Consistent Hierarchical Classification with A Generalized Metric

Yuzhou Cao[1] **Lei Feng**[1] **Bo An**[1,2]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Skywork AI, Singapore
(correspondence to lfengqaq@gmail.com)

Abstract

In multi-class hierarchical classification, a natural evaluation metric is the tree distance loss that takes the value of two labels' distance on the pre-defined tree hierarchy. This metric is motivated by that its Bayes optimal solution is the deepest label on the tree whose induced superclass (subtree rooted at it) includes the true label with probability at least $\frac{1}{2}$. However, it can hardly handle the risk sensitivity of different tasks since its accuracy requirement for induced superclasses is fixed at $\frac{1}{2}$. In this paper, we first introduce a new evaluation metric that generalizes the tree distance loss, whose solution's accuracy constraint $\frac{1+c}{2}$ can be controlled by a penalty value c tailored for different tasks: a higher c indicates the emphasis on prediction's accuracy and a lower one indicates that on specificity. Then, we propose a novel class of consistent surrogate losses based on an intuitive presentation of our generalized metric and its regret, which can be compatible with various binary losses. Finally, we theoretically derive the regret transfer bounds for our proposed surrogates and empirically validate their usefulness on benchmark datasets.

1 Introduction

Label hierarchies widely exist in the scenario of multiclass classification. For example, the labels of news documents obey the hierarchy determined by their topics (Lang, 1995); a natural hierarchy also exists in the task of species classification (Van Horn et al., 2018), where each image has the annotation of its species that has the hierarchy of biological taxonomy. To better utilize the hierarchy information, the task of hierarchical classification is studied and many efforts have been contributed to this area (Valmadre, 2022; Giunchiglia and Lukasiewicz, 2020; Dekel, 2009; Cesa-Bianchi et al., 2004; Babbar et al., 2013; Cesa-Bianchi et al., 2006; Wehrmann et al., 2018).

Given the label hierarchy, which is often a tree graph whose nodes are the class labels, it is able to design evaluation metrics other than the misclassification error that can better reflect the nature of hierarchical classification tasks. A natural evaluation is the tree distance between the prediction and the ground truth, which has been theoretically and empirically studied in previous works (Dekel et al., 2004; Sun and Lim, 2001; Ramaswamy et al., 2015; Bertinetto et al., 2020). A consistent surrogate loss was also proposed in Ramaswamy et al. (2015) to enable efficient optimization of this non-continuous metric.

Intuitively, the knowledge about label hierarchy provides us with an option of giving intermediate predictions (Bertinetto et al., 2020). i.e., predicting a non-leaf label on the tree that denotes the superclass consists of itself and all its descendant labels. Ramaswamy et al. (2015) proved that the tree distance loss fits well with this purpose by showing that its Bayes optimal solution is the class label of the highest level among those labels whose induced superclass includes the ground-truth label with a probability of at least 0.5.

While the Bayes optimal solution of the tree distance loss aims to achieve a balance between the predictions' semantic clarity and safeness, its safeness guarantee is not flexible enough to cope with different risk sensitivities of practical tasks: when a prediction with accuracy higher than 0.5 is needed, the tree distance loss fails to recognize a solution that meets this requirement. Such a requirement is natural since the accuracy guarantee of 0.5 is quite a weak one which only means the superclass is more likely to include the true label than not. Furthermore, the encouraged consistent surrogate in Ramaswamy et al. (2015) takes a hinge-like formulation

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

that is only verified on the linear model class, whose performance on popular deep models remains unclear.

In this paper, we tackle both problems of the design of a risk-sensitive evaluation metric and efficient surrogate losses. The main contributions of this paper can be summarized as follows:

- We propose a generalized version of the popular tree distance loss whose Bayes optimal solution has a flexible accuracy guarantee of $\frac{1+c}{2}$, where $c > 0$ can be set to proper values that conform to the need of practical tasks. An illustration can be found in Figure 1.
- We derive an intuitive representation of the proposed loss and use it to rewrite its risk formulation into the sum of binary classification risks, which enlightens a consistent problem reduction from hierarchical classification to ordinary classification.
- We further delve into the regret formulation of our proposed generalized tree distance loss and provide a condition on the coherency of the model we use, which can capture the essence of hierarchical classification and further reduce the gap between hierarchical classification and binary classification.
- Based on these findings, we propose a loss formulation that can integrate various losses for binary classification to construct consistent surrogates for our generalized tree distance loss and further induce regret transfer bounds for them to better characterize their behavior.

Experimental results on benchmark datasets of hierarchical classification clearly demonstrate the effectiveness of our proposed method.

2 Background: Tree Distance Loss, Consistency, and Surrogates

In this section, we review the problem setting of hierarchical classification *w.r.t.* the evaluation metric of the tree distance loss, and its existing surrogate losses. Before reviewing the problem formulation, we first introduce some notations about the structure of the tree as shown in Table 1, which is necessary for the definition of hierarchical classification.

Notations: Given a tree $H = ([K], E)$ with node set $[K]$, edge set E, and root node $r = 1$, we list the notations for each $y \in [K]$ in Table 1. We further define $U_{\boldsymbol{u}}(\epsilon) = \{y | \max_{y' \in C_y} W_y(\boldsymbol{u}) = \epsilon\}$, which means the collection of y whose children's subtrees' maximum weight is equal to ϵ. Let us denote by $T_y = T_y^1$ and its complement set $\overline{T}_y = T_y^0$, and we will use both notations according to the context. We further define the function $s_{y_1}(y_2) : [K] \to \{0, 1\}$:

$$s_{y_1}(y_2) = \begin{cases} 1, & y_2 \in T_{y_1}, \\ 0, & \text{else.} \end{cases}$$

Table 1: Notations used in tree structure.

$\mathrm{Lev}(y)$	Level of y
P_y	Parent of y
C_y	Children of y
D_y	Descendants of y
T_y	y-induced superclass/subtree
$W_y(\boldsymbol{u})$	Weight of subtree with root y, $\sum_{y' \in T_y} u_{y'}$

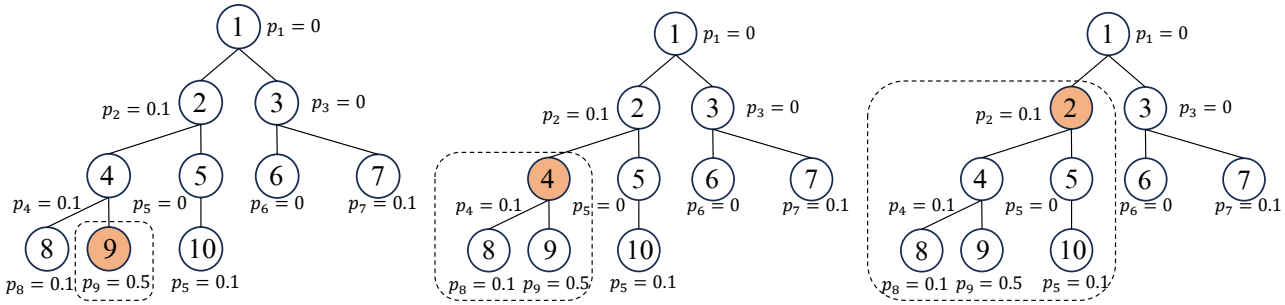
Tree-Distance Loss and Bayes Optimality: Denote by $\mathcal{X}$ and $\mathcal{Y} = [K]$ the input and label spaces, $X \times Y$ is the input-label random variable tuple and $\boldsymbol{x} \times y$ is their realization. In the setting of hierarchical classification, we have access to *i.i.d.* data pairs $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ drawn from the distribution $\mathcal{D}$ with density $p(\boldsymbol{x}, y)$. The goal is to obtain a classifier $f \in \mathcal{X} \to \mathcal{Y}$, which is often required to minimize the risk $R_{\mathcal{D}}^{\ell_H}(f)$, i.e., the expectation of the **tree distance loss** $\ell_H : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$:

$$R_{\mathcal{D}}^{\ell_H}(f) = \mathbb{E}_{p(\boldsymbol{x},y)} \left[\ell_H(f(\boldsymbol{x}), y) \right], \tag{1}$$

where the tree distance loss $\ell_H(y, y')$ is defined as the length of the path between y and y'. The tree distance loss can measure the discrepancy between the predicted label and the ground-truth label according to the tree H. Furthermore, the essence of risk minimization with tree distance loss can be better demonstrated with the following characterization of its **Bayes optimal solution** $f^* = \operatorname{argmin}_f R_{\mathcal{D}}^{\ell_H}(f)$ as stated in Theorem 1 of Ramaswamy et al. (2015). Suppose $\boldsymbol{\eta}(\boldsymbol{x}) = \{p(y|\boldsymbol{x})\}_{y=1}^K$ is the class posterior probability, then we have

$$f^*(\boldsymbol{x}) \in \left(\operatorname*{argmax}_{W_y(\boldsymbol{\eta}(\boldsymbol{x})) \geq 0.5} \mathrm{Lev}(y) \right) \cup U_{\boldsymbol{\eta}(\boldsymbol{x})}(0.5). \tag{2}$$

