Efficient Online Set-valued Classification with Bandit Feedback

Zhou Wang [1] **Xingye Qiao** [1]

Abstract

Conformal prediction is a distribution-free method that wraps a given machine learning model and returns a set of plausible labels that contain the true label with a prescribed coverage rate. In practice, the empirical coverage achieved highly relies on fully observed label information from data both in the training phase for model fitting and the calibration phase for quantile estimation. This dependency poses a challenge in the context of online learning with bandit feedback, where a learner only has access to the correctness of actions (i.e., pulled an arm) but not the full information of the true label. In particular, when the pulled arm is incorrect, the learner only knows that the pulled one is not the true class label, but does not know which label is true. Additionally, bandit feedback further results in a smaller labeled dataset for calibration, limited to instances with correct actions, thereby affecting the accuracy of quantile estimation. To address these limitations, we propose Bandit Class-specific Conformal Prediction (BCCP), offering coverage guarantees on a class-specific granularity. Using an unbiased estimation of an estimand involving the true label, BCCP trains the model and makes set-valued inferences through stochastic gradient descent. Our approach overcomes the challenges of sparsely labeled data in each iteration and generalizes the reliability and applicability of conformal prediction to online decision-making environments.

1. Introduction

Machine learning models, while highly effective, can fail in complicated scenarios due to inherent uncertainties and hence lead to irreversible consequences, particularly in high-stake applications. For instance, in autonomous vehicle systems, misidentifying real obstacles as harmless shadows on the road potentially causes abrupt braking or even dangerous maneuvers. In medical diagnostics, the challenge of differentiating between benign and malignant tumors in ambiguous cases can result in critical misdiagnoses, influencing treatment decisions. Such scenarios underscore the need for models capable of cautiously handling those observations with high uncertainty.

Quantifying the uncertainty associated with each observation can be addressed by reporting a prediction set, which can be realized by some set-valued classification paradigms such as Classification with the Reject Option (Herbei & Wegkamp, 2006; Bartlett & Wegkamp, 2008; Charoenphakdee et al., 2021; Zhang et al., 2018) and Conformal Prediction (Vovk et al., 2005; Shafer & Vovk, 2008; Balasubramanian et al., 2014). Intuitively speaking, an observation with a large prediction set indicates its intrinsic difficulty and it is hard to be correctly classified. Unlike Classification with the Reject Option, the Conformal Prediction method particularly yields a set with valid prediction coverage, i.e., a prediction set includes the true label with a user-prescribed coverage rate $1-\alpha, \alpha \in [0,1]$.

The literature on Conformal Prediction (and other set-valued classification methods) covers various aspects. For instance, Lei et al. (2013); Lei (2014); Lei et al. (2015; 2018); Sadinle et al. (2019); Wang & Qiao (2018; 2022) consider the coverage guarantees conditional on each class instead of the standard marginal coverage (Vovk et al., 2005). Romano et al. (2020); Angelopoulos et al. (2021) explore different (un)conformity scores to output informative conformal prediction sets. Tibshirani et al. (2019) introduces weighted conformal prediction in the situation of covariate distribution shift, while Hechtlinger et al. (2018); Guan & Tibshirani (2022); Wang & Qiao (2023; 2024) generalize set-valued predictions to the realm of out-of-distribution detection due to the semantic distribution shift by admitting an empty prediction set. However, these studies predominantly focus on the setting with access to full-label information and offline training, limiting their applicability in real-world scenarios.

Recent extensions of Conformal Prediction to online learning settings, (1) address arbitrary distribution shifts (Gibbs & Candes, 2021; Gibbs & Candès, 2022; Zaffran et al., 2022; Bhatnagar et al., 2023), and (2) apply the principles on the off-policy evaluation problem (Taufiq et al., 2022;

[1] Department of Mathematics and Statistics, Binghamton University, New York, USA. Correspondence to: Zhou Wang <zwang198@binghamton.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Zhang et al., 2023; Stanton et al., 2023) in reinforcement learning. Yet, these works require significantly more label information than what bandit feedback affords: in the distribution shift problem with full feedback, a learner knows the true label regardless of its decision's correctness; in the policy evaluation problem, the learner receives a reward that reflects the optimality of the pulled arm. In contrast, in the bandit feedback setting (Langford & Zhang, 2007; Kakade et al., 2008; Wang et al., 2010) a learner only receives feedback about the correctness of predictions rather than the ground truth of label information. For instance, a learner in TikTok can correctly capture a positive attitude toward the video recommendation through a user's click, whereas the user's preferences remain uncertain if the presented recommendation is disliked by the user (it does not know what the user likes). Similarly, in personalized medicine, a medical system adjusts chemotherapy treatments based on partial feedback, such as tumor response, without full knowledge of how other treatments might have worked for that patient.

Motivated by the limited literature on Conformal Prediction within the context of online bandit feedback, we introduce the Bandit Class-specific Conformal Prediction (BCCP) framework for the multi-class classification problem. To the best of our knowledge, this is the first effort in applying conformal prediction to this particular context. Our key contributions are as follows: (1) BCCP leverages an unbiased estimator for accurate ground truth inference of label information, allowing the use of those data instances for which the wrong arm was pulled in both model fitting and quantile estimation; (2) Our method capitalizes on the efficiency of stochastic gradient descent for dynamically updating the quantile estimation, which differentiates itself from the traditional split conformal method in which sample quantiles based on a sufficiently large calibration dataset are used; (3) We theoretically prove that both the class-specific coverage and the excess risk with respect to the check loss converge at a rate of $\mathcal{O}(T^{-1/2})$ under certain conditions; (4) Recognizing the practical challenge of selecting an optimal learning rate for updating the quantile estimation, we use an ensemble approach to update the estimation with a range of learning rates; (5) The effectiveness of BCCP is empirically validated using three different score functions and two policies (for pulling arm) across three datasets, demonstrating the versatility and efficacy of our proposed framework.

The rest of the paper is organized as follows. In Section 2, we begin with a review of the related work. This is followed by Section 3, where we introduce our methodology complemented by a series of associated theorems. In Section 4, we present experiments to demonstrate the effectiveness of our method. The conclusions to our work are given in Section 5 and proofs are attached in Appendix A.

2. Preliminary

In this section, we review some key concepts of Conformal Prediction and the Multi-armed Bandit Problem.

2.1. Conformal Prediction

Conformal prediction (Vovk et al., 2005; Lei et al., 2015) is a distribution-free methodology that can complement various machine learning models, such as neural networks, support vector machines (SVMs), and random forests. It is utilized to produce set-valued predictions with a theoretically guaranteed coverage rate prescribed by users.

Consider a labeled training dataset $\mathcal{D} = \{(\boldsymbol{X}_i, Y_i)\}_{i\in\mathcal{I}}$ ($\mathcal{I}$ denotes the index set) and a test instance $\boldsymbol{X}$ with unknown label Y, where both are assumed to be i.i.d. from an unknown distribution over the domain $\mathcal{X} \times \mathcal{Y}$. In the classification problem, the Standard Conformal Prediction employs a mapping (depending on the dataset $\mathcal{D}$) $\widehat{\mathcal{C}} : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ and returns a prediction set $\widehat{\mathcal{C}}(\boldsymbol{X})$ for the test point $\boldsymbol{X}$, ensuring the marginal coverage rate

$$\mathbb{P}(Y \in \widehat{\mathcal{C}}(\boldsymbol{X})) \geq 1 - \alpha, \tag{1}$$

where $\alpha \in [0, 1]$ represents the pre-specified nominal non-coverage rate by practitioners. Notice that the probability is taken over the training dataset $\mathcal{D}$ and the test point $(\boldsymbol{X}, Y)$.

Considering that the marginal coverage guarantee in Standard Conformal Prediction may not be adequate for certain specific classes, Lei et al. (2013; 2015; 2018); Sadinle et al. (2019) explored Class-conditional Conformal Prediction, which offers class-specific coverage

$$\mathbb{P}(Y \in \widehat{\mathcal{C}}(\boldsymbol{X}) \mid Y = k) \geq 1 - \alpha, \; \forall\, k \in \mathcal{Y}. \tag{2}$$

The same paradigm is also considered in Wang & Qiao (2018; 2022; 2023). It is crucial to understand that while (2) implies (1), the converse is not necessarily true. On the other hand, compared to the marginal coverage, the class-specific coverage may yield larger prediction sets when practitioners have limited data for each class. Motivated by this limitation, Ding et al. (2024) proposed Clustered Conformal Prediction to navigate this trade-off between marginal and class-specific coverage in the low-data regime, while Romano et al. (2020); Angelopoulos et al. (2021) proposed different score functions to improve the prediction set size especially when there are many classes.

In general, Conformal Prediction starts with a (conformity) score function $s : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$. It is employed to gauge the proximity of an observation $\boldsymbol{X}$ to any class $k \in \mathcal{Y}$. Intuitively speaking, the larger the conformity score $s(\boldsymbol{X}, k)$, the higher the likelihood that the observation $\boldsymbol{X}$ belongs to the class k. This score function can manifest in various forms, such as the softmax probability in neural networks, the functional margin in SVMs, or the average predicted class probabilities of trees in random forests.

In the split conformal method (Papadopoulos et al., 2002; Lei et al., 2013), the index set $\mathcal{I}$ associated with the original dataset $\mathcal{D}$ is partitioned into two disjoint subsets: the training part $\mathcal{I}_{tr}$ and the calibration part $\mathcal{I}_{cal}$. The former is used to fit a model $\boldsymbol{f}$ in the training phase, such as training a neural network to minimize cross-entropy loss (Romano et al., 2020; Angelopoulos et al., 2021), training an SVM to minimize hinge loss (Wang & Qiao, 2018; 2022), or growing a random forest based on Gini-impurity (Guan & Tibshirani, 2022). This model $\boldsymbol{f}$ is then utilized to customize the aforementioned conformity score function s. For example, s could be directly taken as $\boldsymbol{f}$, or a monotonic function of $\boldsymbol{f}$, e.g., softmax score. With the conformity score established, the next step involves identifying score thresholds $\tau_k, k \in \mathcal{Y}$ within the calibration part $\mathcal{I}_{cal}$, thereby enabling decision-making for the upcoming test points. In summary, a prediction set for a query $\boldsymbol{X}$ with the class-specific coverage guarantee (2) is defined as

$$\widehat{\mathcal{C}}(\boldsymbol{X}) := \{k \in \mathcal{Y} : s(\boldsymbol{X}, k) \geq \tau_k\},$$

where the threshold τ_k is determined as the $100 \times \alpha\%$ sample quantile of the conformity scores for the calibration set, i.e., the $(\lfloor |\mathcal{I}_{cal}|\alpha \rfloor + 1)$-th smallest value in $\{s(\boldsymbol{X}_i, k)\}_{i \in \mathcal{I}_{cal}}$ (Romano et al., 2019). Throughout this article, $|\cdot|$ being applied on a set denotes the size or cardinality of the set.

2.2. Multi-armed Bandit and Multi-class Classification

The Multi-armed Bandit Problem (Lai & Robbins, 1985; Auer et al., 2002) is a fundamental concept in reinforcement learning. It presents a scenario where a learner aims to optimize rewards or minimize regrets (cumulatively assessed from feedback) by pulling an "arm" (or taking an action), A, from a set of available arms denoted as $\{1, \cdots, K\}$, where K represents the total number of arms. The selection of an arm is guided by a policy π, tailored to maximize expected gains over time. The policy π could be a probability distribution to generate an arm to pull, or deterministic.

