FMAPLS: Bayesian Label Shift Estimation Based on Dynamic Dirichlet Parameter Adaptation

Jiawei Hu, *Student Member, IEEE*, Javier A. Barria, *Member, IEEE*

***Abstract*—Label shift, a critical challenge in supervised learning, occurs when the class prior distribution of test data deviates from that of training data, leading to significant degradation in classifier performance. This issue is particularly impactful in real-world applications such as medical diagnosis and satellite imaging, where a reduction in classification accuracy can cause severe consequences. Existing methods, such as Maximum A Posterior Label Shift (MAPLS), rely on strict Dirichlet hyperparameter constraints, thereby imposing a limitation on their adaptability to dynamic learning environments and class-imbalanced scenarios. To address this problem, we propose Full Maximum A Posterior Label Shift (FMAPLS), a Bayesian label shift estimation framework that dynamically co-optimizes Dirichlet hyperparameters and class priors through Expectation-Maximization (EM) iterations. By eliminating rigid hyperparameter constraints and introducing a linear surrogate function, FMAPLS enhances both expressivity and robustness in prior estimation while reducing computational complexity. Extensive experiments on CIFAR100 under shuffled long-tail and Dirichlet imbalanced test priors demonstrate FMAPLS's superiority over state-of-the-art baselines, achieving up to 50% lower KL divergence and maintaining accuracy improvements by up to 0.5% in extreme imbalance settings. The results validate the efficacy of dynamic hyperparameter adaptation in estimating label shift, particularly under high class imbalance and distributional uncertainty.**

***Index Terms*—Label shift, imbalance data, expectation-maximization algorithm**

I. Introduction

DUE to the complexity of large-scale datasets, the training data distribution in supervised learning tasks often fails to represent the test data distribution. This distribution deviation, also called *distribution shift*, will cause a degradation in the performance of the supervised classification model when it is implemented in real-world scenarios. As a type of distribution shift, *label shift* appears when the prior probability $P(Y)$ differs from training to test datasets, while the likelihood probability $P(X|Y)$ remains invariant [1]. In applications such as medical diagnosis and satellite image analysis, label shift-induced classifier inaccuracies can have far-reaching consequences, including adverse effects on human healthcare and economic efficiency [2]. Therefore, accurately estimating label shift is of critical importance.

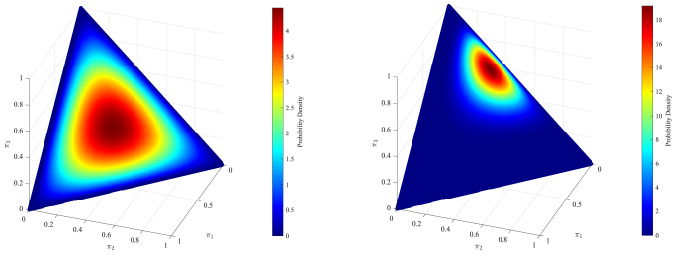
In 2002, Saerens *et al.* [3] proposed an Expectation-Maximization (EM) approach to iteratively update prior probabilities of the test dataset using classifier outputs. Lipton *et al.* [4] further introduced an algorithm called Black Box Shift Estimation (BBSE), which uses the confusion matrix of the held-out portion of the training set to detect and correct the label shift. Based on BBSE, Azizzadenesheli *et al.* [5] proposed Regularized Learning of Label Shift (RLLS), introducing a hyperparameter regularization framework to improve importance weight estimation across labeled training and unlabeled test data. Experimental results show the superiority of RLLS over BBSE in both weight estimation precision and classification accuracy. Later in 2020, Alexandari *et al.* [6] and Garg *et al.* [7] demonstrated that the EM algorithm outperforms both BBSE and RLLS in limited-data scenarios, achieving superior performance on benchmark datasets including CIFAR10 and MNIST. However, existing methods have focused primarily on small datasets and have not considered the influence of class-imbalanced training data on label shift estimation.

