SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2025, Vol. 68, Iss. 5, 152102:1–152102:14
https://doi.org/10.1007/s11432-023-3954-7

Explaining the better generalization of label distribution learning for classification

Jing WANG[1,2*] & Xin GENG[1,2*]

[1]*School of Computer Science and Engineering, Southeast University, Nanjing 211189, China*
[2]*Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing 211189, China*

Received 22 April 2023/Revised 10 September 2023/Accepted 21 September 2023/Published online 17 January 2025

Abstract Label distribution learning (LDL) has shown advantages over traditional single-label learning (SLL) in many real-world applications, but its superiority has not been theoretically understood. In this paper, we attempt to explain why LDL generalizes better than SLL. Label distribution has rich supervision information such that an LDL method can still choose the sub-optimal label from label distribution even if it neglects the optimal one. In comparison, an SLL method has no information to choose from when it fails to predict the optimal label. The better generalization of LDL can be credited to the rich information of label distribution. We further establish the label distribution margin theory to prove this explanation; inspired by the theory, we put forward a novel LDL approach called LDL-LDML. In the experiments, the LDL baselines outperform the SLL ones, and LDL-LDML achieves competitive performance against existing LDL methods, which support our explanation and theories in this paper.

Keywords label distribution learning, classification, generalization, label distribution margin, label ambiguity, neural network

Citation Wang J, Geng X. Explaining the better generalization of label distribution learning for classification. Sci China Inf Sci, 2025, 68(5): 152102, https://doi.org/10.1007/s11432-023-3954-7

1 Introduction

Label distribution learning (LDL) [1] is a recently proposed learning paradigm. Different from the traditional single-label learning (SLL) and multi-label learning (MLL) [2,3] that use 1 and 0 to represent relevant and irrelevant labels, respectively, LDL considers the relationship between instances and labels as real-values. Precisely, LDL assigns each instance with a label distribution, whose elements are called the label description degrees and model the relevant degrees of labels. For example, Figure 1 shows an image of natural scene [4] with the ground-truth labels "Sky" and "Desert", but "Desert" is more relevant than "Sky". LDL can explicitly express that by assigning a label distribution (the label distribution is from [5]) to the image, in which "Sky" and "Desert" have label description degrees of 0.296 and 0.704, respectively, reflecting their different relevance degrees. Therefore, LDL helps solve label ambiguity [6], which is more suitable for some applications in real-world scenarios.

LDL aims to learn a multivariate function from the training set with label distributions [1]. Meanwhile, researchers have found it effective for classification and have applied it to extensive applications, such as facial age estimation [7–9], emotion recognition [10–13], head-pose estimation [14], sentiment analysis [15,16], skin disease grading [17], facial beauty perception [18–20], and noisy label learning [21,22]. These applications work as follows. First, an LDL function is learned from the datasets with label distributions. Second, the learned LDL function is regarded as a classifier; for an unknown instance, the label that has the optimal predicted label description degree by the learned LDL function is treated as the predicted label [23]. For example, in facial age estimation, Geng et al. [8] first learned an LDL function from the facial images that are described by label distributions covering all possible age labels. Then, for an unknown image, they regarded the age label that has the optimal predicted label description degree by the learned function as the predicted age. LDL has shown advantages over traditional SLL in these applications, but its superiority has not been theoretically understood.

* Corresponding author (email: wangjing91@seu.edu.cn, xgeng@seu.edu.cn)

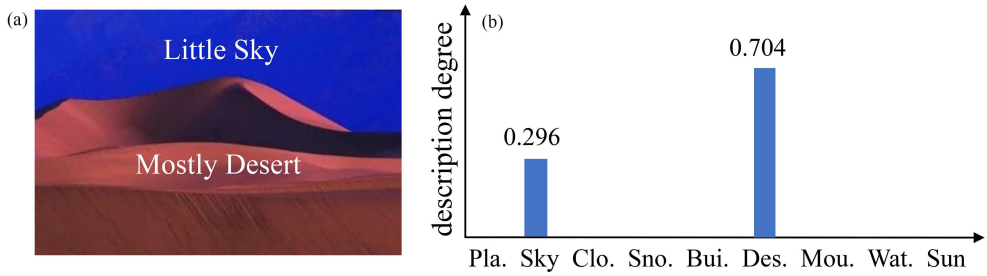


Figure 1 (Color online) Illustration of (a) a natural scene image and (b) its label distribution. The image has the ground-truth labels "Sky" and "Desert", but "Desert" is more relevant than "Sky". In the label distribution, "Sky" and "Desert" have degrees of 0.296 and 0.704, respectively, indicating their different relevance degrees.

Recently, Refs. [23,24] has been devoted to providing some theoretical understanding of LDL, especially from a classification perspective. Our previous work [23] studied the generalization of LDL for classification. We found that the expected L_1-norm loss of LDL bounds the classification error, which suggests that the classification process generalizes as long as label distribution is learned well. We also showed in [24] the relationship between the predicted label and the predicted label distribution. It presents a sufficient condition under which LDL guarantees to predict the optimal label in a label distribution [24]. Essentially, these studies answer why LDL generalizes for classification but cannot explain why it generalizes better, and in particular, why it generalizes better than SLL.

In this paper, we investigate the better generalization of LDL and attempt to explain why LDL generalizes better than traditional SLL. We point out that label distribution has rich supervision information such that LDL methods can choose from a label distribution—if not the optimal label—the sub-optimal label. Thus, LDL is able to generalize better than SLL because SLL algorithms have nothing to choose from a single-label if they miss the optimal label. That is, we can attribute the better generalization of LDL to the rich information of label distribution. In addition, we establish the label distribution margin theory (see Theorem 2) and apply it to justify the above explanation (see Theorem 3). According to the theories, we further put forward a new LDL method called LDL-LDML. In the experiments, the LDL baselines outperform the SLL ones, justifying that the key to the better generalization of LDL lies in label distribution. Besides, LDL-LDML achieves competitive performance compared with existing LDL methods, which further validates our theories.

Our major contributions are summarized as follows. First, we present, to the best of our knowledge, the first explanation for why LDL generalizes better than SLL. Second, we develop the label distribution margin theory to prove the explanation and put forward the LDL-LDML approach according to the theories. Third, we conduct extensive experiments to support our theories.

The rest of this paper is structured as follows. Section 2 briefly reviews the related work. Section 3 presents the main results. Next, Section 4 elaborates on the LDL-LDML method. Section 5 reports the experimental results. Finally, Section 6 concludes this paper.

2 Related work

2.1 Label distribution learning

Geng [1] first proposed LDL and formalized it as a new learning paradigm. He also suggested three strategies for designing LDL approaches, including problem transformation (PT), algorithm adaptation (AA), and specialized algorithms (SA), and put forward several representative LDL baselines, such as PT-support vector machine (SVM), AA-k nearest neighbors (NN), and SA-BFGS [1].

Since then, many specialized LDL algorithms have been proposed. Geng and Hou [25] viewed LDL as a regression problem and proposed LDL-support vector regression (SVR), which employs the multivariate SVR to learn label distribution. Xing et al. [26] applied the logistic boosting to LDL. Gao et al. [6] combined deep learning and LDL and designed the first deep LDL model DLDL. It can learn representation and label distribution in an end-to-end way. Shen et al. [7] discarded the assumption that label distribution can be represented by the maximum entropy model [1] and proposed LDLFs. It uses the differentiable trees [27] and can learn any general form of label distributions [7]. Shen et al. [28] further improved LDLFs and proposed DLDLF. Ren et al. [29] studied feature selection for LDL and designed

LDLSF that learns common and label-specific features. Xu and Zhou [30] proposed the incomplete LDL with some missing entries in the label distribution, and Zhang et al. [31] designed a safe incomplete LDL approach SILDL.

2.2 LDL for classification

So far, LDL has seen many classification applications. In age estimation, Geng et al. [8] used a label distribution to describe an image to model the slow and smooth changes of facial appearance; an image contributes to its chronological age and adjacent ages as well. They learned an LDL function from such distributions; for an unknown image, they regarded the age with the optimal predicted label description degree as the predicted age [8]. Since then, LDL has become a prominent method for facial age estimation [6, 7, 9, 28]. In head-pose estimation, Geng et al. [14] assigned, instead of an inaccurate pose label, a multivariate label distribution (MLD) to an image. It covers not only the given pose but also the neighborhood poses of an image. They learned mapping from an image to an MLD. For a test image, they treated the pose label having the optimal predicted label description degree as the prediction [14]. In facial beauty perception, Liang et al. [18] adopted a label distribution to describe a face image, which covers all ratings from human annotators. In skin disease analysis, Wu et al. [17] utilized two label distributions to model the number of lesions and acne severity because of the ambiguity of counting lesions and grading severity.

Our previous work [23] found that existing general-purpose LDL methods may face the challenge of objective inconsistency when adopted for classification. Concretely, LDL is proposed to learn the whole label distribution, but classification aims to learn the optimal label in a label distribution. LDL may neglect the ground-truth label in order to learn the whole label distribution, which may reduce the classification performance. Gao et al. [32] first noticed such inconsistency in age estimation and then proposed to jointly learn the chronological age and label distribution. We proposed a specially designed LDL method RWLM-LDL in [23] for classification. It addresses the objective inconsistency by re-weighting and large margin and shows a remarkable improvement over the general-purpose LDL approaches [23].

