Learning Generalized Label Distributions

Haitao Wu[1] · Weiwei Li[2] · Kun Yue[3] · Xiuyi Jia[1]
[1]Nanjing University of Science and Technology
[2]Nanjing University of Aeronautics and Astronautics · [3]Yunnan University

1 Motivation

Single-label learning (SL) cannot model label ambiguity,
which motivates a variety of alternative representations:

- Logical Label (LL);
- Label Ranking (LR);
- Ternary Label (TL);
- Label Distribution (LD).

Among them, LD has attracted considerable attention.
However, in this work, we argue that LD suffers from three inherent limitations.

To address these issues, we propose an alternative representation:
Generalized Label Distribution (GLD).

2- Assumptions & Definitions

- Label Distribution (LD):

Assumption 2.1 (Projection-based label distribution)**.** Let $\boldsymbol{q} \in \prod_{j\in[c]}[a_j,\, b_j]_{\mathbb{R}}$ denote the raw data, where c is the number of labels and $[a_j,\, b_j]_{\mathbb{R}}$ is the value range of each label j. The corresponding LD $\boldsymbol{d} \in \Delta^{c-1}$ is obtained via a projection operator $\mathrm{proj}(\cdot)$, i.e., $\boldsymbol{d} = \mathrm{proj}(\boldsymbol{q})$, where

$$\Delta^{c-1} \triangleq \left\{ \boldsymbol{v} \in \mathbb{R}^c \,\middle|\, \mathbf{1}^\top \boldsymbol{v} = 1,\, \boldsymbol{v} \geq 0 \right\} \tag{1}$$

is the $(c-1)$-dimensional probability simplex and $\mathbf{1}$ is a c-dimensional vector of all ones. Each d_j of $\boldsymbol{d}$, namely *description degree*, indicates the degree to which the j-th label describes the sample.

Assumption 2.2 (Proportion-based label distribution)**.** Building on Assumption 2.1, the raw data $\boldsymbol{q}$ satisfies $\forall_{j\in[c]}(q_j \geq 0)$ and $\exists_{j\in[c]}(q_j > 0)$, then the LD is obtained by $\boldsymbol{d} = \boldsymbol{q}/\sum_{j\in[c]} q_j$.

- Logical Label (LL):

Definition 2.3 (Logical label)**.** The LL is a binary vector $\boldsymbol{l} \in \{0,\, 1\}^c$, where $l_j = 1$ indicates that the j-th label $y_j \in \mathcal{Y}$ is relevant to the sample, and $l_j = 0$ otherwise.

- Ternary Label (TL):

Definition B.1 (Ternary label)**.** The TL is a ternary vector $\boldsymbol{t} \in \{-1,\, 0,\, 1\}^c$, where $t_j = 1$ indicates that the j-th label $y_j \in \mathcal{Y}$ is positively associated with the sample, $t_j = -1$ indicates a negative association, and $t_j = 0$ means the label is irrelevant.

- Label Ranking (LR):

Definition B.2 (Label ranking)**.** The LR is defined as a label vector $\boldsymbol{r} \in \mathcal{Y}^c$, where each element corresponds to a label $y \in \mathcal{Y}$ and the order reflects their relative preference or relevance to the sample, and $r_j \neq r_k$ for all $j \neq k$.

2 Limitations of the Label Distribution

2.1 Lack of Fidelity to the Raw Data

Taking the conversion from LD to LL as an example, there are two common strategies:

- Top-k-Based:

$$\mathrm{bin}_1(\boldsymbol{d};\, k)_j \triangleq \left[\!\!\left[y_j \in \arg\operatorname*{top\text{-}k}_{y \in \mathcal{Y}}(\boldsymbol{d}) \right]\!\!\right], \tag{3}$$

- τ-Thresholding-Based:

$$\mathrm{bin}_2(\boldsymbol{d};\, \tau) \triangleq \mathrm{bin}_1(\boldsymbol{d};\, k^*), \tag{4}$$

where $k^* = \mathrm{card}(\min\{k \in [c] \mid \sum_{j \in [k]} \rho_j \geq \tau\})$,
$\boldsymbol{\rho}$ is the sorted LD in *descending order*.

2 Limitations of the Label Distribution

2.1 Lack of Fidelity to the Raw Data

Expected subset error rate between the logical labels derived from the raw data and those derived from LD:

Proposition 2.4. *Let $\mathfrak{L}^{(k)}$ denote the LL derived from the LD via the conversion rule in Equation (3); let $\mathfrak{L}^{*}$ denote the LL directly derived from the raw data random vector $\mathfrak{P}$ by thresholding at τ', i.e., $\mathfrak{L}_j = [\![\mathfrak{P}_j \geq \tau']\!]$ for all $j \in [c]$. The raw data are independently drawn from a uniform distribution over $[a, b]_{\mathbb{R}}$, and the LD is obtained by its projection. Then, we have the following expectation:*

$$\mathbb{E}[\texttt{S. Err.}(\mathfrak{L}^{*}, \mathfrak{L}^{(k)})] = 1 - \binom{c}{k}\left(\frac{b-\tau'}{b-a}\right)^{k}\left(\frac{\tau'-a}{b-a}\right)^{c-k} \geq 50\%. \tag{5}$$

The equality holds if and only if $c = 2$ and $k = 1$.

