SELECTIVE CLASSIFIER ENSEMBLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective classification allows a machine learning model to abstain on some hard inputs and thus improve the safety of its predictions. In this paper, we study the ensemble of selective classifiers, i.e. *selective classifier ensemble*, which combines several weak selective classifiers to obtain a more powerful model. We prove that under some assumptions, the ensemble has a lower *selective risk* than the individual model under a range of coverage. This is nontrivial since the selective risk is a non-convex function of the model prediction. The assumptions and the theoretical result are supported by systematic experiments on both computer vision and natural language processing tasks. A surprising empirical result is that a simple selective classifier ensemble, namely, the *ensemble model with maximum probability as confidence*, is the state-of-the-art selective classifier. For instance, on CIFAR-10, using the same VGG-16 backbone model, this ensemble reduces the AURC (Area Under Risk-Coverage Curve) by about 24%, relative to the previous state-of-the-art method.

1 INTRODUCTION

Although recent years have witnessed the broad applications of deep learning models, their securities have not been fully guaranteed, which gives rise to the study of selective classification. For any given deep learning classifier, there might be inputs that the model is not able to classify in practical applications, for which the model might make unpredictable errors. To prevent this kind of error, we need to delimit the application scope of the deep learning classifier accurately. This gives rise to the study of *selective classification* that learns a *selective classifier* (f, g), where f is a conventional classifier and g is a *selective function* that decides whether the selective classifier should abstain from prediction. Since the classifier is well studied, the study of selective classification focuses on the design of the selective function.

A standard approach to designing the selective function is to design a confidence score function with a threshold, and several confidence score functions have been developed. A simple confidence score function is the maximum predictive probability of the classifier (Hendrycks & Gimpel, 2017). More advanced methods modify either the model architecture (Geifman & El-Yaniv, 2019) or the loss function (Liu et al., 2019; Huang et al., 2020) of the classifier to train the confidence score function and the classifier simultaneously. For example, Deep Gambler (Liu et al., 2019) regards the selective classification problem as gambling and proposes a novel loss function to train the classifier and the confidence score function. Although there are various individual models for the selective classifier, there has been no systematic study of the ensemble method in selective classification.

It is well known that the ensemble method, which combines the individual models to obtain a more powerful model, can improve predictive performances of machine learning models (see Zhou (2012) for a review), but there has been no systematic study of ensemble methods in selective classification. Although the ensemble of Softmax Response (Hendrycks & Gimpel, 2017) has been empirically studied by Lakshminarayanan et al. (2017), ensembles of other kinds of selective classifiers and the theoretical foundation of the ensemble in selective classification have not been studied yet. In this paper, we first demonstrate the theoretical foundation of the ensemble on selective classifiers, that is, with some assumptions, the ensemble has a lower selective risk than the individual model under a range of coverage. This is nontrivial since the selective risk (with the 0/1 loss) are non-convex. Second, we show the experimental results of the performance of the ensemble in selective classification. The contributions of this paper are summarized as follows.

- We are the first to theoretically demonstrate that based on several reasonable assumptions, the ensemble has a lower selective risk than the individual model under a range of coverage. We verify this by systematic experiments on the tasks of image classification and text classification.
- We show a surprising experimental result that two simple methods, the SR ensemble and the Reg-curr ensemble, which can be summarized as *the ensemble model with maximum probability as confidence*, are the state-of-the-art selective classifiers.

2 Problem Formulation of Selective Classification

A selective classifier is composed of a standard classifier and a selective function. Considering a standard classification problem, $\mathbb{X}$ is a feature space, $\mathbb{Y} = \{1, 2, ..., K\}$ is a finite label set, and a classifier f is a function $f : \mathbb{X} \to \mathbb{Y}$. A labeled dataset $\mathbb{D} = \{(x_i, y_i)\}_{i=0}^{N} \subseteq \mathbb{X} \times \mathbb{Y}$ is sampled from a distribution $p_{\mathrm{X,Y}}$. Our goal is to learn a *selective classifier* where f is a standard classifier and $g : \mathbb{X} \to \{0, 1\}$ is a selective function that estimates the correctness of f's prediction. Given input x, the output of selective classifier (f, g) is

$$(f, g)(x) = \begin{cases} f(x), & \text{if } g(x) = 1 \\ \text{Abstain}, & \text{if } g(x) = 0 \end{cases}. \tag{1}$$

Usually, g is realized by a *confidence score* $\hat{\kappa} : \mathbb{X} \to \mathbb{R}^+$ with a threshold τ (Geifman & El-Yaniv, 2017), namely

$$g(x) = \mathbb{I}\{\hat{\kappa}(x) > \tau\}, \tag{2}$$

where $\mathbb{I}$ is the indicator function.