When extended to multi-class classification with bandit feedback, this concept incorporates contextual information or features, $\boldsymbol{X}$, effectively transforming it into a contextual bandit problem. Particularly in online learning settings, at time point t, the learner selects an arm $A_t \sim \pi$ for a given query context $\boldsymbol{X}_t$, and subsequently receives binary feedback $\mathbb{1}\{A_t = Y_t\}$. This feedback, indicating whether the pulled arm (class) matches the true label Y_t, introduces uncertainty regarding the true label, complicating the learner's updating process. For example, different from the full feedback setting (Gibbs & Candes, 2021; Gibbs & Candès, 2022; Bhatnagar et al., 2023), the learner here has no idea upon the true label for the query $\boldsymbol{X}_t$ if the value of feedback is 0. Several studies have explored the domain of contextual bandits, where the hypothesis space comprises linear predictors (Kakade et al., 2008; Wang et al., 2010; Crammer & Gentile, 2013; Abbasi-Yadkori et al., 2011; Gollapudi et al., 2021; van der Hoeven et al., 2021). These works focus on the efficacy of linear models in capturing the relationship between context and action rewards. However, the linear representation has its limitations in capturing complex relationships.

In response to these limitations, recent studies have delved into neural contextual bandits (Zhou et al., 2020; Jin et al., 2021; Zhang et al., 2021; Xu et al., 2022). These approaches leverage the expressive power of deep neural networks to model the context-action relationship more effectively. There are various policies proposed, including Thompson sampling and Upper Confidence Bound algorithms, to navigate the bandit problem in more complex and non-linear environments.

Despite these advancements in reinforcement learning, the existing literature primarily focuses on point prediction and lacks mechanisms for set-valued prediction and coverage control. This gap is particularly concerning in critical domains, as discussed in Section 1. The issue is partially addressed by recent works (Taufiq et al., 2022; Zhang et al., 2023; Stanton et al., 2023), which apply Conformal Prediction to off-policy evaluation problems, thereby returning prediction sets. However, these researches diverge from our work, which specifically addresses the bandit problem setting. Our focus lies in integrating set-valued predictions with the bandit feedback framework, an area that has not been extensively explored, presenting both novel challenges and opportunities for advancing the field.

2.3. Set-valued Classification with Bandit Feedback

The proposed BCCP method (summarized in Algorithm 1) aims to make set-valued decisions with a coverage guarantee for instances from the same distribution as the training data in the bandit feedback setting. Particularly, given a query $\boldsymbol{X}_t$, the learner pulls an arm A_t and receives the feedback $\mathbb{1}\{A_t = Y_t\}$. With this feedback, the learner updates the model and thresholds in conformal prediction (lines 4–5 in Algorithm 1). During the test phase, the learner returns the prediction set based on the trained model and thresholds (line 3 in Algorithm 1).

Take healthcare as an example. Due to cost and safety concerns, insurance companies may only allow the healthcare provider to prescribe one diagnostic test (e.g., X-ray, followed by CT, followed by cancer biomarker blood test, etc.) at a time to the patient (this may be viewed as pulling a single arm). When a diagnostic test turns out negative for a suspect cause, it is still unknown what the cause really is (this is consistent with our setting in which the learner only receives a bandit feedback that confirms the correctness of the pulled arm but does not necessarily reveal the true label). After a series of training over a large number of patients has been conducted, we have a diagnostic system that can make predictions for a new patient based on the patient's profile. Unless for clear-cut cases, often it is much safer for the provider to consider a set of most plausible causes and design the treatment plan that considers all plausible diseases, as opposed to treating the patient based on one single predicted disease.

3. Towards the Bandit Conformal

In this section, we introduce our method: Bandit Class-specific Conformal Prediction, specifically designed for set-valued multi-class classification problems in an online bandit feedback setting: let $\{(\boldsymbol{X}_t, Y_t)\}_{t=1}^{T}$ be a sequence of i.i.d. points from the domain $\mathcal{X} \times \mathcal{Y}$, where a leaner cannot observe the label Y_t and receives the non-zero feedback only when an arm is correctly pulled. We aim to report a prediction set $\widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_t)$ (the learner only uses the information up to time $t-1$) with a class-specific coverage guarantee.

Our methodology entails three pivotal steps: (1) estimating a ground truth based on a policy and feedback, (2) training the model with this estimation, and (3) estimating the $100 \times \alpha\%$ quantile τ_k for each class $k \in \mathcal{Y}$.

3.1. Estimating $\mathbb{1}\{Y_t = k\}$

In the bandit feedback context, for each query instance $\boldsymbol{X}_t$, the learner pulls an arm $A_t \in \mathcal{Y}$ based on a given policy $\pi_t := \pi_t(\cdot \mid \boldsymbol{X}_t)$, effectively making an educated guess about the potential true label. The environment then provides binary feedback indicating the correctness of the chosen arm, i.e., $\mathbb{1}\{A_t = Y_t\}$. As a direct observation of Y_t is not available, we rely on the following estimation to $\mathbb{1}\{Y_t = k\}$, i.e.,

$$\Delta_{t,k} := \frac{\mathbb{1}\{A_t = k\}}{\pi_t(k \mid \boldsymbol{X}_t)} \mathbb{1}\{A_t = Y_t\}.$$

Proposition 3.1. *$\Delta_{t,k}$ serves as an unbiased estimator of $\mathbb{1}\{Y_t = k\}$. This is substantiated by the equation*

$$\begin{aligned}\mathbb{E}_{\pi_t}\left[\Delta_{t,k}\right] &= \mathbb{E}_{\pi_t}\left[\Delta_{t,k} \mid A_t = k\right] \cdot \pi_t(k \mid \boldsymbol{X}_t) \\ &\quad + \mathbb{E}_{\pi_t}\left[\Delta_{t,k} \mid A_t \neq k\right] \cdot \left[1 - \pi_t(k \mid \boldsymbol{X}_t)\right] \\ &= \frac{\mathbb{1}\{k = Y_t\}}{\pi_t(k \mid \boldsymbol{X}_t)} \cdot \pi_t(k \mid \boldsymbol{X}_t) + 0 = \mathbb{1}\{Y_t = k\},\end{aligned}$$

where the expectation is taken with respect to policy π_t, conditioning on all previous information and the point $(\boldsymbol{X}_t, Y_t)$.

This estimation framework lays the groundwork for subsequent tasks in our study. It allows us to effectively utilize the policy's capability to learn the real data-generating process without explicit knowledge about the true label Y_t.

Policy design can be a flexible process, influenced by specific preferences such as the pursuit of simplicity or the goal of minimizing estimation variance. In our research, we theoretically analyze the performance of certain policies characterized by the associated properties, as detailed in Corollaries 3.3 and 3.5. Additionally, we conduct empirical evaluations and compare the performances of two distinct policies: the softmax policy (softmax probability output from a neural network as defined in (4)) and the uniform policy (uniform distribution). See Section 4.

3.2. The Cross-entropy Loss with Bandit Feedback

Throughout this article, we train a neural network model $\boldsymbol{f}_{\mathcal{W}}(\boldsymbol{X}) = (f_{\mathcal{W}}^{1}(\boldsymbol{X}), \cdots, f_{\mathcal{W}}^{|\mathcal{Y}|}(\boldsymbol{X}))^{\top} \in \mathbb{R}^{|\mathcal{Y}|}$, which is parameterized by a set of matrices collectively represented by $\mathcal{W}$. Our primary objective in the training phase, particularly within the bandit feedback context, is to minimize a modified version of the cross-entropy loss for each input query $\boldsymbol{X}_t$, formulated as follows:

$$\mathcal{L}(\boldsymbol{X}_t; \mathcal{W}) = -\sum_{k \in \mathcal{Y}} \Delta_{t,k} \cdot \log\left(\hat{p}(k \mid \boldsymbol{X}_t)\right). \tag{3}$$

By substituting $\Delta_{t,k}$ for the ground-truth label indicator $\mathbb{1}\{Y_t = k\}$, the loss function becomes an unbiased estimator of the traditional cross-entropy loss with full feedback $-\log\left(\hat{p}(Y_t \mid \boldsymbol{X}_t)\right)$ by following a similar derivation in Proposition 3.1. This allows using information in those instances where the true label Y_t is not explicitly available.

The estimated probability mass function $\hat{p}(k \mid \boldsymbol{X}_t)$ for each class k is derived from the outputs of the neural network. Specifically, it is modeled by applying the softmax function to the logits $f_{\mathcal{W}}^{k}(\boldsymbol{X}_t)$ produced by the neural network:

$$\hat{p}(k \mid \boldsymbol{X}_t) := \frac{\exp(f_{\mathcal{W}}^{k}(\boldsymbol{X}_t))}{\sum_{\tilde{k} \in \mathcal{Y}} \exp(f_{\mathcal{W}}^{\tilde{k}}(\boldsymbol{X}_t))}, \; k \in \mathcal{Y}. \tag{4}$$

By integrating the estimator $\Delta_{t,k}$ with the softmax output, our model can update efficiently by optimizing the tailored loss function (3) with stochastic gradient descent. It is important to note that one may employ other loss functions, such as the hinge loss in SVMs (Kakade et al., 2008).

Figure 1 presents a clear visualization of the cross-entropy loss across three real datasets in the bandit feedback setting. It shows the model fitting performance with the softmax and uniform policies. The plots illustrate that during the model training phase, the softmax policy consistently achieves a more rapid reduction in loss compared to the uniform policy. This superior performance can be attributed to the context-aware nature of the softmax policy, which strategically pulls arms based on the specific context of each query. This approach not only leads to a higher frequency of accurate predictions but also ensures better utilization of data points,

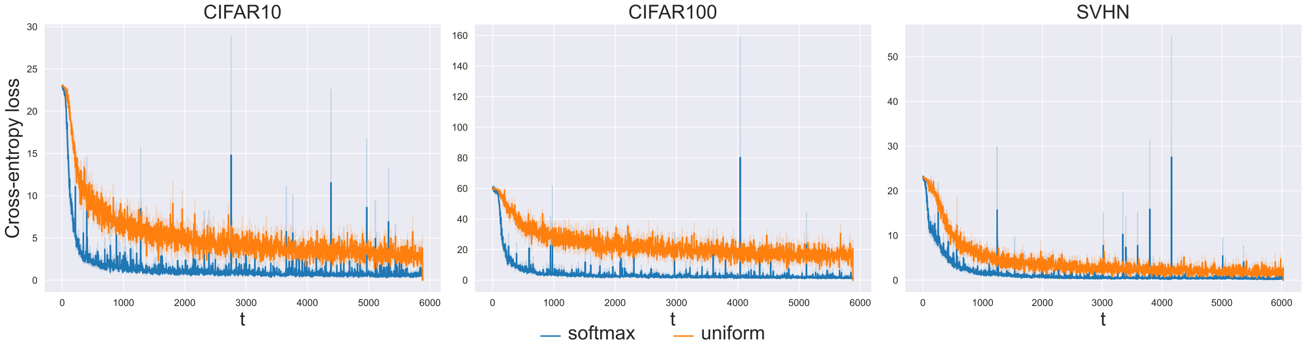

Figure 1. Accumulative cross-entropy loss under softmax policy and uniform policy.

thereby enhancing the overall efficiency and effectiveness of the model training process.

3.3. The Quantile of the Conformity Score

To control the class-specific coverage, our approach leverages thresholds/quantiles associated with a given conformity score function $s(\boldsymbol{X}, k)$, such as softmax, APS (Romano et al., 2020), or RAPS (Angelopoulos et al., 2021) score (see the definitions in Appendix A). Particularly, the primary goal in this phase is to determine a $100 \times \alpha\%$ quantile τ_k of the distribution of $s(\boldsymbol{X}, k)$. To this end, the traditional split conformal method (Papadopoulos et al., 2002; Lei et al., 2013) involves partitioning available labeled data into training and calibration sets. However, in the online setting, since we only have access to a limited dataset at each iteration, split conformal may lead to two primary issues: (1) reduced data for model training, and (2) large prediction sets due to limited labeled calibration data (Ding et al., 2024). These two issues are further aggravated in the bandit feedback setting because only those data whose correct arms are pulled are considered labeled.

