the *Airfoil* regression task, we fix $K = 20$ and vary λ over a range of values, aggregating results over 100 runs under both the i.i.d. and covariate-shift set-ups. For the *ImageNet* classification task, we fix $K = 200$ (a coarser grid than the main choice $K = 1{,}000$ to reduce computational cost), vary $\lambda \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$, and aggregate over 50 runs under the i.i.d. set-up.

In both datasets, empirical coverage for `EXP3.P-CP` and `EXP3.P-CP-SEMI` remains very stable across the tested values of λ, varying only within a narrow range around the target level. The average prediction set size exhibits only modest changes and tends to decrease mildly as λ increases, indicating that larger values of λ mainly act to refine efficiency once coverage has been stabilized. These ablations show that our bandit-based methods are quite robust to the choice of λ, which justifies fixing $\lambda = 1.0$ for `EXP3.P-CP` and `EXP3.P-CP-SEMI` throughout the main experiments. To further streamline the design of this trade-off, our main algorithm `EXP3.P-CP-UNLOCK` is constructed so that its internal trade-off parameter is given in closed form as a function of the user-specified T (time horizon) and α (target miscoverage rate), while enjoying the same theoretical guarantees.



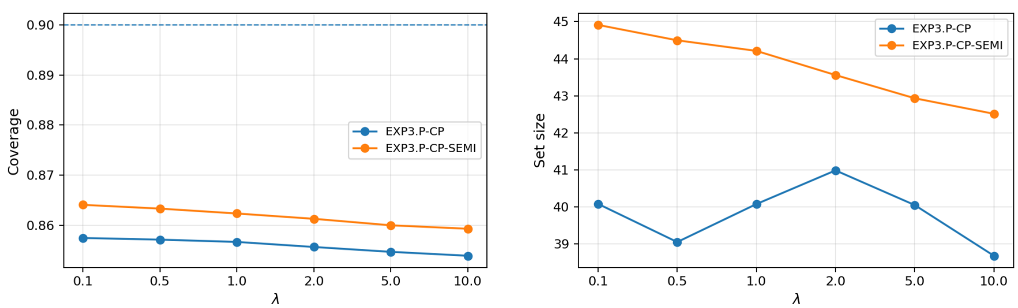
(a) *Airfoil* (i.i.d.): Average coverage

(b) *Airfoil* (i.i.d.): Average set size

(c) *Airfoil* (shifted): Average coverage

(d) *Airfoil* (shifted): Average set size

Figure 5: Ablation on the trade-off parameter λ for the *Airfoil* regression task under the i.i.d. (top) and covariate-shift (bottom) set-ups, averaged over 100 independent runs.



(a) *ImageNet* (i.i.d.): Average coverage

(b) *ImageNet* (i.i.d.): Average set size

Figure 6: Ablation on the trade-off parameter λ for the *ImageNet* classification task under the i.i.d. set-up, averaged over 50 independent runs.

G PROOF OF THEOREM 4 FOR EXP3.P-CP-SEMI

G.1 BIASED UNLOCKING ESTIMATOR AND ITS PROPERTIES

Definition. First of all, we consider the following biased unlocking estimator $\tilde{g}_t(\pi \mid \Pi_t(\pi_t))$ under the semi-bandit feedback scenario as the following:

$$\tilde{g}_t\left(\pi \mid \Pi_t(\pi_t)\right) \coloneqq \underbrace{\mathbb{1}(m_t(\pi_t) = 0)}_{\text{full unlocking}} \times (A) + \underbrace{\mathbb{1}(m_t(\pi_t) = 1)}_{\text{partial unlocking}} \times (B). \tag{18}$$

Letting $\Pi_t^* \coloneqq \{\tilde{\pi} \in \Pi : \tilde{\pi} \leq f_t(x_t, y_t)\}$,

$$(A) \coloneqq \mathbb{1}(\pi \in \Pi_t^*) \left\{ \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})} + \frac{\beta}{p_t(\pi)} \right\} + \mathbb{1}(\pi \in (\Pi_t^*)^c) \frac{\beta}{p_t(\pi)}$$

and

$$(B) \coloneqq \mathbb{1}(\pi \in \Pi_t(\pi_t)) \left\{ \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t(\pi_t)} p_t(\tilde{\pi})} + \frac{\beta}{p_t(\pi)} \right\} + \mathbb{1}(\pi \in \Pi_t(\pi_t)^c) \frac{\beta}{p_t(\pi)}$$

In addition, the unlocking set $\Pi_t(\pi_t) \subset \Pi$ is defined as follows:

$$\Pi_t(\pi_t) = \begin{cases} \Pi_t^* & \text{if } m_t(\pi_t) = 0 \\ \{\tilde{\pi} \in \Pi : \tilde{\pi} \geq \pi_t\} & \text{if } m_t(\pi_t) = 1. \end{cases}$$

Note that $m_t(\pi_1) \leq m_t(\pi_2) \; \forall \, \pi_1 \leq \pi_2$ due to the monotonicity property of the conformal set with respect to the miscoverage, $i.e.$, $\mathbb{1}(y_t \notin C_{\pi_1}(x_t)) \leq \mathbb{1}(y_t \notin C_{\pi_2}(x_t))$ whenever $\pi_1 \leq \pi_2$.

Properties of the Unlocking Set. The followings are properties of the unlocking set $\Pi_t(\tilde{\pi})$:

- $\tilde{\pi} \in \Pi_t(\tilde{\pi})$
- When $m_t(\tilde{\pi}) = 0$ (full feedback), the unlocking set is $\tilde{\pi}$-independent, $\Pi = \Pi_t^* \cup (\Pi_t^*)^c$.

G.2 LOSS FUNCTION AND ITS PROPERTIES

Definition. Now, we introduce our loss estimator $\ell_t(\pi, c) = d_t(\pi, c) + \frac{\lambda\alpha}{\lambda+2} a_t(\pi)$ ($c \in (0, 0.5), \lambda > 0$) and its intuition behind. First, $d_t(\pi, c) \coloneqq |\mathbb{1}(y_t \notin C_\pi(x_t)) - c|$ is defined as the **miscoverage loss**. Note that the conversion lemma (Lemma 1) ensures the convergence to target coverage $1 - \alpha$ when $c = \frac{\alpha}{\lambda+2}$. Second, $a_t(\pi)$ is the **penalty term to optimize the set size**.

Rationale for the Design of $a_t(\pi)$. Recalling that (1) the miscoverage loss is of primary importance in conformal prediction and (2) the binary search-type algorithm is implemented in the batch learning set-up (ref.), we define $a_t(\pi)$ as the following:

$$a_t(\pi) \coloneqq \mathbb{1}(m_t(\pi) = 0) \exp\left(-\frac{\pi}{o(T)}\right) + \mathbb{1}(m_t(\pi) = 1) \exp\left(-\frac{1-\pi}{o(T)}\right)$$

Here, we set denominator inside the exponential to be $\sqrt{T}$, which can be any of the form $o(T) = T^k \; \forall k \in [0.5, 1)$ such that $\frac{o(T)}{T} \to 0$ as $T \to \infty$. Such denominator is necessary for the regret analysis, which will be described in detail in subsequent sections.

Intuitively, if $\pi \in \Pi_t^*$, *i.e.*, $y_t \in C_\pi(x_t)$, $a_t(\pi)$ takes small value as the size of the conformal set is **small** ($|C_\pi(x_t)| \downarrow$). This has the effect to maintain the conformal set to be as small as possible as long as $y_t \in C_\pi(x_t)$. On the other hand, if $\pi \in (\Pi_t^*)^c$, *i.e.*, $y_t \notin C_\pi(x_t)$, $a_t(\pi)$ takes small value as the size of the conformal set is **large** ($|C_\pi(x_t)| \uparrow$). This ensures the penalty term to take the miscoverage loss, instead of the set size efficiency, of priority importance when $y_t \notin C_\pi(x_t)$.

Properties of the Loss Estimator. Then, the followings are properties of our loss estimator $\ell_t(\pi, c) = d_t(\pi, c) + \frac{\lambda\alpha}{\lambda+2} a_t(\pi)$.

- By the definition of $d_t(\pi, c)$, $d_t(\pi, c) = c \; \forall \pi \in \Pi_t^*$ and $d_t(\pi, c) = 1 - c \; \forall \pi \in (\Pi_t^*)^c$. Note that as long as $c \in (0, 0.5)$,
$$c < 1 - c. \tag{19}$$
- For all $\pi_1, \pi_2 \in \Pi_t^*$ such that $\pi_1 \leq \pi_2$,
$$a_t(\pi_1) \geq a_t(\pi_2) \; (\because \; \exp(-\frac{\pi}{\sqrt{T}}) \text{ is monotonically decreasing}).$$
Therefore, for all $\pi_1, \pi_2 \in \Pi_t^*$ such that $\pi_1 \leq \pi_2$,
$$\ell_t(\pi_1, c) \geq \ell_t(\pi_2, c). \tag{20}$$
- For all $\pi_1, \pi_2 \in (\Pi_t^*)^c$ such that $\pi_1 \leq \pi_2$,
$$a_t(\pi_1) \leq a_t(\pi_2) \; (\because \; \exp(-\frac{1-\pi}{\sqrt{T}}) \text{ is monotonically increasing}).$$
Therefore, for all $\pi_1, \pi_2 \in \Pi_t^*$ such that $\pi_1 \leq \pi_2$,
$$\ell_t(\pi_1, c) \leq \ell_t(\pi_2, c). \tag{21}$$
- Let $\ell_{t,0} \coloneqq \max_{\tilde{\pi} \in \Pi_t^*} \ell_t(\tilde{\pi}, c)$ and $\ell_{t,1} \coloneqq \min_{\tilde{\pi} \in (\Pi_t^*)^c} \ell_t(\tilde{\pi}, c)$. Due to the two properties above (Eq. 20-21), by letting $\pi_{t,0} \coloneqq \min_{\tilde{\pi} \in \Pi_t^*} \tilde{\pi} = 0$ and $\pi_{t,1} \coloneqq \min_{\tilde{\pi} \in (\Pi_t^*)^c} \tilde{\pi}$,
$$\begin{aligned} \ell_{t,0} &= \ell_t(\pi_{t,0}, c), \\ \ell_{t,1} &= \ell_t(\pi_{t,1}, c). \end{aligned}$$
Since controlling the miscoverage is of utmost importance before optimizing the set size in conformal prediction, the loss estimator will be most satisfactory when $\ell_{t,0} \leq \ell_{t,1} \; \forall t \in [T]$. This is true when $\ell_{t,1} - \ell_{t,0} = (1 - 2c) + \frac{\lambda\alpha}{\lambda+2} \left\{ \exp(-\frac{1-\pi_{t,1}}{\sqrt{T}}) - \exp(0) \right\} \geq 0$. Note that it holds if we set $c = \frac{\alpha}{\lambda+2}$ as our proposed algorithm does:
$$\begin{aligned} \ell_{t,1} - \ell_{t,0} &= (1 - 2\frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2} \left\{ \exp(-\frac{1-\pi_{t,1}}{\sqrt{T}}) - \exp(0) \right\} \\ &\geq (1 - 2\frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2} \left\{ -\exp(0) \right\} \\ &= 1 - \alpha \\ &\geq 0. \end{aligned}$$
Therefore,
$$\ell_t(\pi_1, \frac{\alpha}{\lambda+2}) \leq \ell_t(\pi_2, \frac{\alpha}{\lambda+2}) \; \forall \pi_1 \in \Pi_t^*, \; \forall \pi_2 \in (\Pi_t^*)^c. \tag{22}$$

The properties of our proposed loss estimator $\ell_t(\pi, c)$ above hold for all $t \in [T]$ and $c \in (0, 0.5)$. However, we only consider the case where $c = \frac{\alpha}{\lambda+2}$ henceforth, since it is the condition that the conversion lemma requires for the coverage guarantee (Lemma 1).

