
Calibration-Aware Online Adaptation under Label Shift

Jiun Jeong¹

Byeongwoo An²

Gi-Soo Kim^{1,2,†}

Kyubo Shin^{3,†}

¹Graduate School of Artificial Intelligence, UNIST, Ulsan, Republic of Korea

²Department of Industrial Engineering, UNIST, Ulsan, Republic of Korea

³R&D Division, THYROSCOPE INC., Ulsan, Republic of Korea

Abstract

We study online adaptation of a pre-trained base classifier to streaming unlabeled data under label shift, where the marginal label proportions in the stream differ from those in the offline training data. We consider the case where the base classifier’s model class may be misspecified, motivating a separate, calibrated auxiliary classifier used solely to estimate the target label proportions. While many works have studied this setting, it is less understood how calibration quality affects the performance of the adapted base classifier in an online setting. In this paper, we thoroughly analyze the estimation error of the adapted base classifier after the deployment of a novel algorithm that estimates the target label proportions in an online fashion and dynamically adapts the base classifier using the estimated proportions. We decompose the error into a term that vanishes over time and a term determined by calibration quality. Moreover, we characterize an explicit trade-off between calibration granularity and finite-sample calibration error, and propose a novel calibration strategy which effectively balances this trade-off.

1 INTRODUCTION

The mismatch between training and test data distributions, known as dataset shift, is a common issue in real-world supervised learning that often degrades predictive model performance after deployment. Among various forms of shift, this paper focuses on label shift, where the marginal label distribution $P(y)$ changes while the class-conditional feature distribution $P(x|y)$ remains unchanged.

This scenario is frequently encountered in healthcare do-

main. For instance, disease prevalence varies due to seasonality or demographics, while the symptoms remain consistent [Zhang et al., 2023, Lipton et al., 2018]. Moreover, clinical datasets are often collected from patients rather than healthy individuals. This naturally over-represents positive cases in the training phase relative to real-world deployment.

Adaptation to label shift is usually performed by estimating the target label distribution $P_{\text{te}}(y)$ and then reweighting base predictions or optimizing the weighted risk estimates. The reweighting approach is motivated by the fact that under $P_{\text{tr}}(x|y) = P_{\text{te}}(x|y)$,

$$P_{\text{te}}(y|x) = \frac{P_{\text{tr}}(y|x) \frac{P_{\text{te}}(y)}{P_{\text{tr}}(y)}}{\sum_{y'=1}^K P_{\text{tr}}(y'|x) \frac{P_{\text{te}}(y')}{P_{\text{tr}}(y')}} \quad (1)$$

where K is the total number of classes. Provided that the base classifier $P_{\text{tr}}(y|x)$ is known or can be accurately approximated, the reweighting approach simply adjusts these base predictions using the importance weights $w_y^* = P_{\text{te}}(y)/P_{\text{tr}}(y)$ for $y = 1, \dots, K$. On the other hand, the second approach usually assumes a parametric form for the classifier under the test distribution, say $P_{\text{te}}(y|x) = p_{\theta}(y|x)$, and aims to optimize θ by minimizing the following test risk:

$$\mathcal{L}_{\text{te}}(\theta) = \mathbb{E}_{(x,y) \sim P_{\text{te}}}[\ell(p_{\theta}(\cdot|x), y)]. \quad (2)$$

Under $P_{\text{tr}}(x|y) = P_{\text{te}}(x|y)$, since we have $P_{\text{te}}(x, y) = P_{\text{tr}}(x|y)P_{\text{te}}(y) = P_{\text{tr}}(x, y)w_y^*$, this objective can be rewritten as the following weighted risk,

$$\mathcal{L}_{\text{tr}}(\theta; w^*) = \mathbb{E}_{(x,y) \sim P_{\text{tr}}}[\ell(w_y^* p_{\theta}(\cdot|x), y)], \quad (3)$$

which can be unbiasedly estimated using samples from the source distribution.

Most of the existing approaches estimate $P_{\text{te}}(y)$ (or importance weight w_y^*) using either confusion matrices (CM) [Lipton et al., 2018, Azizzadenesheli et al., 2019] or by maximizing the marginal likelihood (ML) of unlabeled test features [Saerens et al., 2002, McLachlan, 2005, Ye et al.,

[†]Corresponding authors.

2024, Sulc and Matas, 2019]. Let $\hat{y}(x)$ be either the output of hard, black-box classifier or a random label sampled from a soft, black-box classifier. The confusion matrix C of $\hat{y}$ is a $K \times K$ matrix with $C_{kk'} = P_{\text{tr}}(\hat{y}(x) = k, y = k')$. It can be easily verified that the importance weight vector $w^* = [w_1^*, w_2^*, \dots, w_K^*]^\top$ satisfies the equation

$$Cw^* = \mathbb{E}_{\text{te}}[e_{\hat{y}(x)}], \quad (4)$$

where $e_{\hat{y}(x)} \in \mathbb{R}^K$ is the one-hot vector of $\hat{y}(x)$ and the right-hand side represents the marginal distribution of the black-box classifier's outputs on the test set. Lipton et al. [2018] solves equation (4) after replacing C and $\mathbb{E}_{\text{te}}[\hat{y}(x)]$ with their empirical moment estimates, referred to as the BBSE method. The advantage of the CM method is that it works for any black-box classifier $\hat{y}(\cdot)$. On the other hand, the ML approach solves the following optimization problem $w = \arg\max_{w \in \mathcal{W}} \mathbb{E}_{\text{te}}[\log \sum_{y=1}^K P_{\text{tr}}(y|x)w_y]$ where $\mathcal{W} = \{w \mid \forall y, w_y \geq 0 \text{ and } \sum_{y=1}^K w_y P_{\text{tr}}(y) = 1\}$ and $\mathbb{E}_{\text{te}}(\cdot)$ is replaced by its empirical moment estimate. Unlike the CM method, the ML approach requires access to $P_{\text{tr}}(y|x)$. A closely related alternative to the ML approach is the Bayesian approach, which maximizes the posterior of w rather than the likelihood of unlabeled test features after imposing a Dirichlet prior on the test label proportions [Sulc and Matas, 2019, Ye et al., 2024].

The seminal work of Garg et al. [2020] shows that both the CM and ML methods are special cases of a single framework which estimates the importance weights by performing distribution matching between the empirical estimate of $P_{\text{te}}(z)$ and $\sum_y P_{\text{tr}}(z, y)w_y$ for some representation z of the feature x . In this framework, the weight vector w is estimated as the minimizer of the KL-divergence from $P_{\text{te}}(z)$ to $\sum_y P_{\text{tr}}(z, y)w_y$ which can be formulated as follows:

$$w = \arg\max_{w \in \mathcal{W}} \mathbb{E}_{\text{te}} \left[\log \sum_{y=1}^K P_{\text{tr}}(y|z)w_y \right], \quad (5)$$

where $\mathbb{E}_{\text{te}}(\cdot)$ is replaced by its empirical moment estimate. It is straightforward that the ML method corresponds to (5) with $z = x$ or $z = P_{\text{tr}}(y|x)$. We can also show that the CM method in (4) is equivalent to (5) with $z = \hat{y}(x)$, provided that the true quantities C and $\mathbb{E}_{\text{te}}(\cdot)$ are used. (Detailed proof can be found in Appendix A.) Garg et al. [2020] also derived regularity conditions under which the solution of (5) is unique and coincides with the true importance weight vector. In particular, it suffices that the set of distributions $\{P_{\text{tr}}(z|y) : y = 1, \dots, K\}$ are linearly independent. The formulation in (5) further implies that if we have a calibrated classifier $h_{\text{tr}}(y|x)$, not necessarily equal to the true conditional distribution $P_{\text{tr}}(y|x)$, but satisfying *canonical calibration* as defined in Definition 1.1, then performing ML maximization using $h_{\text{tr}}(y|x)$ instead of $P_{\text{tr}}(y|x)$ is equivalent to solving (5) with $z = \{h_{\text{tr}}(y'|x)\}_{y'=1}^K$. The $P_{\text{tr}}(\hat{y}(x))$ is also a valid candidate for $h_{\text{tr}}(y|x)$.

Definition 1.1 (Canonical Calibration). A predictive model $h_{\text{tr}}(y|x)$ is canonically calibrated on the train distribution P_{tr} if $P_{\text{tr}}(y \mid h_{\text{tr}}(\cdot|x)) = h_{\text{tr}}(y|x)$ holds for all x and y .

Some recent works explored variants of CM and ML methods for label shift estimation. Du Plessis and Sugiyama [2014] investigated other divergence measures besides the KL divergence, eg., the f-divergence or the PE divergence, for quantifying the distance between $P_{\text{te}}(x)$ and $\sum_y P_{\text{tr}}(x, y)w_y$. Wen et al. [2024] proposed to match the class distribution instead of the distribution of x . These methods, just like the ML method, require access to accurate estimate of $P_{\text{tr}}(y|x)$. In another line of work, Tian et al. [2023] employed semiparametric theory to address the label shift problem and proposed to perform moment matching (instead of distribution matching) on a specific function z of x . Fan et al. [2025] proposed to jointly learn the predictive model $P_{\text{te}}(y|x)$ and the importance weights via a bi-level optimization, where the upper-level minimizes a semi-supervised loss to train the classifier, and the importance weights are estimated in lower-level as a function of $P_{\text{te}}(y|x)$.

Most of the aforementioned methods operate on a fixed offline test set. However, real-world deployments often involve streaming data, where unlabeled target inputs arrive sequentially. For instance, a diagnostic model is required to make prediction on the fly without waiting to collect a dataset. [Wu et al., 2021]. To address this, recent works have explored online adaptation strategies [Wu et al., 2021, Bai et al., 2022, Baby et al., 2024, Dai et al., 2025]. Wu et al. [2021] applied online gradient descent for optimizing the cumulative sum of (3), where the sum is taken over the stream sequence and the label marginals along the stream are allowed to be nonstationary. The values of w_y 's at each time point were estimated using the BBSE method. Wu et al. [2021] set $\theta \in \mathbb{R}^K$ and

$$p_\theta(y|x) = \hat{P}_{\text{tr}}(y|x)\theta_y / \left\{ \sum_{y'=1}^K \hat{P}_{\text{tr}}(y'|x)\theta_{y'} \right\}, \quad (6)$$

motivated from the form in (1), where $\hat{P}_{\text{tr}}(y|x)$ denotes an estimate of the source conditional distribution. While this formulation permits arbitrary black-box classifiers for $\hat{P}_{\text{tr}}(y|x)$ and requires learning only the reweighting parameters θ_y 's, the resulting function f may be suboptimal due to the restrictiveness of its function class, especially when $\hat{P}_{\text{tr}}(y|x)$ is inaccurate. Bai et al. [2022] targeted a more challenging objective of competing with the per-round optimal hypotheses, instead of the single best hypothesis, in the same nonstationary environment. They also estimated the time-varying w_y at each time point using the BBSE method. However, instead of adopting the functional form in (6), they imposed a fully parametric form on $p_\theta(y|x)$, and ensured that the objective (3) is convex in θ and thus amenable to online convex optimization (OCO) methods. Baby et al. [2024]

aimed to achieve best-of-both-worlds of Wu et al. [2021] and Bai et al. [2022]. Their algorithm learns per-round optimal reweighting parameters in nonstationary environments, and those parameters can be used to adjust any black-box classifier. However, instead of optimizing θ_y 's as in Wu et al. [2021], they rather set $\theta_y = w_y$. Instead, they learned w_y 's by employing online regression oracles on the estimates of the BBSE method so as to effectively control their variance. While their approach effectively learns the time-varying w_y 's, setting $\theta_y = w_y$ becomes questionable when $\hat{P}_{\text{tr}}(y|x)$ is inaccurate, since even the true importance weights may not minimize the loss of f in this case.

The aforementioned online methods all rely on the BBSE method, which allows unbiased estimation of w when C is known, even with only a single test sample used to estimate $\mathbb{E}_{\text{te}}[\hat{y}(x)]$. This enables unbiased estimation of the time-varying risk in nonstationary environments. This unbiasedness plays a crucial role—especially in Bai et al. [2022]—in ensuring that the expectation of $\mathcal{L}_{\text{tr}}(\theta; w)$ in (3) remains convex in θ , whereas $\mathcal{L}_{\text{tr}}(\theta; w)$ itself may be nonconvex due to possibly negative estimates of w . However, none of the three methods explicitly analyzes the influence of the estimation error of C on their final regret bounds. Bai et al. [2022] does analyze, for a fixed θ , the difference in the risk estimator $\mathcal{L}_{\text{tr}}(\theta; w)$ when w is computed using the true C versus when w is computed using an estimate of C based on a finite number of training samples. However, they do not translate this analysis into a final regret bound, which would additionally require covering arguments over the parameter space of θ .

Recently, Dai et al. [2025] proposed estimating w^* by solving the constrained optimization problem in (5) with $z = \hat{y}(x)$ using projected gradient ascent (PGA), building on Šipka et al. [2022]. The constraint $w \in \mathcal{W}$ enforces positivity of w , thereby guaranteeing the convexity of the objective in (3) even when $P_{\text{tr}}(y|\hat{y}(x))$ is replaced by an estimate. Apart from modifying the estimation procedure for w , Dai et al. [2025] largely follows the framework of Bai et al. [2022] by imposing a fully parametric form on $f(x; \theta)$ and targeting dynamic regret minimization via OCO in nonstationary environments. However, Dai et al. [2025] does not analyze the estimation error of the w arising from replacing $\mathbb{E}_{\text{te}}(\cdot)$ and $P_{\text{tr}}(y|\hat{y}(x))$ with their empirical estimates. Moreover, as noted earlier, only a single sample is available at each time step to estimate the time-varying $\mathbb{E}_{\text{te}}(\cdot)$ in a nonstationary environment. While a single sample suffices for unbiased estimation of w via the BBSE method (albeit with potentially high variance), solving (5) with one sample forces the importance weights to concentrate on a single label—specifically, the label y that maximizes $\hat{P}_{\text{tr}}(y|z)/P_{\text{tr}}(y)$. Finally, the dynamic regret optimized in Dai et al. [2025] is defined with respect to a risk function constructed using the estimated w , rather than the true w , and thus does not correspond to the true dynamic re-

gret. Also, the non-stationarity term in their bound depends on estimator variability rather than actual environmental changes.

As discussed above, previous works have considered various methods for estimating w under both offline and online label shift, most of which can be understood within the unified framework of Garg et al. [2020]. The ML method corresponds to the case $z = x$. A caveat of the ML method is that $P_{\text{tr}}(y|x)$ is generally more difficult to estimate accurately than $P_{\text{tr}}(y|\hat{y}(x))$. Consequently, many works have proposed calibration strategies for $P_{\text{tr}}(y|x)$, such as temperature scaling and bias-corrected temperature scaling [Alexandari et al., 2020]. However, these approaches are largely heuristic, as they restrict the calibration function class to temperature-scaled variants of the original classifier and do not explicitly characterize how miscalibration error scales with finite sample size. Meanwhile, across multiple empirical studies, it has been consistently observed that the ML method often outperforms BBSE, despite the presence of potential miscalibration errors [Alexandari et al., 2020, Garg et al., 2020]. Garg et al. [2020] analyzed the estimation error in w under their unified framework and showed that it depends on the eigenvalue of the soft confusion matrix of the classifier $P_{\text{tr}}(y|z)$. This result implies that a coarse representation z , such as $z = \hat{y}(x)$, leads to a smaller eigenvalue and hence larger estimation error, whereas a finer representation z , such as $z = x$, yields a larger eigenvalue and thus smaller estimation error. Despite this observation, relatively few works have explored alternative choices for the representation z beyond $z = \hat{y}(x)$ and $z = x$.

2 CONTRIBUTIONS

In this paper, we develop a theoretical framework that clarifies the role of calibration in online label shift adaptation. We also propose a two-stage online procedure: (i) online estimation of the target prior $P_{\text{te}}(y)$ from unlabeled inputs using a newly proposed calibrated classifier, and (ii) online adaptation of the predictive model by minimizing the weighted cross-entropy objective using the weight estimates from (i). Our analysis establishes a population-level error bound that explicitly decomposes the error into a term that vanishes over time and a term governed by calibration quality, revealing a trade-off between calibration granularity and finite-sample calibration error. Our proposed calibration procedure effectively balances this trade-off, minimizing the population-level error bound of the adapted classifier.

Our contributions and distinctions from prior work are summarized as follows:

- We propose a new online estimation procedure of $P_{\text{te}}(y)$ that applies Soft-Bayes algorithm Orseau et al. [2017] to optimize (5). We importantly note that the PGA approach in Dai et al. [2025] performs multiple

inner iterations at each time step until convergence, and thus is not fully online in the strict per-sample update sense. Moreover, PGA requires explicit projection onto the constraint set at each update, which can be computationally expensive. In contrast, we employ Soft-Bayes in a fully online manner and avoid costly projection steps. Finally, we provide a rigorous per-time-step analysis of the estimation error of $P_{te}(y)$, which is not explicitly derived in Dai et al. [2025].

- In the optimization of (5), we replace $P_{tr}(y|\hat{y}(x))$, as used by Šipka et al. [2022] and Dai et al. [2025], with a newly proposed calibrated classifier $h_{tr}(y|x)$, which balances the trade-off between calibration granularity and finite-sample calibration error. This new calibrated classifier is constructed via histogram binning within the label-partitioned regions induced by $\hat{y}(x)$.
- We importantly note that even in the stationary online setting, dynamic regret analysis is required, as the weights defining the objective in (3) are unknown and are updated at each time step. We leverage dynamic online convex optimization (OCO) theory [Mokhtari et al., 2016] to analyze dynamic regret of the adapted classifier. Moreover, we establish a direct link between the convergence of the model parameters and the temporal variability of the estimated importance weights.
- In this work, we focus on the online setting in which the target prior $P_{te}(y)$ remains fixed over time. This contrasts with recent approaches to online label shift that directly address nonstationary environments with time-varying priors. Notably, even when applied to the stationary regime, these methods still inherit the limitations discussed above. We therefore turn our attention to the comparatively underexplored stationary online setting and provide a principled analysis that addresses these shortcomings.

3 PROBLEM SETUP AND OVERVIEW

3.1 DEFINITIONS AND ASSUMPTIONS

Notations We define $P(\cdot|E) = [P(y = 1|E), \dots, P(y = K|E)]^\top$ for any event E . For a set A , we define $|A|$ as the cardinality of A . We denote the minimum eigenvalue of a matrix M as $\lambda_{min}(M)$. We define $[K] = \{1, 2, \dots, K\}$ for any natural number K .

Definition 3.1 (Spaces and Distributions.). We define the input feature space as $\mathcal{X} \subset \mathbb{R}^d$ and the label space as $\mathcal{Y} = [K]$ where $[K] = \{1, \dots, K\}$. We denote the training distribution as $P_{tr}(x, y)$ and the test distribution as $P_{te}(x, y)$. The class probability vectors for the training distribution and test distribution are denoted by $\pi_{tr} = [P_{tr}(y = 1), \dots, P_{tr}(y = K)]^\top$ and $\pi_{te} = [P_{te}(y = 1), \dots, P_{te}(y = K)]^\top$ respectively, and belong to the

$(K - 1)$ -dimensional probability simplex $\Delta^{K-1} = \{\pi \in \mathbb{R}^K : \sum_k \pi_k = 1, \pi_k \geq 0\}$. We also define the true importance weight vector w^* as $w^* = \pi_{te} \oslash \pi_{tr}$, where $\oslash$ denotes element-wise division.

Definition 3.2 (Predictive Models.). We define two predictive models with distinct roles. The first is an adaptive predictor, $p_\theta(y | x)$ parameterized by θ , which is optimized to minimize a supervised loss under the test distribution. The second is a fixed predictor $f : \mathcal{X} \rightarrow \Delta^{K-1}$ used solely to estimate the test label distribution π_{te} . The construction of f is described in Section 4.3. In this work, we use multinomial logistic model for $p_\theta(y|x)$.

We make following basic assumptions following literature on label shift. Additional assumptions will appear in the course of the development of our theory.

Assumption 3.3 (Label Shift). $P_{te}(x | y) = P_{tr}(x | y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Assumption 3.4 (Label Distributions). Every class probability in the train distribution and the test distribution is strictly positive, i.e., $\pi_{tr, \min} := \min_k \pi_{tr, k} > 0$ and $\pi_{te, \min} := \min_k \pi_{te, k} > 0$. We also assume that π_{tr} is known.

3.2 PROBLEM FORMULATION

Consider an online setting where test data points $x_1^*, x_2^*, \dots, x_T^*$ arrive sequentially from P_{te} . In this setting, we have two main goals: (1) to adapt the model parameter θ to minimize the expected loss on the test distribution, and (2) to accurately estimate the true test label distribution π_{te} required for that process.

Goal 1: Model Adaptation (θ). The primary goal is to minimize the test risk in (2), with the loss set as the cross entropy, $\ell(p_\theta(\cdot | x), y) = -\log p_\theta(y | x)$. Under Assumption 3.3, this is equivalent to minimizing (3),

$$\mathcal{L}_{tr}(\theta; w) = \mathbb{E}_{(x, y) \sim P_{tr}} [-w_y \log p_\theta(y | x)].$$

The optimal parameter θ^* is defined as

$$\theta^* = \arg \min_{\theta} \mathcal{L}_{tr}(\theta; w^*).$$

We assume that θ^* is unique and that $\|\theta^*\|_2 \leq S$ for some $S > 0$.