To overcome these challenges, we adaptively update a quantile estimate τ_k by utilizing the check loss function (Takeuchi et al., 2006; Koenker & Bassett Jr, 1978; Romano et al., 2019; Gibbs & Candes, 2021) for quantile estimation:

$$\rho_\alpha(s, \tau) = (s - \tau) \cdot \big(\alpha - \mathbb{1}\{s < \tau\}\big).$$

More concretely, a class-specific $100 \times \alpha\%$ quantile $\tau_k, k \in \mathcal{Y}$ is obtained by solving the below optimization problem:

$$\begin{aligned} &\operatorname*{argmin}_{\tau} \mathbb{E}\big[\rho_\alpha(s(\boldsymbol{X}, k), \tau) \mid Y = k\big] \\ &= \operatorname*{argmin}_{\tau} \frac{\mathbb{E}\big[\mathbb{1}\{Y = k\} \cdot \rho_\alpha(s(\boldsymbol{X}, k), \tau)\big]}{\mathbb{E}\big[\mathbb{1}\{Y = k\}\big]} \\ &= \operatorname*{argmin}_{\tau} \mathbb{E}\big[\mathbb{1}\{Y = k\} \cdot \rho_\alpha(s(\boldsymbol{X}, k), \tau)\big], \end{aligned} \tag{5}$$

where the second equality holds due to the fact that $\mathbb{E}\big[\mathbb{1}\{Y = k\}\big] = \mathbb{P}(Y = k)$ does not rely on the quantile estimation. Given that the true joint density function $p(\boldsymbol{x}, y)$ is unknown, we instead employ a data-driven approach for quantile estimation: for each data point consider the loss

$$\Delta_{t,k} \cdot \rho_\alpha(s(\boldsymbol{X}_t, k), \tau), \tag{6}$$

which is an empirical counterpart of the population loss (5). Consequently, τ_k can be dynamically updated through stochastic gradient descent by computing the gradient, $-\Delta_{t,k} \cdot \big(\alpha - \mathbb{1}\{s(\boldsymbol{X}_t, k) < \tau\}\big)$, of the weighted loss (6). The updated quantiles $\tau_k, k \in \mathcal{Y}$ are then applied as the thresholds for the upcoming data in the next iteration only. The complete process, including the model training and quantile estimation in an online learning context, is outlined in Algorithm 1 and Figure 2.

Algorithm 1 Bandit Conformal

```
Require: Initialize weight matrices $\mathcal{W}^0$ and class-specific quantiles $\tau_k^0 = 0, k \in \mathcal{Y}$. Provide a score function $s^t(\cdot,\cdot)$[1], a policy $\pi_t$ and learning rates $\eta_1, \eta_2$.
for $t = 1, 2, 3, \cdots, T$ do
    Learner receives a query $\boldsymbol{X}_t$
    Generates a prediction set for the query:
        $\widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_t) := \big\{k \in \mathcal{Y} : s^{t-1}(\boldsymbol{X}_t, k) \geq \tau_k^{t-1}\big\}$
    Learner pulls an arm $A_t \sim \pi_t$, receives the feedback $\mathbb{1}\{A_t = Y_t\}$, and computes $\Delta_{t,k}$
    Update the network weight matrices and quantiles:
        $\begin{cases} \mathcal{W}^t = \mathcal{W}^{t-1} - \eta_1 \nabla_{\mathcal{W}} \mathcal{L}(\boldsymbol{X}_t; \mathcal{W}^{t-1}) \\ \tau_k^t = \tau_k^{t-1} + \eta_2 \Delta_{t,k}\big(\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t, k) < \tau_k^{t-1}\}\big) \end{cases}$
end for
```

When comparing with Gibbs & Candes (2021); Gibbs & Candès (2022); Zaffran et al. (2022); Bhatnagar et al. (2023),

[1] We add the superscript t on the score function to explicitly impress that it depends on the neural network updated up to t-th iteration. The same argument is applied to other notations.

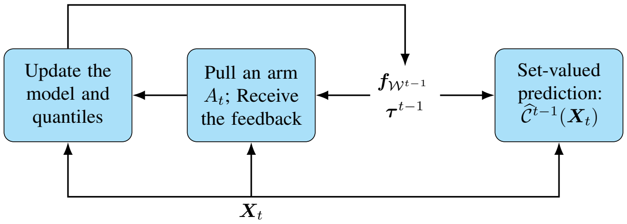
Update the model and quantiles
Pull an arm A_t; Receive the feedback
$\boldsymbol{f}_{\mathcal{W}^{t-1}}$
$\boldsymbol{\tau}^{t-1}$
Set-valued prediction: $\widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_t)$
$\boldsymbol{X}_t$

Figure 2. Flowchart of the online learning with bandit feedback. Here $\boldsymbol{\tau}^{t-1} = (\tau_1^{t-1}, \cdots, \tau_{|\mathcal{Y}|}^{t-1})^\top$.

a critical aspect differentiating our method lies in the quantile updating process in addition to the model training in the bandit feedback context as elucidated in Section 3.2. In particular, the aforementioned studies predominantly work with the unweighted quantile estimation and require verification of whether the true label Y_t falls in its prediction set $\widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_t)$ in their updating rules. This verification is typically achieved either by directly utilizing explicit label information or through multiple arm pulls until the true label is ascertained with absolute certainty. Such methodologies are not feasible in our setting for two primary reasons: (1) we lack direct access to the true label information, and (2) our framework does not permit multiple arm pulls for a single decision instance. In contrast, our approach (see the updating rule in Algorithm 1) involves computing the gradient of the weighted check loss (6) in the bandit feedback setting, which is an unbiased estimator of the gradient of unweighted check loss in the full feedback. This process is tailored to bandit feedback environments where each query allows only a single arm pull.

The below theorem implies the empirical coverage converges to the prescribed coverage.

Theorem 3.2. *Define the filtration $\mathcal{F}_t := (\sigma(\boldsymbol{X}_t, Y_t) \times \sigma(\pi_t)) \cup \mathcal{F}_{t-1}$. Assume $\pi_t(k \mid \boldsymbol{X}_t) \geq c_k > 0$ for all $t \in [T]$ and $\mathbb{E}[\frac{\mathbb{1}\{Y_t=k\}}{\pi_t(k|\boldsymbol{X}_t)} \mid \mathcal{F}_{t-1}] = b_k^t$. With probability at least $1-\delta$ taken over all the randomness, for all class $k \in \mathcal{Y}$, Algorithm 1 yields the empirical coverage gap*

$$\begin{aligned} \mathit{CvgGap}_k &:= \left| \alpha - \frac{1}{T_k} \sum_{t=1}^{T} \mathbb{1}\{Y_t = k\} \cdot \mathbb{1}\{Y_t \notin \widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_t)\} \right| \\ &\leq \frac{\tau_k^T}{\eta_2 T_k} + \frac{\zeta_k(T, \delta/|\mathcal{Y}|)}{T_k}, \end{aligned}$$

where $\zeta_k(T,\delta) = \frac{2}{3c_k} \log \frac{2}{\delta} + \sqrt{2 \log \frac{2}{\delta} \cdot \sum_{t=1}^T b_k^t}$, and $T_k = \sum_{t=1}^T \mathbb{1}\{Y_t = k\}$.

Theorem 3.2 implies the convergence rate of the class-specific coverage guarantee mainly depends on the learning rate η_2 and the sample size T_k of class k. Besides the policy should be bounded strictly below by 0, the additional assumption on $\mathbb{E}[\frac{\mathbb{1}\{Y_t=k\}}{\pi_t(k|\boldsymbol{X}_t)} \mid \mathcal{F}_{t-1}]$ further suggests that the policy should not overly underestimate the proportion of a class; otherwise the empirical coverage gap may increase.

To some extent, Theorem 3.2 ensures that the algorithm yields prediction sets with small sizes. This is because an algorithm with a large prediction set size often comes with inflated coverage, yet the theorem states that the empirical non-coverage must not deviate much away from the desired non-coverage of α. In particular, Theorem 3.2 precludes the trivial case $\widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_t) = \mathcal{Y}$ for all $t \in [T]$.

The below corollary highlights the impact of different policies on the convergence rate.

Corollary 3.3. *Assume the learning rate has the order $\eta_2 = \mathcal{O}(T^{-1/2})$. (1) If the policy π_t aligns with the Bayes posterior probability, i.e., $\pi_t(k \mid \boldsymbol{X}_t) = \mathbb{P}(Y_t = k \mid \boldsymbol{X}_t)$, then we have $\mathbb{E}[\frac{\mathbb{1}\{Y_t=k\}}{\pi_t(k|\boldsymbol{X}_t)} \mid \mathcal{F}_{t-1}] = b_k^t \leq 1$, and hence $\mathit{CvgGap}_k = \mathcal{O}(\frac{\sqrt{T}}{T_k})$. (2) If the policy is the uniform distribution, i.e., $\pi_t(k \mid \boldsymbol{X}_t) = \frac{1}{|\mathcal{Y}|}$, then $b_k^t \leq |\mathcal{Y}| p_k$ (here p_k denotes the prior probability of class k), and hence $\mathit{CvgGap}_k = \mathcal{O}(\frac{\sqrt{T|\mathcal{Y}|p_k}}{T_k})$.*

Corollary 3.3 implies a convergence rate of $\mathrm{CvgGap}_k = \mathcal{O}(T^{-1/2})$ when the learning rate $\eta_2 = \mathcal{O}(T^{-1/2})$ and sample size $T_k = \mathcal{O}(T)$, $k \in \mathcal{Y}$ under both Bayes posterior probability and uniform probability policies. In our experiments, due to the lack of access to the precise data distribution, we instead use the softmax policy, i.e., $\pi_t(k \mid \boldsymbol{X}_t) = \hat{p}(k \mid \boldsymbol{X}_t)$ as defined in (4), to estimate the Bayes posterior probability. As noted by Tibshirani et al. (2019), there are alternative methods for probability estimation, such as moment matching and Kullback-Leibler Divergence minimization. We refer to related work (Sugiyama et al., 2012) for a comprehensive review.

Theorem 3.4. *Let p_k be the prior probability of class $k \in \mathcal{Y}$, and $\tau_k^* = \operatorname{argmin}_\tau \frac{1}{T} \sum_{t=1}^T \mathbb{1}\{Y_t = k\} \rho_\alpha(s^{t-1}(\boldsymbol{X}_t), \tau)$ be the quantile estimate using all the data instances. Define the empirical regret associated with the check loss in the bandit feedback setting as $\mathit{Reg}_{k,\rho_\alpha}(T) := \frac{1}{T} \sum_{t=1}^T \Delta_{t,k} \rho_\alpha(s^{t-1}(\boldsymbol{X}_t), \tau_k^{t-1}) - \frac{1}{T} \sum_{t=1}^T \mathbb{1}\{Y_t = k\} \rho_\alpha(s^{t-1}(\boldsymbol{X}_t), \tau_k^*)$. By choosing $\eta_2 = \tau_k^* p_k^{1/2} \big(\sum_{t=1}^T \mathbb{E}[\frac{\mathbb{1}\{Y_t=k\}}{\pi_t^2(k|\boldsymbol{X}_t)}]\big)^{-1/2}$, Algorithm 1 yields an expected regret*

$$\mathbb{E}[\mathit{Reg}_{k,\rho_\alpha}(T)] \leq \frac{\tau_k^*}{T} \sqrt{p_k \sum_{t=1}^{T} \mathbb{E}\left[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t^2(k \mid \boldsymbol{X}_t)}\right]}.$$

The above expectation is taken over over all the randomness, including the data and algorithm. Note that τ_k^* is bounded, and hence the upper bound converges to 0.