Here, we define the normalized gain $g_t(\pi)$, which is the one used to define the biased unlocking estimator (Eq. 18) as follows:
$$g_t(\pi) = \frac{\ell_{\text{max}} - \ell_t(\pi, \frac{\alpha}{\lambda+2})}{\ell_{\text{max}} - \ell_{\text{min}}} \; \forall \pi \in \Pi,$$
where $\ell_{\text{max}} \coloneqq (1 - \frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2}$ and $\ell_{\text{min}} \coloneqq \frac{\alpha}{\lambda+2}$. Since we only consider the case where $c = \frac{\alpha}{\lambda+2}$ in the algorithm, we use the notation $g_t(\pi)$ instead of $g_t(\pi, \frac{\alpha}{\lambda+2})$ for brevity.

Useful Inequalities. Based on these properties of our proposed loss estimator, by letting $c = \frac{\alpha}{\lambda+2}$ as our algorithm, the following inequalities hold for each case:

- Case 1 ($m_t(\pi_t) = 0$)
  1. $\pi \in \Pi_t^*$ and $\pi \leq \pi_t$: $\ell_t(\pi, \frac{\alpha}{\lambda+2}) \geq \ell_t(\pi_t, \frac{\alpha}{\lambda+2}) \Leftrightarrow g_t(\pi) \leq g_t(\pi_t)$ ($\because$ Eq. 20)
  2. $\pi \in \Pi_t^*$ and $\pi > \pi_t$: $\ell_t(\pi, \frac{\alpha}{\lambda+2}) \leq \ell_t(\pi_t, \frac{\alpha}{\lambda+2}) \Leftrightarrow g_t(\pi) \geq g_t(\pi_t)$, where $\ell_t(\pi_t, \frac{\alpha}{\lambda+2}) - \ell_t(\pi, \frac{\alpha}{\lambda+2}) = \frac{\lambda\alpha}{\lambda+2}\left\{\exp(-\frac{\pi_t}{o(T)}) - \exp(-\frac{\pi}{o(T)})\right\} = o(T)^{-1}$ ($\because$ Taylor expansion). Therefore, since $g_t(\pi) - g_t(\pi_t) = \frac{\frac{\lambda\alpha}{\lambda+2}\left\{\exp(-\frac{\pi_t}{o(T)}) - \exp(-\frac{\pi}{o(T)})\right\}}{\ell_{\max} - \ell_{\min}}$,

$$g_t(\pi) = g_t(\pi_t) + o(T)^{-1} \ (\because \text{ Eq. 20}).$$

  Therefore,

$$g_t(\pi) \leq g_t(\pi_t) + o(T)^{-1} \ \forall \pi \in \Pi_t^*. \tag{23}$$

- Case 2 ($m_t(\pi_t) = 1$)
  1. $\pi \in \Pi_t(\pi_t) \subset (\Pi_t^*)^c$: $\ell_t(\pi) \geq \ell_t(\pi_t) \Leftrightarrow g_t(\pi) \leq g_t(\pi_t)$ ($\because$ Eq. 21)

  Therefore,

$$g_t(\pi) \leq g_t(\pi_t) \ \forall \pi \in \Pi(\pi_t). \tag{24}$$

G.3 Expectation and Arm-wise Bounds for the Unlocking Estimator

Based on our preceding results, our biased unlocking estimator satisfies the following inequality:

- Case 1 ($m_t(\pi_t) = 0$)

$$\begin{aligned}
\mathbb{E}_{\pi \sim p_t} \tilde{g}_t(\pi | \Pi_t(\pi_t)) &= \textstyle\sum_{\pi \in \Pi} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) \\
&= \textstyle\sum_{\pi \in \Pi_t^*} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) + \textstyle\sum_{\pi \in (\Pi_t^*)^c} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) \\
&= \textstyle\sum_{\pi \in \Pi_t^*} p_t(\pi) \left\{ \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})} + \frac{\beta}{p_t(\pi)} \right\} + \textstyle\sum_{\pi \in (\Pi_t^*)^c} p_t(\pi) \left\{ \frac{\beta}{p_t(\pi)} \right\} \\
&\leq g_t(\pi_t) + o(T)^{-1} + K\beta \ (\because \text{ Eq. 23})
\end{aligned} \tag{25}$$

- Case 2 ($m_t(\pi_t) = 1$)

$$\begin{aligned}
\mathbb{E}_{\pi \sim p_t} \tilde{g}_t(\pi | \Pi_t(\pi_t)) &= \textstyle\sum_{\pi \in \Pi} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) \\
&= \textstyle\sum_{\pi \in \Pi_t(\pi_t)} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) + \textstyle\sum_{\pi \in \Pi_t(\pi_t)^c} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) \\
&= \textstyle\sum_{\pi \in \Pi_t(\pi_t)} p_t(\pi) \left\{ \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t(\pi_t)} p_t(\tilde{\pi})} + \frac{\beta}{p_t(\pi)} \right\} + \textstyle\sum_{\pi \in \Pi_t(\pi_t)^c} p_t(\pi) \left\{ \frac{\beta}{p_t(\pi)} \right\} \\
&\leq g_t(\pi_t) + K\beta \ (\because \text{ Eq. 24})
\end{aligned} \tag{26}$$

Combining Eq. 25 and Eq. 26, the following upper bound holds irrespective of the choice of π_t:

$$\mathbb{E}_{\pi \sim p_t} \tilde{g}_t(\pi | \Pi_t(\pi_t)) \leq g_t(\pi_t) + o(T)^{-1} + K\beta. \tag{27}$$

Arm-wise High Probability Bound. Before moving on to the regret analysis, we provide the following lemma, a variant of Lemma 2, which characterizes the property of our biased gain estimator under the semi-bandit feedback scenario.

Lemma 3. *For $\beta \leq 1$ and $g_t(\cdot) \in [0, 1]$, define $\tilde{g}_t(\pi) \in [0, \infty)$ as Eq. 18. Then, for each $\pi \in \Pi$, the following holds with probability at least $1 - \delta$:*

$$\sum_{t=1}^{T} g_t(\pi) \leq \sum_{t=1}^{T} \tilde{g}_t(\pi) + \frac{\ln(\delta^{-1})}{\beta}.$$

Proof. **Step 1: Useful Decomposition.**
Let $\mathbb{E}_t$ be the expectation conditioned on $\pi_1, ...\pi_{t-1}$. First, we decompose the estimator in Eq. 18 as

the following:

$$\begin{aligned}\tilde{g}_t(\pi \mid \Pi_t(\pi_t)) &\coloneqq \underbrace{\mathbb{1}(m_t(\pi_t)=0)}_{\text{full unlocking}} \times (A) + \underbrace{\mathbb{1}(m_t(\pi_t)=1)}_{\text{partial feedback}} \times (B) \\ &= \underbrace{\mathbb{1}(m_t(\pi_t)=0)}_{\text{full unlocking}} \times \{(A1)+(A2)\} + \underbrace{\mathbb{1}(m_t(\pi_t)=1)}_{\text{partial unlocking}} \times \{(B1)+(B2)\} \\ &= \underbrace{\{\mathbb{1}(m_t(\pi_t)=0)\times(A1)+\mathbb{1}(m_t(\pi_t)=1)\times(B1)\}}_{(C1)} + \underbrace{\{\mathbb{1}(m_t(\pi_t)=0)\times(A2)+\mathbb{1}(m_t(\pi_t)=1)\times(B2)\}}_{(C2)}\end{aligned}$$

where

$$(A) = \underbrace{\mathbb{1}(\pi \in \Pi_t^*) \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})}}_{(A1)} + \underbrace{\frac{\beta}{p_t(\pi)}}_{(A2)}$$

and

$$(B) = \underbrace{\mathbb{1}(\pi \in \Pi_t(\pi_t)) \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t(\pi_t)} p_t(\tilde{\pi})}}_{(B1)} + \underbrace{\frac{\beta}{p_t(\pi)}}_{(B2)}.$$

Step 2: Show $\Delta_t(\pi_t) \le 1$.
Next, we show the following two claims are true:

- Claim 1 ($m_t(\pi_t) = 0$)

$$\beta g_t(\pi) - \beta \times (A1) = \beta g_t(\pi) - \beta \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})} \le 1$$

- Claim 2 ($m_t(\pi_t) = 1$):

$$\beta g_t(\pi) - \beta \times (B1) = \beta g_t(\pi) - \beta \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t(\pi_t)} p_t(\tilde{\pi})} \le 1$$

Given $\pi \in \Pi$, Claim 1 and Claim 2 always hold irrespective of the choice of π_t by the algorithm. Then, by letting $\Delta_t(\pi_t) \coloneqq \beta g_t(\pi) - \beta \times (C1)$, $\Delta_t(\pi_t) \le 1$.

Step 3: Show $\mathbb{E} \exp \left[\sum_{t=1}^T (\Delta_t(\pi_t) - \beta \times (C2))\right] \le 1$.
Therefore, since (1) $\exp(x) \le 1 + x + x^2$ for $x \le 1$ and (2) $\Delta_t(\pi_t) \le 1$, for $\beta \le 1$, we have

$$\begin{aligned}\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t) - \beta \times (C2)\right)\right] &\le \mathbb{E}_t\Big[(1 + \Delta_t(\pi_t) + \Delta_t(\pi_t)^2) \times \exp(-\beta \times (C2))\Big] \\ &= \mathbb{E}_t\Big[(1 + \Delta_t(\pi_t) + \Delta_t(\pi_t)^2)\Big] \times \exp\left(\frac{-\beta^2}{p_t(\pi)}\right)\end{aligned}$$

Since π is fixed, we consider each case where $\pi \in \Pi_t^*$ and $\pi \in (\Pi_t^*)^c$.

