Label Shift Estimation for Class-Imbalance Problem: A Bayesian Approach

Changkun Ye[1,2], Russell Tsuchida[2], Lars Petersson[2], and Nick Barnes[1]

[1]Australian National University, Canberra ACT Australia
[2]Data61 CSIRO, Acton ACT Australia

Abstract

As a type of distribution shift, label shift occurs when the source and target domains have different label distributions $\mathbb{P}(Y)$ but identical conditional distributions of data given labels $\mathbb{P}(X|Y)$. Under a Bayesian framework, we propose a novel Maximum A Posteriori (MAP) model and a novel posterior sampling model for the label shift problem. We prove the MAP objective admits a unique optimum and derive an EM algorithm that converges to the global optimum. We propose a novel Adaptive Prior Learning (APL) model to adaptively select prior parameters given data. We use the Markov Chain Monte Carlo (MCMC) method in our posterior sampling model to estimate and correct for label shift. Our methods can effectively resolve class imbalance problems on large-scale datasets without fine-tuning the classifier. Experiments show that our model outperforms existing methods on a variety of label shift settings. Our code is available at https://github.com/ChangkunYe/MAPLS/.

1. Introduction

In supervised learning tasks, the performance of a Neural Network classifier can decrease considerably under distribution shift between source and target domain [20]. As a type of distribution shift, label shift occurs when label distributions $\mathbb{P}(Y = \cdot)$ are different in the source and target domain while the data distribution conditioned on the label $\mathbb{P}(X = x|Y = \cdot)$ is preserved [22].

Under label shift, an optimal classifier on the source domain may no longer be optimal on the target domain [10]. Class imbalance problems can be modelled as label shift problems [36]. One extreme case is Long-Tailed classification, where $\mathbb{P}(X = x|Y = \cdot)$ is preserved while the train set has a Long-Tailed label distribution and the test set has an unknown label distribution. The classifier trained on the source domain has to be adjusted for optimal performance on the target domain [6, 28, 36]. Label shift studies the general case of arbitrary source and target label distributions, including the Long-Tailed classification case.

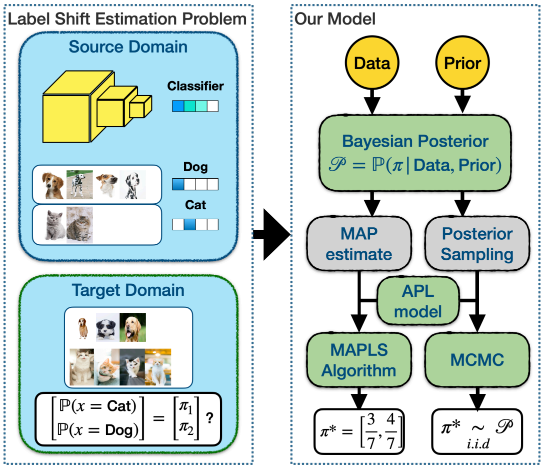
Label Shift Estimation Problem
Source Domain
Classifier
Dog
Cat
Target Domain
$\begin{bmatrix}\mathbb{P}(x = \text{Cat}) \\ \mathbb{P}(x = \text{Dog})\end{bmatrix} = \begin{bmatrix}\pi_1 \\ \pi_2\end{bmatrix}$?
Our Model
Data
Prior
Bayesian Posterior
$\mathscr{P} = \mathbb{P}(\pi \,|\, \text{Data, Prior})$
MAP estimate
Posterior Sampling
APL model
MAPLS Algorithm
MCMC
$\pi^* = \left[\frac{3}{7}, \frac{4}{7}\right]$
$\pi^* \underset{i.i.d}{\sim} \mathscr{P}$

Figure 1. The label shift estimation problem (left) and our proposed Bayesian approach (right). We construct the analytical Bayesian posterior of target label distribution given data and prior. Then based on our proposed APL model that adaptively learns prior parameters given data, we derive a MAPLS algorithm to obtain a MAP estimate of π and propose a posterior sampling model that uses MCMC to obtain samples from the posterior.

Three important problems arise due to label shift: *detection* — detect if label shift has occurred, *estimation* — estimate the target label distribution and *correction* — align the classifier to the target domain [10]. Here, we focus on estimation of the target domain label distribution. Estimation is usually based on labelled data from the source domain, a blackbox classifier and unlabelled target domain data [22].

Despite the good performance of existing label shift estimation approaches [2, 22, 32], two shortcomings hinder their application to real world problems. (1) Existing models are usually tested using source domains with uniform label distribution [1], the more realistic settings of long-tailed distributions [24] are rarely analyzed. (2) Effectiveness of existing models on large-scale datasets with many classes is rarely studied. Most evaluate on small scale datasets like CIFAR10 or MNIST [22, 34].

In this work, we observe that existing label shift estimation models may not perform well on large-scale datasets with large numbers of classes or highly imbalanced label distributions. To tackle this problem, we propose a novel label shift estimation model under a Bayesian framework. We construct the Bayesian posterior of the target label distribution parameters given data and a prior. We derive a novel EM algorithm to obtain an MAP estimate of the target label distribution. We further propose: a novel Adaptive Prior Learning (APL) model that adaptively chooses the prior parameters given data; and a posterior sampling model that uses MCMC to draw i.i.d samples from the posterior. To the best of our knowledge, Bayesian analysis has never been used in previous label shift estimation works.

We conduct extensive experiments on train and test set pairs with different label distributions. In contrast to previous methods that mainly focus on MNIST and CIFAR10, we evaluate our model on the CIFAR100, ImageNet, Places datasets and Long-Tailed versions of each dataset. For target label distributions, as well as evaluating under previous label shift estimation settings [22] with Dirichlet shift, we also introduce Long-Tailed benchmark test set shifts proposed in Long-Tailed classification [16]. Experimental results show that our model consistently outperforms existing models, particularly, obtaining better accuracy when the train set is highly imbalanced. These results demonstrate the applicability of our model to real world label shift tasks.

The contributions of our paper are as follows:

1. We propose a novel label shift model under a Bayesian framework to estimate and correct label shift without retraining the classifier. We construct the posterior of the target label distribution given data and a prior.
2. We derive a novel EM algorithm that computes the maxima of the posterior (MAP estimate) by minimising a strictly convex objective. We propose a novel Adaptive Prior Learning (APL) model to determine the parameters of the prior adaptively given data.
3. We propose a novel posterior sampling model to estimate and correct label shift based on i.i.d samples drawn from the posterior via MCMC.
4. Experiments show that our model consistently outperforms previous label shift estimation models in a variety of label shift settings on CIFAR100, ImageNet, Places and the Long-Tailed version of each dataset.

