Analysis for Abductive Learning and Neural-Symbolic Reasoning Shortcuts

Xiao-Wen Yang [1 2] **Wen-Da Wei** [1 2] **Jie-Jing Shao** [1] **Yu-Feng Li** [1 2] **Zhi-Hua Zhou** [1 2]

Abstract

Abductive learning models (ABL) and neural-symbolic predictive models (NeSy) have been recently shown effective, as they allow us to infer labels that are consistent with some prior knowledge by reasoning over high-level concepts extracted from sub-symbolic inputs. However, their generalization ability is affected by *reasoning shortcuts*: high accuracy on given targets but leveraging intermediate concepts with unintended semantics. Although there have been techniques to alleviate reasoning shortcuts, theoretical efforts on this issue remain to be limited. This paper proposes a simple and effective analysis to quantify harm caused by it and how can mitigate it. We quantify three main factors in how NeSy algorithms are affected by reasoning shortcuts: the complexity of the knowledge base, the sample size, and the hypothesis space. In addition, we demonstrate that ABL can reduce shortcut risk by selecting specific distance functions in consistency optimization, thereby demonstrating its potential and approach to solving shortcut problems. Empirical studies demonstrate the rationality of the analysis. Moreover, the proposal is suitable for many ABL and NeSy algorithms and can be easily extended to handle other cases of reasoning shortcuts.

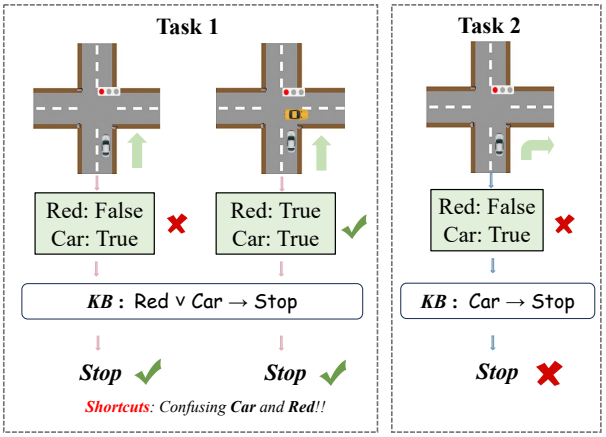

Figure 1. **Reasoning shortcuts.** In Task 1, an autonomous vehicle needs to decide whether to move forward or stop based on the given rule: `Red ∨ Car → Stop`, which means it should stop if there is a red light or a vehicle ahead. The model trained on this task could correctly classify the target, but it acquires a *reasoning shortcut* by confusing the presence of a vehicle and the red light. When the perception model learned in Task 1 is applied to Task 2, which involves determining whether turning right is permissible, the autonomous vehicle mistakenly decides it should stop when the perception model incorrectly predicts a vehicle ahead, possibly resulting in a dangerous situation.

1. Introduction

Recently, neural-symbolic learning (NeSy) (d'Avila Garcez et al., 2019) and abductive learning (ABL) (Zhou, 2019) have been shown effective for integration of raw data and symbolic rules. Neural-symbolic learning (d'Avila Garcez et al., 2019; Sarker et al., 2021; Cunnington et al., 2022) focuses on integrating logical reasoning into the neural networks in an end-to-end manner. Deep neural networks (LeCun et al., 2015) serve as low-level perception models to translate raw inputs into symbolic concepts of practical meaning, while the symbolic knowledge constrains both the intermediate symbolic concepts and the final target using logical rules. Abductive learning (Zhou, 2019) is one recent generic and effective framework that bridges any kind of machine learning algorithms and logical reasoning by using inconsistency minimization to construct pseudo-labels of the intermediate symbolic concepts. The inconsistency value is calculated by a designed distance function (Huang et al., 2021b; Cai et al., 2021). Unlike most NeSy algorithms, ABL does not attempt to make symbolic knowledge differentiable. Instead, it fully utilizes the symbolic reasoning capability within the symbolic knowledge by sampling pseudo-labels of intermediate symbols using abductive reasoning. Both methods combine the interpretability of symbolic knowledge with the learning capabilities and flexibility of neural

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China. [2]School of Artificial Intelligence, Nanjing University, China.. Correspondence to: Yu-Feng Li <liyf@lamda.nju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

networks (d'Avila Garcez et al., 2019), resulting in a system that can effectively adapt to new tasks while remaining comprehensible to human users.

However, numerous researchers have highlighted potential issues in current NeSy systems (Marconato et al., 2023a; Li et al., 2023), particularly *reasoning shortcuts*. It means that neural networks may acquire inaccurate semantics (i.e., overfitting to specific assignments of symbolic concepts) due to the absence of grounding labels for intermediate symbolic concepts at the training stage. While this may achieve high performance on the training task, it can compromise the network's capacity for generalization across new tasks and its interpretability. Figure 1 demonstrates an example of reasoning shortcuts. Many efforts (Marconato et al., 2023b; Li et al., 2023; He et al., 2024) have been proposed to mitigate reasoning shortcuts, such as providing a pre-trained model (Zhou & Huang, 2021; Manhaeve et al., 2019), smoothing labels (Li et al., 2020; Müller et al., 2019), and incorporating semi-supervised data (Huang et al., 2020), among others. Nevertheless, theoretical efforts to quantity the effectiveness of these methods remain to be limited.

In this paper, we propose a simple and effective analysis to quantify the harm caused by the reasoning shortcuts and how we can mitigate it theoretically. We first formalize the severity of reasoning shortcuts. Different from Marconato et al. (2023b)'s definition of reasoning shortcuts, wherein the model achieves maximal log-likelihood on the training set but does not match the ground-truth intermediate concept distribution, we introduce the shortcut risk R_s to measure the severity of reasoning shortcuts. Our definition allows for a more granular quantification of the caused by reasoning shortcuts. Furthermore, we present a formalized definition of the complexity of the symbolic knowledge base, denoted as D_{KB}, based on two basic properties of the knowledge base: data dependence and rule dependence. Data dependence implies that the same symbolic knowledge base yields various effects across diverse data distributions, while rule dependence suggests that the knowledge base under different rules (e.g., $\vee$ or $\oplus$) will yield different impacts on the same data distribution. Based on the above two properties, we prove that the R_s is unbounded if there is no assumption for the hypothesis space. Then we find the upper bounds of R_s if the hypothesis space is under the label smooth assumption or the pre-training assumption. The asymptotic rate of the upper bound can be expressed as $\mathcal{O}(\ln(C - D_{KB}) + 1/\sqrt{N} + \gamma)$, where N represents the size of the training dataset and γ is a constant associated with the characteristics of the hypothesis space. This implies that as the complexity of the knowledge base increases and the number of training samples grows, the shortcut risk will decrease. Besides, our analysis indicates that smoothing labels or providing pre-training models can effectively alleviate the reasoning shortcut problem for NeSy algorithms.

Moreover, we analyze the reasoning shortcut problem of the ABL framework. We prove that the shortcut risk of the ABL algorithm, denoted as R_s^{ABL}, is consistently smaller than that of the NeSy algorithm, and if we can construct a reasonable distance function, R_s^{ABL} will have an upper bound of asymptotic rate $\mathcal{O}(\kappa)$ where κ represents the error rate of the distance function. This means that the reasoning shortcuts may be greatly alleviated, showcasing its potential to address the shortcut problem. Empirical studies demonstrate the rationality of our analyses.

We summarize the contributions of our proposed analysis: (i). We first formalize the reasoning shortcut risk and the complexity of the symbolic knowledge base. (ii). We quantitatively analyze three main factors in how typical NeSy algorithms are affected by reasoning shortcuts. (iii). We find that ABL is more robust to reasoning risks and demonstrates its effectiveness through a large range of empirical studies.

2. Problem Setting and Preliminary

2.1. Problem Setting

A neural-symbolic system usually consists of two parts: a concept perception model f, and a symbolic knowledge base KB. The concept perception model is typically implemented with a neural network to characterize the conditional distribution of the intermediate concept $\mathbf{z}$ given the raw input data $\mathbf{x}$. The intermediate symbol $\mathbf{z}$, which takes on a finite number of values, has a precise interpretation comprehensible to humans. A symbolic knowledge base KB represents a set of logical rules provided by experts, which enables the derivation of the final target label $\mathbf{y}$ satisfying that $\mathbf{z}, \text{KB} \models \mathbf{y}$. More formally, we define an *input space* $\mathcal{X}$, a discrete *target space* $\mathcal{Y}$ of size K, and a discrete *symbol space* $\mathcal{Z}$ of size C. The concept perception model can be defined as $f(\mathbf{z}|\mathbf{x})$ corresponding to the conditional distribution of the intermediate concept given the input, which is in a *hypothesis space* $\mathcal{F} \subseteq \mathcal{X} \to \mathcal{Z}$. For the sake of simplification, we denote $f(\mathbf{x})$ as the predicted label and $f(\mathbf{z}|\mathbf{x})$ as the predicted probability of $\mathbf{z}$ given $\mathbf{x}$.

A distribution $\mathcal{S}$ is defined on space $\mathcal{X} \times \mathcal{Y}$. We sample a training dataset $S = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ from this distribution where N is the size of the dataset. Moreover, we define a joint distribution $\mathcal{P}$ on space $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$. The dataset S can also be regarded as a sample drawn from the distribution $\mathcal{P}$, where the variable $\mathbf{z}$ remains unobserved. The whole task is a K-classification task.