Corollary 3.5. *For the uniform policy and an appropriately* *chosen η_2 as specified in Theorem 3.4, the expected regret* $\mathbb{E}[\mathit{Reg}_{k,\rho_\alpha}(T)] \le \frac{\tau^*|\mathcal{Y}|p_k}{\sqrt{T}}$.

Corollary 3.5 indicates that the expected regret adheres to a theoretical convergence rate of $\mathcal{O}(T^{-1/2})$, under the condition that the learning rate $\eta_2 = \mathcal{O}(T^{-1/2})$ (it can be achieved when the policy is bounded strictly below by 0). This condition aligns with the findings in Corollary 3.3.

Both Theorem 3.4 and Corollary 3.5 provide theoretical guarantees for the convergence behavior of Algorithm 1 in a parametric rate, indicating its potential effectiveness. This result shows that there exists such a learning rate η_2 leading to an optimal convergence rate. How to practically obtain such a precise learning rate is a challenging problem. In practice, as discussed in the work of Gibbs & Candes (2021), the chosen value of η_2 leads to two distinct scenarios. A larger value of η_2 may lead to unstable quantile estimations, causing oscillations in prediction set sizes. Over time, this could result in increasingly larger prediction sets in the online learning process. Conversely, a smaller value of η_2 slows the convergence rate of the coverage, necessitating more iterations to achieve desired coverage levels.

Algorithm 2 Bandit Conformal with Experts

```
Require: Initialize weight matrices $\mathcal{W}^0$, class-specific quantiles $\tau^0_{j,k} = 0$, and experts weights $\omega^0_{j,k} = 1,\ j \in [J], k \in \mathcal{Y}$. A score function $s^t(\cdot,\cdot)$, a policy $\pi_t$ and learning rates $\eta_1, \eta_{2,j}, j \in [J]$.
for $t = 1, 2, 3, \cdots, T$ do
    Learner receives a query $\boldsymbol{X}_t$
    Generates a prediction set for the query:
        $\widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_t) := \left\{k \in \mathcal{Y} : s^{t-1}(\boldsymbol{X}_t, k) \ge \bar{\tau}^{t-1}_k\right\},$
    where $\bar{\tau}^{t-1}_k = \sum_j \omega^{t-1}_{j,k}\tau^{t-1}_{j,k} / \sum_i \omega^{t-1}_{i,k}$
    Learner pulls an arm $A_t \sim \pi_t$, receives the feedback $\mathbb{1}\{A_t = Y_t\}$, and computes $\Delta_{t,k}$
    Update all weights and quantiles:
        $\begin{cases} \mathcal{W}^t = \mathcal{W}^{t-1} - \eta_1 \nabla_{\mathcal{W}} \mathcal{L}(\boldsymbol{X}_t; \mathcal{W}^{t-1}) \\ \tau^t_{j,k} = \tau^{t-1}_{j,k} + \eta_{2,j}\Delta_{t,k}\big(\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t, k) < \tau^{t-1}_{j,k}\}\big) \\ \omega^t_{j,k} = \exp(-\frac{1}{\sqrt{t+1}} \sum_{t' \le t} \Delta_{t',k} \cdot \rho_\alpha(s^{t'-1}(\boldsymbol{X}_{t'}, k), \tau^{t'-1}_{j,k})) \end{cases}$
end for
```

To mitigate the above limitation due to the choice of η_2, we draw inspiration from the adaptive control method in its full feedback setting (Zaffran et al., 2022). We introduce an alternative algorithm, Bandit Conformal with Experts (outlined in Algorithm 2), which eliminates the need for manual tuning of η_2. Specifically, given a grid of learning rate values $\eta_{2,j}, j \in [J]$, it employs an ensemble methodology to aggregate estimated quantiles associated with $\eta_{2,j}$'s based on past performance. The guiding principle is that as the accumulated check loss decreases, the attention placed on the corresponding estimated quantile grows.

Theorem 3.6 below shows that the aggregated quantile through the experts converges to the optimal quantile estimate among the experts. Specifically, an increase in the number of experts, while maintaining the order $J = \mathcal{O}(1)$, can enhance the chance of achieving an improved learning rate, along with more accurate quantile estimations. This finding underscores the importance of expert integration in improving algorithmic performance if one has no prior idea of the optimal learning rate.

Theorem 3.6. *Consider $\bar{\tau}^{t-1}_k$ as the aggregated quantile across $J(\ge 2)$ experts as defined in Algorithm 2, and the same c_k defined in Theorem 3.2. Then, Algorithm 2 yields*

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} \Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t), \bar{\tau}^{t-1}_k) \\ &- \min_{j \in [J]} \frac{1}{T}\sum_{t=1}^{T} \Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t), \tau^{t-1}_{j,k}) \\ &\le \frac{1}{4c_k^2\sqrt{T}} + \frac{2\ln J}{\sqrt{T}}. \end{aligned}$$

Here the assumption for c_k is reasonably flexible as it can be achieved through the policy design.

Notice that, theoretically, the optimal choice of learning rate should vary depending on the class as indicated in Theorem 3.4. However, for the ease of practical implementation, the same value of η_2 (or $\eta_{2,j}$) is applied across all classes.

4. Experiments

Set-up: To assess the effectiveness of our proposed approach, we employ the ResNet50 architecture (He et al., 2016) for model fitting. Our experimental setup includes the CIFAR10, CIFAR100 (with 20 coarser labels), and SVHN datasets, each undergoing 5 replications. Consistently throughout the study, we maintain a non-coverage rate $\alpha = 0.05$. For computational efficiency, the model training is performed on data batches of size 256, utilizing the ADAM optimizer with a learning rate of $\eta_1 = 10^{-4}$ in the model training phase. The entire online learning process spans $T = 6000$ iterations around. We evaluate online classification performance using three score functions: softmax, APS, and RAPS (see their definition in Appendix A) for both the softmax policy and the uniform policy.

Metrics: To examine the performance during online prediction for $t \in [T]$, we report both the minimum and maximum accumulative coverage, defined as:

$$\text{Acum\_cvg\_min}(t) = \min_{k\in\mathcal{Y}} \text{Acum\_cvg}(t,k),$$
$$\text{Acum\_cvg\_max}(t) = \max_{k\in\mathcal{Y}} \text{Acum\_cvg}(t,k),$$

where Acum_cvg(t,k) is defined as

$$\frac{\sum_{s=1}^{t}\sum_{\boldsymbol{X}_i\in\mathcal{B}_s}\mathbb{1}\{Y_i = k \;\&\; Y_i \in \widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_i)\}}{\sum_{s=1}^{t}\sum_{\boldsymbol{X}_i\in\mathcal{B}_s}\mathbb{1}\{Y_i = k\}},$$

with $\mathcal{B}_s$ representing the batch of the dataset at time point s. We include the accumulative prediction set size,

$$\text{Acum\_size}(t) = \frac{\sum_{s=1}^{t}\sum_{\boldsymbol{X}_i\in\mathcal{B}_s}|\widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_i)|}{\sum_{s=1}^{t}|\mathcal{B}_s|},$$

to assess the informativeness of the set-valued classification.

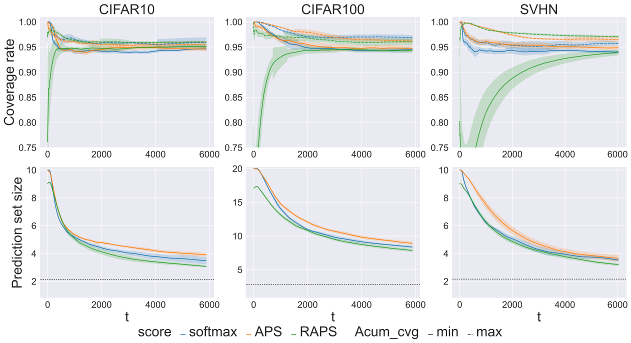

Figure 3. Performances under Algorithm 1 with softmax policy. The black dotted lines in the bottom panel denote the oracle performance of the model with access to the full labels.

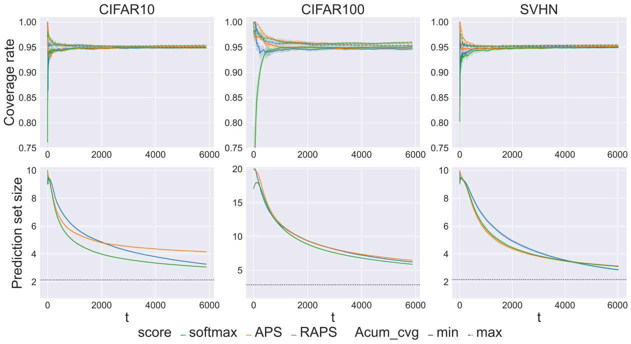

Figure 4. Performances under Algorithm 1 with uniform policy. The black dotted lines in the bottom panel denote the oracle performance of the model with access to the full labels.

Results: Figures 3 and 4 present the set-valued classification with BCCP in the bandit feedback setting under softmax and uniform policies, respectively. The black dotted lines in the bottom panel for each figure denote the final result of a network after sufficiently many iterations with access to the full labels and the usage of the RAPS score function. As the number of iterations increases, the top panels in Figures 3 and 4 reveal that Algorithm 1 effectively approaches the prescribed class-specific coverage of 95%. Additionally, the bottom panels in these figures indicate a trend towards smaller prediction sets.

The choice of learning rate η_1 indeed affects the performance of the model training phase and hence the subsequent quantile estimation. However, in our study, we mainly focus on the role of η_2 instead of particularly optimizing for η_1. For example, the CIFAR100 experiments utilizing the softmax policy and softmax score are presented with a fine-tuned $\eta_2 = 5 \times 10^{-4}$ (see the tuning strategy and sensitivity studies about η_2 in Appendix B). As discussed below Corollary 3.5, an inappropriate selection of the hyperparameter η_2 can result in enlarged prediction set sizes or prolonged convergence times, which hinders the practical applicability of Algorithm 1 in more dynamic settings.

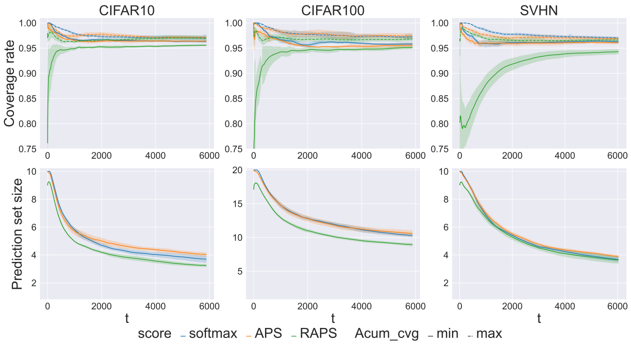

Figure 5. Performances under Algorithm 2 with softmax policy.

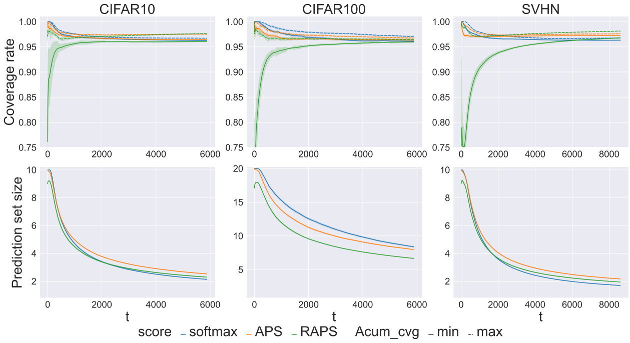

Figure 6. Performances under Algorithm 2 with uniform policy.