Recent work by Ye *et al.* [8] introduced a Bayesian label shift estimation framework that models class priors via a high-dimensional Dirichlet distribution, developing a Maximum A Posterior Label Shift (MAPLS) method through EM-based offline optimization. Although experimental results demonstrate MAPLS's superiority over BBSE and RLLS on large-scale datasets (e.g., CIFAR100, ImageNet) with class imbalance, its Adaptive Prior Learning (APL) module relies on two key rigid constraints: (1) equality of Dirichlet hyperparameter values, and (2) empirical formulation-driven hyperparameter updates. These constraints hinder the framework's adaptability and diversity, especially in complex and dynamic learning environments [9], [10].

To address the above issues, we propose a Full Maximum A Posterior Label Shift (FMAPLS) method to enhance the performance of label shift estimation. Unlike MAPLS, FMAPLS introduces dynamic hyperparameter adaptation in the Dirichlet framework by decoupling hyperparameter constraints and jointly optimizing priors with hyperparameters through EM iterations. This co-optimization can better represent the prior probability of the test dataset. In order to solve the problem that the closed-form solution cannot be obtained during the hyperparameter iteration process, we introduce a linear surrogate function with demonstrated maximum value congruence to the original formulation. This substitution reduces algorithmic complexity while preserving optimization maximum points. In summary, our contributions can be concluded as follows:

- A dynamic Dirichlet hyperparameter adaptation process is proposed within the EM framework, which relaxes the original MAPLS framework constraints and enhances the capture of complex label shifts in test datasets.
- A linear surrogate function is implemented to replace the Dirichlet hyperparameter update formulation, so that the updates can be computed using a closed-form solution expression.

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: j.hu24@imperial.ac.uk; j.barria@imperial.ac.uk).



(a) $\boldsymbol{\alpha} = [2, 2, 2]$ (b) $\boldsymbol{\alpha} = [2, 5, 9]$

Fig. 1: The probability density of Dirichlet distribution with different $\boldsymbol{\alpha}$.

II. Problem Setting

We first introduce the important notation and the general problem of label shift estimation. Consider a dataset with an input space $\boldsymbol{X}$ and a label space $\boldsymbol{Y} = \{1, 2, ..., K\}$, where K is the number of classes. A neural network-based classifier $f : \boldsymbol{X} \to \boldsymbol{Y}$ is used to predict the posterior probability vector $\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})$ of each input data X. The classifier obtains the label Y corresponding to the input data X by maximizing the probability vector.

Under the label shift setting, it is assumed that the source and test prior probabilities are different, while the likelihood probability remains unchanged. It can be mathematically expressed as:

$$\begin{cases} P(X_s = x|Y_s = y) = P(X_t = x|Y_t = y) \\ P(Y_s = y) \neq P(Y_t = y) \end{cases} \tag{1}$$

where X_s and Y_s denote the source domain input data and labels, respectively, while X_t and Y_t represent the test domain counterparts.

To simulate the source-target prior discrepancy, we adopt the standard long-tailed distribution with an imbalance ratio ρ from [11] for training, while evaluating under shuffled long-tail and Dirichlet imbalanced test priors. The class frequency vector of the source dataset $\boldsymbol{n_s}$ and test dataset $\boldsymbol{n_t}$ can be written as:

$$\boldsymbol{n}_s = f_{LT}(\boldsymbol{N_s}, \rho, K) = \lfloor \boldsymbol{N_s} \rho^{\frac{i}{K}} \rfloor \tag{2}$$

$$\boldsymbol{n}_t^{(l)} = \begin{cases} \sigma(f_{LT}(\boldsymbol{N_t}, \rho, K)) & l = \text{shuffle} \\ \lfloor \boldsymbol{N_t} \circ Dir(\boldsymbol{\alpha}) \rfloor & l = \text{Dirichlet} \end{cases} \tag{3}$$

where $\boldsymbol{N_s}$ and $\boldsymbol{N_t}$, respectively, denote the class sizes in the original source and target domains, $\sigma(\cdot)$ and $Dir(\boldsymbol{\alpha})$ represent the shuffle function and the Dirichlet distribution with a K-dimensional hyperparameter $\boldsymbol{\alpha}$, $\circ$ is the Hadamard product and $\lfloor \cdot \rfloor$ is floor function. The further discussion of Dirichlet distribution will be shown in Section III.A.