Goal 2: Label Distribution Estimation (π). Since the true weights w^* are unknown, we need to estimate them. We aim to solve the following optimization problem,

$$w_f = \arg \max_{w \in \mathcal{W}} \mathbb{E}_{te} \left[\log \sum_{k=1}^K f_k(x^*) w_y \right], \quad (7)$$

where $f(\cdot)$ is defined in Definition 3.2. If $f(\cdot)$ is canonically calibrated on the training distribution as in Definition 1.1,

(7) is equivalent to (5) with $z = f(x^*)$, as aforementioned in Section 1. Furthermore, under regularity assumptions (Assumption 5.1), w_f is unique and $w_f = w^*$.

Since π_{tr} are known, we express the negative of the objective in (7) as a function of π , $L(\pi) = E_{te}[\tilde{\ell}_s(x^*; \pi)]$, where

$$\tilde{\ell}_s(x^*; \pi) = -\log \left(\sum_{k=1}^K f_k(x^*) \pi_k / \pi_{tr,k} \right).$$

For notational convenience, let $q(x) := f(x) \odot \pi_{tr}$, we can rewrite $\tilde{\ell}_s(x^*; \pi) = -\log(q(x^*)^\top \pi)$. We also define its minimizer over a constrained set $\Pi_c = \{\pi \in \Delta^{K-1} : \min_k \pi_k \geq \min(\pi_{te, \min}, \pi_{tr, \min}/2T)\}$ as

$$\pi_f := \arg \min_{\pi \in \Pi_c} L(\pi). \quad (8)$$

We note that $\pi_{te} \in \Pi_c$ as well. We construct f in Section 4.2 so that it estimates $P_{tr}(y|z)$ for some carefully chosen representation z . A gap arises between π_f and π_{te} , which depends on the calibration error $\mathcal{E}(f) := (\mathbb{E}_{tr}[\sum_{y=1}^K (P_{tr}(y|z) - f_y(x))^2])^{1/2}$. We analyze this gap in Section 5, and propose a representation z which minimizes an upper bound of the gap.

4 PROPOSED ALGORITHM

4.1 OVERVIEW

We propose an algorithm consisting of an offline construction of f followed by two online stages: estimation of π_f and adaptation of θ . The predictor f is obtained by refining the partition of the input space induced by another hard classifier $\hat{y} : \mathcal{X} \rightarrow [K]$, which need not be accurate and may be any black-box classifier. Then π_f is estimated in Stage 1 in an online fashion. Stage 2 adapts the model parameters θ using the estimated importance weights. The algorithm is initialized with the label distribution estimate $\hat{\pi}^{(1)} = \pi_{tr}$ and the model parameters $\hat{\theta}^{(1)}$ pre-trained on P_{tr} . For convenience, we set $\hat{y}(x) = \arg \max_{k \in [K]} p_{\hat{\theta}^{(1)}}(y = k|x)$. Algorithm 1 summarizes the full two-stage online procedure. Given the unlabeled test input x_t^* at time t , Stage 1 outputs and estimate $\hat{\pi}^{(t+1)}$ and computes the current weights $\hat{w}^{(t)} = \hat{\pi}^{(t)} \odot \pi_{tr}$. Then Stage 2 performs OGD update from $\hat{\theta}^{(t)}$ to $\hat{\theta}^{(t+1)}$.

4.2 CONSTRUCTION OF f VIA HISTOGRAM BINNING.

Setting $f_k(x) = P_{tr}(y = k|\hat{y}(x))$ for $k = 1, \dots, K$ and solving (8) gives $\pi_f = \pi_{te}$ under regularity assumption (Assumption 5.1). However, a gap arises between π_f and π_{te} as $P_{tr}(y = k|\hat{y}(x))$ is replaced by an empirical estimate, i.e., the empirical proportion of the k -th class in an held out

Algorithm 1 MALO

- 1: **Require:** $K, \pi_{tr}, \mathcal{D}_{cal} \sim P_{tr}, \mathcal{B} \sim P_{tr}, \hat{\theta}^{(1)}, \hat{y}(x)$
 - 2: **Require:** Soft-Bayes stepsize $\eta_t = \sqrt{\frac{\log K}{2K^t}}$, Stochastic OGD stepsize γ
 - 3: **Input:** test stream $\{x_t^*\}_{t=1}^T$
 - 4: Construct $f \leftarrow \text{HistBin}(\hat{y}, \mathcal{D}_{cal}, s)$
 - 5: Initialize $\hat{\pi}^{(1)} \leftarrow \pi_{tr}$
 - 6: **for** $t = 1, 2, \dots, T$ **do**
 - 7: Observe x_t^*
 - 8: **Predict:** $\hat{p}_t(\cdot | x_t^*)$
 - 9: $q(x_t^*) \leftarrow f(x_t^*) \odot \pi_{tr}$
 - 10: $v_{t,k} \leftarrow \frac{\hat{\pi}_k^{(t)} q_k(x_t^*)}{q(x_t^*)^\top \hat{\pi}^{(t)}} \quad \forall k \in \{1, \dots, K\}$
 - 11: $\alpha_t \leftarrow \eta_{t+1}/\eta_t$
 - 12: $\hat{\pi}^{(t+1)} \leftarrow \alpha_t ((1 - \eta_t)\hat{\pi}^{(t)} + \eta_t v_t) + (1 - \alpha_t)\hat{\pi}^{(1)}$
 - 13: $\hat{w}^{(t)} \leftarrow \hat{\pi}^{(t)} \odot \pi_{tr}$
 - 14: $\hat{\theta}^{(t+1)} \leftarrow \hat{\theta}^{(t)} - \gamma \nabla_{\theta} \hat{\mathcal{L}}_{tr}(\hat{\theta}^{(t)}; \hat{w}^{(t)}, \mathcal{B})$
 - 15: **end for**
 - 16: **Output:** $\hat{\theta}^{(t)}$
-

dataset from the source distribution. Garg et al. [2020] made a conjecture that this gap depends in part on the granularity of the representation $z = \hat{y}(x)$. In this paper, we build on this observation and propose a classifier f that utilizes a finer representation z . We let $\mathcal{D}_{cal} = \{(x_i^{cal}, y_i^{cal})\}_{i=1}^{n_{cal}} \sim P_{tr}$ be a held-out dataset from the source distribution and call it a *calibration set*. We first divide the set $\{x_i^{cal}\}_{i=1}^{n_{cal}}$ into K regions according to the value of $\hat{y}(x)$. Then for each $k = 1, \dots, K$, we further divide the k -th region into s equal-mass bins by performing histogram binning on the values of $p_{\hat{\theta}^{(1)}}(y = k|x_i)$ for the x_i 's in that region, where x_i 's are from another, independent held-out dataset $\mathcal{Q}$ from the source distribution. Then we define f as $f(x) = \hat{p}_{h_B(x)}$ where $h_B(x)$ denotes the bin index of x and $\hat{p}_b$ is the empirical class distribution of samples from $\mathcal{D}_{cal}$ in bin b and $B = Ks$ denotes the total number of bins. We note that $f_y(x)$ is an unbiased estimator of $P_{tr}(y|z)$ with $z = h_B(x)$, which is a finer representation than $z = \hat{y}(x)$ when $s > 1$. We denote $f_c(x) = P_{tr}(y|h_B(x))$, where c stands for 'calibrated'. The choice of s controls a trade-off between calibration granularity and finite-sample estimation error, which we formalize in Section 5.3.

4.3 STAGE 1: ONLINE ESTIMATION OF π

In the offline setting, projected gradient descent can be applied to solve the empirical version of (8). On the other hand, several works view the $-L(\pi)$ as the negative marginal log-likelihood of x^* with the unknown $P_{tr}(y = k | x^*)$ replaced by $f_k(x^*)$, and employ the EM algorithm [Alexandari et al., 2020, Sulc and Matas, 2019, Miller and Futoma, 2023]. Royer and Lampert [2015] devised an online version of the EM algorithm, but without theoretical analysis.

In this work, we use the Soft-Bayes algorithm. When test sample $x_t^* \sim P_{te}$ arrives, the algorithm observes the loss $\tilde{\ell}_s(x_t^*; \hat{\pi}^{(t)}) = -\log(q(x_t^*)^\top \hat{\pi}^{(t)})$ based on the current estimate $\hat{\pi}^{(t)}$. It then outputs the next estimate $\hat{\pi}^{(t+1)}$ using the Soft-Bayes algorithm.

The Soft-Bayes algorithm computes $v_t \in \Delta^{K-1}$ using $\hat{\pi}^{(t)}$ as the prior and $q(x_t^*)$ as the observation. Its k -th component is defined as $v_{t,k} = \hat{\pi}_k^{(t)} q_k(x_t^*) / \{\sum_{j=1}^K \hat{\pi}_j^{(t)} q_j(x_t^*)\}$. Then computes the next estimate $\hat{\pi}^{(t+1)}$ as a combination of current estimate $\hat{\pi}^{(t)}$, the posterior v_t , and the initial estimate $\hat{\pi}^{(1)}$:

$$\hat{\pi}^{(t+1)} = \alpha_t \left((1 - \eta_t) \hat{\pi}^{(t)} + \eta_t v_t \right) + (1 - \alpha_t) \hat{\pi}^{(1)}, \quad (9)$$

where η_t denotes the learning rate at time t , and $\alpha_t = \eta_{t+1}/\eta_t \in (0, 1]$. Then, the regret bound for the Soft-Bayes algorithm is as follows:

Theorem 4.1 (Orseau et al. [2017], Theorem 10). *Suppose we set $\eta_t = \sqrt{\frac{\ln K}{2Kt}}$. Then, the cumulative surrogate regret is bounded by:*

$$\begin{aligned} \tilde{R}(T) &:= \sum_{t=1}^T \tilde{\ell}_s(x_t^*; \hat{\pi}^{(t)}) - \sum_{t=1}^T \tilde{\ell}_s(x_t^*; \pi_f) \\ &\leq 2\sqrt{2(T+1)K \ln K} + \left(\frac{K}{2} + \ln K \right) \ln(T+1) + \ln K. \end{aligned}$$

Consequently, the regret satisfies $\tilde{R}(T) = \mathcal{O}(\sqrt{TK \ln K})$.

Since this regret bound is independent of the upper bound of gradient of $\tilde{\ell}_s$, Soft-Bayes ensures the regret bound does not become too loose even if some rounds induce extremely large gradients. This contrasts with exponentiated gradient and online gradient descent, where the regret depends on the upper bound of the gradient. Moreover, Soft-Bayes guarantees sublinear boundedness of the path length $\hat{V}_T = \sum_{t=1}^{T-1} \|\hat{\pi}^{(t+1)} - \hat{\pi}^{(t)}\|_2$, which is requisite for the dynamic OCO analysis in Stage 2.

4.4 STAGE 2: ONLINE ADAPTATION OF θ

At time t , the importance weights are computed from the current estimate $\hat{\pi}^{(t)}$ as $\hat{w}^{(t)} = \hat{\pi}^{(t)} \odot \pi_{tr}$ and a time-varying objective function for θ is defined as:

$$F_t(\theta) := \mathcal{L}_{tr}(\theta; \hat{w}^{(t)}) = \mathbb{E}_{(x,y) \sim P_{tr}} [-\hat{w}_y^{(t)} \log p_\theta(y | x)].$$

An empirical estimate of $F_t(\theta)$ is constructed using a batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^m$ sampled i.i.d. from P_{tr} as

$$\hat{F}_t(\theta) := \hat{\mathcal{L}}_{tr}(\theta; \hat{w}^{(t)}, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,y) \in \mathcal{B}} [-\hat{w}_y^{(t)} \log p_\theta(y | x)].$$

With learning rate γ and batch gradient $g_t(\theta) = \nabla_\theta \hat{F}_t(\theta)$, the parameters are updated as $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - \gamma g_t(\hat{\theta}^{(t)})$. Let

$\hat{\theta}_t^* := \arg \min_{\theta \in \Theta} \hat{F}_t(\theta)$ denote the instantaneous empirical optimum at time t , where $\Theta = \{\theta \in \mathbb{R}^{Kd} \mid \|\theta\|_2 \leq S'\}$ for some known, finite upper bound S' for S . Tracking this sequence of shifting optima $\hat{\theta}_t^*$ can be viewed as a Dynamic Online Convex Optimization (DOCO) problem, whose performance is evaluated by the dynamic regret $\hat{R}_{dyn}(T) = \sum_{t=1}^T \hat{F}_t(\hat{\theta}^{(t)}) - \sum_{t=1}^T \hat{F}_t(\hat{\theta}_t^*)$. The regret bounds $\tilde{R}(T)$ and $\hat{R}_{dyn}(T)$ will later be translated into error bounds between the respective online estimates and the target estimands.

5 THEORETICAL ANALYSIS

Our analysis has two steps. In Section 5.1, we bound the label distribution estimation error $\|\hat{\pi}^{(t)} - \pi_{te}\|_2$. In Section 5.2, we characterize how this error propagates to the model parameters θ via dynamic online convex optimization. We adapt the analysis of Garg et al. [2020] to the online setting. We also identify a term that is determined by calibration quality, $\mathcal{E}(f)^2/(\lambda_f^{te})^2$, where $\mathcal{E}(f)$ is the calibration error and λ_f^{te} is the minimum eigenvalue of $E_{te}[f(x)f(x)^\top]$, which similarly appears in the analysis of Garg et al. [2020]. Since histogram binning controls both quantities through the number of bins B , Section 5.3 analyzes this trade-off and derives the optimal choice of B .

5.1 STEP I: ONLINE ESTIMATION OF π

Assumption 5.1 (Regularity Condition). Let $f_{0,c}(x) = P_{tr}(\cdot | \hat{y}(x))$. The expected outer product $\mathbb{E}_{tr}[f_{0,c}(x)f_{0,c}(x)^\top]$ under the training distribution has minimum eigenvalue $\lambda_{f_{0,c}}^{tr} > 0$.

Let $\lambda_{f_c}^{tr} = \lambda_{min}(\mathbb{E}_{tr}[f_c(x)f_c(x)^\top])$, where f_c is the classifier based on finer representation than f_0 . We later show in Lemma 5.8(i) that $\lambda_{f_c}^{tr} \geq \lambda_{f_{0,c}}^{tr}$, with equality holding when $s = 1$. We also relate $\lambda_{f_c}^{tr}$ and $\lambda_f^{tr} = \lambda_{min}(\mathbb{E}_{tr}[f(x)f(x)^\top])$ in Lemma 5.8(i) and Proposition B.1, where f is the classifier that we use in estimation of π and which is unbiased estimator of f_c . Its positivity is guaranteed with high probability.

The online estimation error of $\hat{\pi}^{(t)}$ can be decomposed as $\|\hat{\pi}^{(t)} - \pi_{te}\|_2 \leq \|\hat{\pi}^{(t)} - \pi_f\|_2 + \|\pi_f - \pi_{te}\|_2$, where $\pi_f = \arg \min_{\pi \in \Pi_c} L(\pi)$. Under Assumptions 3.3, 3.4, 5.1, 5.6 and 5.7, $L(\pi)$ is strongly convex (Proposition B.1). Combined with online-to-batch conversion, the Soft-Bayes regret $\tilde{R}(T) = \mathcal{O}(\sqrt{TK \log K})$ yields an $\mathcal{O}(1/(\sqrt{T}))$ rate to π_f (Theorem B.2). Moreover, the calibration gap satisfies $\|\pi_f - \pi_{te}\|_2 \leq C' \mathcal{E}(f)/\lambda_f^{tr}$, where C' depends on $\pi_{tr,min}, \pi_{tr,max}$, and w_{min}^* (Lemma B.4). Combining yields the following bound.

Theorem 5.2 (Online Error Bound for Label Estimation). *Under Assumptions 3.3, 3.4, 5.1, 5.6 and 5.7, the total av-*

erage squared error with respect to π_{te} is bounded with probability at least $1 - \delta$ by:

$$\frac{1}{T} \sum_{t=1}^T \|\hat{\pi}^{(t)} - \pi_{te}\|_2^2 \leq \mathcal{O}\left(\frac{\sqrt{\log(1/\delta)}}{\lambda_f^{\text{tr}} \sqrt{T}}\right) + \mathcal{O}\left(\frac{\mathcal{E}(f)^2}{(\lambda_f^{\text{tr}})^2}\right).$$

The precise bound and proof are given in Section B.4. We remark that we can project the estimates $\hat{\pi}^{(t)}$ for timesteps $t \in [T]$ such that $\hat{\pi}^{(t)} \notin \Pi'_c$ into Π'_c , where $\Pi'_c = \{\pi \in \Delta^{K-1} : \min_k \pi_k \geq \epsilon_\pi\}$, where $\epsilon_\pi > 0$ is some known lower bound of $\pi_{te, \min}$. The resulting estimates will still satisfy the bounds above due to Projection Inequality.

5.2 STEP II: MODEL-SIDE DYNAMIC OCO FOR θ

The following assumptions ensure smoothness and strong convexity of $\hat{F}_t(\theta)$.

Assumption 5.3 (Feature Space Properties). Feature vectors are bounded: $\|x\|_2 \leq L_x$.

Assumption 5.4 (Positive Definite Second Moment). The population second-moment matrix $M_x := \mathbb{E}_{x \sim P_{\text{tr}}}[xx^\top]$ is positive definite: $\lambda_{\min}(M_x) \geq \lambda_x > 0$.

The average squared error $\frac{1}{T} \sum_t \mathbb{E}[\|\hat{\theta}^{(t)} - \theta^*\|_2^2]$ can be decomposed into three terms: (a) an optimization tracking error from tracking the $\hat{\theta}_t^+$, (b) a weight-estimation error due to inaccuracies in $\hat{w}^{(t)}$, and (c) a finite-sample error due to the source validation batch $\mathcal{B}$ (Section D.1). Term (a) is controlled by dynamic regret and depends on the path length $\hat{V}_T := \sum_{t=1}^{T-1} \|\hat{\theta}_{t+1}^+ - \hat{\theta}_t^+\|_2$. Since $\hat{F}_t$ depends on $\hat{w}^{(t)}$, this path length can be bounded in terms of the movement of $\hat{\pi}^{(t)}$ under the Soft-Bayes update (Lemma C.7 and Theorem C.8). Term (b) is bounded by the label estimation error from Section 5.1 through the Lipschitz continuity of the optimal parameter (Corollary C.6). Combining these bounds with term (c) yields the following convergence guarantee.

Theorem 5.5 (Online Error Bound for Model Adaptation). Let $\delta_B \in (0, 1)$ and suppose $|\mathcal{B}| \geq C \log(d/\delta_B)$ for some constant C defined in Section C.3. Then, with probability $1 - 2\delta - \delta_B$, under Assumptions 3.3, 3.4, 5.1, 5.3, 5.4, 5.6 and 5.7, the expected average squared error converges as:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\hat{\theta}^{(t)} - \theta^*\|_2^2] &\leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{\sqrt{\log(1/\delta)}}{\lambda_f^{\text{tr}} \sqrt{T}}\right) \\ &\quad + \mathcal{O}\left(\frac{\mathcal{E}(f)^2}{(\lambda_f^{\text{tr}})^2}\right) + \mathcal{O}\left(\frac{1}{|\mathcal{B}|}\right). \end{aligned}$$

The precise bound and proof are given in Section D.1. We remark that we can project the estimates $\hat{\theta}^{(t)}$ for timesteps $t \in [T]$ such that $\hat{\theta}^{(t)} \notin \Theta$ into Θ , where $\Theta = \{\theta \in \mathbb{R}^{Kd} \mid \|\theta\|_2 \leq S'\}$ for some known, finite upper bound S' for S . The resulting estimates will still satisfy the bounds above due to Projection Inequality.

As $T \rightarrow \infty$, the error converges to $\mathcal{O}(\mathcal{E}(f)^2/(\lambda_f^{\text{tr}})^2 + 1/|\mathcal{B}|)$. Since the $1/|\mathcal{B}|$ term can be made small with a large source validation batch, the asymptotic performance is dominated by $\mathcal{E}(f)^2/(\lambda_f^{\text{tr}})^2$. For histogram-binning calibration, both $\mathcal{E}(f)$ and λ_f^{tr} depend on the number of bins $B = Ks$, which we analyze next.

5.3 ANALYSIS OF CALIBRATION GAP

In this section, we analyze the term $\mathcal{E}(f)^2/(\lambda_f^{\text{tr}})^2$ that appears in Theorem 5.5. Define the binning function $h_B : \mathcal{X} \rightarrow \{1, \dots, B\}$, and the population and empirical bin-conditional label distributions of bin b as p_b and $\hat{p}_b$ respectively. We calculate $\hat{p}_b$ using a calibration set $\mathcal{D}_{\text{cal}}$. We make the following assumptions.

Assumption 5.6 (Bin mass). For each bin b created by the proposed binning scheme h_B , denote the bin mass $P_b := P_{\text{tr}}(h_B(X) = b)$ and let $P_{\min} = \min_{b \in [B]} P_b$. We assume that $P_{\min} = \mathcal{O}(1/B) > 0$.