Proof sketch. $\mathfrak{L}^{*}$ arises as a Bernoulli vector determined by the uniform thresholding, while $\mathfrak{L}^{(k)}$ corresponds to selecting the k largest order statistics, which directly yield the stated formula. □

Corollary 2.5. *The following equation holds:*

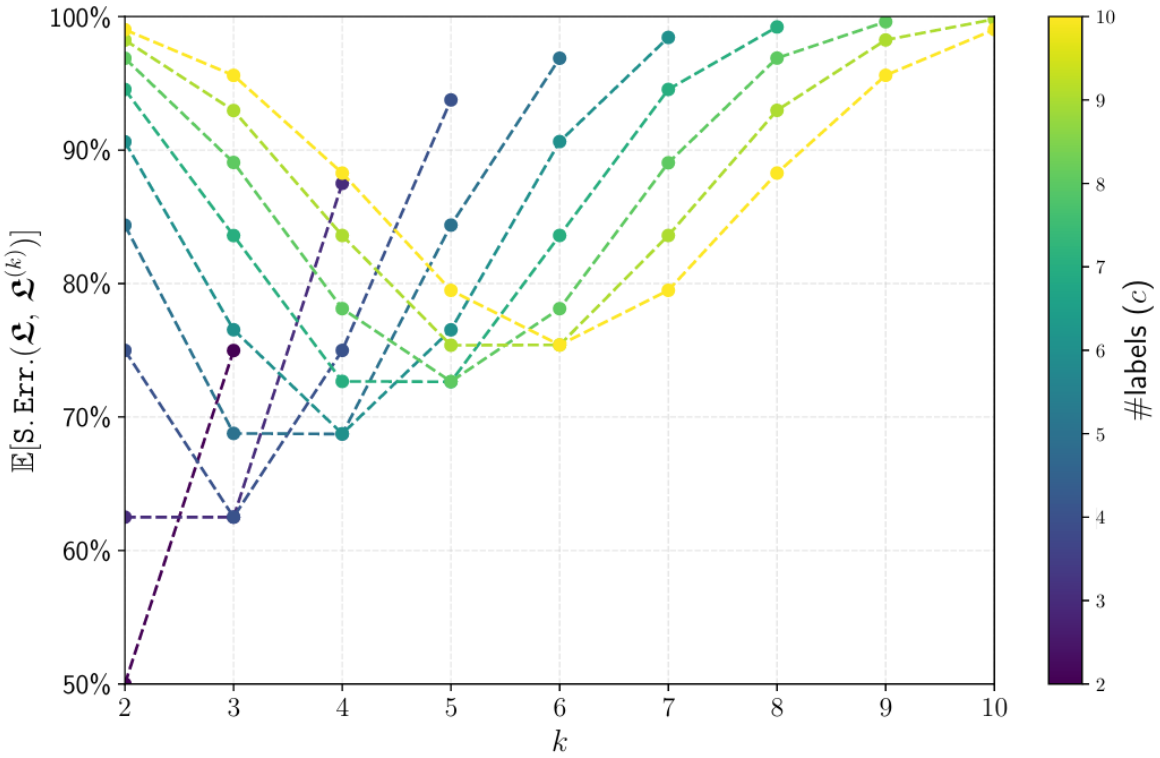
$$\arg\min_{k} \mathbb{E}[\texttt{S. Err.}(\mathfrak{L}^{*}, \mathfrak{L}^{(k)})] = \left\lfloor \frac{c}{2} \right\rfloor. \tag{6}$$

2 Limitations of the Label Distribution

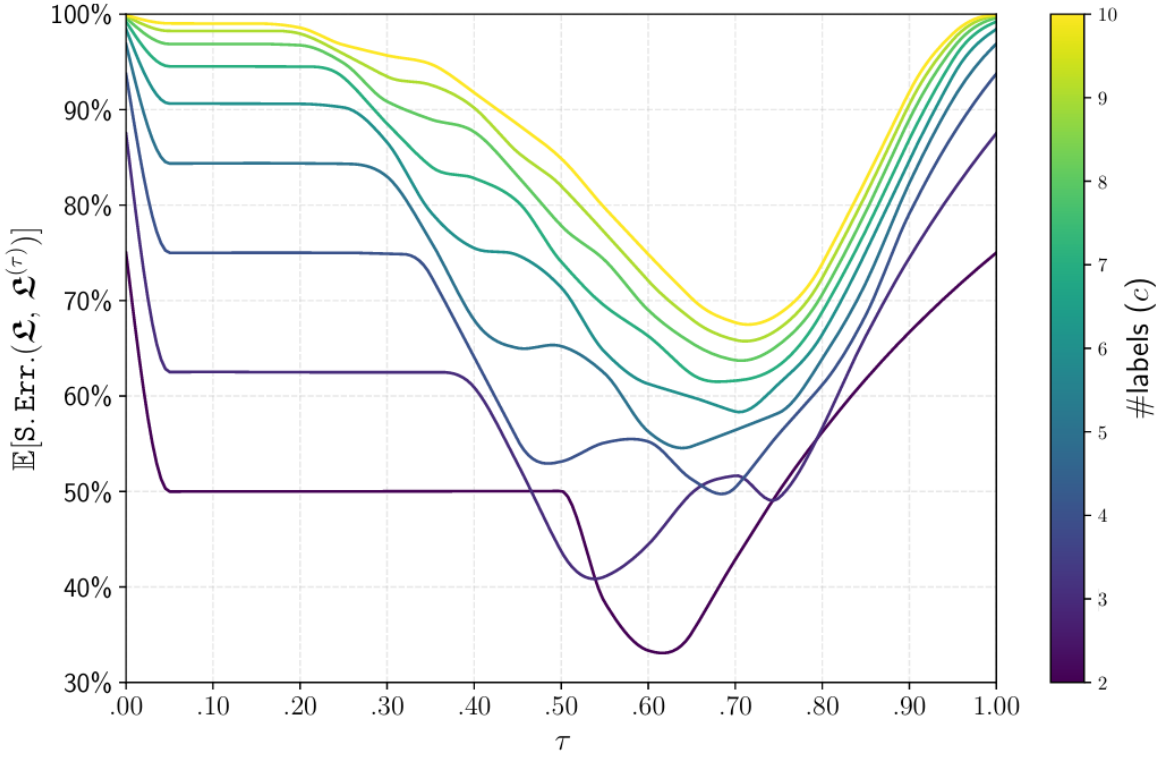
(Cont'd)