III. Proposed Method

Building upon the MAPLS framework [8], we formulate the test data prior probability as a Dirichlet distribution and relax its fixed hyperparameter constraints. This enables joint estimation of Dirichlet hyperparameters $\boldsymbol{\alpha}$ and test data priors $\boldsymbol{\pi}$ through an EM iterative optimization process, thus solving the problem that cannot obtain a direct closed-form solution via standard analytical techniques [13], [14]. In order to overcome the difficulty of obtaining a closed-form solution for $\boldsymbol{\alpha}$ update, we further introduce a linear surrogate function to reduce computational complexity while ensuring the same optimization results.

A. Dirichlet distribution

The Dirichlet distribution with K-dimensional hyperparameters $\boldsymbol{\alpha} \in \mathbb{R}^K_{>0}$ is supported on the $(K-1)$-dimensional standard simplex $\Delta^{K-1} = \{\boldsymbol{\pi} \in \mathbb{R}^K | \pi_i \geq 0, \sum_{i=1}^K \pi_i = 1\}$ [12]. The distribution is given by:

$$P(\boldsymbol{\pi}; \boldsymbol{\alpha}) = Dir(\boldsymbol{\pi}; \boldsymbol{\alpha}) = \frac{1}{\boldsymbol{B}(\boldsymbol{\alpha})} \prod_{j=1}^{K} \pi_j^{\alpha_j - 1} \tag{4}$$

$$\boldsymbol{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{j=1}^K \alpha_j)} \tag{5}$$

where $\Gamma(\cdot)$ is Gamma function. Due to its simplex support and being the conjugate prior of the categorical and multinomial distributions, the Dirichlet distribution is often modeled as a prior probability in Bayesian statistics [15], [16], [17]. Fig. 1 visualizes the probability density of Dirichlet distributions under different $\boldsymbol{\alpha}$. A symmetric setting yields a nearly uniform prior, while an asymmetric one concentrates probability on specific classes. This property motivates $\boldsymbol{\alpha}$ dynamic adaptation in FMAPLS, enabling the model to capture diverse and imbalanced label priors.

B. FMAPLS algorithm

Assume the neural network-based classifier have N input data $\boldsymbol{X} = \{x_i, x_2, ..., x_N\}$, the complete posterior probability of parameter $\boldsymbol{\theta} = \{\boldsymbol{\pi}; \boldsymbol{\alpha}\}$ can be written as:

$$\begin{aligned} P(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) &= \frac{P(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) P(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} P(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) d\boldsymbol{\theta}} \\ &= \frac{1}{Z} P(\boldsymbol{\theta}) \prod_{i=1}^{N} P(X_t = x_i, Y_t = y_i|\boldsymbol{\theta}) \\ &= \frac{1}{Z} P(\boldsymbol{\theta}) \prod_{i=1}^{N} P(X_s = x_i|Y_s = y_i) P(Y_t = y_i|\boldsymbol{\theta}) \end{aligned} \tag{6}$$

where Z is a constant value. Leveraging the posterior, EM algorithm is employed to iteratively refine parameter $\boldsymbol{\theta}$.

For the **E step**, the expectation of complete-data logarithm likelihood function conditioned on $\boldsymbol{\theta}^{(t)}$ at tth iteration can be expressed as equation (7), where $\delta_{y_i,j}$ is the Kronecker delta function, $Const$ term includes all terms that are irrelevant to $\boldsymbol{\theta}$, such as $1/Z$ and $\sum_{i=1}^N P(X_s = x_i|Y_s = y_i)$.