Assumption 5.7 (Calibration Setup). The calibration data $\mathcal{D}_{\text{cal}}$ are drawn i.i.d. from P_{tr} , and we let $n_{\text{cal}} := |\mathcal{D}_{\text{cal}}|$, which is assumed to be sufficiently large so that

$$n_{\text{cal}} \geq \max \left\{ \frac{16}{P_{\min}^2} \log \left(\frac{2}{\delta} \right), \frac{128B}{(\lambda_{f_{0,c}}^{\text{tr}})^2} \log \left(\frac{2B}{\delta} \right) \right\},$$

for confidence level $\delta \in (0, 1)$ and the value of B that is proposed later.

Then, the following lemma characterizes how the number of bins B affects $\lambda_f^{\text{tr}} := \lambda_{\min}(E_{\text{tr}}[f(x)f(x)^\top])$ and $\mathcal{E}(f)$.

Lemma 5.8 (Calibration Trade-off). (i) *Geometric Improvement*: Let h_{B_1} and h_{B_2} be two binning schemes such that the partition induced by h_{B_2} is a refinement of that of h_{B_1} (i.e., $\sigma(h_{B_1}(X)) \subset \sigma(h_{B_2}(X))$). Then for $f_{B_1,c}(x) = p_{h_{B_1}(x)}$ and $f_{B_2,c}(x) = p_{h_{B_2}(x)}$, the expected outer products satisfy:

$$\mathbb{E}_{\text{tr}}[f_{B_2,c} f_{B_2,c}^\top] \succeq \mathbb{E}_{\text{tr}}[f_{B_1,c} f_{B_1,c}^\top].$$

Hence, $\lambda_{f_{B_2,c}}^{\text{tr}} \geq \lambda_{f_{B_1,c}}^{\text{tr}}$. Moreover, under Assumptions 5.1, 5.6 and 5.7 and for $f_B(x) = \hat{p}_{h_B(x)}$, we have with probability at least $1 - \delta$,

$$\lambda_{f_B}^{\text{tr}} \geq \lambda_{f_{B,c}}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} > 0,$$

provided that $n_{\text{cal}} \geq \frac{128B}{(\lambda_{f_{0,c}}^{\text{tr}})^2} \log \left(\frac{2B}{\delta} \right)$.

(ii) **Statistical Cost**: Under Assumptions 5.6 and 5.7, with probability at least $1 - \delta$, the calibration error is bounded by $\mathcal{E}(f_B)^2 \leq \frac{8B}{n_{\text{cal}}} \log \left(\frac{2B}{\delta} \right)$. (Proof in Sections E.1 and E.2).

From Lemma 5.8 (i), it is evident that finer binning (i.e., larger B) improves the eigenvalue λ_f^{tr} but increases $\mathcal{E}(f_B)$

as well. In our framework, we have $B = Ks$. To determine the optimal s^* , we analyze the term $\mathcal{E}(f)^2/(\lambda_f^{\text{tr}})^2$ using the bounds in Lemma 5.8(i) and (ii).

An upper bound of $\mathcal{E}(f)^2/(\lambda_f^{\text{tr}})^2$ is $cJ(B)$, where c is irrelevant of B and

$$J(B) := \frac{B \log(2B/\delta)}{n_{\text{cal}} \left(\lambda_{f,c}^{\text{tr}} - \frac{\lambda_{f_0,c}^{\text{tr}}}{2} \right)^2} = \frac{Ks \log(2Ks/\delta)}{n_{\text{cal}} \left(\lambda_{f,c}^{\text{tr}} - \frac{\lambda_{f_0,c}^{\text{tr}}}{2} \right)^2}.$$

We aim to minimize $J(B)$ with respect to s .

Theorem 5.9 (Optimal Binning Strategy). *Under Assumptions 3.3, 3.4, 5.1, 5.3 and 5.7, the optimal s^* is given by*

$$s^* = \sqrt{5C_1''/12(\lambda_\infty^{\text{tr}} - \lambda_{f_0,c}^{\text{tr}}/2)},$$

where values of C_1'' and $\lambda_\infty^{\text{tr}}$ are presented in Section E.3. (Proof in Section E.3).

6 EXPERIMENTS

We empirically evaluate the proposed two-stage online procedure on two datasets under constant label shift.

6.1 SETUP

We use two datasets: (i) Toy-10, a synthetic 10-class classification task with gaussian features, with an MLP as predictor, (ii) CIFAR-10 [Krizhevsky et al., 2009] under a constant label shift where one majority class is over-represented following Wu et al. [2021], with ResNet-18 [He et al., 2016] as predictor. We compare our method with six baselines. NoAdapt is the fixed base classifier without online adaptation, ROGD [Wu et al., 2021], UOGD [Bai et al., 2022], ATLAS [Bai et al., 2022], FLH-FTL [Baby et al., 2024], LASU [Dai et al., 2025], and the Oracle sets the importance weights using the true target label distribution π_{te} . Learning rates follow the guided in each respective paper. For our method, we set $\gamma = 0.1$ and $s = 7$ sub-bins per class. In each round, Stage 1 sequentially updates the label-marginal estimate $\hat{\pi}$ from the batch samples, after which Stage 2 performs a single gradient step on the model parameters using the updated estimate. We report the average cross-entropy and the average squared ℓ_2 -norm error of the label-marginal estimate, $\frac{1}{T} \sum_{t=1}^T \|\hat{\pi}_t - \pi_{\text{te}}\|_2^2$. All experiments use $|\mathcal{B}| = 32$ samples per round and are repeated over 5 random seeds. Further details are given in Appendix F.

6.2 EXPERIMENTAL RESULTS

Table 1 reports the average cross-entropy and the squared ℓ_2 error of the label-marginal estimate on Toy-10 (mean $\pm$ std) and Table 2 reports the corresponding results on CIFAR-10.

Table 1: Toy-10: Average Cross-Entropy and label-marginal squared error with mean $\pm$ std across 5 seeds.

Method	Average Cross-Entropy	$\frac{1}{T} \sum \ \hat{\pi} - \pi_{\text{te}}\ _2^2$
NoAdapt	0.364 $\pm$ 0.001	0.042
ROGD	0.337 $\pm$ 0.001	0.872 $\pm$ 9.0e ⁻⁵
UOGD	0.319 $\pm$ 0.002	0.034 $\pm$ 1.358e ⁻³
ATLAS	0.325 $\pm$ 0.002	0.034 $\pm$ 1.358e ⁻³
FLH-FTL	0.359 $\pm$ 0.028	0.003 $\pm$ 3.05e ⁻⁴
LASU	0.340 $\pm$ 0.004	0.032 $\pm$ 1.196e ⁻³
Ours	0.319 $\pm$ 0.002	0.005 $\pm$ 3.57e ⁻⁴
Oracle	0.316 $\pm$ 0.001	—

Table 2: CIFAR-10: Average cross-entropy and label-marginal squared error with mean $\pm$ std over 5 seeds.

Method	Average Cross-Entropy	$\frac{1}{T} \sum \ \hat{\pi} - \pi_{\text{te}}\ _2^2$
NoAdapt	0.413 $\pm$ 0.000	0.222
ROGD	0.291 $\pm$ 0.001	0.766 $\pm$ 2.85e ⁻⁴
UOGD	0.260 $\pm$ 0.000	0.032 $\pm$ 1.311e ⁻³
ATLAS	0.260 $\pm$ 0.000	0.032 $\pm$ 1.311e ⁻³
FLH-FTL	0.277 $\pm$ 0.012	9.28e ⁻⁴ $\pm$ 8.9e ⁻⁵
LASU	0.309 $\pm$ 0.006	0.028 $\pm$ 1.062e ⁻³
Ours	0.259 $\pm$ 0.000	0.003 $\pm$ 1.98e ⁻⁴
Oracle	0.259 $\pm$ 0.000	—

Tables 1 and 2 show that our proposed method achieves the lowest cross-entropy on both datasets. UOGD and ATLAS share the same BBSE-based $\hat{\pi}$ estimation, yielding identical $\|\hat{\pi} - \pi_{\text{te}}\|_2^2$. ROGD exhibits a large $\|\hat{\pi} - \pi_{\text{te}}\|_2^2$ because its per-sample BBSE estimate at each time step leading to high average squared error despite reasonable cross-entropy. Theorem 5.9 predicts an optimal sub-bin count s^* that balances the geometric approximation error against the statistical estimation error. Figure 1 validates this prediction on CIFAR-10: cross-entropy decreases as s increases from $s=1$, reaching a minimum around $s=7$, then rises slightly for larger s as bin counts become sparse. Figure 2 shows the average cross-entropy on CIFAR-10 with different online batch sizes. Methods that rely on batch-level estimation of π_{te} typically suffer from unstable estimates at small batch sizes, degrading downstream performance. In contrast, our Soft-Bayes updates produce stable per-sample estimates regardless of batch size, so reducing the batch size simply increases the number of model updates, yielding lower cross-entropy.

7 CONCLUSION

We proposed an online procedure for label shift adaptation that combines histogram binning, online label distribution estimation via Soft-Bayes, and model parameter adaptation via online gradient descent. Our theoretical analysis decom-

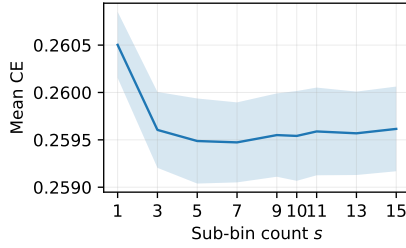


Figure 1: Effect of sub-bin count s on average cross-entropy (CIFAR-10).

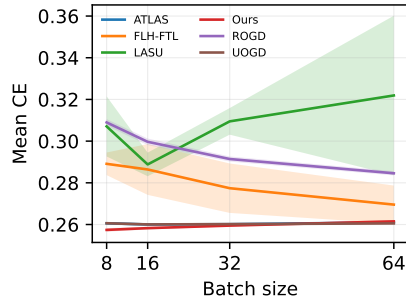


Figure 2: Average cross-entropy on CIFAR-10 as a function of online batch size.

poses the model-parameter estimation error into a vanishing term of order $\mathcal{O}(1/\sqrt{T})$ and a term governed by the ratio $\mathcal{E}(f)^2/(\lambda_f^{\text{tr}})^2$, and derives the optimal bin count B^* that minimizes this term. Experiments showed that the proposed method achieves cross-entropy that outperforms baselines.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2022-0-00469, Development of Core Technologies for Task-oriented Reinforcement Learning for Commercialization of Autonomous Drones; No. RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST)).

References

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 222–232. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/alexandari20a.html>.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animeshree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rJl0r3R9KX>.

Dheeraj Baby, Saurabh Garg, Tzu-Ching Yen, Sivaraman Balakrishnan, Zachary Lipton, and Yu-Xiang Wang. Online label shift: Optimal dynamic regret meets practical algorithms. *Advances in Neural Information Processing Systems*, 36, 2024.

Yong Bai, Yu-Jie Zhang, Peng Zhao, Masashi Sugiyama, and Zhi-Hua Zhou. Adapting to online label shift with provable guarantees. *Advances in Neural Information Processing Systems*, 35:29960–29974, 2022.

Yucong Dai, Shilin Gu, Ruidong Fan, Chao Xu, and Chenping Hou. Label shift meets online learning: Ensuring consistent adaptation with universal dynamic regret. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15392–15401, June 2025.

Marthinus Christoffel Du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

Ruidong Fan, Xiao Ouyang, Tingjin Luo, Lijun Zhang, and Chenping Hou. One-step label shift adaptation via robust weight estimation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI '25*, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/565. URL <https://doi.org/10.24963/ijcai.2025/565>.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3290–3300. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/219e052492f4008818b8adb6366c7ed6-Paper.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Jennifer Dy and Andreas Krause, editors,

- Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3122–3130. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/lipton18a.html>.
- Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2005.
- Andrew C Miller and Joseph Futoma. Label shift estimators for non-ignorable missing data. *arXiv preprint arXiv:2310.18261*, 2023.
- Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7195–7201, 2016. doi: 10.1109/CDC.2016.7799379.
- Francesco Orabona. Online learning: A modern introduction using convex optimization. *arXiv:1912.13213v6 [cs.LG]*, 2023. URL <https://arxiv.org/abs/1912.13213v6>.
- Laurent Orseau, Tor Lattimore, and Shane Legg. Soft-bayes: Prod for mixtures of experts with log-loss. In Steve Hanneke and Lev Reyzin, editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 372–399. PMLR, 15–17 Oct 2017. URL <https://proceedings.mlr.press/v76/orseau17a.html>.
- Amelie Royer and Christoph H Lampert. Classifier adaptation at prediction time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1401–1409, 2015.
- Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 01 2002. ISSN 0899-7667. doi: 10.1162/089976602753284446. URL <https://doi.org/10.1162/089976602753284446>.
- Milan Sulc and Jiri Matas. Improving cnn classifiers by estimating test-time priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
- Qinglong Tian, Xin Zhang, and Jiwei Zhao. Elsa: Efficient label shift adaptation through the lens of semiparametric models. In *International Conference on Machine Learning*, pages 34120–34142. PMLR, 2023.
- Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015. doi: 10.1561/22000000048. URL <https://arxiv.org/abs/1501.01571>.
- Tomáš Šipka, Milan Šulc, and Jiří Matas. The hitchhiker’s guide to prior-shift adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1524, January 2022.
- Hongwei Wen, Annika Betken, and Hanyuan Hang. Class probability matching with calibrated networks for label shift adaption. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Representation Learning*, volume 2024, pages 29337–29367, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/7e3767db483c942b883eb4f8cfb74e31-Paper-Conference.pdf.
- Ruihan Wu, Chuan Guo, Yi Su, and Kilian Q Weinberger. Online adaptation to label distribution shift. *Advances in Neural Information Processing Systems*, 34:11340–11351, 2021.
- Changkun Ye, Russell Tsuchida, Lars Petersson, and Nick Barnes. Label shift estimation for class-imbalance problem: A bayesian approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1073–1082, January 2024.
- Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. <https://D2L.ai>.

Calibration-Aware Online Adaptation under Label Shift (Supplementary Material)

Jiun Jeong¹

Byeongwoo An²

Gi-Soo Kim^{1,2,†}

Kyubo Shin^{3,†}

¹Graduate School of Artificial Intelligence, UNIST, Ulsan, Republic of Korea

²Department of Industrial Engineering, UNIST, Ulsan, Republic of Korea

³R&D Division, THYROSCOPE INC., Ulsan, Republic of Korea

A EQUIVALENCE OF (4) AND (5) WITH $z = \hat{y}(x)$

The objective in (5) can be written as

$$\sum_{k=1}^K P_{\text{te}}(\hat{y}(x) = k) \log \left(\sum_{k'=1}^K P_{\text{tr}}(y = k' \mid \hat{y}(x) = k) w_{k'} \right).$$

By definition of C , we have $P_{\text{tr}}(y = k' \mid \hat{y}(x) = k) = C_{kk'}/P_{\text{tr}}(\hat{y}(x) = k)$, and therefore

$$\sum_{k=1}^K P_{\text{te}}(\hat{y}(x) = k) \log \left(\sum_{k'=1}^K P_{\text{tr}}(y = k' \mid \hat{y}(x) = k) w_{k'} \right) = \sum_{k=1}^K P_{\text{te}}(\hat{y}(x) = k) \log ((Cw)_k) - \sum_{k=1}^K P_{\text{te}}(\hat{y}(x) = k) \log P_{\text{tr}}(\hat{y}(x) = k),$$

and the second term does not depend on w . Moreover, since $w \in \mathcal{W}$,

$$\sum_{k=1}^K (Cw)_k = \sum_{k'=1}^K \left(\sum_{k=1}^K C_{kk'} \right) w_{k'} = \sum_{k'=1}^K P_{\text{tr}}(y = k') w_{k'} = 1,$$

and $(Cw)_k \geq 0$ for all k , so Cw lies in the probability simplex. Hence,

$$\sum_{k=1}^K P_{\text{te}}(\hat{y}(x) = k) \log ((Cw)_k) = \sum_{k=1}^K P_{\text{te}}(\hat{y}(x) = k) \log P_{\text{te}}(\hat{y}(x) = k) - \text{KL}(\mathbb{E}_{te}[e_{\hat{y}(x)}] \parallel Cw),$$

which is maximized if and only if $\text{KL}(\mathbb{E}_{te}[e_{\hat{y}(x)}] \parallel Cw) = 0$, equivalently $Cw = \mathbb{E}_{te}[e_{\hat{y}(x)}]$.

B PROOFS FOR SECTION 5.1: ONLINE ESTIMATION OF π

This section provides detailed derivations and proofs for the theoretical results presented in Section 5.1.

B.1 PROPOSITION B.1: STRONG CONVEXITY OF $L(\pi)$

Proposition B.1 (Strong Convexity of $L(\pi)$). *Under Assumptions 3.3, 3.4, 5.1, 5.6 and 5.7, $L(\pi)$ is μ_f -strongly convex with probability at least $1 - \delta$, where*

$$\mu_f := \frac{\pi_{\text{tr},\min}^2}{\pi_{\text{tr},\max}^2} \lambda_f^{\text{te}} > 0.$$

Proof. Under Assumptions 5.1, 5.6 and 5.7, Lemma 5.8 states that $\lambda_f^{\text{tr}} > 0$ with probability at least $1 - \delta$. Under Assumption 3.3,

$$\mathbb{E}_{\text{te}}[f(x^*)f(x^*)^\top] = \sum_{k=1}^K w_k^* \pi_{\text{tr},k} \mathbb{E}_{\text{tr}}[f(x)f(x)^\top \mid y = k] \succeq \frac{\pi_{\text{te},\min}}{\pi_{\text{tr},\max}} \mathbb{E}_{\text{tr}}[f(x)f(x)^\top].$$

Therefore,

$$\lambda_f^{\text{te}} \geq \frac{\pi_{\text{te},\min}}{\pi_{\text{tr},\max}} \lambda_f^{\text{tr}} > 0. \quad (\text{B.1.1})$$

We now show that $L(\pi)$ is μ_f -strongly convex, where μ_f is defined in the Proposition statement. The Hessian of $L(\pi)$ is

$$\nabla^2 L(\pi) = \mathbb{E}_{\text{te}} \left[\frac{q(x^*)q(x^*)^\top}{(q(x^*)^\top \pi)^2} \right].$$

Since $q(x) = f(x) \odot \pi_{\text{tr}}$ and $\pi_{\text{tr},\min} > 0$ under Assumption 3.4,

$$q(x^*)^\top \pi = \sum_{k=1}^K \frac{f_k(x^*)}{\pi_{\text{tr},k}} \pi_k \leq \frac{1}{\pi_{\text{tr},\min}} \sum_{k=1}^K f_k(x^*) \pi_k \leq \frac{1}{\pi_{\text{tr},\min}}.$$

Therefore,

$$\nabla^2 L(\pi) \succeq \pi_{\text{tr},\min}^2 \mathbb{E}_{\text{te}}[q(x^*)q(x^*)^\top].$$

Using again that $q(x) = f(x) \odot \pi_{\text{tr}}$,

$$\mathbb{E}_{\text{te}}[q(x^*)q(x^*)^\top] = \text{diag}(\pi_{\text{tr}})^{-1} \mathbb{E}_{\text{te}}[f(x^*)f(x^*)^\top] \text{diag}(\pi_{\text{tr}})^{-1}.$$

Since $\lambda_{\min}(\text{diag}(\pi_{\text{tr}})^{-1}) = 1/\pi_{\text{tr},\max}$, we have

$$\lambda_{\min}(\mathbb{E}_{\text{te}}[q(x^*)q(x^*)^\top]) \geq \frac{\lambda_f^{\text{te}}}{\pi_{\text{tr},\max}^2}.$$

Hence,

$$\nabla^2 L(\pi) \succeq \frac{\pi_{\text{tr},\min}^2}{\pi_{\text{tr},\max}^2} \lambda_f^{\text{te}} I.$$

Therefore, $L(\pi)$ is μ_f -strongly convex with

$$\mu_f := \frac{\pi_{\text{tr},\min}^2}{\pi_{\text{tr},\max}^2} \lambda_f^{\text{te}} > 0,$$

where the positivity follows from (B.1.1) and Assumption 3.4. $\square$

B.2 THEOREM B.2: CONVERGENCE OF $\hat{\pi}^{(t)}$ TO π_f

Theorem B.2 (Convergence of $\hat{\pi}^{(t)}$ to π_f). *Under Assumptions 3.3, 3.4, 5.1, 5.6 and 5.7, for a sequence $\hat{\pi}^{(1)}, \dots, \hat{\pi}^{(T)}$ generated by the Soft-Bayes algorithm with regret $\tilde{R}(T)$, the following holds with probability at least $1 - \delta$:*

$$\frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_f \right\|_2^2 \leq \frac{2}{\mu_f} \left(\frac{\tilde{R}(T)}{T} + 2 \log \left(\frac{2T}{\pi_{\text{tr},\min}^2} \right) \sqrt{\frac{2 \log(2/\delta)}{T}} \right).$$

Proof. The proof employs the strong convexity of the surrogate population risk $L(\pi)$ with the online-to-batch conversion theorem.

Step 1: Online-to-Batch Conversion. We first present the online-to-batch conversion method of Orabona [2023].