The preceding studies have empirically validated the advantages of LDL over traditional SLL but have not provided any theoretical understanding. In this study, we attempt to explain why LDL wins SLL.

2.3 Generalization of LDL

Recently, researchers have noticed the generalization study of LDL. Zhao and Zhou [33] introduced the optimal transport (OT) distance to LDL and proposed LALOT. They derived a data-dependent generalization bound for it. Our earlier work [34] analyzed the generalization of some representative LDL baselines. We further studied the generalization of LDL for classification and disclosed that the expected L_1-norm loss of LDL bounds the classification error [23]. In another word, the classification process generalizes as long as LDL generalizes. Moreover, we proved a sufficient condition, under which the classification process guarantees to select the optimal label in a label distribution [24]. These researches explained the generalization of LDL. In comparison, this study investigates why LDL generalizes better, and in particular, why it generalizes better than traditional SLL.

3 Main results

In this section, we first introduce some preliminaries. Then, we develop the label distribution margin theory and apply it to explain why LDL has better generalization than SLL.

3.1 Preliminaries

Let $\mathcal{X} \in \mathbb{R}^q$ be the input space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$ stand for the label space having m candidate labels. Let $\mathcal{D}$ denote the underlying distribution over $\mathcal{X}$. In the settings of LDL, each $\boldsymbol{x}$ is assigned with a label distribution $D = \{d_{\boldsymbol{x}}^{y_1}, d_{\boldsymbol{x}}^{y_2}, \ldots, d_{\boldsymbol{x}}^{y_m}\}$; $d_{\boldsymbol{x}}^{y_j}$ is called the label description degree satisfying $d_{\boldsymbol{x}}^{y_j} \geqslant 0$ and $\sum_j d_{\boldsymbol{x}}^{y_j} = 1$ [1]. Let $S = \{(\boldsymbol{x}_1, D_1), \ldots, (\boldsymbol{x}_n, D_n)\}$ be a training set, where $D_i = \{d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \ldots, d_{\boldsymbol{x}_i}^{y_m}\}$ is the label distribution of $\boldsymbol{x}_i$. Table 1 summarizes the mainly adopted notations.

Table 1 Description of some important notations.

Symbol	Definition	Symbol	Definition
$\mathcal{X}$	Feature space	$\mathbb{I}(\cdot)$	The indicator function
$\mathcal{Y}$	Label space	$\mathbb{E}[\cdot]$	The expectation function
$\boldsymbol{x}$	Instance variable	$y_{\boldsymbol{x}}^j$	The jth optimal label of $\boldsymbol{x}$
D	The label distribution of $\boldsymbol{x}$	$\mathcal{D}$	The underlying distribution over $\mathcal{X}$
$d_{\boldsymbol{x}}^{y_j}$	The label description degree of y_j	y	The random label variable
$\Delta_{\boldsymbol{x}}^k$	The kth label distribution margin	$L(f)$	The error of f

We can formalize LDL for classification as follows. First, given a loss function ℓ, an LDL function p is learned from S by solving the following problem:

$$\min_{\boldsymbol{W}} \sum_{i=1}^{n} \ell(D_i, p(\boldsymbol{x}_i; \boldsymbol{W})),$$

where p is typically a parametric model, such as the maximum entropy model [1] and deep neural network [6]. Second, an LDL classifier f can be defined according to the learned LDL function p by

$$f(\boldsymbol{x}) = \arg\max_{\bar{y} \in \mathcal{Y}} p_{\boldsymbol{x}}^{\bar{y}}, \tag{1}$$

where $p_{\boldsymbol{x}}^{\bar{y}}$ is the predicted label description degree of $\bar{y}$ to $\boldsymbol{x}$. That is, f regards the label with the optimal predicted label description degree as the predicted label. Our previous work [23] has studied the generalization of LDL for classification. Suppose the label distribution is the conditional probability distribution. Let y be the random label variable and $L(f)$ be the error of f defined by

$$L(f) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}, y} \left[\mathbb{I}(f(\boldsymbol{x}) \neq y) \right],$$

where $\mathbb{I}(\cdot)$ is the indicator function. Let L^* be the Bayes error [35]. Ref. [23] proved Theorem 1.

Theorem 1 ([23])**.** Let p be a learned LDL function and f be the classifier defined as (1). Then the error of f satisfies the following inequality:

$$L(f) - L^* \leqslant \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left[\sum_j |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}| \right]. \tag{2}$$

In (2), the left-hand side is the difference between the error of f and the Bayes error; the right-hand side is the expected L_1-norm loss of p. That is, LDL dominates classification [23]. It explains the generalization of LDL for classification but not the better generalization of LDL.

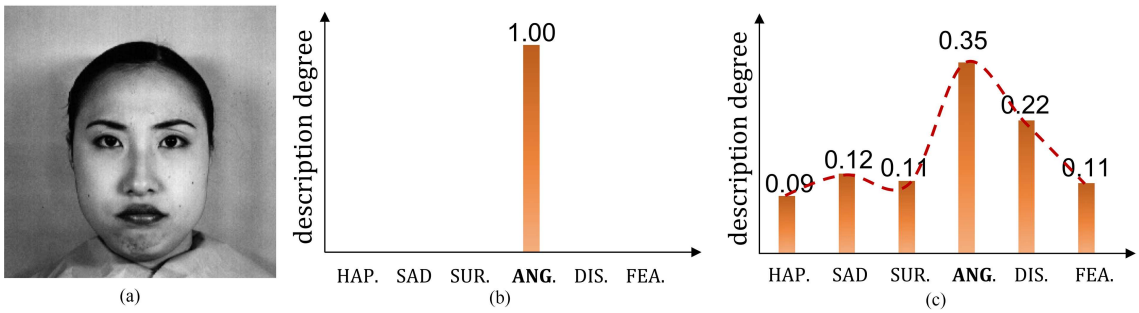
3.2 Label distribution margin theory

Label distribution margin was first introduced in our previous work [24]. It specifies a sufficient condition for LDL to select the optimal label in a label distribution. A label distribution contains rich information that includes not only the optimal label but also the sub-optimal label and even the kth optimal label. In this study, we generalize [24] and develop the label distribution margin theory.

For $\boldsymbol{x}$, let $\{y_{\boldsymbol{x}}^1, y_{\boldsymbol{x}}^2, \ldots, y_{\boldsymbol{x}}^m\}$ denote the sorted label set according to the label description degree of $\boldsymbol{x}$ in descending order[1)]. Generally, $y_{\boldsymbol{x}}^1$ should be the best choice for any classifier on $\boldsymbol{x}$ since it has the optimal label description degree, i.e., $y_{\boldsymbol{x}}^1 = \arg\max_{\bar{y} \in \mathcal{Y}} d_{\boldsymbol{x}}^{\bar{y}}$. For a classifier f that $f(\boldsymbol{x}) \neq y_{\boldsymbol{x}}^1$, its best decision on $\boldsymbol{x}$ should be $y_{\boldsymbol{x}}^2$ because $y_{\boldsymbol{x}}^2$ has the sub-optimal label description degree, i.e., $y_{\boldsymbol{x}}^2 = \arg\max_{\bar{y} \neq y_{\boldsymbol{x}}^1} d_{\boldsymbol{x}}^{\bar{y}}$. We can generalize this idea. For a classifier f, if $f(\boldsymbol{x}) \notin \{y_{\boldsymbol{x}}^j\}_{j \in [k-1]}$ ($k \geqslant 1$ and $k = 1$ implies an empty set), its best decision on $\boldsymbol{x}$ would be $y_{\boldsymbol{x}}^k$ that has the kth optimal label description degree, i.e.,

$$y_{\boldsymbol{x}}^k = \arg\max_{\bar{y} \in \mathcal{Y} \setminus \{y_{\boldsymbol{x}}^j\}_{j \in [k-1]}} d_{\boldsymbol{x}}^{\bar{y}}.$$

1) For example, for a label distribution {0.3, 0.1, 0.4, 0.2}, we have $y_{\boldsymbol{x}}^1 = y_3, y_{\boldsymbol{x}}^2 = y_1, y_{\boldsymbol{x}}^3 = y_4$, and $y_{\boldsymbol{x}}^4 = y_2$.



(a) (b) (c)

Figure 2 (Color online) Illustration of the difference between single-label and label distribution. (a) Image from the famous JAFFE [36] database with the ground-truth label "ANG.". (b) Single-label annotation, where only "ANG." has a degree of 1.0, and other labels all have degrees of 0's. (c) Annotation by a label distribution of $\{0.09, 0.12, 0.11, 0.35, 0.22, 0.11\}$, where all labels have label description degrees.

Our theory generalizes [24] to the kth optimal label. It proves a sufficient condition for an LDL classifier to predict $y_{\boldsymbol{x}}^k$ if its output is not in $\{y_{\boldsymbol{x}}^j\}_{j\in[k-1]}$. For each $\boldsymbol{x}$ and $1 \leqslant k \leqslant m-1$, define the kth label distribution margin by

$$\Delta_{\boldsymbol{x}}^k = d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}^k} - d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}^{k+1}}, \tag{3}$$

which is the difference between the kth optimal and the $(k+1)$th optimal label description degrees[2)]. We can prove the following theorem.