Now, our goal is to show that

$$\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t) - \beta \times (C2)\right)\right] \le 1 \; \forall t \in [T].$$

Step 3-1: $\pi \in \Pi_t^*$.
Step 3-1-1: $\mathbb{E}_t\Big[\Delta_t(\pi_t)\Big]$.

$$\begin{aligned}\sum_{\tilde{\pi} \in \Pi} p_t(\tilde{\pi}) \Delta_t(\tilde{\pi}) &= \beta g_t(\pi) - \sum_{\tilde{\pi} \in \Pi} p_t(\tilde{\pi}) \times \beta (C1) \\ &= \beta g_t(\pi) - \beta \sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi}) \frac{g_t(\pi)}{\sum_{\pi' \in \Pi_t^*} p_t(\pi')} \\ &= 0.\end{aligned}$$

Step 3-1-2: $\mathbb{E}_t\Big[\Delta_t(\pi_t)^2\Big]$**.**

$$\begin{aligned}
\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 &= \sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 \\
&= (\beta g_t(\pi))^2 \sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi}) \left(1 - \frac{1}{\sum_{\pi'\in\Pi_t^*} p_t(\pi')}\right)^2 + (\beta g_t(\pi))^2 \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi}) \\
&\leq \frac{\beta^2}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})} \; (\because \; g_t(\pi), \tilde{g}_t(\pi) \in [0,1]) \\
&\leq \frac{\beta^2}{p_t(\pi)}
\end{aligned}$$

Step 3-1-3: Combine. Combining the above results and applying the fact that $1 + x \leq \exp(x)$ is suffice as follows:

$$\begin{aligned}
\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t) - \beta\times(C2)\right)\right] &\leq \mathbb{E}_t\Big[(1 + \Delta_t(\pi_t) + \Delta_t(\pi_t)^2) \times \exp(-\beta\times(C2))\Big] \\
&\leq \exp\left(-\frac{\beta^2}{p_t(\pi)}\right) \mathbb{E}_t\Big[(1 + \Delta_t(\pi_t) + \Delta_t(\pi_t)^2)\Big] \\
&\leq \exp\left(-\frac{\beta^2}{p_t(\pi)}\right)\left(1 + \frac{\beta^2}{p_t(\pi)}\right) \\
&\leq 1 \; (\because \; 1 + x \leq \exp(x)).
\end{aligned}$$

Step 3-2: $\pi \in (\Pi_t^*)^c$**.**
Step 3-2-1: $\mathbb{E}_t\Big[\Delta_t(\pi_t)\Big]$**.**

$$\begin{aligned}
\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\Delta_t(\tilde{\pi}) &= \beta g_t(\pi) - \sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi}) \times \beta(C1) \\
&= \beta g_t(\pi) - \beta \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi}) \frac{g_t(\pi)}{\sum_{\pi'\in\Pi_t(\tilde{\pi})} p_t(\pi')} \\
&\leq \beta g_t(\pi) - \beta \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi}) \frac{g_t(\pi)}{\sum_{\pi'\in(\Pi_t^*)^c} p_t(\pi')} \\
&= 0.
\end{aligned}$$

Step 3-2-2: $\mathbb{E}_t\Big[\Delta_t(\pi_t)^2\Big]$**.**

$$\begin{aligned}
\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 \\
&= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 \\
&= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})(\beta g_t(\pi))^2 + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})(\beta g_t(\pi) - \frac{\beta g_t(\pi)}{\sum_{\pi'\in\Pi_t(\tilde{\pi})} p_t(\pi')})^2 \\
&\qquad \textstyle+ \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})(\beta g_t(\pi))^2 \\
&\leq \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})(\beta g_t(\pi))^2 + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})(\beta g_t(\pi) - \frac{\beta g_t(\pi)}{p_t(\tilde{\pi})})^2 \\
&\qquad \textstyle+ \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})(\beta g_t(\pi))^2 \\
&\leq \textstyle\frac{\beta^2}{p_t(\pi)} \; (\because g_t(\pi), \tilde{g}_t(\pi) \in [0,1]).
\end{aligned}$$

Step 3-2-3: Combine. Combining the above results,

$$\begin{aligned}\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t)-\beta\times(C2)\right)\right] &\le \mathbb{E}_t\Big[(1+\Delta_t(\pi_t)+\Delta_t(\pi_t)^2)\times\exp(-\beta\times(C2))\Big]\\ &\le \exp(-\frac{\beta^2}{p_t(\pi)})\mathbb{E}_t\Big[(1+\Delta_t(\pi_t)+\Delta_t(\pi_t)^2)\Big]\\ &\le \exp\left(-\frac{\beta^2}{p_t(\pi)}\right)\left(1+\frac{\beta^2}{p_t(\pi)}\right)\\ &\le 1\ (\because\ 1+x\le\exp(x)).\end{aligned}$$

By sequentially applying the double expectation rule for $t=T,\dots,1$,

$$\mathbb{E}\exp\left[\sum_{t=1}^{T}\left(\Delta_t(\pi_t)-\beta\times(C2)\right)\right]\le 1. \tag{28}$$

Moreover, from the Markov's inequality, we have $\mathbb{P}\left(X>\ln(1/\delta)\right)=\mathbb{P}\left(\exp(X)>1/\delta\right)\le\delta\mathbb{E}\exp(X)$. Combined with Eq. 28, we have

$$\beta\sum_{t=1}^{T}g_t(\pi)\le\beta\sum_{t=1}^{T}\tilde{g}_t(\pi)+\ln(\delta^{-1})$$

with probability at least $1-\delta$. This completes the proof. □

Proof of the Theorem. We first provide the proof sketch of Theorem 1 as the following.

Proof Sketch. The proof on Theorem 1 is complete, by substituting the above inequality (Eq. 27) to (Eq. 9) in the proof of `EXP3.P` (Theorem 2).

First, we re-express Eq. 27 for the simplicity of proof as the following:

$$-g_t(\pi_t)\le-\mathbb{E}_{\pi\sim p_t}\tilde{g}_t(\pi|\Pi_t(\pi_t))+o(T)^{-1}+K\beta \tag{29}$$

Now, we show the proof on the regret bound of Algorithm 4, which consists of four steps.

First, our goal is to show that, if $\gamma\le\frac{1}{2}$ and $(1+\beta)K\eta\le\gamma$,

$$\mathbf{Reg}(T,\alpha)\le(\ell_{\text{max}}-\ell_{\text{min}})\left(K\beta T+o(T)^{-1}T+\gamma T+(1+\beta)\eta KT+\frac{\ln(K\delta^{-1})}{\beta}+\frac{\ln K}{\eta}\right). \tag{30}$$

Irrespective of the hyperparameter setup, note that Eq. 30 always holds if $T\ge 5.15\sqrt{TK\ln(K\delta^{-1})}$. If $T<5.15\sqrt{TK\ln(K\delta^{-1})}$, this implies that $\gamma\le\frac{1}{2}$ and $(1+\beta)K\eta\le\gamma$, which makes it suffice to show that Eq. 30 holds for $\gamma\le\frac{1}{2}$ and $(1+\beta)K\eta\le\gamma$.

Step 1: Simple equalities. Recall that the gain $g_t(\pi)\in[0,1]$ is defined with respect to the loss $\ell_t(\pi,\frac{\alpha}{\lambda+2})\in[\ell_{\text{min}},\ell_{\text{max}}]$ as the following:

$$g_t(\pi)=\frac{\ell_{\text{max}}-\ell_t(\pi,\frac{\alpha}{\lambda+2})}{\ell_{\text{max}}-\ell_{\text{min}}}.$$

Then, for all $\pi\in\Pi$, the following equality holds:

$$\begin{aligned}R_\pi(T,\alpha)&=\sum_{t=1}^{T}\ell_t(\pi_t,\frac{\alpha}{\lambda+2})-\sum_{t=1}^{T}\ell_t(\pi,\frac{\alpha}{\lambda+2})\\ &=(\ell_{\text{max}}-\ell_{\text{min}})\left(\sum_{t=1}^{T}g_t(\pi)-\sum_{t=1}^{T}g_t(\pi_t)\right)\ (\because\ \text{Definition of } g_t(\cdot))\\ &\le(\ell_{\text{max}}-\ell_{\text{min}})\left(K\beta T+o(T)^{-1}T+\sum_{t=1}^{T}g_t(\pi)-\sum_{t=1}^{T}\mathbb{E}_{\tilde{\pi}\sim p_t}\tilde{g}_t(\tilde{\pi}\mid\Pi_t(\pi_t))\right)\ (\because\ \text{Eq. 29}).\end{aligned} \tag{31}$$

Using the definition of cumulant generating function and the relationship that $p_t = (1-\gamma)\omega_t + \gamma u$ where $\omega_t(\pi) = \frac{\exp\left(\eta \tilde{G}_{t-1}(\pi)\right)}{\sum_{\tilde{\pi}\in\Pi} \exp\left(\eta \tilde{G}_{t-1}(\tilde{\pi})\right)}$ and u is the uniform distribution over K arms, the following holds:

$$\begin{aligned}
-\mathbb{E}_{\tilde{\pi}\sim p_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) &= -(1-\gamma)\mathbb{E}_{\tilde{\pi}\sim\omega_t}\tilde{g}_t(\pi_i \mid \Pi_t(\pi_t)) - \gamma\mathbb{E}_{\tilde{\pi}\sim u}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \; (\because p_t = (1-\gamma)\omega_t + \gamma u) \\
&= (1-\gamma)\Bigg[\frac{1}{\eta}\ln\mathbb{E}_{\tilde{\pi}\sim\omega_t}\exp\left(\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) - \mathbb{E}_{\tilde{\pi}\sim\omega_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))\right) - \\
&\qquad \frac{1}{\eta}\ln\mathbb{E}_{\tilde{\pi}\sim\omega_t}\exp(\eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)))\Bigg] - \gamma\mathbb{E}_{\tilde{\pi}\sim u}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \quad (32)
\end{aligned}$$

Step 2: Bounding the first term of Eq. 32. First, we show that irrespective of the choice of π_t,

$$\eta\tilde{g}_t(\pi|\Pi_t(\pi_t)) \leq 1 \; \forall\pi \in \Pi.$$

- $m(\pi_t) = 0,\, \pi \in \Pi_t^*$

$$\begin{aligned}
\eta\tilde{g}_t(\pi|\Pi_t(\pi_t)) &= \eta\frac{g_t(\pi)}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})} + \eta\frac{\beta}{p_t(\pi)} \\
&\leq \eta\left(\frac{g_t(\pi)+\beta}{p_t(\pi)}\right) \\
&\leq \frac{\eta(1+\beta)}{(1-\gamma)w_t(\tilde{\pi}) + \gamma\frac{1}{K}} \; (\because (1+\beta)\eta K \leq \gamma) \\
&\leq 1.
\end{aligned}$$

- $m(\pi_t) = 1,\, \pi \in \Pi_t(\pi_t)$

$$\begin{aligned}
\eta\tilde{g}_t(\pi|\Pi_t(\pi_t)) &= \eta\left(\frac{g_t(\pi)}{\sum_{\tilde{\pi}\in\Pi_t(\pi_t)} p_t(\tilde{\pi})} + \frac{\beta}{p_t(\pi)}\right) \\
&\leq \eta\left(\frac{g_t(\pi)+\beta}{p_t(\pi)}\right) \\
&\leq \frac{\eta(1+\beta)}{(1-\gamma)w_t(\tilde{\pi}) + \gamma\frac{1}{K}} \; (\because (1+\beta)\eta K \leq \gamma) \\
&\leq 1.
\end{aligned}$$

- $m(\pi_t) = 0,\, \pi \in (\Pi_t^*)^c;\, m(\pi_t) = 1,\, \pi \in \Pi_t(\pi_t)^c$

$$\begin{aligned}
\eta\tilde{g}_t(\pi|\Pi_t(\pi_t)) &= \eta\frac{\beta}{p_t(\pi)} \\
&\leq \eta\left(\frac{1+\beta}{p_t(\pi)}\right) \\
&\leq \frac{\eta(1+\beta)}{(1-\gamma)w_t(\pi) + \gamma\frac{1}{K}} \\
&\leq \frac{\gamma\frac{1}{K}}{(1-\gamma)w_t(\pi) + \gamma\frac{1}{K}} \; (\because (1+\beta)\eta K \leq \gamma) \\
&\leq 1.
\end{aligned}$$