Coverage and *selective risk* are two basic evaluation metrics of selective classifiers, and the goal of selective classifiers is to minimize the selective risk for a target coverage. The *coverage* of (f, g) is defined to be the probability of (f, g) not abstaining from prediction (Geifman & El-Yaniv, 2017), i.e.

$$\phi(f, g) := \mathbb{E}_{p(x)}[g(x)], \tag{3}$$

where $p(x)$ is the probability density function of input x. The *selective risk* (Geifman & El-Yaniv, 2017) of (f, g) is

$$R(f, g) := \frac{\mathbb{E}_{p(x)}[\ell(f(x), y)g(x)]}{\mathbb{E}_{p(x)}(g(x))}, \tag{4}$$

where $\ell : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$ is a given loss function. Usually, ℓ is the 0/1 loss (Geifman & El-Yaniv, 2017; 2019; Liu et al., 2019; Huang et al., 2020). Based on these definitions, the objective of selective classifiers is formalized as

$$\min R(f, g), \text{ s.t. } \phi(f, g) \geq c_{\text{target}},$$

where c_{target} is a given target coverage.

When the selective function g is developed as (2), the confidence threshold τ controls the trade-off between coverage and selective risk. With different values of τ, (f, g) has different pairs of coverage and selective risk $(\phi(f, g; \tau), R(f, g; \tau))$, which forms the *risk-coverage curve* (Geifman & El-Yaniv, 2017) of (f, g). The risk-coverage curve specifies the entire performance profile of a selective classifier, and it is easy to see that the selective classifier with a lower risk-coverage curve is better. To evaluate selective classifiers more concisely, the *area under the risk-coverage curve* (*AURC*) is introduced as a metric of selective classifiers (Xin et al., 2021), and the selective classifier with a lower AURC is better.

3 Related Work

Here, we summarize the previous studies on selective classification and ensemble methods. We also discuss the difference between selective classification and out-of-distribution detection.

3.1 SELECTIVE CLASSIFICATION

The critical problem of designing a selective classifier is to design its selective function, and there are two types of the selective function g. One is the *implicit selective function*, which is derived from the classifier. The other is the *explicit selective function*, which is a neural network trained with the classifier f simultaneously. Previous works for selective classifiers are listed as follows.

Selective classifiers with implicit selective functions include SR (Softmax Response) (Hendrycks & Gimpel, 2017), MC-Dropout (Monte Carlo-Dropout) (Gal & Ghahramani, 2016), and Reg-curr (Xin et al., 2021). In SR, the selective classifier is a vanilla classifier with its maximum predictive probability as confidence (i.e. the maximum of the output of the softmax layer). In MC-Dropout, it enables the dropout layer of the classifier and runs multiple feed-forward iterations at inference time to obtain the variance of the maximum probability output of f's softmax layer, whose negative value is used as the confidence score. Reg-curr behaves the same as the SR at inference time but uses an RPP-based regularizer at training time, where RPP (Reversed Pair Proportion) (Xin et al., 2021) is the proportion of reversed pairs of confidence scores.

Selective classifiers with explicit selective functions include SN (SelectiveNet) (Geifman & El-Yaniv, 2019), Gambler (Deep Gambler) (Liu et al., 2019), and SAT (Self-Adaptive Training) (Huang et al., 2020). SN is a neural network that combines f and g, where f and g share convolutional layers and have their separate fully-connected layers. The loss function of the model is the selective risk with some regularizers. A hyperparameter c is needed to specify the target coverage. Gambler adds the abstention option to the classifier as an extra class, that is, for a given input x, the predictive probability of the extra class is the confidence of abstention, i.e, $1 - \hat{\kappa}(x)$. At training time, it regards the selective classification problem as gambling and is trained to maximize the gambling reward. Similar to Gambler, SAT adds the abstention option to the classifier as an extra class. However, SAT has a different training procedure, which is trained with a soft label that tells the model on which sample to abstain. SAT is the previous state-of-the-art selective classifier for image classification tasks.