To address this limitation, in this study, we employed a range of learning rate values, i.e., [0.1, 0.01, 0.001, 0.0001], through an expert-based approach in Algorithm 2. The results are shown in Figures 5 and 6. Notably, while using the softmax policy, the results from Figure 5 indicate that the prediction set sizes from Algorithm 2 are only marginally larger compared to those from Algorithm 1 with carefully tuned η_2. With the uniform policy, Algorithm 2 demonstrates more efficient performance, yielding smaller prediction sets for the CIFAR10 and SVHN datasets. Notably, the RAPS score function outperforms the other scores in producing smaller prediction sets on the dataset when there are many classes, i.e., CIFAR100.

5. Conclusion

In this article, we extend Conformal Prediction to the framework of online bandit feedback, where a learner is only told whether or not a pulled arm is correct in a dynamic multi-class classification problem. We make use of an unbiased indicator function estimation of the ground truth to overcome the incomplete information in the feedback, allowing the proposed Bandit Class-specific Conformal Prediction (BCCP) to effectively make set-valued inferences and adaptively fit the model accordingly. Particularly, the indicator function estimation allows us to utilize stochastic gradient descent to efficiently achieve the quantile estimation instead of the traditional split conformal, which requires sufficient labeled calibration data and might not be realistic in the setting of bandit feedback. Theoretically, we show the $\mathcal{O}(T^{-1/2})$ convergence rate for both the coverage guarantee and the regret of the check loss under certain conditions. Empirically, the experiments conducted on three datasets with three score functions and two policies demonstrated the effectiveness of BCCP.

Our research opens several promising avenues for future exploration. One potential direction is the investigation of alternative indicator function estimations or policy designs that could offer improved theoretical or empirical performance. Additionally, refining the coverage guarantee within specific fixed-size time windows (Bhatnagar et al., 2023) instead of the full-time horizon in our work could further bolster the reliability of BCCP over different time scales. Moreover, expanding the scope of BCCP to address challenges such as covariate shift (Tibshirani et al., 2019) and semantic shift (Wang & Qiao, 2023) could significantly broaden its applicability.

In conclusion, our work not only contributes a novel and provable solution to the problem of online multi-class classification with bandit feedback but also sets another new direction in conformal prediction. It opens up possibilities for real-world applications and lays a foundation for further research domains.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Balasubramanian, V., Ho, S.-S., and Vovk, V. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Bartlett, P. L. and Wegkamp, M. H. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840, 2008.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 2337–2363. PMLR, 2023.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Charoenphakdee, N., Cui, Z., Zhang, Y., and Sugiyama, M. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pp. 1507–1517. PMLR, 2021.

Crammer, K. and Gentile, C. Multiclass classification with bandit feedback using adaptive regularization. *Machine learning*, 90(3):347–383, 2013.

Ding, T., Angelopoulos, A., Bates, S., Jordan, M., and Tibshirani, R. J. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Gibbs, I. and Candès, E. Conformal inference for online prediction with arbitrary distribution shifts. *arXiv preprint arXiv:2208.08401*, 2022.

Gollapudi, S., Guruganesh, G., Kollias, K., Manurangsi, P., Leme, R., and Schneider, J. Contextual recommendations and low-regret cutting-plane algorithms. *Advances in Neural Information Processing Systems*, 34:22498–22508, 2021.

Guan, L. and Tibshirani, R. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 84(2):524, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hechtlinger, Y., Póczos, B., and Wasserman, L. Cautious deep learning. *arXiv preprint arXiv:1805.09460*, 2018.

Herbei, R. and Wegkamp, M. H. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pp. 709–721, 2006.

Jin, T., Xu, P., Shi, J., Xiao, X., and Gu, Q. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pp. 440–447, 2008.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Lei, J. Classification with confidence. *Biometrika*, 101(4): 755–769, 2014.

Lei, J., Robins, J., and Wasserman, L. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Lei, J., Rinaldo, A., and Wasserman, L. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):29–43, 2015.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pp. 345–356. Springer, 2002.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

Stanton, S., Maddox, W., and Wilson, A. G. Bayesian optimization with conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pp. 959–986. PMLR, 2023.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(45):1231–1264, 2006. URL http://jmlr.org/papers/v7/takeuchi06a.html.

Taufiq, M. F., Ton, J.-F., Cornish, R., Teh, Y. W., and Doucet, A. Conformal off-policy prediction in contextual bandits. *Advances in Neural Information Processing Systems*, 35: 31512–31524, 2022.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

van der Hoeven, D., Fusco, F., and Cesa-Bianchi, N. Beyond bandit feedback in online multiclass classification. *Advances in neural information processing systems*, 34: 13280–13291, 2021.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Wang, S., Jin, R., and Valizadegan, H. A potential-based framework for online multi-class learning with partial feedback. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 900–907. JMLR Workshop and Conference Proceedings, 2010.

Wang, W. and Qiao, X. Learning confidence sets using support vector machines. In *Advances in Neural Information Processing Systems*, pp. 4929–4938, 2018.

Wang, W. and Qiao, X. Set-valued support vector machine with bounded error rates. *Journal of the American Statistical Association*, pp. 1–13, 2022.

Wang, Z. and Qiao, X. Set-valued classification with out-of-distribution detection for many classes. *Journal of Machine Learning Research*, 24(375):1–39, 2023.

Wang, Z. and Qiao, X. Deep generalized prediction set classifier and its theoretical guarantees. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=H7gLN5nqVF. Featured Certification.

Xu, P., Wen, Z., Zhao, H., and Gu, Q. Neural contextual bandits with deep representation and shallow exploration. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xnYACQquaGV.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR, 2022.

Zhang, C., Wang, W., and Qiao, X. On reject and refine options in multicategory classification. *Journal of the American Statistical Association*, 113(522):730–745, 2018.

Zhang, W., Zhou, D., Li, L., and Gu, Q. Neural thompson sampling. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tkAtoZkcUnm.

Zhang, Y., Shi, C., and Luo, S. Conformal off-policy prediction. In *International Conference on Artificial Intelligence and Statistics*, pp. 2751–2768. PMLR, 2023.

Zhou, D., Li, L., and Gu, Q. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

A. Conformity Scores

Let $\hat{p}(k \mid \boldsymbol{X}), k \in \mathcal{Y}$ as defined in (4) be the estimated posterior probability based on the neural network $\boldsymbol{f}_{\mathcal{W}}(\boldsymbol{X})$. Thus, for the test point $(\boldsymbol{X}, Y)$, the softmax score is defined as

$$s(\boldsymbol{X}, k) = \hat{p}(k \mid \boldsymbol{X}).$$

Sort the estimated posterior probabilities $\hat{p}(k \mid \boldsymbol{X}), k \in \mathcal{Y}$ with the ascending order such that $\hat{p}(k_1 \mid \boldsymbol{X}) \leq \hat{p}(k_2 \mid \boldsymbol{X}) \leq \cdots \leq \hat{p}(k_{|\mathcal{Y}|} \mid \boldsymbol{X})$. Additionally, denote r as the index such that $k_r = Y$. Thus, the APS score is defined as

$$s(\boldsymbol{X}, k) = 1 - \sum_{l=1}^{r-1} \hat{p}(k_l \mid \boldsymbol{X}) - U \cdot \hat{p}(k_r \mid \boldsymbol{X}),$$

where U is a random variable sampled from the uniform distribution on the interval $[0, 1]$.

Let k_{reg} be the number above which the prediction set size will be penalized with the penalty λ. Thus, the RAPS is defined as

$$s(\boldsymbol{X}, k) = 1 - \sum_{l=1}^{r-1} \hat{p}(k_l \mid \boldsymbol{X}) - U \cdot \hat{p}(k_r \mid \boldsymbol{X}) - \lambda \cdot [r - k_{reg}]_+,$$

where $[\cdot]_+ = \max(0, \cdot)$. By following the similar routine in Ding et al. (2024), in our experiments, we pick $\lambda = 0.01$ and $k_{reg} = 5$ for CIFAR100 while $k_{reg} = 1$ for the remaining two less difficult datasets.

B. Extra Studies on η_2

In our study, we adopt the learning rate tuning approach as described by Gibbs & Candes (2021), selecting a value that ensures a stable learning trajectory characterized by a balance between smaller prediction set sizes and satisfactory coverage convergence. However, this tuning strategy presents challenges in practical applications. Specifically, different datasets require distinct optimal learning rate values, and identifying these values through manual tuning is both time-consuming and less adaptive. To illustrate these challenges, we conducted sensitivity analyses on the impact of varying η_2 in Algorithm 1. These studies underscore the limitations of manually tuning a single η_2 value.

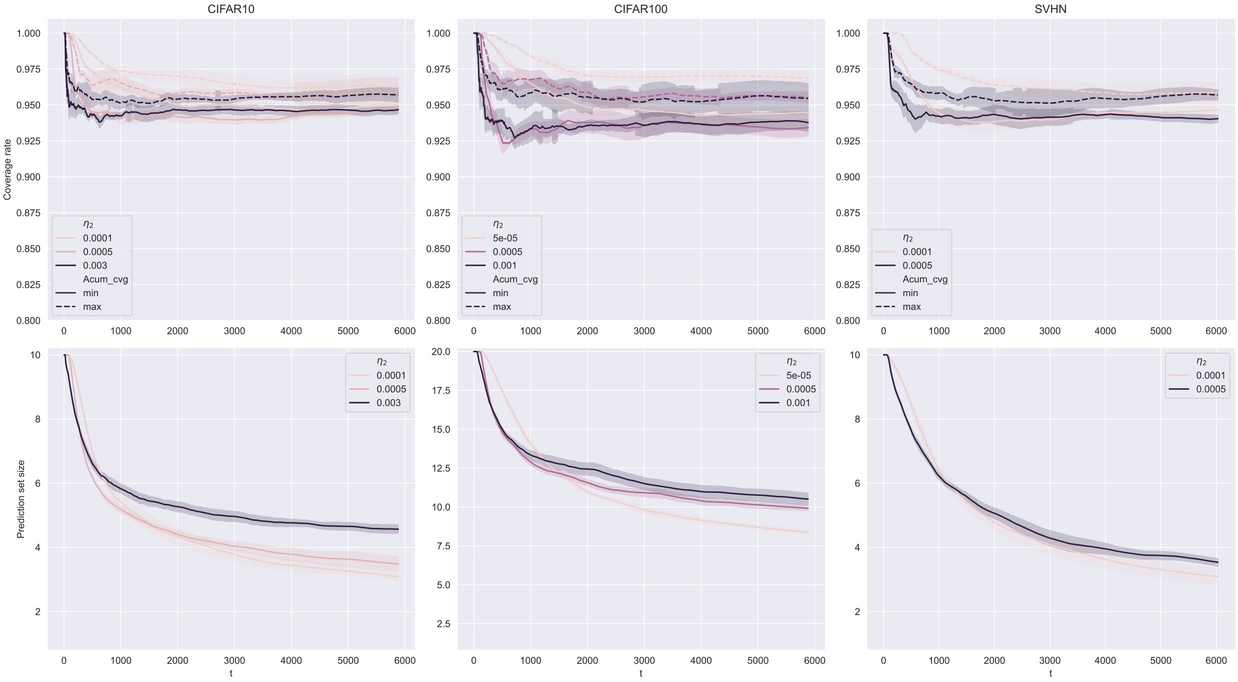

Figure 7. Performances under Algorithm 1 with softmax policy and softmax score.