According to the Bayes Formula and the label shift defined in equation (1), $P(Y_t = y_i = j|X_t = x_i, \boldsymbol{\theta}^{(t)})$ can be further written as [3]:

$$\begin{aligned} &P(Y_t = y_i = j|X_t = x_i, \boldsymbol{\theta}^{(t)}) \\ &= \frac{\frac{\pi_j^{(t)}}{\varepsilon_j} P(Y_s = y_i = j|X_s = x_i)}{\sum_{k=1}^K \frac{\pi_k^{(t)}}{\varepsilon_k} P(Y_s = y_i = k|X_s = x_i)} = B_i^{(t)} \pi_j^{(t)} \end{aligned} \tag{8}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_K)$ denotes the training dataset's class prior probabilities, estimated via class proportions.

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}\left[\log P(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y})\right] \\
&= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}\left[Const + \log\frac{1}{\boldsymbol{B}(\boldsymbol{\alpha})} + \sum_{j=1}^{K}(\alpha_j - 1)\log\pi_j + \sum_{i=1}^{N}\sum_{j=1}^{K}\delta_{y_i,j}\log\pi_j\right] \\
&= Const + \log\frac{1}{\boldsymbol{B}(\boldsymbol{\alpha})} + \sum_{j=1}^{K}(\alpha_j - 1)\log\pi_j + \sum_{j=1}^{K}\sum_{i=1}^{N}P(Y_t = y_i = j|X_t = x_i, \boldsymbol{\theta}^{(t)})\log\pi_j
\end{aligned} \tag{7}
$$

For the update of $\boldsymbol{\pi}$ in the **M step**, the goal is to maximize terms that related to $\boldsymbol{\pi}$ in the equation (7) with normalization constraint:

$$
\begin{cases}
\arg\max\limits_{\boldsymbol{\pi}} \sum\limits_{j=1}^{K}(\alpha_j^{(t)} - 1)\log\pi_j + \sum\limits_{j=1}^{K}\sum\limits_{i=1}^{N}B_i^{(t)}\pi_j^{(t)}\log\pi_j \\
\text{s.t.}: \quad \sum\limits_{j=1}^{K}\pi_j = 1 \ \text{ and } \ \pi_j > 0
\end{cases} \tag{9}
$$

Using Lagrange Multiplier Method, the update equation can be obtained via:

$$
\pi_j^{(t+1)} = \frac{\alpha_j^{(t)} - 1 + \sum_{i=1}^{N}B_i^{(t)}\pi_j^{(t)}}{\sum_{k=1}^{K}\left(\alpha_k^{(t)} - 1 + \sum_{i=1}^{N}B_i^{(t)}\pi_k^{(t)}\right)} \tag{10}
$$

As for the update of $\boldsymbol{\alpha}$ in the **M step**, the maximize function and the partial derivative function of $\boldsymbol{\alpha}$ can be respectively expressed as:

$$
\arg\max_{\boldsymbol{\alpha}} \ L_{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \ \log\frac{1}{\boldsymbol{B}(\boldsymbol{\alpha})} + \sum_{j=1}^{K}(\alpha_j - 1)\log\pi_j^{(t+1)} \tag{11}
$$

$$
\frac{\partial L_{\boldsymbol{\alpha}}}{\partial\alpha_j} = -\psi(\alpha_j) + \psi(\sum_{n=1}^{K}\alpha_n) + \log\pi_j^{(t+1)} \tag{12}
$$

Due to the presence of the digamma function $\psi(\cdot)$ in equation (12), it is difficult to give an analytical solution for α_j. In order to achieve the continuous iteration of α_j, the first-order gradient ascent method is used to update parameter $\boldsymbol{\alpha}$:

$$
\alpha_j^{(t+1)} = \alpha_j^{(t)} + \mu\left[-\psi(\alpha_j^{(t)}) + \psi(\sum_{n=1}^{K}\alpha_n^{(t)}) + \log\pi_j^{(t+1)}\right] \tag{13}
$$

where μ is the learning rate.

C. Convergence of parameter updates

To guarantee the updates of $\boldsymbol{\pi}$ and $\boldsymbol{\alpha}$ converge to the global maximum at each iteration, it is essential to prove that the corresponding objective functions are concave. Due to the concavity of the logarithmic function, equation (9) is clearly concave. As for the equation (11), due to the second term of $L_{\boldsymbol{\alpha}}$ is a linear combination of $\boldsymbol{\alpha}$, so we only need to prove $\log\boldsymbol{B}(\boldsymbol{\alpha})$ is convex.