2.1 Lack of Fidelity to the Raw Data

The subset error rate w.r.t. the conversion parameters & the number of labels:



(a) $\mathbb{E}[\text{S. Err.}(\mathfrak{L}^*, \mathfrak{L}^{(k)})]$ w.r.t. k & c.



(b) $\mathbb{E}[\text{S. Err.}(\mathfrak{L}^*, \mathfrak{L}^{(\tau)})]$ w.r.t. τ & c.

2 Limitations of the Label Distribution

2.2 Disruption of Inter-Sample Order

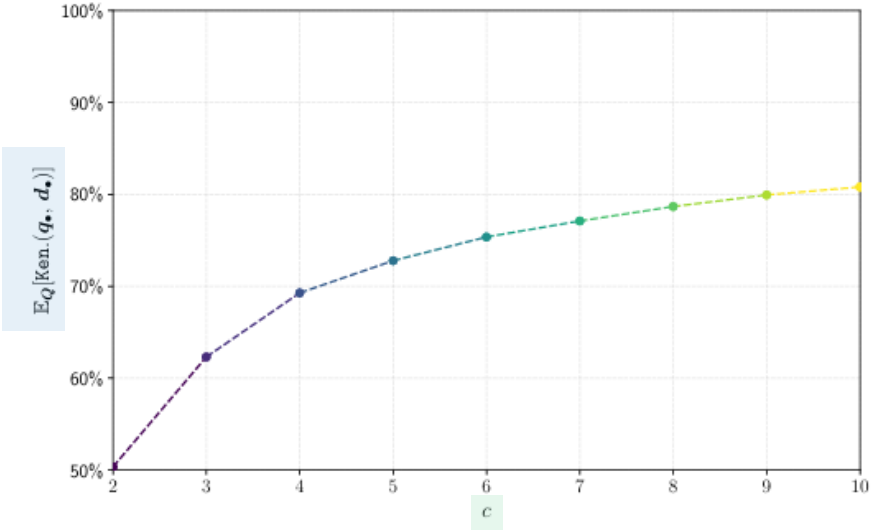
Kendall's Tau: $$\texttt{Ken.}(\boldsymbol{u}, \boldsymbol{v}) \triangleq \frac{2\sum_{i<j} \operatorname{sign}(u_i - u_j)\operatorname{sign}(v_i - v_j)}{n(n-1)}. \tag{7}$$

Expected inter-sample order consistency:

Proposition 2.7. *Assume that the raw data matrix $\boldsymbol{Q} \in [0, 1]_{\mathbb{R}}^{c\times n}$ has entries drawn independently from the uniform distribution. The corresponding LD matrix is obtained by $\boldsymbol{D} = \operatorname{proj}(\boldsymbol{Q})$. Treating each row as a sequence across samples, the inter-sample order consistency can be quantified by*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Q}}[\texttt{Ken.}(\boldsymbol{q}_\bullet, \boldsymbol{d}_\bullet)] &= \mathbb{E}_{\boldsymbol{Q}}[\operatorname{sign}((\boldsymbol{q}_\bullet - \boldsymbol{q}'_\bullet)^\top(\boldsymbol{q}_\bullet \oslash \boldsymbol{s} - \boldsymbol{q}'_\bullet \oslash \boldsymbol{s}))] \\ &= 2\mathbb{P}((\boldsymbol{q}_\bullet - \boldsymbol{q}'_\bullet)^\top(\boldsymbol{q}_\bullet \oslash \boldsymbol{s} - \boldsymbol{q}'_\bullet \oslash \boldsymbol{s}) > 0) - 1,\end{aligned} \tag{8}$$

where $\boldsymbol{q}_\bullet$ and $\boldsymbol{q}'_\bullet$ are two independent sequences drawn from the same distribution as the rows of $\boldsymbol{Q}$, and $s_i = \sum_{j\in[c]}(\boldsymbol{q}_i)_j$ is the LD normalization factor for each sample i.



Number of Labels

2 Limitations of the Label Distribution

2.3 Limited Practical Applicability

The practical limitations of LD mainly stem from two aspects:

- Inability to reconstruct the raw data;
- Difficulty in handling out-of-distribution (OOD) samples.