2.2. Neural-symbolic Learning

Many researchers (Marconato et al., 2023a; Li et al., 2023) pointed out that the optimization of representative neural-symbolic algorithms such as DeepProblog (Manhaeve et al., 2019), LTN (Badreddine et al., 2022), and Semantic Loss

(Xu et al., 2018) can have a general form, that is given a training dataset S, find $f \in \mathcal{F}$ that minimizing:

$$\hat{\mathcal{L}}_{nesy} = -\frac{1}{N}\sum_{\mathbf{x}_i,\mathbf{y}_i} \ln(\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \models \mathbf{y}_i) \cdot f(\mathbf{z}|\mathbf{x}_i)) \quad (1)$$

Owing to the intractability of the aforementioned loss function, precise optimization is unattainable. Deepproblog (Manhaeve et al., 2019) leverages knowledge compilation while LTN (Badreddine et al., 2022) leverages fuzzy logic to make it differentiable. In light of the fact that the incorrect intermediate concept satisfying the symbolic knowledge may also occupy a term within the loss function, this objective does not sufficiently guarantee the correct prediction of intermediate symbolic concepts, thereby giving rise to the issue of reasoning shortcuts.

2.3. Abductive Learning

Abductive learning (ABL) is a new framework that bridges neural networks and logical reasoning by using inconsistency minimization to construct pseudo-labels of the intermediate symbolic concepts. Concretely, abductive Learning consists of a perception model, denoted as f, and a reasoning model. The perception model serves the purpose of mapping the raw input $\mathbf{x}$ to an intermediate concept $\mathbf{z}$, similar to neural-symbolic algorithms. We denote the prediction of the perception model as $f(\mathbf{x})$. The reasoning model takes $f(\mathbf{x})$ as input and utilizes abductive reasoning to identify the most similar $\overline{\mathbf{z}}$ that satisfies the knowledge base. $\overline{\mathbf{z}}$ is subsequently treated as the pseudo-label for retraining the perception model. The entire process iterates iteratively. The acquisition of $\overline{\mathbf{z}}$ given $(\mathbf{x}, \mathbf{y})$ by the reasoning part can be formalized as the following optimization problem:

$$\begin{aligned} \overline{\mathbf{z}} &= \arg\min_{\mathbf{z}\in\mathcal{Z}} \text{Dis}(\mathbf{z}, f(\mathbf{x}))) \\ \text{s.t.} \quad & \mathbf{z}, \text{KB} \models \mathbf{y} \end{aligned} \quad (2)$$

The function Dis denotes a pre-defined distance metric. And the re-training for the perception model is to minimize such loss: $\hat{\mathcal{L}}_{ABL} = -\frac{1}{N}\sum_{(\mathbf{x}_i,\mathbf{y}_i)} \ln(f(\overline{\mathbf{z}}_i|\mathbf{x}_i))$. To facilitate further analysis, we denote the acquisition of $\overline{\mathbf{z}}$ as sampling from a distribution $\Phi_{(\mathbf{x},\mathbf{y},f)}(\mathbf{z})$ (It depends on both the data and the current perception model and its support set is $\{\mathbf{z}|\mathbf{z}, \text{KB} \models \mathbf{y}\}$). Thus, the loss function of ABL can be rewritten as follows:

$$\hat{\mathcal{L}}_{ABL} = -\frac{1}{N}\sum_{(\mathbf{x}_i,\mathbf{y}_i)} \mathbb{E}_{\overline{\mathbf{z}}_i\sim\Phi_{(\mathbf{x}_i,\mathbf{y}_i,f)}(\mathbf{z})} \ln(f(\overline{\mathbf{z}}_i|\mathbf{x}_i)) \quad (3)$$

Since the optimization goal of ABL is not to directly optimize the correct intermediate concepts, the reasoning shortcut problem also occurs in ABL.

3. Reasoning Shortcuts and Knowledge Base

In this section, we give detailed definitions of both the shortcut risk and the complexity of the knowledge base KB.

3.1. Reasoning Shortcut Risk

Reasoning shortcuts occur when the perception model overfits an erroneous concept given the raw input. Despite the perception model's misconstruction of the concept, the target prediction may still be correct. Consequently, this phenomenon leads to a high prediction accuracy on the training data but fails to generalize to novel, unseen tasks. We find that the occurrence of shortcut problems can be attributed to the disparity between the optimization objective of neural-symbolic algorithms and the objective of directly supervised learning on intermediate concepts. Considering a supervised learning task whose target is to directly learn f given grounding labels of intermediate concepts. The objective of this task is to minimize the cross-entropy loss $\mathcal{L}$ under the joint distribution of $(\mathbf{x}, \mathbf{z}, \mathbf{y})$, i.e.,

$$\mathcal{L} = -\mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}} \ln(f(\mathbf{z}|\mathbf{x})) \quad (4)$$

We believe that if f can minimize $\mathcal{L}$, then there would be no occurrence of shortcut problems. However, in real neural-symbolic tasks, we do not optimize along the same objective but optimize $\hat{\mathcal{L}}_{nesy}$ using a training dataset of limited size. This leads to the emergence of the reasoning shortcut problem. Hence, we define the severity of reasoning shortcuts as the disparity between our desired objectives and the attainable objectives within a finite dataset. Formally, we express this definition as follows:

Definition 3.1 (Shortcut Risk)**.** The shortcut risk R_s is defined as:

$$R_s \triangleq \mathcal{L} - \hat{\mathcal{L}}_{nesy} \quad (5)$$

The shortcut risk represents the severity of the reasoning shortcuts. The larger R_s, the more severe the issue of reasoning shortcuts. We have $\mathbb{E}[R_s] \geq 0$, so we only need to consider the upper bound of R_s.

3.2. Complexity of the Knowledge Base

Based on existing findings (Marconato et al., 2023a), the complexity of the symbolic knowledge base is a key factor that influences the severity of the shortcut risk. The complexity of the symbolic knowledge base represents the strength of its contribution to the overall neural-symbolic system. The more complex symbolic knowledge we have, the less prone to reasoning shortcuts. We observe that two properties highly affect the definition of the complexity of the symbolic knowledge base: data dependence and rule dependence. Below we will provide two intuitive examples to explain how each property affects the complexity of the symbolic knowledge separately.

Example 3.1. *Considering two MINIST-Addition tasks that have different input data. In Task 1, the input data contains one sample:*  $= 0$. *In Task 2, the input data also contains one sample:*  $= 1$. *Task 1 can uniquely determine 0 correctly, but Task 2 tends to present ambiguity in distinguishing between 0 and 1.*

This example demonstrates that with different input data, the contribution of symbolic knowledge to the neural-symbolic system can be different. In the extreme case, if the input samples have nothing to do with the symbolic knowledge, then the knowledge base is not useful for the task, so we should define its complexity as 0. Such analysis indicates that data dependence is an important property for us to formally define symbolic knowledge's complexity.

Example 3.2. *Considering two tasks both given a pair of images* $(\mathbf{x}_1, \mathbf{x}_2)$ *representing* $true$ *or* $false$. *The KB for Task 1 is* $\mathbf{y} = \mathbf{z}_1 \vee \mathbf{z}_2$*; while the KB for Task 2 is* $\mathbf{y} = \mathbf{z}_1 \oplus \mathbf{z}_2$. *When given* $\mathbf{y} = false$, *Task 1 can easily get that* $\mathbf{z}_1 = false$ *and* $\mathbf{z}_2 = false$ *but Task 2 can be confused to determine the value of* $\mathbf{z}_1$ *or* $\mathbf{z}_2$.

This example demonstrates that different rules in the knowledge base can have different contributions to the neural-symbolic system. Such analysis indicates that rule dependence property should also be considered to define the knowledge base's complexity. Based on these two properties, we give a formal definition for the complexity of a symbolic knowledge base.

Definition 3.2 (Complexity of KB)**.** The complexity of KB, denoted as D_{KB}, is defined as:

$$D_{KB} \triangleq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \nvDash \mathbf{y})] \tag{6}$$

D_{KB} is defined as the expected count of instances $\mathbf{z}$ that conflict with the knowledge base. This measure decides upon the distribution of task data and the specific set of rules contained within KB, thereby aligning with the aforementioned two properties. Intuitively, for any $\mathbf{z}$, if the complexity of KB is sufficiently high, it is more likely to encounter contradictions with $\mathbf{z}$, thus resulting in a large D_{KB}.

Considering that only training samples can be obtained in the real training process, we can define the empirical complexity of a symbolic knowledge base.

Definition 3.3 (Empirical Complexity of KB)**.** Empirical Complexity of KB given training dataset S is defined as:

$$\hat{D}_{KB} \triangleq \frac{1}{N} \sum_{(\mathbf{x}_i,\mathbf{y}_i)\in S} \sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \nvDash \mathbf{y}_i)] \tag{7}$$

Lemma 3.4 (Relationship between D_{KB} and $\hat{D}_{KB}$)**.** *According to the Hoeffding's inequality, we can obtain the following conclusion: For* $\forall\delta \in (0, 1)$, *with the probability of at least* $1 - \delta$, *we have*

$$|D_{KB} - \hat{D}_{KB}| \leqslant (C-1)\sqrt{\frac{\ln 2/\delta}{2N}} \tag{8}$$

This lemma demonstrates that as the training dataset is sufficiently large, the difference between D_{KB} and $\hat{D}_{KB}$ is tightly bounded with a large probability. The proof is provided in Appendix C.

4. Analysis of NeSy Reasoning Shortcut Risk

In this section, we give analyses of the upper bounds of R_s. The upper bound of the shortcut risk quantifies the utmost severity of shortcuts that a given algorithm can potentially exhibit in the worst-case scenario. This measure has theoretical significance for the design of neural-symbolic algorithms. We first show that R_s is unbounded when the knowledge base is not complex enough.

Theorem 4.1 (Unbounded property of R_s)**.** *Suppose that the hypothesis space* $\mathcal{F}$ *is comprised of the set of all mappings from* $\mathcal{X}$ *to* $\mathcal{Z}$, *i.e.,* $\mathcal{F} : \mathcal{X} \to \mathcal{Z}$. *When* $D_{KB} < C - 1$, $\exists f \in \mathcal{F}$, *such that* $\mathcal{L}_{nesy} = 0$ *and* $\mathcal{L} \to +\infty$, *so that* $R_s \to +\infty$.