Our findings, presented in Figures 7 to 10, explore the sensitivity of the learning rate η_2. We observed that a higher η_2 value accelerates coverage control, as indicated by the darker lines in the top panels of each figure. However, this generally comes at the cost of enlarged prediction set sizes, evident from the darker lines in the bottom panels of the figures. Moreover, the prediction set size shows considerable sensitivity to variations in η_2. This highlights the practical limitations of Algorithm 1 and underscores the necessity of implementing Algorithm 2, which utilizes an expert-based method to aggregate results across multiple learning rates $\eta_{2,j}, j \in [J]$. This approach not only addresses the challenges of manual tuning but also enhances the algorithm's adaptability and effectiveness across diverse datasets.

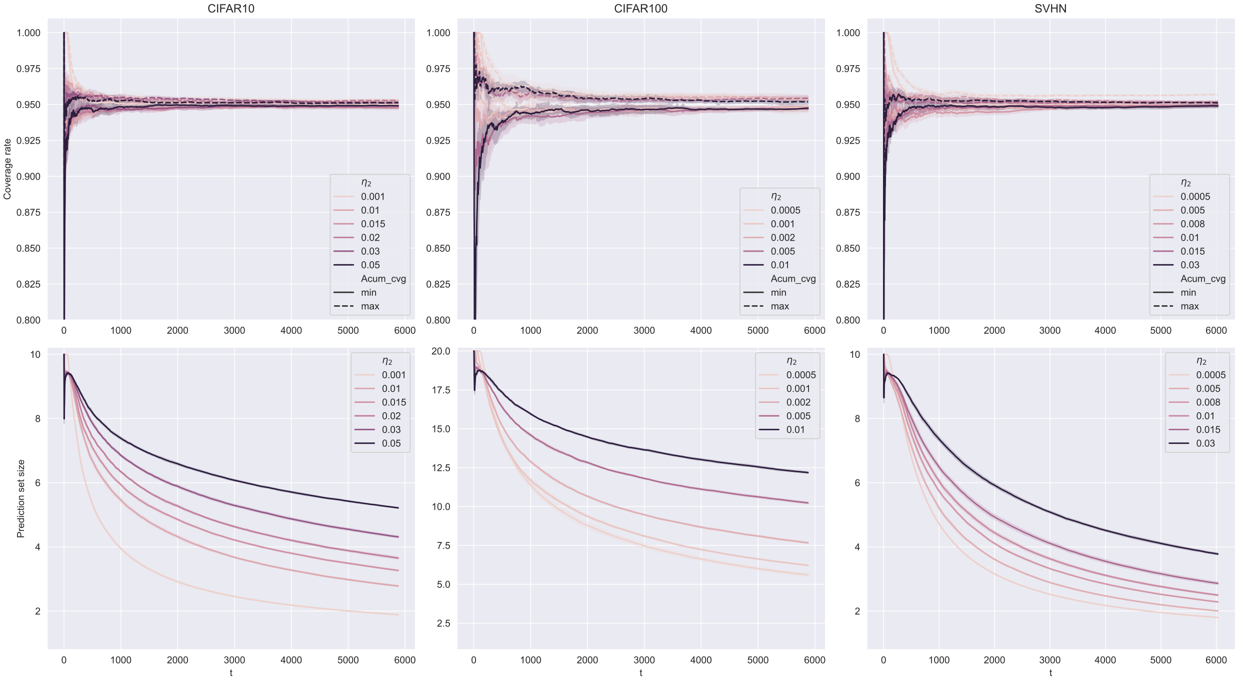

Figure 8. Performances under Algorithm 1 with uniform policy and softmax score.

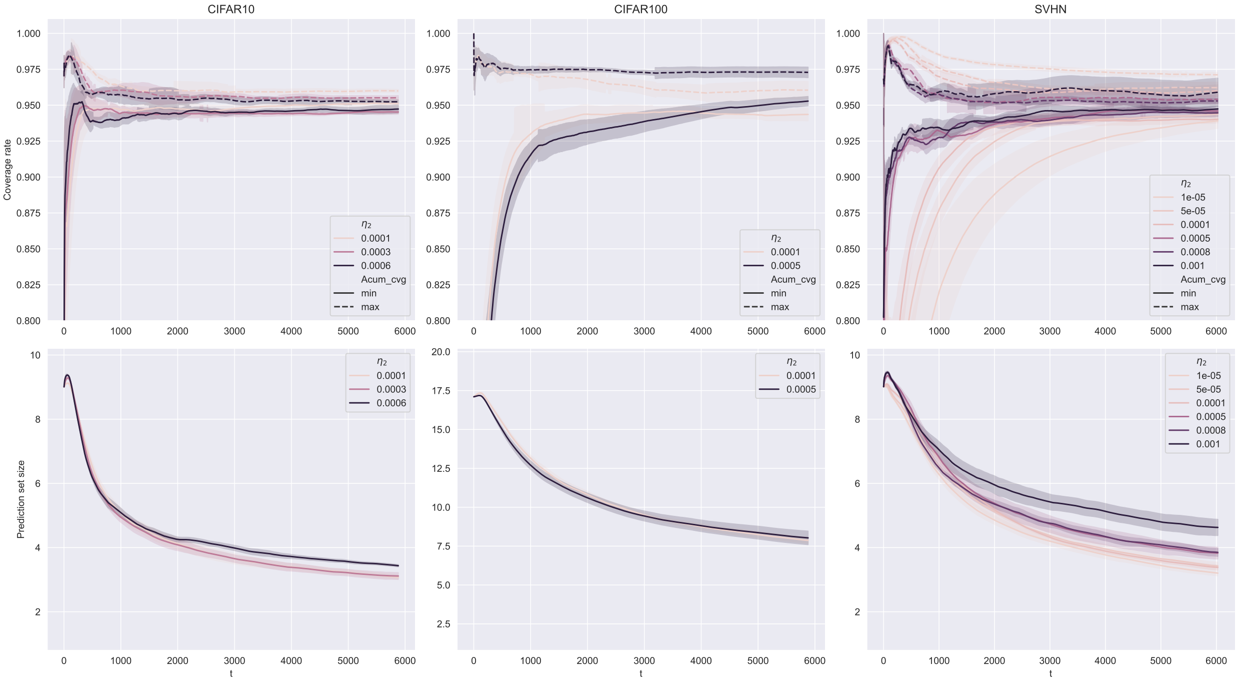

Figure 9. Performances under Algorithm 1 with softmax policy and RAPS score.

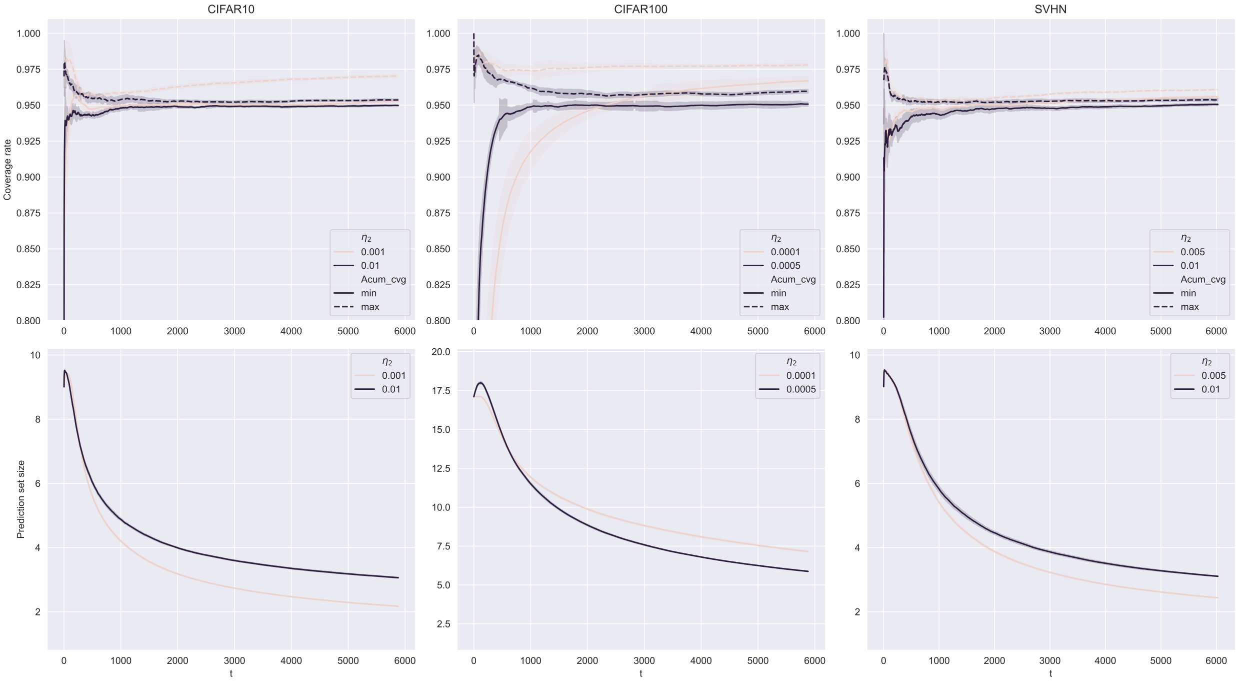

Figure 10. Performances under Algorithm 1 with uniform policy and RAPS score.

C. Discussion of Policy π_t

In this section, for each class, we show the effectiveness of different policies on the correctness of arm pulling, i.e., $\mathbb{P}(A_t = Y_t \mid Y_t = k),\ k \in \mathcal{Y}$. In Figure 11, under the softmax policy (top panel) and the uniform policy (bottom panel), we

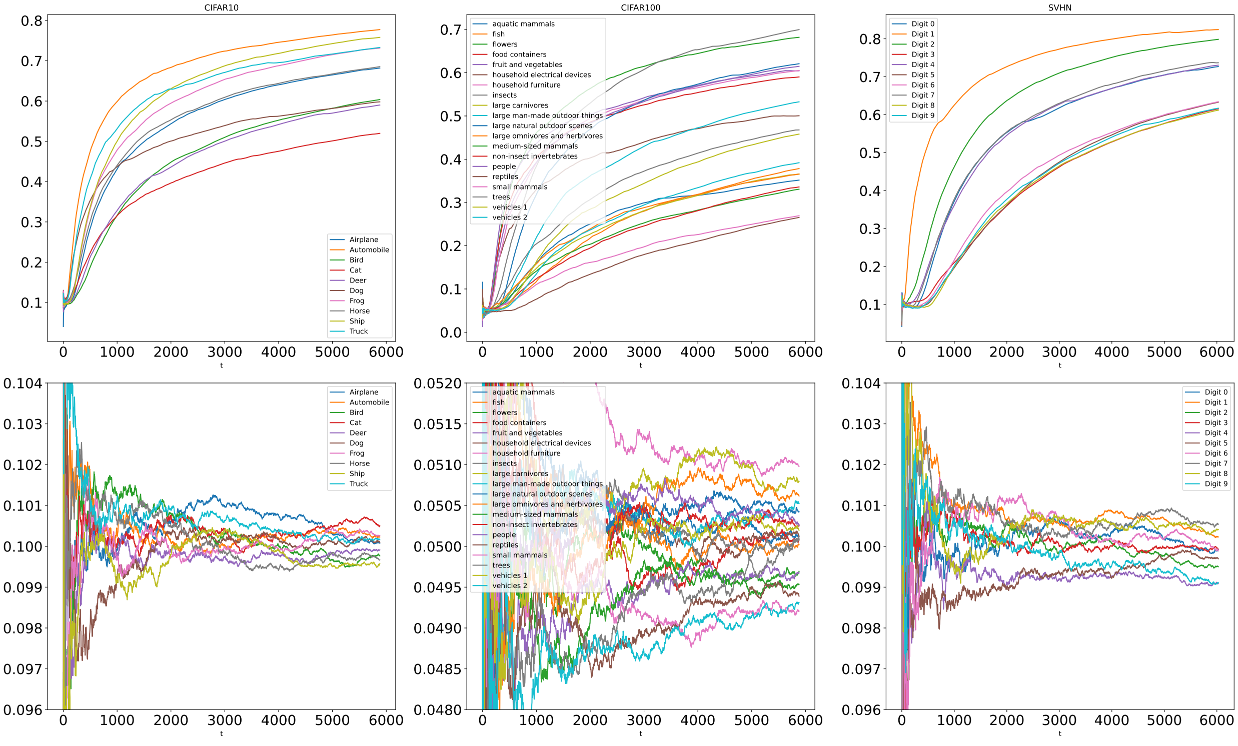

Figure 11. Proportion of correctly pulled arm with RAPS score under softmax (top) and uniform (bottom) policy.