Since (1) $\ln x \leq x - 1$, (2) $\exp(x) \leq 1 + x + x^2$ for all $x \leq 1$, and (3) $\eta\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) \leq 1$,

$$\begin{aligned}
&\ln\mathbb{E}_{\tilde{\pi}\sim\omega_t}\exp\left(\eta(\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) - \mathbb{E}_{\tilde{\pi}\sim\omega_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)))\right) \\
&\quad = \ln\mathbb{E}_{\tilde{\pi}\sim\omega_t}\exp(\eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))) - \eta\mathbb{E}_{\tilde{\pi}\sim\omega_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \\
&\quad \leq \mathbb{E}_{\tilde{\pi}\sim\omega_t}\left\{\exp(\eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))) - 1 - \eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))\right\} \; (\because \ln x \leq x - 1) \\
&\quad \leq \mathbb{E}_{\tilde{\pi}\sim\omega_t}\eta^2\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))^2 \; (\because \exp(x) \leq 1 + x + x^2) \\
&\quad \leq \eta^2\frac{1+\beta}{1-\gamma}\sum_{\tilde{\pi}\in\Pi}\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)),
\end{aligned} \quad (33)$$

where the last inequality holds due to the following:

- $m(\pi_t) = 0$

$$\begin{aligned}\mathbb{E}_{\tilde{\pi}\sim\omega_t}\eta^2\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))^2 &= \eta^2\textstyle\sum_{\tilde{\pi}\in\Pi} w_t(\tilde{\pi})\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))^2 \\ &\leq \tfrac{\eta^2}{1-\gamma}\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))^2 \ \left(\because \tfrac{w_t(\tilde{\pi})}{p_t(\tilde{\pi})} \leq \tfrac{1}{1-\gamma}\right) \\ &= \tfrac{\eta^2}{1-\gamma}\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\left(\tfrac{g_t(\tilde{\pi})\mathbb{1}(\tilde{\pi}\in\Pi_t^*)}{\sum_{\pi'\in\Pi_t^*} p_t(\pi')} + \tfrac{\beta}{p_t(\tilde{\pi})}\right)\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) \\ &\leq \tfrac{\eta^2}{1-\gamma}\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\left(\tfrac{g_t(\tilde{\pi})\mathbb{1}(\tilde{\pi}\in\Pi_t^*)+\beta}{p_t(\tilde{\pi})}\right)\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) \\ &\leq \eta^2\tfrac{1+\beta}{1-\gamma}\textstyle\sum_{\tilde{\pi}\in\Pi}\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))\end{aligned}$$

- $m(\pi_t) = 1$

$$\begin{aligned}\mathbb{E}_{\tilde{\pi}\sim\omega_t}\eta^2\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))^2 &= \eta^2\textstyle\sum_{\tilde{\pi}\in\Pi} w_t(\tilde{\pi})\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))^2 \\ &\leq \tfrac{\eta^2}{1-\gamma}\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))^2 \ \left(\because \tfrac{w_t(\tilde{\pi})}{p_t(\tilde{\pi})} \leq \tfrac{1}{1-\gamma}\right) \\ &= \tfrac{\eta^2}{1-\gamma}\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\left(\tfrac{g_t(\tilde{\pi})\mathbb{1}(\tilde{\pi}\in\Pi_t(\pi_t))}{\sum_{\Pi_t(\pi_t)} p_t(\pi')} + \tfrac{\beta}{p_t(\tilde{\pi})}\right)\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) \\ &\leq \tfrac{\eta^2}{1-\gamma}\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\left(\tfrac{g_t(\tilde{\pi})\mathbb{1}(\tilde{\pi}\in\Pi_t^*)+\beta}{p_t(\tilde{\pi})}\right)\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) \\ &\leq \eta^2\tfrac{1+\beta}{1-\gamma}\textstyle\sum_{\tilde{\pi}\in\Pi}\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))\end{aligned}$$

Step 3: Summing. Let $\tilde{G}_0(\tilde{\pi}) = 0$. Then, combining Eq. 32-Eq. 33 and summing over t yield

$$\begin{aligned}&-\sum_{t=1}^{T}\mathbb{E}_{\tilde{\pi}\sim p_t}\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) \\ &\leq (1+\beta)\eta\sum_{t=1}^{T}\sum_{\tilde{\pi}\in\Pi}\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) - \frac{1-\gamma}{\eta}\sum_{t=1}^{T}\ln\left(\sum_{\tilde{\pi}\in\Pi} w_t(\tilde{\pi})\exp(\eta\tilde{g}_t(\tilde{\pi}\mid\Pi_t(\pi_t)))\right) \\ &= (1+\beta)\eta\sum_{t=1}^{T}\sum_{\tilde{\pi}\in\Pi}\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) - \frac{1-\gamma}{\eta}\ln\left(\frac{\sum_{\tilde{\pi}\in\Pi}\exp(\eta\tilde{G}_T(\tilde{\pi}))}{\sum_{\tilde{\pi}\in\Pi}\exp(\eta\tilde{G}_0(\tilde{\pi}))}\right) \ (\because \text{ Definition of } \omega_t(\tilde{\pi}), \tilde{G}_t(\tilde{\pi})) \\ &\leq (1+\beta)\eta K\max_{\tilde{\pi}\in\Pi}\tilde{G}_T(\tilde{\pi}) + \frac{\ln K}{\eta} - \frac{1-\gamma}{\eta}\ln\left(\sum_{\tilde{\pi}\in\Pi}\exp(\eta\tilde{G}_T(\tilde{\pi}))\right) \ (\because 1-\gamma\leq 1 \text{ and } \tilde{G}_0(\tilde{\pi})=0) \\ &\leq -(1-\gamma-(1+\beta)\eta K)\max_{\tilde{\pi}\in\Pi}\tilde{G}_T(\tilde{\pi}) + \frac{\ln K}{\eta} \\ &\leq -(1-\gamma-(1+\beta)\eta K)\max_{\tilde{\pi}\in\Pi}\sum_{t=1}^{T} g_t(\tilde{\pi}) + \frac{\ln(K\delta^{-1})}{\beta} + \frac{\ln K}{\eta},\end{aligned} \tag{34}$$

where the last inequality holds due to the Lemma 3, union bound (the reason for using the confidence term of $\frac{\delta}{K}$), and the initial assumption that $\gamma \leq \frac{1}{2}$ and $(1+\beta)K\eta \leq \gamma$. Plugging Eq. 34 into Eq. 31, the following holds with probability $1 - \frac{\delta}{K}$ for all $\pi \in \Pi$:

$$\begin{aligned}R_\pi(T,\alpha) &\leq (\ell_{\max}-\ell_{\min})\left(K\beta T + o(T)^{-1}T + \sum_{t=1}^{T} g_t(\pi) - \sum_{t=1}^{T}\mathbb{E}_{\tilde{\pi}\sim p_t}\tilde{g}_t(\tilde{\pi}\mid\Pi_t(\pi_t))\right) \\ &\leq (\ell_{\max}-\ell_{\min})\left(K\beta T + o(T)^{-1}T + \sum_{t=1}^{T} g_t(\pi) - (1-\gamma-(1+\beta)\eta K)\max_{\tilde{\pi}\in\Pi}\sum_{t=1}^{T} g_t(\tilde{\pi}) + \frac{\ln(K\delta^{-1})}{\beta} + \frac{\ln K}{\eta}\right) \\ &\leq (\ell_{\max}-\ell_{\min})\left(K\beta T + o(T)^{-1}T + \gamma T + (1+\beta)\eta KT + \frac{\ln(K\delta^{-1})}{\beta} + \frac{\ln K}{\eta}\right).\end{aligned}$$

The last inequality holds when we set $\lambda > 0$ to be the one such that $\varepsilon(\lambda,\alpha) = o(T)^{-1}$, which we set $\frac{1}{\sqrt{T}}$ in our algorithm. Since $\textbf{Reg}(T,\alpha) := \max R_\pi(T,\alpha)$, this completes the proof by taking the union bound.

H PROOF OF THEOREM 5 FOR EXP3.P-CP-UNLOCK

H.1 BIASED UNLOCKING ESTIMATOR AND ITS PROPERTIES

Definition. First of all, we consider the following biased unlocking estimator $\tilde{g}_t(\pi \mid \Pi_t(\pi_t))$ under the semi-bandit feedback scenario as the following:

$$\tilde{g}_t(\pi \mid \Pi_t(\pi_t)) \coloneqq \underbrace{\mathbb{1}(m_t(\pi_t) = 0)}_{\text{full unlocking}} \times (A) + \underbrace{\mathbb{1}(m_t(\pi_t) = 1)}_{\text{partial unlocking}} \times (B). \tag{35}$$

Letting $\Pi_t^* \coloneqq \{\tilde{\pi} \in \Pi : \tilde{\pi} \le f_t(x_t, y_t)\}$,

$$(A) \coloneqq \mathbb{1}(\pi \in \Pi_t^*) \left\{ \frac{g_t(\pi)}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})} + \left(1 + \frac{1}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})}\right) \beta \right\} + \mathbb{1}(\pi \in (\Pi_t^*)^c) \left\{ g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi} \le \pi} p_t(\tilde{\pi})} \right\}$$

and

$$(B) \coloneqq \mathbb{1}(\pi \in \Pi_t(\pi_t)) \left\{ g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi} \le \pi} p_t(\tilde{\pi})} \right\} + \mathbb{1}(\pi \in \Pi_t(\pi_t)^c) \left\{ \tilde{g}_t(\pi) + \left(1 + \frac{1}{p_t(\pi)}\right) \beta \right\}.$$

In addition, the unlocking set $\Pi_t(\pi_t) \subset \Pi$ is defined as follows:

$$\Pi_t(\pi_t) = \begin{cases} \Pi & \text{if } m_t(\pi_t) = 0 \\ \{\tilde{\pi} \in \Pi : \tilde{\pi} \ge \pi_t\} & \text{if } m_t(\pi_t) = 1. \end{cases}$$

Note that $m_t(\pi_1) \le m_t(\pi_2) \; \forall \, \pi_1 \le \pi_2$ due to the monotonicity property of the conformal set with respect to the miscoverage, *i.e.*, $\mathbb{1}(y_t \notin C_{\pi_1}(x_t)) \le \mathbb{1}(y_t \notin C_{\pi_2}(x_t))$ whenever $\pi_1 \le \pi_2$.

Besides, we use a **pseudo-gain** $\tilde{g}_t(\pi) \; \forall \pi \in \Pi_t(\pi_t)^c$, since we only observe the true gain only when $\pi \in \Pi_t(\pi_t) \subset (\Pi_t^*)^c$ if $m_t(\pi_t) = 1$, *i.e.*, $y_t \notin C_{\pi_t}(x_t)$. We will formally define the term in subsequent sections.

Properties of the Unlocking Set. The followings are properties of the unlocking set $\Pi_t(\tilde{\pi})$:

- $\tilde{\pi} \in \Pi_t(\tilde{\pi})$
- When $m_t(\tilde{\pi}) = 0$ (full feedback), the unlocking set is $\tilde{\pi}$-independent, $\Pi = \Pi_t^* \cup (\Pi_t^*)^c$.

H.2 LOSS FUNCTION AND ITS PROPERTIES

Definition. Now, we introduce our loss estimator $\ell_t(\pi, c) = d_t(\pi, c) + \frac{\lambda\alpha}{\lambda+2} a_t(\pi)$ $(c \in (0, 0.5), \lambda > 0)$ and its intuition behind. First, $d_t(\pi, c) \coloneqq |\mathbb{1}(y_t \notin C_\pi(x_t)) - c|$ is defined as the **miscoverage loss**. Note that the conversion lemma (Lemma 1) ensures the convergence to target coverage $1 - \alpha$ when $c = \frac{\alpha}{\lambda+2}$. Second, $a_t(\pi)$ is the **penalty term to optimize the set size**.