Theorem 2 (Label distribution margin theory)**.** Let p be a learned LDL function and f be the classifier defined as (1). For each $\boldsymbol{x}$ with $f(\boldsymbol{x}) \notin \{y_{\boldsymbol{x}}^j\}_{j\in[k-1]}$, if p satisfies

$$\sum_{\bar{y}\in\{y_{\boldsymbol{x}}^k,\ldots,y_{\boldsymbol{x}}^m\}} |p_{\boldsymbol{x}}^{\bar{y}} - d_{\boldsymbol{x}}^{\bar{y}}| \leqslant \Delta_{\boldsymbol{x}}^k, \tag{4}$$

we must have $f(\boldsymbol{x}) = y_{\boldsymbol{x}}^k$ for $k \geqslant 1$.

For the learned LDL function p, the kth label distribution margin $\Delta_{\boldsymbol{x}}^k$ provides a sufficient condition, under which the LDL classifier f guarantees to predict $y_{\boldsymbol{x}}^k$ on $\boldsymbol{x}$ if $f(\boldsymbol{x}) \notin \{y_{\boldsymbol{x}}^j\}_{j\in[k-1]}$. Especially, for $k=1$, $\{y_{\boldsymbol{x}}^j\}_{j\in[k-1]}$ is empty, that is, $\Delta_{\boldsymbol{x}}^1$ gives a sufficient condition for f outputting $y_{\boldsymbol{x}}^1$. Obviously, Ref. [24] is a special case of our label distribution margin theory of $k=1$.

3.3 Better generalization of LDL

Next, we explain the better generalization of LDL. We start with an example in Figure 2. Figure 2(a) demonstrates an image from the famous JAFFE [36] database with the ground-truth label "ANG."[3)]. Figure 2(b) displays the single-label annotation, where only the label "ANG." has a degree of 1.0 and other labels have degrees of 0's. Figure 2(c) shows the annotation by a label distribution[4)], where all labels have label description degrees. According to Figure 2, we can conduct the following analyses.

• For the label distribution, all labels have supervision information. Even if an LDL classifier neglects the optimal label "ANG.", it can still select the sub-optimal label "DIS." from the label distribution as the prediction.

• For the single-label, there is no supervision information for all labels except the optimal label "ANG.". As a result, if an SLL classifier misses "ANG.", it has no information and may even output the worst label "HAP." as the prediction.

The better generalization of LDL lies in the rich information of label distribution. Label distribution can guide an LDL classifier to choose the sub-optimal label even if it misses the optimal one, as explained in Figure 3. However, if an SLL classifier neglects the optimal label, it would have no information. The next theorem formalizes the above analyses and proves the better generalization of LDL.

2) Given a label distribution $\{0.5, 0.4, 0.1, 0.0, 0.0\}$ for $\boldsymbol{x}$, $\Delta_{\boldsymbol{x}}^1 = 0.5-0.4 = 0.1$, $\Delta_{\boldsymbol{x}}^2 = 0.4-0.1 = 0.3$, and $\Delta_{\boldsymbol{x}}^3 = 0.1-0.0 = 0.1$.

3) The filename of the image is "'KR.AN1.83", indicating the ground-truth label "ANG.".

4) The mean ratings for six emotions from 60 human annotators are 1.39, 1.87, 1.73, 4.58, 3.19, and 1.68 [36], which are then normalized into a label distribution [1].

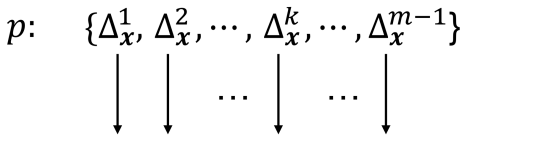


Figure 3 Explanation of Theorem 2. For the learned LDL function p, $\Delta_{\boldsymbol{x}}^k$ sets a sufficient condition under which the classifier f selects $y_{\boldsymbol{x}}^k$ on $\boldsymbol{x}$ if $f(\boldsymbol{x}) \notin \{y_{\boldsymbol{x}}^1, y_{\boldsymbol{x}}^2, \ldots, y_{\boldsymbol{x}}^{k-1}\}$.

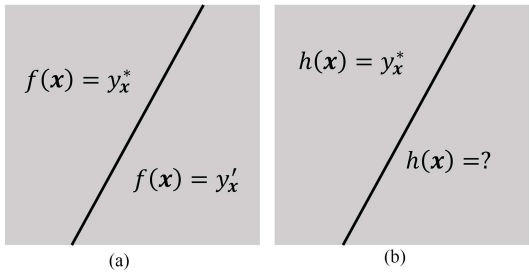


Figure 4 Interpretation of Theorem 3. (a) An LDL classifier f and (b) an SLL classifier h. The decision of f on the right region is $y_{\boldsymbol{x}}^2$ if Eq. (5) holds. However, for h, the prediction on the right region is not guaranteed.

Theorem 3. Let p be a learned LDL function, f be the LDL classifier, and h be an SLL classifier. Suppose (i) f and h use the same model[5)] and $\{\boldsymbol{x} : f(\boldsymbol{x}) = y_{\boldsymbol{x}}^1\} = \{\boldsymbol{x} : h(\boldsymbol{x}) = y_{\boldsymbol{x}}^1\}$; and (ii) p satisfies

$$\sum_{j=1}^{m} |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}| \leqslant \Delta_{\boldsymbol{x}}^2 \tag{5}$$

for all $\boldsymbol{x}$ in $\{\boldsymbol{x} : f(\boldsymbol{x}) \neq y_{\boldsymbol{x}}^1\}$. Then, we have $L(f) \leqslant L(h)$.

In Theorem 3, the first assumption says that $\{\boldsymbol{x} : f(\boldsymbol{x}) = y_{\boldsymbol{x}}^1\} = \{\boldsymbol{x} : h(\boldsymbol{x}) = y_{\boldsymbol{x}}^1\}$, and the second assumption says that in the complementary set, i.e., $\{\boldsymbol{x} : f(\boldsymbol{x}) \neq y_{\boldsymbol{x}}^1\}$, Eq. (5) always holds. Figure 4 interprets Theorem 3. The solid lines are the decision boundaries of $f(x) = y_{\boldsymbol{x}}^1$ and $h(x) = y_{\boldsymbol{x}}^1$. For an LDL classifier f, its decision in the right region is $y_{\boldsymbol{x}}^2$ if Eq. (5) holds. For an SLL classifier h, its prediction in the right region is not guaranteed. Suppose the left region of Figure 4(a) equals that of Figure 4(b) and Eq. (5) holds; then f has a smaller error than h.

4 The LDL-LDML method

4.1 Algorithm formulation

Next, we refer to Theorem 3 and present the details of the LDL-LDML approach. First, we consider assumption (i) of Theorem 3; suppose both LDL and SLL use a neural network as the learning model and apply the cross-entropy loss to learn the optimal label. Second, to satisfy assumption (ii) of Theorem 3, we design the label distribution margin loss (LDML) as

$$\ell_{\mathrm{LDML}}(p, \boldsymbol{x}) = \begin{cases} 0, & \text{if } \sum_j |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}| \leqslant \Delta_{\boldsymbol{x}}^2, \\ \sum_j |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|, & \text{otherwise.} \end{cases} \tag{6}$$

It is noteworthy that when assumption (ii) holds, LDML equals 0; otherwise, LDML equals the L_1-norm loss and raises a positive penalty. As a result, optimizing this loss function will encourage p to meet assumption (ii). We combine these two loss functions, i.e., $\ell = \ell_{\mathrm{CE}} + \lambda \ell_{\mathrm{LDML}}$, where ℓ_{CE} denotes the cross-entropy loss, and λ is a trade-off parameter. The optimization objective for LDL-LDML is

$$\mathcal{L} = \sum_{i=1}^{n} -\ln p_{\boldsymbol{x}_i}^{y_{\boldsymbol{x}_i}^1} + \lambda \cdot \sum_{i=1}^{n} \ell_{\mathrm{LDML}}(p, \boldsymbol{x}_i). \tag{7}$$

In the objective function (7), the first term learns the optimal label with the cross-entropy loss in order to satisfy assumption (i) of Theorem 3. It also helps address the objective inconsistency [23] (refer to Subsection 2.2) by focusing on the optimal label. The second term optimizes LDML for the sake of satisfying assumption (ii) of Theorem 3. Thus, LDL-LDML implements Theorem 3 from an algorithmic perspective.

5) This assumption avoids the difference of generalization caused by different models. Note that $y_{\boldsymbol{x}}^1$ has the highest degree in both label distribution and single-label, so we assume $\{\boldsymbol{x} : f(\boldsymbol{x}) = y_{\boldsymbol{x}}^1\} = \{\boldsymbol{x} : h(\boldsymbol{x}) = y_{\boldsymbol{x}}^1\}$. More discussion is left to Section 6.

Table 2 Characteristics of the experimental datasets.