Notice that the weight term $W_y(\boldsymbol{\eta}(\boldsymbol{x}))$ is exactly the likelihood that the ground-truth label is included in the induced superclass T_y, i.e., the **accuracy** of T_y, while a prediction y with higher level can induce a more compact and exact T_y than its ancestors. In a nutshell, the Bayes optimal prediction $f^*(\boldsymbol{x})$ is the label of the highest level among those labels whose induced superclasses include the ground-truth label with probabilities greater than 0.5. The set $U_{\boldsymbol{\eta}(\boldsymbol{x})}(0.5)$ can be seen as a further trade-off: since it is hard to judge if a superclass with accuracy equals 0.5 is a good solution, we can take into consideration its parent class to diminish the ambiguity. Nevertheless, the constraint for the accuracy of superclasses is fixed



(a) $c = 0$ (b) $c = 0.4$ (c) $c = 0.8$

Figure 1: Illustration of the Bayes optimal solution for ℓ_H^c under different c. The highlighted label is the optimal prediction and the labels enclosed by the dashed line make up the induced superclass.

at 0.5, which can be too loose and restrictive due to the risk-sensitivity nature of different tasks.

Consistency and Surrogates: It is worth noting that the tree distance loss ℓ_H is discontinuous in general, which makes its minimization problem NP-hard (Feldman et al., 2012). To make the optimization problem tractable, an effective method is to substitute the discontinuous target losses with continuous surrogates. This method has been systematically studied in various fields of statistical machine learning, including but not limited to multiclass classification (Bartlett et al., 2006; Zhang, 2004; Tewari and Bartlett, 2007; Ramaswamy and Agarwal, 2016; Pires and Szepesvári, 2016; Finocchiaro et al., 2019; Mao et al., 2023a; Bao, 2023), multilabel classification (Gao and Zhou, 2013; Zhang et al., 2020; Koyejo et al., 2015; Wu et al., 2021), linear-fractional utility and AUC maximization (Gao and Zhou, 2015; Menon and Williamson, 2014; Bao and Sugiyama, 2020; Mao et al., 2023b),top-K classification (Lapin et al., 2018; Yang and Koyejo, 2020), adversarially robust classification Bao et al. (2020); Awasthi et al. (2021b,a, 2023), and classification with rejection (Cortes et al., 2016a,b; Ni et al., 2019; Charoenphakdee et al., 2021; Cao et al., 2022).

To have a guaranteed performance both theoretically and practically, the **consistency** of surrogate losses is often required. In the field of hierarchical classification, we are also interested in the design of consistent surrogates for the tree distance loss. Let $\mathcal{C} \subset \mathbb{R}^d$ and $\Phi : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}_+$ is a surrogate loss. We aim to learn a model $\boldsymbol{g}^* : \mathcal{X} \to \mathcal{C}$ via the surrogate risk minimization described as follows:

$$\boldsymbol{g}^* \in \operatorname{argmin}_{\boldsymbol{g}} R_{\mathcal{D}}^{\Phi}(\boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x},y)}[\Phi(\boldsymbol{g}(\boldsymbol{x}), y)], \quad (3)$$

and then design a link function $\varphi : \mathcal{C} \to \mathcal{Y}$ so that $\varphi \circ \boldsymbol{g}^*$ can be used for prediction. A consistent surrogate should fulfill the following conditions for any $\mathcal{D}$:

$$\varphi \circ \boldsymbol{g}^* \in \operatorname{argmin}_f R_{\mathcal{D}}^{\ell_H}(f), \ \forall \boldsymbol{g}^* \in \operatorname{argmin}_{\boldsymbol{g}} R_{\mathcal{D}}^{\Phi}(\boldsymbol{g}),$$

which immediately indicates that the minimization of the surrogate risk $R_{\mathcal{D}}^{\Phi}(\boldsymbol{g})$ can lead to that of $R_{\mathcal{D}}^{\ell_H}(f)$.

Inspired by the characterization of the Bayes optimal solution (2), a natural surrogate was proposed in Ramaswamy et al. (2015), which directly estimates the class posterior probability and then traverses the tree to find the optimal label according to the estimated probability. A problem reduction based on the hinge-like loss functions (Ramaswamy et al., 2018) for classification with rejection was further proposed (Ramaswamy et al., 2015), which can better utilize the label hierarchy and achieve a tight regret bound. In Sections 4-6, we will give a novel consistent surrogate formulation that can be constructed using more kinds of loss functions while allowing simple regret analyses based on fruitful results from the field of binary classification.

3 Risk Sensitive Metric: Generalized Tree Distance Loss

It is noticeable that the original tree distance loss is unable to handle the increased risk sensitivity in the sense that the accuracy constraint of its optimal prediction's induced superclass is fixed at 0.5. In this section, we propose a generalized tree distance loss that allows us to manually set the accuracy constraint according to practical requirements.

By inspecting the Bayes optimal solution (2) of the tree distance loss, we can learn that the level of a prediction y' plays a crucial role in evaluating its exactness. Meanwhile, the level of y' is also closely related to the accuracy of its induced superclass $T_{y'}$. To achieve a trade-off between the exactness and accuracy of the prediction according to the risk sensitivity, it is a natural idea to pay more attention to the level term. To this end, we propose a straightforward implementation of this trade-off (i.e., a generalized tree distance loss), which is formulated as follows:

Definition 1. (Generalized tree distance loss) Our generalized tree distance loss is defined as follows:

$$\ell_H^c(y', y) = \ell_H(y', y) + c * \mathrm{Lev}(y'), \tag{4}$$

where $c \in [0, 1]$ is the trade-off parameter.

The generalized tree distance loss ℓ_H^c only differs from the original one ℓ_H on the extra term $c * \mathrm{Lev}(y')$, which can be simply obtained. The extra term $\mathrm{Lev}(y')$ serves as a penalty on the prediction's level, while c can control the degree of penalty. As a result, predictions of lower levels, i.e., those labels whose induced superclass are of higher accuracy, can be preferable under this new evaluation metric.

Though it seems heuristic at first glance due to its simplicity, the generalized tree distance loss has a strong theoretical guarantee, which can be characterized by its Bayes optimality shown below:

Theorem 1. (Bayes optimality of ℓ_H^c) Denote by $R_{\mathcal{D}}^{\ell_H^c}(f) = \mathbb{E}_{p(\boldsymbol{x},y)}[\ell_H^c(f(\boldsymbol{x}), y)]$. $f_c^* \in \mathrm{argmin}_f R_{\mathcal{D}}^{\ell_H^c}(f)$ *i.f.f* it meets the following condition almost surely:

$$f_c^*(\boldsymbol{x}) \in \left(\operatorname*{argmax}_{W_y(\boldsymbol{\eta}(\boldsymbol{x})) \geq \frac{1+c}{2}} \mathrm{Lev}(y) \right) \cup U_{\boldsymbol{\eta}(\boldsymbol{x})}\left(\frac{1+c}{2}\right). \tag{5}$$

The proof of Theorem 1 is provided in Appendix A.

This Bayes optimal solution of ℓ_H^c theoretically validates the rationality of our generalized metric: by adding a level-related term $c * \mathrm{Lev}(y')$, the accuracy constraint for the desired prediction can be raised by $\frac{c}{2}$ compared with the original one and thus generate a solution with higher accuracy, which shows that our generalized tree distance loss can conform to the task's risk sensitivity by switching to a proper value of c. Aside from the risk $R^{\ell_H^c}(f)$, we are also interested in the **regret**, which is also the key object of our study:

$$\mathtt{Regret}_{\mathcal{D}}^{\ell_H^c}(f) = R_{\mathcal{D}}^{\ell_H^c}(f) - R_{\mathcal{D}}^{\ell_H^c}(f_c^*),$$

which is equal to 0 *i.f.f.* $f = f_c^*$ almost surely.

After validating the effectiveness of our proposed generalized metric, we should notice that its optimization faces the same difficulties as the original one due to its discontinuity. Furthermore, the complexity caused by its hierarchy also makes the design of surrogates and regret analysis a challenge. In the following sections, we will first propose a node-wise risk representation to demystify the problem structure of risk minimization with ℓ_H^c (and ℓ_H), and then show that it can strongly inspire the design of consistent surrogates for ℓ_H^c and, furthermore, enable the regret analysis.