Theorem B.3 (Orabona [2023], Theorem 3.14). Define the population risk of $\pi \in \Pi_c$ as

$$L(\pi) = \mathbb{E}_{\text{te}}[\tilde{\ell}_s(x^*; \pi)],$$

where the expectation is taken with respect to x^* drawn from distribution P_{te} .

Suppose we have an i.i.d. sequence $x_1^*, \dots, x_T^*$ drawn from P_{te} and an online learning algorithm that outputs $\hat{\pi}^{(1)}, \dots, \hat{\pi}^{(T)}$ satisfying the following regret bound:

$$\tilde{R}(T) = \sum_{t=1}^T \left\{ \tilde{\ell}_s(x_t^*; \hat{\pi}^{(t)}) - \tilde{\ell}_s(x_t^*; \pi) \right\} < \infty$$

for some $\pi \in \Pi_c$.

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

$$\frac{1}{T} \sum_{t=1}^T L(\hat{\pi}^{(t)}) \leq L(\pi) + \frac{\tilde{R}(T)}{T} + 2B_{\tilde{\ell}} \sqrt{\frac{2 \log \frac{2}{\delta}}{T}},$$

where $B_{\tilde{\ell}}$ is a constant such that $|\tilde{\ell}_s(x_t^*; \hat{\pi}^{(t)}) - L(\hat{\pi}^{(t)})| \leq B_{\tilde{\ell}}$ for all $t \in [T]$ and $|\tilde{\ell}_s(x_t^*; \pi) - L(\pi)| \leq B_{\tilde{\ell}}$ for all $\pi \in \Pi_c$.

In our setting, we can easily derive $|\tilde{\ell}_s(x_t^*; \hat{\pi}^{(t)}) - L(\hat{\pi}^{(t)})| \leq -\log(\pi_{\text{tr}, \min} \hat{\pi}_{\min}^{(t)})$. Due to the update mechanism of Soft-Bayes, we have

$$\hat{\pi}_{\min}^{(t)} \geq (1 - \alpha_{t-1}) \pi_{\text{tr}, \min} = \left(1 - \frac{\eta_t}{\eta_{t-1}}\right) \pi_{\text{tr}, \min} = \left(1 - \sqrt{\frac{t-1}{t}}\right) \pi_{\text{tr}, \min} \geq \frac{1}{2t} \pi_{\text{tr}, \min}$$

for all $t \geq 1$. Hence, $|\tilde{\ell}_s(x_t^*; \hat{\pi}^{(t)}) - L(\hat{\pi}^{(t)})| \leq \log\left(\frac{2t}{\pi_{\text{tr}, \min}^2}\right) \leq \log\left(\frac{2T}{\pi_{\text{tr}, \min}^2}\right) = B_{\tilde{\ell}}$. For the same value of $B_{\tilde{\ell}}$, we also have $|\tilde{\ell}_s(x_t^*; \pi) - L(\pi)| \leq B_{\tilde{\ell}}$ for all $\pi \in \Pi_c$ due to constraint of Π_c .

Applying Theorem B.3, with probability at least $1 - \delta$, we have:

$$\frac{1}{T} \sum_{t=1}^T L(\hat{\pi}^{(t)}) - L(\pi_f) \leq \frac{\tilde{R}(T)}{T} + 2 \log\left(\frac{2T}{\pi_{\text{tr}, \min}^2}\right) \sqrt{\frac{2 \log(2/\delta)}{T}}. \quad (\text{B.2.1})$$

Step 2: Utilizing Strong Convexity. By Proposition B.1, under 3.3, Assumptions 3.4 and 5.1, the population risk $L(\pi)$ is μ_f -strongly convex with $\mu_f > 0$. The definition of strong convexity implies that for any $\pi \in \Pi_c$:

$$L(\pi) \geq L(\pi_f) + \nabla L(\pi_f)^\top (\pi - \pi_f) + \frac{\mu_f}{2} \|\pi - \pi_f\|_2^2.$$

Since π_f is the minimizer of $L(\pi)$ over the convex set Π_c , the first-order optimality condition states that $\nabla L(\pi_f)^\top (\pi - \pi_f) \geq 0$ for all $\pi \in \Pi_c$. Therefore, we obtain a lower bound on the excess risk for all $\pi \in \Pi_c$:

$$L(\pi) - L(\pi_f) \geq \frac{\mu_f}{2} \|\pi - \pi_f\|_2^2.$$

Step 3: Combining the Bounds.

Note that $\hat{\pi}^{(t)} \in \Pi_c$ for all $t \in [T]$. Apply Step 2's inequality to each $\hat{\pi}^{(t)}$ and average over $t = 1$ to T :

$$\frac{1}{T} \sum_{t=1}^T \left(L(\hat{\pi}^{(t)}) - L(\pi_f) \right) \geq \frac{\mu_f}{2} \cdot \frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_f \right\|_2^2.$$

By online-to-batch conversion (B.2.1):

$$\frac{\mu_f}{2} \cdot \frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_f \right\|_2^2 \leq \frac{\tilde{R}(T)}{T} + 2 \log \left(\frac{2T}{\pi_{\text{tr},\min}^2} \right) \sqrt{\frac{2 \log(2/\delta)}{T}}.$$

Rearranging the terms yields the final result:

$$\frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_f \right\|_2^2 \leq \frac{2}{\mu_f} \left(\frac{\tilde{R}(T)}{T} + 2 \log \left(\frac{2T}{\pi_{\text{tr},\min}^2} \right) \sqrt{\frac{2 \log(2/\delta)}{T}} \right).$$

□

B.3 LEMMA B.4: BOUND ON CALIBRATION GAP FOR π

Lemma B.4 (Bound on Calibration Gap for π). *Under conditions of Proposition B.1, the distance between π_f and π_{te} is bounded by:*

$$\|\pi_f - \pi_{\text{te}}\|_2 \leq \frac{C'}{\mu_f} \mathcal{E}(f),$$

where

$$C' = \left(\frac{\pi_{\text{tr},\max}}{\pi_{\text{te},\min}} + \frac{1}{\pi_{\text{tr},\min}} \frac{\pi_{\text{tr},\max}^2}{\pi_{\text{te},\min}^2} \right) \frac{1}{\pi_{\text{tr},\min}} \frac{\pi_{\text{te},\max}}{\pi_{\text{tr},\min}}.$$

Proof. Step 1: Bounding Distance to Gradient Norm via Strong Convexity. By Proposition B.1, $L(\pi)$ is μ_f -strongly convex. This implies the following inequality holds:

$$\langle \nabla L(\pi_f) - \nabla L(\pi_{\text{te}}), \pi_f - \pi_{\text{te}} \rangle \geq \mu \|\pi_f - \pi_{\text{te}}\|_2^2.$$

Since π_f is the minimizer of $L(\pi)$ in Π_c and since $\pi_{\text{te}} \in \Pi_c$, the first-order optimality condition implies $\langle \nabla L(\pi_f), \pi_{\text{te}} - \pi_f \rangle \geq 0$. Rearranging the strong convexity inequality gives:

$$\begin{aligned} \mu_f \|\pi_f - \pi_{\text{te}}\|_2^2 &\leq \langle \nabla L(\pi_{\text{te}}) - \nabla L(\pi_f), \pi_{\text{te}} - \pi_f \rangle \\ &= \langle \nabla L(\pi_{\text{te}}), \pi_{\text{te}} - \pi_f \rangle - \underbrace{\langle \nabla L(\pi_f), \pi_{\text{te}} - \pi_f \rangle}_{\geq 0} \\ &\leq \langle \nabla L(\pi_{\text{te}}), \pi_{\text{te}} - \pi_f \rangle. \end{aligned}$$

Applying Cauchy-Schwarz inequality to the right-hand side:

$$\mu_f \|\pi_f - \pi_{\text{te}}\|_2^2 \leq \|\nabla L(\pi_{\text{te}})\|_2 \|\pi_{\text{te}} - \pi_f\|_2.$$

Dividing by $\|\pi_f - \pi_{\text{te}}\|_2$ when $\pi_f \neq \pi_{\text{te}}$:

$$\|\pi_f - \pi_{\text{te}}\|_2 \leq \frac{1}{\mu_f} \|\nabla L(\pi_{\text{te}})\|_2. \quad (\text{B.3.1})$$

Step 2: Bounding the Gradient Norm.

Let $f_c(x) = P_{\text{tr}}(\cdot \mid h_B(x))$ denote the canonically calibrated classifier, and define $q_c(x) = f_c(x) \odot \pi_{\text{tr}}$ and $L_c(\pi) = \mathbb{E}_{\text{te}}[-\log(q_c(x^*)^\top \pi)]$. Since $\pi_{\text{te}} \in \Pi_c$ minimizes $L_c(\pi)$ over Δ^{K-1} , $\nabla L_c(\pi_{\text{te}}) = 0$.

Rewrite the gradient norm as:

$$\|\nabla L(\pi_{\text{te}})\|_2 = \|\nabla L(\pi_{\text{te}}) - \nabla L_c(\pi_{\text{te}})\|_2.$$

Substituting the gradient expressions (derived in Appendix B.1):

$$\|\nabla L(\pi_{\text{te}})\|_2 = \left\| \mathbb{E}_{\text{te}} \left[\frac{q_c(x^*)}{q_c(x^*)^\top \pi_{\text{te}}} - \frac{q(x^*)}{q(x^*)^\top \pi_{\text{te}}} \right] \right\|_2.$$

Applying Jensen's inequality ($\|\mathbb{E}[Z]\|_2 \leq \mathbb{E}[\|Z\|_2]$):

$$\|\nabla L(\pi_{\text{te}})\|_2 \leq \mathbb{E}_{\text{te}} \left[\left\| \frac{q_c(x^*)}{q_c(x^*)^\top \pi_{\text{te}}} - \frac{q(x^*)}{q(x^*)^\top \pi_{\text{te}}} \right\|_2 \right]. \quad (\text{B.3.2})$$

Step 3: Bounding the Gradient Difference. Due to Assumption 3.4, we have for all x^* :

$$q(x^*)^\top \pi_{\text{te}} \geq \frac{\pi_{\text{te},\min}}{\pi_{\text{tr},\max}} \quad \text{and} \quad q_c(x^*)^\top \pi_{\text{te}} \geq \frac{\pi_{\text{te},\min}}{\pi_{\text{tr},\max}}.$$

Furthermore, $\|q_c(x^*)\|_2 \leq \|D_{\text{tr}}^{-1}\|_2 \|f_c(x^*)\|_2 \leq 1/\pi_{\text{tr},\min}$.

Now we can bound the difference term (omitting x^* for brevity):

$$\begin{aligned} \frac{q_c}{q_c^\top \pi_{\text{te}}} - \frac{q}{q^\top \pi_{\text{te}}} &= \left(\frac{q_c}{q^\top \pi_{\text{te}}} - \frac{q}{q^\top \pi_{\text{te}}} \right) + \left(\frac{q_c}{q_c^\top \pi_{\text{te}}} - \frac{q}{q^\top \pi_{\text{te}}} \right) \\ &= \frac{q_c - q}{q^\top \pi_{\text{te}}} + q_c \left(\frac{q^\top \pi_{\text{te}} - q_c^\top \pi_{\text{te}}}{(q_c^\top \pi_{\text{te}})(q^\top \pi_{\text{te}})} \right). \end{aligned}$$

By triangle inequality for the L2-norm:

$$\left\| \frac{q_c}{q_c^\top \pi_{\text{te}}} - \frac{q}{q^\top \pi_{\text{te}}} \right\|_2 \leq \left\| \frac{q_c - q}{q^\top \pi_{\text{te}}} \right\|_2 + \left\| q_c \frac{(q - q_c)^\top \pi_{\text{te}}}{(q_c^\top \pi_{\text{te}})(q^\top \pi_{\text{te}})} \right\|_2.$$

For the first term, using the lower bound $q^\top \pi_{\text{te}} \geq \frac{\pi_{\text{te},\min}}{\pi_{\text{tr},\max}}$:

$$\left\| \frac{q_c - q}{q^\top \pi_{\text{te}}} \right\|_2 \leq \frac{\pi_{\text{tr},\max}}{\pi_{\text{te},\min}} \|q_c - q\|_2.$$

For the second term:

$$\left\| q_c \frac{(q - q_c)^\top \pi_{\text{te}}}{(q_c^\top \pi_{\text{te}})(q^\top \pi_{\text{te}})} \right\|_2 = \|q_c\|_2 \frac{|(q - q_c)^\top \pi_{\text{te}}|}{|(q_c^\top \pi_{\text{te}})(q^\top \pi_{\text{te}})|}.$$

By Cauchy-Schwarz and $\|\pi_{\text{te}}\|_2 \leq 1$, the numerator is $|(q - q_c)^\top \pi_{\text{te}}| \leq \|q - q_c\|_2$. Thus, the second term is bounded by:

$$\frac{1}{\pi_{\text{tr},\min}} \frac{\pi_{\text{tr},\max}^2}{\pi_{\text{te},\min}^2} \|q_c - q\|_2.$$

Combining the bounds for both terms and define the Lipschitz constant C_L :

$$C_L := \frac{\pi_{\text{tr},\max}}{\pi_{\text{te},\min}} + \frac{1}{\pi_{\text{tr},\min}} \frac{\pi_{\text{tr},\max}^2}{\pi_{\text{te},\min}^2}.$$

Then,

$$\left\| \frac{q_c(x^*)}{q_c(x^*)^\top \pi_{\text{te}}} - \frac{q(x^*)}{q(x^*)^\top \pi_{\text{te}}} \right\|_2 \leq C_L \|q_c(x^*) - q(x^*)\|_2.$$

Step 4: Linking to Calibration Error $\mathcal{E}(f)$. We have

$$\|\nabla L(\pi_{\text{te}})\|_2 \leq C_L \mathbb{E}_{\text{te}}[\|q_c(x^*) - q(x^*)\|_2].$$

Convert the difference in q back to the difference in f :

$$\|q_c(x^*) - q(x^*)\|_2 \leq \|D_{\text{tr}}^{-1}\|_2 \|f_c(x^*) - f(x^*)\|_2 = \frac{1}{\pi_{\text{tr},\min}} \|f_c(x^*) - f(x^*)\|_2.$$

Thus,

$$\|\nabla L(\pi_{\text{te}})\|_2 \leq C_L \frac{1}{\pi_{\text{tr},\min}} \mathbb{E}_{\text{te}}[\|f_c(x^*) - f(x^*)\|_2].$$

By Assumptions 3.3 and 3.4,

$$\begin{aligned}
\mathbb{E}_{\text{te}}[\|f_c(x^*) - f(x^*)\|_2] &= \sum_{k=1}^K P_{\text{te}}(y = k) \mathbb{E}_{\text{te}}[\|f_c(x) - f(x)\|_2 \mid y = k] \\
&= \sum_{k=1}^K w_k^* P_{\text{tr}}(y = k) \mathbb{E}_{\text{tr}}[\|f_c(x) - f(x)\|_2 \mid y = k] \quad (\text{By label shift assumption}) \\
&\leq \frac{\pi_{\text{te}, \max}}{\pi_{\text{tr}, \min}} \sum_{k=1}^K P_{\text{tr}}(y = k) \mathbb{E}_{\text{tr}}[\|f_c(x) - f(x)\|_2 \mid y = k] \\
&= \frac{\pi_{\text{te}, \max}}{\pi_{\text{tr}, \min}} \mathbb{E}_{\text{tr}}[\|f_c(x) - f(x)\|_2].
\end{aligned}$$

Jensen's inequality gives:

$$\mathbb{E}_{\text{tr}}[\|f_c(x) - f(x)\|_2] \leq \left(\mathbb{E}_{\text{tr}}[\|f_c(x) - f(x)\|_2^2] \right)^{1/2} = \mathcal{E}(f).$$

Step 5: Final Bound. Combining the inequalities:

$$\|\nabla L(\pi_{\text{te}})\|_2 \leq C_L \frac{1}{\pi_{\text{tr}, \min}} \frac{\pi_{\text{te}, \max}}{\pi_{\text{tr}, \min}} \mathcal{E}(f).$$

Substituting this back into (B.3.1):

$$\|\pi_f - \pi_{\text{te}}\|_2 \leq \frac{C_L \frac{1}{\pi_{\text{tr}, \min}} \frac{\pi_{\text{te}, \max}}{\pi_{\text{tr}, \min}}}{\mu_f} \mathcal{E}(f).$$

Defining the constant $C' = C_L \frac{1}{\pi_{\text{tr}, \min}} \frac{\pi_{\text{te}, \max}}{\pi_{\text{tr}, \min}}$ completes the proof. □

B.4 THEOREM 5.2: ONLINE ERROR BOUND FOR LABEL ESTIMATION

Theorem 5.2 (Online Error Bound for Label Estimation). Under the conditions of Theorem B.2 and Lemma B.4, the total average squared error with respect to the true label distribution π_{te} is bounded with probability at least $1 - \delta$ by:

$$\frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_{\text{te}} \right\|_2^2 \leq \frac{4}{\mu_f} \left(\frac{\tilde{R}(T)}{T} + 2 \log \left(\frac{2T}{\pi_{\text{tr}, \min}^2} \right) \sqrt{\frac{2 \log(2/\delta)}{T}} \right) + \frac{2(C')^2}{\mu_f^2} \mathcal{E}(f)^2.$$

Proof. The proof combines the bounds on the estimation error in Theorem B.2 and the calibration gap in Lemma B.4.

Step 1: Error Decomposition. Total average squared error can be decompose the error at time t around surrogate optimum π_f :

$$\hat{\pi}^{(t)} - \pi_{\text{te}} = \left(\hat{\pi}^{(t)} - \pi_f \right) + \left(\pi_f - \pi_{\text{te}} \right).$$

For any vectors a, b , the inequality $\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$ holds.

$$\left\| \hat{\pi}^{(t)} - \pi_{\text{te}} \right\|_2^2 \leq 2 \left\| \hat{\pi}^{(t)} - \pi_f \right\|_2^2 + 2 \left\| \pi_f - \pi_{\text{te}} \right\|_2^2.$$

Step 2: Averaging and Substitution. Averaging over $t = 1, \dots, T$, we obtain

$$\frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_{\text{te}} \right\|_2^2 \leq \underbrace{2 \left(\frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_f \right\|_2^2 \right)}_{\text{Estimation error}} + 2 \underbrace{\left\| \pi_f - \pi_{\text{te}} \right\|_2^2}_{\text{Calibration gap}}.$$

By Theorem B.2, the estimation error is bounded with probability at least $1 - \delta$ as:

$$\frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_f \right\|_2^2 \leq \frac{2}{\mu_f} \left(\frac{\tilde{R}(T)}{T} + 2 \log \left(\frac{2T}{\pi_{\text{tr},\min}^2} \right) \sqrt{\frac{2 \log(2/\delta)}{T}} \right).$$

By Lemma B.4, the calibration gap is bounded by

$$\left\| \pi_f - \pi_{\text{te}} \right\|_2^2 \leq \left(\frac{C'}{\mu_f} \mathcal{E}(f) \right)^2 = \frac{(C')^2}{\mu_f^2} \mathcal{E}(f)^2.$$

Substituting these bounds, with probability at least $1 - \delta$, we obtain the result

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_{\text{te}} \right\|_2^2 &\leq 2 \left[\frac{2}{\mu_f} \left(\frac{\tilde{R}(T)}{T} + 2 \log \left(\frac{2T}{\pi_{\text{tr},\min}^2} \right) \sqrt{\frac{2 \log(2/\delta)}{T}} \right) \right] + 2 \left[\frac{(C')^2}{\mu_f^2} \mathcal{E}(f)^2 \right] \\ &= \frac{4}{\mu_f} \left(\frac{\tilde{R}(T)}{T} + 2 \log \left(\frac{2T}{\pi_{\text{tr},\min}^2} \right) \sqrt{\frac{2 \log(2/\delta)}{T}} \right) + \frac{2(C')^2}{\mu_f^2} \mathcal{E}(f)^2. \end{aligned}$$

□

Since $\mu_f := \frac{\pi_{\text{tr},\min}^2}{\pi_{\text{tr},\max}^2} \lambda_f^{\text{te}}$ (see Proposition B.1), substituting gives

$$\frac{1}{T} \sum_{t=1}^T \left\| \hat{\pi}^{(t)} - \pi_{\text{te}} \right\|_2^2 \leq \frac{4\pi_{\text{tr},\max}^2}{\pi_{\text{tr},\min}^2 \lambda_f^{\text{te}}} \left(\frac{\tilde{R}(T)}{T} + 2 \log \left(\frac{2T}{\pi_{\text{tr},\min}^2} \right) \sqrt{\frac{2 \log(2/\delta)}{T}} \right) + \frac{2(C')^2 \pi_{\text{tr},\max}^4}{\pi_{\text{tr},\min}^4 (\lambda_f^{\text{te}})^2} \mathcal{E}(f)^2.$$

C PROOFS FOR SECTION 5.2: MODEL-SIDE DYNAMIC OCO

This section provides detailed derivations and proofs for the theoretical results presented in Section 5.2. The analysis relies on the geometric properties of the empirical objective $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$.