Rationale for the Design of $a_t(\pi)$. Recalling that (1) the miscoverage loss is of primary importance in conformal prediction and (2) the binary search-type algorithm is implemented in the batch learning set-up (ref.), we define $a_t(\pi)$ as the following:

$$a_t(\pi) \coloneqq \mathbb{1}(m_t(\pi) = 0) \exp\left(-\frac{\pi}{o(T)}\right) + \mathbb{1}(m_t(\pi) = 1) \exp\left(-\frac{1 - \pi}{o(T)}\right)$$

Here, we set denominator inside the exponential to be $\sqrt{T}$, which can be any of the form $o(T) = T^k \; \forall k \in [0.5, 1)$ such that $\frac{o(T)}{T} \to 0$ as $T \to \infty$. Such denominator is necessary for the regret analysis, which will be described in detail in subsequent sections.

Intuitively, if $\pi \in \Pi_t^*$, *i.e.*, $y_t \in C_\pi(x_t)$, $a_t(\pi)$ takes small value as the size of the conformal set is **small** ($|C_\pi(x_t)| \downarrow$). This has the effect to maintain the conformal set to be as small as possible as long as $y_t \in C_\pi(x_t)$. On the other hand, if $\pi \in (\Pi_t^*)^c$, *i.e.*, $y_t \notin C_\pi(x_t)$, $a_t(\pi)$ takes small value as the size of the conformal set is **large** ($|C_\pi(x_t)| \uparrow$). This ensures the penalty term to take the miscoverage loss, instead of the set size efficiency, of priority importance when $y_t \notin C_\pi(x_t)$.

Actually, the same intuition is considered in the design of the biased unlocking estimator $\tilde{g}_t(\pi|\Pi_t(\pi_t))$ (Eq. 35). Specifically, if we look at cases where $(\pi \in \Pi_t^*)^c$, $\mathbb{1}(\pi \in (\Pi_t^*)^c) \left\{ \frac{\beta}{\sum_{\tilde{\pi} \le \pi} p_t(\tilde{\pi})} \right\}$ in (A) and $\mathbb{1}(\pi \in \Pi_t(\pi_t)) \left\{ \frac{\beta}{\sum_{\tilde{\pi} \le \pi} p_t(\tilde{\pi})} \right\}$ in (B), we observe that the denominator of the bonus term β is defined as $\sum_{\tilde{\pi} \le \pi} p_t(\tilde{\pi})$. Since it is a increasing function in π, the bonus term is modeled to be large when $\pi \in (\Pi_t^*)^c$ ($y_t \notin C_\pi(x_t)$) and π is small ($|C_\pi(x_t)| \uparrow$).

Properties of the Loss Estimator. Then, the followings are properties of our loss estimator $\ell_t(\pi, c) = d_t(\pi, c) + \frac{\lambda\alpha}{\lambda+2} a_t(\pi)$.

- By the definition of $d_t(\pi, c)$, $d_t(\pi, c) = c \, \forall \pi \in \Pi_t^*$ and $d_t(\pi, c) = 1 - c \, \forall \pi \in (\Pi_t^*)^c$. Note that as long as $c \in (0, 0.5)$,

$$c < 1 - c. \tag{36}$$

- For all $\pi_1, \pi_2 \in \Pi_t^*$ such that $\pi_1 \le \pi_2$,

$$a_t(\pi_1) \ge a_t(\pi_2) \ (\because \ \exp(-\frac{\pi}{\sqrt{T}}) \text{ is monotonically decreasing}).$$

  Therefore, for all $\pi_1, \pi_2 \in \Pi_t^*$ such that $\pi_1 \le \pi_2$,

$$\ell_t(\pi_1, c) \ge \ell_t(\pi_2, c). \tag{37}$$

- For all $\pi_1, \pi_2 \in (\Pi_t^*)^c$ such that $\pi_1 \le \pi_2$,

$$a_t(\pi_1) \le a_t(\pi_2) \ (\because \ \exp(-\frac{1-\pi}{\sqrt{T}}) \text{ is monotonically increasing}).$$

  Therefore, for all $\pi_1, \pi_2 \in \Pi_t^*$ such that $\pi_1 \le \pi_2$,

$$\ell_t(\pi_1, c) \le \ell_t(\pi_2, c). \tag{38}$$

- Let $\ell_{t,0} \coloneqq \max_{\tilde{\pi} \in \Pi_t^*} \ell_t(\tilde{\pi}, c)$ and $\ell_{t,1} \coloneqq \min_{\tilde{\pi} \in (\Pi_t^*)^c} \ell_t(\tilde{\pi}, c)$. Due to the two properties above (Eq. 37-38), by letting $\pi_{t,0} \coloneqq \min_{\tilde{\pi} \in \Pi_t^*} \tilde{\pi} = 0$ and $\pi_{t,1} \coloneqq \min_{\tilde{\pi} \in (\Pi_t^*)^c} \tilde{\pi}$,

$$\ell_{t,0} = \ell_t(\pi_{t,0}, c),$$
$$\ell_{t,1} = \ell_t(\pi_{t,1}, c).$$

  Since controlling the miscoverage is of utmost importance before optimizing the set size in conformal prediction, the loss estimator will be most satisfactory when $\ell_{t,0} \le \ell_{t,1} \, \forall t \in [T]$. This is true when $\ell_{t,1} - \ell_{t,0} = (1 - 2c) + \frac{\lambda\alpha}{\lambda+2} \left\{ \exp(-\frac{1-\pi_{t,1}}{\sqrt{T}}) - \exp(0) \right\} \ge 0$. Note that it holds if we set $c = \frac{\alpha}{\lambda+2}$ as our proposed algorithm does:

$$\begin{aligned}
\ell_{t,1} - \ell_{t,0} &= (1 - 2\frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2} \left\{ \exp(-\frac{1-\pi_{t,1}}{\sqrt{T}}) - \exp(0) \right\} \\
&\ge (1 - 2\frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2} \left\{ -\exp(0) \right\} \\
&= 1 - \alpha \\
&\ge 0.
\end{aligned}$$

  Therefore,

$$\ell_t(\pi_1, \frac{\alpha}{\lambda+2}) \le \ell_t(\pi_2, \frac{\alpha}{\lambda+2}) \ \forall \pi_1 \in \Pi_t^*, \ \forall \pi_2 \in (\Pi_t^*)^c. \tag{39}$$

The properties of our proposed loss estimator $\ell_t(\pi, c)$ above hold for all $t \in [T]$ and $c \in (0, 0.5)$. However, we only consider the case where $c = \frac{\alpha}{\lambda+2}$ henceforth, since it is the condition that the conversion lemma requires for the coverage guarantee (Lemma 1).

Here, we define the normalized gain $g_t(\pi)$, which is the one used to define the biased unlocking estimator (Eq. 35) as follows:

$$g_t(\pi) = \frac{\ell_{\text{max}} - \ell_t(\pi, \frac{\alpha}{\lambda+2})}{\ell_{\text{max}} - \ell_{\text{min}}} \ \forall \pi \in \Pi,$$

where $\ell_{\text{max}} \coloneqq (1 - \frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2}$ and $\ell_{\text{min}} \coloneqq \frac{\alpha}{\lambda+2}$. Since we only consider the case where $c = \frac{\alpha}{\lambda+2}$ in the algorithm, we use the notation $g_t(\pi)$ instead of $g_t(\pi, \frac{\alpha}{\lambda+2})$ for brevity.

Useful Inequalities. Based on these properties of our proposed loss estimator, by letting $c = \frac{\alpha}{\lambda+2}$ as our algorithm, the following inequalities hold for each case:

- Case 1 ($m_t(\pi_t) = 0$)
    1. $\pi \in \Pi_t^*$ and $\pi \leq \pi_t$: $\ell_t(\pi, \frac{\alpha}{\lambda+2}) \geq \ell_t(\pi_t, \frac{\alpha}{\lambda+2}) \Leftrightarrow g_t(\pi) \leq g_t(\pi_t)$ ($\because$ Eq. 37)
    2. $\pi \in \Pi_t^*$ and $\pi > \pi_t$: $\ell_t(\pi, \frac{\alpha}{\lambda+2}) \leq \ell_t(\pi_t, \frac{\alpha}{\lambda+2}) \Leftrightarrow g_t(\pi) \geq g_t(\pi_t)$, where $\ell_t(\pi_t, \frac{\alpha}{\lambda+2}) - \ell_t(\pi, \frac{\alpha}{\lambda+2}) = \frac{\lambda\alpha}{\lambda+2}\left\{\exp(-\frac{\pi_t}{o(T)}) - \exp(-\frac{\pi}{o(T)})\right\} = o(T)^{-1}$ ($\because$ Taylor expansion). Therefore, since $g_t(\pi) - g_t(\pi_t) = \frac{\frac{\lambda\alpha}{\lambda+2}\left\{\exp(-\frac{\pi_t}{o(T)}) - \exp(-\frac{\pi}{o(T)})\right\}}{\ell_{\text{max}} - \ell_{\text{min}}}$,
    $$g_t(\pi) = g_t(\pi_t) + o(T)^{-1} \ (\because \text{ Eq. 37}).$$
    3. $\pi \in (\Pi_t^*)^c$: $\ell_t(\pi) \geq \ell_t(\pi_t) \Leftrightarrow g_t(\pi) \leq g_t(\pi_t)$ ($\because$ Eq. 39).
        - Additionally, for all $\pi \in (\Pi_t^*)^c$,
        $$\begin{aligned} \frac{g_t(\pi)}{g_t(\pi_t)} &= \frac{\ell_{\text{max}} - \ell_t(\pi, \frac{\alpha}{\lambda+2})}{\ell_{\text{max}} - \ell_t(\pi_t, \frac{\alpha}{\lambda+2})} \\ &= \frac{\{(1 - \frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2}\} - \{(1 - \frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2}\exp(-\frac{1-\pi}{\sqrt{T}})\}}{\{(1 - \frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2}\} - \{\frac{\alpha}{\lambda+2} + \frac{\lambda\alpha}{\lambda+2}\exp(-\frac{\pi}{\sqrt{T}})\}} \\ &\leq \frac{\frac{\lambda\alpha}{\lambda+2}}{1 - \frac{2\alpha}{\lambda+2}} \\ &= \frac{\lambda\alpha}{(\lambda+2) - 2\alpha} \eqqcolon \varepsilon(\lambda, \alpha). \end{aligned}$$
        Therefore,
        $$g_t(\pi) = \varepsilon(\lambda, \alpha) g_t(\pi_t) \ \forall \pi \in (\Pi_t^*)^c.$$