3.2 RANDOMIZATION-BASED ENSEMBLE

In the *randomization-based* ensemble method, each member model is trained independently with the same architecture and training procedure but with different randomization seeds for the random initialization of parameters and the random shuffling of training data for each training epoch. At inference time, the predictive probability of the ensemble for each class is the average of those of member models. Lakshminarayanan et al. (2017) applies this ensemble method to deep neural networks (Deep Ensemble) and has achieved state-of-the-art performance in uncertainty estimation. The difference between Lakshminarayanan et al. (2017) and our work is that they only empirically study the ensemble of vanilla classifiers (the SR ensemble), but we not only empirically study ensembles of multiple outstanding selective classifiers but also provide the theoretical foundation of the ensemble in selective classification. As for the theoretical works, Krogh & Vedelsby (1994) proposes the *error-ambiguity decomposition* to explain the better performance of the randomization-based ensemble in regression tasks. However, for classification tasks, there is no such simple and elegant analysis, since the evaluation metrics are non-convex (Zhou, 2021). Thus, the corresponding analysis for classification tasks needs additional assumptions, e.g. unbiased, uncorrelated, and identically distributed estimation errors for the posterior probability distribution (Tumer & Ghosh, 1996; Fumera & Roli, 2005). Nevertheless, these assumptions are impractical (Fumera & Roli, 2005). As far as we know, there is no systematic study of the ensemble in the context of selective classification.

3.3 OUT-OF-DISTRIBUTION DETECTION

A related topic of selective classification is out-of-distribution (OOD) detection (Lakshminarayanan et al., 2017)) (also called as open set recognition (Scheirer et al., 2012), or novelty detection (Schölkopf et al., 2001)), which detects samples different greatly from a given dataset, i.e. OOD samples. The essential difference between selective classification and OOD detection lies in their different goals. The goal of the former is to detect samples where the classifier predicts incorrectly, which depends on both the classifier and samples, while that of the latter is to detect samples different greatly from a given dataset, which depends on samples only. In addition, at present, selective classification assumes that test data and training data are sampled from the same distribution (Geifman

& El-Yaniv, 2017), instead of using OOD test data as OOD detection. Thus, selective classification and OOD detection are complementary in preventing error predictions of machine learning models, as Figure 3 shows.

4 METHOD

With the randomization-based ensemble method, we propose the *selective classifier ensemble*. The basic idea is that each predictive probability (as well as the confidence score in the case of explicit selective functions) of the ensemble should be the average of those of the member models. Formally, we assume that for an input sample x, a classifier f at first provides the predictive probability distribution $\hat{\boldsymbol{\pi}}_\theta = (\hat{\pi}_\theta^1, \cdots, \hat{\pi}_\theta^K)$ and then makes prediction $f(x;\theta) = \arg\max_{1\leq k\leq K} \hat{\pi}_\theta^k(x)$, where K is the number of classes, θ denotes the parameters of f, and $\hat{\pi}_\theta^k(x)$ is the predictive probability for class k (the superscript is not an exponent). Then, the predictive probability distribution of the ensemble classifier of M member models is

$$\hat{\boldsymbol{\pi}}_{\text{ens}}(x) := \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{\pi}}_m(x). \tag{5}$$

The ensemble of the selective function is defined as follows. For implicit selective functions (e.g. SR), to keep the ensemble the same kind of selective classifier as the individual model (for example, the ensemble of SR should still be an SR model), the confidence score of the ensemble is derived from $\hat{\boldsymbol{\pi}}_{ens}$ in the same way as the individual model. For example, the confidence score of the *SR ensemble* is

$$\hat{\kappa}_{\text{ens}}(x) = \max_k \hat{\pi}_{\text{ens}}^k(x). \tag{6}$$

For explicit selective functions (e.g. SAT), the confidence score of the ensemble is the average of those of member models,

$$\hat{\kappa}_{\text{ens}}(x) = \frac{1}{M}\sum_{m=1}^{M} \hat{\kappa}_m(x). \tag{7}$$

5 THEORETICAL ANALYSIS OF SELECTIVE CLASSIFIER ENSEMBLE

In this section, we analyze the selective risk (with the 0/1 loss) of the ensemble of a simple selective classifier, the SR ensemble (see Section 4 for its definition). If the selective risk is a convex function of the predictive probability distribution, then according to the definition of the convex function, (6) implies that the selective risk of the ensemble is less than or equal to that of the individual model. However, the selective risk is non-convex, because the 0/1 loss is a step function. Thus, the analysis is not easy. To prove a lower selective risk of the ensemble, we need some assumptions. We introduce the assumptions in Section 5.1 (verified in Section 6.1) and the theoretical results in Section 5.2. The analysis for the other selective classifiers is left for future study.