2. Related Works

2.1. Label Shift

Label Shift Estimation The problem of estimating target label distributions based on source domain data and unlabelled target domain samples is called label shift estimation [10]. Earlier works [8,33,42] require explicit modelling of the conditional probability $\mathbb{P}(X = x|Y = y)$, which is not feasible for high-dimensional data like images. Guo *et al*. [11] propose to construct a marginal distribution of data $\mathbb{P}(X = x)$, the target label distribution is estimated by matching the constructed distribution with a ground truth target domain distribution estimated by the unlabelled data.

In 2002, for high dimensional datasets, Saerens *et al*. [32] proposed an EM algorithm to obtain Maximum Likelihood Estimates (MLE) of the target label distribution, referred as Maximum Likelihood Label Shift (MLLS) [10]. Although MLLS is an old method, recently, Alexandari *et al*. [1] and Garg *et al*. [10] have shown its effectiveness over BBSE related methods on CIFAR10 and MNIST datasets.

More recently, Lipton *et al*. [22] proposed BlackBox Shift Estimation (BBSE) to first model the correlation between predicted labels from a blackbox classifier and the ground truth labels. The target label distribution is then predicted with the correlation and the unlabelled target samples. Based on BBSE, Tachet *et al*. [34] add non-negative constraints to the optimization objective of BBSE. Similarly, Azizzadeneshel *et al*. [2] developed Regularized Learning under Label Shift (RLLS) as a constrained BBSE model. Wu *et al*. [40] extend BBSE to a continuous learning setting with a target label distribution evolving with time.

Existing label shift estimation models focus more on the theoretical perspective of the problem rather than real world applications. Moreover, none of these works utilize Bayesian analysis in their models.

Label Shift Correction If the target domain label distribution is given, label shift correction models help align the existing classifier to the target domain. The approach can be performed either online during training or offline without retraining. Saerens *et al*. [32] propose an offline label shift correction (LSC) method to adjust the decision boundary of the classifier and correct for label shift avoiding the need for retraining. On the other hand, BBSE [22] and related methods also adopt an importance-weighted Empirical Risk Minimization (ERM) approach to retrain a new classifier for the target domain.

2.2. Class-Imbalance Problem

Class imbalance can lead to a decrease in classification performance if the source domain has an imbalanced label distribution [5, 39]. Recent works on class-imbalance usually aim to correct label shift for a Neural Network classifier with an imbalanced source label distribution and uniform target label distribution.

Re-Weighting and Re-Sampling Earlier works [12, 13, 17] re-weight the training loss or up-sample rare classes to create a class-balanced train set. The re-weighting approach is similar to importance-weighted ERM proposed by BBSE [22]. However, these methods have been shown to overfit rare classes [9] on highly imbalanced train sets.

Offline Correction Recently, several works introduce

LSC to correct the classifier for the target domain. For example, Tian *et al.* [36] proposed to combine the original classifier and an LSC corrected classifier for class-imbalance. LADE [16] proposed to learn a better classifier and the correct classifier for a uniform test set with LSC.

Other Advanced Models Other recent works on the class-imbalance problem propose more complicated mechanisms to obtain better performance. However, these models usually have less flexibility to adjust for different target label distributions. For example, LDAM [6] proposes a loss that aims to minimize the generalization error bound on a uniform test set. OT [28] proposes an optimal transport algorithm to optimize a classifier for a uniform test set. These models require retraining or algorithmic adjustment for a different target label distribution.

3. Preliminaries

3.1. Problem Setup

We use similar notation to [10]. We denote the input image space as $\mathcal{X} \subseteq \mathbb{R}^{H \times W \times C}$ where H, W, C are the height, width and channels of the image, and the corresponding label space as $\mathcal{Y} = \{1, 2, ...K\}$, where K is the number of classes. The random variable of image and label pairs on source and target domains are denoted as $(X_s, Y_s) \sim P_s$ and $(X_t, Y_t) \sim P_t$ respectively.

Under the label shift setting, it is assumed that the source and target domain have a different label distribution $\mathbb{P}(Y_s = i) \neq \mathbb{P}(Y_t = i)$. The conditional probability of an image given its label is identical [22]:

$$\mathbb{P}(X_s = x | Y_s = i) = \mathbb{P}(X_t = x | Y_t = i). \tag{1}$$

Three main problems are usually discussed in label shift, namely *detection*, *estimation* and *correction* [10]. In this work, we focus on label shift estimation. To tackle label shift estimation for classification, a blackbox classifier $f : \mathcal{X} \to \Delta^{K-1}$ is usually assumed to be available, which is sometimes required to perform well on the source domain [10, 23]. Here Δ^{K-1} is the space of a K dimensional probability simplex.

Note that the source label $Y_s \sim \mathrm{Cat}(K, \mathbf{c})$ and target label $Y_t \sim \mathrm{Cat}(K, \boldsymbol{\pi})$ each follow a categorical distribution over K classes. $\mathbf{c}, \boldsymbol{\pi}$ are parameters of the two distributions respectively, with $\mathbf{c}, \boldsymbol{\pi} \in \Delta^{K-1}$. Estimating the target label distribution $\mathbb{P}(Y_t = \cdot) = \boldsymbol{\pi}$ is equivalent to estimating the parameters $\boldsymbol{\pi}$ of the categorical distribution $\mathrm{Cat}(K, \boldsymbol{\pi})$.

3.2. MLLS Label Shift Estimation

Saerens *et al.* [32] derive MLLS by assuming the classifier reflects the true conditional probability $f(x)_j = \mathbb{P}(Y_s = j | X_s = x), j = 1, 2...K$. The source domain label distribution $\mathbb{P}(Y_t = i) = \mathbf{c}$ can be estimated using labelled data. With unlabelled target domain data $\mathbb{X}$, MLLS estimates the target label distribution $\mathbb{P}(Y_t = i) = \boldsymbol{\pi}$ by maximizing the log likelihood (2):

$$\log L(\boldsymbol{\pi}; \mathbb{X}) := \log \left(\prod_{i=1}^{N} \mathbb{P}(X_t = x_i | \boldsymbol{\pi}) \right) \tag{2}$$

using the EM algorithm. We provide more discussions on basics and advantages of EM algorithm in Appendix A.

Label Shift Problem Setup	
Given	$\{x_i^s, y_i^s\}_{i=1}^{N^s}$, where $(x_i^s, y_i^s) \sim_{i.i.d} P_s$ $f : \mathcal{X} \to \Delta^{K-1}$ $\mathbb{X} = \{x_i\}_{i=1}^{N}$, where $(x_i, \cdot) \sim_{i.i.d} P_t$
Detection	If $\mathbb{P}(Y_s = \cdot) \neq \mathbb{P}(Y_t = \cdot)$?
Estimation	$\mathbb{P}(Y_t = \cdot) = ?$
Correction	$\arg\max_g \mathbb{P}(g(X_t; f) = Y_t)$

Table 1. **Label Shift problem setup.** The available data is: labelled samples from source domain P_s, a classifier f, and unlabelled target domain data $\mathbb{X}$. Label shift *detection* tests if label shift occurs, *estimation* estimates the target label distribution and *correction* adopt the classifier for the target domain.