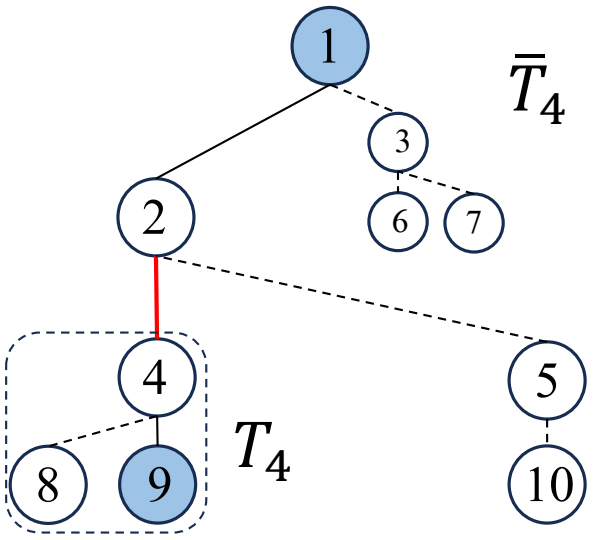


Figure 2: Since labels 1 and 9 are not included in T_4 at the same time, we can learn that the edge between 4 and its parent must be a part of the route from 1 to 9.

4 Node-wise Representation and Problem Reduction

According to the previous sections, our task is to obtain the model that can minimize risk $R^{\ell_H^c}(f)$, i.e., achieve zero regret $\mathtt{Regret}_{\mathcal{D}}^{\ell_H^c}(f) = 0$. An indispensable step toward solving the problem is the design of surrogate loss, which is not as clear as in the ordinary classification scenario due to the complexity of our target loss. In this case, a promising key is to find a proper **representation** of our target loss, which has been shown to be helpful for the design of losses and regret analysis (Reid and Williamson, 2009; Yoshida et al., 2021). In this section, we give a node-wise representation of our target loss ℓ_H^c and then generalize it to its risk, which finally leads to an intuitive problem reduction to binary classification. We further conduct regret analysis and show that this problem reduction also possesses an intuitive regret representation if the modeling method captures the intrinsic structure of hierarchical classification.

4.1 Loss Representation and Binary Classification Reduction

Recalling the definition of ℓ_H^c, we can find that it consists of two components, including the path length between y' and y and the level of y', which can be seen as the path length between y' and root node 1. To calculate the length of the two paths in the tree H, a viable method is to enumerate each non-root node of the tree, i.e., $y \neq 1$, and count the number of nodes whose edge to its parent is a part of the path, which is exactly the path length. A direct realization of this

idea is implemented below:

Theorem 2. (Node-wise Representation of ℓ_H^c)

$$\ell_H^c(y', y) = \sum_{i=2}^{K} \left[\mathbb{I}(y' \notin T_i^{s_i(y)}) + c*\mathbb{I}(y' \notin T_i^0)\right]. \quad (6)$$

The proof of Theorem 2 is provided in Appendix B.

A straightforward explanation for this representation is that if y and y' are not included in the same subset induced by the node i, the path between them must overlap with the edge from i to its parent. An illustration of this explanation is shown in Figure 2. The second term uses T_i^0 directly since the root node 1 is always in $\overline{T}_i$. Given this representation, we are able to obtain the following intuitive representation of point-wise risk by taking the expectation over $p(\boldsymbol{x}, y)$:

Lemma 1. (Node-wise Representation of $R_{\mathcal{D}}^{\ell_H^c}$) Denote by $\mathcal{D}_i$ a dummy distribution on $\mathcal{X} \times \{0, 1\}$ with density $p_i(\boldsymbol{x}, \gamma)$, we can obtain the following representation:

$$R_{\mathcal{D}}^{\ell_H^c}(f) = (1 + c) \sum_{i=2}^{K} \mathbb{E}_{p_i(\boldsymbol{x},\gamma)} \left[\mathbb{I}(f(\boldsymbol{x}) \notin T_i^{\gamma})\right], \quad (7)$$

where $p_i(\gamma = 1|\boldsymbol{x}) = \eta^i(\boldsymbol{x}) = \frac{\Pr(Y \in T_i^1|\boldsymbol{x})}{1+c}$ and the marginal density $p_i(\boldsymbol{x}) = p(\boldsymbol{x})$.

The proof of Lemma 1 is provided in Appendix C.

Considering the expectation terms in (7) as $(K - 1)$ sub-problems, it can be seen that each of them is similar to a binary classification problem that aims to classify samples into one of its two induced subsets. To be detailed, for each non-root node i, we can construct a binary classifier $f^i : \mathcal{X} \to \{0, 1\}$ to predict the set $S_i^{f^i(\boldsymbol{x})}$ that the ground-truth label of $\boldsymbol{x}$ belongs to. This problem reduction can be formulated as the risk minimization problem below for $i \in [2, K]$:

$$\min_{f^i} R_{\mathcal{D}_i}^{\ell_{01}}(f^i) = \mathbb{E}_{p_i(\boldsymbol{x},\gamma)}[\ell_{01}(f^i(\boldsymbol{x}), \gamma)], \quad (8)$$

where $\ell_{01}(f^i(\boldsymbol{x}), \gamma) = \mathbb{I}(f^i(\boldsymbol{x}) \neq \gamma)$ is the celebrated zero-one loss in binary classification. The following theorem further reveals the strong connection between the binary sub-problems and our generalized hierarchical classification problem:

Theorem 3. Denote by $\mathcal{F} = \{f^i\}_{i=2}^K$ the sequence of binary classifiers and $\mathcal{F}^* = \{f^{i*}\}_{i=2}^K$ the solutions of (8)[1]. Then there exists a binary classifier f_c^* such that the following equation holds almost surely:

$$s_i(f_c^*(\boldsymbol{x})) = f^{i*}(\boldsymbol{x}). \quad (9)$$

The proof of Theorem 3 is provided in Appendix D.

[1]Assume $f^{i*}(\boldsymbol{x}) = 1$ when $\eta^i(\boldsymbol{x}) = 0.5$.

As a result, we can get f_c^* with the following operation:

$$f_c^*(\boldsymbol{x}) = \varphi \circ \mathcal{F}^*(\boldsymbol{x}), \quad (10)$$

where $\varphi \circ \mathcal{F}(\boldsymbol{x}) = \cap_{i=2}^K T_i^{f^i(\boldsymbol{x})}$.

According to (9), the solutions of binary classification sub-problems are closely connected to f_c^*: $f^{i*}(\boldsymbol{x})$ can accurately reflect if the Bayes optimal prediction $f_c^*(\boldsymbol{x})$ is in the subtree T_i or not. Since each $f^{i*}(\boldsymbol{x})$ provides a set that includes $f_c^*(\boldsymbol{x})$, we can finally get $f_c^*(\boldsymbol{x})$ by taking the intersection of all these sets as in (10).

Given Theorem 3, it seems that the rationality of problem reduction (8) is justified since we can finally get the Bayes optimal solution of $R_{\mathcal{D}}^{\ell_H^c}(f)$ via the solutions of (8). However, it is noticeable that we have no access to the exact value of $p(\boldsymbol{x}, y)$ in practical scenarios, and we may only obtain an approximation of optimal solutions $\tilde{\mathcal{F}} \approx \mathcal{F}^*$. Due to the deviation of $\tilde{\mathcal{F}}$, we can be easily confused by **contradictions** when trying to assemble the prediction of $\tilde{f}^i$. Let us consider the tree structure in Figure 1 with $\tilde{f}^3(\boldsymbol{x}) = \tilde{f}^4(\boldsymbol{x}) = 1$. According to these results, the ground-truth label is in both $\{3, 6, 7\}$ and $\{4, 8, 9\}$, which is contradictory since $\{3, 6, 7\} \cap \{4, 8, 9\} = \emptyset$, and thus it is hard to obtain a prediction in $[K]$, which makes it hard to be put into final deployment. In the next subsection, we give a crucial property to characterize a family of models that can avoid such a contradiction, and further show its helpfulness with regret analysis.

4.2 Finer Modeling and Regret Analysis

In the last part of the previous section, we have shown that separately constructing binary classifiers for sub-problems (8) can cause serious contradictions that confuse the final prediction. To achieve contradiction-free prediction, we introduce a precise property that characterizes the coherency of $\mathcal{F}$ *w.r.t.* H:

Definition 2. (H-Coherency) A sequence of classifiers $\mathcal{F} = [f^i]_{i=2}^K$ is H-coherent if $|\varphi \circ \mathcal{F}(\boldsymbol{x})| = 1$, $\forall \boldsymbol{x} \in \mathcal{X}$.

This property requires that the binary classifiers finally reach a consensus without a veto to any of them, which means there is no contradiction as described before.

Given this property, we can finally bind the risk of each binary sub-problem together: inversely, we can learn that $\mathbb{I}(\varphi \circ \mathcal{F}(\boldsymbol{x}) \notin T_i^y) = \mathbb{I}(f^i(\boldsymbol{x}) \neq y)$ according to $\mathcal{F}$'s coherency, and thus we can substitute the expectation term in (7) with the binary problems' risks (8). Such a connection makes it possible to finally decompose the regret of our hierarchical classification problem into the sum of binary problems' regrets:

Theorem 4. (Node-wise Representation of Regret) Denote by $\texttt{Regret}_{\mathcal{D}_i}^{\ell_{01}}(f^i) = R_{\mathcal{D}_i}^{\ell_{01}}(f^i) - R_{\mathcal{D}_i}^{\ell_{01}}(f^{i*})$. If

$\mathcal{F} = \{f^i\}_{i=2}^K$ is H-coherent, then:

$$\texttt{Regret}_{\mathcal{D}}^{\ell_H^c}(\varphi \circ \mathcal{F}) = (1+c)\sum_{i=2}^{K} \texttt{Regret}_{\mathcal{D}_i}^{\ell_{01}}(f^i). \quad (11)$$

The proof of Theorem 4 is provided in Appendix E.