C.1 DETAILS FOR MULTINOMIAL LOGISTIC MODEL

We instantiate the adaptive predictor $p_\theta(y | x)$ as a Multinomial Logistic Regression model. In this section, we detail the model setup, derive the gradient and Hessian of the weighted cross-entropy loss, and prove that the gradient is bounded. These results are later used to establish geometric properties of $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$ (Lemma C.1).

C.1.1 Model Setup and Notation

We consider a K -class classification problem and set the K -th class as the baseline category. Let $K' = K - 1$. The model parameters are

$$\theta = \left[(\theta^1)^\top, \dots, (\theta^{K'})^\top \right]^\top \in \mathbb{R}^{dK'}, \quad \theta^k \in \mathbb{R}^d.$$

The conditional probability $p_\theta(k | x)$ is

$$p_\theta(k | x) = \begin{cases} \frac{\exp((\theta^k)^\top x)}{1 + \sum_{j=1}^{K'} \exp((\theta^j)^\top x)} & \text{for } 1 \leq k \leq K', \\ \frac{1}{1 + \sum_{j=1}^{K'} \exp((\theta^j)^\top x)} & \text{for } k = K. \end{cases}$$

The objective function is the expected weighted cross-entropy loss

$$\mathcal{L}_{\text{tr}}(\theta; w) = \mathbb{E}_{(x,y) \sim P_{\text{tr}}} [-w_y \log p_\theta(y | x)].$$

For a single data point (x, y) , denote $\ell(\theta; x, y) = -w_y \log p_\theta(y | x)$.

We introduce the following compact notation:

- $s_k(x) = p_\theta(k|x)$.
- $\mathbf{s}(x) = [s_1(x), \dots, s_{K'}(x)]^\top \in \mathbb{R}^{K'}$.
- $S(x) = \text{diag}(\mathbf{s}(x)) \in \mathbb{R}^{K' \times K'}$.
- $\Phi(x) = I_{K'} \otimes x \in \mathbb{R}^{d_{K'} \times K'}$ (Kronecker product).
- $e_y \in \mathbb{R}^{K'}$ is the standard basis vector for class y if $1 \leq y \leq K'$, and the zero vector if $y = K$.

C.1.2 Gradient Calculation

Gradient of ℓ with respect to θ^j , $j \in [K']$, is:

$$\frac{\partial \ell}{\partial \theta^j} = -w_y \frac{1}{s_y(x)} \frac{\partial s_y(x)}{\partial \theta^j}.$$

For the softmax function, the derivative with respect to θ^j is given by

$$\frac{\partial s_k(x)}{\partial \theta^j} = s_k(x) (\mathbf{1}\{k = j\} - s_j(x)) x$$

Substituting this into the gradient of ℓ with $k = y$ yields

$$\begin{aligned} \frac{\partial \ell}{\partial \theta^j} &= -w_y \frac{1}{s_y(x)} [s_y(x) (\mathbf{1}\{y = j\} - s_j(x)) x] \\ &= w_y (s_j(x) - \mathbf{1}\{y = j\}) x \end{aligned}$$

Stacking the blocks $\partial \ell / \partial \theta^j$ for $j = 1, \dots, K'$, the full gradient $\nabla_\theta \ell \in \mathbb{R}^{d_{K'}}$ can be written compactly as

$$\nabla_\theta \ell = w_y \Phi(x) (\mathbf{s}(x) - e_y). \quad (\text{C.1.1})$$

C.1.3 Hessian Calculation

Now compute the Hessian matrix $\nabla_\theta^2 \ell$. Differentiating (C.1.1) with respect to θ' , we obtain

$$\nabla_\theta^2 \ell = \frac{\partial}{\partial \theta^\top} (\nabla_\theta \ell) = w_y \Phi(x) \frac{\partial \mathbf{s}(x)}{\partial \theta^\top} = w_y \Phi(x) (S(x) - \mathbf{s}(x) \mathbf{s}(x)^\top) \Phi(x)^\top.$$

Taking expectation over $P_{\text{tr}}(x, y)$, the Hessian of the population risk $\mathcal{L}_{\text{tr}}(\theta; w)$ is

$$\begin{aligned} \nabla_\theta^2 \mathcal{L}_{\text{tr}}(\theta; w) &= \mathbb{E}_{(x, y) \sim P_{\text{tr}}} [\nabla_\theta^2 \ell] \\ &= \mathbb{E}_{x \sim P_{\text{tr}}} [\mathbb{E}_{y|x} [w_y] \Phi(x) (S(x) - \mathbf{s}(x) \mathbf{s}(x)^\top) \Phi(x)^\top] \end{aligned}$$

C.1.4 Verification of Bounded Gradient

There exists a constant $G > 0$ such that $\|\nabla_\theta \log p_\theta(y | x)\|_2 \leq G$ for all x, y, θ .

We verify that this assumption holds under Assumption 5.3 (Bounded Features, $\|x\|_2 \leq L_x$).

From (C.1.1) with $w_y = 1$, we have

$$\ell(\theta; x, y) = -\log p_\theta(y | x), \quad \nabla_\theta \ell = \Phi(x) (\mathbf{s}(x) - e_y).$$

Therefore,

$$\nabla_\theta \log p_\theta(y | x) = -\nabla_\theta \ell = -\Phi(x) (\mathbf{s}(x) - e_y)$$

and hence

$$\|\nabla_\theta \log p_\theta(y | x)\|_2 = \|\Phi(x) (\mathbf{s}(x) - e_y)\|_2.$$

We bound this norm as follows. First,

$$\|\Phi(x)(\mathbf{s}(x) - e_y)\|_2 \leq \|\Phi(x)\|_2 \|\mathbf{s}(x) - e_y\|_2.$$

Since $\Phi(x) = I_{K'} \otimes x$, for any $v \in \mathbb{R}^{K'}$,

$$\|\Phi(x)v\|_2^2 = \sum_{k=1}^{K'} v_k^2 \|x\|_2^2 = \|x\|_2^2 \|v\|_2^2$$

so the spectral norm of $\Phi(x)$ is

$$\|\Phi(x)\|_2 = \|x\|_2.$$

Next, $\mathbf{s}(x)$ is a sub-probability vector in $[0, 1]^{K'}$ with $\sum_{k=1}^{K'} s_k(x) \leq 1$, and e_y is either a standard basis vector (for $y \leq K'$) or the zero vector (for $y = K$). The maximum Euclidean distance between any two such points is at most $\sqrt{2}$, so

$$\|\mathbf{s}(x) - e_y\|_2 \leq \sqrt{2}.$$

Combining these bounds and using Assumption 5.3, $\|x\|_2 \leq L_x$, we obtain

$$\|\nabla_\theta \log p_\theta(y | x)\|_2 \leq \|x\|_2 \cdot \sqrt{2} \leq \sqrt{2} L_x.$$

C.2 LEMMA C.1: PROPERTIES OF THE EMPIRICAL OBJECTIVE $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$

Lemma C.1 (Properties of the empirical objective). *With batch $\mathcal{B}$ and any weight vector w such that $0 < w_{\min} \leq w_k \leq w_{\max} < \infty$ for all k , under Assumption 5.3, the empirical objective $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$ is L_θ -smooth with*

$$L_\theta = \frac{w_{\max}}{2} L_x^2.$$

Note that for $w = \hat{w}^{(t)}$, we have $w_k \geq w_{\min} = \frac{\epsilon_\pi}{\pi_{\text{tr}, \max}} > 0$ and $w_k \leq w_{\max} = \frac{1}{\pi_{\text{tr}, \min}} < \infty$ due to update mechanism of $\hat{w}^{(t)}$ and Assumption 3.4. Moreover, suppose that $|\mathcal{B}| \geq \frac{8L_x^2}{\lambda_x} \log(d/\delta_B)$ for $\delta_B \in (0, 1)$. Under Assumption 5.4, $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$ is $\hat{\mu}_\theta$ -strongly convex for all $\theta \in \Theta$ with

$$\hat{\mu}_\theta := w_{\min} \delta_c^2 \lambda_{\min}(\hat{M}_{\mathcal{B}}),$$

where $\hat{\mu}_\theta \geq \frac{1}{2} \mu_\theta$ with probability $1 - \delta_B$ via Lemma C.2, where $\mu_\theta := w_{\min} \delta_c^2 \lambda_x$, and $\delta_c = \min\{\delta_l, \delta_u\}$ with $\delta_l := \min\{p_\theta(k | x) : k \in [K], \theta \in \Theta, \|x\|_2 \leq L_x\} > 0$ and $\delta_u := \min\{1 - p_\theta(k | x) : k \in [K], \theta \in \Theta, \|x\|_2 \leq L_x\} > 0$.

Proof. With a batch $\mathcal{B}$ and a any weight vector w , recall that the empirical objective is

$$\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) := \frac{1}{|\mathcal{B}|} \sum_{(x_i, y_i) \in \mathcal{B}} [-w_{y_i} \log p_\theta(y_i | x_i)].$$

Smoothness. In this section, an upper bound on the largest eigenvalue of the Hessian is derived, which implies smoothness of $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$.

Step 1: Bounding the Hessian spectral norm. Let $m := |\mathcal{B}|$ and write $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) = \frac{1}{m} \sum_{i=1}^m \ell_i(\theta)$, where $\ell_i(\theta) := -w_{y_i} \log p_\theta(y_i | x_i)$. Recall the notation from Appendix C.1: $\Phi(x) = I_{K'} \otimes x$, $\mathbf{s}(x) = [s_1(x), \dots, s_{K'}(x)]^\top$, $S(x) = \text{diag}(\mathbf{s}(x))$, and $J(x) := S(x) - \mathbf{s}(x)\mathbf{s}(x)^\top$. By Appendix C.1.3, the Hessian of each ℓ_i is

$$\nabla_\theta^2 \ell_i(\theta) = w_{y_i} \Phi(x_i) J(x_i) \Phi(x_i)^\top.$$

Therefore,

$$\nabla_\theta^2 \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) = \frac{1}{m} \sum_{i=1}^m w_{y_i} \Phi(x_i) J(x_i) \Phi(x_i)^\top. \quad (\text{C.2.1})$$

By Jensen's inequality,

$$\left\| \nabla_\theta^2 \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) \right\|_2 \leq \frac{1}{m} \sum_{i=1}^m \left\| w_{y_i} \Phi(x_i) J(x_i) \Phi(x_i)^\top \right\|_2.$$

Sub-multiplicative gives:

$$\left\| w_{y_i} \Phi(x_i) J(x_i) \Phi(x_i)^\top \right\|_2 \leq w_{y_i} \|\Phi(x_i)\|_2^2 \|J(x_i)\|_2.$$

Step 2: Bounding each factor. Three factors separately bounded.

1. Weights. Since $w_k \leq w_{\max}$ for all k , we have $w_{y_i} \leq w_{\max}$ for all i .

2. Feature block. Since $\Phi(x_i) = I_{K'} \otimes x_i$, we have

$$\|\Phi(x_i)\|_2 = \|I_{K'}\|_2 \|x_i\|_2 = \|x_i\|_2.$$

Assumption 5.3 states that $\|x_i\|_2 \leq L_x$, hence

$$\|\Phi(x_i)\|_2^2 \leq L_x^2.$$

3. Jacobian. Employ Gershgorin circle theorem to bound $\|J(x)\|_2 = \lambda_{\max}(J(x))$. The entries of $J(x)$ are

$$J_{kk} = s_k(x) (1 - s_k(x)), \quad J_{kj} = -s_k(x) s_j(x) \quad (k \neq j),$$

for $k, j = 1, \dots, K'$.

For the k -th row, the Gershgorin radius is

$$R_k = \sum_{j \neq k} |J_{kj}| = \sum_{j \neq k} s_k(x) s_j(x) = s_k(x) \sum_{j \neq k} s_j(x).$$

Since $\sum_{j=1}^{K'} s_j(x) = 1 - s_K(x)$, we have

$$\sum_{j \neq k} s_j(x) = (1 - s_K(x)) - s_k(x),$$

and

$$R_k = s_k(x) (1 - s_K(x) - s_k(x)).$$

Gershgorin's theorem implies that every eigenvalue λ of $J(x)$ lies in at least one disc

$$\lambda \in [J_{kk} - R_k, J_{kk} + R_k] = [s_k(x) (1 - s_k(x)) - R_k, s_k(x) (1 - s_k(x)) + R_k].$$

Therefore,

$$\begin{aligned} \lambda_{\max}(J(x)) &\leq \max_k \{J_{kk} + R_k\} \\ &= \max_k \{s_k(x) (1 - s_k(x)) + s_k(x) (1 - s_K(x) - s_k(x))\} \\ &= \max_k \{2s_k(x) (1 - s_k(x)) - s_k(x) s_K(x)\}. \end{aligned}$$

Since $s_k(x) \geq 0$ and $s_K(x) \geq 0$, we have

$$\lambda_{\max}(J(x)) \leq \max_k \{2s_k(x) (1 - s_k(x))\}.$$

The scalar function $g(s) = 2s(1 - s)$ is maximized at $s = 1/2$ with $g(1/2) = 1/2$. Hence,

$$\|J(x)\|_2 = \lambda_{\max}(J(x)) \leq \frac{1}{2}.$$

Step 3: Final bound. Putting the previous bounds together,

$$\begin{aligned} \left\| \nabla_{\theta}^2 \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) \right\|_2 &\leq \frac{1}{m} \sum_{i=1}^m w_{y_i} \|\Phi(x_i)\|_2^2 \|J(x_i)\|_2 \leq \frac{1}{m} \sum_{i=1}^m \left[w_{\max} \cdot L_x^2 \cdot \frac{1}{2} \right] \\ &= \frac{w_{\max} L_x^2}{2}. \end{aligned}$$

Therefore, $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$ is L_{θ} -smooth with $L_{\theta} = w_{\max} L_x^2 / 2$.

Strong convexity. We now show that $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$ is $\hat{\mu}_{\theta}$ -strongly convex.

Step 1: Lower bound on the weights. Since $w_k \geq w_{\min}$ for all k , for every sample (x_i, y_i) ,

$$w_{y_i} \geq w_{\min}.$$

Applying this lower bound in (C.2.1) yields following inequality:

$$\nabla_{\theta}^2 \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) \succeq w_{\min} \cdot \frac{1}{m} \sum_{i=1}^m \Phi(x_i) J(x_i) \Phi(x_i)^{\top}. \quad (\text{C.2.2})$$

Step 2: Lower bound on the Jacobian. Since we have

$$\delta_c \leq p_{\theta}(k \mid x) \leq 1 - \delta_c \quad \text{for all } k, \theta \in \Theta, \text{ and } x \text{ with } \|x\|_2 \leq L_x,$$

$s_k(x) \geq \delta_c$ for all $k = 1, \dots, K$.

Let $v \in \mathbb{R}^{K'}$ be arbitrary. Let analyze the following quadratic form

$$v^{\top} J(x) v = v^{\top} (S(x) - \mathbf{s}(x) \mathbf{s}(x)^{\top}) v = \sum_{j=1}^{K'} s_j(x) v_j^2 - \left(\sum_{j=1}^{K'} s_j(x) v_j \right)^2.$$

Cauchy-Schwarz inequality in the form $\left(\sum_j a_j b_j \right)^2 \leq \left(\sum_j a_j^2 \right) \left(\sum_j b_j^2 \right)$ with

$$a_j = \sqrt{s_j(x)}, \quad b_j = \sqrt{s_j(x)} v_j$$

gives

$$\left(\sum_{j=1}^{K'} s_j(x) v_j \right)^2 \leq \left(\sum_{j=1}^{K'} s_j(x) \right) \left(\sum_{j=1}^{K'} s_j(x) v_j^2 \right).$$

Let $s_K(x) = p_{\theta}(K \mid x)$. Then, we have

$$\sum_{j=1}^{K'} s_j(x) = 1 - s_K(x).$$

Substituting this back gives

$$\begin{aligned} v^{\top} J(x) v &\geq \left(\sum_{j=1}^{K'} s_j(x) v_j^2 \right) - (1 - s_K(x)) \left(\sum_{j=1}^{K'} s_j(x) v_j^2 \right) \\ &= s_K(x) \sum_{j=1}^{K'} s_j(x) v_j^2. \end{aligned}$$

We have $s_K(x) \geq \delta_c$ and $s_j(x) \geq \delta_c$ for all $j = 1, \dots, K'$. Hence,

$$v^{\top} J(x) v \geq \delta_c \sum_{j=1}^{K'} s_j(x) v_j^2 \geq \delta_c \sum_{j=1}^{K'} \delta_c v_j^2 = \delta_c^2 \|v\|_2^2.$$

Therefore $\lambda_{\min}(J(x)) \geq \delta_c^2$ for all x .

Step 3: Combining with the feature block. Let $u \in \mathbb{R}^{dK'}$ be arbitrary. By (C.2.2),

$$\begin{aligned} u^{\top} \nabla_{\theta}^2 \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) u &\geq w_{\min} \cdot \frac{1}{m} \sum_{i=1}^m u^{\top} \Phi(x_i) J(x_i) \Phi(x_i)^{\top} u \\ &= w_{\min} \cdot \frac{1}{m} \sum_{i=1}^m (\Phi(x_i)^{\top} u)^{\top} J(x_i) (\Phi(x_i)^{\top} u). \end{aligned}$$

Define $v(x_i) := \Phi(x_i)^\top u \in \mathbb{R}^{K'}$. From Step 2, for each x_i ,

$$v(x_i)^\top J(x_i) v(x_i) \geq \delta_c^2 \|v(x_i)\|_2^2.$$

Substituting this inequality,

$$\begin{aligned} u^\top \nabla_\theta^2 \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) u &\geq w_{\min} \cdot \frac{1}{m} \sum_{i=1}^m \delta_c^2 \|v(x_i)\|_2^2 \\ &= w_{\min} \delta_c^2 u^\top \left(\frac{1}{m} \sum_{i=1}^m \Phi(x_i) \Phi(x_i)^\top \right) u. \end{aligned}$$

Step 4: Linking to the empirical second moment. Recall that $\Phi(x) = I_{K'} \otimes x$. By properties of the Kronecker product,

$$\Phi(x_i) \Phi(x_i)^\top = (I_{K'} \otimes x_i) (I_{K'} \otimes x_i^\top) = I_{K'} \otimes (x_i x_i^\top).$$

Therefore,

$$\frac{1}{m} \sum_{i=1}^m \Phi(x_i) \Phi(x_i)^\top = I_{K'} \otimes \left(\frac{1}{m} \sum_{i=1}^m x_i x_i^\top \right) = I_{K'} \otimes \hat{M}_{\mathcal{B}}.$$

Combining with the previous step gives

$$u^\top \nabla_\theta^2 \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) u \geq w_{\min} \delta_c^2 u^\top (I_{K'} \otimes \hat{M}_{\mathcal{B}}) u.$$

The minimal eigenvalue of a Kronecker product satisfies

$$\lambda_{\min}(A \otimes B) = \lambda_{\min}(A) \lambda_{\min}(B).$$

Thus,

$$\lambda_{\min}(I_{K'} \otimes \hat{M}_{\mathcal{B}}) = \lambda_{\min}(I_{K'}) \lambda_{\min}(\hat{M}_{\mathcal{B}}) = \lambda_{\min}(\hat{M}_{\mathcal{B}}).$$

Therefore,

$$u^\top \nabla_\theta^2 \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}) u \geq w_{\min} \delta_c^2 \lambda_{\min}(\hat{M}_{\mathcal{B}}) \|u\|_2^2.$$

This shows that $\hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B})$ is $\hat{\mu}_\theta$ -strongly convex with

$$\hat{\mu}_\theta = w_{\min} \delta_c^2 \lambda_{\min}(\hat{M}_{\mathcal{B}}).$$

□

C.3 PROOF OF LEMMA C.2

Lemma C.2 (Lower bound on the empirical second moment). *Under Assumptions 5.3 and 5.4, let $\mathcal{B}$ be an i.i.d. batch of size m drawn from P_{tr} . Define the empirical second-moment matrix*

$$\hat{M}_{\mathcal{B}} := \frac{1}{|\mathcal{B}|} \sum_{(x_i, y_i) \in \mathcal{B}} x_i x_i^\top.$$

Then, for any $\delta_{\mathcal{B}} \in (0, 1)$,

$$\lambda_{\min}(\hat{M}_{\mathcal{B}}) \geq \frac{1}{2} \lambda_x$$

holds with probability at least $1 - \delta_{\mathcal{B}}$ provided that

$$m \geq \frac{8L_x^2}{\lambda_x} \log\left(\frac{d}{\delta_{\mathcal{B}}}\right).$$

Proof. Let $m := |\mathcal{B}|$ and define i.i.d. random PSD matrices $A_i := x_i x_i^\top \in \mathbb{R}^{d \times d}$ so that $\hat{M}_{\mathcal{B}} = \frac{1}{m} \sum_{i=1}^m A_i$. Each A_i is positive semidefinite, hence $\lambda_{\min}(A_i) \geq 0$. By Assumption 5.3, $\lambda_{\max}(A_i) = \|x_i\|_2^2 \leq L_x^2$.