As established in [18], the Dirichlet distribution belongs to the *exponential family*. Its *log normalizer* $A(\eta) = \log\boldsymbol{B}(\boldsymbol{\alpha})$ (with *natural parameter* $\boldsymbol{\eta} = \boldsymbol{\alpha} - 1$) is a differentiable real-valued convex function of $\boldsymbol{\eta}$ over the open convex domain $\boldsymbol{\alpha} > 0$ [19]. Consequently, equation (11) is concave.

The concavity of M-step subproblems ensures that each M-step finds the global maximum for its respective parameters given the current expected sufficient statistics. This property ensures that the EM generates a parameter sequence $\{(\boldsymbol{\pi}^{(t)}, \boldsymbol{\alpha}^{(t)})\}$ corresponding to monotonically non-decreasing log-likelihood values until the change in the parameter is less than a threshold. After FMAPLS converges, the final estimated test prior probability $\boldsymbol{\pi}^*$ is applied to reweight the classifier's predicted probabilities according to equation (8).

D. Linear surrogate function

While gradient ascent can optimize $\boldsymbol{\alpha}$, the method's convergence rate and stability critically depend on the initialization and the learning rate μ. To overcome initialization dependence and learning rate sensitivity problems, FMAPLS applied a linear surrogate function to replace equation (13):

$$
\frac{\alpha_1}{\pi_1^{(t+1)}} = \frac{\alpha_2}{\pi_2^{(t+1)}} = \cdots = \frac{\alpha_K}{\pi_K^{(t+1)}} \tag{14}
$$

Assume there is a constant c such that:

$$
\alpha_j = \frac{c}{\max(\boldsymbol{\pi}^{(t+1)})}\pi_j^{(t+1)} = \hat{c}\pi_j^{(t+1)} \tag{15}
$$

Thus, the partial derivative equation (12) can be expressed as:

$$
\begin{aligned}
\frac{\partial L_{\boldsymbol{\alpha}}}{\partial\alpha_j} &= -\psi(\hat{c}\pi_j^{t+1)}) + \psi(\hat{c}) + \log\pi_j^{(t+1)} \\
&= \left[\log(\hat{c}\pi_j^{(t+1)}) - \psi(\hat{c}\pi_j^{(t+1)})\right] + \left[\psi(\hat{c}) - \log(\hat{c})\right]
\end{aligned} \tag{16}
$$

According to the asymptotic expansion of the digamma function [20], equation (16) approaches 0 as $c \to +\infty$, making it asymptotically equivalent to gradient ascent in locating the maximum. For practical implementation with finite c, however, the first and second term in equation (16) are approximated by $-\mathcal{O}(\frac{1}{\hat{c}\pi_j^{(t+1)}})$ and $\mathcal{O}(\frac{1}{\hat{c}})$, respectively. Consequently, for finite c, the condition $\boldsymbol{\alpha} \propto \boldsymbol{\pi}^{(t+1)}$ yields a solution where $|\nabla_{\boldsymbol{\alpha}}L_{\boldsymbol{\alpha}}| = \mathcal{O}\left(\frac{1}{c}\right)$, placing it within an $\mathcal{O}\left(\frac{1}{c}\right)$-neighborhood of the true maximum due to the concavity of $L_{\boldsymbol{\alpha}}$.

Unlike traditional gradient ascent, which requires $\mathcal{O}(T_{\text{grad}}K^2)$ complexity per EM iteration and is sensitive to initialization and μ, our linear surrogate provides closed-form $\mathcal{O}(K)$ updates. Because $L_{\boldsymbol{\alpha}}$ lacks a closed-form maximizer, gradient ascent is computationally expensive and unstable in high dimensions, motivating the adoption of a surrogate function. This alternative removes parameter tuning while ensuring maximization in the $\boldsymbol{\alpha}$-update step.