Remark 2.9. Under Assumption 2.1, the practical limitations of LD can be summarized as follows:

- Inability to reconstruct the raw data: LD can *not* recover the original information, such as absolute magnitude and label annotation (Xu et al., 2020). For example, different colors (e.g., Magenta $\boldsymbol{q} = \langle 255, 0, 255 \rangle^\top$ and Patriarch $\boldsymbol{q}' = \langle 128, 0, 128 \rangle^\top$ in the RGB color space) may share the same LD ($\boldsymbol{d} = \langle 0.5, 0, 0.5 \rangle^\top$) in an LD color space.
- Lack of mechanism to handle OOD data: LD assumes that every label is relevant to the sample to some extent, which may *not* hold in real-world scenarios (Wu et al., 2025). For example, in a emotion recognition task, an image (e.g., a neutral face) may not convey any of the predefined emotions (e.g., the primary emotions like happiness, sadness, anger, fear, surprise, and disgust), but LD would still assign non-zero description degrees to all emotions.

3 Generalized Label Distribution

To design a new representation for label ambiguity/polysemy,
it should simultaneously address the three aforementioned limitations.

Remark 3.1. Assume a GLD mapping $\eta(\cdot)$. To address the limitations inherent in traditional LD, a GLD should satisfy the following properties:

- Regarding Remark 2.6, GLD should allow conversion to other forms of label representations, e.g., LD, LL, TL, LR, and SL, without any information loss. For example, in the case of LL, there should exist a mapping function $\mathrm{GLD2LL}(\cdot)$ such that, for any LL $\boldsymbol{l}^*$ derived from the raw data $\boldsymbol{q}$, the equation $\mathtt{S.Err.}(\boldsymbol{l}^*, \mathrm{GLD2LL}(\eta(\boldsymbol{q}))) = 0$ holds.
- Regarding Remark 2.8, GLD should preserve the inter-sample order, i.e., for the j-th *row* of any raw data matrix $\boldsymbol{Q}$, the equation $\mathtt{Ken.}(\boldsymbol{q}_{\bullet j}, \eta(\boldsymbol{Q})_{\bullet j}) = 1$ holds.
- Regarding Remark 2.9, GLD should be capable of handling OOD data while providing a clear characterization of both positive and negative correlations; $\eta(\cdot)$ should be bijective to ensure that the raw data can be accurately reconstructed, i.e., $\boldsymbol{q} = \eta^{-1}(\eta(\boldsymbol{q}))$ for any raw data $\boldsymbol{q}$.

3 Generalized Label Distribution

(Cont'd)

Def.:

Definition 3.2 (Generalized label distribution)**.** Given the raw data $\boldsymbol{q} \in \prod_{j\in[c]}[a_j,\, b_j]_{\mathbb{R}}$, the GLD is defined as

$$[-1,\, 1]^c_{\mathbb{R}} \ni \boldsymbol{g} = \eta(\boldsymbol{q};\, \boldsymbol{a},\, \boldsymbol{b}) \triangleq 2(\boldsymbol{q} - \boldsymbol{a}) \oslash (\boldsymbol{b} - \boldsymbol{a}) - 1. \tag{9}$$

Conversely, the raw data can be recovered from the GLD by

$$\prod_{j\in[c]} [a_j,\, b_j]_{\mathbb{R}} \ni \boldsymbol{q} = \eta^{-1}(\boldsymbol{g};\, \boldsymbol{a},\, \boldsymbol{b}) \triangleq \frac{(\boldsymbol{g} + 1) \odot (\boldsymbol{b} - \boldsymbol{a})}{2} + \boldsymbol{a}. \tag{10}$$

Theoretical Evidence that GLD is More Expressive and More Faithful to the Raw Data:

Based on Definition 3.2, we have the following theorem regarding the expressiveness of GLD.

Theorem 3.3. *Let $\mathfrak{P} \sim \mathcal{Q}$ be the random vector of the raw data, $\mathfrak{D} = \mathrm{proj}(\mathfrak{P})$, and $\mathfrak{G} = \eta(\mathfrak{P})$. Under Assumption 2.1, we have*

$$I(\mathfrak{P};\, \mathfrak{D}) \leq I(\mathfrak{P};\, \mathfrak{G}), \tag{11}$$

where $I(\cdot;\, \cdot)$ is the mutual information. The equality holds if $\mathcal{Q} \subseteq \psi(\Delta^{c-1})$, where $\psi(\cdot)$ is an invertible transformation and $\mathcal{Q}$ is the distribution of the raw data.

Proof. See Appendix A.1 for details. □

The equality condition is highly restrictive.

In practical, the mapping $\psi(\Delta^{c-1})$ can be a scaled simplex $\kappa\Delta^{c-1}$ ($\kappa > 0$), a.k.a. the *Aitchison simplex* (Aitchison, 1994), i.e., the raw data are compositional data constrained to a fixed sum κ. It should be noted that this is a rather stringent condition, rarely satisfied in general scenarios. Theorem 3.3 indicates that GLD provides a more expressive and faithful representation compared to conventional LD, supporting tasks that rely on nuanced label correlations.