Let $M_x := \mathbb{E}[A_i] = \mathbb{E}_{x \sim P_{\text{tr}}}[xx^\top]$. By Assumption 5.4, $\lambda_{\min}(M_x) \geq \lambda_x$, hence

$$\lambda_{\min}\left(\mathbb{E}\left[\sum_{i=1}^m A_i\right]\right) = \lambda_{\min}(mM_x) \geq m\lambda_x.$$

Let $\mu_{\min} := \lambda_{\min}(\sum_{i=1}^m \mathbb{E}[A_i]) = \lambda_{\min}(\mathbb{E}[\sum_{i=1}^m A_i])$ so that $\mu_{\min} \geq m\lambda_x$. Here, we introduce the matrix inequality from Tropp [2015]

Theorem C.3 (Tropp [2015], Theorem 5.1.1 (Matrix Chernoff)). *Consider a finite sequence $\{X_k\}$ of independent, random, Hermitian matrices with common dimension d . Assume that*

$$0 \leq \lambda_{\min}(X_k) \quad \text{and} \quad \lambda_{\max}(X_k) \leq L \quad \text{for each index } k.$$

Introduce the random matrix

$$Y = \sum_k X_k.$$

Define the minimum eigenvalue $\mu_{\min}$ and maximum eigenvalue $\mu_{\max}$ of the expectation $\mathbb{E}Y$:

$$\begin{aligned} \mu_{\min} &= \lambda_{\min}(\mathbb{E}Y) = \lambda_{\min}\left(\sum_k \mathbb{E}X_k\right), \quad \text{and} \\ \mu_{\max} &= \lambda_{\max}(\mathbb{E}Y) = \lambda_{\max}\left(\sum_k \mathbb{E}X_k\right). \end{aligned}$$

Then, for $\theta > 0$,

$$\begin{aligned} \mathbb{E}\lambda_{\min}(Y) &\geq \frac{1 - e^{-\theta}}{\theta} \mu_{\min} - \frac{1}{\theta} L \log d, \quad \text{and} \\ \mathbb{E}\lambda_{\max}(Y) &\leq \frac{e^{\theta} - 1}{\theta} \mu_{\max} + \frac{1}{\theta} L \log d. \end{aligned}$$

Furthermore,

$$\begin{aligned} \mathbb{P}\{\lambda_{\min}(Y) \leq (1 - \varepsilon)\mu_{\min}\} &\leq d \left[\frac{e^{-\varepsilon}}{(1 - \varepsilon)^{1-\varepsilon}} \right]^{\mu_{\min}/L} \quad \text{for } \varepsilon \in [0, 1), \quad \text{and} \\ \mathbb{P}\{\lambda_{\max}(Y) \geq (1 + \varepsilon)\mu_{\max}\} &\leq d \left[\frac{e^{\varepsilon}}{(1 + \varepsilon)^{1+\varepsilon}} \right]^{\mu_{\max}/L} \quad \text{for } \varepsilon \geq 0. \end{aligned}$$

Applying Theorem C.3 with $X_i = A_i$ and $L = L_x^2$, for any $\varepsilon \in (0, 1)$,

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{i=1}^m A_i\right) \leq (1 - \varepsilon)\mu_{\min}\right) \leq d \left[\frac{\exp(-\varepsilon)}{(1 - \varepsilon)^{1-\varepsilon}} \right]^{\mu_{\min}/L_x^2}.$$

Using the standard inequality $\frac{\exp(-\varepsilon)}{(1 - \varepsilon)^{1-\varepsilon}} \leq \exp(-\varepsilon^2/2)$ for $\varepsilon \in (0, 1)$, we obtain

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{i=1}^m A_i\right) \leq (1 - \varepsilon)\mu_{\min}\right) \leq d \exp\left(-\frac{\varepsilon^2 \mu_{\min}}{2L_x^2}\right) \leq d \exp\left(-\frac{\varepsilon^2 m \lambda_x}{2L_x^2}\right).$$

Since $\mu_{\min} \geq m\lambda_x$, we have $(1 - \varepsilon)m\lambda_x \leq (1 - \varepsilon)\mu_{\min}$, and therefore the same tail bound applies to the event $\lambda_{\min}(\sum_{i=1}^m A_i) \leq (1 - \varepsilon)m\lambda_x$. Choose $\varepsilon = \frac{1}{2}$ and set the right-hand side to be at most $\delta_{\mathcal{B}}$. It suffices that $m \geq \frac{8L_x^2}{\lambda_x} \log\left(\frac{d}{\delta_{\mathcal{B}}}\right)$. On this event,

$$\lambda_{\min}(\hat{M}_{\mathcal{B}}) = \frac{1}{m} \lambda_{\min}\left(\sum_{i=1}^m A_i\right) \geq \frac{1}{2} \lambda_x.$$

□

C.4 PROOF OF LEMMA C.4

Lemma C.4 (Finite-sample error of the empirical optimum). *Under conditions of Lemma C.1, the empirical optimum $\hat{\theta}^* = \arg \min_{\theta \in \Theta} \hat{\mathcal{L}}_{\text{tr}}(\theta; w^*, \mathcal{B})$ satisfies*

$$\mathbb{E} \|\hat{\theta}^* - \theta^*\|_2^2 \leq \frac{w_{\max}^2 G^2}{\hat{\mu}_\theta^2 |\mathcal{B}|}.$$

Proof. Since $w_{\min} \leq w_k^* \leq w_{\max}$ for all k due to Assumption 3.4, Lemma C.1 implies that $\hat{\mathcal{L}}_{\text{tr}}(\cdot; w^*, \mathcal{B})$ is $\hat{\mu}_\theta$ -strongly convex for all $\theta \in \Theta$. Therefore, we have

$$(\nabla_\theta \hat{\mathcal{L}}_{\text{tr}}(\hat{\theta}^*; w^*, \mathcal{B}) - \nabla_\theta \hat{\mathcal{L}}_{\text{tr}}(\theta^*; w^*, \mathcal{B}))^\top (\hat{\theta}^* - \theta^*) \geq \hat{\mu}_\theta \|\hat{\theta}^* - \theta^*\|_2^2.$$

By the first-order optimality condition, $\nabla_\theta \hat{\mathcal{L}}_{\text{tr}}(\hat{\theta}^*; w^*, \mathcal{B})^\top (\hat{\theta}^* - \theta^*) \leq 0$. Using Cauchy–Schwarz, this yields

$$\|\hat{\theta}^* - \theta^*\|_2 \leq \frac{1}{\hat{\mu}_\theta} \|\nabla_\theta \hat{\mathcal{L}}_{\text{tr}}(\hat{\theta}^*; w^*, \mathcal{B})\|_2.$$

Now note that

$$\nabla_\theta \hat{\mathcal{L}}_{\text{tr}}(\theta^*; w^*, \mathcal{B}) = \frac{1}{m} \sum_{i=1}^m X_i,$$

where we set $m := |\mathcal{B}|$ and define $X_i := \nabla_\theta [-w_{y_i}^* \log p_{\theta^*}(y_i | x_i)]$ for $(x_i, y_i) \in \mathcal{B}$. These are i.i.d. random vectors with mean zero because θ^* minimizes the population objective $\mathcal{L}_{\text{tr}}(\theta; w^*)$. Moreover, we have $\|X_i\|_2 \leq w_{\max} G$ under Assumption 5.3. Therefore,

$$\mathbb{E} \left[\|\nabla_\theta \hat{\mathcal{L}}_{\text{tr}}(\theta^*; w^*, \mathcal{B})\|_2^2 \right] = \frac{1}{m^2} \sum_{i=1}^m \mathbb{E} \|X_i\|_2^2 \leq \frac{w_{\max}^2 G^2}{m} = \frac{w_{\max}^2 G^2}{|\mathcal{B}|},$$

Substituting the previous bound yields

$$\mathbb{E} \|\hat{\theta}^* - \theta^*\|_2^2 \leq \frac{1}{\hat{\mu}_\theta^2} \mathbb{E} \left[\|\nabla_\theta \hat{\mathcal{L}}_{\text{tr}}(\theta^*; w^*, \mathcal{B})\|_2^2 \right] \leq \frac{w_{\max}^2 G^2}{\hat{\mu}_\theta^2 |\mathcal{B}|}.$$

□

C.5 PROOF OF LEMMA C.5

Lemma C.5 (Lipschitz Continuity of Optimal Parameter). *Under conditions of Lemma C.1, define the empirical optimizer map*

$$\hat{\theta}(w) := \arg \min_{\theta \in \Theta} \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}).$$

Then, for any weight vectors w_1, w_2 such that $w_{\min} \leq (w_1)_k, (w_2)_k \leq w_{\max}$ for all k , it satisfies

$$\|\hat{\theta}(w_1) - \hat{\theta}(w_2)\|_2 \leq \frac{G}{\hat{\mu}_\theta} \|w_1 - w_2\|_2.$$

Proof. Let $\hat{\theta}_1 := \hat{\theta}(w_1)$ and $\hat{\theta}_2 := \hat{\theta}(w_2)$. For conciseness, define the empirical objective

$$\hat{L}(\theta; w) := \hat{\mathcal{L}}_{\text{tr}}(\theta; w, \mathcal{B}), \quad m := |\mathcal{B}|.$$

Step 1 (Using strong convexity). By Lemma C.1, the function $\hat{L}(\cdot; w_2)$ is $\hat{\mu}_\theta$ -strongly convex for all $\theta \in \Theta$. Therefore,

$$\hat{\mu}_\theta \|\hat{\theta}_1 - \hat{\theta}_2\|_2^2 \leq (\nabla_\theta \hat{L}(\hat{\theta}_1; w_2) - \nabla_\theta \hat{L}(\hat{\theta}_2; w_2))^\top (\hat{\theta}_1 - \hat{\theta}_2).$$

Step 2 (Reducing to the gradient difference in w). Using first order optimality condition,

$$\hat{\mu}_\theta \|\hat{\theta}_1 - \hat{\theta}_2\|_2^2 \leq (\nabla_\theta \hat{L}(\hat{\theta}_1; w_2))^\top (\hat{\theta}_1 - \hat{\theta}_2).$$

Using first order optimality condition once more, we have

$$\hat{\mu}_\theta \|\hat{\theta}_1 - \hat{\theta}_2\|_2^2 \leq (\nabla_\theta \hat{L}(\hat{\theta}_1; w_2) - \nabla_\theta \hat{L}(\hat{\theta}_1; w_1))^\top (\hat{\theta}_1 - \hat{\theta}_2).$$

Then by Cauchy-Schwarz,

$$\|\hat{\theta}_1 - \hat{\theta}_2\|_2 \leq \frac{1}{\hat{\mu}_\theta} \|\nabla_\theta \hat{L}(\hat{\theta}_1; w_1) - \nabla_\theta \hat{L}(\hat{\theta}_1; w_2)\|_2. \quad (\text{C.5.1})$$

Since $\hat{L}$ is linear in w , for any θ ,

$$\nabla_\theta \hat{L}(\theta; w_2) - \nabla_\theta \hat{L}(\theta; w_1) = \frac{1}{m} \sum_{(x,y) \in \mathcal{B}} -(w_{2,y} - w_{1,y}) \nabla_\theta \log p_\theta(y | x).$$

Step 3 (Bounding the gradient difference). By the triangle inequality and bounded gradient,

$$\begin{aligned} \|\nabla_\theta \hat{L}(\hat{\theta}_1; w_2) - \nabla_\theta \hat{L}(\hat{\theta}_1; w_1)\|_2 &\leq \frac{1}{m} \sum_{(x,y) \in \mathcal{B}} |w_{2,y} - w_{1,y}| \cdot \|\nabla_\theta \log p_{\hat{\theta}_1}(y | x)\|_2 \\ &\leq \frac{G}{m} \sum_{(x,y) \in \mathcal{B}} |w_{2,y} - w_{1,y}|. \end{aligned}$$

For each sample (x, y) , we have

$$|w_{2,y} - w_{1,y}| = |\langle e_y, w_2 - w_1 \rangle| \leq \|e_y\|_2 \|w_2 - w_1\|_2 = \|w_2 - w_1\|_2,$$

where $e_y \in \mathbb{R}^K$ is the y -th standard basis vector. Then, we can obtain

$$\|\nabla_\theta \hat{L}(\hat{\theta}_1; w_2) - \nabla_\theta \hat{L}(\hat{\theta}_1; w_1)\|_2 \leq G \|w_2 - w_1\|_2. \quad (\text{C.5.2})$$

Combining (C.5.1) and (C.5.2) gives

$$\|\hat{\theta}_1 - \hat{\theta}_2\|_2 \leq \frac{G}{\hat{\mu}_\theta} \|w_2 - w_1\|_2,$$

which completes the proof. $\square$

C.6 PROOF OF COROLLARY C.6

Corollary C.6. *Under Assumption 3.4 and conditions of Lemma C.5, the cumulative expected average squared error between the instantaneous optimum $\hat{\theta}_t^\dagger$ and the empirical optimum $\hat{\theta}^\star$ is bounded by:*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\hat{\theta}_t^\dagger - \hat{\theta}^\star\|_2^2] \leq C_W \cdot \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\hat{\pi}^{(t)} - \pi_{\text{te}}\|_2^2],$$

where $C_W = \left(\frac{G}{\hat{\mu}_\theta \pi_{\text{tr}, \min}}\right)^2$ is a problem-dependent constant.

Proof. Both $\hat{w}^{(t)}$ and $w^\star$ satisfy the bounds required by Lemma C.5. Therefore,

$$\|\hat{\theta}_t^\dagger - \hat{\theta}^\star\|_2 \leq \frac{G}{\hat{\mu}_\theta} \|\hat{w}^{(t)} - w^\star\|_2.$$

Moreover,

$$\|\hat{w}^{(t)} - w^\star\|_2 = \|(\hat{\pi}^{(t)} - \pi_{\text{te}}) \oslash \pi_{\text{tr}}\|_2 \leq \frac{1}{\pi_{\text{tr}, \min}} \|\hat{\pi}^{(t)} - \pi_{\text{te}}\|_2.$$

Then, we have

$$\|\hat{\theta}_t^\dagger - \hat{\theta}^\star\|_2^2 \leq C_W \|\hat{\pi}^{(t)} - \pi_{\text{te}}\|_2^2,$$

where $C_W = \left(\frac{G}{\hat{\mu}_\theta \pi_{\text{tr}, \min}}\right)^2$. Taking expectation and averaging over $t = 1, \dots, T$ proves the corollary. $\square$

C.7 PROOF OF LEMMA C.7 (BOUNDED PATH LENGTH $\hat{V}_T$)

Lemma C.7 (Bounded Path Length $\hat{V}_T$). *Under Assumption 3.4 and conditions of Lemma C.5, when $\hat{\pi}^{(t)}$ is updated by the Soft-Bayes algorithm with learning rate $\eta_t = \sqrt{\frac{\log K}{2Kt}}$, the path length $\hat{V}_T = \sum_{t=1}^{T-1} \|\hat{\theta}_{t+1}^\dagger - \hat{\theta}_t^\dagger\|_2$ is bounded as:*

$$\hat{V}_T \leq \frac{G}{\hat{\mu}_\theta \pi_{\text{tr}, \min}} \left(2\sqrt{\frac{\log K}{K}}(\sqrt{T} - 1) + \frac{\sqrt{2}}{2} \log T \right).$$

Proof. The proof proceeds by making a relationship between the path length $\hat{V}_T$ and the path length of the estimated label distributions. Subsequently, it employs the properties of the Soft-Bayes update rule to bound the latter.

Step 1: Analyzing $\hat{V}_T$ in terms of the path length of $\hat{\pi}^{(t)}$. Since both $\hat{w}^{(t+1)}$ and $\hat{w}^{(t)}$ satisfy the bounds required by Lemma C.5, we have

$$\|\hat{\theta}_{t+1}^\dagger - \hat{\theta}_t^\dagger\|_2 \leq \frac{G}{\hat{\mu}_\theta} \|\hat{w}^{(t+1)} - \hat{w}^{(t)}\|_2.$$

Using $\hat{w}^{(t)} = \hat{\pi}^{(t)} \oslash \pi_{\text{tr}}$ and Assumption 3.4,

$$\|\hat{\theta}_{t+1}^\dagger - \hat{\theta}_t^\dagger\|_2 \leq \frac{G}{\hat{\mu}_\theta \pi_{\text{tr}, \min}} \|\hat{\pi}^{(t+1)} - \hat{\pi}^{(t)}\|_2.$$

Sum over $t = 1, \dots, T-1$,

$$\hat{V}_T \leq \frac{G}{\hat{\mu}_\theta \pi_{\text{tr}, \min}} \sum_{t=1}^{T-1} \|\hat{\pi}^{(t+1)} - \hat{\pi}^{(t)}\|_2. \quad (\text{C.7.1})$$

Step 2: Bounding the path length of $\hat{\pi}^{(t)}$ under Soft-Bayes. Recall that the update rule of Soft-Bayes is

$$\hat{\pi}^{(t+1)} = \alpha_t \left((1 - \eta_t) \hat{\pi}^{(t)} + \eta_t v_t \right) + (1 - \alpha_t) \hat{\pi}^{(1)}.$$

Rearranging for analyze the difference $\hat{\pi}^{(t+1)} - \hat{\pi}^{(t)}$:

$$\hat{\pi}^{(t+1)} - \hat{\pi}^{(t)} = \eta_{t+1} \left(v_t - \hat{\pi}^{(t)} \right) + (1 - \alpha_t) \left(\hat{\pi}^{(1)} - \hat{\pi}^{(t)} \right).$$

Taking the ℓ_2 norm and by the triangle inequality,

$$\|\hat{\pi}^{(t+1)} - \hat{\pi}^{(t)}\|_2 \leq \eta_{t+1} \|v_t - \hat{\pi}^{(t)}\|_2 + (1 - \alpha_t) \|\hat{\pi}^{(1)} - \hat{\pi}^{(t)}\|_2.$$

Since the vectors v_t , $\hat{\pi}^{(t)}$, and $\hat{\pi}^{(1)}$ all lie in the probability simplex Δ^{K-1} , we can get

$$\|\hat{\pi}^{(t+1)} - \hat{\pi}^{(t)}\|_2 \leq \sqrt{2} (\eta_{t+1} + 1 - \alpha_t).$$

Sum over $t = 1, \dots, T-1$,

$$\sum_{t=1}^{T-1} \|\hat{\pi}^{(t+1)} - \hat{\pi}^{(t)}\|_2 \leq \sqrt{2} \sum_{t=1}^{T-1} (\eta_{t+1} + 1 - \alpha_t) = \sqrt{2} \left(\sum_{t=2}^T \eta_t + \sum_{t=1}^{T-1} \left(1 - \frac{\eta_{t+1}}{\eta_t} \right) \right). \quad (\text{C.7.2})$$

Step 3: Bounding the two sums for η_t . First, analyze the sum of learning rates. Let $\eta_t = \sqrt{\frac{\log K}{2Kt}}$ and $C_\eta = \sqrt{\frac{\log K}{2K}}$. By the integral bound $\sum_{t=2}^T t^{-1/2} \leq \int_1^T x^{-1/2} dx = 2(\sqrt{T} - 1)$, we have

$$\sum_{t=2}^T \eta_t = C_\eta \sum_{t=2}^T \frac{1}{\sqrt{t}} \leq 2C_\eta(\sqrt{T} - 1) = \sqrt{\frac{2 \log K}{K}}(\sqrt{T} - 1).$$

Next, bound the ratio of learning rates. Since $\eta_{t+1}/\eta_t \in (0, 1)$, $1 - (\eta_{t+1}/\eta_t) \leq -\ln(\eta_{t+1}/\eta_t)$. It gives:

$$\sum_{t=1}^{T-1} \left(1 - \frac{\eta_{t+1}}{\eta_t}\right) \leq -\sum_{t=1}^{T-1} \ln\left(\frac{\eta_{t+1}}{\eta_t}\right) = \sum_{t=1}^{T-1} (\ln \eta_t - \ln \eta_{t+1}) = \ln\left(\frac{\eta_1}{\eta_T}\right).$$

With $\eta_t = C_\eta t^{-1/2}$,

$$\ln\left(\frac{\eta_1}{\eta_T}\right) = \ln\left(\frac{C_\eta}{C_\eta T^{-1/2}}\right) = \ln(\sqrt{T}) = \frac{1}{2} \log T.$$

Step 4: Combine the bounds. Substituting these two bounds obtained in Step 3 into (C.7.2),

$$\begin{aligned} \sum_{t=1}^{T-1} \left\| \hat{\pi}^{(t+1)} - \hat{\pi}^{(t)} \right\|_2 &\leq \sqrt{2} \left(\sqrt{\frac{2 \log K}{K}} (\sqrt{T} - 1) + \frac{1}{2} \log T \right) \\ &= 2 \sqrt{\frac{\log K}{K}} (\sqrt{T} - 1) + \frac{\sqrt{2}}{2} \log T. \end{aligned}$$

Combine the bound with (C.7.1),

$$\hat{V}_T \leq \frac{G}{\hat{\mu}_\theta \pi_{\text{tr}, \min}} \left(2 \sqrt{\frac{\log K}{K}} (\sqrt{T} - 1) + \frac{\sqrt{2}}{2} \log T \right),$$

which proves the lemma. $\square$

C.8 PROOF OF THEOREM C.8

Theorem C.8 (Dynamic Regret of Online Model Adaptation). *Suppose conditions of Lemma C.1 hold. With probability $1 - \delta$, the empirical dynamic regret against the instantaneous optimum $\hat{\theta}_t^\dagger$ satisfies*

$$\hat{R}_{\text{dyn}}(T) \leq GK_1 \hat{V}_T + GK_2,$$

where

$$K_1 = \frac{\|\hat{\theta}^{(1)} - \hat{\theta}_1^\dagger\|_2 - \rho \|\hat{\theta}^{(T)} - \hat{\theta}_T^\dagger\|_2}{1 - \rho}, \quad K_2 = \frac{1}{1 - \rho},$$

with $\rho = \sqrt{1 - \frac{h\hat{\mu}_\theta}{\gamma}} < 1$ for any $\gamma \geq L_\theta$ and $h \in (0, 1]$. Consequently, when Lemma C.7 holds,

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\hat{\theta}^{(t)} - \hat{\theta}_t^\dagger\|_2^2] \leq \mathcal{O}\left(\sqrt{\frac{\log K}{KT}}\right).$$

Proof. We first restate the corollary from Mokhtari et al. [2016] that we use.