5.1 ASSUMPTIONS

Given an SR ensemble with M $(M > 1)$ members, we assume that there are samples on which all member models provide almost the same predictive probability distributions. Furthermore, we idealize them as *definite samples*, for which all member models provide exactly the same predictive probability distribution. Then, the rest of samples are referred to as *ambiguous samples*. Let D be the event that the input sample is definite and A be the event that the input sample is ambiguous. Considering the case that the input sample is randomly drawn from a dataset, the predictive probability for class k $(1 \leq k \leq K)$, and the confidence of the SR model are random variables. We denote these random variables as Π^k and C respectively and use π^k and κ to denote their values respectively. Generally, for any continuous random variable Z, p_{Z} denotes the *probability density function* (*PDF*) of Z, and for any real variable z, $z \to 1^-$ denotes that z approaches 1 from the left. Based on the idealization and notations above, we introduce the following assumptions.

Assumption 1. For any individual SR model, with its confidence score denoted as C, we have

$$\lim_{\tau\to 1^-} \Pr(Err|A, \text{C} \geq \tau) > \lim_{\tau\to 1^-} \Pr(Err|D, \text{C} \geq \tau), \tag{8}$$

where Err is the event that the model makes an error prediction.

In Assumption 1, $\Pr(Err|A, \mathrm{C} \geq \tau)$ is the selective risk of the individual model with a confidence threshold of τ for ambiguous samples, and $\Pr(Err|D, \mathrm{C} \geq \tau)$ is that for definite samples. The motivation of Assumption 1 is that ambiguous samples seem more difficult to classify than definite samples.

Assumption 2. For any individual SR model, both $\lim_{\kappa \to 1^-} p_{\mathrm{C}}(\kappa|D)$ and $\lim_{\kappa \to 1^-} p_{\mathrm{C}}(\kappa|A)$ exist and are non-zero, i.e. positive, where C is the confidence score of the SR model.

Assumption 2 claims that for both ambiguous samples and definite samples, when the confidence score of the individual model approaches 1, its PDF approaches a non-zero value. This means individual model might assign high confidence to ambiguous samples with a considerable probability, which might lead to the overconfident problem, that is, the model assign high confidence to error predictions. In Section 5.2, we prove that the ensemble avoids this problem and thus has a better performance than the individual model.

Assumption 3. For any SR ensemble of M member models, let Π_i^k be the predictive probability of the i-th member model for class k. Then the joint probability density function $p_{\Pi_1^k, \Pi_2^k, \ldots, \Pi_M^k}(\cdot|A), \forall k \in \{1, \ldots, K\}$, is bounded.

Assumption 3 guarantees that the joint distribution of $\Pi_1^k, \Pi_2^k, \ldots, \Pi_M^k$ for ambiguous samples is not so extreme as that for definite samples. For definite samples, given the predictive probability distribution of one member model (denoted as $\boldsymbol{\pi}_0$), the predictive probability distributions of other members are certain, so the joint probability density function goes infinite at $(\boldsymbol{\pi}_0, ..., \boldsymbol{\pi}_0)$. Assumption 3 forbids such extreme case for ambiguous samples.

5.2 Analysis Results

With the assumptions above, we obtain Proposition 1 and Theorem 2 (see Appendix B for details of proof). Proposition 1 provides the key insight into the better performance of the ensemble, which shows that the ensemble does not assign high confidence to ambiguous samples and thus overcomes the problem of the individual model.

Proposition 1. *If Assumption 3 holds, then* $\lim_{\kappa_{\mathrm{ens}} \to 1^-} p_{\mathrm{C}_{\mathrm{ens}}}(\kappa_{\mathrm{ens}}|A) = 0$*, where* $\mathrm{C}_{\mathrm{ens}}$ *is the confidence score of the ensemble.*

Proposition 1 suggests that given the input sample being ambiguous, when the confidence of the ensemble approaches one, its PDF approaches zero, so the ensemble does not provide high (close to one) confidence for ambiguous samples in a probabilistic sense. Based on this, Theorem 2 completes the proof of the better selective classification performance of the SR ensemble than an SR individual model.

Theorem 2. *If Assumption 1-3 holds, then* $\exists \phi_0 \in (0, 1)$*, such that* $\forall \phi \in (0, \phi_0)$*,*

$$R_{\mathrm{ens}}(\phi) < R_{\mathrm{ind}}(\phi), \tag{9}$$

where $R_{\mathrm{ens}}(\phi)$ *and* $R_{\mathrm{ind}}(\phi)$ *are the selective risks given the coverage* ϕ *of the SR ensemble and the individual SR model respectively.*

6 Experiments

Datasets. We conduct experiments on multiple datasets for image classification and text classification tasks. Following Geifman & El-Yaniv (2017; 2019); Liu et al. (2019); Huang et al. (2020), we use CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and SVHN (Netzer et al., 2011) for image classification tasks, and following Xin et al. (2021), we use MRPC (Dolan & Brockett, 2005), MNLI (Williams et al., 2018), and QNLI (Wang et al., 2018) for text classification tasks. The usage of the training set and test set in selective classification are the same as the standard classification, because current selective classification focuses on in-domain data (i.e., data from the same distribution as the training set) (Geifman & El-Yaniv, 2017). For example, if the selective classifier is trained on the training set of CIFAR-10 (Krizhevsky, 2009), then it will be tested on the test set of CIFAR-10. Furthermore, MNLI's development set and test set are divided into *matched* and *mismatched* parts.