ID.	Dataset	#Instances	#Features	#Labels	ID.	Dataset	#Instances	#Features	#Labels
1	Alpha	2465	24	18	9	Spo5	2465	24	3
2	Cdc	2465	24	15	10	Scene	2000	294	9
3	Cold	2465	24	4	11	SBU_3DFE	2500	243	6
4	Diau	2465	24	7	12	SJAFFE	213	243	6
5	Dtt	2465	24	4	13	Movie	7755	1869	5
6	Elu	2465	24	14	14	SCUT-FBP	1500	300	5
7	Heat	2465	24	6	15	FBP5500	5,500	512	5
8	Spo	2465	24	6	16	M2B	1240	250	5

4.2 Theoretical analysis

Define $\mathcal{C} = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \neq y^1_{\boldsymbol{x}}\}$. Let $\mathcal{D}_{\mathcal{C}}$ be the underlying distribution over $\mathcal{C}$. Define the error of f on $\mathcal{C}$ by $L_{\mathcal{C}}(f) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{C}},y}[\mathbb{I}(f(\boldsymbol{x}) \neq y)]$. Let f^* denote the classifier that always outputs the sub-optimal label, i.e., $f^*(\boldsymbol{x}) = y^2_{\boldsymbol{x}}$, and let $L_{\mathcal{C}}(f^*)$ stand for the error of f^* on $\mathcal{C}$. $L_{\mathcal{C}}(f^*)$ is the possible minimal error for f on $\mathcal{C}$, and $L_{\mathcal{C}}(f) = L_{\mathcal{C}}(f^*)$ if Eq. (5) holds for all $\boldsymbol{x} \in \mathcal{C}$. Next, we analyze the error of f on $\mathcal{C}$ and show the relationship between $L_{\mathcal{C}}(f)$ and $L_{\mathcal{C}}(f^*)$.

Theorem 4. Let p be a learned LDL function and f be the LDL classifier defined as (1). The error of f on $\mathcal{C}$ satisfies the following inequality:

$$L_{\mathcal{C}}(f) - L_{\mathcal{C}}(f^*) \leqslant \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{C}}}[\ell_{\mathrm{LDML}}(p, \boldsymbol{x})]. \tag{8}$$

In (8), the left-hand side is the difference between the error of f on $\mathcal{C}$ and that of f^*, and the right-hand side is the expected label distribution margin loss on $\mathcal{C}$. That is, the error of f on $\mathcal{C}$ is close to that of f^* if the expected label distribution margin loss approaches 0. To minimize the error of f on $\mathcal{C}$, it suffices to minimize the label distribution margin loss. We can make the following analyses.

- According to Theorem 1, for an LDL classifier f, we have $L_{\mathcal{C}}(f) - L_{\mathcal{C}}(f^*) \leqslant \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{C}}}[\sum_j |p^{y_j}_{\boldsymbol{x}} - d^{y_j}_{\boldsymbol{x}}|]$ [23]. By the definition of LDML, $\ell_{\mathrm{LDML}}(p, \boldsymbol{x}) \leqslant \sum_j |p^{y_j}_{\boldsymbol{x}} - d^{y_j}_{\boldsymbol{x}}|$. Thus, Eq. (8) is tighter, which implies that LDML helps improve generalization.
- For an SLL classifier h, $L_{\mathcal{C}}(h)$ is not bounded, as discussed above. In comparison, for an LDL classifier f, $L_{\mathcal{C}}(f)$ is bounded by the expected label distribution margin loss (or L_1-norm loss). As a result, LDL may have better generalization than SLL.

5 Experiments

In this section, we conduct experiments to justify our theories and the competitive performance of LDL-LDML. We carry out the experiments on a Linux server with 2.7 GHz CPU and 62 GB memory.

5.1 Experimental configurations

5.1.1 *Experimental datasets*

We conduct the experiments extensively on sixteen datasets with label distributions, characteristics of which are listed in Table 2. Let #Instances, #Features, and #Labels denote the numbers of instances, features, and labels, respectively.

The first thirteen datasets[6)] are collected by Geng [1]. Specifically, the first nine datasets (from Alpha to Spo5) are taken from the genome-wide expression data of yeast Saccharomyces cerevisiae [37]. The Scene is a dataset of natural images with ranking annotations, which are transformed into compatible label distributions [5]. The SBU_3DFE and SJAFFE are from two facial expression databases BU_3DFE [38] and JAFFE [36], respectively, where the mean ratings are normalized into a label distribution. The Movie is about user ratings on movies [25]. The last three datasets are about facial beauty perception. We preprocess SCUT-FBP[7)] [19] and M2B [39] as [40]. For FBP5500[8)] [18], we borrow the trained ResNet [41] model provided by the authors to extract 512-dimensional features.

6) http://palm.seu.edu.cn/xgeng/LDL/index.htm.
7) http://www.hcii-lab.net/data/SCUT-FBP/EN/download.html.
8) https://github.com/HCIILAB/SCUT-FBP5500-Database-Release.

Table 3 Comparison results (mean±std.%) between LR and SA-BFGS in terms of error probability loss. We highlight the best results in boldface and use •/◦ to indicate whether SA-BFGS is statistically better/worse than LR (at a confidence level of 0.05).

Algorithm	Alpha	Cdc	Diau	Elu	Heat	Cold	Dtt	Spo
LR	94.44±0.02•	93.04±0.02•	84.64±0.09•	92.86±0.02•	82.58±0.05	73.14±0.06•	74.20±0.07	81.02±0.15•
SA-BFGS	**94.29±0.01**	**92.90±0.01**	**84.30±0.04**	**92.63±0.02**	**82.51±0.06**	**73.04±0.09**	**74.18±0.05**	**81.00±0.13**
Algorithm	Spo5	SJAFFE	SBU_3DFE	Scene	Movie	SCU-FBP	FBP5500	M2B
LR	65.38±0.26	83.12±1.45•	**79.55±0.30**	68.45±0.98•	**69.04±0.08**	**63.22±0.79**◦	47.49±0.30•	59.89±1.18
SA-BFGS	**65.46±0.24**	**81.89±0.31**	79.73±0.26	**66.98±0.96**	69.32±0.23	67.15±1.17	**44.71±0.23**	**59.47±1.08**

5.1.2 *Evaluation metrics*

Our previous work [23] uses the 0/1 loss to estimate the classification performance of LDL methods, where the optimal label in the label distribution is regarded as the ground-truth label. However, the 0/1 loss cannot further estimate the generalization of a classifier when its prediction is not the optimal label. For example, given a label distribution $\{0.4, 0.3, 0.2, 0.1, 0.0\}$, the optimal label is y_1. If the predicted label is y_1, the 0/1 loss is 0. However, the 0/1 losses for predicting y_2, y_3, y_4, and y_5 are all 0's, incapable of distinguishing them. To justify our theories and evaluate the generalization of the comparing methods, we use the error probability loss [42] defined as

$$\ell_{\mathrm{EP}}(y, f(\boldsymbol{x})) = \mathbb{E}_y\left[\mathbb{I}(f(\boldsymbol{x}) \neq y)\right] = 1 - \mathbb{P}(y = f(\boldsymbol{x}) \mid \boldsymbol{x}) = 1 - d_{\boldsymbol{x}}^{f(\boldsymbol{x})},$$

where y is a random variable. The last equation holds because the label distribution is assumed to be the conditional probability distribution. For the preceding example, the error probability loss of predicting y_1 is the smallest, i.e., $1 - 0.4 = 0.6$; the error probabilities for predicting y_2, y_3, y_4, and y_5 are 0.7, 0.8, 0.9, and 1.0, respectively, which can distinguish the sub-optimal label y_2 from other labels.

5.2 Comparison between LDL and SLL algorithms

As we have claimed, the better generalization of LDL is due to the rich supervision information of label distribution. To justify that, we compare LDL methods with SLL ones:

- AA-kNN [1] against kNN. AA-kNN adopts the kNN algorithm to LDL. For each unknown instance, AA-kNN calculates the mean of the label distributions of its k nearest neighbors as the prediction.
- SA-BFGS [1] against logistic regression (LR). SA-BFGS employs the maximum entropy model to learn label distribution. Essentially, SA-BFGS can be viewed as a multivariate LR.

For AA-kNN and kNN, k is set to 5. For SA-BFGS and LR, the regularization parameter λ is set to 0.01. Notice that each LDL approach has the same model and parameters as its SLL counterpart. The only difference is that the former learns label distribution while the latter learns single-label. We evaluate their generalization performance in terms of error probability loss. We run each method for ten times random data partitions with half the samples for training and the other half for testing.

Tables 3 and 4 report the comparison results (mean±std.%) between LDL and SLL algorithms. We conduct the pairwise t-tests for each LDL method against the SLL one and let •/◦ denote whether the former is statistically better/worse than the latter (at a significance level of 0.05). According to Tables 5 and 6, LR achieves statistically better performance than SA-BFGS on SCU-FBP. As discussed in Subsection 2.2, the reason lies in that SA-BFGS is a general-purpose LDL method and faces the challenge of objective inconsistency [23]. Moreover, SA-BFGS and AA-kNN outperform LR and kNN in 75.00% (12 out of 16) and 93.75% (15 out of 16) of the cases, respectively. Besides, SA-BFGS and AA-kNN achieve win/tie/loss counts of 9/6/1 and 15/1/0 against LR and kNN, respectively. The LDL methods statistically outperform the SLL ones in 78.13% of the cases, which justifies the better generalization of the LDL methods. Since each LDL method only differs from its SLL counterpart in learning label distribution, the results validate the key to the better generalization of LDL lying in label distribution.