Theorem 4.1 showcases the unbounded nature of R_s when the hypothesis space $\mathcal{F}$ exhibits sufficient complexity but the complexity of the knowledge base is insufficient. In such cases, the learned function f satisfies the knowledge base for all training samples, thus $\hat{\mathcal{L}}_{nesy} = 0$. Simultaneously, it produces erroneous predictions for the intermediate concept $\mathbf{z}$ across all samples, leading to the desired objective loss $\mathcal{L}$ towards infinity. Consequently, the absence of constraints on the hypothesis space $\mathcal{F}$ greatly increases the risk of reasoning shortcuts.

Numerous approaches (Marconato et al., 2023b; Li et al., 2023) have been introduced to alleviate the issue of reasoning shortcuts, including the utilization of pre-trained models, label smoothing, extra-supervised data, and various other strategies. We posit that the aforementioned approaches have the potential to mitigate reasoning shortcuts due to their assumption-making for the hypothesis space $\mathcal{F}$, which will simplify it. Based on existing solutions to mitigate the shortcut problem, we define two assumptions regarding $\mathcal{F}$: the label smoothness assumption and the pre-training assumption.

Definition 4.2 (Label smoothness assumption.)**.** The hypothesis space $\mathcal{F}_\eta$ satisfying label smooth assumption is defined as:

$$\mathcal{F}_\eta = \{f | f \in \mathcal{F} \wedge \forall \mathbf{x} \in \mathcal{X}, \frac{\max_{\mathbf{z}} f(\mathbf{z}|\mathbf{x})}{\min_{\mathbf{z}} f(\mathbf{z}|\mathbf{x})} \leq \eta\} \tag{9}$$

Definition 4.3 (Pre-training assumption.)**.** The hypothesis space $\mathcal{F}_\epsilon$ satisfying pre-training assumption is defined as:

$$\mathcal{F}_\epsilon = \{f | f \in \mathcal{F} \wedge \forall \mathbf{x} \in \mathcal{X}, \forall \mathbf{z} \neq \mathbf{g}, f(\mathbf{z}|\mathbf{x}) \leq \epsilon\}, \quad (10)$$

where $\mathbf{g}$ is the grounding label of $\mathbf{x}$.

The label smoothness assumption requires that the model f should not exhibit excessive confidence in its predictions, meaning that the ratio between the maximum and minimum probabilities of a prediction should not exceed η. This assumption can accurately reflect the goal of the label smoothing method which softens the one-hot labels and has been proved to be an effective way to mitigate reasoning shortcuts. The pre-training assumption states that a model f, following pre-training on annotated data, exhibits a certain level of discernment regarding incorrect labels. In other words, the probability of predicting any erroneous label should not surpass ϵ. This assumption is consistent with existing methodologies that employ pre-trained models or integrate supplementary supervised data as regularization.

Given the definitions of the two assumptions, we proceed to derive upper bounds for R_s under the constrained hypothesis space. Firstly, we prove the upper bound of expected R_s.

Theorem 4.4. *We have the upper bounds of $\mathbb{E}[R_s]$ under two assumptions of the hypothesis space given the data distribution $\mathcal{S}$.*

(i). When $\mathcal{F}_\eta$ satisfies the label smooth assumption.

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[R_s] \leq \frac{1}{2}\ln(C - D_{KB}) + \gamma_\eta, \quad \forall f \in \mathcal{F}_\eta, \quad (11)$$

where γ_η is a constant about $\mathcal{F}_\eta$, $\gamma_\eta = \ln\eta + \frac{1}{2}\ln C$.

(ii). When $\mathcal{F}_\epsilon$ satisfies the pre-training assumption.

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[R_s] \leq \frac{1}{2}\ln(C - D_{KB}) + \gamma_\epsilon, \quad \forall f \in \mathcal{F}_\epsilon, \quad (12)$$

where γ_ϵ is a constant about $\mathcal{F}_\epsilon$, $\gamma_\epsilon = \frac{(C-1)\epsilon^2}{2(C-1)\epsilon^2 - 4(C-1)\epsilon + 2}$.

Theorem 4.4 demonstrates that the upper bound for the expected R_s consists of two components: one is related to the complexity of the knowledge base, and the other is a constant about the characteristics of the hypothesis space. As the upper bound on the knowledge base complexity increases, the shortcut risk in the expected sense diminishes. It is insufficient to solely obtain an upper bound for the expected R_s because, during the actual training process, only $\hat{\mathcal{L}}_{nesy}$ and $\hat{D}_{KB}$ are accessible. Below, we present the upper bounds for R_s.

Theorem 4.5. *Given training dataset S of size N, for any $0 \leq \delta_1 \leq 1$, $0 \leq \delta_2 \leq 1$, with the probability of at least $(1-\delta_1)(1-\delta_2)$:*

(i). When $\mathcal{F}_\eta$ satisfies label smooth assumption.

$$\begin{aligned} R_s \leq \frac{1}{2}\ln\left(C + (C-1)\sqrt{\frac{\ln 2/\delta_1}{2N}} - \hat{D}_{KB}\right) \\ + 3B_\eta\sqrt{\frac{\ln 2/\delta_2}{2N}} + 2\hat{R}_m(\mathcal{F}_\eta) + \gamma_\eta, \quad \forall f \in \mathcal{F}_\eta \end{aligned} \quad (13)$$

where B_η is the bound of $\hat{\mathcal{L}}_{nesy}$, $B_\eta = \ln C + \ln\eta$.

(ii). When $\mathcal{F}_\epsilon$ satisfies pre-training assumption.

$$\begin{aligned} R_s \leq \frac{1}{2}\ln\left(C + (C-1)\sqrt{\frac{\ln 2/\delta_1}{2N}} - \hat{D}_{KB}\right) \\ + 3B_\epsilon\sqrt{\frac{\ln 2/\delta_2}{2N}} + 2\hat{R}_m(\mathcal{F}_\epsilon) + \gamma_\epsilon, \quad \forall f \in \mathcal{F}_\epsilon \end{aligned} \quad (14)$$

where B_ϵ is the bound of $\hat{\mathcal{L}}_{nesy}$, $B_\epsilon = -\ln(1-(C-1)\epsilon)$.

$\hat{R}_m(\mathcal{F})$ represents the empirical Rademacher complexity of $\mathcal{F}$. Theorem 4.5 establishes that, with a high probability, R_s can be bounded by three terms. The first term relates to the complexity of the knowledge base, where a more complex KB leads to a tighter upper bound for R_s. The second term corresponds to the convergence of the number of training samples, following a rate of $\mathcal{O}(1/\sqrt{N})$. The third term comprises the empirical Rademacher complexity of the hypothesis space and a constant factor, which characterizes the properties inherent to the hypothesis space itself. The asymptotic complexity of the upper bound can be generally denoted as $\mathcal{O}(\ln(C - D_{KB}) + 1/\sqrt{N} + \gamma)$. All the proofs are in Appendix C.

5. Analysis of ABL

Similar to the definition of shortcut risk in the domain of neural-symbolic learning, the shortcut risk associated with ABL, denoted as R_s^{ABL}, can be formulated in the following manner:

Definition 5.1 (Shortcut Risk of ABL)**.** The shortcut risk R_s^{ABL} is defined as:

$$R_s^{ABL} \triangleq \mathcal{L} - \hat{\mathcal{L}}_{ABL} \quad (15)$$

Based on the aforementioned formalization of ABL, it can be observed that if the distance metric can be appropriately designed, it can provide more suitable feedback $\overline{\mathbf{z}}$ compared to the ordinary neural-symbolic system to train the concept perception model, thereby enabling it to achieve better training outcomes and mitigate the reasoning shortcuts to some extent. To formally analyze the reasoning shortcuts of ABL, we first contemplate the simplest setting of the distance function, specifically $\text{Dis}(\cdot,\cdot) = 0$ in this paper.

Theorem 5.2. *When $Dis(\cdot,\cdot) = 0$, we have (i): Φ is the uniform distribution on the $\mathbf{z}$ which satisfies the knowledge*

base.

$$\Phi_{(\mathbf{x}_i,\mathbf{y}_i,f)}(\mathbf{z}) = \Phi_{(\mathbf{y}_i)}(\mathbf{z}) \propto \begin{cases} 1 & \mathbf{z}, KB \models \mathbf{y}_i, \\ 0 & \mathbf{z}, KB \not\models \mathbf{y}_i. \end{cases} \tag{16}$$

(ii): R_s^{ABL} has an upper bound.

$$R_s^{ABL} \leq R_s - \frac{\ln C}{C-1}\left(C - \hat{D}_{KB} - 1\right) \tag{17}$$

Theorem 5.2 demonstrates that R_s^{ABL} is upper bounded by R_s. The rationale for this theorem lies in the fact that ABL only assigns a single pseudo-label at a time, whereas the NeSy algorithm integrates all potential pseudo-labels consistent with the knowledge base into the loss function. This causes $\mathcal{L}_{nesy}$ to become smaller, resulting in a larger R_s than R_s^{ABL}. And when the knowledge base's complexity $\hat{D}_{KB}$ is smaller, then R_s^{ABL} will have a tighter bound. As a result, NeSy can be easier to fall into reasoning shortcuts. This finding indicates that ABL exhibits superior efficacy in mitigating the occurrence of reasoning shortcuts when compared to typical neural-symbolic algorithms.

Theorem 5.3. *If the sampling distribution Φ derived from the Dis function satisfies that $\forall \mathbf{x} \in \mathcal{X}, \forall c \neq \mathbf{g}, \Phi_{(\mathbf{x},\mathbf{y},f)}(c) \leq \kappa$ ($\mathbf{g}$ is the grounding label of $\mathbf{x}$).*

(i). When $\mathcal{F}_\eta$ satisfies the label smooth assumption.

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[R_s^{ABL}] \leq \eta^2(C - D_{KB} - 1) \cdot \kappa \tag{18}$$

(ii). When $\mathcal{F}_\epsilon$ satisfies the pre-training assumption.