    Therefore,
    $$\begin{aligned} g_t(\pi) &\leq g_t(\pi_t) + o(T)^{-1} \ \forall \pi \in \Pi_t^*, \\ g_t(\pi) &\leq \varepsilon(\lambda, \alpha) g_t(\pi_t) \ \forall \pi \in (\Pi_t^*)^c. \end{aligned} \tag{40}$$
- Case 2 ($m_t(\pi_t) = 1$)
    1. $\pi \in \Pi_t(\pi_t) \subset (\Pi_t^*)^c$: $\ell_t(\pi) \geq \ell_t(\pi_t) \Leftrightarrow g_t(\pi) \leq g_t(\pi_t)$ ($\because$ Eq. 38)
    2. We use a **pseudo-gain** $\tilde{g}_t(\pi) \ \forall \pi \in \Pi_t(\pi_t)^c$, since we only observe the true gain only when $\pi \in \Pi_t(\pi_t) \subset (\Pi_t^*)^c$ if $m_t(\pi_t) = 1$. First, note that
    $$\Pi_t(\pi_t)^c = \Pi_t^* \cup \left[(\Pi_t^*)^c - \Pi_t(\pi_t)\right].$$
    Second, we define the **pseudo-gain** $\tilde{g}_t(\pi) \ \forall \pi \in \Pi_t(\pi_t)^c$ in our algorithm as follows:
    $$\tilde{g}_t(\pi) \coloneqq \frac{\ell_{\text{max}} - \tilde{\ell}_t(\pi)}{\ell_{\text{max}} - \ell_{\text{min}}} \tag{41}$$
    $$= \frac{\ell_{\text{max}} - \{(1 - \frac{\alpha}{\lambda+2}) + \frac{\lambda\alpha}{\lambda+2}\exp(-\frac{1-\pi}{\sqrt{T}})\}}{\ell_{\text{max}} - \ell_{\text{min}}}, \tag{42}$$
    which satisfies the following properties: For all $\pi \in \Pi_t(\pi_t)^c$,
    $$\begin{aligned} \tilde{g}(\pi) - g_t(\pi_t) &= \frac{\frac{\lambda\alpha}{\lambda+2}\{\exp(-\frac{1-\pi}{\sqrt{T}}) - \exp(-\frac{1-\pi_t}{\sqrt{T}})\}}{\ell_{\text{max}} - \ell_{\text{min}}} \\ &\leq o(T)^{-1} \ (\because \text{ Taylor expansion}). \end{aligned}$$

Therefore,

$$
\begin{aligned}
&g_t(\pi) \leq g_t(\pi_t) \; \forall \pi \in \Pi(\pi_t), \\
&\tilde{g}_t(\pi) \leq g_t(\pi_t) + o(T)^{-1} \; \forall \pi \in \Pi(\pi_t)^c.
\end{aligned} \tag{43}
$$

H.3 Expectation and Arm-wise Bounds for the Unlocking Estimator

Based on our preceding results, our biased unlocking estimator satisfies the following inequality:

- Case 1 ($m_t(\pi_t) = 0$)

$$
\begin{aligned}
\mathbb{E}_{\pi \sim p_t} \tilde{g}_t(\pi | \Pi_t(\pi_t)) &= \textstyle\sum_{\pi \in \Pi} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) \\
&= \textstyle\sum_{\pi \in \Pi_t^*} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) + \sum_{\pi \in (\Pi_t^*)^c} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) \\
&= \textstyle\sum_{\pi \in \Pi_t^*} p_t(\pi) \left\{ \frac{g_t(\pi) + \beta}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})} + \beta \right\} + \sum_{\pi \in (\Pi_t^*)^c} p_t(\pi) \left\{ g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi} \leq \pi} p_t(\tilde{\pi})} \right\} \\
&\leq (1 + \varepsilon(\lambda, \alpha)) g_t(\pi_t) + o(T)^{-1} + \left(2 + \frac{\sum_{\tilde{\pi} \in (\Pi_t^*)^c} p_t(\tilde{\pi})}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})} \right) \beta \; (\because \text{Eq. } 40)
\end{aligned} \tag{44}
$$

- Case 2 ($m_t(\pi_t) = 1$)

$$
\begin{aligned}
\mathbb{E}_{\pi \sim p_t} \tilde{g}_t(\pi | \Pi_t(\pi_t)) &= \textstyle\sum_{\pi \in \Pi} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) \\
&= \textstyle\sum_{\pi \in \Pi_t(\pi_t)} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) + \sum_{\pi \in \Pi_t(\pi_t)^c} p_t(\pi) \tilde{g}_t(\pi | \Pi_t(\pi_t)) \\
&= \textstyle\sum_{\pi \in \Pi_t(\pi_t)} p_t(\pi) \left\{ g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi} \leq \pi} p_t(\tilde{\pi})} \right\} + \sum_{\pi \in \Pi_t(\pi_t)^c} p_t(\pi) \left\{ \tilde{g}_t(\pi) + \frac{\beta}{p_t(\pi)} + \beta \right\} \\
&\leq g_t(\pi_t) + o(T)^{-1} + \left(1 + |\Pi_t(\pi_t)^c| + \frac{\sum_{\tilde{\pi} \in \Pi_t(\pi_t)} p_t(\tilde{\pi})}{\sum_{\tilde{\pi} \leq \pi} p_t(\tilde{\pi})} \right) \beta \; (\because \text{Eq. } 43)
\end{aligned} \tag{45}
$$

Combining Eq. 44 and Eq. 45, the following upper bound holds irrespective of the choice of π_t:

$$
\begin{aligned}
\mathbb{E}_{\pi \sim p_t} \tilde{g}_t(\pi | \Pi_t(\pi_t)) \leq\; & (1 + \varepsilon(\lambda, \alpha)) g_t(\pi_t) + o(T)^{-1} \\
& + \underbrace{\left(1 + \{ \mathbb{1}(m_t(\pi_t) = 0) 1 + \mathbb{1}(m_t(\pi_t) = 1) |\Pi_t(\pi_t)^c| \} + \frac{\sum_{\tilde{\pi} \in (\Pi_t^*)^c} p_t(\tilde{\pi})}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})} \right)}_{C_t} \beta.
\end{aligned} \tag{46}
$$

Arm-wise High Probability Bound. Before moving on to the regret analysis, we provide the following lemma, a variant of Lemma 2, which characterizes the property of our biased gain estimator under the semi-bandit feedback scenario.

Lemma 4. *For $\beta \leq 1$ and $g_t(\cdot) \in [0, 1]$, define $\tilde{g}_t(\pi) \in [0, \infty)$ as Eq. 35. Then, for each $\pi \in \Pi$, the following holds with probability at least $1 - \delta$:*

$$
\sum_{t=1}^{T} g_t(\pi) \leq \sum_{t=1}^{T} \tilde{g}_t(\pi) + \frac{\ln(\delta^{-1})}{\beta}.
$$

Proof. **Step 1: Useful Decomposition.**
Let $\mathbb{E}_t$ be the expectation conditioned on $\pi_1, ... \pi_{t-1}$. First, we decompose the estimator in Eq. 35 as the following:

$$
\begin{aligned}
\tilde{g}_t(\pi \mid \Pi_t(\pi_t)) &:= \underbrace{\mathbb{1}(m_t(\pi_t) = 0)}_{\text{full unlocking}} \times (A) + \underbrace{\mathbb{1}(m_t(\pi_t) = 1)}_{\text{partial unlocking}} \times (B) \\
&= \underbrace{\mathbb{1}(m_t(\pi_t) = 0)}_{\text{full unlocking}} \times \{(A1) + (A2)\} + \underbrace{\mathbb{1}(m_t(\pi_t) = 1)}_{\text{partial unlocking}} \times \{(B1) + (B2)\} \\
&= \underbrace{\{\mathbb{1}(m_t(\pi_t)=0) \times (A1) + \mathbb{1}(m_t(\pi_t)=1) \times (B1)\}}_{(C1)} + \underbrace{\{\mathbb{1}(m_t(\pi_t)=0) \times (A2) + \mathbb{1}(m_t(\pi_t)=1) \times (B2)\}}_{(C2)}
\end{aligned}
$$

where

$$(A)=\underbrace{\mathbb{1}(\pi\in\Pi_t^*)\frac{g_t(\pi)}{\sum_{\tilde{\pi}\in\Pi_t^*}p_t(\tilde{\pi})}+\mathbb{1}(\pi\in(\Pi_t^*)^c)g_t(\pi)}_{(A1)}+\underbrace{\frac{\beta}{\mathbb{1}(\pi\in\Pi_t^*)\sum_{\tilde{\pi}\in\Pi_t^*}p_t(\tilde{\pi})+\mathbb{1}(\pi\in(\Pi_t^*)^c)\sum_{\tilde{\pi}\leq\pi}p_t(\tilde{\pi})}+\mathbb{1}(\pi\in\Pi_t^*)\beta}_{(A2)}$$

and

$$(B)=\underbrace{\mathbb{1}(\pi\in\Pi_t(\pi_t))g_t(\pi)+\mathbb{1}(\pi\in\Pi_t(\pi_t)^c)\tilde{g}_t(\pi)}_{(B1)}+\underbrace{\frac{\beta}{\mathbb{1}(\pi\in\Pi_t(\pi_t))\sum_{\tilde{\pi}\leq\pi}p_t(\tilde{\pi})+\mathbb{1}(\pi\in\Pi_t(\pi_t)^c)p_t(\pi)}+\mathbb{1}(\pi\in\Pi_t(\pi_t)^c)\beta}_{(B2)}.$$

Step 2: Show $\Delta_t(\pi_t) \leq 1$.
Next, we show the following two claims are true:

- Claim 1 ($m_t(\pi_t) = 0$)
  - $\pi \in \Pi_t^*$
  $$\beta g_t(\pi) - \beta \times (A1) = \beta g_t(\pi) - \beta \frac{g_t(\pi)}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})} \leq 0$$
  - $\pi \in (\Pi_t^*)^c$
  $$\beta g_t(\pi) - \beta \times (A1) = \beta g_t(\pi) - \beta g_t(\pi) = 0$$
- Claim 2 ($m_t(\pi_t) = 1$): We consider two cases where $\pi \in \Pi_t(\pi_t)$ and $\pi \in \Pi_t(\pi_t)^c$.
  - $\pi \in \Pi_t(\pi_t)$
  $$\beta g_t(\pi) - \beta \times (B1) = \beta g_t(\pi) - \beta g_t(\pi) = 0$$
  - $\pi \in \Pi_t(\pi_t)^c$
  $$\beta g_t(\pi) - \beta \times (B1) = \beta g_t(\pi) - \beta \tilde{g}_t(\pi) \leq 1 \ (\because\ \beta \leq 1 \text{ and } g_t(\pi), \tilde{g}_t(\pi) \in [0, 1])$$

Given $\pi \in \Pi$, Claim 1 and Claim 2 always hold irrespective of the choice of π_t by the algorithm. Then, by letting $\Delta_t(\pi_t) := \beta g_t(\pi) - \beta \times (C1)$, $\Delta_t(\pi_t) \leq 1$.

Step 3: Show $\mathbb{E}$ exp $\left[\sum_{t=1}^{T} (\Delta_t(\pi_t) - \beta \times (C2))\right] \leq 1$.
Therefore, since (1) $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ and (2) $\Delta_t(\pi_t) \leq 1$, for $\beta \leq 1$, we have

$$\begin{aligned}\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t) - \beta \times (C2)\right)\right] &\leq \mathbb{E}_t\left[(1 + \Delta_t(\pi_t) + \Delta_t(\pi_t)^2) \times \exp(-\beta \times (C2))\right] \\ &\leq \left\{\mathbb{1}(\pi\in\Pi_t^*)\exp\left(-\frac{\beta^2}{\sum_{\tilde{\pi}\in\Pi_t^*}p_t(\tilde{\pi})}-\beta^2\right)+\mathbb{1}(\pi\in(\Pi_t^*)^c)\exp\left(-\frac{\beta^2}{\sum_{\tilde{\pi}\leq\pi}p_t(\tilde{\pi})}\right)\right\} \\ &\qquad \times\mathbb{E}_t\left[(1+\Delta_t(\pi_t)+\Delta_t(\pi_t)^2)\right]\end{aligned}$$

Since π is fixed, we consider each case where $\pi \in \Pi_t^*$ and $\pi \in (\Pi_t^*)^c$.