The proof of this regret representation is straightforward since Theorem 3 indicates that $\mathcal{F}^*$ is also H-coherent. Given this regret representation, we can learn that a group of binary classifiers $\mathcal{F}$ that is close to the optimal ones $\mathcal{F}^*$ can also induce a hierarchical classifier $\varphi \circ \mathcal{F}$ that is close to f_c^*. In Section 6, we will further show that it can help the derivation of the regret transfer bounds for various surrogates.

5 Coherent and Consistent Surrogates

In the previous sections, we have shown that the minimization of target risk $R_{\mathcal{D}}^{\ell_H^c}(f)$ is equivalent to solving $(K-1)$ binary classification problems if we choose the model appropriately. Then there exists a problem before we can get the consistent surrogates for ℓ_H^c: how can we design a modeling method that meets the condition of H-coherency? In this section, we first solve this issue with the following conclusion:

Lemma 2. Given $\hat{\boldsymbol{\eta}}(\boldsymbol{x}) \in \Delta^K$, we denote by $\bar{\eta}_i(\boldsymbol{x}) = \sum_{j \in T_i} \hat{\eta}_j(\boldsymbol{x})$. When $f_{\hat{\boldsymbol{\eta}}}^i(\boldsymbol{x}) = \mathbb{I}(\bar{\eta}_i(\boldsymbol{x}) > 0.5)$, $\mathcal{F}_{\hat{\boldsymbol{\eta}}} = \{f_{\hat{\boldsymbol{\eta}}}^i\}_{i=1}^K$ is H-coherent.

The proof of Lemma 2 is provided in Appendix F.

The proof can be completed by checking that $\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x})$ is neither empty nor having multiple elements. Compared with directly modeling the dummy distribution $\eta^i(\boldsymbol{x})$ with a function in $\mathcal{X} \to [0,1]$, which is not H-coherent in general[2], our method instead models $\boldsymbol{\eta}(\boldsymbol{x})$ first and then uses it to construct $\eta^i(\boldsymbol{x})$, which captures the intrinsic structure of the setting of this problem. In practical applications, $\hat{\boldsymbol{\eta}}$ can be the composite of the softmax function a K-dimensional scoring function, which can be easily implemented.

Based on the proposed coherent modeling, we can proceed to the design of surrogate losses. Inspired by our discussions on the binary classification reduction, we can substitute the indicator functions in (6) with binary classification losses to construct the surrogates for ℓ_H^c. Meanwhile, our coherent modeling use $\overline{\boldsymbol{\eta}} \in [0,1]$ to model $\eta^i(\boldsymbol{x})$, it is promising that we can find a suitable binary loss function as the component of our surrogate from the family of losses that focus on inputs in the range of $[0,1]$, e.g., proper losses (Reid and Williamson, 2010; Williamson et al., 2016). This idea is supported by the following theorem:

[2]We defer the proof of this claim to Appendix G.

Theorem 5. (Consistency Result) Let us denote by $R_{\mathcal{D}}^{\Phi_{\phi^c}}(\hat{\boldsymbol{\eta}}) = \mathbb{E}_{p(\boldsymbol{x},y)}[\Phi_\phi^c(\hat{\boldsymbol{\eta}}(\boldsymbol{x}), y)]$. When ϕ is a binary strictly proper loss or mean absolute loss, the following loss formulation is a consistent surrogate *w.r.t.* ℓ_H^c:

$$\Phi_\phi^c(\hat{\boldsymbol{\eta}}(\boldsymbol{x}), y) = \sum_{i=2}^{K} [\phi(\bar{\eta}_i(\boldsymbol{x}), s_i(y)) + c * \phi(\bar{\eta}_i(\boldsymbol{x}), 0)] \quad (12)$$

i.e., for any $\hat{\boldsymbol{\eta}}^* \in \operatorname{argmin}_{\mathcal{X} \to \Delta^K} R_{\mathcal{D}}^{\Phi_{\phi^c}}(\hat{\boldsymbol{\eta}})$, we have $f_{\hat{\boldsymbol{\eta}}}^i = f^{i*}$, and thus $\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}^*} \in \operatorname{argmin}_f R_{\mathcal{D}}^{\ell_H^c}(f)$.

The proof of Theorem 5 is provided in Appendix H.

According to the theorem above, by plugging the Log loss and the Mean Absolute Error (MAE) into our formulation, we can get the following realizations of consistent surrogates:

Example 1. (The realization by the Log loss.)

$$\Phi_{\text{Log}}^c(\hat{\boldsymbol{\eta}}(\boldsymbol{x}), y) = \sum_{i=2}^{K} [-c \log(1 - \bar{\eta}_i(\boldsymbol{x})) - (1 - s_i(y)) \log(1 - \bar{\eta}_i(\boldsymbol{x})) - s_i(y) \log(\bar{\eta}_i(\boldsymbol{x}))]. \quad (13)$$

Example 2. (The realization by MAE.)

$$\Phi_{\text{MAE}}^c(\hat{\boldsymbol{\eta}}(\boldsymbol{x}), y) = \sum_{i=2}^{K} [-c\bar{\eta}_i(\boldsymbol{x}) - (1 - s_i(y))\bar{\eta}_i(\boldsymbol{x}) - s_i(y)(1 - \bar{\eta}_i(\boldsymbol{x}))]. \quad (14)$$

With these instantiations, we finally turn our findings into consistent surrogate formulations. We will empirically validate these surrogates in Section 7.

6 Regret Transfer Bound

While the previous section provides infinite-sample consistency for our loss formulation, we are also interested in the performance guarantee of solutions that are close to the optimal ones, since we often approximate the data distribution with the sample mean and the empirically optimal solutions usually differ from the Bayes optimal ones. The following regret transfer bound provides such guarantee for the consistent surrogates considered in this paper:

Theorem 6. (A data-dependent bound) Denote by $\texttt{Regret}_{\mathcal{D}}^{\Phi_\phi^c}(\hat{\boldsymbol{\eta}}) = R_{\mathcal{D}}^{\Phi_{\phi^c}}(\hat{\boldsymbol{\eta}}) - R_{\mathcal{D}}^{\Phi_{\phi^c}}(\hat{\boldsymbol{\eta}}^*)$. For ϕ considered in Theorem 5 with regret transfer bound $\mathcal{O}(\epsilon^\alpha)$ for $\alpha \in (0,1]$, the following regret transfer bound holds:

$$\texttt{Regret}_{\mathcal{D}}^{\ell_H^c}(\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}) \leq k \overline{L}_{\mathcal{D}}(\hat{\boldsymbol{\eta}})^{1-\alpha} \left(\texttt{Regret}_{\mathcal{D}}^{\Phi_\phi^c}(\hat{\boldsymbol{\eta}})\right)^\alpha,$$

where $k > 0$ only depends on ϕ and $\overline{L}_{\mathcal{D}}(\hat{\boldsymbol{\eta}}) = \mathbb{E}_{p(\boldsymbol{x})}[\ell_H(\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}), f_c^*(\boldsymbol{x}))]$ is the averaged distance

between the estimated and Bayes optimal predictions (holds for any valid f_c^*).

The proof of Theorem 6 is provided in Appendix I.

Notice that we can immediately get a looser but data-independent bound by substitute $\overline{L}_{\mathcal{D}}(\hat{\boldsymbol{\eta}})$ with the diameter of the tree $d(H)$. Given this conclusion, we directly use it to provide a more detailed bound for the widely used log loss and MAE loss:

Corollary 1. Regret transfer bound when the Log loss is used:

$$\texttt{Regret}_{\mathcal{D}}^{\ell_H^c}(\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}) \leq 2\sqrt{\overline{L}_{\mathcal{D}}(\hat{\boldsymbol{\eta}})\texttt{Regret}_{\mathcal{D}}^{\Phi_{\text{Log}}^c}(\hat{\boldsymbol{\eta}})}.$$

Regret transfer bound when MAE is used:

$$\texttt{Regret}_{\mathcal{D}}^{\ell_H^c}(\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}) \leq 2\texttt{Regret}_{\mathcal{D}}^{\Phi_{\text{MAE}}^c}(\hat{\boldsymbol{\eta}}).$$

The proof of Corollary 1 is provided in Appendix J.