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[R_s^{ABL}] \leq \frac{\epsilon(C - D_{KB} - 1)}{1 - (C-1)\epsilon} \cdot \kappa \tag{19}$$

Theorem 5.3 demonstrates that when Φ is capable of sampling incorrect intermediate labels with a small error rate κ, the expected shortcut risk of ABL remains small. Furthermore, when the complexity of the knowledge base is fixed, the upper bound on the shortcut risk approaches zero with a rate of $\mathcal{O}(\kappa)$. The result implies that if ABL can select or train a good distance function in some way, then the shortcut risk of ABL will not be significant, highlighting its potential to address the shortcut problem and the advantage of ABL algorithms over typical NeSy algorithms. All the proofs are in Appendix C.

6. Related Work

Neural-symbolic Learning Neural-symbolic learning (Besold et al., 2021; De Raedt et al., 2020) has received considerable attention in recent years. Typical methods (Yang et al., 2022; Xu et al., 2018; Fischer et al., 2019; Huang et al., 2021a) regard logical rules as constraints that act as effective regularization during the training of neural networks. Among them, Huang et al. (2021a) devised a loss function that compels the network's output to adhere more closely to logic constraints. Yang et al. (2022); Fischer et al. (2019) adopt similar strategies to address various types of logical constraints. Furthermore, several techniques (Badreddine et al., 2022; Manhaeve et al., 2019) have specifically emphasized the integration of neural networks with established tools for logical reasoning. Badreddine et al. (2022) extend fuzzy logic with neural predicates to create a fully differentiable logical language known as Real Logic. Similarly, DeepProbLog (Manhaeve et al., 2019) is a probabilistic logic programming language that integrates deep learning through the use of neural predicates. The objective of these approaches is to construct comprehensive logical systems in a differentiable manner. In fact, regardless of the perspective from which neural-symbolic methods are designed, they may encounter shortcut problems.

Abductive Learning Abductive learning (Zhou, 2019) facilitates the simultaneous optimization of machine learning and logical reasoning through the minimization of the inconsistency. The main emphasis of this approach is to deal with intermediate symbolic concepts that serve as pseudo-labels during the learning process and as variables for abductive reasoning. By utilizing pseudo-labels, the machine learning model can be iteratively updated, while abduction searches for the most appropriate revised pseudo-label to minimize any inconsistencies between the raw data and knowledge base. There are various variants of ABL. Cai et al. (2021) extend the ABL framework by leveraging the logical domain knowledge base, which is represented through groundings. Huang et al. (2021b) employs a similarity-based consistency metric to determine the most suitable pseudo-label among all possible abduction results, thereby enhancing the optimization process of the ABL framework in terms of speed and stability. Some researchers have incorporated the ABL framework into weakly supervised scenarios (Shi & Li, 2024; Zhou et al., 2024), including semi-supervised learning (Huang et al., 2020) and transfer learning (Zhou et al., 2023). The shortcut problem similarly occurs for ABL algorithms. Yang et al. (2024) consider when the knowledge base contains inaccurate rules, the severity of reasoning shortcuts increases. In this paper, we show that by designing a suitable distance function, ABL can have lower shortcut risk, which demonstrates its potential to mitigate reasoning shortcuts.

Reasoning Shortcuts The issue of shortcuts is a significant and challenging problem in the field of neural-symbolic, and only a few articles have paid attention to this problem. Stammer et al. (2022) investigate the approach of injecting additional knowledge into the model to address input-level shortcuts. In the context of NeSy, Marconato et al. (2023a)

introduce the concept of the reasoning shortcut and proposes a method that combines concept supervision and concept-level rehearsal to address shortcut problems in the context of continual learning. Li et al. (2023) propose a minimax objective that ensures the concepts learned by the model satisfy the knowledge base and have less shortcuts. However, all these methods lack corresponding in-depth theoretical analysis results. Moreover, the work proposed by Marconato et al. (2023b) theoretically analyzes the generality of the reasoning shortcut problem and extracts the key factors that may have an impact on it, thereby proposing several mitigation strategies. One limitation of the work is that the characterization of the complexity of the knowledge base may not be sufficiently accurate, which could potentially affect the thoroughness of the analysis regarding the impact of the knowledge base on reasoning shortcuts.

7. Empirical Study

In this section, we empirically corroborate two principal findings that have previously been supported by theoretical evidence: (1). As the complexity of the knowledge base increases, both the neural-symbolic approaches and the ABL algorithms exhibit lower shortcut risk. (2). The selection of the distance function Dis influences the performance of the ABL algorithm significantly, where a good choice of Dis can assist in alleviating the reasoning shortcuts.

7.1. MNIST-Addition

Task Description We consider the MNIST-Addition experiment (Manhaeve et al., 2019). The input of this task is a pair of MNIST images (LeCun et al., 1998), and the output is the sum of the individual digits. The knowledge base given by this task is the addition rule. We construct six datasets of different levels from the original task. We first split target labels into an 'easy' pool and a 'hard' pool. Then we sample input data from these two distinct pools with different ratios. Therefore, the constructed datasets will have different levels of difficulty. Correspondingly, descending through the levels precipitates an escalation in the complexity of the empirical knowledge base. To ensure equity for all the experiments, each dataset has a fixed training sample size of 30,000 and the test dataset remains unchanged. Detailed information is provided in Appendix B.1.

Competing baselines We do experiments on three representative neural-symbolic learning algorithms: semantic loss (SL) (Xu et al., 2018), DeepProbLog (DPL) (Manhaeve et al., 2019) and LTN (Badreddine et al., 2022). For ABL algorithms, we instantiate the trivial distance function $\mathrm{Dis}(\cdot,\cdot) = 0$ as the baseline, denoted as ABL. In order to verify whether the provided two assumptions on the hypothesis space are effective for abductive learning, we additionally compare the two methods ABL+L and ABL+P, which respectively represent the use of label smoothing and a pre-trained model.

Distance functions To verify the impact of the selection of the distance function Dis on the performance of ABL, we additionally select several non-trivial distance functions. Firstly, we consider Hamming distance which calculates the number of positions at which the corresponding symbols are different of two strings. In the context of the MNIST-Addition task, the digit pair $(1, 2)$ exhibits a distance of one from $(1, 1)$ and a distance of two from $(3, 4)$. We denote the ABL algorithm with such distance function as ABL(H). Secondly, we follow Huang et al. (2021b) which adopts the metric $\mathrm{Dis}(\mathbf{z}, f(\mathbf{x})) = 1 - f(\mathbf{z}|\mathbf{x})$, signifying the utilization of the contemporary model's confidence to construct the sampler Φ. We denote this method of such distance function within ABL algorithms as ABL(C). Moreover, we consider a pre-training model g, and the metric $\mathrm{Dis}(\mathbf{z}, f(\mathbf{x})) = 1 - g(\mathbf{z}|\mathbf{x})$. We denote it as ABL(P). Different from ABL(C), the extra model g is not updated during training.

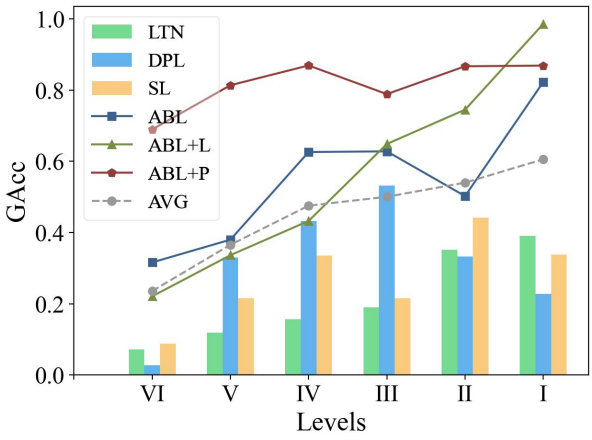

Figure 2. The grid accuracy (GAcc) of different methods under different levels of MNIST-Addition task.

Experimental results We report the results of typical NeSy and ABL algorithms in Figure 2 whey varying the difficulty level of the training dataset. Given that the shortcut risk R_s is not amenable to direct computation, we deploy the metric of Grid accuracy (GAcc) to evaluate the model's degree of mitigation on reasoning shortcuts. Grid accuracy represents the prediction accuracy of the model for the probability of intermediate symbols, specifically, in the MINST-Addition experiment, it is the classification accuracy of MINST images. As the neuro-symbolic system primarily aims for the precise classification of intermediate concepts, it follows logically that GAcc should be considered a reasonable metric for evaluation purposes. The results indicate

Table 1. Grid accuracy (GAcc) (%mean ± std) on different levels of MNIST-Addition tasks.

Method (**GAcc**)	MNIST-Addition Datasets of Different Levels						
	I	**II**	**III**	**IV**	**V**	**VI**	**AVG**
SL	$33.81_{\pm 25.30}$	$44.19_{\pm 20.34}$	$21.56_{\pm 23.61}$	$33.60_{\pm 25.50}$	$21.53_{\pm 23.57}$	$8.85_{\pm 3.57}$	$27.26_{\pm 20.32}$
LTN	$39.09_{\pm 16.70}$	$35.17_{\pm 23.56}$	$18.97_{\pm 18.07}$	$15.63_{\pm 19.49}$	$11.87_{\pm 11.62}$	$7.24_{\pm 8.34}$	$21.33_{\pm 16.30}$
DPL	$22.85_{\pm 24.90}$	$33.27_{\pm 25.03}$	$53.22_{\pm 0.25}$	$43.25_{\pm 20.04}$	$34.97_{\pm 22.93}$	$2.75_{\pm 1.01}$	$31.72_{\pm 15.69}$
ABL	$82.24_{\pm 32.02}$	$50.23_{\pm 39.28}$	$62.80_{\pm 43.25}$	$\underline{62.60_{\pm 43.46}}$	$34.73_{\pm 34.59}$	$\underline{31.62_{\pm 34.03}}$	$54.04_{\pm 37.77}$
ABL(H)	$80.64_{\pm 35.41}$	$64.50_{\pm 41.67}$	$62.84_{\pm 43.29}$	$46.69_{\pm 42.22}$	$\underline{36.59_{\pm 33.54}}$	$19.48_{\pm 14.13}$	$51.79_{\pm 35.04}$
ABL(C)	$\underline{89.61_{\pm 13.27}}$	$\underline{88.44_{\pm 20.31}}$	$\underline{82.05_{\pm 12.69}}$	$53.25_{\pm 43.83}$	$30.45_{\pm 25.31}$	$4.04_{\pm 7.71}$	$\underline{57.97_{\pm 20.52}}$
ABL(P)	$\mathbf{98.63_{\pm 0.08}}$	$\mathbf{98.60_{\pm 0.11}}$	$\mathbf{98.55_{\pm 0.10}}$	$\mathbf{98.53_{\pm 0.08}}$	$\mathbf{98.54_{\pm 0.07}}$	$\mathbf{98.57_{\pm 0.13}}$	$\mathbf{98.57_{\pm 0.10}}$

Table 2. GAcc (%) and Acc (%) on the BDD-OIA task.