In the algorithm, the parameter $\boldsymbol{\pi}$ is first initialized as $\boldsymbol{\pi}^{(0)}$. The EM algorithm then proceeds by repeatedly applying two alternating steps, the E-Step and the M-Step. **E-Step:** the model evaluates the conditional probability $g(x_i; \boldsymbol{\pi}^{(t)})_j := \mathbb{P}(Y_t = j | X_t = x_i, \boldsymbol{\pi}^{(t)})$ under label shift with:

$$g(x_i; \boldsymbol{\pi}^{(t)})_j = \frac{\frac{\pi_j^{(t)}}{c_j} f(x_i)_j}{\sum_{l=1}^{K} \frac{\pi_l^{(t)}}{c_l} f(x_i)_l}. \tag{3}$$

Equality has been proved by Saeren *et al.* [32] under label shift along with the assumptions of MLLS.

M-Step: $\boldsymbol{\pi}^{(t+1)}$ for next iteration can be obtained via:

$$\pi_j^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} g(x_i; \boldsymbol{\pi}^{(t)})_j. \tag{4}$$

Given $\boldsymbol{\pi}^{(0)}$, the iterative procedure of the EM algorithm is repeated until numerical convergence to obtain the MLE of the target label distribution $\mathbb{P}(Y_t = i) = \pi_i$.

3.3. Offline Label Shift Correction

For classifier, f, trained on the source domain P_s, Saerens *et al.* [32] proposed to construct a new target domain classifier $g : \mathcal{X} \to \Delta^{K-1}$ to correct for label shift:

$$g(x)_j = \frac{\frac{\pi_j}{c_j} f(x)_j}{\sum_{l=1}^{K} \frac{\pi_l}{c_l} f(x)_l} \tag{5}$$

where $c_j, \pi_j, j = 1, 2...K$ are parameters of source and target label distributions respectively.

The advantage of this model is that adjustment does not require retraining of f — see [29, 34, 36] for more theoretical discussions.

4. Proposed Method

In this work, we propose a novel Bayesian approach for the label shift estimation problem. By employing a prior distribution over target label distribution $\mathbb{P}(Y_t = \cdot) = \boldsymbol{\pi}$, we obtain the posterior of $\boldsymbol{\pi}$ given available data $\mathbb{X}$. Based on the posterior, we derive an EM algorithm to obtain the Maximum *A Posteriori* (MAP) estimate of $\boldsymbol{\pi}$. To utilize the information of the entire posterior, we also propose to use Hamiltonian Monte-Carlo (HMC) method to obtain i.i.d samples from the posterior.

The categorical distribution $Y_t \sim \text{Cat}(K, \boldsymbol{\pi})$ requires that the prior distribution over $\boldsymbol{\pi}$ is supported on Δ^{K-1}. K dimensional Dirichlet distributions satisfy this constraint, and are often used as a prior over parameters of categorical distributions [19, 37]. Therefore, we employ a Dirichlet prior over the parameters $\boldsymbol{\pi} \sim \text{Dir}(K, \boldsymbol{\alpha})$ of the target label distribution $\text{Cat}(K, \boldsymbol{\pi})$, where $\boldsymbol{\alpha} \in \mathbb{R}^K_{>1}$. With the Dirichlet prior as $\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha})$ and unlabelled target domain samples $\mathbb{X}$, we can write the posterior of $\boldsymbol{\pi}$ given $\mathbb{X}$ and $\boldsymbol{\alpha}$ as:

$$\mathbb{P}(\boldsymbol{\pi}|\mathbb{X}, \boldsymbol{\alpha}) = \frac{1}{Z}\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_{i=1}^{N} \mathbb{P}(X_t = x_i|\boldsymbol{\pi}), \tag{6}$$

where $Z = \int \mathbb{P}(\mathbb{X}|\boldsymbol{\pi}')\mathbb{P}(\boldsymbol{\pi}'|\boldsymbol{\alpha}))d\boldsymbol{\pi}'$ is a constant w.r.t $\boldsymbol{\pi}$.

The marginal distribution $\mathbb{P}(X_t = x|\boldsymbol{\pi})$ can be rewritten as a combination of known expressions. Given the source domain labelled data, we can estimate the source domain label distribution $\mathbb{P}(Y_s = j) = c_j$ in $Y_s \sim \text{Cat}(K, \mathbf{c})$, which is also a categorical distribution. $\mathbb{P}(Y_s = j|X_s = x_i)$ on the source domain can be modelled by the blackbox classifier f, and the target label distribution is $\mathbb{P}(Y_t = j|\boldsymbol{\pi}) = \pi_j$. Formally we are given:

$$\begin{aligned} \mathbb{P}(Y_s = j) &= c_j > 0 \\ \mathbb{P}(Y_s = j|X_s = x) &= f(x)_j \\ \mathbb{P}(Y_t = j|\boldsymbol{\pi}) &= \pi_j, \end{aligned} \tag{7}$$

where $c_i > 0, i = 1, 2...K$ because each class has non-zero sample frequency on the source domain.

In (7), we assume the classifier f is well-specified to model $\mathbb{P}(Y_s = \cdot|X_s = x)$. We further discuss in Section 4.2 when this may not be the case in practice. With (7) available, utilizing Bayes rule, we can rewrite the posterior (6) as:

$$\mathbb{P}(\boldsymbol{\pi}|\mathbb{X}, \boldsymbol{\alpha}) = \frac{1}{Z}\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_{i=1}^{N}\sum_{j=1}^{K} \mathbb{P}(X_t = x_i)\frac{\pi_j}{c_j} f(x_i)_j. \tag{8}$$

Note that $\mathbb{P}(X_t = x_i)$ and Z are constants w.r.t $\boldsymbol{\pi}$ and $\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha})$ is the Dirichlet prior. Therefore the analytical expression for the un-normalized posterior $\mathbb{P}(\boldsymbol{\pi}|\mathbb{X}, \boldsymbol{\alpha})$ can be obtained from (8).

4.1. MAP estimate

We first derive an EM algorithm to obtain MAP estimate of $\boldsymbol{\pi}$. By definition, any MAP estimate $\boldsymbol{\pi}^*$ minimizes the negative log posterior:

$$\boldsymbol{\pi}^* \in \underset{\boldsymbol{\pi} \in \Delta^{K-1}}{\arg\min} - \log \mathbb{P}(\boldsymbol{\pi}|\mathbb{X}, \boldsymbol{\alpha}) \tag{9}$$

We prove that the optimization problem (9) is strictly convex in $\boldsymbol{\pi}$ and propose a novel EM algorithm to find $\boldsymbol{\pi}^*$. We name our proposed algorithm: Maximum *a Posteriori* Label Shift (MAPLS).