It is noticeable that while our MAE surrogate and the OvA hinge surrogate in Ramaswamy et al. (2015) both achieve linear regret. However, the hinge-like surrogate's regret transfer bound depends linearly on a constant that is determined by the diameter of the tree, which can be large when the class number increases. According to the results above, our MAE surrogate's regret transfer bound does not rely on the tree's structure, which indicates that our method benefits from its coherent model design.

7 Experiments

In this section, we conduct experiments to empirically evaluate the performance of our proposed method by comparing it with existing baselines on the CIFAR-100 dataset and a deep model.

7.1 Experimental Setup

We first compare all the methods according to the error measured by our proposed generalized tree distance loss with $c \in \{0.0, 0.2, 0.5, 0.8\}$, and further provide the averaged level of the predictions and the induced superclass's misclassification rate for further reference. The three statistics are shown in Table 2, 3, and 4, respectively. All the experiments are conducted with 8 NVIDIA GeForce 3090 GPUs. More details and results can be found in the appendix.

Baselines. We compare our consistent implementations LOG (1) and MAE (2) with following baselines:

- FLAT: the plug-in classifier method that works by directly estimating class-posterior probability with cross-entropy loss and then generating the prediction according to the characterization of Bayes optimal solution (2).
- OvA: the OvA hinge loss based multiclass classification with rejection reduction proposed in Ramaswamy et al. (2015). To make it capable of dealing with $\frac{1+c}{2} > 0.5$, we generalize it by using its corresponding formulation with a rejection threshold larger than 0.5 in Ramaswamy et al. (2018).

Dataset, model, and optimizers. We provide the details about our used dataset with the corresponding model and optimizers as follows:

- We conduct experiments based on the dataset of CIFAR-100 (Krizhevsky, 2012). We process it into a dataset with label hierarchy by integrating its 20 coarse labels and finally get a label tree with 2 levels, 21 non-leaf labels, and 100 leaf labels.
- To validate all the methods' performance when combined with deep models, we use a 28-layer WideResNet Zagoruyko and Komodakis (2016).
- For the method of LOG and FLAT, we use SGD with cosine annealing as the optimizer, and the epoch number, learning rate, weight decay, and batch size are set to 400, 1e-1. 5e-4, and 128. For MAE and OvA, Adam (Kingma and Ba, 2015) is used as the optimizer. The learning rate is set to 1e-3 and other parameters are the same as in the setting of SGD.

7.2 Experimental Results

Combining Table 2-4, we can learn that:

- Our method LOG outperforms FLAT consistently over all the selections of penalty cost c. To explain this observation, we can refer to Table 3 and 4. Notice that the averaged level of prediction for method FLAT is larger than that for LOG, while the misclassification error of its induced superclass is always higher than LOG's. These observations indicate that FLAT generated predictions that are higher than necessary on average. The cause of this phenomenon can be that it omits the label hierarchy in the training process, which leads to a biased estimation of prediction.
- Though MAE and OvA achieve linear regret transfer bounds, they both suffer from the problem of underfitting: the level of their prediction is obviously lower than necessary. This phenomenon can

Table 2: Experimental results of generalized tree distance loss of predictions on the tree structure of CIFAR-100 for 5 trails. The best performance is highlighted in bold.

Method		Penalty cost c			
		0.0	0.2	0.5	0.8
Previously proposed	FLAT	0.70 (0.02)	1.03 (0.01)	1.50 (0.02)	1.91 (0.01)
	OvA	1.11 (0.04)	1.73 (0.16)	2.01 (0.07)	2.42 (0.09)
Our proposed	MAE	0.93 (0.03)	1.25 (0.02)	1.84 (0.03)	2.16 (0.01)
	LOG	**0.68 (0.01)**	**0.98 (0.01)**	**1.41 (0.02)**	**1.88 (0.01)**

Table 3: Experimental results of the averaged level of predictions on the tree structure of CIFAR-100 for 5 trails.

Method		Penalty cost c			
		0.0	0.2	0.5	0.8
Previously proposed	FLAT	1.81 (0.03)	1.73 (0.02)	1.60 (0.02)	1.43 (0.02)
	OvA	1.21 (0.10)	1.245 (0.05)	1.08 (0.07)	0.68 (0.14)
Our proposed	MAE	1.50 (0.01)	1.52 (0.01)	1.10 (0.03)	0.97 (0.17)
	LOG	1.79 (0.01)	1.66 (0.02)	1.54 (0.01)	1.39 (0.01)

Table 4: Experimental results of the induced superclass's misclassification error of predictions on the tree structure of CIFAR-100 for 5 trails. The results are rescaled to 0-100.

Method		Penalty cost c			
		0.0	0.2	0.5	0.8
Previously proposed	FLAT	17.34 (0.07)	13.87 (0.16)	10.06 (0.14)	7.42 (0.18)
	OvA	28.18 (1.38)	25.99 (0.61)	24.48 (1.13)	23.78 (0.92)
Our proposed	MAE	17.26 (0.38)	16.67 (0.34)	12.66 (0.26)	11.86 (0.96)
	LOG	16.76 (0.09)	11.45 (0.15)	8.45 (0.12)	6.11 (0.06)

be caused by the empirical finding (Zhang and Sabuncu, 2018; Feng et al., 2020) that the MAE loss and hinge loss are all hard to optimize due to their sparse/zero gradients. Furthermore, the performance of OvA is outperformed by MAE, which can be attributed to that hinge loss is not differentiable, which can be troublesome in gradient-based optimization.

In conclusion, our proposed method LOG consistently outperforms the baseline methods due to its, which again validates the efficacy of our formulation.

8 Conclusion

In this paper, we studied the problem of the design of evaluation metrics and novel loss functions for hierarchical classification. We first generalized the popular evaluation metric (i.e., tree distance loss) to make it able to reflect the risk sensitivity of different tasks. Then we gave an intuitive representation of our proposed generalized loss and used it to induce a problem reduction (more specifically, from hierarchical classification to binary classification). A more detailed analysis

of the used model was conducted to further justify the rationality of this problem reduction from the aspect of regret. Finally, we derived consistent surrogates for the proposed generalized tree distance loss that can be compatible with various binary losses and showed that the regret transfer bounds can further characterize the property of proposed methods. Experimental results demonstrate the efficacy of our methods.

Acknowledgements

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISGAward No: AISG2-GC-2023-009).

References

Awasthi, P., Frank, N., Mao, A., Mohri, M., and Zhong, Y. (2021a). Calibration and consistency of adversarial surrogate losses. *CoRR*, abs/2104.09658.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. (2021b). A finer calibration analysis for adversarial robustness. *CoRR*, abs/2105.01550.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. (2023). Theoretically grounded loss functions and algorithms for adversarial robustness. In *AISTATS*.

Babbar, R., Partalas, I., Gaussier, É., and Amini, M. (2013). On flat versus hierarchical classification in large-scale taxonomies. In *NeurIPS*.

Bao, H. (2023). Proper losses, moduli of convexity, and surrogate regret bounds. In Neu, G. and Rosasco, L., editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 525–547. PMLR.

Bao, H., Scott, C., and Sugiyama, M. (2020). Calibrated surrogate losses for adversarially robust classification. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 408–451. PMLR.

Bao, H. and Sugiyama, M. (2020). Calibrated surrogate maximization of linear-fractional utility in binary classification. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 2337–2347. PMLR.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., and Lord, N. A. (2020). Making better mistakes: Leveraging class hierarchies with deep networks. In *CVPR*.

Cao, Y., Cai, T., Feng, L., Gu, L., Gu, J., An, B., Niu, G., and Sugiyama, M. (2022). Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In *NeurIPS*.

Cesa-Bianchi, N., Gentile, C., Tironi, A., and Zaniboni, L. (2004). Incremental algorithms for hierarchical classification. In *NeurIPS*.

Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. (2006). Hierarchical classification: combining bayes with SVM. In *ICML*.

Charoenphakdee, N., Cui, Z., Zhang, Y., and Sugiyama, M. (2021). Classification with rejection based on cost-sensitive classification. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 1507–1517. PMLR.

Cortes, C., DeSalvo, G., and Mohri, M. (2016a). Boosting with abstention. In *NeurIPS*, pages 1660–1668.

Cortes, C., DeSalvo, G., and Mohri, M. (2016b). Learning with rejection. In *ALT*, volume 9925, pages 67–82.

Dekel, O. (2009). Distribution-calibrated hierarchical classification. In *NeurIPS*.

Dekel, O., Keshet, J., and Singer, Y. (2004). Large margin hierarchical classification. In *ICML*.

Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. (2012). Agnostic learning of monomials by halfspaces is hard. *SIAM J. Comput.*, 41(6):1558–1590.

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. (2020). Can cross entropy loss be robust to label noise? In *IJCAI*.

Finocchiaro, J., Frongillo, R. M., and Waggoner, B. (2019). An embedding framework for consistent polyhedral surrogates. In *NeurIPS*, pages 10780–10790.