IV. Experimental results

We conduct experiments on the CIFAR100 benchmark [21] using a ResNet32 classifier, training on standard long-tail imbalanced variants and evaluating under both shuffle

TABLE I: Performance of KL divergence and classification accuracy ("KL|Acc.") on shuffle long-tail test distributions

Dataset	CIFAR100-0.5-LT										CIFAR100-0.2-LT									
Imbalance ratio ρ	0.02		0.025		0.05		0.1		0.5		0.02		0.025		0.05		0.1		0.5	
MLLS	0.054	71.664	0.052	71.654	0.044	70.206	0.036	68.942	0.027	67.250	0.057	68.668	0.052	68.520	0.041	67.454	0.033	65.832	0.020	64.486
BBSE	0.079	71.522	0.071	71.564	0.048	70.234	0.035	69.062	0.018	67.474	0.083	68.462	0.072	68.348	0.054	67.384	0.039	65.794	0.020	64.542
RLLS	0.652	67.206	0.588	67.520	0.411	67.366	0.254	67.258	0.049	67.322	0.767	62.846	0.699	63.308	0.501	63.340	0.361	62.796	0.159	63.142
MAPLS	0.071	**71.686**	0.063	71.698	0.043	**70.308**	0.032	**69.160**	**0.013**	**67.620**	0.099	68.360	0.090	68.248	0.066	67.294	0.046	65.786	**0.011**	**64.748**
FMAPLS	**0.048**	**71.686**	**0.047**	**71.712**	**0.038**	70.264	**0.031**	69.014	0.022	67.352	**0.052**	**68.714**	**0.047**	**68.540**	**0.038**	**67.500**	**0.031**	**65.856**	0.019	64.524

Dataset	CIFAR100-0.1-LT										CIFAR100-0.05-LT									
Imbalance ratio ρ	0.02		0.025		0.05		0.1		0.5		0.02		0.025		0.05		0.1		0.5	
MLLS	0.104	63.658	0.091	63.334	0.070	62.134	0.054	60.680	0.041	59.066	0.162	60.898	0.151	60.660	0.135	59.052	0.101	57.948	0.076	56.012
BBSE	0.114	63.466	0.102	63.136	0.074	62.150	0.063	60.740	0.041	59.200	0.146	60.430	0.133	60.202	0.107	58.690	0.079	57.676	0.059	55.858
RLLS	0.859	57.292	0.804	57.516	0.602	57.684	0.481	57.512	0.269	57.636	0.998	51.804	0.962	52.276	0.737	52.004	0.623	52.134	0.412	52.102
MAPLS	0.149	63.318	0.134	62.996	0.096	62.094	0.070	**60.826**	**0.014**	**59.950**	0.185	60.262	0.175	60.040	0.131	58.714	0.088	57.882	**0.015**	**56.758**
FMAPLS	**0.085**	**63.716**	**0.074**	**63.422**	**0.062**	**62.226**	**0.052**	60.756	0.039	59.126	**0.095**	**60.966**	**0.084**	**60.728**	**0.065**	**59.130**	**0.050**	**58.000**	0.039	56.088

TABLE II: Performance of KL divergence and classification accuracy ("KL|Acc.") on Dirichlet test distributions

Dataset	CIFAR100-0.5-LT								CIFAR100-0.2-LT							
Dirichlet $\alpha_{\max}$	1		2		3		4		1		2		3		4	
MLLS	0.054	74.309	0.043	72.067	0.043	71.221	0.041	70.961	0.066	71.415	0.052	69.207	0.045	67.949	0.043	67.889
BBSE	0.134	73.724	0.091	71.781	0.077	71.010	0.069	70.843	0.140	70.802	0.100	68.778	0.081	67.709	0.077	67.634
RLLS	1.127	67.176	0.778	67.251	0.651	67.320	0.575	67.589	1.230	62.910	0.935	62.909	0.823	62.681	0.716	63.153
MAPLS	0.123	73.945	0.088	71.900	0.078	71.069	0.072	70.894	0.177	70.666	0.139	68.647	0.117	67.543	0.112	67.440
FMAPLS	**0.052**	**74.314**	**0.041**	**72.080**	**0.041**	**71.237**	**0.039**	**70.984**	**0.065**	**71.421**	**0.051**	**69.211**	**0.044**	**67.953**	**0.042**	**67.894**