Proposition 1 *Under label shift defined in* (1), *suppose* (7) *holds for* $\forall(x, i) \in \mathcal{X} \times \mathcal{Y}$. *Let* $\boldsymbol{\pi} \sim Dir(K, \boldsymbol{\alpha})$ *with* $\boldsymbol{\alpha} \in \mathbb{R}^K_{>1}$. *Then in* (9), *the objective is strictly convex in* $\boldsymbol{\pi}$, $\boldsymbol{\pi}^*$ *is unique and EM Algorithm 1 converges to* $\boldsymbol{\pi}^*$.

```
Algorithm 1 MAPLS
  Input: Target domain $\{x_i|i = 1, 2, ...N, \{x_i, \cdot\} \sim P_t\}$, source domain $\mathbb{P}(Y_s = j) = c_j$, classifier $f(x)$ and Dirichlet prior $\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha})$.
  Initialize: $\boldsymbol{\pi}^{(0)} \in \Delta^{K-1}_{>0}$.
  for $t = 0$ to $T$ do
    E-step Evaluate $g(x_i; \boldsymbol{\pi}^{(t)})_j$:
        $g(x_i; \boldsymbol{\pi}^{(t)})_j = \frac{\frac{\pi_j^{(t)}}{c_j} f(x_i)_j}{\sum_{l=1}^{K} \frac{\pi_l^{(t)}}{c_l} f(x_i)_l}.$   (10)
    M-step Obtain $\boldsymbol{\pi}^{(t+1)}$ with:
        $\pi_j^{(t+1)} = \frac{\sum_{i=1}^{N} g(x_i; \boldsymbol{\pi}^{(t)})_j + \alpha_j - 1}{N + \sum_{l=1}^{K}(\alpha_l - 1)}.$   (11)
  end for
  Output: $\mathbb{P}(Y_t = \cdot) = \boldsymbol{\pi}^{(T+1)}$
```

The detailed proof can be found in Appendix B.1, B.2. Algorithm 1 can seen as a generalization of MLLS. In the M-Step, we can rewrite (11) as:

$$\pi_j^{(t+1)} = \lambda \underbrace{\frac{\sum_{i=1}^{N} g(x_i; \boldsymbol{\pi}^{(t)})_j}{N}}_{\text{Data contribution}} + (1 - \lambda) \underbrace{\frac{\alpha_j - 1}{\sum_{l=1}^{K}(\alpha_l - 1)}}_{\text{Prior contribution}} \tag{12}$$

where $\lambda \in (0, 1)$ has the form:

$$\lambda = \frac{N}{N + \sum_{l=1}^{K}(\alpha_l - 1)}. \tag{13}$$

As $\lambda \to 1^-$, the algorithm degenerates to MLLS. As $\lambda \to 0^+$, the MAP estimate will converge to the Dirichlet prior $\text{Dir}(K, \boldsymbol{\alpha})$. In this manner, λ can be seen as our confidence in our label distribution estimation.

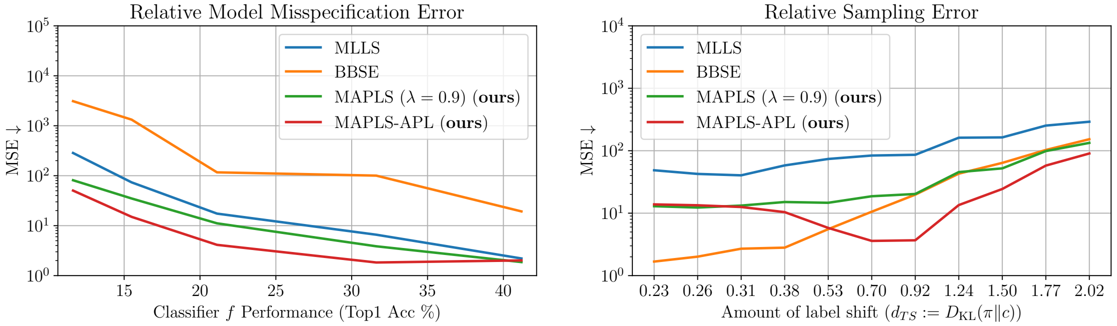

Figure 2. **Label shift estimation error analysis.** The Mean Square Error (MSE, see Section 5.3) increases when: (1) the model is misspecificed, i.e. $\mathbb{P}(Y_s = \cdot | X_s = x) = f(x)$ is not satisfied (left); (2) the sampling error gets magnified when source and target domains have large label shift (right). Our MAPLS with fixed prior ($\lambda = 0.9$ in (14)) can reduce both errors compared with MLLS. Our MAPLS-APL model with prior parameters learned given data can further reduce MSE and outperform BBSE under large label shift (right).

The choice of $\boldsymbol{\alpha}$ and corresponding λ affect the MAP estimate $\boldsymbol{\pi}^*$. In practice, it is important to determine an appropriate $\boldsymbol{\alpha}$ and λ to give a good MAP estimate $\boldsymbol{\pi}^*$ for the target label distribution $\mathbb{P}(Y_t = \cdot)$.

After obtaining $\boldsymbol{\pi}^*$, we can use (5) to correct the source domain classifier f to the target domain under label shift.

Symmetric Dirichlet Prior: The Dirichlet prior possesses K parameters in $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_K] \in \mathbb{R}^K_{>1}$. When no information about the target domain label distribution is available, we may set $\alpha_j = \alpha_0$. This has the advantage of reducing the number of parameters to be chosen, at the cost of limiting expressivity.

- *The Dirichlet prior satisfies* $\boldsymbol{\pi} \sim Dir(K, \alpha_0 \mathbf{1})$.

Then the M-Step of the MAPLS algorithm in the form of (12) can be further simplified as:

$$\pi_j^{(t+1)} = \lambda \frac{\sum_{i=1}^N g(x_i; \boldsymbol{\pi}^{(t)})_j}{N} + (1 - \lambda) \frac{1}{K} \tag{14}$$

where $\lambda = N/(N + K(\alpha_0 - 1))$ also has a simpler form.

The MAPLS algorithm with $\lambda \to 0^+$ will converge to a uniform categorical distribution with $\boldsymbol{\pi} = \mathbf{1}/K$ in $Y_t \sim \text{Cat}(K, \boldsymbol{\pi})$. Note that now $\boldsymbol{\alpha} = \alpha_0 \mathbf{1}$ is fully determined by λ, and we can determine parameter α_0 in the prior by selecting a value for λ. In this case, $1 - \lambda$ represents the strength of regularization in the MAP estimation procedure.

4.2. Adaptive Prior Learning Model

In our MAPLS algorithm 1, the prior parameter $\boldsymbol{\alpha}$ should be determined before the estimation of $\boldsymbol{\pi}$. In this work, based on the analysis of the possible estimation error, we propose a novel Adaptive Prior Learning (APL) model to adaptively learn $\boldsymbol{\alpha}$ given available data. Our model is inspired by the empirical Bayesian [7, 30] approach.