Method	**GAcc**	**Acc**
DPL	57.74 ± 6.70	69.00 ± 0.74
ABL	76.96 ± 0.04	75.17 ± 0.18
ABL(H)	76.94 ± 0.07	$\mathbf{75.31 \pm 0.21}$
ABL(C)	78.54 ± 1.43	75.30 ± 0.12
ABL(P)	$\mathbf{85.13 \pm 0.15}$	74.03 ± 0.10

that as the level decreases (i.e., $\hat{D}_{KB}$ increases), most methods show an upward trend in predictive performance for intermediate concepts. This demonstrates that the shortcut problem is mitigated as the knowledge base becomes more complex. We find that the ABL algorithm performs better when using label smoothing, especially when the knowledge base complexity is high. Additionally, we observe that the ABL algorithm has a higher low bound performance and overall performance after incorporating pre-training models. It is worth noting that the DPL algorithm experiences a performance decline at levels **I** and **II**, which we attribute to the instability of the DPL algorithm. Overall, the experimental results can be effectively regarded as supporting evidence for our Theorem 4.4 and Theorem 4.5.

Table 1 shows the grid accuracy of the ABL algorithms with different distance functions. **Bold** and underline indicate the optimal and sub-optimal performance, respectively. The results indicate that using a distance function constructed from a pre-trained model can correctly identify labels of the correct intermediate concepts with a high probability by minimizing inconsistency. As a result, ABL(P) achieves high performance and largely mitigates the reasoning shortcut problem, which experimentally supports the result of Theorem 5.3. Additionally, we observe that ABL(C) exhibits a greater shortcut risk on higher-level datasets, with only 4.04% GAcc at Level **VI**. However, it performs better on lower-level tasks. We attribute this to the heavy reliance on the distance function, which uses the confidence of the current model. When the complexity of the knowledge base is not high enough, this distance measure is more prone to obtaining incorrect pseudo-labels, thereby deepening reasoning shortcuts. Overall, our experiments demonstrate that the selection of Dis significantly influences the performance of the ABL algorithms and a good choice of Dis can assist in alleviating reasoning shortcuts. Our experimental details and additional results are provided in Appendix B.1.

7.2. BDD-OIA

Task description We also conducted experiments on an autonomous driving task. BDD-OIA (Xu et al., 2020) is a commonly used dataset, which predicts the current feasible actions ($\mathcal{Y} = \{$`move_forward`, `stop`, `turn_left`, `turn_right`$\}$). (Marconato et al., 2023a) has demonstrated by experiments that reasoning shortcuts exist in this task, however, the shortcut risk reduction of ABL algorithms has not been well studied. There are 21 symbolic concepts for this task, such as whether there is a vehicle ahead (`vehicle_ahead`), and rules in the knowledge base, such as `vehicle_ahead` $\rightarrow \neg$`move_forward`. Similar to the setting of the MNIST-Addition task, we compare the performances of ABL algorithms under different distance functions. In addition, we compare with the typical method DPL as a baseline. More detailed information about this dataset and experimental details are provided in Appendix B.2.

Experimental results The experimental results are shown in Table 2. As the task involves multi-label classification, GAcc and Acc represent the average accuracy on intermediate concepts and targets, respectively. The results demonstrate that ABL(P) achieves the optimal GAcc at a slight cost to the target Acc, indicating its minimal harm from reasoning shortcuts. Furthermore, our approach utilizing the model's confidence, ABL(C), outperforms the trivial distance method (ABL) in terms of performance. Overall, this experiment demonstrates a significant improvement over the baseline by leveraging confidence and pre-trained models as

the distance functions, inspiring us to consider incorporating machine learning into the design of Dis for future ABL algorithms. Additionally, we observed that ABL(H) does not exhibit a noticeable improvement compared to ABL. We attribute this to the fact that the Hamming distance does not align well with the structure of this task, resulting in negligible differences from the trivial distance method.

8. Conclusion

In this paper, we focus on providing a rigorous theoretical analysis of the reasoning shortcut within the framework of neural-symbolic learning and abductive learning. We first formalize the definition of shortcut risk and the complexity of the knowledge base. Secondly, we prove the upper bound of the shortcut risk under two assumptions of the hypothesis space. Finally, we theoretically show the potential of ABL in reducing the risk of reasoning shortcuts. Empirical studies support the above theoretical results.

In the future, our efforts will be focused on expanding the complexity of the knowledge base to a higher-order form, with the aim of deepening the understanding of reasoning shortcuts. Furthermore, we intend to investigate the influence of sample distribution on reasoning shortcuts.

Acknowledgements

This research was supported by Leading-edge Technology Program of Jiangsu Science Foundation (BK20232003) and the National Science Foundation of China (62176118).

Impact Statement

This paper aims at advancing the field of neural-symbolic integration. Our work includes experiments related to autonomous driving, which can raise ethical considerations and societal implications. While our primary focus is within the domain of Machine Learning, we acknowledge the broader implications on transportation infrastructure, job markets, and privacy. We emphasize the importance of continued research and discussions to tackle challenges in autonomous driving by NeSy algorithms and ensure the responsible deployment of autonomous driving systems.

References

Badreddine, S., d'Avila Garcez, A. S., Serafini, L., and Spranger, M. Logic tensor networks. *Artif. Intell.*, pp. 103649, 2022.

Besold, T. R., d'Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P. M., Hitzler, P., Kühnberger, K., Lamb, L. C., Lima, P. M. V., de Penning, L., Pinkas, G., Poon, H., and Zaverucha, G. Neural-symbolic learning and reasoning: A survey and interpretation. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pp. 1–51. 2021.

Cai, L.-W., Dai, W.-Z., Huang, Y.-X., Li, Y.-F., Muggleton, S. H., and Jiang, Y. Abductive learning with ground knowledge base. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1815–1821, 2021.

Cunnington, D., Law, M., Lobo, J., and Russo, A. Inductive learning of complex knowledge from raw data. *CoRR*, 2022.

d'Avila Garcez, A. S., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., and Tran, S. N. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, pp. 611–632, 2019.

De Raedt, L., Dumancic, S., Manhaeve, R., and Marra, G. From statistical relational to neuro-symbolic artificial intelligence. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4943–4950, 2020.

Fischer, M., Balunovic, M., Drachsler-Cohen, D., Gehr, T., Zhang, C., and Vechev, M. T. DL2: training and querying neural networks with logic. In *Proceedings of the International Conference on Machine Learning*, pp. 1931–1941, 2019.

He, H.-Y., Dai, W.-Z., and Li, M. Reduced implication-bias logic loss for neuro-symbolic learning. *Machine Learning*, 113:3357–3377, 2024.

Huang, J., Li, Z., Chen, B., Samel, K., Naik, M., Song, L., and Si, X. Scallop: From probabilistic deductive databases to scalable differentiable reasoning. In *Advances in Neural Information Processing Systems*, pp. 25134–25145, 2021a.

Huang, Y.-X., Dai, W.-Z., Yang, J., Cai, L.-W., Cheng, S., Huang, R., Li, Y.-F., and Zhou, Z.-H. Semi-supervised abductive learning and its application to theft judicial sentencing. In *Proceedings of the International Conference on Data Mining*, pp. 1070–1075, 2020.

Huang, Y.-X., Dai, W.-Z., Cai, L.-W., Muggleton, S. H., and Jiang, Y. Fast abductive learning by similarity-based consistency optimization. In *Advances in Neural Information Processing Systems*, pp. 26574–26584, 2021b.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pp. 2278–2324, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. E. Deep learning. *Nature*, pp. 436–444, 2015.

Li, Q., Huang, S., Hong, Y., Chen, Y., Wu, Y. N., and Zhu, S. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning. In *Proceedings of the International Conference on Machine Learning*, pp. 5884–5894, 2020.

Li, Z., Liu, Z., Yao, Y., Xu, J., Chen, T., Ma, X., and Lü, J. Learning with logical constraints but without shortcut satisfaction. In *Proceedings of the International Conference on Learning Representations*, 2023.

Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., and Raedt, L. D. Deepproblog: Neural probabilistic logic programming. In *Proceedings of the Benelux Conference on Artificial Intelligence*, pp. 3753–3763, 2019.

Marconato, E., Bontempo, G., Ficarra, E., Calderara, S., Passerini, A., and Teso, S. Neuro-symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal. In *Proceedings of the International Conference on Machine Learning*, pp. 23915–23936, 2023a.

Marconato, E., Teso, S., Vergari, A., and Passerini, A. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. *CoRR*, 2023b.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pp. 4696–4705, 2019.

Sarker, M. K., Zhou, L., Eberhart, A., and Hitzler, P. Neuro-symbolic artificial intelligence: Current trends. *CoRR*, 2021.

Sawada, Y. and Nakamura, K. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, pp. 41758–41765, 2022.