Gao, W. and Zhou, Z. (2013). On the consistency of multi-label learning. *Artif. Intell.*, 199-200:22–44.

Gao, W. and Zhou, Z. (2015). On the consistency of AUC pairwise optimization. In *IJCAI*, pages 939–945.

Giunchiglia, E. and Lukasiewicz, T. (2020). Coherent hierarchical multi-label classification networks. In *NeurIPS*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Koyejo, O., Natarajan, N., Ravikumar, P., and Dhillon, I. S. (2015). Consistent multilabel classification. In *NeurIPS*, pages 3321–3329.

Krizhevsky, A. (2012). Learning multiple layers of features from tiny images. *University of Toronto*.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *ICML*.

Lapin, M., Hein, M., and Schiele, B. (2018). Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(7):1533–1554.

Mao, A., Mohri, M., and Zhong, Y. (2023a). Cross-entropy loss functions: Theoretical analysis and applications. In *ICML*.

Mao, A., Mohri, M., and Zhong, Y. (2023b). H-consistency bounds for pairwise misranking loss surrogates. In *ICML*.

Menon, A. K. and Williamson, R. C. (2014). Bayes-optimal scorers for bipartite ranking. In *COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 68–106. JMLR.org.

Ni, C., Charoenphakdee, N., Honda, J., and Sugiyama, M. (2019). On the calibration of multiclass classification with rejection. In *NeurIPS*, pages 2582–2592.

Pires, B. Á. and Szepesvári, C. (2016). Multiclass classification calibration functions. *CoRR*, abs/1609.06385.

Ramaswamy, H. G. and Agarwal, S. (2016). Convex calibration dimension for multiclass loss matrices. *J. Mach. Learn. Res.*, 17:14:1–14:45.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2015). Convex calibrated surrogates for hierarchical classification. In *ICML*.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2018). Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530–554.

Reid, M. D. and Williamson, R. C. (2009). Surrogate regret bounds for proper losses. In *ICML*.

Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *J. Mach. Learn. Res.*, 11:2387–2422.

Sun, A. and Lim, E. (2001). Hierarchical text classification and evaluation. In *ICDM*, pages 521–528.

Tewari, A. and Bartlett, P. L. (2007). On the consistency of multiclass classification methods. *J. Mach. Learn. Res.*, 8:1007–1025.

Valmadre, J. (2022). Hierarchical classification at multiple operating points. In *NeurIPS*.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The inaturalist species classification and detection dataset. In *CVPR*.

Wehrmann, J., Cerri, R., and Barros, R. C. (2018). Hierarchical multi-label classification networks. In *ICML*.

Williamson, R. C., Vernet, E., and Reid, M. D. (2016). Composite multiclass losses. *J. Mach. Learn. Res.*, 17:223:1–223:52.

Wu, G., Li, C., Xu, K., and Zhu, J. (2021). Rethinking and reweighting the univariate losses for multi-label ranking: Consistency and generalization. *CoRR*, abs/2105.05026.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.

Yang, F. and Koyejo, S. (2020). On the consistency of top-k surrogate losses. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 10727–10735. PMLR.

Yoshida, S. M., Takenouchi, T., and Sugiyama, M. (2021). Lower-bounded proper losses for weakly supervised classification. In *ICML*.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *BMVC*.

Zhang, H., Zhan, T., Basu, S., and Davidson, I. (2021). A framework for deep constrained clustering. *Data Mining and Knowledge Discovery*.

Zhang, M., Ramaswamy, H. G., and Agarwal, S. (2020). Convex calibrated surrogates for the multi-label f-measure. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 11246–11255. PMLR.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251.

Zhang, Z. and Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pages 8792–8802.

A Proof of Theorem 1

Proof. We focus on the case of $c > 0$. In this proof, we denote by $R^{\ell_H^c}(y', \boldsymbol{\eta}(\boldsymbol{x})) = \sum_{y\in[K]} \eta_y(\boldsymbol{x})\ell_H^c(y', y)$ the inner risk. Since $[K]$ is finite, the existence of the solution is proven. We first show that any y' that does not meet the condition in the theorem cannot be the solution, and then show that all the y that meet the condition in Theorem 1 have the same value of inner risk, which concludes the proof.

(A). $\Leftarrow$:

Fixing a $\boldsymbol{\eta}(\boldsymbol{x}) \in \Delta^K$ and $c \in [0, 1]$, we assume that there exist a $y' = f_c^*(\boldsymbol{x})$ which does not meet the condition in Theorem 1. Then $y' \notin \operatorname*{argmax}_{W_y(\boldsymbol{\eta}(\boldsymbol{x}))\geq 0.5+\frac{c}{2}} \mathrm{Lev}(y)$ and $y' \notin U_{\boldsymbol{\eta}(\boldsymbol{x})}(0.5 + \frac{c}{2})$. There is two cases:

①. $W_{y'}(\boldsymbol{\eta}(\boldsymbol{x})) < 0.5 + \frac{c}{2}$.

In this case, we can learn that y' must not be the root node 1. Denote by $\tilde{y}$ its parent. We can learn that:

$$\begin{aligned}
&R^{\ell_H^c}(y', \boldsymbol{\eta}(\boldsymbol{x})) - R^{\ell_H^c}(\tilde{y}, \boldsymbol{\eta}(\boldsymbol{x})) \\
&= c - W_{y'}(\boldsymbol{\eta}(\boldsymbol{x})) + (1 - W_{y'})(\boldsymbol{\eta}(\boldsymbol{x})) \\
&= c - 2W_{y'}(\boldsymbol{\eta}(\boldsymbol{x}) + 1 \\
&> c - 1 - c + 1 \\
&= 0,
\end{aligned}$$

which means $R^{\ell_H^c}(y', \boldsymbol{\eta}(\boldsymbol{x})) > R^{\ell_H^c}(\tilde{y}, \boldsymbol{\eta}(\boldsymbol{x}))$ and thus leads to a contradiction.

②. $W_{y'}(\boldsymbol{\eta}(\boldsymbol{x})) \geq 0.5 + \frac{c}{2}$,

Denote by $y^* \in \operatorname{argmax}_{W_y(\boldsymbol{\eta}(\boldsymbol{x}))\geq 0.5+\frac{c}{2}} \mathrm{Lev}(y)$. We can learn that $\mathrm{Lev}(y') < \mathrm{Lev}(y^*)$. In this case, y^* must be y''s descendant, otherwise $\sum_{y\in[K]} p(y|\boldsymbol{x}) > 1 + c$. Furthermore, y''s children y'', which is the ancestor of y^*, has $W_{y''}(\boldsymbol{\eta}(\boldsymbol{x})) > 0.5 + \frac{c}{2}$ since $y' \notin U_{\boldsymbol{\eta}(\boldsymbol{x})}(0.5 + \frac{c}{2})$. Then we can learn:

$$\begin{aligned}
&R^{\ell_H^c}(y', \boldsymbol{\eta}(\boldsymbol{x})) - R^{\ell_H^c}(y'', \boldsymbol{\eta}(\boldsymbol{x})) \\
&= W_{y''}(\boldsymbol{\eta}(\boldsymbol{x})) - (1 - W_{y''})(\boldsymbol{\eta}(\boldsymbol{x})) - c \\
&= 2W_{y''}(\boldsymbol{\eta}(\boldsymbol{x})) - 1 - c \\
&> 0
\end{aligned}$$

which means $R^{\ell_H^c}(y', \boldsymbol{\eta}(\boldsymbol{x})) > R^{\ell_H^c}(y'', \boldsymbol{\eta}(\boldsymbol{x}))$ and thus leads to a contradiction.

(B). $\Rightarrow$

We focus on the set $\operatorname{argmax}_{W_y(\boldsymbol{\eta}(\boldsymbol{x}))\geq 0.5+\frac{c}{2}} \mathrm{Lev}(y)$. It is a **singleton**: if there exist $y' \neq y''$ that are both in this set, $1 = \sum_{y\in[K]} p(y|\boldsymbol{x}) \geq W_{y'}(\eta(\boldsymbol{x})) + W_{y''}(\eta(\boldsymbol{x})) \geq 1 + c$, which leads to a contradiction.