Estimation Error Analysis: Intuitively, two factors can induce estimation error in our posterior. Firstly, we use a classifier $f(x)$ to model ground truth $\mathbb{P}(Y_s = \cdot | X_s = x)$ in (7), when the classifier fails to represent the ground truth, the model is subject to misspecification error. Secondly, even if (7) is satisfied, our MAPLS model will have an associated sampling error due to using a finite number of samples, like other models [10].

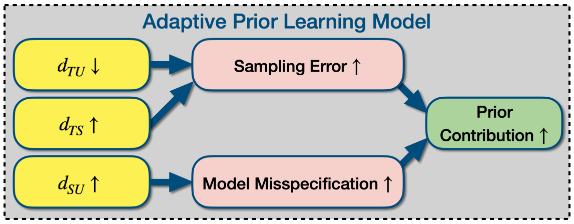

Figure 3. **Structure of our Adaptive Prior Learning model.** The parameter in the prior is adaptively determined by the available data with a heuristic based on d_{TU}, d_{TS}, d_{SU} defined in (15).

Adaptive Prior Learning: In our APL model, we propose to use a heuristic to loosely evaluate the magnitude of model misspecification error. The sampling error of the label shift estimation model can be magnified with large label shift between the target and source domains (e.g. Fig. 2). Therefore, our APL model also includes a heuristic to mitigate sampling error.

Practically, we first run MAPLS with $\lambda = 1$ to obtain an initial MLE of the target label distribution $\boldsymbol{\pi}^{MLE}$. Then our APL model quantifies the two estimation errors based on the three KL-divergences below:

$$\begin{cases} d_{SU} := D_{\text{KL}}(\mathbf{c} \| \mathbf{1}/K) \\ d_{TU} := D_{\text{KL}}(\boldsymbol{\pi}^{MLE} \| \mathbf{1}/K) \\ d_{TS} := D_{\text{KL}}(\boldsymbol{\pi}^{MLE} \| \mathbf{c}) \end{cases} \tag{15}$$

where $\mathbf{1}/K$ denotes a uniform label distribution and $\mathbf{c}$ is the parameter of source label distribution. S, T, U represents source, target and uniform label distribution respectively.

Adapt to model misspecification: A Neural Network classifier f trained on the source domain usually has poor performance when the source domain has a highly imbalanced label distribution ($d_{SU} \gg 0$) [6]. In this case, the classifier is more likely to be subject to model misspecification error when estimating label shift. Hence we increase prior contribution in (14) with higher d_{SU}.

Adapt to sampling error: We propose two approaches to mitigate the problem that sampling error tends to increase given large label shift. Firstly, we use d_{TS} to approximate the amount of shift between target and source label distribution. A higher d_{TS} implies larger label shift, which will lead to more severe sampling error. Thus our APL model should increase the contribution of prior in (14) with higher d_{TS}. Secondly, when $\boldsymbol{\pi}^{MLE}$ is close to a uniform label distribution $\mathbf{1}/K$, we also propose to increase the prior contribution so that (14) can push the estimate more towards $\mathbf{1}/K$.

Overall APL model: By defining a normalization function $F(x) = x/(1+x)$, our APL model determines λ via:

$$\lambda = a \cdot F(\gamma \cdot d_{TU}) + (1-a) \cdot (1 - F(\gamma \cdot d_{TS})), \tag{16}$$

where $\gamma = 1 - F(b \cdot d_{SU})$ takes into account the model misspecification error and d_{TU}, d_{TS} evaluates the sampling error. Here $a \in [0,1]$ represents the trade-off between the two approaches to reduce sampling error and $b \in [0,1]$ represents the strength of misspecification error. We provide more discussion of our APL model in Appendix D.

4.3. Sampling from the Bayesian posterior

Apart from the point estimate $\boldsymbol{\pi}^*$, we also use Bayesian analysis to utilize the entire posterior $\mathbb{P}(\boldsymbol{\pi}|\mathbb{X},\boldsymbol{\alpha})$ as our estimated target label distribution. In this work, we propose to use the Markov Chain Monte Carlo (MCMC) method to obtain i.i.d samples of the posterior. The samples can then be used for downstream label shift correction tasks.

Based on (8), we can rewrite $\mathbb{P}(\boldsymbol{\pi}|\mathbb{X},\boldsymbol{\alpha})$ as:

$$\mathbb{P}(\boldsymbol{\pi}|\mathbb{X},\boldsymbol{\alpha}) = \frac{1}{Z}\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha})\prod_{i=1}^{N}\sum_{j=1}^{K}\frac{\pi_j}{c_j}f(x_i)_j, \tag{17}$$

where Z contains $\int \mathbb{P}(\mathbb{X}|\boldsymbol{\pi}')\mathbb{P}(\boldsymbol{\pi}'|\boldsymbol{\alpha}))d\boldsymbol{\pi}'$ and $\mathbb{P}(X_t = x_i)$, which are constant w.r.t $\boldsymbol{\pi}$ and are usually intractable.

To avoid evaluation of Z, we adopt the MCMC method to obtain samples of the posterior. Because the Hamiltonian Monte-Carlo (HMC) sampler can be more efficient than other MCMC methods in high dimensional space [3, 26] , we adopt the HMC to obtain i.i.d samples of the posterior:

$$\Pi = \{\boldsymbol{\pi}^i\}_{i=1}^{L}, \text{ where } \boldsymbol{\pi}^i \sim_{i.i.d} \mathbb{P}(\boldsymbol{\pi}|\mathbb{X},\boldsymbol{\alpha}), \tag{18}$$

where $\boldsymbol{\alpha}$ is determined by our APL model.

After collecting Π, each $\boldsymbol{\pi}^i$ is used as a point estimate of $\boldsymbol{\pi}$ for the down stream task. For example, for the label shift correction problem, we use every $\boldsymbol{\pi}^i \in \Pi$ to correct the source domain classifier $f(x)$ to the target domain $g_i(x)$ under label shift based on (5). The target domain average SoftMax classifier can then be constructed as:

$$g(x)_j = \sum_{i=1}^{L}\frac{1}{L}\frac{\frac{\pi^i_j}{c_j}f(x)_j}{\sum_{l=1}^{K}\frac{\pi^i_l}{c_l}f(x)_l}. \tag{19}$$

With samples of the posterior, the uncertainty of our estimated $\boldsymbol{\pi}$ given data can also be analyzed. Comparing with Algorithm 1, this approach utilizes the entire posterior at the cost of computation resources.

Remark: MCMC can be computationally expensive in high dimensional space, because sufficient warm up steps are required if the Markov chain is initialized randomly in value space [25]. Fortunately, since the posterior in our model is strictly log concave (Proposition 1) with known maximal point $\boldsymbol{\pi}^*$ obtained by MAPLS, we can initialize the Markov chain at $\boldsymbol{\pi}^*$ and the HMC sampler can then collect i.i.d samples more efficiently without warm up steps.