Now, our goal is to show that

$$\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t) - \beta \times (C2)\right)\right] \leq 1 \ \forall t \in [T].$$

Step 3-1: $\pi \in \Pi_t^*$.
Step 3-1-1: $\mathbb{E}_t\left[\Delta_t(\pi_t)\right]$.

$$\begin{aligned}\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\Delta_t(\tilde{\pi}) &= \beta g_t(\pi) - \sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi}) \times \beta(C1) \\ &\leq \beta g_t(\pi) - \beta \sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi}) \frac{g_t(\pi)}{\sum_{\pi'\in\Pi_t^*} p_t(\pi')} \\ &= 0.\end{aligned}$$

Step 3-1-2: $\mathbb{E}_t\Big[\Delta_t(\pi_t)^2\Big]$**.**

$$\begin{aligned}\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 &= \sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 \\ &= (\beta g_t(\pi))^2 \sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\left(1-\frac{1}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})}\right)^2 + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta\tilde{g}_t(\pi))^2 \\ &\le \frac{\beta^2}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})} + \beta^2 \; (\because \; g_t(\pi),\tilde{g}_t(\pi)\in[0,1])\end{aligned}$$

Step 3-1-3: Combine. Combining the above results and applying the fact that $1+x\le\exp(x)$ is suffice as follows:

$$\begin{aligned}\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t)-\beta\times(C2)\right)\right] &\le \mathbb{E}_t\Big[(1+\Delta_t(\pi_t)+\Delta_t(\pi_t)^2)\times\exp(-\beta\times(C2))\Big] \\ &= \exp\left(-\frac{\beta^2}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})}-\beta^2\right)\mathbb{E}_t\Big[(1+\Delta_t(\pi_t)+\Delta_t(\pi_t)^2)\Big] \\ &\le \exp\left(-\frac{\beta^2}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})}-\beta^2\right)\left(1+\frac{\beta^2}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})}+\beta^2\right) \\ &\le 1 \; (\because \; 1+x\le\exp(x)).\end{aligned}$$

Step 3-2: $\pi\in(\Pi_t^*)^c$**.**

Step 3-2-1: $\mathbb{E}_t\Big[\Delta_t(\pi_t)\Big]$**.**

$$\begin{aligned}\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\Delta_t(\tilde{\pi}) &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\Delta_t(\tilde{\pi}) + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi})\Delta_t(\tilde{\pi}) \\ &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\Delta_t(\tilde{\pi}) + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})\Delta_t(\tilde{\pi}) + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})\Delta_t(\tilde{\pi}) \\ &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi)) + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi)) \\ &\qquad \textstyle+ \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta\tilde{g}_t(\pi)) \\ &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi)) + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi)) \\ &\qquad \textstyle+ \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi)) \; (\because \text{Eq. 41}) \\ &= 0.\end{aligned}$$

Step 3-2-2: $\mathbb{E}_t\Big[\Delta_t(\pi_t)^2\Big]$**.**

$$\begin{aligned}\textstyle\sum_{\tilde{\pi}\in\Pi} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 \\ &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})\Delta_t(\tilde{\pi})^2 \\ &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi))^2 + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi))^2 \\ &\qquad \textstyle+ \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta\tilde{g}_t(\pi))^2 \\ &= \textstyle\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi))^2 + \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi))^2 \\ &\qquad \textstyle+ \sum_{\tilde{\pi}:\pi\in\Pi_t(\tilde{\pi})^c} p_t(\tilde{\pi})(\beta g_t(\pi)-\beta g_t(\pi))^2 \; (\because \text{Eq. 41}) \\ &= 0.\end{aligned}$$

Step 3-2-3: Combine. Combining the above results,

$$\begin{aligned}\mathbb{E}_t\left[\exp\left(\Delta_t(\pi_t)-\beta\times(C2)\right)\right] &\le \mathbb{E}_t\Big[(1+\Delta_t(\pi_t)+\Delta_t(\pi_t)^2)\times\exp(-\beta\times(C2))\Big] \\ &= \exp(-\frac{\beta^2}{\sum_{\tilde{\pi}\le\pi} p_t(\tilde{\pi})})\mathbb{E}_t\Big[(1+\Delta_t(\pi_t)+\Delta_t(\pi_t)^2)\Big] \\ &= \exp(-\frac{\beta^2}{\sum_{\tilde{\pi}\le\pi} p_t(\tilde{\pi})}) \\ &\le 1.\end{aligned}$$

By sequentially applying the double expectation rule for $t = T, \ldots, 1$,

$$\mathbb{E} \exp\left[\sum_{t=1}^{T} \left(\Delta_t(\pi_t) - \beta \times (C2)\right)\right] \leq 1. \tag{47}$$

Moreover, from the Markov's inequality, we have $\mathbb{P}\left(X > \ln(1/\delta)\right) = \mathbb{P}\left(\exp(X) > 1/\delta\right) \leq \delta \mathbb{E}\exp(X)$. Combined with Eq. 47, we have

$$\beta \sum_{t=1}^{T} g_t(\pi) \leq \beta \sum_{t=1}^{T} \tilde{g}_t(\pi) + \ln(\delta^{-1})$$

with probability at least $1 - \delta$. This completes the proof. □

Proof of the Theorem. We first provide the proof sketch of Theorem 1 as the following.

Proof Sketch. The proof on Theorem 1 is complete, by substituting the above inequality (Eq. 46) to (Eq. 9) in the proof of `EXP3.P` (Theorem 2).

First, we re-express Eq. 46 for the simplicity of proof as the following:

$$-g_t(\pi_t) \leq \frac{1}{1+\varepsilon(\lambda,\alpha)} \Bigg\{ -\mathbb{E}_{\pi \sim p_t} \tilde{g}_t(\pi | \Pi_t(\pi_t)) + o(T)^{-1} + \underbrace{\left(1 + \{\mathbb{1}(m_t(\pi_t) = 0) 1 + \mathbb{1}(m_t(\pi_t) = 1)|\Pi_t(\pi_t)^c|\} + \frac{\sum_{\tilde{\pi} \in (\Pi_t^*)^c} p_t(\tilde{\pi})}{\sum_{\tilde{\pi} \in \Pi_t^*} p_t(\tilde{\pi})}\right)}_{C_t} \beta \Bigg\}. \tag{48}$$

Now, we show the proof on the regret bound of Algorithm 1, which consists of four steps.

First, our goal is to show that, if $\gamma \leq \frac{1}{2}$ and $(1 + 2\beta)K\eta \leq \gamma$,

$$\mathbf{Reg}(T, \alpha) \leq \frac{\ell_{\max} - \ell_{\min}}{1 + \varepsilon(\lambda, \alpha)} \left(C\beta T + 3o(T)^{-1}T + \gamma T + (1 + 2\beta)\eta KT + \frac{\ln(K\delta^{-1})}{\beta} + \frac{\ln K}{\eta} \right), \tag{49}$$

where $C = \min\left(\frac{\sum_{t=1}^{T} C_t}{T}, 4\right)$.

Irrespective of the hyperparameter setup, note that Eq. 49 always holds if $T \geq 5.15\sqrt{TK\ln(K\delta^{-1})}$. If $T < 5.15\sqrt{TK\ln(K\delta^{-1})}$, this implies that $\gamma \leq \frac{1}{2}$ and $(1 + 2\beta)K\eta \leq \gamma$, which makes it suffice to show that Eq. 49 holds for $\gamma \leq \frac{1}{2}$ and $(1 + 2\beta)K\eta \leq \gamma$.

Step 1: Simple equalities. Recall that the gain $g_t(\pi) \in [0, 1]$ is defined with respect to the loss $\ell_t(\pi, \frac{\alpha}{\lambda+2}) \in [\ell_{\min}, \ell_{\max}]$ as the following:

$$g_t(\pi) = \frac{\ell_{\max} - \ell_t(\pi, \frac{\alpha}{\lambda+2})}{\ell_{\max} - \ell_{\min}}.$$

Then, for all $\pi \in \Pi$, the following equality holds:

$$\begin{aligned}
R_\pi(T, \alpha) &= \sum_{t=1}^{T} \ell_t(\pi_t, \frac{\alpha}{\lambda+2}) - \sum_{t=1}^{T} \ell_t(\pi, \frac{\alpha}{\lambda+2}) \\
&= (\ell_{\max} - \ell_{\min}) \left(\sum_{t=1}^{T} g_t(\pi) - \sum_{t=1}^{T} g_t(\pi_t) \right) \ (\because \text{ Definition of } g_t(\cdot)) \\
&\leq \frac{\ell_{\max} - \ell_{\min}}{1 + \varepsilon(\lambda, \alpha)} \left(C\beta T + o(T)^{-1}T + (1 + \varepsilon(\lambda, \alpha)) \sum_{t=1}^{T} g_t(\pi) - \sum_{t=1}^{T} \mathbb{E}_{\tilde{\pi} \sim p_t} \tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \right) \ (\because \text{ Eq. 48}).
\end{aligned} \tag{50}$$

Using the definition of cumulant generating function and the relationship that $p_t = (1-\gamma)\omega_t + \gamma u$ where $\omega_t(\pi) = \frac{\exp\left(\eta \tilde{G}_{t-1}(\pi)\right)}{\sum_{\tilde{\pi}\in\Pi} \exp\left(\eta \tilde{G}_{t-1}(\tilde{\pi})\right)}$ and u is the uniform distribution over K arms, the following holds:

$$
\begin{aligned}
-\mathbb{E}_{\tilde{\pi}\sim p_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) &= -(1-\gamma)\mathbb{E}_{\tilde{\pi}\sim\omega_t}\tilde{g}_t(\pi_i \mid \Pi_t(\pi_t)) - \gamma\mathbb{E}_{\tilde{\pi}\sim u}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \ (\because p_t = (1-\gamma)\omega_t + \gamma u) \\
&= (1-\gamma)\Bigg[\frac{1}{\eta}\ln\mathbb{E}_{\tilde{\pi}\sim\omega_t}\exp\left(\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) - \mathbb{E}_{\tilde{\pi}\sim\omega_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))\right) - \\
&\qquad \frac{1}{\eta}\ln\mathbb{E}_{\tilde{\pi}\sim\omega_t}\exp(\eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)))\Bigg] - \gamma\mathbb{E}_{\tilde{\pi}\sim u}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \quad (51)
\end{aligned}
$$

Step 2: Bounding the first term of Eq. 51. First, we show that irrespective of the choice of π_t,

$$\eta\tilde{g}_t(\pi|\Pi_t(\pi_t)) \leq 1 \ \forall \pi \in \Pi.$$

- $m(\pi_t) = 0$, $\pi \in \Pi_t^*$

$$
\begin{aligned}
\eta\tilde{g}_t(\pi|\Pi_t(\pi_t)) &= \eta\frac{g_t(\pi)+\beta}{\sum_{\tilde{\pi}\in\Pi_t^*} p_t(\tilde{\pi})} + \eta\beta \\
&\leq \frac{\eta(1+2\beta)}{\sum_{\tilde{\pi}\in\Pi_t^*}((1-\gamma)w_t(\tilde{\pi}) + \gamma\frac{1}{K})} \ (\because (1+2\beta)\eta K \leq \gamma) \\
&\leq 1.
\end{aligned}
$$