3 Generalized Label Distribution

(Cont'd)

Relationships and Transformations among Existing Label Ambiguity/Polysemy Representations and Learning Paradigms

From \ To	GLD	LD	TL	LL	LR	SL
GLD	Ours	Eq. (50)	Eq. (51)	Eq. (52)	Eq. (53)	Eq. (54)
LD		LDL	✗	✗	Eq. (55)	Eq. (56)
TL		Lu & Jia (NeurIPS 22)		✗	✗	✗
LL		LE		MLL	✗	✗
LR		Lu & Jia (NeurIPS 24)		Single-Positive Multi-Label Learning		Eq. (57)
SL						Classification

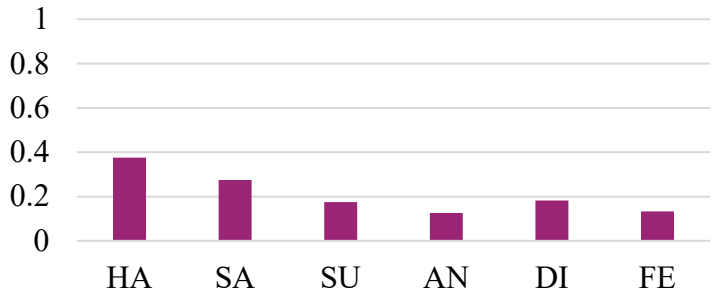
3 Generalized Label Distribution

(Cont'd)

Semantic Interpretation & Underlying Philosophy of GLD

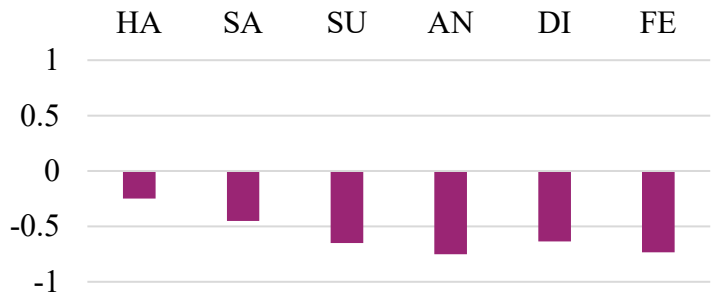
Underlying philosophy The essence of GLD lies in the concept of *net probability* from a statistical perspective. It reflects the intrinsic association between samples and labels, analogous to *odds*, but within a more interpretable range that can express both positive and negative correlations. Whereas previous work in LD focused on the description degrees, i.e., answering "*To what extent does a label describe a sample?*", GLD acts as a set of correlation coefficients, i.e., answering "*To what extent is a sample associated with a label?*". In this sense, we name the values in GLD as *relative degrees*.

Net Probability (analogous to odds, but with a more interpretable range)



Description Degree

"To what extent does a label describe a sample?"



Relative Degree

"To what extent is a sample associated with a label?"

4 Learning Generalized Label Distributions

Problem formulation Let $\mathcal{X} = \mathbb{R}^m$ denote the input space, and $\mathcal{G} \subseteq [-1, 1]^c_{\mathbb{R}}$ the output space. Given the training set $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i)\}_{i \in [n]}$, the goal is to find a mapping $f : \mathcal{X} \mapsto \mathcal{G}$, where $\boldsymbol{x}_i \in \mathcal{X}$ and $\boldsymbol{g}_i \in \mathcal{G}$ for all $i \in [n]$. The function f is not only required to accurately predict the GLD $\boldsymbol{g}$ for any unseen instance $\boldsymbol{x}$, but also to perform well on other paradigms, e.g., LDL, MLL, etc.

Solutions:

- **Strategy 1:** Problem Transformation;
- **Strategy 2:** Algorithm Adaptation;

Problem transformation The problem of learning GLDs can be reformulated as a series of constrained regression tasks, which bears similarity to classical methods in LDL and multi-target regression (Liu et al., 2009; Geng & Hou, 2015). Like them, we employ the support vector regression (SVR) algorithm as the underlying model. Note that we do *not* adopt the same problem transformation strategy used in (Geng, 2016), as it would compromise the intrinsic structure of the target data. To address the bounded nature of GLDs, we incorporate a forward-regressor-inverse transformation framework. Specifically, before training, the target values are first mapped to the real line via $\operatorname{arctanh}(\operatorname{clip}(\cdot;\, -1+\varepsilon,\, 1-\varepsilon))$, where $\operatorname{clip}(\cdot;\, a,\, b)$ restricts the input to the specified range $[a,\, b]_{\mathbb{R}}$, and ε is a small positive constant to avoid numerical issues. After training, the predictions are mapped back to the original space using $\tanh(\cdot)$. The resulting method, referred to as GLD-SVR, is straightforward and intuitive.

Algorithm adaptation We can also naturally adapt some existing algorithms to deal with GLDs, among which we consider the k-nearest neighbor algorithm as an example, since there are already some extended variants for LDL and MLL (Zhang & Zhou, 2007; Geng, 2016). For a given test sample $\boldsymbol{x}$, we first determine the set $\mathcal{N}(\boldsymbol{x})$ of its k nearest neighbors from the training set. The GLD prediction is then obtained by averaging the GLDs of these neighbors, i.e., $1/k \sum_{\boldsymbol{x}_i \in \mathcal{N}(\boldsymbol{x})} \boldsymbol{g}_i$. The resulting method, referred to as GLD-kNN, is also simple yet effective.

4 Learning Generalized Label Distributions

(Cont'd)

- **Strategy 3:** Specialized Algorithm.