The matched parts are sampled from the same source as the training set (so they are in-domain data), while the mismatched parts are sampled from different sources. In our experiments, only the matched parts are used in experiments, since the current selective classification only considers in-domain data. In addition, test sets of MRPC, QNLI, and MNLI are not accessible, so we use their development sets as test sets. Following Liu et al. (2019); Huang et al. (2020), since CIFAR-10, CIFAR-100 and SVHN originally had no development set, their development sets were 2000 samples randomly split from corresponding test sets. More details of all the datasets in our experiments are described in Appendix C.1.

Evaluation Metrics. The evaluation metrics are AURC and selective risk (the lower the better for both). AURC is a comprehensive metric of selective classifier, and selective risk is a common metric in previous works (Geifman & El-Yaniv, 2019; Liu et al., 2019; Huang et al., 2020). In this paper, given a selective classifier, the result of selective risk is shown in the form of risk-coverage curves, which shows the selective risk of the selective classifier against its coverage. According to the object of selective classification, a selective classifier with a lower risk-coverage curve is better.

Networks. Following Huang et al. (2020); Xin et al. (2021), for image classification and text classification, we use VGG-16 (Simonyan & Zisserman, 2014) and BERT-base (Devlin et al., 2019) as backbones of selective classifiers respectively. More details of the backbone models and their training procedures are provided in Appendix C.2.

Baselines. We use SR (Geifman & El-Yaniv, 2017), SN (SelectiveNet) (Geifman & El-Yaniv, 2019), Gambler (Liu et al., 2019), SAT (Huang et al., 2020), and Reg-curr (Xin et al., 2021) for both image classification and text classification. Note that the SN is optimized for a fixed coverage, so the comprehensive metrics AURC and risk-coverage curve, which summarizes performances under different coverages, are not suitable for evaluating the SN. Thus, we only evaluate the selective risk for a fixed coverage of the SN ensemble and provide the results in Appendix D. To ensure a fair comparison, for tasks of image classification, all the baselines are re-implemented based on the open resource code of SAT (Huang et al., 2020), and for tasks of text classification, they are re-implemented based on the open resource code of Reg-curr (Xin et al., 2021). The details of hyperparameters of each baseline are provided in Appendix C.3.

6.1 Verification of the Assumptions

We examine Assumption 1 and Assumption 2 (*only* for the baseline of SR) on datasets for both image classification and text classification. Since the definite samples are the idealization of samples on which member models provide almost the same predictive probability distributions, we take samples with a *standard deviation of predictive probability distributions of member models* (or STD for short) less than a small positive number ϵ as definite samples and the other samples as ambiguous samples in experiments. Formally, STD $:= \sqrt{\frac{\sum_{j=1}^{M}(\hat{\boldsymbol{\pi}}_j - \frac{1}{M}\sum_{i=1}^{M}\hat{\boldsymbol{\pi}}_i)^2}{M-1}}$, where $\hat{\boldsymbol{\pi}}_j$ is the predictive probability distribution vector of the j-th member model, and samples with STD $< \epsilon$ approximates definite samples in experiments. We choose $\epsilon = 10^{-3}$ for datasets of image classification and $\epsilon = 10^{-2}$ for datasets of text classification. Figure 1(a) and Figure 1(e) show the selective error rates (selective risks) of samples with STD $< 10^{-3}$ and samples with STD $\geq 10^{-3}$ given a range of confidence thresholds (which approximate $\Pr(Err|D, \mathrm{C} \geq \tau)$ and $\Pr(Err|A, \mathrm{C} \geq \tau)$) on the test set of each dataset. The results show that the selective risk of samples with STD $< 10^{-3}$ is lower than that of samples with STD $\geq 10^{-3}$ for all confidence thresholds near 1 on all datasets except MRPC, which verifies Assumption 1. Figure 1(b)-1(d) and Figure 1(f)-1(h) shows the histogram of confidence scores of samples with STD $< \epsilon$ and that of other samples, which approximate $p_{\mathrm{C}}(\kappa|D)$ and $p_{\mathrm{C}}(\kappa|A)$, on the test set of each dataset. The results show that the number of samples with STD $< 10^{-3}$ is non-zero in the top bin on all datasets except MRPC, which verifies Assumption 2. In summary, Assumption 1 and 2 holds on all datasets except MRPC.