report the accumulative performance of arm pulling for each class, i.e.,

$$\frac{\sum_{s=1}^{t} \sum_{\boldsymbol{X}_i \in \mathcal{B}_s} \mathbb{1}\{A_i = k\}}{\sum_{s=1}^{t} \sum_{\boldsymbol{X}_i \in \mathcal{B}_s} \mathbb{1}\{Y_i = k\}}, \ k \in \mathcal{Y}.$$

Due to the usage of context $\boldsymbol{X}_i$ in each batch $\mathcal{B}_s, s \leq t$, softmax policy leads to higher accuracy for arm pulling. In contrast, the uniform policy's correctness is close to $\frac{1}{|\mathcal{Y}|}$. These behaviors align with the one in cross-entropy loss minimization in Figure 1, where the softmax policy quickly decreases the loss compared to the uniform policy. On the other hand, when it comes to the performance of set-valued classification in Figures 3 and 5, the uniform policy both converges faster to the desired coverage rate and gets slightly smaller prediction sets on average than the softmax policy.

The above interesting phenomenon may mirror the exploration-exploitation dilemma in reinforcement learning. Specifically, the softmax policy capitalizes on more known information characterized by $\hat{p}(k \mid \boldsymbol{X}_t)$ as defined in (4) and hence "guesses" labels with higher frequent success. Such a policy can greedily and quickly decrease the cross-entropy loss but sacrifices the performance of the set-valued prediction. In contrast, the uniform policy has a higher capability of exploration, possibly leading to the fast empirical convergence of coverage rate and smaller prediction sets, even though it has an inferior capability to reduce the cross-entropy loss in each iteration.

D. Proofs

Proof of Theorem 3.2. Define $M_{t,k} := [\Delta_{t,k} - \mathbb{1}\{Y_t = k\}] \cdot [\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t, k) < \tau_k^{t-1}\}]$ and

$$\begin{aligned}
\mathbb{V}[M_{t,k} \mid \mathcal{F}_{t-1}] &:= \mathbb{E}_{(\boldsymbol{X}_t, Y_t)}\left[\mathbb{E}\left[\frac{\mathbb{1}\{A_t = k\} \cdot \mathbb{1}\{A_t = Y_t\}}{\pi_t^2(k \mid \boldsymbol{X}_t)}[\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t, k) < \tau_k^{t-1}\}]^2 \mid \mathcal{F}_{t-1}, (\boldsymbol{X}_t, Y_t)\right]\right] \\
&= \mathbb{E}_{(\boldsymbol{X}_t, Y_t)}\left[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t(k \mid \boldsymbol{X}_t)}[\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t, k) < \tau_k^{t-1}\}]^2 \mid \mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}_{(\boldsymbol{X}_t, Y_t)}\left[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t(k \mid \boldsymbol{X}_t)} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t(k \mid \boldsymbol{X}_t)} \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

Additionally, we have

$$|M_{t,k}| \leq \frac{1}{c_k} \text{ and } \mathbb{E}[M_{t,k} \mid \mathcal{F}_{t-1}] = \mathbb{E}_{(\boldsymbol{X}_t, Y_t)}[\mathbb{E}[M_{t,k} \mid \mathcal{F}_{t-1}, (\boldsymbol{X}_t, Y_t)]] = 0. \tag{7}$$

Then, by utilizing the Chernoff bound, for any $\xi > 0$, we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{t=1}^{T} M_{t,k} \geq \varepsilon\right] &\leq \exp(-\xi\varepsilon) \cdot \mathbb{E}\left[\exp(\xi \sum_{t=1}^{T} M_{t,k})\right] \\
&= \exp(-\xi\varepsilon) \cdot \mathbb{E}\left[\mathbb{E}\left[\exp(\xi \sum_{t=1}^{T-1} M_{t,k} + \xi M_{T,k}) \mid \mathcal{F}_{T-1}\right]\right] \\
&= \exp(-\xi\varepsilon) \cdot \mathbb{E}\left[\exp(\xi \sum_{t=1}^{T-1} M_{t,k}) \cdot \mathbb{E}\left[\exp(\xi M_{T,k}) \mid \mathcal{F}_{T-1}\right]\right] \\
&\leq \exp(-\xi\varepsilon) \cdot \mathbb{E}\left[\exp(\xi \sum_{t=1}^{T-1} M_{t,k}) \cdot \exp\left(\mathbb{V}\left[M_{T,k} \mid \mathcal{F}_{T-1}\right] c_k^2(\exp(\xi/c_k) - c_k^2 - c_k\xi)\right)\right] \qquad (8) \\
&\leq \exp(-\xi\varepsilon) \cdot \exp\left(b_k^T \cdot c_k^2(\exp(\xi/c_k) - c_k^2 - c_k\xi)\right) \cdot \mathbb{E}\left[\exp(\xi \sum_{t=1}^{T-1} M_{t,k})\right] \\
&\leq \exp\left(c_k^2(\exp(\xi/c_k) - c_k^2 - c_k\xi) \sum_{t=1}^{T} b_k^t - \xi\varepsilon\right) \\
&= \exp\left(-c_k^2 \sum_{t=1}^{T} b_k^t \cdot \left[-\frac{\varepsilon/c_k}{\sum_{t=1}^{T} b_k^t} + (\frac{\varepsilon/c_k}{\sum_{t=1}^{T} b_k^t} + 1) \cdot \log(\frac{\varepsilon/c_k}{\sum_{t=1}^{T} b_k^t} + 1)\right]\right), \qquad (9)
\end{aligned}$$

where (8) holds due to

$$\begin{aligned}
\mathbb{E}\left[\exp(\xi M_{t,k}) \mid \mathcal{F}_{t-1}\right] &= 1 + \mathbb{E}\left[\xi M_{t,k} \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[\sum_{n=2}^{\infty} \frac{\xi^n M_{t,k}^n}{n!} \mid \mathcal{F}_{t-1}\right] \\
&\leq 1 + \mathbb{E}\left[M_{t,k}^2 \sum_{n=2}^{\infty} \frac{\xi^n |M_{t,k}|^{n-2}}{n!} \mid \mathcal{F}_{t-1}\right] \\
&\leq 1 + \mathbb{V}[M_{t,k} \mid \mathcal{F}_{t-1}] \sum_{n=2}^{\infty} \frac{\xi^n}{c_k^{n-2} n!} \\
&= 1 + \mathbb{V}[M_{t,k} \mid \mathcal{F}_{t-1}] c_k^2(\exp(\xi/c_k) - c_k^2 - c_k\xi) \\
&\leq \exp\left(\mathbb{V}[M_{t,k} \mid \mathcal{F}_{t-1}] c_k^2(\exp(\xi/c_k) - c_k^2 - c_k\xi)\right),
\end{aligned}$$

and (9) holds since we set $\xi = c_k \log(\frac{\varepsilon/c_k}{\sum_{t=1}^{T} b_k^t} + 1)$.

By applying the fact of $(1+u)\log(1+u)-u \geq \frac{u^2}{2+2u/3},\ u \geq 0$ on (9), we have $\mathbb{P}\Big[\sum_{t=1}^T M_{t,k} \geq \varepsilon\Big] \leq \exp(-\frac{\varepsilon^2}{2\sum_{t=1}^T b_k^t + 2\varepsilon/(3c_k)})$, and hence

$$\mathbb{P}\left[\left|\sum_{t=1}^T M_{t,k}\right| \geq \varepsilon\right] \leq 2\exp(-\frac{\varepsilon^2}{2\sum_{t=1}^T b_k^t + 2\varepsilon/(3c_k)}).$$

Thus, with the probability at least $1-\delta$, we have

$$\begin{aligned}
&\left|\sum_{t=1}^T \Delta_{t,k}\cdot[\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t,k) < \tau_k^{t-1}\}] - \mathbb{1}\{Y_t = k\}\cdot[\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t,k) < \tau_k^{t-1}\}]\right| \\
=& \left|\sum_{t=1}^T M_{t,k}\right| \\
\leq& \frac{1}{3c_k}\log\frac{2}{\delta} + \sqrt{(\frac{1}{3c_k}\log\frac{2}{\delta})^2 + 2\sum_{t=1}^T b_k^t \log\frac{2}{\delta}} \\
\leq& \frac{2}{3c_k}\log\frac{2}{\delta} + \sqrt{2\log\frac{2}{\delta}\sum_{t=1}^T b_k^t} := \zeta_k(T,\delta)
\end{aligned} \tag{10}$$

Deriving from the updating rule for the quantile estimation in Algorithm 1, we have

$$\begin{aligned}
&\tau_k^T = \tau_k^0 + \eta_2\sum_{t=1}^T \Delta_{t,k}\cdot\left[\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t,k) < \tau_k^{t-1}\}\right] \\
\Longrightarrow\ & \sum_{t=1}^T \Delta_{t,k}\cdot[\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t,k) < \tau_k^{t-1}\}] = \frac{\tau_k^T}{\eta_2}.
\end{aligned} \tag{11}$$

Therefore, combing (10) with (11), with probability at least $1-\delta$, we have

$$\begin{aligned}
&\frac{\tau_k^T}{\eta_2} - \zeta_k(T,\delta) \leq \sum_{t=1}^T \mathbb{1}\{Y_t = k\}\cdot[\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t,k) < \tau_k^{t-1}\}] \leq \frac{\tau_k^T}{\eta_2} + \zeta_k(T,\delta) \\
\Rightarrow\ & \frac{\tau_k^T}{\eta_2 T_k} - \frac{\zeta_k(T,\delta)}{T_k} \leq \alpha - \sum_{t=1}^T \frac{\mathbb{1}\{Y_t = k\}}{T_k}\cdot\mathbb{1}\{Y_t \notin \widehat{\mathcal{C}}^{t-1}(\boldsymbol{X}_t)\} \leq \frac{\tau_k^T}{\eta_2 T_k} + \frac{\zeta_k(T,\delta)}{T_k}
\end{aligned}$$