Shi, J.-X., W. T. and Li, Y.-F. Residual diverse ensemble for long-tailed multi-label text classification. *SCIENCE CHINA Information Sciences*, 2024.

Stammer, W., Memmel, M., Schramowski, P., and Kersting, K. Interactive disentanglement: Learning concepts by interacting with their prototype representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10307–10318, 2022.

Xu, J., Zhang, Z., Friedman, T., Liang, Y., and den Broeck, G. V. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the International Conference on Machine Learning*, pp. 5498–5507, 2018.

Xu, Y., Yang, X., Gong, L., Lin, H.-C., Wu, T.-Y., Li, Y., and Vasconcelos, N. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yang, X.-W., Shao, J.-J., Tu, W.-W., Li, Y.-F., Dai, W.-Z., and Zhou, Z.-H. Safe abductive learning in the presence of inaccurate rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 16361–16369, 2024.

Yang, Z., Lee, J., and Park, C. Injecting logical constraints into neural networks via straight-through estimators. In *Proceedings of International Conference on Machine Learning*, pp. 25096–25122, 2022.

Zhou, Z., Guo, L.-Z., Jia, L.-H., Zhang, D., and Li, Y.-F. Ods: Test-time adaptation in the presence of open-world data shift. In *Proceedings of the International Conference on Machine Learning*, pp. 42574–42588, 2023.

Zhou, Z., Yang, M., Shi, J.-X., Guo, L.-Z., and Li, Y.-F. Decoop: Robust prompt tuning with out-of-distribution detection. In *Proceedings of the International Conference on Machine Learning*, 2024.

Zhou, Z.-H. Abductive learning: towards bridging machine learning and logical reasoning. *SCIENCE CHINA Information Sciences*, pp. 76101:1–76101:3, 2019.

Zhou, Z.-H. and Huang, Y.-X. Abductive learning. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pp. 353–369. 2021.

A. Limitations

Here are some limitations of our analysis. Firstly, the current definition of the complexity of the knowledge base is in the expectation form which does not consider the variance of the data distribution. Secondly, we only focus on the general form of NeSy algorithms but need further analysis for the shortcut risk of specific algorithms.

B. Experimental Details and Additional Results

B.1. MNIST-Addition Tasks

$\hat{D}_{KB}$ of the datasets We report the value of $C - \hat{D}_{KB}$ for each level dataset in Table 3. As the level increases, the complexity of the knowledge base decreases.

Table 3. The value of $C - \hat{D}_{KB}$ for the datasets of different levels.

Levels	**I**	**II**	**III**	**IV**	**V**	**VI**
$C - \hat{D}_{kB}$	7.40	7.68	7.96	8.24	8.52	8.80

Training details For all our experiments, we use LeNet-5(LeCun et al., 1998) as the perception model. For ABL algorithms, we use the Adam optimizer (Kingma & Ba, 2015) with the learning rate of $3e-4$ to train our networks. Both methods ABL+P and ABL(P) use a pre-trained network that can achieve 38.24% accuracy on the MNIST test dataset. All our experiments are implemented by Pytorch and are conducted on an NVIDIA A800. We repeated each experiment five times.

Additional results We also report the target accuracy on different levels of MNIST-Addition tasks. Results are shown in Table 4.

Table 4. Final prediction accuracy (Acc) (%mean $\pm$ std) on different levels of MNIST-Addition tasks.

Method (**Acc**)	MNIST-Addition Datasets of Different Levels						
	I	**II**	**III**	**IV**	**V**	**VI**	**AVG**
SL	$78.48_{\pm 24.19}$	$\underline{88.01_{\pm 19.55}}$	$63.65_{\pm 19.45}$	$78.54_{\pm 23.91}$	$63.11_{\pm 19.62}$	$\underline{39.35_{\pm 8.19}}$	$68.52_{\pm 19.15}$
LTN	$64.78_{\pm 35.55}$	$69.48_{\pm 37.34}$	$29.22_{\pm 34.69}$	$27.89_{\pm 35.10}$	$19.84_{\pm 13.09}$	$30.96_{\pm 5.07}$	$40.36_{\pm 26.80}$
DPL	$66.16_{\pm 22.44}$	$76.22_{\pm 22.93}$	$\underline{93.87_{\pm 1.00}}$	$\underline{83.16_{\pm 21.44}}$	$\underline{73.12_{\pm 26.41}}$	$25.36_{\pm 5.55}$	$\underline{69.65_{\pm 16.63}}$
ABL	$79.46_{\pm 34.20}$	$45.50_{\pm 41.79}$	$61.20_{\pm 42.96}$	$61.31_{\pm 42.75}$	$29.32_{\pm 33.91}$	$25.20_{\pm 35.73}$	$50.33_{\pm 38.56}$
ABL(H)	$79.10_{\pm 35.26}$	$62.20_{\pm 42.48}$	$61.28_{\pm 43.03}$	$49.15_{\pm 39.82}$	$31.31_{\pm 33.88}$	$12.15_{\pm 11.38}$	$49.20_{\pm 34.31}$
ABL(C)	$\underline{83.22_{\pm 19.76}}$	$84.95_{\pm 24.52}$	$70.80_{\pm 18.47}$	$57.52_{\pm 13.34}$	$33.52_{\pm 11.70}$	$33.20_{\pm 9.35}$	$60.54_{\pm 16.19}$
ABL(P)	$\mathbf{97.27_{\pm 0.15}}$	$\mathbf{97.20_{\pm 0.22}}$	$\mathbf{97.12_{\pm 0.22}}$	$\mathbf{97.08_{\pm 0.15}}$	$\mathbf{97.10_{\pm 0.12}}$	$\mathbf{97.19_{\pm 0.26}}$	$\mathbf{97.16_{\pm 0.19}}$

B.2. BDD-OIA Task

Dataset details This dataset comprises frames extracted from driving scene videos, which are utilized for autonomous predictions (Xu et al., 2020). The objective is to predict four actions for each frame, namely $\mathcal{Y} = (\texttt{move\_forward}, \texttt{stop}, \texttt{turn\_left}, \texttt{turn\_right})$. Each frame is annotated with 21 intermediate concepts denoted as $\mathbf{z}$, and further details can be found in Table 5. The training set consists of 16,000 frames, while the test set contains 4,500 annotated data points. Figure B.2 presents some examples from this dataset. Before usage, the dataset was pre-processed by Marconato et al. (2023b) using a pretrained Faster-RCNN model on BDD-100k, in conjunction with the first module in CBM-AUC (Sawada & Nakamura, 2022), resulting in embeddings of dimension 2048.

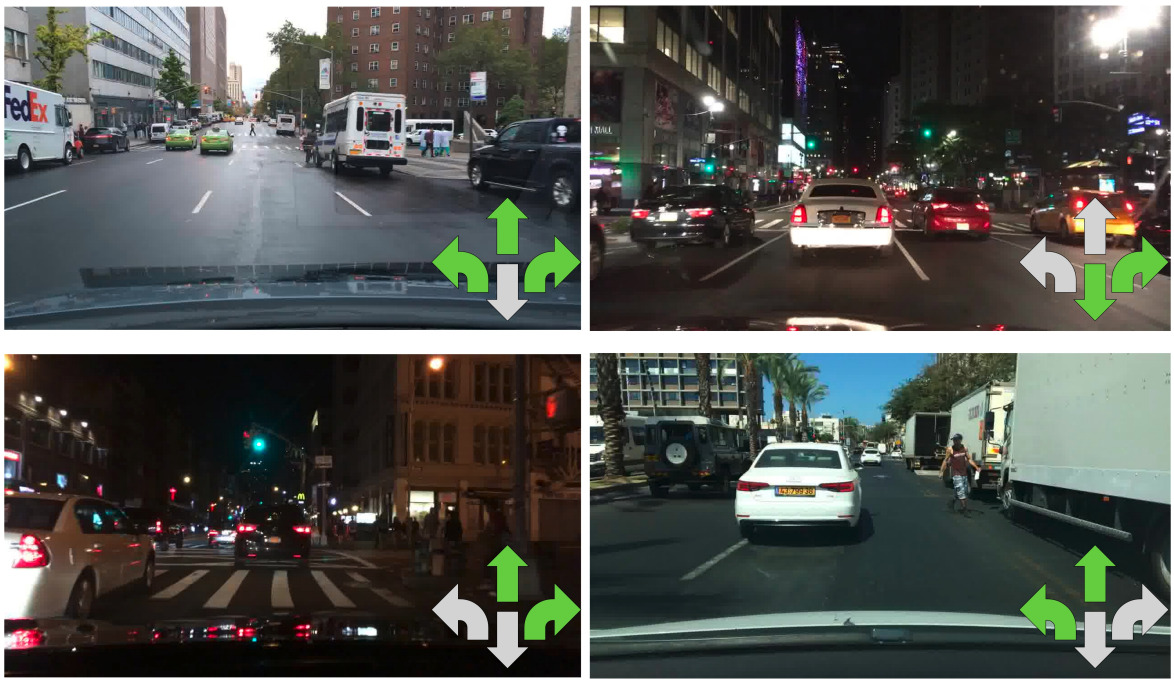

Figure 3. Examples of BDD-OIA. Figure is taken from (Xu et al., 2020).