6.2 Evaluation of Selective Classifier Ensembles

Table 1 shows the AURCs of the individual model and the ensembles of 2-5 member models for each baseline on each dataset, where o is the hyperparameter of Deep Gambler. The results show that the AURC of the ensemble is significantly lower than that of the individual model in each case except the case of Gambler on MRPC. On CIFAR-10, CIFAR-100, SVHN, MRPC, QNLI, and MNLI, the

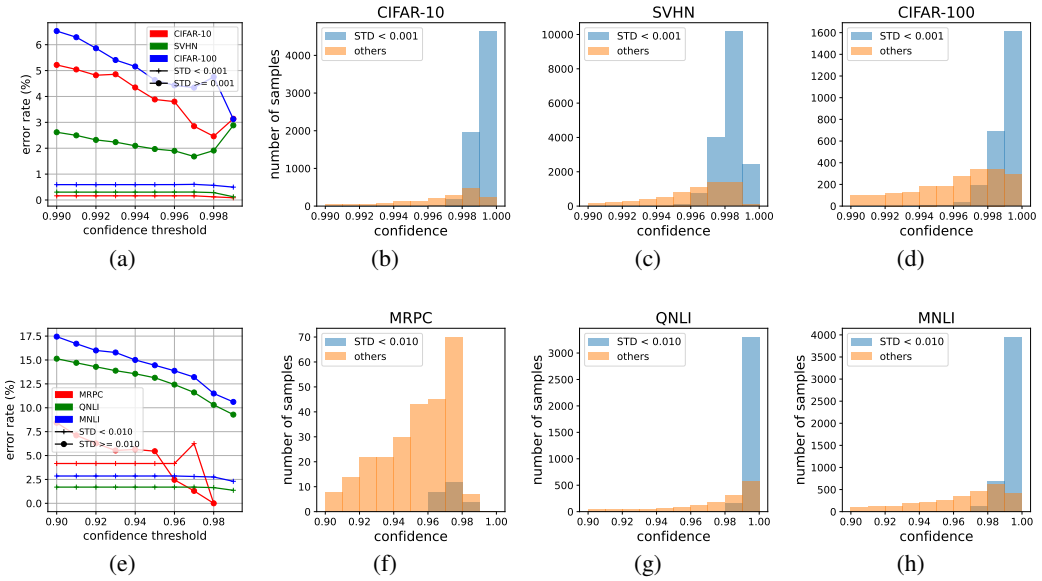

Figure 1: (a)/(e): the selective error rates (selective risks) of definite samples and ambiguous samples given a range of confidence thresholds on the test set of each dataset for image/text classification. (b)-(d): the histogram of confidence scores of samples with STD $< 10^{-3}$ and that of other samples on the test set of each dataset for image classification. (f)-(h): the histogram of confidence scores of samples with STD $< 10^{-2}$ and that of other samples on the test set of each dataset for text classification.

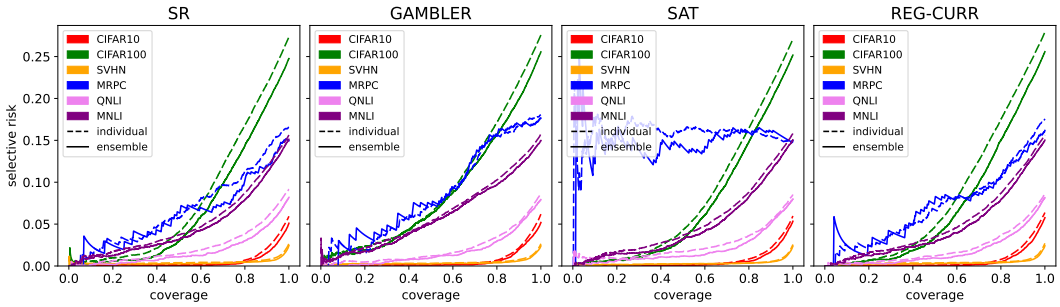

Figure 2: Risk-coverage curves of ensembles and individual models of each baseline on multiple datasets, where all ensembles consist of 5 member models.