- $m(\pi_t) = 0$, $\pi \in (\Pi_t^*)^c$; $m(\pi_t) = 1$, $\pi \in \Pi_t(\pi_t)$

$$
\begin{aligned}
\eta\tilde{g}_t(\pi|\Pi_t(\pi_t)) &= \eta\left(g_t(\pi) + \frac{\beta}{\sum_{\tilde{\pi}\leq\pi} p_t(\tilde{\pi})}\right) \\
&\leq \eta\left(\frac{g_t(\pi)+\beta}{\sum_{\tilde{\pi}\leq\pi} p_t(\tilde{\pi})}\right) \\
&\leq \frac{\eta(1+\beta)}{\sum_{\tilde{\pi}\leq\pi}((1-\gamma)w_t(\tilde{\pi}) + \gamma\frac{1}{K})} \\
&\leq \frac{\gamma\frac{1}{K}}{\sum_{\tilde{\pi}\leq\pi}((1-\gamma)w_t(\tilde{\pi}) + \gamma\frac{1}{K})} \ (\because (1+2\beta)\eta K \leq \gamma) \\
&\leq 1.
\end{aligned}
$$

- $m(\pi_t) = 1$, $\pi \in \Pi_t(\pi_t)^c$

$$
\begin{aligned}
\eta\tilde{g}_t(\pi|\Pi_t(\pi_t)) &= \eta\left(\tilde{g}_t(\pi) + \frac{\beta}{p_t(\pi)}\right) \\
&\leq \eta\left(\frac{g_t(\pi)+\beta}{p_t(\pi)}\right) \\
&\leq \frac{\eta(1+\beta)}{(1-\gamma)w_t(\pi) + \gamma\frac{1}{K}} \\
&\leq \frac{\gamma\frac{1}{K}}{(1-\gamma)w_t(\pi) + \gamma\frac{1}{K}} \ (\because (1+2\beta)\eta K \leq \gamma) \\
&\leq 1.
\end{aligned}
$$

Since (1) $\ln x \leq x - 1$, (2) $\exp(x) \leq 1 + x + x^2$ for all $x \leq 1$, and (3) $\eta\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t)) \leq 1$,

$$
\begin{aligned}
&\ln \mathbb{E}_{\tilde{\pi}\sim\omega_t} \exp\left(\eta(\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) - \mathbb{E}_{\tilde{\pi}\sim\omega_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)))\right) \\
&\qquad = \ln \mathbb{E}_{\tilde{\pi}\sim\omega_t} \exp(\eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))) - \eta\mathbb{E}_{\tilde{\pi}\sim\omega_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \\
&\qquad \leq \mathbb{E}_{\tilde{\pi}\sim\omega_t}\left\{\exp(\eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))) - 1 - \eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))\right\} \ (\because \ln x \leq x - 1) \\
&\qquad \leq \mathbb{E}_{\tilde{\pi}\sim\omega_t}\eta^2\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t))^2 \ (\because \exp(x) \leq 1 + x + x^2) \\
&\qquad \leq \eta^2 \frac{1+2\beta}{1-\gamma} K\left(\tilde{g}_t(0|\Pi_t(\pi_t)) + o(T)^{-1}\right),
\end{aligned} \tag{52}
$$

where the last inequality holds due to the following:

- $m(\pi_t) = 0$

$$
\begin{aligned}
\mathbb{E}_{\tilde{\pi}\sim\omega_t}\eta^2\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))^2 &= \eta^2\left\{\sum_{\tilde{\pi}\in\Pi_t^*} w_t(\tilde{\pi})\left(\frac{g_t(\tilde{\pi})+\beta}{\sum_{\pi'\in\Pi_t^*} p_t(\pi')}+\beta\right)^2 + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} w_t(\tilde{\pi})\left(g_t(\tilde{\pi})+\frac{\beta}{\sum_{\pi'\leq\tilde{\pi}} p_t(\pi')}\right)^2\right\} \\
&\leq \eta^2\left\{\sum_{\tilde{\pi}\in\Pi_t^*} w_t(\tilde{\pi})\left(\frac{g_t(\tilde{\pi})+2\beta}{\sum_{\pi'\in\Pi_t^*} p_t(\pi')}\right)^2 + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} w_t(\tilde{\pi})\left(g_t(\tilde{\pi})+\frac{\beta}{\sum_{\pi'\leq\tilde{\pi}} p_t(\pi')}\right)^2\right\} \\
&\leq \eta^2(1+2\beta)\left\{\sum_{\tilde{\pi}\in\Pi_t^*} \frac{w_t(\tilde{\pi})\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))}{\sum_{\pi'\in\Pi_t^*} p_t(\pi')} + \sum_{\tilde{\pi}\in(\Pi_t^*)^c} \frac{w_t(\tilde{\pi})\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))}{p_t(\tilde{\pi})}\right\} \\
&\leq \eta^2 \frac{1+2\beta}{1-\gamma} \underbrace{\left(1+\left|(\Pi_t^*)^c\right|\right)}_{\leq K}\left(\tilde{g}_t(0|\Pi_t(\pi_t))+o(T)^{-1}\right) \ \left(\because \frac{w_t(\tilde{\pi})}{p_t(\tilde{\pi})} \leq \frac{1}{1-\gamma}, \text{Eq. 40}\right)
\end{aligned}
$$

- $m(\pi_t) = 1$

$$
\begin{aligned}
\mathbb{E}_{\tilde{\pi}\sim\omega_t}\eta^2\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))^2 &= \eta^2\left\{\sum_{\tilde{\pi}\in\Pi_t(\pi_t)} w_t(\tilde{\pi})\left(g_t(\tilde{\pi})+\frac{\beta}{\sum_{\pi'\leq\tilde{\pi}} p_t(\pi')}\right)^2 + \sum_{\tilde{\pi}\in\Pi_t(\pi_t)^c} w_t(\tilde{\pi})\left(\tilde{g}_t(\tilde{\pi})+\frac{\beta}{p_t(\tilde{\pi})}+\beta\right)^2\right\} \\
&\leq \eta^2\left\{\sum_{\tilde{\pi}\in\Pi_t(\pi_t)} w_t(\tilde{\pi})\left(g_t(\tilde{\pi})+\frac{\beta}{\sum_{\pi'\leq\tilde{\pi}} p_t(\pi')}\right)^2 + \sum_{\tilde{\pi}\in\Pi_t(\pi_t)^c} w_t(\tilde{\pi})\left(\tilde{g}_t(\tilde{\pi})+\frac{2\beta}{p_t(\tilde{\pi})}\right)^2\right\} \\
&\leq \eta^2(1+2\beta)\left\{\sum_{\tilde{\pi}\in\Pi_t(\pi_t)} \frac{w_t(\tilde{\pi})\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))}{p_t(\tilde{\pi})} + \sum_{\tilde{\pi}\in\Pi_t(\pi_t)^c} \frac{w_t(\tilde{\pi})\tilde{g}_t(\tilde{\pi}|\Pi_t(\pi_t))}{p_t(\tilde{\pi})}\right\} \\
&\leq \eta^2 \frac{1+2\beta}{1-\gamma} K\left(\tilde{g}_t(0|\Pi_t(\pi_t))+o(T)^{-1}\right) \ \left(\because \frac{w_t(\tilde{\pi})}{p_t(\tilde{\pi})} \leq \frac{1}{1-\gamma}, \text{Eq. 43}\right)
\end{aligned}
$$

Step 3: Summing. Let $\tilde{G}_0(\tilde{\pi}) = 0$. Then, combining Eq. 51-Eq. 52 and summing over t yield

$$
\begin{aligned}
&-\sum_{t=1}^{T} \mathbb{E}_{\tilde{\pi}\sim p_t}\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \\
&\qquad \leq (1+2\beta)\eta K \sum_{t=1}^{T}(\tilde{g}_t(0 \mid \Pi_t(\pi_t)) + o(T)^{-1}) - \frac{1-\gamma}{\eta}\sum_{t=1}^{T} \ln\left(\sum_{\tilde{\pi}\in\Pi} w_t(\tilde{\pi})\exp(\eta\tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)))\right) \\
&\qquad = (1+2\beta)\eta K(\tilde{G}_T(0) + To(T)^{-1}) - \frac{1-\gamma}{\eta}\ln\left(\frac{\sum_{\tilde{\pi}\in\Pi}\exp(\eta\tilde{G}_T(\tilde{\pi}))}{\sum_{\tilde{\pi}\in\Pi}\exp(\eta\tilde{G}_0(\tilde{\pi}))}\right) \ (\because \text{ Definition of } \omega_t(\tilde{\pi}), \tilde{G}_t(\tilde{\pi})) \\
&\qquad \leq (1+2\beta)\eta K(\tilde{G}_T(0) + To(T)^{-1}) + \frac{\ln K}{\eta} - \frac{1-\gamma}{\eta}\ln\left(\sum_{\tilde{\pi}\in\Pi}\exp(\eta\tilde{G}_T(\tilde{\pi}))\right) \ (\because 1-\gamma \leq 1 \text{ and } \tilde{G}_0(\tilde{\pi}) = 0) \\
&\qquad \leq -(1 - r - (1+2\beta)\eta K)(\tilde{G}_T(0) + To(T)^{-1}) + \frac{\ln K}{\eta} \ (\because \text{ Property of log-sum-exponential}) \\
&\qquad \leq -(1 - r - (1+2\beta)\eta K)\sum_{t=1}^{T} g_t(0) + \frac{\ln(K\delta^{-1})}{\beta} + \frac{\ln K}{\eta},
\end{aligned} \tag{53}
$$

where the last inequality holds due to the Lemma 4, union bound (the reason for using the confidence term of $\frac{\delta}{K}$), and the initial assumption that $\gamma \leq \frac{1}{2}$ and $(1+2\beta)K\eta \leq \gamma$. Plugging Eq. 53 into Eq. 50, the following holds with probability $1 - \frac{\delta}{K}$ for all $\pi \in \Pi$:

$$
\begin{aligned}
R_\pi(T, \alpha) &\leq \frac{\ell_{\text{max}} - \ell_{\text{min}}}{1 + \varepsilon(\lambda, \alpha)} \left(C\beta T + o(T)^{-1}T + (1 + \varepsilon(\lambda, \alpha)) \sum_{t=1}^{T} g_t(\pi) - \sum_{t=1}^{T} \mathbb{E}_{\tilde{\pi} \sim p_t} \tilde{g}_t(\tilde{\pi} \mid \Pi_t(\pi_t)) \right) \\
&\leq \frac{\ell_{\text{max}} - \ell_{\text{min}}}{1 + \varepsilon(\lambda, \alpha)} \left(C\beta T + 2o(T)^{-1}T + \varepsilon(\lambda, \alpha)T + \gamma T + (1 + 2\beta)\eta K T + \frac{\ln(K\delta^{-1})}{\beta} + \frac{\ln K}{\eta} \right) \\
&\leq (\ell_{\text{max}} - \ell_{\text{min}}) \left(C\beta T + 3o(T)^{-1}T + \gamma T + (1 + 2\beta)\eta K T + \frac{\ln(K\delta^{-1})}{\beta} + \frac{\ln K}{\eta} \right).
\end{aligned}
$$

The last inequality holds when we set $\lambda > 0$ to be the one such that $\varepsilon(\lambda, \alpha) = o(T)^{-1}$, which we set $\frac{1}{\sqrt{T}}$ in our algorithm. Since $\textbf{Reg}(T, \alpha) \coloneqq \max R_\pi(T, \alpha)$, this completes the proof by taking the union bound.