4.4. Estimation of Source Label Distribution

Given source domain data $\{x_i^s, y_i^s\}_{i=1}^{N^s}$ and blackbox classifier f, there are two known methods to estimate the source domain label distribution $\mathbb{P}(Y_s = \cdot) = \mathbf{c}$. MLE is the standard method to estimate $\mathbf{c}$ with source domain ground truth labels y_i^s. On the other hand, when classifier f is calibrated on the source domain, Alexandari *et al.* [1] also proposed to estimate $\mathbf{c}$ with source domain images in $\{x_i^s, y_i^s\}_{i=1}^{N^s}$ and classifier f with

$$c_j = \frac{1}{N^s}\sum_{i=1}^{N^s} f(x_i^s)_j. \tag{20}$$

In this work, we adopt both approaches to estimate $\mathbf{c}$. We name the MLE approach as the “hard” method and (20) as the “soft” method.

4.5. Overall Method

We propose to estimate and correct label shift as follows:

Algorithm 2 Overall Method

```
Input: Source domain data $\{x_i^s, y_i^s\}_i^{N^s}$, classifier $f : \mathcal{X} \to \Delta^{K-1}$ and target domain data $\mathbb{X} = \{x_i|(x_i, \cdot) \sim P_t, i = 1, 2, ...N\}$.
Parameter Determination:
  • c: Use MLE or (20) to estimate $P(Y_s = \cdot) = \mathbf{c}$.
  • $\boldsymbol{\pi}^{MLE}$: Use MAPLS 1 to obtain $\boldsymbol{\pi}^{MLE}$ ($\alpha_0 = 1$).
  • $\alpha_0$: Determine $\lambda$ with APL model.
With $\alpha_0$ and (14), use MAPLS 1 to obtain $\boldsymbol{\pi}^*$.
if Point Estimate then
  Correct Label Shift: Obtain $g$ with (5).
else if Posterior Sampling then
  Use HMC (initialized with $\boldsymbol{\pi}^*$) to obtain $\Pi$ in (18).
  Correct Label Shift: Obtain $g$ with (19).
end if
Output: Target domain classifier $g$.
```

We name the model that uses the MAP estimate MAPLS-APL and the model that uses posterior sampling PSLS-APL, where the “APL” indicates that parameter $\boldsymbol{\alpha}$ in the prior distribution is learned with our APL model.

5. Experiments

5.1. Datasets

We evaluate our model on the CIFAR100 [21], ImageNet 2012 [31] and Places2 [44] datasets. Following common use in Long-Tailed research [6, 38, 43], we also use Long-Tail versions of ImageNet, Places [24] and CIFAR100.

We test the models on test sets with Dirichlet shift proposed by previous label shift estimation models [1, 22]. Dirichlet Shift generates a random test set label distribution from a K dimensional Dirichlet distribution. We also adopt the ordered Long-Tailed shifted test set used in LADE [16], which has the same or inverse order of the Long-Tailed distributed train set. We further extend this setting to a shuffled Long-Tailed test set, where the test set still has a Long-Tailed label distribution but with random class order.

	Dataset	**Setup**
Train Set	CIFAR100 [21]	Original, Long-Tailed with $R = \{2, 5, 10, 20, 50, 100, 200\}$
	ImageNet [31]	Original, Long-Tailed [24]
	Places [44]	Original, Long-Tailed [24]
	Test Shift Type	**Params**
Test Set	Original	None
	Dirichlet [22]	$\alpha = 1.0, 10$
	Ordered Long-Tail [16]	$R = \{2, 5, 10, 50\}$ Order = "Forward", "Backward"
	Shuffled Long-Tail	$R = \{2, 5, 10, 50\}$

Table 2. **Label shift experiment settings.** R is referred to as the imbalance ratio — the ratio of maximum and minimum sample number per class respectively in test set. α is the parameter of the Dirichlet distribution.

5.2. Model Setup

Both our MAPLS/MAPLS-APL algorithm and previous MLLS algorithm are initialized with $\boldsymbol{\pi}^{(0)} = \mathbf{c}$ and run for 100 epochs to ensure convergence. Because $\boldsymbol{\pi}^*$ is unique as proved in Proposition 1, our MAPLS is guaranteed to converge to a single MAP estimate. In our APL model, we empirically set $a = 0.9, b = 0.5$ in (16) for all the label shift settings in all datasets. For our PSLS model, use a HMC sampler called No-U-Turn Sampler [15] provided by Pyro [4] to collect 5000 samples from the posterior.

We implement the Neural Network classifiers using PyTorch [27]. We use the ResNet32 [18] classifier for CIFAR100 and every CIFAR100-LT dataset. We use pre-trained ResNet50 [14] and pre-trained Resnet152 for ImageNet and Places datasets respectively. We train a ResNet50 and ResNet152 for ImageNet-LT and Places-LT datsets respectively. More details of classifier implementations can be found in Appendix E.1.

5.3. Evaluation Metrics

We follow previous methods [1, 2, 22] to evaluate label shift estimation performance with $(\mathbf{w} - \hat{\mathbf{w}})^2/K$, where $w_i = \mathbb{P}(Y_t = i)/\mathbb{P}(Y_s = i), i = 1, 2...K$ is the target over the source label distribution ratio. $\mathbf{w}$ is the ground truth ratio estimated by the source and target labels. $\hat{\mathbf{w}}$ is the predicted ratio with $\mathbb{P}(Y_t = \cdot)$ estimated by each model.

Test Set Train Set	Ordered LT	Shuffled LT	Dirichlet
CIFAR100/CIFAR100-LT	55%	50%	52%
ImageNet/ImageNet-LT	82%	90%	75%
Places/Places-LT	68%	90%	92%

Table 3. **SOTA comparison summary of estimation error.** The percentage of settings that our MAPLS-APL model outperforms SOTA models (MLLS, BBSE, RLLS) in terms of $(w - \hat{w})^2/K$.

Test Set Train Set	Ordered LT	Shuffled LT	Dirichlet
CIFAR100/CIFAR100-LT	56%	58%	58%
ImageNet/ImageNet-LT	59%	80%	58%
Places/Places-LT	59%	80%	75%

Table 4. **SOTA comparison summary of Top1 Accuracy.** The percentage of settings that our MAPLS-APL model outperforms SOTA models and the baseline classifier in terms of accuracy.

We also provide Top1 accuracy for different label shift estimation models with LSC (5) on all datasets. The result summary is available in Tab. 4 and detailed results are reported in Appendix F,G,H, for CIFAR100/CIFAR100-LT, ImageNet/ImageNet-LT and Places/Places-LT respectively.