AURCs of the best ensemble models of 5 members are respectively 23%, 15%, 7%, 11%, 14%, and 9% lower than those of the best individual models. Among all ensembles, SR ensemble has the lowest AURCs on CIFAR-10 and CIFAR-100, SAT ensemble has the lowest AURCs on SVHN [1], and Reg-curr has the lowest AURCs on all datasets for text classification. Risk-coverage curves of the ensembles and the individual models of each baseline on each dataset are shown in Figure 2, which shows that for each baseline on each dataset except MRPC, the risk-coverage curve of the ensemble is almost completely below that of the individual model. In other words, on each dataset except MRPC, the ensemble has a lower selective risk than the individual model under any coverage.

[1]The result that only on SVHN is SAT the state-of-the-art contradicts that of Huang et al. (2020). This may be because Huang et al. (2020) directly cites the evaluation results of other methods from different papers, which may have subtle differences in the training settings of the backbone model (e.g. data augmentation). By contrast, we re-implement all selective classification methods based on the open resource code of SAT (Huang et al., 2020) for image classification in our experiments, so the fairness of comparison is guaranteed.

Table 1: AURC$/10^{-4}$ on each dataset, where MNLI-(m) is the matched part of the MNLI development set. The best entries are marked in bold.

Dataset	#Member	SR	Gambler	SAT	Reg-curr
CIFAR-10	1	**66.1±5.6**	$^{o=2.20}$ 70.7±2.6	67.1±0.3	67.1±2.0
	2	**55.5±1.7**	$^{o=2.20}$ 63.2±1.8	63.0±0.4	59.2±1.1
	3	**53.0±1.7**	$^{o=2.20}$ 60.8±1.2	59.1±1.1	56.0±1.6
	4	**51.3±2.2**	$^{o=2.20}$ 58.9±2.0	58.3±0.5	54.6±1.0
	5	**51.2±2.0**	$^{o=2.20}$ 57.4±1.0	57.6±0.6	53.5±0.5
CIFAR-100	1	**793.1±9.0**	$^{o=4.60}$ 930.6±10.2	807.3±6.0	851.9±12.8
	2	**722.6±3.5**	$^{o=4.60}$ 880.7±7.0	756.7±1.6	767.6±1.1
	3	**695.9±2.6**	$^{o=4.60}$ 871.3±4.3	741.2±3.8	739.5±5.7
	4	**681.8±4.4**	$^{o=4.60}$ 859.0±4.6	729.2±6.9	721.8±7.9
	5	**672.5±1.9**	$^{o=4.60}$ 857.1±4.7	715.5±12.2	714.2±6.3
SVHN	1	46.7±0.2	$^{o=2.60}$ 44.2±2.5	**35.0±0.5**	43.8±1.0
	2	42.1±1.8	$^{o=2.60}$ 41.3±2.9	**33.5±0.8**	39.9±1.4
	3	39.4±0.8	$^{o=2.60}$ 38.2±0.8	**32.9±1.1**	38.2±0.3
	4	38.1±0.9	$^{o=2.60}$ 37.2±1.0	**32.7±0.9**	37.0±1.4
	5	37.1±0.3	$^{o=2.60}$ 37.0±0.8	**32.4±0.9**	36.6±1.2
MRPC	1	**683.8 ±95.6**	$^{o=1.80}$691.2 ±100.5	1725.5 ±102.9	718.1 ±19.6
	2	**617.6 ±2.5**	$^{o=1.80}$683.8 ±93.1	1495.1 ±139.7	639.7 ±19.6
	3	634.8 ±14.7	$^{o=1.80}$678.9 ±105.4	1382.4 ±137.3	**629.9 ±22.1**
	4	**615.2 ±31.9**	$^{o=1.80}$678.9 ±110.3	1377.5 ±110.3	617.6 ±66.2
	5	627.5 ±9.8	$^{o=1.80}$693.6 ±127.5	1414.2 ±85.8	**610.3 ±78.4**
QNLI	1	221.5 ±8.1	$^{o=1.60}$235.8 ±35.1	246.6 ±4.2	**201.0 ±1.3**
	2	196.4 ±7.0	$^{o=1.60}$195.5 ±12.8	220.0 ±1.1	**186.9 ±1.5**
	3	180.1 ±4.4	$^{o=1.60}$188.4 ±9.0	201.5 ±4.6	**176.6 ±3.3**
	4	175.7 ±3.5	$^{o=1.60}$181.0 ±6.2	196.0 ±5.5	**174.3 ±3.5**
	5	173.3 ±1.8	$^{o=1.60}$177.7 ±5.3	192.9 ±1.8	**171.9 ±2.4**
MNLI-(m)	1	515.3 ±8.6	$^{o=2.80}$607.6 ±19.3	519.8 ±8.0	**496.8 ±9.5**
	2	482.3 ±9.3	$^{o=2.80}$590.6 ±19.4	482.8 ±16.5	**466.9 ±7.7**
	3	475.0 ±8.4	$^{o=2.80}$563.0 ±11.4	470.1 ±6.5	**458.5 ±5.3**
	4	467.1 ±2.8	$^{o=2.80}$567.1 ±14.9	460.1 ±6.4	**453.0 ±1.0**
	5	465.2 ±3.8	$^{o=2.80}$569.4 ±11.4	454.4 ±3.1	**451.9 ±2.1**