Same as (Marconato et al., 2023b), the rules of the knowledge base are as follows:

For $\texttt{move\_forward}/\texttt{stop}$:

$$\begin{cases} \texttt{red\_light} \rightarrow \neg\texttt{green\_light} \\ \texttt{obstacle} = \texttt{car} \vee \texttt{person} \vee \texttt{rider} \vee \texttt{other\_obstacle} \\ \texttt{road\_clear} \leftrightarrow \neg\texttt{obstacle} \\ \texttt{green\_light} \vee \texttt{follow} \vee \texttt{clear} \rightarrow \texttt{move\_forward} \\ \texttt{red\_light} \vee \texttt{stop\_sign} \vee \texttt{obstacle} \rightarrow \texttt{stop} \\ \texttt{stop} \rightarrow \neg\texttt{move\_forward} \end{cases}$$

For $\texttt{turn\_left}$, and similarly for $\texttt{turn\_right}$, we have:

$$\begin{cases} \texttt{can\_turn} = \texttt{left\_lane} \vee \texttt{left\_green\_light} \vee \texttt{left\_follow} \\ \texttt{cannot\_turn} = \texttt{no\_left\_lane} \vee \texttt{left\_obstacle} \vee \texttt{left\_solid\_line} \\ \texttt{can\_turn} \wedge \neg\texttt{cannot\_turn} \rightarrow \texttt{turn\_left} \end{cases}$$

Experimental details For all our experiments, we use a linear layer as the perception model given the embeddings of 2048 dimensions. We use the Adam optimizer (Kingma & Ba, 2015) with the learning rate of $5e-3$ to train our networks. The loss function to train the ABL algorithms is BCELoss. All our experiments are implemented by Pytorch and are conducted on an NVIDIA A800. We repeated each experiment five times.

Table 5. Concepts annotated in BDD-OIA. Table taken from (Xu et al., 2020)

Action Category	Concepts	Count
move_forward	green_light	7805
	follow	3489
	road_clear	4838
stop	red_light	5381
	traffic_sign	1539
	car	233
	person	163
	rider	5255
	other_obstacle	455
turn_left	left_lane	154
	left_green_light	885
	left_follow	365
	no_left_lane	150
	left_obstacle	666
	letf_solid_line	316
turn_right	right_lane	6081
	right_green_light	4022
	right_follow	2161
	no_right_lane	4503
	right_obstacle	4514
	right_solid_line	3660

C. Theorem Proof

C.1. Proof of Lemma 3.4

According to the Hoeffding's inequality, we can obtain that $\forall \epsilon > 0$

$$P(|D_{KB} - \hat{D}_{KB}| \leq \epsilon) \geq 1 - 2e^{\frac{-2N^2\epsilon^2}{N(C-1)^2}} \tag{20}$$

By substituting $\delta = 2e^{\frac{-2N^2\epsilon^2}{N(C-1)^2}}$, we can subsequently obtain

$$\epsilon = \sqrt{\frac{(C-1)^2 \ln(2/\delta)}{2N}} = (C-1)\sqrt{\frac{\ln(2/\delta)}{2N}} \tag{21}$$

Therefore, we can conclude that for $\forall \delta \in (0, 1)$, with the probability of at least $1 - \delta$:

$$|D_{KB} - \hat{D}_{KB}| \leqslant (C-1)\sqrt{\frac{\ln 2/\delta}{2N}} \tag{22}$$

C.2. Proof of Theorem 4.1

Since $D_{KB} < C - 1$ and the hypothesis space is complex enough, it is always possible to find a model f that predicts wrong intermediate concepts on all inputs, but the wrong concepts conform to the knowledge base. At this time there are $\hat{\mathcal{L}}_{nesy} = 0$ and $\mathcal{L} = +\infty$. This leads to the unbounded of R_s.

C.3. Proof of Theorem 4.4: the Upper Bounds of Expected R_s

Lemma 2. For $\forall f \in \mathcal{F}$, we have that

$$\mathcal{L}_{nesy} \geq -\frac{1}{2}\ln\left(C - D_{KB}\right) - \frac{1}{2}\ln C - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln \max_{\mathbf{z}} f(\mathbf{z}|\mathbf{x})\right)$$

Proof.

$$\begin{aligned}
\mathcal{L}_{nesy} &= -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln\left(\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \models \mathbf{y}) \cdot f(\mathbf{z}|\mathbf{x})\right)\right) \\
&\geq -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln \sqrt{\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}^2(\mathbf{z}, \text{KB} \models \mathbf{y})}\sqrt{\sum_{\mathbf{z}\in\mathcal{Z}} f^2(\mathbf{z}|\mathbf{x})}\right) \\
&= -\frac{1}{2}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln \sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \models \mathbf{y})\right) - \frac{1}{2}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln \sum_{\mathbf{z}\in\mathcal{Z}} f^2(\mathbf{z}|\mathbf{x})\right) \\
&\geq -\frac{1}{2}\ln \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \models \mathbf{y})\right) - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln \sqrt{\sum_{\mathbf{z}\in\mathcal{Z}} f^2(\mathbf{z}|\mathbf{x})}\right) \\
&= -\frac{1}{2}\ln\left(C - D_{KB}\right) - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln \sqrt{\sum_{\mathbf{z}\in\mathcal{Z}} f^2(\mathbf{z}|\mathbf{x})}\right) \\
&\geq -\frac{1}{2}\ln\left(C - D_{KB}\right) - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln\left(\sqrt{C} \cdot \max_{\mathbf{z}} f(\mathbf{z}|\mathbf{x})\right)\right) \\
&= -\frac{1}{2}\ln\left(C - D_{KB}\right) - \frac{1}{2}\ln C - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln\left(\max_{\mathbf{z}} f(\mathbf{z}|\mathbf{x})\right)\right)
\end{aligned}$$

C.3.1. PROOF OF THEOREM 4.4.(I)

Based on the label smoothness assumption, we have that $\max_{\mathbf{z}} f(\mathbf{z}|\mathbf{x}) \leq \eta \min_{\mathbf{z}} f(\mathbf{z}|\mathbf{x}) \leq \eta f(\mathbf{g}|\mathbf{x})$ when$f \in \mathcal{F}_\eta$. Combining Lemma 2,we can obtain that

$$\begin{aligned}\mathcal{L}_{nesy} &\geq -\frac{1}{2}\ln\left(C - D_{KB}\right) - \frac{1}{2}\ln C - \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\ln(\eta f(\mathbf{z}|\mathbf{x})) \\ &= -\frac{1}{2}\ln\left(C - D_{KB}\right) - \frac{1}{2}\ln C - \ln\eta + \mathcal{L}\end{aligned}$$

Thus, it is evident that

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[R_s] \leq \frac{1}{2}\ln(C - D_{KB}) + \gamma_\eta, \quad \forall f \in \mathcal{F}_\eta,$$

where $\gamma_\eta = \ln\eta + \frac{1}{2}\ln C$.

C.3.2. PROOF OF THEOREM 4.4.(II)

According to the proof of the Lemma 2, we get that

$$\mathcal{L}_{nesy} \geq -\frac{1}{2}\ln\left(C - D_{KB}\right) - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\ln\sqrt{\sum_{\mathbf{z}\in\mathcal{Z}} f^2(\mathbf{z}|\mathbf{x})}\right) \tag{23}$$

Based on the pre-training assumption, we show that

$$f(\mathbf{g}|\mathbf{x}) \geq 1 - (C-1)\epsilon, \quad \forall f \in \mathcal{F}_\epsilon$$

Thus, we infer that

$$\begin{aligned}\ln\sqrt{\sum_{\mathbf{z}\in\mathcal{Z}} f^2(\mathbf{z}|\mathbf{x})} &\leq \ln\sqrt{f^2(\mathbf{g}|\mathbf{x}) + (C-1)\epsilon^2} \\ &= \ln\left(\frac{\sqrt{f^2(\mathbf{g}|\mathbf{x}) + (C-1)\epsilon^2}}{f(\mathbf{g}|\mathbf{x})}\right) + \ln f(\mathbf{g}|\mathbf{x}) \\ &= \ln\left(\sqrt{1 + \frac{(C-1)\epsilon^2}{f^2(\mathbf{g}|\mathbf{x})}}\right) + \ln f(\mathbf{g}|\mathbf{x}) \\ &= \frac{1}{2}\ln\left(1 + \frac{(C-1)\epsilon^2}{f^2(\mathbf{g}|\mathbf{x})}\right) + \ln f(\mathbf{g}|\mathbf{x}) \\ &\leq \frac{(C-1)\epsilon^2}{2f^2(\mathbf{g}|\mathbf{x})} + \ln f(\mathbf{g}|\mathbf{x}) \\ &\leq \frac{(C-1)\epsilon^2}{2(C-1)\epsilon^2 - 4(C-1)\epsilon + 2} + \ln f(\mathbf{g}|\mathbf{x})\end{aligned}$$

Following the Eq(23), we obtain that

$$\mathcal{L}_{nesy} \geq -\frac{1}{2}\ln\left(C - D_{KB}\right) - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}\left(\frac{(C-1)\epsilon^2}{2(C-1)\epsilon^2 - 4(C-1)\epsilon + 2} + \ln f(\mathbf{g}|\mathbf{x})\right)$$

By decomposing the expectation simplistically, we can draw the conclusion that

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[R_s] \leq \frac{1}{2}\ln(C - D_{KB}) + \gamma_\epsilon, \quad \forall f \in \mathcal{F}_\epsilon,$$

where $\gamma_\epsilon = \frac{(C-1)\epsilon^2}{2(C-1)\epsilon^2 - 4(C-1)\epsilon + 2}$.

C.4. Proof of Theorem 4.5: the Upper Bounds of R_s

Lemma 3 $\forall f \in \mathcal{F}_\eta$, $\mathcal{L}_{nesy}$ is bounded.

Proof. It is evident that $\mathcal{L}_{nesy} \geq 0$. Therefore, it suffices to demonstrate that $\mathcal{L}_{nesy}$ has an upper bound.

Based on the label smoothness assumption, we can easily infer that

$$\forall \mathbf{x}, \mathbf{z} \quad f(\mathbf{z}|\mathbf{x}) \geq \min_{\mathbf{z}} f(\mathbf{z}|\mathbf{x}) \geq \frac{1}{C\eta}$$

Therefore we have that, $\forall \mathbf{x}$,

$$\begin{aligned}
&-\ln\left(\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \models \mathbf{y}) \cdot f(\mathbf{z}|\mathbf{x})\right) \\
&\leq -\ln\left(\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \models \mathbf{y}) \cdot \frac{1}{C\eta}\right) \\
&\leq -\ln\left(\frac{1}{C\eta}\right) \\
&= \ln C + \ln \eta
\end{aligned}$$

In other words, $\mathcal{L}_{nesy}$ has an upper bound $B_\eta = \ln C + \ln \eta$.