5.4. State-of-the-art Comparison

We compare the performance of our method with several state-of-the-art (SOTA) label shift estimation methods, including MLLS [1, 32], BBSE [22] and RLLS [2]. In BBSE and RLLS, there are also "soft" and "hard" versions of each model. We evaluate performance of these models with previously available implementation (details in Appendix E.2).

In the setting of large-scale datasets, methods that require retraining the classifier on the source domain will suffer from high computational cost. Therefore we have not reproduced and reported Tachet *et al*. [34] in our results.

We provide the SOTA comparison of our MAPLS-APL model in terms of $(\mathbf{w} - \hat{\mathbf{w}})^2/K$ in Tab. 3 and Top1 Accuracy in Tab. 4. More details are discussed in Appendix C. Note that unlike SOTA models that obtain a point estimate of $\boldsymbol{\pi}$, our PSLS-APL model obtains samples of $\boldsymbol{\pi}$ from the posterior instead. Thus only Top1 Accuracy is compared for our PSLS-APL model (Tab. 6) instead of both metrics.

As shown in Tab. 3, our MAPLS-APL model outperforms SOTA models in at least 50% of the label shift and dataset settings. As an example on ImageNet in Tab. 5, our model outperforms other models by a large margin for the highly imbalanced train set ImageNet-LT.

As shown in Tab. 4, in terms of Top1 Accuracy, our MAPLS-APL model outperforms SOTA models and baseline in at least 50% of the settings. As an example in Tab. 6, our MAPLS-APL and PSLS-APL model have similar performance and outperform SOTA models in most settings.

By analyzing the performance in Tab. 5, 6, one obvious advantage of our model is its robustness to the source label

Dataset	ImageNet									ImageNet-LT								
Shift Type	Shuffled LT					Dirichlet				Shuffled LT					Dirichlet			
Params	50	25	10	5	2	$\alpha = 10.0$		$\alpha = 1.0$		50	25	10	5	2	$\alpha = 10.0$		$\alpha = 1.0$	
Test sample No.	fixed	fixed	fixed	fixed	fixed	12500	25000	12500	25000	fixed	fixed	fixed	fixed	fixed	12500	25000	12500	25000
MLLS-hard	0.1210	0.1102	0.1001	0.0868	0.0766	0.1111	0.0848	0.1299	0.1113	36.09	34.49	30.57	26.90	24.42	28.44	26.03	38.21	36.18
MLLS-soft	**0.1121**	0.0972	0.0868	0.0721	0.0637	0.0981	0.0721	**0.1154**	**0.0977**	80.66	82.10	84.92	81.54	76.59	91.28	83.62	82.62	84.23
BBSE-hard	0.1285	0.1020	0.0871	0.0699	0.0581	0.0869	0.0661	0.1285	0.1173	$3.2e^5$	$1.8e^6$	$1.4e^6$	$2.0e^7$	$4.8e^5$	$4.8e^5$	$1.0e^7$	$1.7e^6$	$1.2e^{10}$
BBSE-soft	0.1305	0.1086	0.0969	0.0790	0.0671	0.1052	0.0769	0.1366	0.1177	28.00	25.48	18.04	15.86	12.07	13.84	12.89	28.30	27.75
RLLS-hard	1.1450	0.7160	0.4436	0.2244	0.0473	0.1159	0.1122	1.1020	1.0607	45.00	38.77	29.99	24.18	19.96	21.98	21.05	46.05	45.75
RLLS-soft	1.1450	0.7160	0.4436	0.2244	0.0473	0.1159	0.1122	1.1020	1.0607	45.00	38.77	29.98	24.18	19.96	21.98	21.05	46.05	45.75
MAPLS-APL-hard (**Ours**)	0.1236	0.1006	0.0816	0.0633	0.0482	0.0736	0.0570	0.1283	0.1142	20.25	16.62	**10.26**	**6.18**	**2.62**	**4.72**	**3.86**	21.16	20.68
MAPLS-APL-soft (**Ours**)	0.1144	**0.0904**	**0.0710**	**0.0521**	**0.0370**	**0.0628**	**0.0465**	0.1160	0.1025	**19.48**	**16.39**	11.23	7.58	4.43	6.62	5.66	**18.94**	**18.75**

Table 5. **Performance of $(\mathbf{w} - \hat{\mathbf{w}})^2/K$ (↓) on the ImageNet and ImageNet-LT datasets**, with shuffled Long-Tailed test set that have an imbalance ratio $\{50, 10, 5, 2\}$ and Dirichlet test set that have $\alpha = \{1, 10\}$ and total test sample number $\{12500, 25000\}$ in each setting. Best performances are in bold face and second best are in blue. Our PSLS-APL model is only suitable for Top1 Accuracy comparison.

Dataset	ImageNet-LT									Places-LT								
Order	Forward				Uniform	Backward				Forward				Uniform	Backward			
Imbalance Ratio	25	10	5	2	1	2	5	10	25	25	10	5	2	1	2	5	10	25
Baseline	**62.55**	**58.48**	54.97	49.59	45.31	40.94	35.22	31.31	26.56	41.25	38.04	35.11	31.06	27.92	24.76	20.89	18.26	15.49
MLLS-hard	59.10	55.42	52.70	48.93	46.47	44.04	41.27	39.82	38.30	40.46	37.78	35.67	32.95	31.08	29.22	26.85	25.30	23.70
MLLS-soft	58.45	54.70	52.13	48.56	46.34	44.11	41.66	40.41	39.30	39.90	37.20	35.06	32.43	30.53	28.72	26.58	25.50	24.11
BBSE-hard	33.20	25.93	24.53	19.03	26.15	23.99	16.85	28.03	15.67	28.65	28.39	27.83	26.37	26.79	24.51	23.09	16.69	17.60
BBSE-soft	60.95	57.47	54.86	51.03	48.23	45.67	42.47	40.42	37.94	41.12	38.32	36.18	33.16	30.94	28.75	26.16	24.34	22.31
RLLS-hard	**62.55**	**58.48**	54.97	49.59	45.31	40.94	35.22	31.31	26.56	41.25	38.04	35.11	31.06	27.92	24.76	20.89	18.26	15.72
RLLS-soft	**62.55**	**58.48**	54.97	49.59	45.31	40.94	35.22	31.31	26.56	41.25	38.04	35.11	31.06	27.92	24.76	20.89	18.26	15.49
MAPLS-APL-hard (**ours**)	60.67	57.72	**55.56**	52.51	50.31	48.05	45.09	43.33	41.31	41.34	**39.55**	38.01	**36.04**	34.48	32.80	30.49	28.68	26.63
MAPLS-APL-soft (**ours**)	60.34	57.58	55.44	52.50	50.32	**48.33**	**45.69**	**44.21**	**42.50**	41.15	39.32	37.78	36.04	**34.58**	32.97	**30.87**	**29.35**	**27.41**
PSLS-APL-hard (**ours**)	60.80	58.00	55.49	52.65	50.34	47.88	45.07	43.27	41.33	**41.61**	39.19	**38.14**	36.01	34.49	**32.99**	30.44	28.64	26.74
PSLS-APL-soft (**ours**)	60.61	57.95	55.46	**52.76**	**50.45**	48.00	45.47	43.86	42.22	41.44	39.11	38.11	36.01	34.52	32.97	30.51	28.89	27.17

Table 6. **Performance of Top1 Accuracy (↑) on ImageNet-LT and Place-LT dataset**, with Ordered Long-Tailed test set that have imbalance ratio $R = \{25, 10, 5, 2\}$. Best performances are in bold face and second best are in blue.