Corollary 1 of Mokhtari et al. [2016] (restated). Let $\mathcal{X}$ be a nonempty convex subset of a Euclidean space, $f_t : \mathcal{X} \rightarrow \mathbb{R}$, $t = 1, \dots, T$, and $\mathbf{x}_t^* := \arg \min_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x})$. Assume that for all t : (i) f_t is μ -strongly convex on $\mathcal{X}$ with $\mu > 0$; (ii) ∇f_t is L -Lipschitz continuous on $\mathcal{X}$; and (iii) $\sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x})\|_2 \leq G$. Given $\mathbf{x}_1 \in \mathcal{X}$, consider the online gradient descent (OGD) scheme

$$\tilde{\mathbf{x}}_t = \Pi_{\mathcal{X}}\left(\mathbf{x}_t - \frac{1}{\gamma} \nabla f_t(\mathbf{x}_t)\right), \quad \mathbf{x}_{t+1} = \mathbf{x}_t + h(\tilde{\mathbf{x}}_t - \mathbf{x}_t),$$

where $\Pi_{\mathcal{X}}$ is the Euclidean projection onto $\mathcal{X}$, $h \in (0, 1]$, and $\gamma \geq L$. Then, with

$$\rho = \sqrt{1 - \frac{h\mu}{\gamma}} < 1, \quad K_1 = \frac{\|\mathbf{x}_1 - \mathbf{x}_1^*\|_2 - \rho \|\mathbf{x}_T - \mathbf{x}_T^*\|_2}{1 - \rho}, \quad K_2 = \frac{1}{1 - \rho},$$

the dynamic regret satisfies

$$\text{Reg}_T^d := \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_t^*) \leq GK_1 \sum_{t=2}^T \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|_2 + GK_2.$$

We apply this corollary with $f_t = \hat{F}_t$, $\mathbf{x}_t = \hat{\theta}^{(t)}$, $\mathbf{x}_t^* = \hat{\theta}_t^\dagger$, and $\mathcal{X} = \Theta$. Its assumptions are satisfied by conditions of Lemma C.1, each $\hat{F}_t$ is $\hat{\mu}_\theta$ -strongly convex and L_θ -smooth, and by (C.1.1), $\|\nabla \hat{F}_t\|_2 \leq w_{\max} \sqrt{2} L_x =: G$. Since the Stage 2 update is this OGD scheme with $h = 1$ and stepsize $1/\gamma \leq 1/L_\theta$, the the corollary yields the stated bound

$$\hat{R}_{\text{dyn}}(T) \leq GK_1 \sum_{t=2}^T \|\hat{\theta}_t^\dagger - \hat{\theta}_{t-1}^\dagger\|_2 + GK_2 = GK_1 \hat{V}_T + GK_2.$$

Finally, $\hat{\mu}_\theta$ -strong convexity implies that for each t ,

$$\hat{F}_t(\hat{\theta}^{(t)}) - \hat{F}_t(\hat{\theta}_t^\dagger) \geq \frac{\hat{\mu}_\theta}{2} \|\hat{\theta}^{(t)} - \hat{\theta}_t^\dagger\|_2^2.$$

Summing over t and dividing by T gives

$$\frac{1}{T} \sum_{t=1}^T \|\hat{\theta}^{(t)} - \hat{\theta}_t^\dagger\|_2^2 \leq \frac{2}{\hat{\mu}_\theta T} \hat{R}_{\text{dyn}}(T) \leq \frac{2G}{\hat{\mu}_\theta T} (K_1 \hat{V}_T + K_2).$$

Taking expectations and substituting the path-length bound from Lemma C.7 yields the claimed rate. $\square$

D PROOFS FOR SECTION 5.2: CONVERGENCE OF θ

D.1 PROOF OF THEOREM 5.5

Theorem 5.5 (Online Error Bound for Model Adaptation). Let $\delta_{\mathcal{B}} \in (0, 1)$ and suppose $|\mathcal{B}| \geq C \log(d/\delta_{\mathcal{B}})$ for some constant C defined in Section C.3. Suppose Assumptions 3.3, 3.4, 5.1, 5.3, 5.4, 5.6 and 5.7 hold. Then, with probability $1 - 2\delta - \delta_{\mathcal{B}}$, the expected average squared error converges as:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\hat{\theta}^{(t)} - \theta^*\|_2^2] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{\log(1/\delta)}{\mu_f \sqrt{T}}\right) + \mathcal{O}\left(\frac{\mathcal{E}(f)^2}{\mu_f^2}\right) + \mathcal{O}\left(\frac{1}{|\mathcal{B}|}\right).$$

Proof. Let $\hat{\theta}^* := \arg \min_{\theta \in \Theta} \hat{\mathcal{L}}_{\text{tr}}(\theta; w^*, \mathcal{B})$ denote the empirical optimum under the true weights. By the triangle inequality and $(a + b + c)^2 \leq 3a^2 + 3b^2 + 3c^2$, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\hat{\theta}^{(t)} - \theta^*\|_2^2] &\leq 3 \cdot \underbrace{\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\hat{\theta}^{(t)} - \hat{\theta}_t^\dagger\|_2^2]}_{\text{(a) Estimated branch}} + 3 \cdot \underbrace{\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\hat{\theta}_t^\dagger - \hat{\theta}^*\|_2^2]}_{\text{(b) Estimated-vs-true branch}} \\ &\quad + 3 \cdot \underbrace{\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\hat{\theta}^* - \theta^*\|_2^2]}_{\text{(c) True branch}}. \end{aligned}$$

The term (a) can be bounded by Theorem C.8, the term (b) can be bounded by Corollary C.6 and Theorem 5.2, and the term (c) can be bounded by Lemma C.4. Combining these bounds yields

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\hat{\theta}^{(t)} - \theta^*\|_2^2] &\leq \underbrace{\frac{6G_{\text{tr}}K_1}{\hat{\mu}_\theta T} \hat{V}_T + \frac{6G_{\text{tr}}K_2}{\hat{\mu}_\theta T}}_{\text{(Optimization Error: Dynamic OCO)}} \\ &\quad + \underbrace{\frac{12C_W}{\mu_f} \frac{\tilde{R}(T)}{T} + \frac{24C_W B_{\tilde{\ell}}}{\mu_f} \sqrt{\frac{2 \log(2/\delta)}{T}}}_{\text{(Finite-Sample Label Estimation Error)}} \\ &\quad + \underbrace{\frac{6C_W (C')^2}{\mu_f^2} \mathcal{E}(f)^2}_{\text{(Calibration Gap)}} + \underbrace{\frac{3w_{\max}^2 G^2}{\hat{\mu}_\theta^2 |\mathcal{B}|}}_{\text{(Finite-Sample Model Estimation Error)}}. \end{aligned}$$

Substituting the path-length bound from Lemma C.7 into the first term and applying $\mu_f = \frac{\pi_{\text{tr},\min}^2}{\pi_{\text{tr},\max}^2} \lambda_f^{\text{te}}$ (Proposition B.1) to replace $1/\mu$ and $1/\mu^2$ yields

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\|\hat{\theta}^{(t)} - \theta^*\|_2^2 \right] &\leq \frac{6G_{\text{tr}}K_1}{\hat{\mu}_\theta T} \hat{V}_T + \frac{6G_{\text{tr}}K_2}{\hat{\mu}_\theta T} \\ &\quad + \frac{12C_W \pi_{\text{tr},\max}^2}{\pi_{\text{tr},\min}^2 \lambda_f^{\text{te}}} \frac{\tilde{R}(T)}{T} + \frac{24C_W B_{\tilde{\ell}} \pi_{\text{tr},\max}^2}{\pi_{\text{tr},\min}^2 \lambda_f^{\text{te}}} \sqrt{\frac{2 \log(2/\delta)}{T}} \\ &\quad + \frac{6C_W (C')^2 \pi_{\text{tr},\max}^4}{\pi_{\text{tr},\min}^4 (\lambda_f^{\text{te}})^2} \mathcal{E}(f)^2 + \frac{3w_{\max}^2 G^2}{\hat{\mu}_\theta^2 |\mathcal{B}|}. \end{aligned}$$

□

E PROOFS FOR SECTION 5.3: CALIBRATION STRATEGIES

This appendix provides the proofs of the binning-based calibration results in Section 5.3.

E.1 PROOF OF LEMMA 5.8(I)

Lemma 5.8(i) (Calibration Trade-off). (i) Geometric Improvement: Let h_{B_1} and h_{B_2} be two binning schemes such that the partition induced by h_{B_2} is a refinement of that of h_{B_1} (i.e., $\sigma(h_{B_1}(X)) \subset \sigma(h_{B_2}(X))$). Then for $f_{B_1,c}(x) = p_{h_{B_1}}(x)$ and $f_{B_2,c}(x) = p_{h_{B_2}}(x)$, the expected outer products satisfy:

$$\mathbb{E}_{\text{tr}}[f_{B_2,c} f_{B_2,c}^\top] \succeq \mathbb{E}_{\text{tr}}[f_{B_1,c} f_{B_1,c}^\top].$$

Hence, $\lambda_{f_{B_2,c}}^{\text{tr}} \geq \lambda_{f_{B_1,c}}^{\text{tr}}$. Moreover, under Assumptions 5.1, 5.6 and 5.7 and for $f_B(x) = \hat{p}_{h_B}(x)$, we have with probability at least $1 - \delta$,

$$\lambda_{f_B}^{\text{tr}} \geq \lambda_{f_{B,c}}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} > 0,$$

provided that $n_{\text{cal}} \geq \frac{128B}{(\lambda_{f_{0,c}}^{\text{tr}})^2} \log\left(\frac{2B}{\delta}\right)$.

Proof. Let $\mathcal{H}_1 = \sigma(h_{B_1}(X))$ and $\mathcal{H}_2 = \sigma(h_{B_2}(X))$. Since h_{B_2} is a refinement of h_{B_1} , we have $\mathcal{H}_1 \subset \mathcal{H}_2$. By the tower property of conditional expectation:

$$\begin{aligned} f_{B_1,c}(X) &= \mathbb{E}_{\text{tr}}[Y | h_{B_1}(X)] = \mathbb{E}_{\text{tr}}[\mathbb{E}_{\text{tr}}[Y | h_{B_2}(X)] | h_{B_1}(X)] \\ &= \mathbb{E}_{\text{tr}}[f_{B_2,c}(X) | h_{B_1}(X)]. \end{aligned}$$

Apply the multivariate law of total variance to $f_{B_2,c}(X)$ conditioned on $\mathcal{H}_1$:

$$\begin{aligned} \mathbb{E}_{\text{tr}}[f_{B_2,c} f_{B_2,c}^\top] &= \mathbb{E}_{\text{tr}}[\mathbb{E}_{\text{tr}}[f_{B_2,c} f_{B_2,c}^\top | h_{B_1}(X)]] \\ &= \mathbb{E}_{\text{tr}}\left[\mathbb{E}_{\text{tr}}[f_{B_2,c} | h_{B_1}(X)] \mathbb{E}_{\text{tr}}[f_{B_2,c} | h_{B_1}(X)]^\top + \text{Var}_{\text{tr}}(f_{B_2,c} | h_{B_1}(X))\right]. \end{aligned}$$

Substituting $\mathbb{E}_{\text{tr}}[f_{B_2,c}(X) | h_{B_1}(X)] = f_{B_1,c}(X)$ gives

$$\mathbb{E}_{\text{tr}}[f_{B_2,c} f_{B_2,c}^\top] = \mathbb{E}_{\text{tr}}[f_{B_1,c} f_{B_1,c}^\top] + \mathbb{E}_{\text{tr}}[\text{Var}_{\text{tr}}(f_{B_2,c} | \mathcal{H}_1)].$$

Since $\text{Var}_{\text{tr}}(f_{B_2,c} | \mathcal{H}_1)$ is positive semi-definite, we obtain

$$\mathbb{E}_{\text{tr}}[f_{B_2,c} f_{B_2,c}^\top] \succeq \mathbb{E}_{\text{tr}}[f_{B_1,c} f_{B_1,c}^\top].$$

Since the Loewner order is preserved, we have $\lambda_{f_{B_2}}^{\text{tr}} \geq \lambda_{f_{B_1}}^{\text{tr}}$. We now analyze the difference between $\lambda_{f_B}^{\text{tr}}$ and $\lambda_{f_{B,c}}^{\text{tr}}$ for any B . We use the fact that when A and B are Hermitian matrices,

$$|\lambda_{\min}(A + B) - \lambda_{\min}(A)| \leq \|B\|_{\text{op}}$$

by Weyl's inequality, where $\|\cdot\|_{op}$ denotes the operator norm. Hence we have,

$$\lambda_{\min}(\mathbb{E}_{\text{tr}}[f_B f_B^\top]) \geq \lambda_{\min}(\mathbb{E}_{\text{tr}}[f_{B,c} f_{B,c}^\top]) - \|\mathbb{E}_{\text{tr}}[f_B f_B^\top - f_{B,c} f_{B,c}^\top]\|_{op}.$$

Here,

$$\begin{aligned} \|\mathbb{E}_{\text{tr}}[f_B f_B^\top - f_{B,c} f_{B,c}^\top]\|_{op} &\leq \mathbb{E}_{\text{tr}}[\|f_B f_B^\top - f_{B,c} f_{B,c}^\top\|_{op}] \\ &\leq \mathbb{E}_{\text{tr}}[\|f_B(f_B - f_{B,c})^\top + (f_B - f_{B,c})f_{B,c}^\top\|_{op}] \\ &\leq \mathbb{E}_{\text{tr}}[(\|f_B\|_2 + \|f_{B,c}\|_2)\|f_B - f_{B,c}\|_2] \\ &\leq 2\mathbb{E}_{\text{tr}}[\|f_B - f_{B,c}\|_2] \\ &\leq 2\sqrt{\mathbb{E}_{\text{tr}}[\|f_B - f_{B,c}\|_2^2]} = 2\mathcal{E}(f_B), \end{aligned}$$

where the first and last inequalities are due to Jensen's inequality. According to Section E.2, we have with probability at least $1 - \delta$, $\mathcal{E}(f_B)^2 \leq \frac{8B}{n_{cal}} \log\left(\frac{2B}{\delta}\right)$. Hence, provided that $n_{cal} \geq \frac{128B}{(\lambda_{f_0,c}^{\text{tr}})^2} \log\left(\frac{2B}{\delta}\right)$, we have with probability at least $1 - \delta$,

$$\lambda_{f_B}^{\text{tr}} \geq \lambda_{f_{B,c}}^{\text{tr}} - \frac{\lambda_{f_0,c}^{\text{tr}}}{2} > 0,$$

where the last inequality deterministically holds due to $\lambda_{f_{B,c}}^{\text{tr}} \geq \lambda_{f_0,c}^{\text{tr}}$ established above and $\lambda_{f_0,c}^{\text{tr}} > 0$ in Assumption 5.1. $\square$

E.2 PROOF OF LEMMA 5.8(II)

Lemma 5.8(ii) (Calibration Trade-off). Under Assumptions 5.6 and 5.7, With probability at least $1 - \delta$, the squared expected canonical calibration error is bounded by:

$$\mathcal{E}(f_B)^2 \leq \frac{8B}{n_{cal}} \log\left(\frac{2B}{\delta}\right),$$

provided that $n_{cal} \geq 16\log(2/\delta)/P_{\min}^2$.

Proof. For each bin b , denote the bin mass

$$P_b := P_{\text{tr}}(h_B(X) = b),$$

and let m_b be the number of calibration samples falling into bin b . Define the true label distribution in each bin b as

$$p_b \in \Delta^{K-1}, \quad p_{b,k} = P_{\text{tr}}(Y = k \mid h_B(X) = b).$$

Let $\hat{p}_b$ be the empirical estimate of p_b based on calibration samples in bin b :

$$\hat{p}_{b,k} = \frac{1}{m_b} \sum_{i: h_B(x_i) = b} \mathbf{1}\{y_i = k\}.$$

Then, the predictor f_B can be expressed as

$$f_B(x) := \hat{P}_{\text{tr}}(y = \cdot \mid h_B(x)) = \hat{p}_{h_B(x)},$$

and we analogously define the calibrated predictor $f_{B,c}$ as

$$f_{B,c}(x) := P_{\text{tr}}(y = \cdot \mid h_B(x)) = p_{h_B(x)},$$

The calibration error of f_B is then

$$\begin{aligned} \mathcal{E}(f_B)^2 &= \mathbb{E}_{\text{tr}}[\|f_{B,c}(X) - f_B(X)\|_2^2] \\ &= \sum_{b=1}^B P_b \|\hat{p}_b - p_b\|_2^2. \end{aligned}$$

Given $m_b, m_b \hat{p}_b$ is multinomial random variable, following $\mathcal{M}(m_b, p_b)$. Applying McDiarmid's inequality, the following holds with probability at least $1 - \delta/(2B)$ for each $b \in [B]$:

$$\|\hat{p}_b - p_b\|_2 \leq 2\sqrt{\frac{\log(2B/\delta)}{m_b}}.$$

Hence, with probability at least $1 - \delta/2$,

$$\mathcal{E}(f_B)^2 \leq 4\log\left(\frac{2B}{\delta}\right) \sum_{b=1}^B \frac{P_b}{m_b}.$$

We now observe that $m := [m_1, \dots, m_B]^\top \sim \mathcal{M}(n_{cal}, P)$, where $P := [P_1, \dots, P_B]$. Applying McDiarmid's inequality once more, we have with probability at least $1 - \delta/2$,

$$\left\| \frac{m}{n_{cal}} - P \right\|_2 \leq 2\sqrt{\frac{\log(2/\delta)}{n_{cal}}}.$$

Hence, with probability at least $1 - \delta/2$,

$$m_b \geq n_{cal} \frac{P_b}{2} \quad \text{for all } b \in [B]$$

provided that $n_{cal} \geq 16\log(2/\delta)/P_{\min}^2$, where $P_{\min} = \min_{b \in [B]} P_b$. Combining the inequalities, we have with probability at least $1 - \delta$,

$$\mathcal{E}(f_B)^2 \leq 4\log\left(\frac{2B}{\delta}\right) \sum_{b=1}^B \frac{P_b}{m_b} \leq \frac{8B}{n_{cal}} \log\left(\frac{2B}{\delta}\right).$$

□

E.3 PROOF OF THEOREM 5.9 (OPTIMAL BINNING STRATEGY)

Theorem 5.9 (Optimal Binning Strategy). Suppose Assumptions 3.3, 3.4, 5.1, 5.3, 5.6 and 5.7 hold.

For the class-wise refinement strategy $B = Ks$, we obtain $s^* = \sqrt{\frac{5C_1}{12\bar{\mu}_\infty(K)}}$.

Proof. From Theorem 5.5, the bias term in the model-parameter error is

$$\frac{4C_W C'^2}{\mu_f^2} \mathcal{E}(f)^2,$$

where C_W, C'^2 are independent of B .

Step 1: Identification of B -dependent terms. By Lemma 5.8(ii), for the equal-mass binned predictor f_B ,

$$\mathcal{E}(f_B)^2 \leq \frac{8B}{n_{cal}} \log\left(\frac{2B}{\delta}\right).$$

Also by Lemma 5.8(i) and Proposition B.1, we have

$$\mu_f = \frac{\pi_{\text{tr},\min}^2}{\pi_{\text{tr},\max}^2} \lambda_f^{\text{te}} \geq \frac{\pi_{\text{tr},\min}^2}{\pi_{\text{tr},\max}^2} \frac{\pi_{\text{te},\min}}{\pi_{\text{tr},\max}} \lambda_f^{\text{tr}} \geq \frac{\pi_{\text{tr},\min}^2}{\pi_{\text{tr},\max}^2} \frac{\pi_{\text{te},\min}}{\pi_{\text{tr},\max}} \left(\lambda_{f,c}^{\text{tr}} - \frac{\lambda_{f_0,c}^{\text{tr}}}{2} \right).$$

Therefore, define $J(B)$ with B -dependent terms from the calibration error term in Theorem 5.5:

$$J(B) := \frac{B \log(2B/\delta)}{n_{cal} \left(\lambda_{f,c}^{\text{tr}} - \frac{\lambda_{f_0,c}^{\text{tr}}}{2} \right)^2}.$$

Step 2: Analysis of $\mathbb{E}_{tr}[f_{B,c}f_{B,c}^\top]$ via Decomposition. Let $f^*(X) := P_{tr}(Y | X)$ and $f_{B,c}$ denote the calibrated predictor with B bins. Recall that the region $\mathcal{X}$ is divided into disjoint regions $\mathcal{X}_1, \dots, \mathcal{X}_K$, where $\hat{y}(x) = \operatorname{argmax}_{j \in [K]} f_{0,j}(x) = k$ for $x \in \mathcal{X}_k$. We further divide each region $\mathcal{X}_k$ via histogram binning on the values of $f_{0,k}(x)$, utilizing a held out dataset $\mathcal{Q}_{cal} = \{x_i\}_{i=1}^{|\mathcal{Q}_{cal}|}$ sampled from the training distribution independently of the set D_{cal} .