□

Proof of Theorem 3.4. Recall the definition $\tau_k^* = \min_\tau \frac{1}{T}\sum_{t=1}^T \mathbb{1}\{Y_t = k\}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t),\tau)$. Thus,

$$\begin{aligned}
T\cdot\mathrm{Reg}_{k,\rho_\alpha}(T) &= \sum_{t=1}^T \Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t),\tau_k^{t-1}) - \sum_{t=1}^T \mathbb{1}\{Y_t = k\}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t),\tau_k^*) \\
&= \underbrace{\sum_{t=1}^T \Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t),\tau_k^{t-1}) - \sum_{t=1}^T \Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t),\tau_k^*)}_{\mathrm{Diff}_1} \\
&\quad + \underbrace{\sum_{t=1}^T \Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t),\tau_k^*) - \sum_{t=1}^T \mathbb{1}\{Y_t = k\}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t),\tau_k^*)}_{\mathrm{Diff}_2},
\end{aligned}$$

where $\mathbb{E}[\text{Diff}_2] = 0$ since $\Delta_{t,k}$ is an unbiased estimator of $\mathbb{1}\{Y_t = k\}$ conditional on $\mathcal{F}_{t-1} \cup (\boldsymbol{X}_t, Y_t)$. Additionally, we have

$$\begin{aligned}
\text{Diff}_1 &\leq \sum_{t=1}^{T} \Delta_{t,k} \cdot g_{t-1,k} \cdot (\tau_k^{t-1} - \tau_k^*), \qquad \text{here (sub)gradient } g_{t-1,k} := -\Delta_{t,k}[\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t) < \tau_k^{t-1}\}] \\
&= \sum_{t=1}^{T} \frac{\Delta_{t,k}}{\eta_2} \cdot (\tau_k^{t-1} - \tau_k^t) \cdot (\tau_k^{t-1} - \tau_k^*) \\
&= \sum_{t=1}^{T} \frac{\Delta_{t,k}}{2\eta_2} \cdot [(\tau_k^{t-1} - \tau_k^t)^2 + (\tau_k^{t-1} - \tau_k^*)^2 - (\tau_k^t - \tau_k^*)] \\
&= \sum_{t=1}^{T} \frac{\Delta_{t,k}^3 \eta_2}{2} \cdot [\alpha - \mathbb{1}\{s^{t-1}(\boldsymbol{X}_t) < \tau_k^{t-1}\}]^2 + \sum_{t=1}^{T} \frac{\Delta_{t,k}}{2\eta_2} \cdot [(\tau_k^{t-1} - \tau_k^*)^2 - (\tau_k^t - \tau_k^*)^2],
\end{aligned}$$

which further implies

$$\begin{aligned}
\mathbb{E}[\text{Diff}_1] &\leq \sum_{t=1}^{T} \frac{\eta_2}{2} \mathbb{E}\Big[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t^2(k \mid \boldsymbol{X}_t)}\Big] + \sum_{t=1}^{T} \frac{p_k}{2\eta_2} \cdot \mathbb{E}[(\tau_k^{t-1} - \tau_k^*)^2 - (\tau_k^t - \tau_k^*)^2] \\
&= \frac{\eta_2}{2} \sum_{t=1}^{T} \mathbb{E}\Big[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t^2(k \mid \boldsymbol{X}_t)}\Big] + \frac{p_k}{2\eta_2} \mathbb{E}[(\tau_k^0 - \tau_k^*)^2 - (\tau_k^T - \tau_k^*)^2] \\
&\leq \frac{\eta_2}{2} \sum_{t=1}^{T} \mathbb{E}\Big[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t^2(k \mid \boldsymbol{X}_t)}\Big] + \frac{p_k (\tau_k^*)^2}{2\eta_2} \\
&= \tau_k^* \sqrt{p_k \sum_{t=1}^{T} \mathbb{E}\Big[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t^2(k \mid \boldsymbol{X}_t)}\Big]} \quad \text{by choosing } \eta_2 = \tau_k^* \sqrt{\frac{p_k}{\sum_{t=1}^{T} \mathbb{E}\big[\frac{\mathbb{1}\{Y_t = k\}}{\pi_t^2(k \mid \boldsymbol{X}_t)}\big]}}
\end{aligned}$$

□

To prove Theorem 3.6, we follow a similar argument in Cesa-Bianchi & Lugosi (2006) with two introduced lemmas. Additionally, our proof relies on the assumption that the check loss function ρ_α is bounded. It holds once the score function is bounded, e.g., the softmax, APS, and RAPS scores utilized in our study. Therefore, without loss of generality, we assume $|\rho_\alpha(\cdot, \cdot)| \leq 1$.

Lemma D.1. *Let X be a random variable with $a \leq X \leq b$. Then for any $s \in \mathbb{R}$,*

$$\ln \mathbb{E}[\exp(sX)] \leq s\mathbb{E}[X] + \frac{s^2(b-a)^2}{8}.$$

Lemma D.2. *For all $J \geq 2$, for all $\beta_2 \geq \beta_1 \geq 0$, and for all $d_j \geq 0, j \in [J]$ such that $\sum_{j \in [J]} \exp(-\beta_1 d_j) \geq 1$,*

$$\ln \frac{\sum_{j \in [J]} \exp(-\beta_1 d_j)}{\sum_{j \in [J]} \exp(-\beta_2 d_j)} \leq \frac{\beta_2 - \beta_1}{\beta_1} \ln J.$$

Proof to Theorem 3.6. For the notation simplicity, let $L_{j,k}^t = \sum_{t'=1}^{t} \Delta_{t',k} \rho_\alpha(s^{t'-1}(\boldsymbol{X}_{t'}, k), \tau^{t'-1})$ be the accumulative weighted check loss (up to time t) with j-th expert for class k, and $j_k^t \in \operatorname{argmin}_{j \in [J]} L_{j,k}^t$ denote an expert with the smallest accumulative loss up to time t for class k. After defining the weights

$$\omega_{j,k}^t = \exp(-\frac{1}{\sqrt{t+1}} L_{j,k}^t), \;\; {\omega'}_{j,k}^t = \exp(-\frac{1}{\sqrt{t}} L_{j,k}^t), \;\; \text{and } \bar{\omega}_{j,k}^t = \omega_{j,k}^t / \sum_{i \in [J]} \omega_{i,k}^t,$$

we have the below equation

$$\sqrt{t}\ln\bar{\omega}^{t-1}_{j_k^{t-1},k}-\sqrt{t+1}\ln\bar{\omega}^{t}_{j_k^{t},k}=\underbrace{(\sqrt{t+1}-\sqrt{t})\ln\frac{1}{\bar{\omega}^{t}_{j_k^{t},k}}}_{\textcircled{1}}+\underbrace{\sqrt{t}\ln\frac{\bar{\omega}'^{t}_{j_k^{t},k}}{\bar{\omega}^{t}_{j_k^{t},k}}}_{\textcircled{2}}+\underbrace{\sqrt{t}\ln\frac{\bar{\omega}^{t-1}_{j_k^{t-1},k}}{\bar{\omega}'^{t}_{j_k^{t},k}}}_{\textcircled{3}},\tag{12}$$

where

$$\textcircled{1}\leq(\sqrt{t+1}-\sqrt{t})\ln J\tag{13}$$

since $j_k^t\in\operatorname{argmin}_{j\in[J]}L^t_{j,k}$ and hence $\bar{\omega}^t_{j_k^t,k}\geq\frac{1}{J}$,

$$\begin{aligned}\textcircled{2}=\sqrt{t}\ln\frac{\sum_{j\in[J]}\exp[-\frac{1}{\sqrt{t+1}}(L^t_{j,k}-L^t_{j_k^t,k})]}{\sum_{j\in[J]}\exp[-\frac{1}{\sqrt{t}}(L^t_{j,k}-L^t_{j_k^t,k})]}&\leq\sqrt{t}\frac{\frac{1}{\sqrt{t}}-\frac{1}{\sqrt{t+1}}}{\frac{1}{\sqrt{t+1}}}\ln J\quad\text{(due to Lemma D.2)}\\&=(\sqrt{t+1}-\sqrt{t})\ln J,\end{aligned}\tag{14}$$

and

$$\begin{aligned}\textcircled{3}&=\sqrt{t}\ln\frac{\omega^{t-1}_{j_k^{t-1},k}}{\omega'^{t}_{j_k^{t},k}}+\sqrt{t}\ln\frac{\sum_{j\in[J]}\omega'^{t}_{j,k}}{\sum_{j\in[J]}\omega^{t-1}_{j,k}}\\&=\sqrt{t}\ln\frac{\exp(-\frac{1}{\sqrt{t}}L^{t-1}_{j_k^{t-1},k})}{\exp(-\frac{1}{\sqrt{t}}L^{t}_{j_k^{t},k})}+\sqrt{t}\ln\frac{\sum_{j\in[J]}\omega'^{t}_{j,k}}{\sum_{j\in[J]}\omega^{t-1}_{j,k}}=L^t_{j_k^t,k}-L^{t-1}_{j_k^{t-1},k}+\underbrace{\sqrt{t}\ln\frac{\sum_{j\in[J]}\omega'^{t}_{j,k}}{\sum_{j\in[J]}\omega^{t-1}_{j,k}}}_{\textcircled{4}}.\end{aligned}\tag{15}$$

Additionally,

$$\begin{aligned}\textcircled{4}&=\sqrt{t}\ln\frac{\sum_{j\in[J]}\exp[-\frac{1}{\sqrt{t}}(L^{t-1}_{j,k}+\Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t,k),\tau^{t-1}_{j,k}))]}{\sum_{j\in[J]}\exp(-\frac{1}{\sqrt{t}}L^{t-1}_{j,k})}\\&=\sqrt{t}\ln\frac{\sum_{j\in[J]}\omega^{t-1}_{j,k}\cdot\exp(-\frac{1}{\sqrt{t}}\Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t,k),\tau^{t-1}_{j,k}))}{\sum_{j\in[J]}\omega^{t-1}_{j,k}}\\&=\sqrt{t}\ln\sum_{j\in[J]}\bar{\omega}^{t-1}_{j,k}\cdot\exp(-\frac{1}{\sqrt{t}}\Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t,k),\tau^{t-1}_{j,k}))\\&\leq\sqrt{t}\Big[-\frac{1}{\sqrt{t}}\sum_{j\in[J]}\bar{\omega}^{t-1}_{j,k}\Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t,k),\tau^{t-1}_{j,k}))+\frac{1}{8c_k^2t}\Big]\quad\text{(due to Lemma D.1)}\\&\leq-\Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t,k),\sum_{j\in[J]}\bar{\omega}^{t-1}_{j,k}\tau^{t-1}_{j,k}))+\frac{1}{8c_k^2\sqrt{t}}\\&=-\Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t,k),\bar{\tau}^{t-1}_k))+\frac{1}{8c_k^2\sqrt{t}}.\end{aligned}\tag{16}$$

Thus, by combing (12) to (16), we have

$$\begin{aligned}&\Delta_{t,k}\rho_\alpha(s^{t-1}(\boldsymbol{X}_t,k),\bar{\tau}^{t-1}_k)))-(L^t_{j_k^t,k}-L^{t-1}_{j_k^{t-1},k})\\&\leq\sqrt{t+1}\ln\bar{\omega}^t_{j_k^t,k}-\sqrt{t}\ln\bar{\omega}^{t-1}_{j_k^{t-1},k}+\frac{1}{8c_k^2\sqrt{t}}+2(\sqrt{t+1}-\sqrt{t})\ln J\end{aligned}\tag{17}$$

By taking the sum over $t \in [T]$ for both sides of (17), we have

$$\sum_{t=1}^{T} \Delta_{t,k} \rho_\alpha(s^{t-1}(\boldsymbol{X}_t, k), \bar{\tau}_k^{t-1})) - L_{j_k^T, k}^T \leq \ln J + \frac{1}{8c_k^2} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} + 2(\sqrt{T+1} - 1) \ln J$$

$$\Longrightarrow \sum_{t=1}^{T} \Delta_{t,k} \rho_\alpha(s^{t-1}(\boldsymbol{X}_t, k), \bar{\tau}_k^{t-1})) - \min_{j \in [J]} \sum_{t=1}^{T} \Delta_{t,k} \cdot \rho_\alpha(s(\boldsymbol{X}_t, k), \tau_{j,k}^{t-1}) \leq \frac{1}{4c_k^2} \sqrt{T} + 2\sqrt{T} \ln J$$

□