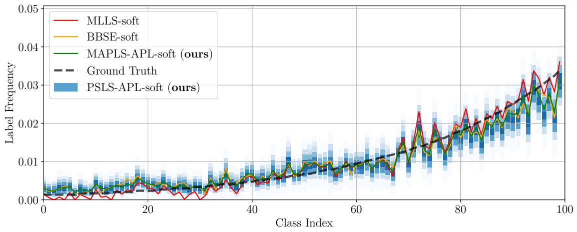

Figure 4. **Illustration of the label shift estimation result ($\boldsymbol{\pi}$).** On the Long-Tailed CIFAR100 dataset with Ordered Long-Tailed test set, our PSLS-APL model uses HMC to obtain samples of the posterior $\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha}, \mathbb{X})$ (posterior sample density histogram plot as blue bar heatmap), which fit nicely with the ground truth.

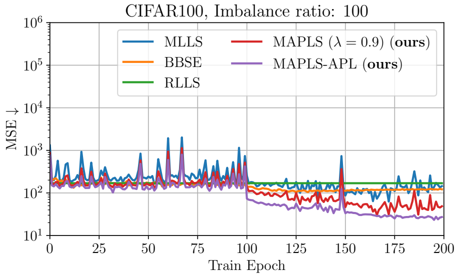

Figure 5. **Ablation study on stability of the MAPLS-APL model.** On the Long-Tailed CIFAR100 dataset with Ordered LT test set, our model is stable during the training of the classifier and performs better than SOTA methods.

distribution. When source domains have highly imbalanced label distributions (e.g. ImageNet-LT, Places-LT), the label shift estimation performance of our model stays relatively stable while previous models degrade significantly.

5.5. Ablation Study

We provide the density histogram of posterior samples Π collected by our PSLS-APL model in Fig. 4, with single value of $\boldsymbol{\pi}$ estimated by other models as well. The posterior $\mathbb{P}(\boldsymbol{\pi}|\mathbb{X}, \boldsymbol{\alpha})$ fits well with the ground truth and is able to provide a sense of uncertainty of our estimation.

We also analyze the estimation stability of our model during the training of classifier f on the source domain. Specifically, we monitor the performance of each label shift estimation model during the training of a Neural Network classifier on Long-Tailed CIFAR100 dataset. The test sets have Ordered Long-Tailed label distribution. As shown in Fig. 5, the performance of BBSE, MLLS and our model improves during the training of the classifier. This observation suggests that label shift estimation performance of these models could be further improved with a better classifier. Our MAPLS ($\lambda = 0.9$) and MAPLS-APL model performs better and stable in the last 50 epochs.

6. Discussion and Conclusion

In this work, we develop label shift estimation methods MAPLS-APL and PSLS-APL under a Bayesian framework that are applicable to large-scale datasets and robust to highly imbalanced source label distributions. In our MAPLS model, we derive an EM algorithm to obtain the MAP estimate of the target label distribution and propose a novel Adaptive Prior Learning model to adaptively adjust the prior parameter. In our PSLS model, we use HMC to sample from the strictly log-concave posterior $\mathbb{P}(\boldsymbol{\pi}|\mathbb{X}, \boldsymbol{\alpha})$.

Unlike previous benchmark evaluations, our experimental settings additionally covers a variety of large-scale datasets (ImageNet, Places) with highly imbalanced label distributions, which provide a more realistic evaluation of SOTA methods. Experiments on these datasets have demonstrated the effectiveness of our model and its potential to be applied in real world label shift problems.

References

[1] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020. 1, 2, 6, 7, 13, 14, 17, 19, 20

[2] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2018. 1, 2, 7, 17, 20

[3] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017. 6

[4] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018. 7

[5] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 2

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1, 3, 5, 7

[7] George Casella. Illustrating empirical bayes methods. *Chemometrics and intelligent laboratory systems*, 16(2):107–125, 1992. 5

[8] Yee Seng Chan and Hwee Tou Ng. Word sense disambiguation with distribution estimation. In *IJCAI*, volume 5, pages 1010–5, 2005. 2

[9] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2

[10] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020. 1, 2, 3, 5, 17, 19

[11] Jiaxian Guo, Mingming Gong, Tongliang Liu, Kun Zhang, and Dacheng Tao. Ltf: A label transformation framework for correcting label shift. In *International Conference on Machine Learning*, pages 3843–3853. PMLR, 2020. 2

[12] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 2

[13] Haibo He and Yunqian Ma. Imbalanced learning: foundations, algorithms, and applications. 2013. 2

[14] 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. 7, 19

[15] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014. 7

[16] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636, 2021. 2, 3, 7, 21

[17] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 2

[18] Yerlan Idelbayev. Proper ResNet implementation for CIFAR10/CIFAR100 in PyTorch. 7, 19

[19] Taejong Joo, Uijung Chung, and Min-Gwan Seo. Being bayesian about categorical probability. In *International Conference on Machine Learning*, pages 4950–4961. PMLR, 2020. 4

[20] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 1

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7, 21

[22] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018. 1, 2, 3, 7, 17, 20, 21

[23] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018. 3

[24] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 7

[25] Oren Mangoubi and Aaron Smith. Rapid mixing of hamiltonian monte carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017. 6

[26] Radford Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2011. 6

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 7, 19

[28] Hanyu Peng, Mingming Sun, and Ping Li. Optimal transport for long-tailed recognition with learnable cost matrix. In *International Conference on Learning Representations*, 2022. 1, 3

[29] Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021. 3

[30] Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in statistics*, pages 388–394. Springer, 1992. 5

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 7, 21

[32] 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, 2002. 1, 2, 3, 7, 14, 20

[33] Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009. 2

[34] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33:19276–19289, 2020. 1, 2, 3, 7

[35] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 19

[36] Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, and Zsolt Kira. Posterior re-calibration for imbalanced datasets. *Advances in Neural Information Processing Systems*, 33:8101–8113, 2020. 1, 3

[37] Stephen Tu. The dirichlet-multinomial and dirichlet-categorical models for bayesian inference. *Computer Science Division, UC Berkeley*, 2, 2014. 4

[38] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2020. 7

[39] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[40] 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. 2

[41] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 19

[42] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR, 2013. 2

[43] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16489–16498, 2021. 7

[44] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 7, 21