As explained in Subsection 3.3, when an LDL classifier misses the optimal label, it can still select the sub-optimal label as the prediction, which helps improve its generalization. To investigate that, we analyze the distribution of the predicted labels by LDL and SLL methods. Figures 5 and 6 report the ratios of the predicted labels by each approach that equal the optimal labels, sub-optimal labels, and others. From Figures 5 and 6, AA-kNN and SA-BFGS are more likely than kNN and LR to select the sub-optimal labels when they miss the optimal ones, and thereby achieve smaller error probability losses. Specifically, SA-BFGS has the same ratio of 22.4% in selecting the optimal labels as LR on SJAFFE,

Table 4 Comparison results (mean±std.%) between kNN and AA-kNN in terms of error probability loss. We highlight the best results in boldface and use •/◦ to indicate whether AA-kNN is statistically better/worse than kNN (at a confidence level of 0.05).

Algorithm	Alpha	Cdc	Diau	Elu	Heat	Cold	Dtt	Spo
kNN	94.40±0.03•	93.16±0.02•	84.71±0.07•	92.83±0.03•	82.72±0.04•	**73.58±0.07**	74.45±0.07•	82.03±0.15•
AA-kNN	**94.35±0.03**	**93.08±0.03**	**84.49±0.08**	**92.70±0.02**	**82.61±0.07**	73.59±0.10	**74.39±0.08**	**81.82±0.10**
Algorithm	Spo5	SJAFFE	SBU_3DFE	Scene	Movie	SCU_3DFE	FBP5500	M2B
kNN	65.47±0.16•	79.85±0.85•	79.87±0.13•	70.01±0.72•	68.99±0.17•	56.74±0.68•	45.33±0.22•	56.20±0.92•
AA-kNN	**65.20±0.11**	**78.64±0.74**	**79.62±0.22**	**68.57±0.72**	**68.31±0.14**	**54.99±0.54**	**44.95±0.18**	**55.62±1.02**

Table 5 Experimental results (mean±std.%) of the comparing methods in terms of 0/1 loss. We highlight the best results in bold and let •/◦ denote whether LDL-LDML is statistically better/worse than each comparing method (at a confidence level of 0.05).

Dataset	AA-kNN	SA-BFGS	LDL-SCL	LDL-LDM	RWLM-LDL	LDL-LDML
Alpha	90.24±0.45•	89.79±0.71•	89.49±1.86•	89.71±0.59•	81.89±0.70•	**79.06±0.65**
Cdc	87.24±0.96•	83.30±0.51	82.60±0.62	83.25±0.52	**82.27±0.89**	83.20±0.89
Diau	71.13±0.94•	70.11±0.81•	71.18±0.68•	70.06±0.79•	70.26±2.11•	68.08±1.98•
Elu	89.39±0.90•	90.44±0.70•	91.10±1.13•	90.41±0.73•	83.98±0.54•	**81.01±0.63**
Heat	72.20±1.01•	70.85±0.66•	74.19±2.09•	70.82±0.61•	70.89±1.58•	**69.33±1.18**
Cold	61.69±1.52•	58.12±0.80•	57.80±0.90•	63.37±0.84•	62.18±1.27•	**57.69±0.77**
Dtt	66.63±1.19•	63.43±0.79•	64.62±1.79•	63.37±0.84•	**63.11±0.93**	64.49±1.05◦
Spo	66.39±0.87•	56.97±0.84•	55.87±0.76•	56.91±0.86•	59.90±0.97•	**54.87±0.76**
Spo5	57.91±0.99•	56.93±0.95•	56.23±2.68•	56.77±0.95•	57.97±2.02•	**54.48±1.20**
SJAFFE	63.93±2.48•	78.97±3.08•	73.18±8.50•	51.87±3.93•	51.59±3.73•	**47.29±3.30**
SBU_3DFE	67.29±1.26•	66.18±1.41•	**54.69±1.27**	60.26±1.78•	60.38±1.03•	55.41±1.36
Scene	62.82±0.86•	61.33±1.04•	72.38±2.48•	59.85±1.23•	55.26±2.08•	**44.26±1.72**
Movie	44.80±0.75•	49.10±1.11•	45.47±0.98•	47.27±0.87•	**43.16±0.85**	43.51±0.49
SCU-FBP	47.17±1.34•	62.07±1.97•	68.27±7.82•	48.01±0.73•	47.15±1.01•	**45.97±1.03**
FBP5500	23.75±0.89	23.46±0.90	23.83±0.94•	23.83±1.07•	23.16±0.75	**23.04±1.01**
M2B	50.40±0.56•	56.95±1.17•	49.48±1.16	56.81±1.45•	51.18±1.63•	**49.45±1.15**
W./T./L. counts	15/1/0	14/2/0	13/3/0	15/1/0	12/4/0	–

but it achieves a higher one of 27.1% in selecting the sub-optimal labels than LR, which explains why SA-BFGS generalizes better than LR.

In summary, the experimental results validate that (1) the better generalization of LDL benefits from the rich information of label distribution, and (2) LDL methods are more likely to select the sub-optimal labels than SLL ones when missing the optimal labels. The results further justify our explanation of the better generalization of LDL.

5.3 Experimental results of LDL-LDML

LDL-LDML is inspired by Theorem 3. To support Theorem 3 and validate the performance of LDL-LDML, we compare it with AA-kNN, SA-BFGS, and three state-of-the-art LDL methods.

• LDL-SCL [43]: It first encodes label correlation as additional features and then jointly learns label distribution and the encoding of label correlation. Its parameters are tuned as suggested by [43].

• LDL-LDM [40]: It learns label distribution manifold to model label correlation in LDL and exploits both global and local label correlations. We tune its parameters as suggested by [40].

• RWLM-LDL [23]: It re-weights instances w.r.t. the entropy of label distribution and employs a large margin to solve the objective inconsistency [23]. It is a specially designed LDL method for classification. We set $\lambda_1 = 0.0001$, $\lambda_2 = 1$, and $\rho = 0.1$, as given by [23].

We implement LDL-LDML by a three-layer-neural network with 64 neurons in the hidden layer, apply the Adam [44] algorithm with a learning rate of 0.01 and batch size of 64, and set $\lambda = 1$. We run each method for ten random data partitions with half the samples for training and the other half for testing.

Tables 5 and 6 tabulate the experimental results (mean±std.%) of the comparing approaches in terms of 0/1 loss and error probability loss, respectively. We conduct the pairwise t-tests for LDL-LDML against each comparing method and let •/◦ denote whether the former is statistically better/worse than the latter (at a confidence level of 0.05). We summarize the win/tie/loss counts for LDL-LDML against each method in the last rows in Tables 5 and 6.

Table 6 Experimental results (mean±std.%) of the comparing methods in terms of error probability loss. We highlight the best results in bold and let •/○ denote whether LDL-LDML is statistically better/worse than each comparing method (at a confidence level of 0.05).

Dataset	AA-kNN	SA-BFGS	LDL-SCL	LDL-LDM	RWLM-LDL	LDL-LDML
Alpha	94.35±0.01•	94.30±0.01•	94.30±0.02•	94.30±0.01•	94.43±0.03•	**94.27±0.02**
Cdc	93.09±0.02•	92.92±0.02	92.88±0.01	**92.92±0.02**	92.95±0.02	92.96±0.06
Diau	84.51±0.06•	84.31±0.03	**84.28±0.04**	84.30±0.03	84.69±0.05•	84.31±0.04
Elu	92.70±0.02•	92.70±0.014•	92.70±0.01•	92.69±0.01•	92.81±0.03•	**92.64±0.03**
Heat	82.58±0.08•	82.48±0.08	82.55±0.11•	82.48±0.08	82.67±0.13•	**82.45±0.07**
Cold	73.54±0.13•	73.02±0.06•	72.98±0.07	74.19±0.06•	73.54±0.10•	**72.95±0.07**
Dtt	74.41±0.05•	74.20±0.06	74.26±0.08	74.19±0.06	**74.17±0.06**○	74.31±0.09
Spo	81.85±0.10•	81.12±0.11•	81.13±0.12•	81.12±0.11•	81.41±0.18•	**81.02±0.12**
Spo5	65.21±0.24•	65.29±0.16•	65.64±0.26•	65.28±0.16•	65.81±0.25•	**65.06±0.20**
SJAFFE	79.29±0.64•	81.69±0.29•	81.07±1.79•	77.07±0.76	77.92±0.83•	**76.72±0.47**
SBU_3DFE	79.69±0.22•	79.85±0.34•	**77.00±0.30**	77.95±0.46•	78.46±0.24•	**77.00±0.28**
Scene	68.14±0.53•	66.86±0.57•	75.03±1.83•	65.94±0.64•	72.11±1.36•	**65.18±0.63**
Movie	68.37±0.17•	69.34±0.22•	68.88±0.35•	68.85±0.20•	**67.92±0.10**○	68.17±0.07
SCU-FBP	55.10±0.62•	66.70±0.90•	70.54±6.34•	55.66±0.45•	55.08±0.32•	**54.05±0.45**
FBP5500	44.96±0.28	44.85±0.27	44.88±0.30	**44.63±0.28**	44.65±0.23	44.87±0.40
M2B	55.58±0.50•	60.21±0.71•	55.76±0.50•	60.14±0.76•	55.70±0.73•	**54.76±0.51**
W./T./L. counts	15/1/0	11/5/0	10/6/0	10/6/0	12/2/2	–