Dataset	CIFAR100-0.1-LT								CIFAR100-0.05-LT							
Dirichlet $\alpha_{\max}$	1		2		3		4		1		2		3		4	
MLLS	0.105	66.787	0.081	64.169	0.074	63.261	0.071	62.497	0.118	63.329	0.098	60.722	0.083	61.046	0.078	60.173
BBSE	0.177	66.301	0.125	63.867	0.112	63.029	0.103	62.312	0.203	62.674	0.167	60.074	0.133	60.641	0.125	59.734
RLLS	1.374	57.625	1.030	57.393	0.949	57.273	0.832	57.444	1.501	51.738	1.149	51.532	1.032	52.542	0.955	52.378
MAPLS	0.249	66.003	0.195	63.526	0.178	62.777	0.163	62.135	0.303	62.229	0.242	59.921	0.218	60.357	0.210	59.478
FMAPLS	**0.103**	**66.802**	**0.080**	**64.189**	**0.072**	**63.280**	**0.070**	**62.518**	**0.115**	**63.345**	**0.095**	**60.731**	**0.080**	**61.072**	**0.076**	**60.196**

and Dirichlet imbalanced to simulate realistic label shifts. The initialization of FMAPLS is set to uniform distribution ($\pi_j^0 = 1/K, \alpha_j^0 = 1$) and c is set to 10^4. Label shift estimation performance is assessed by KL divergence [22] and post-shift classification accuracy, in comparison with MLLS [3], BBSE [4], RLLS [5], and MAPLS [8]. All data in Table I and II are the means of 100 independent Monte Carlo experiments, obtained via repeated random sampling of shuffled and Dirichlet-distributed test priors.

Table I evaluates label shift estimation and classification accuracy under shuffled long-tail test distributions with varying ρ. FMAPLS shows clear advantages in extreme imbalance ($\rho < 0.1$), yielding the lowest KL divergence and comparable or better accuracy than the baselines. Specifically, it reduces KL divergence by up to 50% over MAPLS while improving classification accuracy by up to 0.5% in most cases. These results confirm the effectiveness of FMAPLS's dynamic adaptation under severe class imbalance.

However, the superior performance of FMAPLS diminishes for $\rho \geq 0.1$, where KL or classification accuracy exhibits comparable performance to MAPLS with occasional degradation. This divergence stems from the regularization that the APL approach brings to MAPLS, which enforces equal values for α_i across each iteration. By imposing a symmetric prior, this constraint suppresses overfitting to the sample noise in training data, thereby reducing the variance of the algorithm. Although MAPLS introduces some bias due to these constraints, it has little impact on the estimation results as $\rho \to 1$.

Table II quantifies label shift estimation and classification accuracy performance under Dirichlet distributed test priors. Hyperparameters α_i are constrained within the interval $(0, \alpha_{\max}]$, and a lower $\alpha_{\max}$ will increase the uncertainty of the prior. Compared with baselines, FMAPLS achieves the lowest KL divergence and the highest accuracy in all experimental settings, demonstrating not only superior performance but also remarkable robustness to varying degrees of data imbalance and distributional uncertainty. Notably, by relaxing MAPLS's hyperparameter constraints, FMAPLS achieves lower KL divergence and higher accuracy under increasing uncertainty in test priors. This demonstrates the advantage of dynamic parameter adaptation in handling distributional uncertainty.