Can be derived from the assumption of multivariate Gaussian noise:

$$-\log \prod_{i\in[n]} \varphi(\boldsymbol{x}_i, \boldsymbol{g}_i; \boldsymbol{\Sigma}) = \underbrace{\frac{nc}{2}\log(2\pi) + \frac{n}{2}\log\det(\boldsymbol{\Sigma})}_{\text{constant}} + \frac{1}{2}\sum_{i\in[n]} \underbrace{(\boldsymbol{g}_i - f(\boldsymbol{x}_i))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{g}_i - f(\boldsymbol{x}_i))}_{\text{squared Mahalanobis distance}}. \tag{12}$$

Theoretical Justification for using the Mahalanobis Distance as the Loss Function

$$\ell(\boldsymbol{W}; \{(\boldsymbol{x}_i, \boldsymbol{g}_i)\}_{i\in[n]}) = \frac{1}{n}\sum_{i\in[n]} (\boldsymbol{g}_i - \tanh(\boldsymbol{W}\boldsymbol{x}_i))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{g}_i - \tanh(\boldsymbol{W}\boldsymbol{x}_i)). \tag{13}$$

Optimization We can learn the optimal model parameters $\boldsymbol{W}^*$ by minimizing the empirical loss, i.e., $\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}}(\ell)$, which can be solved using gradient-based optimization methods.

Remark 3.5. ℓ is differentiable and its gradient w.r.t. $\boldsymbol{W}$ is given by

$$\frac{\partial \ell(\boldsymbol{W})}{\partial \boldsymbol{W}} = -\frac{2}{n}\sum_{i\in[n]} ((\boldsymbol{\Sigma}^{-1}(\boldsymbol{g}_i - \tanh(\boldsymbol{W}\boldsymbol{x}_i))) \odot \operatorname{sech}^2(\boldsymbol{W}\boldsymbol{x}_i))\boldsymbol{x}_i^\top, \tag{14}$$

derivation of which is provided in Appendix A.2, where $\operatorname{sech}(\cdot)$ is the hyperbolic secant function.

5 Theoretical Analysis

The variance of GLD-*k*NN is always no larger than that of LD-*k*NN, and is conditionally no larger than that of LL-*k*NN.

Theorem 3.6. *Let $\mathfrak{D}^*$ denote the underlying LD derived from the raw data random vector under Assumption 2.2; let $\mathfrak{D}'$ and $\mathfrak{D}$ be the predictions of GLD-kNN and LD-kNN, respectively. Suppose both $\mathfrak{D}$ and $\mathfrak{D}'$ are unbiased estimators of $\mathfrak{D}^*$. Then, the variance of $\mathfrak{D}'$ with respect to $\mathfrak{D}^*$ is no larger than that of $\mathfrak{D}$, i.e., $\mathbb{V}[\mathfrak{D}'_j - \mathfrak{D}^*_j] \leq \mathbb{V}[\mathfrak{D}_j - \mathfrak{D}^*_j]$ for all $j \in [c]$. The equality holds under the same condition as that required for the equality case in Theorem 3.3.*

Theorem 3.7. *Let $\mathfrak{L}^*$ denote the underlying LL derived from the raw data random vector $\mathfrak{P}$ by thresholding at $\boldsymbol{\tau}'$; let $\mathfrak{L}'$ and $\mathfrak{L}$ be the predictions of GLD-kNN and LL-kNN, respectively. Suppose both $\mathfrak{L}$ and $\mathfrak{L}'$ are unbiased estimators of $\mathfrak{L}^*$. Then, for the j-th label, $\mathbb{V}[\mathfrak{L}'_j - \mathfrak{L}^*_j] \leq \mathbb{V}[\mathfrak{L}_j - \mathfrak{L}^*_j]$ holds when $\sqrt{\mathbb{V}[\mathfrak{P}_j]} \leq 2|\tau'_j|\sqrt{\mathbb{P}(\mathfrak{P}_j \geq \tau'_j)(1 - \mathbb{P}(\mathfrak{P}_j \geq \tau'_j))}$ holds.*

GLD-BFGS enjoys a tight generali-zation bound:

Theorem 3.9. *Define a family of functions $\mathcal{H}$ and the j-th component of the output $\mathcal{H}_j = \{\boldsymbol{x} \mapsto \boldsymbol{w}_j \cdot \boldsymbol{x}\}$, where $\|\boldsymbol{w}_j\|_2 \leq \xi$ for all $j \in [c]$. Let $\mathcal{F}$ be the family of functions for methods corresponding to Section 3.2.2, for any $\delta > 0$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$, we have*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{M}^2(\boldsymbol{g}_{\boldsymbol{x}},\, f(\boldsymbol{x}))\right] \leq \frac{1}{n}\sum_{i\in[n]} \mathrm{M}^2(\boldsymbol{g}_i,\, f(\boldsymbol{x}_i)) + 12c\|\boldsymbol{\Sigma}^{-1}\|\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{8\xi c\sqrt{2c}\|\boldsymbol{\Sigma}^{-1}\|}{\sqrt{n}} \max_{i\in[n]} \|\boldsymbol{x}_i\|, \tag{16}$$

6 Experiments

Datasets We construct an artificial dataset `artf` similar to (Geng, 2016), with only a limited number of training samples to simulate a challenging learning scenario; details are given in Appendix C. The real-world datasets have been examined in previous research (Spyromitros-Xioufis et al., 2016; González et al., 2021), including: a facial emotion recognition dataset of Japanese female faces, denoted as `jaf`; a geochemical composition dataset collected from the topsoil of the Swiss Jura region, denoted as `jura`; a water quality dataset `wq`; two datasets for price prediction, `atp` and `scm`; and a dataset on heating/cooling loads of buildings (i.e., energy efficiency), denoted as `enb`.

Evaluation metrics We do *not* directly evaluate the GLD prediction task itself, as its practical application is unexplored. Instead, we assess performance through metrics in these following versatile tasks, where $\downarrow$ ($\uparrow$) indicates the smaller (larger) the better; see Appendix C for more details.