For $x \in \mathcal{Q}_{cal,k} := \mathcal{X}_k \cap \mathcal{Q}_{cal}$, let $U_k(x) := f_{0,k}(x)$ and let

$$F_{U_k}(u) := \hat{\mathbb{P}}_{tr}(U_k \leq u), \quad u \in [0, 1]$$

be its empirical CDF. By the property of CDF,

$$V_k(X) := F_{U_k}(U_k(X))$$

follows a uniform distribution on $\{0, \frac{1}{|\mathcal{Q}_{cal,k}|}, \frac{2}{|\mathcal{Q}_{cal,k}|}, \dots, 1\}$ for $X \in \mathcal{X}_k$ with $X \sim P_{tr}$, due to the fact that $U_k(X) \stackrel{d}{=} U_k(x_i)$. (We assume $|\mathcal{Q}_{cal,k}|$ is sufficiently large for all $k \in [K]$. Note that we do not need the labels of $x_i \in \mathcal{Q}_{cal}$.) Under equal-mass binning, $[0, 1]$ is partitioned into $s \geq 1$ intervals

$$I_j = [j/s, (j+1)/s), \quad j = 0, \dots, s-1,$$

each with probability $\mathbb{P}_{tr}(V_k(X) \in I_j | X \in \mathcal{X}_k) = 1/s$. Let

$$\Delta := 1/s, \quad v_j := (j + 0.5)/s$$

denote the bin width and the center of bin I_j , respectively. Define the conditional expectation function

$$m_k(v) := \mathbb{E}_{tr}[f^*(X) | V_k(X) = v, X \in \mathcal{X}_k] \in \mathbb{R}^K, \quad k \in [K],$$

and assume that $m_k(v)$ is twice differentiable with a uniform second-derivative bound: there exists $L_2 > 0$ such that

$$\|m_k''(v)\|_2 \leq L_2 \quad \text{for all } v \in [0, 1] \text{ and all } k \in [K].$$

The binned predictor $f_{B,c}$ assigns the class distribution within each bin, i.e., for $X \in \mathcal{X}_k$ such that $V_k(X) \in I_j$,

$$f_{B,c}(X) = \mathbb{E}_{tr}[f^*(X) | V_k(X) \in I_j, X \in \mathcal{X}_k] = \mathbb{E}_{tr}[m(V_k(X)) | V_k(X) \in I_j, X \in \mathcal{X}_k] =: \mu_{k,j}.$$

Then,

$$\mathbb{E}_{tr}[f_{B,c}(X)f_{B,c}(X)^\top | X \in \mathcal{X}_k] = \sum_{j=0}^{s-1} \mathbb{P}_{tr}(V_k(X) \in I_j | X \in \mathcal{X}_k) \mu_{k,j} \mu_{k,j}^\top = \frac{1}{s} \sum_{j=0}^{s-1} \mu_{k,j} \mu_{k,j}^\top.$$

For each bin I_j , second moment of $m_k(V_k)$ decomposes into the squared mean and the covariance:

$$\mathbb{E}_{tr}[m_k(V_k)m_k(V_k)^\top | V_k \in I_j, X \in \mathcal{X}_k] = \mu_{k,j} \mu_{k,j}^\top + \operatorname{Cov}(m_k(V_k) | V_k \in I_j, X \in \mathcal{X}_k).$$

Averaging over all bins gives

$$\mathbb{E}_{tr}[m_k(V_k)m_k(V_k)^\top | X \in \mathcal{X}_k] = \frac{1}{s} \sum_{j=0}^{s-1} \mu_{k,j} \mu_{k,j}^\top + \frac{1}{s} \sum_{j=0}^{s-1} \operatorname{Cov}(m_k(V_k) | V_k \in I_j, X \in \mathcal{X}_k).$$

And we can obtain the following expression:

$$\mathbb{E}_{tr}[f_{B,c}(X)f_{B,c}(X)^\top | X \in \mathcal{X}_k] = \mathbb{E}_{tr}[m_k(V_k)m_k(V_k)^\top | X \in \mathcal{X}_k] - \frac{1}{s} \sum_{j=0}^{s-1} \operatorname{Cov}(m_k(V_k) | V_k \in I_j, X \in \mathcal{X}_k). \quad (\text{E.3.1})$$

Since the first term does not depend on s , we now focus on $\frac{1}{s} \sum_{j=0}^{s-1} \operatorname{Cov}(m_k(V_k) | V_k \in I_j, X \in \mathcal{X}_k)$. (We temporarily drop the subscript k and the conditioning argument $X \in \mathcal{X}_k$ for notational simplicity.) For each bin I_j , define the centered variable

$$Z := V - v_j | V \in I_j.$$

Since $V \mid (V \in I_j)$ is uniform on $[j/s, (j+1)/s)$, we have

$$Z \sim \text{Unif}[-\Delta/2, \Delta/2], \quad \mathbb{E}[Z] = 0, \quad \mathbb{E}[Z^2] = \frac{\Delta^2}{12}, \quad \mathbb{E}[|Z|^3] = \frac{\Delta^3}{32}, \quad \mathbb{E}[Z^4] = \frac{\Delta^4}{80}.$$

Using a first-order Taylor expansion of $m(\cdot)$ around the bin center v_j ,

$$m(v_j + Z) = m(v_j) + m'(v_j)Z + r(Z),$$

where the remainder satisfies

$$\|r(Z)\|_2 \leq \frac{1}{2}L_2Z^2.$$

The covariance within bin I_j becomes:

$$\text{Cov}(m(V) \mid V \in I_j) = \text{Cov}(m(v_j + Z)) = \text{Cov}(m(v_j) + m'(v_j)Z + r(Z)).$$

Since $m(v_j)$ is a deterministic vector,

$$\text{Cov}(m(v_j + Z)) = \text{Cov}(m'(v_j)Z + r(Z)).$$

By $\text{Cov}(X + Y) = \text{Cov}(X) + \text{Cov}(Y) + \text{Cov}(X, Y) + \text{Cov}(Y, X)$ and $\mathbb{E}[Z] = 0$, we obtain three terms:

$$\text{Cov}(m'(v_j)Z + r(Z)) = \text{Cov}(m'(v_j)Z) + \text{Cov}(r(Z)) + \text{Cross}_j,$$

where Cross_j consists of the mixed covariance terms involving $m'(v_j)Z$ and $r(Z)$. Let us analyze each term separately.

1. Main term. Since $m'(v_j)$ is deterministic within the bin,

$$\text{Cov}(m'(v_j)Z) = m'(v_j) \text{Var}(Z) m'(v_j)^\top = m'(v_j) \mathbb{E}[Z^2] m'(v_j)^\top = \frac{\Delta^2}{12} m'(v_j) m'(v_j)^\top.$$

2. Cross term. The exact form of the cross term is

$$\mathbb{E}[m'(v_j)Z r(Z)^\top] + \mathbb{E}[r(Z) m'(v_j)^\top Z] - \mathbb{E}[m'(v_j)Z] \mathbb{E}[r(Z)]^\top - \mathbb{E}[r(Z)] \mathbb{E}[m'(v_j)Z]^\top.$$

The last two terms vanish because $\mathbb{E}[m'(v_j)Z] = m'(v_j) \mathbb{E}[Z] = 0$, so we only need to analyze the first two terms. Using Jensen's inequality and the bound on $r(Z)$,

$$\begin{aligned} \|\mathbb{E}[m'(v_j)Z r(Z)^\top]\|_2 &\leq \mathbb{E}[\|m'(v_j)Z r(Z)^\top\|_2] \\ &= \|m'(v_j)\|_2 \mathbb{E}[|Z| \|r(Z)\|_2] \\ &\leq \|m'(v_j)\|_2 \cdot \frac{1}{2}L_2 \mathbb{E}[|Z|^3] = \frac{1}{64} \|m'(v_j)\|_2 L_2 \Delta^3. \end{aligned}$$

Similarly, the transposed term $\mathbb{E}[r(Z) m'(v_j)^\top Z]$ has the same bound. Then,

$$\|\text{Cross}_j\|_2 \leq \frac{1}{32} \|m'(v_j)\|_2 L_2 \Delta^3.$$

3. Remainder term. Since

$$\text{Cov}(r(Z)) = \mathbb{E}[r(Z)r(Z)^\top] - \mathbb{E}[r(Z)]\mathbb{E}[r(Z)]^\top \preceq \mathbb{E}[r(Z)r(Z)^\top],$$

we have

$$\|\text{Cov}(r(Z))\|_2 \leq \|\mathbb{E}[r(Z)r(Z)^\top]\|_2 \leq \mathbb{E}[\|r(Z)\|_2^2] \leq \frac{1}{4}L_2^2 \mathbb{E}[Z^4] = \frac{L_2^2 \Delta^4}{320}.$$

Collecting these bounds, we can write

$$\text{Cov}(m(V) \mid V \in I_j) = \frac{\Delta^2}{12} m'(v_j) m'(v_j)^\top + R_j,$$

with

$$\|R_j\|_2 \leq \frac{1}{32} \|m'(v_j)\|_2 L_2 \Delta^3 + \frac{L_2^2 \Delta^4}{320}.$$

Substituting $\Delta = 1/s$ and this expression into (E.3.1), we obtain

$$\mathbb{E}_{\text{tr}}[f_{B,c}(X)f_{B,c}(X)^\top | X \in \mathcal{X}_k] = \mathbb{E}_{\text{tr}}[m_k(V_k)m_k(V_k)^\top | X \in \mathcal{X}_k] - \left(\frac{1}{12s^2} \frac{1}{s} \sum_{j=0}^{s-1} m'_k(v_j)m'_k(v_j)^\top + R_s^k \right),$$

where R_s^k satisfies

$$\|R_s^k\|_2 \leq \frac{L_2}{32s^3} \frac{1}{s} \sum_{j=0}^{s-1} \|m'_k(v_j)\|_2 + \frac{L_2^2}{320s^4}.$$

Now averaging over all regions $\mathcal{X}_k$'s, we have,

$$\begin{aligned} \mathbb{E}_{\text{tr}}[f_{B,c}(X)f_{B,c}(X)^\top] &= \sum_{k=1}^K P(X \in \mathcal{X}_k) \mathbb{E}_{\text{tr}}[m_k(V_k)m_k(V_k)^\top | X \in \mathcal{X}_k] \\ &\quad - \underbrace{\sum_{k=1}^K P(X \in \mathcal{X}_k) \left(\frac{1}{12s^2} \frac{1}{s} \sum_{j=0}^{s-1} m'_k(v_j)m'_k(v_j)^\top + R_s^k \right)}_{=:\Delta_{\text{disc}}(s)}, \end{aligned}$$

Step 3: Lower bound on $\lambda_{f,c}^{\text{tr}}$. We have $\lambda_{f,c}^{\text{tr}} = \lambda_{\min}(\mathbb{E}_{\text{tr}}[f_{B,c}(X)f_{B,c}(X)^\top])$. Let also

$$\lambda_\infty^{\text{tr}} = \lambda_{\min} \left(\sum_{k=1}^K P(X \in \mathcal{X}_k) \mathbb{E}_{\text{tr}}[m_k(V_k)m_k(V_k)^\top | X \in \mathcal{X}_k] \right),$$

whose value does not depend on s . By Weyl's inequality and trace inequality, we have

$$\lambda_{f,c}^{\text{tr}} \geq \lambda_\infty^{\text{tr}} - \lambda_{\max}(\Delta_{\text{disc}}(s)) \tag{E.3.2}$$

$$\geq \lambda_\infty^{\text{tr}} - \text{tr}(\Delta_{\text{disc}}(s)). \tag{E.3.3}$$

Define the following supremum constants:

$$C_1 := \sup_{k \in [K], v \in [0,1]} \|m'_k(v)\|_2^2, \quad C'_1 := \sup_{k \in [K], v \in [0,1]} \|m'_k(v)\|_2.$$

Also, using the fact that $\text{tr}(R_B) \leq K\|R_B\|_2$, we have,

$$\text{tr}(\Delta_{\text{disc}}(s)) \leq \frac{C_1}{12s^2} + \frac{KL_2C'_1}{32s^3} + \frac{KL_2^2}{320s^4}.$$

Substituting this into (E.3.3), we obtain

$$\lambda_{f,c}^{\text{tr}} \geq \underline{\lambda}_s^{\text{tr}} := \lambda_\infty^{\text{tr}} - \left(\frac{C_1}{12s^2} + \frac{KL_2C'_1}{32s^3} + \frac{KL_2^2}{320s^4} \right).$$

Step 4: Optimal Refinement Strategy. We aim to minimize the upper bound of the objective function

$$J(B) := \frac{B \log(2B/\delta)}{n_{\text{cal}} \left(\lambda_{f,c}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} \right)^2} = \frac{Ks \log(2Ks/\delta)}{n_{\text{cal}} \left(\lambda_{f,c}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} \right)^2}$$

with respect to s .

Due to analysis in Step 3, we have, for sufficiently large s ,

$$J(B) \leq \frac{Ks \log(2Ks/\delta)}{n_{\text{cal}} \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} - \frac{C_1''}{12s^2} \right)^2}$$

where $C_1'' = C_1 + \frac{12}{32}KL_2C_1' + \frac{12}{320}KL_2^2$. We roughly have,

$$J(B) \leq \frac{2\sqrt{2/\delta}(Ks)^{3/2}}{n_{\text{cal}} \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} - \frac{C_1''}{12s^2} \right)^2} =: \bar{J}(B)$$

Taking the logarithm and differentiating with respect to s :

$$\frac{d}{ds} \log \bar{J}(Ks) = \frac{3}{2s} - 2 \frac{d}{ds} \log \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} - \frac{C_1''}{12s^2} \right).$$

Since the value inside the logarithm cannot be negative, s must satisfy the following range:

$$s > \sqrt{\frac{C_1''}{12 \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} \right)}} \quad (\text{E.3.4})$$

Simplifying the terms:

$$\frac{3}{2s} - 2 \cdot \frac{\frac{C_1''}{6s^3}}{\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} - \frac{C_1''}{12s^2}} = \frac{1}{s} - \frac{\frac{C_1''}{3}}{s^3 \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} - \frac{C_1''}{12s^2} \right)}$$

Thus $\frac{d}{ds} \log \bar{J}(Ks) > 0$ when $\frac{1}{s} - \frac{\frac{C_1''}{3}}{s^3 \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} - \frac{C_1''}{12s^2} \right)} > 0$

Multiplying both sides by the positive quantity $s^3 \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} - \frac{C_1''}{12s^2} \right)$:

$$\begin{aligned} \iff s^2 \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} - \frac{C_1''}{12s^2} \right) - \frac{C_1''}{3} &> 0 \\ \iff \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} \right) s^2 - \frac{5}{12} C_1'' &> 0 \end{aligned}$$

Let s^* be the point where this expression equals zero. Since the LHS is increasing in s , LHS is negative for $s < s^*$ and positive for $s > s^*$. Therefore, s^* is minimum point of $\bar{J}(Ks)$ where

$$s^* = \sqrt{\frac{5C_1''}{12 \left(\lambda_{\infty}^{\text{tr}} - \frac{\lambda_{f_{0,c}}^{\text{tr}}}{2} \right)}}$$

which satisfies condition (E.3.4).

□

F EXPERIMENTAL DETAILS

F.1 DATASET DETAILS

Toy-10. We construct a synthetic $K=10$ class classification task in $d=32$ dimensions. Each class center is placed at radius $r=3$ along its own coordinate axis in $\mathbb{R}^{32}$. Samples are drawn from isotropic Gaussians with noise standard deviation $\sigma=1$ centered at each class mean. The training set ($n_{\text{train}} \approx 7,000$) and the validation set ($n_{\text{val}} \approx 3,000$) follow a uniform label distribution. The validation set is used for calibration and confusion matrix construction. The test stream ($n_{\text{test}}=2,000$) is generated under a constant label shift with target distribution

$$\pi_{\text{te}} = (0.25, 0.03, 0.12, 0.05, 0.10, 0.08, 0.04, 0.15, 0.06, 0.12).$$

The base classifier is a single-hidden-layer MLP with 64 hidden units and ReLU activation, trained by Adam optimizer with learning rate $= 10^{-2}$ and weight decay 10^{-2} for 50 epochs on the training set.

CIFAR-10. We use the standard CIFAR-10 dataset [Krizhevsky et al., 2009] with 50,000 training and 10,000 test images. The training set is split 70/30 into a model-training portion (35,000) and a validation set (15,000). The validation set is used for calibration and confusion matrix construction. The test stream uses all 10,000 test images under a constant label shift where one majority class (`cat`, class index 3) has probability 0.55 and all remaining classes have probability 0.05 each.

The base classifier is ResNet-18 [He et al., 2016] with ImageNet-pretrained weights from `timm`. We apply a two-phase fine-tuning procedure:

1. **Phase 1 (backbone frozen):** Only the final fully connected layer is trained with Adam ($\text{lr} = 10^{-3}$) for 3 epochs.
2. **Phase 2 (full fine-tune):** All layers are unfrozen and trained with Adam ($\text{lr} = 10^{-4}$) for 2 additional epochs.

The best checkpoint across both phases is selected by validation accuracy.

F.2 HISTOGRAM-BINNING CALIBRATION CONSTRUCTION

For our method, we construct a per-class equal-mass histogram binning calibration map on the validation set. For each predicted class $\hat{y} = k$, we sort the corresponding confidence scores $\max_k p_{\theta_0}(k | x)$ and compute s quantile-based bin edges. Duplicate edges are merged via `np.unique`.

F.3 STEP-SIZE SENSITIVITY.

Since our theoretical step-size yields an upper bound rather than a single computable value, and Baby et al. [2024] noted that theory-prescribed rates can be pessimistic, we search the γ from 0.1 downward. Similarly, LASU requires the step size ξ to be manually specified. We therefore conduct a sensitivity analysis for these two methods. Figure 3 shows the mean cross-entropy trajectories on CIFAR-10 under varying step sizes. For LASU, only $\xi = 0.001$ produces a stable trajectory. As discussed in 1, this instability can be attributed to the projected gradient ascent used by LASU: the gradient of the log-likelihood with respect to the label distribution estimate scales as $1/\hat{P}_{\text{te}}(y=k)$, so when a component $\hat{P}_{\text{te}}(y=k)$ becomes small, the gradient grows large and pushes the estimate further away from that class. In contrast, our method exhibits stable performance across the range of γ values tested.

F.4 COMPUTATIONAL RESOURCES

All experiments are implemented in Python 3.11 and run on NVIDIA GeForce RTX 3090 (24 GB) under Linux.

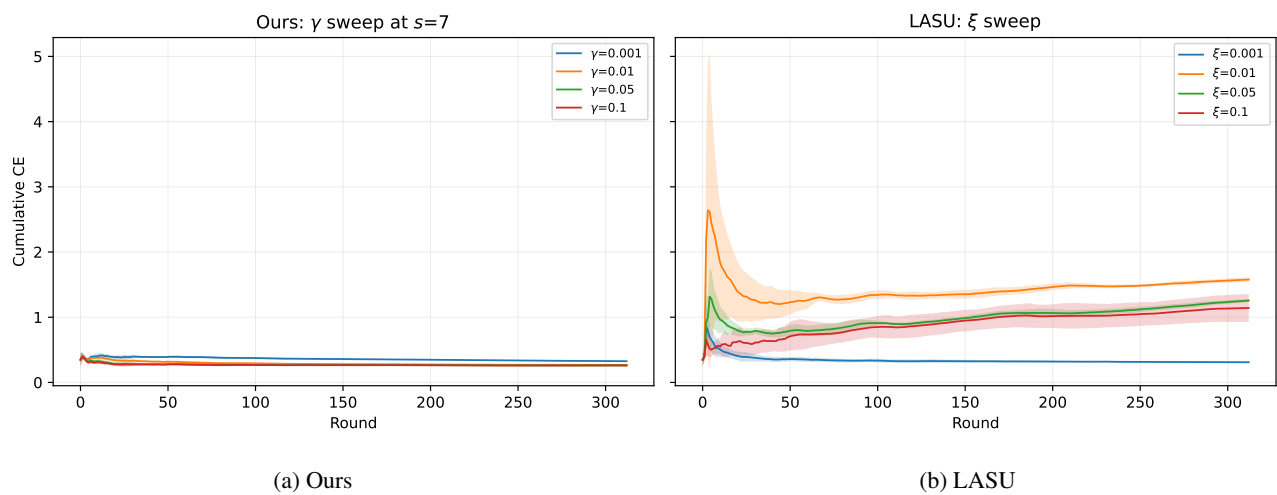


Figure 3: Mean cross-entropy trajectories on CIFAR-10 under varying step sizes.