V. Final remarks

In this letter, we propose FMAPLS, a Bayesian label shift estimation framework that dynamically adapts Dirichlet hyperparameters through EM iterations. Unlike MAPLS, FMAPLS removes strict hyperparameter constraints and jointly optimizes $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ via a co-adaptation process. To address the lack of a closed-form solution for updating $\boldsymbol{\alpha}$, we introduce a linear surrogate function with the same optimality conditions as the original objective. Extensive experiments show that FMAPLS achieves lower KL divergence and higher classification accuracy, especially under severe class imbalance and high prior uncertainty. The proposed dynamic adaptation and surrogate-based optimization effectively overcome the limitations of static Dirichlet assumptions, ensuring robustness across diverse data distributions. Future work includes validating the algorithm on more extensive datasets, exploring parameter posterior estimation with variational inference or Markov Chain Monte Carlo, and using Probably Approximately Correct-Bayesian method to analyze FMAPLS generalization bounds.

REFERENCES

[1] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognit.*, vol. 45, no. 1, pp. 521–530, 2012.

[2] P. W. Koh et al., "Wilds: A benchmark of in-the-wild distribution shifts," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 5637–5664.

[3] M. Saerens, P. Latinne, and C. Decaestecker, "Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure," *Neural Comput.*, vol. 14, no. 1, pp. 21–41, 2002.

[4] Z. Lipton, Y. X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 3122–3130.

[5] K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar, "Regularized learning for domain adaptation under label shifts," in *Proc. 7th Int. Conf. Learn. Representations (ICLR)*, 2019, pp. 1–13.

[6] A. Alexandari, A. Kundaje, and A. Shrikumar, "Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 222–232.

[7] S. Garg, Y. Wu, S. Balakrishnan, and Z. Lipton, "A unified view of label shift estimation," in *Proc. 34th Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 3290–3300.

[8] C. Ye, R. Tsuchida, L. Petersson, and N. Barnes, "Label shift estimation for class-imbalance problem: A bayesian approach," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 1073–1082.

[9] Y. Meng, X. Yao, K. Chen, Y. Wu, L. Zhang, Z. Bing, and A. Knoll, "Pretrained Bayesian non-parametric knowledge prior in robotic long-horizon reinforcement learning," 2025, *arXiv:2503.21975*.

[10] H. Zhang, Q. Lin, Y. Li, L. Cheng, and Y.-C. Wu, "Activity detection for massive connectivity in cell-free networks with unknown large-scale fading, channel statistics, noise variance, and activity probability: A Bayesian approach," *IEEE Trans. Signal Process.*, vol. 72, pp. 942–957, 2024.

[11] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. 33rd Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 1567–1578.

[12] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet distribution and related processes," Dept. Elect. Eng., Univ. Washington, Seattle, WA, USA, Rep. UWEETR-2010-0006, 2010.

[13] E. Efendi, and B. Dulek, "Online EM-based ensemble classification with correlated agents," *IEEE Signal Process. Lett.*, vol. 28, pp. 294–298, 2021.

[14] S. R. Chetupalli, and T. V. Sreenivas, "Joint Bayesian estimation of time-varying LP parameters and excitation for speech," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 357–361, 2017.

[15] H. Steck, and T. Jaakkola, "On the Dirichlet prior and Bayesian regularization," in *Proc. 15th Adv. Neural Inf. Process. Syst. (NIPS)*, 2002.

[16] F. Tahsin, H. Ennajari, and N. Bouguila, "Author dirichlet multinomial allocation model with generalized distribution (ADMAGD)," in *Proc. Int. Symp. Netw. Comput. Commun. (ISNCC)*, 2024, pp. 1–7.

[17] C. Xu, J. Si, Z. Guan, W. Zhao, Y. Wu, and X. Gao, "Reliable conflictive multi-view learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 14, 2024, pp. 16129–16137.

[18] J. Huang, "Maximum likelihood estimation of Dirichlet distribution parameters," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., 2005.

[19] F. Nielsen, and R. Nock, "Entropies and cross-entropies of exponential families," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2010, pp. 3621–3624.

[20] J. M. Bernardo, "Psi (digamma) function," *J. Royal Statist. Soc. Ser. C (Appl. Statist.)*, vol. 25, no. 3, pp. 315–317, 1976.

[21] A. Krizhevsky, and G. Hinton, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[22] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.