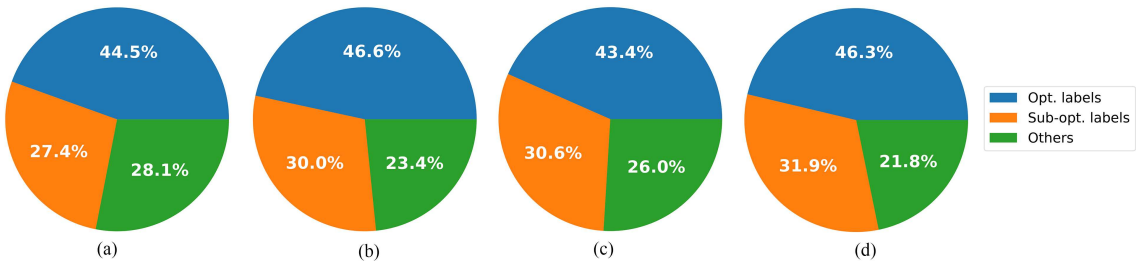


Figure 5 (Color online) Distributions of the predicted labels of (a) kNN ($\ell_{\mathrm{EP}} = 59.59$), (b) AA-$k$NN ($\ell_{\mathrm{EP}} = 58.04$), (c) LR ($\ell_{\mathrm{EP}} = 59.59$), and (d) SA-BFGS on M2B ($\ell_{\mathrm{EP}} = 57.73$).

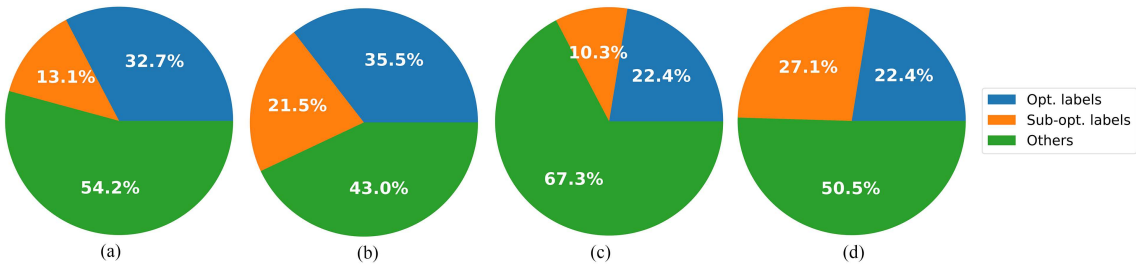


Figure 6 (Color online) Distributions of the predicted labels of (a) kNN ($\ell_{\mathrm{EP}} = 80.36$), (b) AA-$k$NN ($\ell_{\mathrm{EP}} = 79.50$), (c) LR ($\ell_{\mathrm{EP}} = 83.83$), and (d) SA-BFGS ($\ell_{\mathrm{EP}} = 81.65$) on SJAFFE.

From Tables 5 and 6, LDL-LDML statistically outperforms AA-kNN, SA-BFGS, LDL-SCL, and LDL-LDM. These four methods are general-purpose LDL methods that aim to learn label distribution and may neglect the optimal labels [23]. LDL-LDML, in comparison, learns the optimal labels by the cross-entropy loss, which helps improve its classification performance. Moreover, RWLM-LDL is a specially designed LDL method for classification [23] that applies re-weighting and a large margin to learn the optimal label and achieves the optimal results on Dtt and Movie. LDL-LDML achieves statistically better performance than RWLM-LDL on all datasets except Dtt and Movie because it considers the optimal and sub-optimal labels by optimizing LDML, which further brings better classification performance. To summarize, the results validate the competitive performance of LDL-LDML and support Theorems 3 and 4.

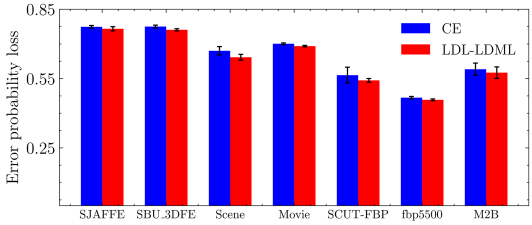


Figure 7 (Color online) Comparison results in terms of error probability loss on the last seven datasets.

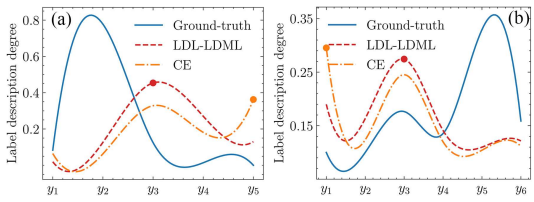


Figure 8 (Color online) Comparisons between LDL-LDML and CE on two typical examples of (a) SCUT-FBP and (b) SBU_3DFE.

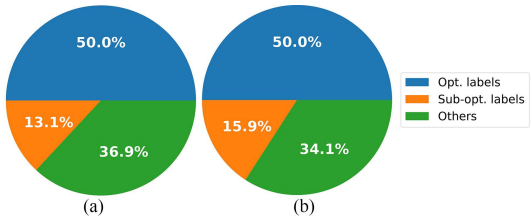


Figure 9 (Color online) Distribution of the predicted labels of (a) CE ($\ell_{\mathrm{EP}} = 77.12$) and (b) LDL-LDML ($\ell_{\mathrm{EP}} = 76.50$) on SJAFFE.

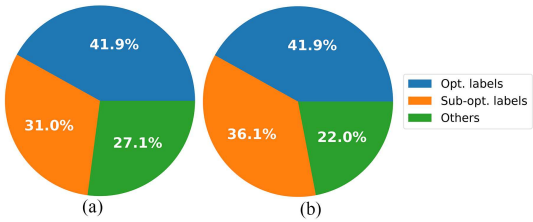


Figure 10 (Color online) Distribution of the predicted labels of (a) CE ($\ell_{\mathrm{EP}} = 61.59$) and (b) LDL-LDML ($\ell_{\mathrm{EP}} = 59.08$) on M2B.

5.4 Ablation study

Optimizing LDML encourages an LDL method to select the optimal labels when it misses the optimal ones, which helps improve its generalization. In this subsection, we conduct an ablation study to investigate the advantages of LDML. First, we set $\lambda = 0$ in LDL-LDML (only with cross-entropy loss) and define it as CE. Notice that LDL-LDML only differs from CE in applying LDML to learn label distribution. We run CE and LDL-LDML on the sixteen datasets for ten times random data partitions (50% for training and 50% for testing). Figure 7 reports the detailed comparison results between CE and LDL-LDML on the last seven datasets in terms of error probability loss. Besides, we conduct the Wilcoxon signed-rank tests [45] for LDL-LDML against CE. The p-value of the tests equals 3.05E−5. That is, LDL-LDML statistically outperforms CE at a confidence level of 0.05, which indicates that LDML can truly improve the generalization of LDL-LDML.

To further elucidate the benefit of LDML, Figure 8 visualizes two typical examples on SCUT-FBP and SBU_3DFE, where the ground-truth label distributions and the predicted label distributions by CE and LDL-LDML are plotted. As shown in Figure 8, when missing the optimal labels, CE takes the labels with the minimal label description degrees as the predicted labels. In comparison, LDL-LDML benefits from LDML and selects the sub-optimal labels as the prediction, resulting in better generalization. Moreover, Figures 9 and 10 analyze the ratios of the predicted labels as the optimal labels, sub-optimal labels, and others. Compared with CE, LDL-LDML tends to select the sub-optimal labels as the prediction when missing the optimal ones, which explains why LDL-LDML generalizes better than CE and justifies the usefulness of LDML.

5.5 Further analysis

5.5.1 *Extension to other LDL methods*

LDML can be extended to other LDL approaches. Here, we extend LDML to two existing LDL methods, AA-BP [1] and LDL-HR [46]. AA-BP applies the back-propagation neural network with the sum-squared loss to learn label distribution; we add LDML to the objective function of AA-BP with a trade-off of 1 (denoted by AA-BP w/ LDML). LDL-HR is a specialized LDL method for classification that learns the highest label and the rest label description degrees [46]. To extend LDML to LDL-HR, we replace the loss function for learning the rest label description degrees with LDML (denoted by LDL-HR w/ LDML). Following the same evaluation protocol as in Subsection 5.3, Table 7 reports the performance of these methods in terms of error probability loss. According to Table 7, AA-BP and LDL-HR with LDML

Table 7 Comparison results (mean±std.%) of LDL methods against the counterparts with LDML. The best results are highlighted in boldface.