Denote by the unique element in the set by y^*. If $U_{\boldsymbol{\eta}(\boldsymbol{x})}(0.5 + \frac{c}{2})$ is empty, the proof is done. When it is non-empty, we can learn that: any element in $U_{\boldsymbol{\eta}(\boldsymbol{x})}(0.5 + \frac{c}{2})$ must be the ancestor of y^* and we can prove it use the same contradiction as in the previous paragraph. For any $y' \in U_{\boldsymbol{\eta}(\boldsymbol{x})}(0.5 + \frac{c}{2})$, denote by y'' the child with $W_{y''}(\boldsymbol{\eta}(\boldsymbol{x})) = 0.5 + \frac{c}{2}$. We can learn that it must be y^* or y^*'s ancestor, and $W_{y^*}(\eta(\boldsymbol{x})) = 0.5 + \frac{c}{2}$. Then for any y', suppose the path from it to y^* is $y' \to y_1 \to \cdots \to y_n \to y^*$. We can learn $W_{y_i}(\boldsymbol{\eta}(\boldsymbol{x}))$ for $i = 1, \cdots, n$. Denote by $y' = y_0,\ y^* = y_{n+1}$, for each step on the path:

$$\Delta^i = R^{\ell_H^c}(y_i, \boldsymbol{\eta}(\boldsymbol{x})) - R^{\ell_H^c}(y_{i+1}, \boldsymbol{\eta}(\boldsymbol{x})) = 0,\ i = 0, \cdots, n.$$

We can learn:

$$R^{\ell_H^c}(y', \boldsymbol{\eta}(\boldsymbol{x})) - R^{\ell_H^c}(y^*, \boldsymbol{\eta}(\boldsymbol{x})) = \sum_{i=1}^{n} \Delta^i = 0,$$

and thus we can conclude the proof. □

B Proof of Theorem 2

Proof. First we prove that $\sum_{i=2}^{K} \mathbb{I}(y' \notin T_i^0) = \text{Lev}(y')$. Then we prove $\ell_H^c(y', y) = \sum_{i=2}^{K} \mathbb{I}(y' \notin T_i^{s_i(y)})$.

Denote by $1 \to y_1 \cdots \to y_{\text{Lev}(y')-1} \to y'$ the path between 1 and y'. We can learn that $y' \notin T_{y_i}^0$ for $i = 1, \cdots$, and $y' \in T_{y''}^0$ for other y''. Then $\sum_{i=2}^{K} \mathbb{I}(y' \notin T_i^0) = \sum_{i=1}^{\text{Lev}(y')-1} \mathbb{I}(y' \notin T_i^0) + \mathbb{I}(y' \notin T_{y'}^0) = \text{Lev}(y')$.

Denote by $\tilde{y}$ the node of the highest level among all the common ancestors of y and y'. Denote the path between y and y' by $y \to y_1 \to \cdots \to y_n \to \tilde{y} \to y_1' \cdots \to y_{n'}' \to y'$. We can learn that $\mathbb{I}(y' \notin T_i^{s_i(y)}) = 1$ if and only if i is on the path except $\tilde{y}$: if an i is not on the path , y and y' will be both in T_i^0 or T_i^1; if $i = \tilde{y}$, y and y' will be both in T_i^1. Then we can learn $\sum_{i=2}^{K} \mathbb{I}(y' \notin T_i^{s_i(y)}) = n + n' + 2 = \ell_H(y', y)$. □

C Proof of Lemma 1

Proof. We can learn that

$$\begin{aligned}
\mathbb{E}_{p(y|\boldsymbol{x})}[\mathbb{I}(y' \notin T_i^{s_i(y)}) + c*\mathbb{I}(y' \notin T_i^0)] &= Pr(Y \in T_i^1|\boldsymbol{x})\mathbb{I}(f(\boldsymbol{x}) \notin T_i^1) + (1 + c - Pr(Y \in T_i^1|\boldsymbol{x}))\mathbb{I}(f(\boldsymbol{x}) \notin T_i^0) \\
&= (1 + c)\left[p_i(1|\boldsymbol{x})\mathbb{I}(f(\boldsymbol{x}) \notin T_i^1) + p_i(0|\boldsymbol{x})\mathbb{I}(f(\boldsymbol{x}) \notin T_i^0)\right] \\
&= (1 + c)\mathbb{E}_{p_i(\gamma|\boldsymbol{x})}[\mathbb{I}(f(\boldsymbol{x}) \notin T_i^{\gamma})]
\end{aligned}$$

Further taking the expectation *w.r.t.* $p(\boldsymbol{x})$ and we can conclude the proof. □

D Proof of Theorem 3

Proof. Given the binary classification problems, we can explicitly give their optimal solutions $\forall i \in [2, K]$:

$$f^{i*}(\boldsymbol{x}) = \mathbb{I}(Pr(Y \in T_i) \geq \frac{1 + c}{2})$$

Then we can learn that $\varphi \circ \mathcal{F}^*(\boldsymbol{x}) = \underset{W_y(\boldsymbol{\eta}(\boldsymbol{x})) \geq \frac{1+c}{2}}{\text{argmax}} \text{Lev}(y)$, which is a Bayes optimal solution for ℓ_H^c. The uniqueness is due to the fact that there cannot be to mutual exclusive $Y_1, Y_2 \subset \mathcal{Y}$ that $\sum_{y \in Y_{1/2}} p(y|\boldsymbol{x}) \geq \frac{1+c}{2}$ otherwise $\sum_{y=1}^{K} p(y|\boldsymbol{x}) \geq 1 + c > 0$. □

E Proof of Theorem 4

Proof. Notice that according to Theorem 3, $\mathcal{F}^*$ is also H-coherent. Meanwhile, $f_c^* = \varphi \circ \mathcal{F}^*$ is the Bayes optimal solution for our target loss. Then we can write:

$$R_{\mathcal{D}}^{\ell_H^c}(f_c^*) = (1 + c)\sum_{i=2}^{K} R_{\mathcal{D}_i^{\ell_{01}}}(f^{i*}),$$

Furthermore, since $\mathcal{F}$ is H-coherent:

$$R_{\mathcal{D}}^{\ell_H^c}(\varphi \circ \mathcal{F}) = (1 + c)\sum_{i=2}^{K} R_{\mathcal{D}_i^{\ell_{01}}}(f^i),$$

and thus we can conclude the proof using the definition of ℓ_{01}^c's regret and the equations above. □

F Proof of Lemma 2

Proof. First, $\varphi \circ \mathcal{F}_{\hat{\eta}}(\boldsymbol{x})$ is non-empty. Denote by A_y the ancestors of y (except 1 and include y). Denote by $\hat{y} = \underset{W_y(\hat{\eta}(\boldsymbol{x})) > \frac{1}{2}}{\text{argmax}} \text{Lev}(y)$. For $y' \in [2, K]/A_{\hat{y}}$, we can learn that $f_{\hat{\eta}}^{y'}(\boldsymbol{x}) = 0$ and $\hat{y} \in T_{y'}^0$. For $y' \in A_{\hat{y}}$, $f_{\hat{\eta}}^{y'}(\boldsymbol{x}) = 1$ and $\hat{y} \in T_{y'}^1$, and thus $\hat{y}$ must be in this set.

Secondly, we show that it is a singleton. Suppose there are two different elements y_1 and y_2 in this set, they cannot be each other's ancestor according to the definition of φ, and thus T_{y_1} and T_{y_2} are mutually exclusive. Then we can learn that $\bar{\eta}_{y_1}(\boldsymbol{x}) > 0.5$ and $\bar{\eta}_{y_2}(\boldsymbol{x}) > 0.5$. Then $\sum_y \hat{\eta}_y(\boldsymbol{x}) \geq \bar{\eta}_{y_1}(\boldsymbol{x}) + \bar{\eta}_{y_1}(\boldsymbol{x}) > 1$, which leads to a contradiction and concludes the proof. □

G Proof of the Non-Coherency of Simple Binary Reduction

Proof. The proof is direct. Suppose we have a label hierarchy $1 \to 2 \to 3 \to 4$ and we directly construct each dummy distribution, suppose there is a sample $\boldsymbol{x}$ that $\eta^2(\boldsymbol{x}) = 0.2$ and $\eta^3(\boldsymbol{x}) = 0.7$, a contradiction can be derived since the derived $f^3(\boldsymbol{x}) = 1$ and $f^2(\boldsymbol{x}) = 0$ and $T_3^1 \cap T_2^0 = \emptyset$. □

H Proof of Theorem 5

Proof. First of all, we study the Bayes optimal solution *w.r.t.* the surrogate loss.

For strictly proper losses, the optimal solution is $\bar{\eta}_i^*(\boldsymbol{x}) = \sum_{y \in T_i} \frac{p(y|\boldsymbol{x})}{1+c}$ for $i \in [2, K]$, and thus we can learn that $\hat{\eta}_i^*(\boldsymbol{x}) = \frac{p(i|\boldsymbol{x})}{1+c}$ for $i \in [2, K]$. Given this characterization, the consistency is directly verified.

For the MAE loss, $\bar{\eta}_i^*(\boldsymbol{x}) = 1$ if $Pr(Y \in T_i|\boldsymbol{x}) > 1 + c$ and 0 otherwise, and thus $\hat{\eta}_i^*(\boldsymbol{x}) = 1$ if and only if $Pr(Y \in T_i|\boldsymbol{x}) > 1 + c$ and $Pr(Y \in T_j|\boldsymbol{x}) \leq 1 + c$ for all of $i's$ children j. This characterization directly lead to the consistency according to the definition of $f_{\hat{\eta}}^i$ and the Bayes optimal solution.