C.4.1. Proof of Theorem 4.5.(i)

Owing to the result of the empirical Rademacher complexity, we can obtain that for any $0 \leq \delta_1 \leq 1$, with the probability of at least $1 - \delta_1$:

$$|\mathcal{L}_{nesy} - \hat{\mathcal{L}}_{nesy}| \leq 2\hat{R}_m(F_\eta) + 3B_\eta\sqrt{\frac{\ln\frac{2}{\delta_1}}{2N}}, \quad \forall f \in \mathcal{F}_\eta$$

Combining the Lemma 1 and the Lemma 3, we can draw the conclusion that given training dataset S of size N, for any $0 \leq \delta_1 \leq 1$, $0 \leq \delta_2 \leq 1$, with the probability of at least $(1-\delta_1)(1-\delta_2)$:

$$\begin{aligned}
R_s \leq \frac{1}{2}\ln\left(C + (C-1)\sqrt{\frac{\ln 2/\delta_1}{2N}} - \hat{D}_{KB}\right) \\
+ 3B_\eta\sqrt{\frac{\ln 2/\delta_2}{2N}} + 2\hat{R}_m(\mathcal{F}_\eta) + \gamma_\eta, \quad \forall f \in \mathcal{F}_\eta
\end{aligned}$$

where $B_\eta = \ln C + \ln \eta$.

Lemma 4 $\forall f \in \mathcal{F}_\epsilon$, $\mathcal{L}_{nesy}$ is bounded.

Proof. Similarly to Lemma 3, we only have got to demonstrate that $\mathcal{L}_{nesy}$ has an upper bound. Based on the conclusion that $\forall f \in \mathcal{F}_\epsilon, f(\mathbf{g}|\mathbf{x}) \geq 1 - (C-1)\epsilon$ as previously demonstrated, we can get that $\forall \mathbf{x}$

$$\begin{aligned}
&-\ln\left(\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{I}(\mathbf{z}, \text{KB} \models \mathbf{y}) \cdot f(\mathbf{z}|\mathbf{x})\right) \\
&\leq -\ln\left(f(\mathbf{g}|\mathbf{x})\right) \\
&\leq -\ln\left(1 - (C-1)\epsilon\right)
\end{aligned}$$

Therefore, $\mathcal{L}_{nesy}$ has an upper bound $B_\epsilon = -\ln\left(1 - (C-1)\epsilon\right)$

C.4.2. Proof of Theorem 4.5.(ii)

Based on the Lemma 4 before and analogously to the proof of theorem 4.5.(i), we utilize the empirical Rademacher complexity and Lemma 1 to draw the conclusion that for any $0 \leq \delta_1 \leq 1$, $0 \leq \delta_2 \leq 1$, with the probability of at least

$(1-\delta_1)(1-\delta_2)$:

$$R_s \leq \frac{1}{2}\ln\left(C+(C-1)\sqrt{\frac{\ln 2/\delta_1}{2N}}-\hat{D}_{KB}\right) + 3B_\epsilon\sqrt{\frac{\ln 2/\delta_2}{2N}}+2\hat{R}_m(\mathcal{F}_\epsilon)+\gamma_\epsilon, \quad \forall f\in\mathcal{F}_\epsilon$$

where $B_\epsilon = -\ln(1-(C-1)\epsilon)$.

C.5. Proof of the upper bounds of R_s^{ABL}

To simplify the notation, we pre-define $A_y = \{\mathbf{z}|\mathbf{z}, \text{KB} \vDash \mathbf{y}\}$, so we can get $D_{KB} = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathcal{S}}|A_y|$.

C.5.1. Proof of Theorem 5.2.(ii)

Next, we provide proof for the upper bound of R_s^{ABL} when Dis $= 0$.

Proof.

$$\begin{aligned}
\hat{\mathcal{L}}_{ABL} &= -\frac{1}{N}\sum_{\mathbf{x}_i,\mathbf{y}_i}\mathbb{E}_{\overline{\mathbf{z}}_i\sim\Phi}\ln\left(f(\overline{\mathbf{z}_i}|\mathbf{x})\right)\\
&\geq -\frac{1}{N}\sum_{\mathbf{x}_i,\mathbf{y}_i}\ln\left(\mathbb{E}_{\overline{\mathbf{z}}_i\sim\Phi}(f(\overline{\mathbf{z}_i}|\mathbf{x}))\right)\\
&= -\frac{1}{N}\sum_{\mathbf{x}_i,\mathbf{y}_i}\ln\left(\sum_{\mathbf{z}\in\mathcal{Z}}\left(\frac{1}{|A_y|}\mathbb{I}(\mathbf{z},\text{KB}\vDash\mathbf{y})\cdot f(\mathbf{z}|\mathbf{x})\right)\right)\\
&= -\frac{1}{N}\sum_{\mathbf{x}_i,\mathbf{y}_i}\left(\mathbb{I}(\mathbf{z},\text{KB}\vDash\mathbf{y})\cdot f(\mathbf{z}|\mathbf{x})\right)+\frac{1}{N}\sum_{\mathbf{x}_i,\mathbf{y}_i}\ln|A_y|\\
&\geq \hat{\mathcal{L}}_{nesy}+\sum_{\mathbf{x}_i,\mathbf{y}_i}\ln|A_y|\\
&\geq \hat{\mathcal{L}}_{nesy}+\frac{1}{N}\sum_{\mathbf{x}_i,\mathbf{y}_i}\left(\frac{\ln C}{C-1}|A_y|-\frac{\ln C}{C-1}\right)\\
&= \hat{\mathcal{L}}_{nesy}+\frac{\ln C}{C-1}\left(\frac{1}{N}\sum_{\mathbf{x}_i,\mathbf{y}_i}|A_y|-1\right)\\
&= \hat{\mathcal{L}}_{nesy}+\frac{\ln C(C-\hat{D}_{KB}-1)}{C-1}
\end{aligned}$$

Thus, we can obtain that

$$\begin{aligned}
R_s^{ABL} &= \mathcal{L}-\hat{\mathcal{L}}_{ABL}\\
&\leq \mathcal{L}-\mathcal{L}_{nesy}-\frac{\ln C(C-\hat{D}_{KB}-1)}{C-1}\\
&= R_s-\frac{\ln C}{C-1}(C-\hat{D}_{KB}-1)
\end{aligned}$$

The conclusion successfully illustrates that R_s^{ABL} is upper bounded by R_s and will have a tighter bound when $\hat{D}_{KB}$ is smaller.

C.5.2. Proof of Theorem 5.3.(i)

Under the label smooth assumption, we have

$$\frac{1}{C\eta}\leq f(\mathbf{z}|\mathbf{x})\leq\frac{\eta}{C}$$

So, we have *Proof.*

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[R_s^{ABL}] &= \mathcal{L} - \mathcal{L}_{ABL} \\
&= \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\mathbb{E}_{\overline{\mathbf{z}}\sim\Phi}\ln f(\overline{\mathbf{z}}|\mathbf{x}) - \ln f(\mathbf{z}|\mathbf{x})\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\frac{\mathbb{E}_{\overline{\mathbf{z}}\sim\Phi}f(\overline{\mathbf{z}}|\mathbf{x})}{f(\mathbf{z}|\mathbf{x})}\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\frac{\Phi(\mathbf{z})\cdot f(\mathbf{z}|\mathbf{x}) + \kappa(|A_y|-1)\cdot\frac{\eta}{C}}{f(\mathbf{z}|\mathbf{x})}\right) \\
&= \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\left(\Phi(\mathbf{z}) + \frac{\kappa(|A_y|-1)\cdot\frac{\eta}{C}}{f(\mathbf{z}|\mathbf{x})}\right)\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\left(\Phi(\mathbf{z}) + \frac{\kappa(|A_y|-1)\cdot\frac{\eta}{C}}{\frac{1}{C\eta}}\right)\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\left(1+\kappa(|A_y|-1)\eta^2\right)\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}(|A_y|-1)\eta^2\kappa \\
&= \eta^2(C - D_{KB} - 1)\cdot\kappa
\end{aligned}$$

C.5.3. Proof of Theorem 5.3.(ii)

Under the pre-training assumption, we have *Proof.*

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{S}}[R_s^{ABL}] &= \mathcal{L} - \mathcal{L}_{ABL} \\
&= \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\mathbb{E}_{\overline{\mathbf{z}}\sim\Phi}\ln f(\overline{\mathbf{z}}|\mathbf{x}) - \ln f(\mathbf{z}|\mathbf{x})\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\frac{\mathbb{E}_{\overline{\mathbf{z}}\sim\Phi}f(\overline{\mathbf{z}}|\mathbf{x})}{f(\mathbf{z}|\mathbf{x})}\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\frac{\Phi(\mathbf{z})\cdot f(\mathbf{z}|\mathbf{x}) + \kappa(|A_y|-1)\cdot\epsilon}{f(\mathbf{z}|\mathbf{x})}\right) \\
&= \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\left(\Phi(\mathbf{z}) + \frac{\kappa(|A_y|-1)\cdot\epsilon}{f(\mathbf{z}|\mathbf{x})}\right)\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\left(\Phi(\mathbf{z}) + \frac{\kappa(|A_y|-1)\cdot\epsilon}{1-(C-1)\epsilon}\right)\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left(\ln\left(1 + \frac{\kappa(|A_y|-1)\cdot\epsilon}{1-(C-1)\epsilon}\right)\right) \\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\frac{\kappa\epsilon(|A_y|-1)}{1-(C-1)\epsilon} \\
&= \frac{\epsilon(C - D_{KB} - 1)}{1-(C-1)\epsilon}\cdot\kappa
\end{aligned}$$