Algorithm	SJAFFE	SBU_3DFE	Scene	Movie	SCUT-FBP	FBP5500	M2B
AA-BP	82.84±1.39	81.97±0.73	77.02±2.08	68.16±0.31	**54.05±1.27**	44.15±0.55	54.36±2.81
AA-BP w/ LDML	**81.86±1.16**	**80.66±0.78**	**65.39±2.62**	**67.72±0.31**	54.07±1.20	**43.88±0.68**	**54.08±2.22**
LDL-HR	83.64±3.03	79.47±0.76	67.39±2.53	67.92±0.26	54.96±1.04	45.54±0.63	54.36±2.81
LDL-HR w/ LDML	**77.16±2.11**	**79.27±0.71**	**64.86±2.55**	**67.90±0.23**	**54.80±0.93**	**44.20±0.71**	**54.17±2.08**

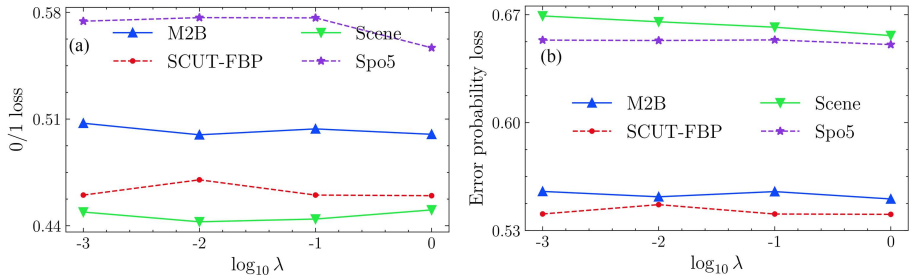


Figure 11 (Color online) Performance of LDL-LDML with λ varying from {0.001, 0.01, 0.1, 1} on Scene, M2B, SCUT-FBP, and Spo5 in terms of (a) 0/1 loss and (b) error probability loss.

outperform AA-BP and LDL-HR by a margin, respectively. That is, leveraging LDML helps improve the generalization of other LDL methods.

5.5.2 *Parameter sensitivity*

LDL-LDML has a trade-off parameter λ, which balances CE and LDML. To study its sensitivity, we run LDL-LDML with λ varying from the candidate set $\{0.001, 0.01, 0.1, 1\}$. Figures 11(a) and (b) report the performance of LDL-LDML in terms of 0/1 loss and error probability loss, respectively. According to Figure 11, $\lambda = 1$ brings better performance. Hence, we simply set it to 1 in the experiments.

6 Conclusion and discussion

This paper studies the better generalization of LDL and answers why LDL generalizes better than traditional SLL. We disclose that label distribution has rich information, so an LDL method can choose, if not the optimal label, at least the sub-optimal label from label distribution. This may bring better generalization than SLL because there is nothing to choose from in single-label if an SLL method misses the optimal label. We develop the label distribution margin theory to prove the explanation and propose a novel LDL method called LDL-LDML. In the experiments, the LDL baselines outperform the SLL methods by a margin, and LDL-LDML achieves competitive performance compared with several state-of-the-art LDL methods, which validate our explanations and theories.

The label distribution margin theory (Theorem 2) presents a condition for an LDL method to select the kth optimal label. Although in this study we only consider the sub-optimal ($k = 2$) label, one may further improve the generalization of LDL by considering the kth optimal label for $k > 2$. In Theorem 3, the first assumption says that f (LDL) and h (SLL) use the same model. We make this assumption to avoid the difference of generalization caused by different models. For example, if h uses a deep neural network (DNN) and f applies a linear model, h probably generalizes better than f due to the learnability of DNN. The second assumption says that Eq. (5) holds for all instances on which f neglects the optimal labels. Although this assumption is strong, it can guide the design of new LDL methods. For example, we propose to optimize LDML to satisfy it as much as possible. In the future, we will explore how to weaken this assumption.

Our explanation and theories only apply to SLL. However, LDL has also shown advantages over traditional MLL [23, 47]. In the future, we will extend our theory to MLL and try to answer why LDL generalizes better than MLL.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62306073, 62125602, 62076063), Natural Science Foundation of Jiangsu Province (Grant No. BK20230832), and China Postdoctoral Science Foundation (Grant No. 2022M720028).

References

1 Geng X. Label distribution learning. IEEE Trans Knowl Data Eng, 2016, 28: 1734–1748
2 Xu M, Guo L-Z. Learning from group supervision: the impact of supervision deficiency on multi-label learning. Sci China Inf Sci, 2021, 64: 130101
3 Liu W W, Wang H B, Shen X B, et al. The emerging trends of multi-label learning. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 7955–7974
4 Zhang M L, Zhou Z H. ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn, 2007, 40: 2038–2048
5 Geng X, Luo L R. Multilabel ranking with inconsistent rankers. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 3742–3747
6 Gao B B, Xing C, Xie C W, et al. Deep label distribution learning with label ambiguity. IEEE Trans Image Process, 2017, 26: 2825–2838
7 Shen W, Zhao K, Guo Y L, et al. Label distribution learning forests. In: Proceedings of Conference on Neural Information Processing Systems, Long Beach, 2017. 834–843
8 Geng X, Yin C, Zhou Z-H. Facial age estimation by learning from label distributions. IEEE Trans Pattern Anal Mach Intell, 2013, 35: 2401–2412
9 Wen X, Li B Y, Guo H Y, et al. Adaptive variance based label distribution learning for facial age estimation. In: Proceedings of European Conference on Computer Vision, 2020. 379–395
10 Li S, Deng W H. Blended emotion in-the-wild: multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. Int J Comput Vis, 2019, 127: 884–906
11 Jia X Y, Zheng X, Li W W, et al. Facial emotion distribution learning by exploiting low-rank label correlations locally. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 9833–9842
12 Shu Y Z, Yang P, Liu N Q, et al. Emotion distribution learning based on peripheral physiological signals. IEEE Trans Affective Comput, 2023, 14: 2470–2483
13 Zhou X Z, Wei Z Q, Xu M, et al. Facial depression recognition by deep joint label distribution and metric learning. IEEE Trans Affective Comput, 2022, 13: 1605–1618
14 Geng X, Qian X, Huo Z W, et al. Head pose estimation based on multivariate label distribution. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 1974–1991
15 Yang J F, Sun M, Sun X X. Learning visual sentiment distributions via augmented conditional probability neural network. In: Proceedings of AAAI Conference on Artificial Intelligence, San Francisco, 2017. 224–230
16 Yang J F, She D Y, Sun M. Joint image emotion classification and distribution learning via deep convolutional neural network. In: Proceedings of International Joint Conference on Artificial Intelligence, Melbourne, 2017. 3266–3272
17 Wu X, Wen N, Liang J, et al. Joint acne image grading and counting via label distribution learning. In: Proceedings of IEEE/CVF International Conference on Computer Vision, Seoul, 2019. 10641–10650
18 Liang L, Lin L, Jin L, et al. SCUT-FBP5500: a diverse benchmark dataset for multi-paradigm facial beauty prediction. In: Proceedings of International Conference on Pattern Recognition, Beijing, 2018. 1598–1603
19 Xie D R, Liang L Y, Jin L W, et al. SCUT-FBP: a benchmark dataset for facial beauty perception. In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Hong Kong, 2015. 1821–1826
20 Fan Y Y, Liu S, Li B, et al. Label distribution-based facial attractiveness computation by deep residual learning. IEEE Trans Multimedia, 2018, 20: 2196–2208
21 Yi K, Wu J X. Probabilistic end-to-end noise correction for learning with noisy labels. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 7017–7025
22 Jiang L X, Zhang H, Tao F N, et al. Learning from crowds with multiple noisy label distribution propagation. IEEE Trans Neural Netw Learn Syst, 2022, 33: 6558–6568
23 Wang J, Geng X, Xue H. Re-weighting Large Margin Label Distribution Learning for Classification. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 5445–5459
24 Wang J, Geng X. Label distribution learning machine. In: Proceedings of International Conference on Machine Learning, 2021. 10749–10759
25 Geng X, Hou P. Pre-release prediction of crowd opinion on movies by label distribution learning. In: Proceedings of International Joint Conference on Artificial Intelligence, Buenos Aires, 2015. 3511–3517
26 Xing C, Geng X, Xue H. Logistic boosting regression for label distribution learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 4489–4497
27 Kontschieder P, Fiterau M, Criminisi A, et al. Deep neural decision forests. In: Proceedings of IEEE/CVF International Conference on Computer Vision, Santiago, 2015. 1467–1475
28 Shen W, Guo Y L, Wang Y, et al. Deep differentiable random forests for age estimation. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 404–419
29 Ren T T, Jia X Y, Li W W, et al. Label distribution learning with label-specific features. In: Proceedings of International Joint Conference on Artificial Intelligence, 2019. 3318–3324
30 Xu M, Zhou Z H. Incomplete label distribution learning. In: Proceedings of International Joint Conference on Artificial Intelligence, Melbourne, 2017. 3175–3181
31 Zhang J, Tao H, Luo T, et al. Safe incomplete label distribution learning. Pattern Recogn, 2022, 125: 108518
32 Gao B B, Zhou H Y, Wu J X, et al. Age estimation using expectation of label distribution learning. In: Proceedings of International Joint Conference on Artificial Intelligence, Stockholm, 2018. 712–718
33 Zhao P, Zhou Z H. Label distribution learning by optimal transport. In: Proceedings of AAAI Conference on Artificial Intelligence, New Orleans, 2018. 4506–4513
34 Wang J, Geng X. Theoretical analysis of label distribution learning. In: Proceedings of AAAI Conference on Artificial Intelligence, Hawaii, 2019. 5256–5263
35 Devroye L, Györfi L, Lugosi G. A Probabilistic Theory of Pattern Recognition. New York: Springer, 1996
36 Lyons M, Akamatsu S, Kamachi M, et al. Coding facial expressions with Gabor wavelets. In: Proceedings of International Conference on Automatic Face and Gesture Recognition, Nara, 1998. 200–205
37 Eisen M B, Spellman P T, Brown P O, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 1998, 95: 14863–14868
38 Yin L J, Wei X Z, Sun Y, et al. A 3D facial expression database for facial behavior research. In: Proceedings of International Conference on Automatic Face and Gesture Recognition, Southampton, 2006. 211–216
39 Nguyen T V, Liu S, Ni B, et al. Sense beauty via face, dressing, and/or voice. In: Proceedings of ACM International Conference on Multimedia, Nara, 2012. 239–248
40 Wang J, Geng X. Label distribution learning by exploiting label distribution manifold. IEEE Trans Neural Netw Learn Syst, 2023, 34: 839–852
41 He K M, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778
42 Wang J, Geng X. Classification with label distribution learning. In: Proceedings of International Joint Conference on Artificial Intelligence, 2019. 3712–3718