□

I Proof of Theorem 6

Proof. According to Theorem 6, we have:

$$\texttt{Regret}_{\mathcal{D}}^{\ell_H^c}(\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}) = (1 + c) \sum_{i=2}^{K} \texttt{Regret}_{\mathcal{D}_i}^{\ell_{01}}(f_{\hat{\boldsymbol{\eta}}}^i).$$

We can further rewrite it into the point-wise regret version, which is equal to the regret with $p(\boldsymbol{x}') = \delta(\boldsymbol{x}' - \boldsymbol{x})$:

$$\texttt{Regret}^{\ell_H^c}(\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x})) = (1 + c) \sum_{i=2}^{K} \texttt{Regret}^{\ell_{01}}(f_{\hat{\boldsymbol{\eta}}}^i(\boldsymbol{x}), \eta^i(\boldsymbol{x})).$$

For any f_c^*, suppose all the nodes on the path from $\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x})$ to f_c^* except their common ancestor consists the set $\mathcal{Y}_{\hat{\boldsymbol{\eta}}}$. We can learn that the size of this set is $\ell_H(\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}), f_c^*(\boldsymbol{x}))$ from the proof of Theorem 2. For those $i \notin \mathcal{Y}_{\hat{\boldsymbol{\eta}}}$, the binary subproblem is correctly solved, i.e., $f_{\hat{\boldsymbol{\eta}}}^i(\boldsymbol{x}) = f^{i*}(\boldsymbol{x})$ and thus $\texttt{Regret}^{\ell_{01}}(f_{\hat{\boldsymbol{\eta}}}^i(\boldsymbol{x}), \eta^i(\boldsymbol{x})) = 0$, then we can learn:

$$\texttt{Regret}^{\ell_H^c}(\varphi \circ \mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x})) = (1 + c) \sum_{i \in \mathcal{Y}_{\hat{\boldsymbol{\eta}}}} \texttt{Regret}^{\ell_{01}}(f_{\hat{\boldsymbol{\eta}}}^i(\boldsymbol{x}), \eta^i(\boldsymbol{x})).$$

Combining the regret transfer bound of ϕ, Jensen's inequality, and our loss formulation, we can learn:

$$
\begin{aligned}
&(1+c)\sum_{i\in\mathcal{Y}_{\hat{\boldsymbol{\eta}}}}\texttt{Regret}^{\ell_{01}}(f^i_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}),\eta^i(\boldsymbol{x}))\\
&\leq (1+c)\sum_{i\in\mathcal{Y}_{\hat{\boldsymbol{\eta}}}}\left(\texttt{Regret}^{\phi}(\bar{\eta}_i(\boldsymbol{x}),\eta^i(\boldsymbol{x}))\right)^{\alpha}\\
&\leq (1+c)k'|\mathcal{Y}_{\hat{\boldsymbol{\eta}}}|\left(\frac{\sum_{i\in\mathcal{Y}_{\hat{\boldsymbol{\eta}}}}\texttt{Regret}^{\phi}(\bar{\eta}_i(\boldsymbol{x}),\eta^i(\boldsymbol{x})}{|\mathcal{Y}_{\hat{\boldsymbol{\eta}}}|}\right)^{\alpha}\\
&= (1+c)k'\ell_H(\varphi\circ\mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}),f^*_c(\boldsymbol{x})\left(\frac{\sum_{i\in\mathcal{Y}_{\hat{\boldsymbol{\eta}}}}\texttt{Regret}^{\phi}(\bar{\eta}_i(\boldsymbol{x}),\eta^i(\boldsymbol{x})}{\ell_H(\varphi\circ\mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}),f^*_c(\boldsymbol{x})}\right)^{\alpha}\\
&= (1+c)k'\ell_H(\varphi\circ\mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}),f^*_c(\boldsymbol{x}))^{1-\alpha}\left(\sum_{i\in\mathcal{Y}_{\hat{\boldsymbol{\eta}}}}\texttt{Regret}^{\phi}(\bar{\eta}_i(\boldsymbol{x}),\eta^i(\boldsymbol{x})\right)^{\alpha}\\
&\leq (1+c)k'\ell_H(\varphi\circ\mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}),f^*_c(\boldsymbol{x}))^{1-\alpha}\left(\sum_{i=2}^{K}\texttt{Regret}^{\phi}(\bar{\eta}_i(\boldsymbol{x}),\eta^i(\boldsymbol{x})\right)^{\alpha}\\
&\leq (1+c)k'\ell_H(\varphi\circ\mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}),f^*_c(\boldsymbol{x}))^{1-\alpha}\left(\texttt{Regret}^{\Phi^c_\phi}(\hat{\boldsymbol{\eta}}(\boldsymbol{x}),\boldsymbol{\eta}(\boldsymbol{x}))/(1+c)\right)^{\alpha}\\
&\leq 2k'\ell_H(\varphi\circ\mathcal{F}_{\hat{\boldsymbol{\eta}}}(\boldsymbol{x}),f^*_c(\boldsymbol{x}))^{1-\alpha}\left(\texttt{Regret}^{\Phi^c_\phi}(\hat{\boldsymbol{\eta}}(\boldsymbol{x}),\boldsymbol{\eta}(\boldsymbol{x}))\right)^{\alpha}
\end{aligned}
$$

Using the expectation version of Holder's inequality and let $k = 2k'$ and we can conclude the proof. □

J Proof of Corollary 1

Proof. The conclusion immediately holds since for log loss, $\alpha = \frac{1}{2}$ and $k = 2$; for MAE, $\alpha = 1$ and $k = 2$. □

K Additional Experimental Results on Fashion-MNIST

In this section, we evaluate our method on the Fashion-MNIST (Xiao et al., 2017) dataset with a WorkNet induced label hierarchy (Zhang et al., 2021). We use a CNN with the same architecture as in Charoenphakdee et al. (2021). Adam is used as the optimizer and the epoch number, batchsize, learning rate are set to 50, 128, and 1e-3, respectively. These experiments are conducted with a NVIDIA GeForce 4090 GPU. We report and compare the same statistics as in the experiments of CIFAR-100, which are listed in Table 5-7. According to the results, it can be seen that our proposed methods outperform baselines, while MAE obviously outperform other methods. It is a natural result since the CNN is a simpler model compared with the WideResNet and the class number is also smaller, which makes the optimization easier for MAE.

Table 5: Experimental results of generalized tree distance loss of predictions for all the methods on the tree structure of Fashion-MNIST for 5 trails. The best performance is highlighted in bold.

Method		Penalty cost c			
		0.0	0.2	0.5	0.8
Previously proposed	FLAT	0.24 (0.05)	1.20 (0.03)	2.60 (0.05)	3.98 (0.02)
	OvA	0.32 (0.03)	1.47 (0.11)	2.80 (0.17)	4.13 (0.38)
Our proposed	MAE	**0.20 (0.05)**	**1.15 (0.01)**	**2.58 (0.01)**	**3.96 (0.03)**
	LOG	0.22 (0.02)	**1.16** (**0.01**)	2.59 (0.07)	4.01 (0.04)

Table 6: Experimental results of the averaged level of predictions for all the methods on the tree structure of Fashion-MNIST for 5 trails.

Method		Penalty cost c			
		0.0	0.2	0.5	0.8
Previously proposed	FLAT	4.77 (0.08)	4.70 (0.03)	4.63 (0.05)	4.30 (0.18)
	OvA	4.70 (0.13)	4.42 (0.11)	4.16 (0.21)	3.65 (0.28)
Our proposed	MAE	4.80 (0.03)	4.75 (0.07)	4.69 (0.05)	4.58 (0.14)
	LOG	4.78 (0.05)	4.71 (0.10)	4.55 (0.08)	4.10 (0.06)

Table 7: Experimental results of the induced superclass's misclassification error of predictions for all the methods on the tree structure of Fashion-MNIST for 5 trails. The results are rescaled to 0-100.

Method		Penalty cost c			
		0.0	0.2	0.5	0.8
Previously proposed	FLAT	6.26 (0.08)	4.62 (0.03)	3.44 (0.05)	0.97 (0.08)
	OvA	6.05 (0.53)	5.90 (1.21)	2.88 (0.56)	2.36 (0.72)
Our proposed	MAE	5.75 (0.05)	4.91 (0.12)	3.47 (0.06)	2.16 (0.04)
	LOG	5.93 (0.19)	3.95 (0.09)	1.94 (0.07)	1.04 (0.11)

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**] Please refer to the definitions and description of each theorem/lemma.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**] We provide the regret transfer bound to further characterize the properties of our method.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**No**]
We will provide a demo for our proposed method in the future version.

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**]
   (b) Complete proofs of all theoretical results. [**Yes**]
   (c) Clear explanations of any assumptions. [**Yes**]

   Please refer to the definitions and description of each theorem/lemma. The proof is provided in the appendix

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**]

   Please refer to the experiment part and the appendix for the detailed experimental setup.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [**Yes**]
   (b) The license information of the assets, if applicable. [**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]
   (d) Information about consent from data providers/curators. [**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]