- OOD classification: For GLD, if all relative degree values are negative, the sample is regarded as an OOD instance. While LD cannot provide explicit OOD information, the situation where all labels are related to a sample is uncommon. Therefore, by leveraging the uniform distribution, we design a new metric `OOD Err.` $\downarrow$ to make a relatively fair comparison.
- Intra/inter-sample ranking: The Spearman rank correlation coefficient `Spear.` $\uparrow$ is commonly used for evaluating label ranking performance (Jia et al., 2023). In addition, we compute Kendall's tau for each *row* of the predicted/ground-truth matrix to assess the consistency of inter-sample ordering, and denote this metric as `Ken.'` $\uparrow$.
- LL prediction: We convert the model outputs into LLs and evaluate them using the Hamming distance `Ham.` $\downarrow$ and subset accuracy `S. Acc.` $\uparrow$ (Zhang & Zhou, 2013). For LD, the binarization algorithm bin_2 is adopted.
- LD prediction: We evaluate using the Clark distance `Clark` $\downarrow$ (Geng, 2016) and a "regularized" K-L divergence (KLD) denoted as μ_{KLD} $\uparrow$ (Li et al., 2025).

6 Experiments

(Cont'd)

Evaluation Metric for OOD Classification

$$\texttt{OOD Err.}(\boldsymbol{g}, \boldsymbol{v}) \triangleq \begin{cases} \begin{cases} [\![\forall_{j\in[c]}(v_j \le 0)]\!], & \text{if } \boldsymbol{v} \text{ is GLD}, \\ [\![\sum_{j\in[c]} v_j \log(cv_j) \le \delta_0/2]\!], & \text{if } \boldsymbol{v} \text{ is LD}, \end{cases} & \forall_{j\in[c]}(g_j \le 0), \\ [\![\arg\max_j(\boldsymbol{g}) = \arg\max_j(\boldsymbol{v})]\!], & \text{o/w}, \end{cases} \tag{69}$$

Three Additional Adapted Baseline Methods

Baselines GLD-SVR, GLD-kNN, and GLD-BFGS in Section 3.2 correspond to baselines proposed in (Geng, 2016), which we denote as LD-SVR, LD-kNN, and LD-BFGS for clarity. In addition, we design the following GLD methods based on the specialized algorithm strategy:

- GLD-DF: This method is based on an ensemble method, denoted as LD-DF, which assigns different data batches to different label pairs (González et al., 2021). In our adaptation, we replace its base estimators with GLD-BFGS models, and substitute the built-in LD-kNN with GLD-kNN.
- GLD-LRR: The reference method LD-LRR designs a loss function that incorporates label ranking relationships (Jia et al., 2023). We convert the predicted GLD into LD to make it compatible with this loss, and replace the original KLD with the squared Mahalanobis distance.
- GLD-Delta: The corresponding method, denoted as LD-Delta, optimizes a "regularized" KLD to ensure that the majority of samples are approximately predicted, thereby mitigating strong interference from outlier data (Li et al., 2025). To adapt it for GLD, both the loss and the worst-case expectation are computed using the squared Mahalanobis distance.

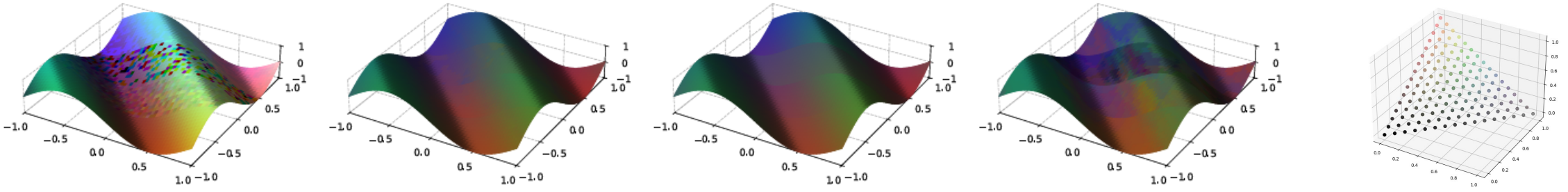
6 Experiments

(Cont'd)