43 Jia X Y, Li Z C, Zheng X, et al. Label distribution learning with label correlations on local samples. IEEE Trans Knowl Data Eng, 2021, 33: 1619–1631
44 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: Proceedings of International Conference on Learning Representations, San Diego, 2015. 1–15
45 Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res, 2006, 7: 1–30
46 Wang J, Geng X. Learn the highest label and rest label description degrees. In: Proceedings of the International Joint Conference on Artificial Intelligence, 2021. 3097–3103
47 Zhang M L, Zhang Q W, Fang J P, et al. Leveraging implicit relative labeling-importance information for effective multi-label learning. IEEE Trans Knowl Data Eng, 2021, 33: 2057–2070

Appendix A Proof of Theorem 2

First, we introduce the following lemma.

Lemma A1 ([34]). Let a, b, c, and d be four real-values. If $a \geqslant b$ and $c \geqslant d$, then $|a-b| \leqslant |a-d| + |c-b|$.

Refer to [34] for the proof. Next, we formally give the proof of Theorem 2.

Proof. We prove this by contradiction. Suppose for the sake of contradiction that there exists one $\boldsymbol{x}$ such that $f(\boldsymbol{x}) \notin \{y_{\boldsymbol{x}}^j\}_{j\in[k-1]}$ and $f(\boldsymbol{x}) \neq y_{\boldsymbol{x}}^k$. Without loss of generality, let $f(\boldsymbol{x}) = y_l$ and $y_{\boldsymbol{x}}^k = y_j$ for $l \neq j$. According to the definition of $y_{\boldsymbol{x}}^k$ and f, it is trivial to have $d_{\boldsymbol{x}}^{y_j} \geqslant d_{\boldsymbol{x}}^{y_l}$ and $p_{\boldsymbol{x}}^{y_l} \geqslant p_{\boldsymbol{x}}^{y_j}$. Next, according to Lemma A1, it follows that

$$d_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_l} \leqslant |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}| + |p_{\boldsymbol{x}}^{y_l} - d_{\boldsymbol{x}}^{y_l}| < \sum_{\bar{y}\in\{y_{\boldsymbol{x}}^k,\ldots,y_{\boldsymbol{x}}^m\}} |p_{\boldsymbol{x}}^{\bar{y}} - d_{\boldsymbol{x}}^{\bar{y}}|.$$

By the definition of $\Delta_{\boldsymbol{x}}^k$, we have $\Delta_{\boldsymbol{x}}^k \leqslant d_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_l}$, which yields $\Delta_{\boldsymbol{x}}^k < \sum_{\bar{y}\in\{y_{\boldsymbol{x}}^k,\ldots,y_{\boldsymbol{x}}^m\}} |p_{\boldsymbol{x}}^{\bar{y}} - d_{\boldsymbol{x}}^{\bar{y}}|$. This contradicts with (4). Thus, we must have $f(\boldsymbol{x}) = y_{\boldsymbol{x}}^k$ for all $\boldsymbol{x}$ with $f(\boldsymbol{x}) \notin \{y_{\boldsymbol{x}}^j\}_{j\in[k-1]}$.

Appendix B Proof of Theorem 3

Proof. Let $\bar{\mathcal{C}}$ be the complementary set of $\mathcal{C}$ and $\mathcal{D}_{\bar{\mathcal{C}}}$ stand for the underlying distribution of $\bar{\mathcal{C}}$. First, by assumption (i), it is trivial to validate

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\bar{\mathcal{C}}},y}\left[\mathbb{I}(f(\boldsymbol{x}) \neq y)\right] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\bar{\mathcal{C}}},y}\left[\mathbb{I}(h(\boldsymbol{x}) \neq y)\right],$$

because both f and h output $y_{\boldsymbol{x}}^1$ as the prediction on $\bar{\mathcal{C}}$. To prove Theorem 3, it suffices to show

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{C}},y}\left[\mathbb{I}(f(\boldsymbol{x}) \neq y)\right] \leqslant \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{C}},y}\left[\mathbb{I}(h(\boldsymbol{x}) \neq y)\right]. \tag{B1}$$

Fix one $\boldsymbol{x} \in \mathcal{C}$. According to Theorem 2, if assumption (ii) holds, we must have $f(\boldsymbol{x}) = y_{\boldsymbol{x}}^2$ (a special case of Theorem 2 with $k = 2$). Then, the error of f on $\boldsymbol{x}$ is

$$\mathbb{P}\left(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x}\right) = 1 - \mathbb{P}(y = f(\boldsymbol{x}) \mid \boldsymbol{x}) = 1 - d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}^2}.$$

Without loss of generality, let $h(\boldsymbol{x}) = y_l$. Similarly, the error of h on $\boldsymbol{x}$ is $\mathbb{P}\left(h(\boldsymbol{x}) \neq y \mid \boldsymbol{x}\right) = 1 - \mathbb{P}(y = h(\boldsymbol{x}) \mid \boldsymbol{x}) = 1 - d_{\boldsymbol{x}}^{y_l}$. Apparently $d_{\boldsymbol{x}}^{y_l} \leqslant d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}^2}$, which yields $\mathbb{P}(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) \leqslant \mathbb{P}(h(\boldsymbol{x}) \neq y \mid \boldsymbol{x})$. Taking expectation on both sides w.r.t. $\mathcal{D}_{\mathcal{C}}$, we prove (B1) and complete the proof of Theorem 3.

Appendix C Proof of Theorem 4

Proof. Fix one $\boldsymbol{x} \in \mathcal{C}$. If $\sum_j |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}| \leqslant \Delta_{\boldsymbol{x}}^2$, $f(\boldsymbol{x}) = f^*(\boldsymbol{x})$ and

$$\mathbb{P}(f(\boldsymbol{x}) \neq y|\boldsymbol{x}) - \mathbb{P}(f^*(\boldsymbol{x}) \neq y|\boldsymbol{x}) = 0. \tag{C1}$$

We proceed with $\sum_j |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}| > \Delta_{\boldsymbol{x}}^2$. Without loss of generality, let $f(\boldsymbol{x}) = y_l$ and $y_{\boldsymbol{x}}^2 = y_u$. If $y_l = y_u$, Eq. (C1) still holds. If $y_l \neq y_u$, by the definitions of $\mathcal{C}$ and $y_{\boldsymbol{x}}^2$, we have $p_{\boldsymbol{x}}^{y_l} \geqslant p_{\boldsymbol{x}}^{y_u}$ and $d_{\boldsymbol{x}}^{y_u} \geqslant d_{\boldsymbol{x}}^{y_l}$. Applying Lemma A1, we have

$$\mathbb{P}(f(\boldsymbol{x}) \neq y|\boldsymbol{x}) - \mathbb{P}(f^*(\boldsymbol{x}) \neq y|\boldsymbol{x}) = d_{\boldsymbol{x}}^{y_u} - d_{\boldsymbol{x}}^{y_l} \leqslant |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}| + |p_{\boldsymbol{x}}^{y_u} - d_{\boldsymbol{x}}^{y_u}| \leqslant \sum_j |p_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|. \tag{C2}$$

Combine (C1) and (C2), yielding $\mathbb{P}(f(\boldsymbol{x}) \neq y|\boldsymbol{x}) - \mathbb{P}(f^*(\boldsymbol{x}) \neq y|\boldsymbol{x}) \leqslant \ell_{\mathrm{LDML}}(p, \boldsymbol{x})$. Taking expectations on both sides of this equation w.r.t. $\mathcal{D}_{\mathcal{C}}$, we complete the proof.