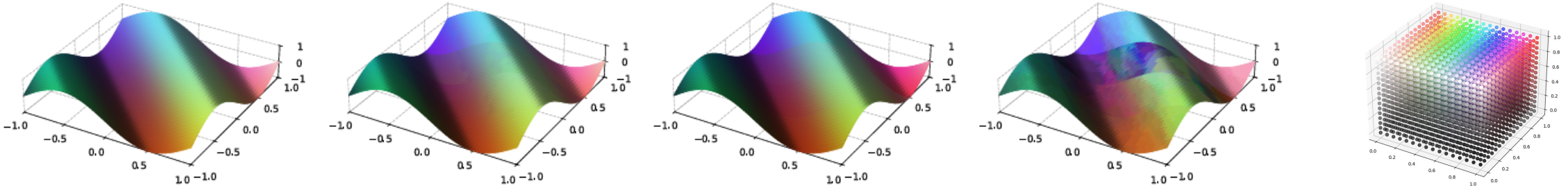
Algorithms	Clark ↓	μ_{KLD} ↑	Ham. ↓	S. Acc. ↑	Spear. ↑	Ken.′ ↑	OOD Err. ↓	Avg. Rank
				artf				
LD-SVR	.0307 ±.005	98.58% ±.005	● .2338 ±.046	● 40.85% ±.097	● .9640 ±.028	● .5083 ±.065	● 12.70% ±.076	6.14
GLD-SVR	.0307 ±.006	98.57% ±.005	.0143 ±.014	95.70% ±.041	**.9763** ±.022	.9502 ±.014	03.35% ±.041	2.86
LD-kNN	.0475 ±.008	96.40% ±.011	● .2340 ±.047	● 40.65% ±.102	.9333 ±.036	● .4942 ±.064	● 13.95% ±.071	9.43
GLD-kNN	.0476 ±.008	96.41% ±.011	.0547 ±.031	84.10% ±.088	.9328 ±.037	.9046 ±.024	08.25% ±.056	6.86
LD-BFGS	● .0783 ±.010	● 90.90% ±.022	● .2532 ±.048	● 36.85% ±.094	● .9470 ±.036	● .5063 ±.065	● 09.30% ±.063	10.3
GLD-BFGS	.0444 ±.009	96.55% ±.010	.0670 ±.030	80.40% ±.088	.9565 ±.030	.9589 ±.014	04.25% ±.042	5.29
LD-DF	● .0492 ±.007	● 96.38% ±.010	● .2497 ±.047	● 37.10% ±.096	● .9200 ±.046	● .5103 ±.066	● 16.35% ±.082	10.1
GLD-DF	**.0237** ±.004	**99.05%** ±.003	**.0118** ±.013	**96.45%** ±.038	.9758 ±.021	**.9648** ±.011	**02.45%** ±.031	**1.14**
LD-LRR	● .0783 ±.010	● 90.90% ±.022	● .2532 ±.048	● 36.85% ±.094	● .9470 ±.036	● .5062 ±.065	● 09.30% ±.063	10.4
GLD-LRR	.0444 ±.009	96.55% ±.010	.0670 ±.030	80.40% ±.088	.9565 ±.030	.9589 ±.014	04.25% ±.042	5.14
LD-Delta	.0285 ±.012	98.41% ±.014	● .2413 ±.048	● 38.95% ±.098	○ .9683 ±.027	● .5103 ±.066	● 10.75% ±.069	6.43
GLD-Delta	.0299 ±.006	98.56% ±.006	.0303 ±.020	90.90% ±.059	.9590 ±.027	.9559 ±.014	06.60% ±.058	3.86
				jura				
LD-SVR	.2709 ±.027	● 91.90% ±.018	● .4687 ±.033	● 02.23% ±.023	.9217 ±.030	● .3160 ±.053	● 61.78% ±.090	9.29
GLD-SVR	.2665 ±.025	92.55% ±.016	.1137 ±.027	68.27% ±.069	.9235 ±.030	.5419 ±.063	17.16% ±.059	5.71
LD-kNN	.2833 ±.027	○ 91.29% ±.019	● .4752 ±.036	● 03.26% ±.025	.9177 ±.031	● .2806 ±.050	● 61.44% ±.091	10.4
GLD-kNN	.2768 ±.029	90.39% ±.024	.1184 ±.029	68.04% ±.073	.9188 ±.031	.4981 ±.060	20.61% ±.059	8.00
LD-BFGS	● .3009 ±.031	92.60% ±.015	● .4803 ±.033	● 03.12% ±.026	.9348 ±.029	● .3120 ±.050	● 57.12% ±.097	8.71
GLD-BFGS	.2530 ±.024	92.86% ±.016	.1155 ±.027	68.38% ±.070	.9301 ±.030	.5772 ±.056	15.43% ±.052	4.29
LD-DF	● .2562 ±.025	93.56% ±.014	● .4795 ±.036	● 03.26% ±.029	.9308 ±.028	● .3149 ±.050	● 59.05% ±.095	7.43
GLD-DF	**.2394** ±.023	93.67% ±.014	.1151 ±.027	69.42% ±.068	.9291 ±.030	**.6028** ±.051	15.91% ±.052	**2.86**
LD-LRR	.2595 ±.025	○ 93.59% ±.014	● .4792 ±.034	● 02.93% ±.027	.9336 ±.031	● .3125 ±.050	● 59.21% ±.097	7.57
GLD-LRR	.2733 ±.099	90.21% ±.129	.1130 ±.026	68.91% ±.070	.9189 ±.070	.5553 ±.099	**15.38%** ±.051	5.86
LD-Delta	.2494 ±.024	○ **94.33%** ±.014	● .4720 ±.037	● 03.71% ±.031	○ **.9403** ±.030	● .3264 ±.052	● 58.96% ±.093	5.00
GLD-Delta	.2483 ±.022	93.42% ±.017	**.1102** ±.027	**70.55%** ±.068	.9319 ±.030	.5950 ±.054	18.08% ±.057	**2.86**

6 Experiments

(Cont'd)



(a) Noise data (b) LD-DF (c) LD-BFGS (d) LD-kNN LD



(e) Original data (f) GLD-DF (g) GLD-BFGS (h) GLD-kNN GLD

Figure 3: (Best viewed in color) Visualized results of the robustness testing.

7 Conclusion

The main contributions of this work are:

- We provide a thorough theoretical analysis of the inherent limitations of label distributions.
- We propose Generalized Label Distribution (GLD) as a unified representation of label ambiguity/polysemy that simultaneously addresses these limitations.
- We develop several algorithms for learning GLDs and extend existing LDL methods to be compatible with GLD learning.
- Both theoretical analyses and extensive experiments demonstrate the effectiveness of the proposed GLD learning framework.