The experimental results for SR in Figure 2 verify the correctness and practicability of Theorem 2. The assumptions hold on each dataset except MRPC. Combining this with Theorem 2, the SR ensemble should have a lower selective risk than the individual model under a range of coverage (from 0 to some ϕ_0) on each dataset except MRPC. This is supported by the experimental results.

It is surprising that the SR ensemble and Reg-curr ensemble are state-of-the-art selective classifiers for image and text classification tasks respectively. They only use the maximum probability as the confidence score, while SAT and SN use more sophisticated and explicit confidence score function. Note that, on SVHN, the SAT ensemble performs better than the SR ensemble. We find that the annotations on SVHN are noisy, and after manually removing some noisy samples in the training set, SR ensemble performs better than SAT ensemble. Therefore, the SVHN result indicates that SAT is better at handling label noise rather than selecting when to abstain. Please refer to Appendix E.3 for more details.

In addition, we conduct experiments to explore further properties of the selective classifier ensemble, including the effect of the number of members and the relationship between the classification performance and the selective classification performance (see Appendix E). The results are that an ensemble with more member models has a better selective performance, and good classification performance of the ensemble does not necessarily imply good selective classification performance.

7 Discussion

The experimental results suggest that the assumptions may capture the essential reason for the better performance of the ensemble than the individual selective classifier. According to the analysis result, the assumptions are a sufficient condition for good performance of the ensemble, but no analysis

shows that they are a necessary condition. However, the experiments show that the assumptions might be a necessary condition, since the ensemble only does not perform well on MRPC, the only dataset where the assumptions do not hold (see Figure 2 and Figure 1). Thus, the assumptions might capture the essential reason why the ensemble works.

Based on this work, there are two possible directions for future work. First, the theoretical result in this paper may motivate the analysis of the randomization-based ensemble for *standard classification* in *practical settings*. The previous analyses of the randomization-based ensemble for standard classification need some impractical assumptions (Tumer & Ghosh, 1996; Fumera & Roli, 2005). On the contrary, the assumptions in this paper are a good approximation of practical settings (see Section 6.1), and more importantly, the standard classification is a special case of selective classification, i.e. selective classification with coverage of 1. Therefore, our analysis (although it does not cover the case of coverage of 1, i.e. the case of standard classification) may motivate the analysis of the randomization-based ensemble for standard classification in practical settings. Second, it is interesting to relax the assumption of the idealization of the definite samples and then show a similar theoretical result. The idealization is a strong assumption that all member models provide the same predictive probability distribution on definite samples. Although it may be a good approximation of practical setting, it is unrealistic anyway. However, we believe that a similar theoretical result holds in the absence of the idealization, given a set of assumptions similar to the assumptions in this paper.

8 Conclusion

We prove that under some assumptions, the ensemble has a lower selective risk than the individual model under a range of coverage. Although the metrics of selective classification are non-convex, we complete the proof with the help of several assumptions motivated by empirical observations. The assumptions and the result are well supported by the experimental results on multiple datasets of image classification tasks and text classification tasks. A surprising empirical result is that two simple methods, SR ensemble and its variant Reg-curr ensemble, (which can be summarized as *the ensemble models with maximum probability as confidence*) are state-of-the-art selective classifiers.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing*, 2005.

Giorgio Fumera and Fabio Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE transactions on pattern analysis and machine intelligence*, 27 (6):942–956, 2005.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representation*, 2017.

Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. *Advances in Neural Information Processing Systems*, 2020.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, 1994.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 2017.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Kagan Tumer and Joydeep Ghosh. Analysis of decision boundaries in linearly combined neural classifiers. *Pattern recognition*, 29(2):341–348, 1996.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1040–1051, 2021.

Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.

A RELATIONSHIP BETWEEN SELECTIVE CLASSIFICATION AND OOD DETECTION

Figure 3 shows the relationship between selective classification and OOD detection (or open set recognition, novelty detection).

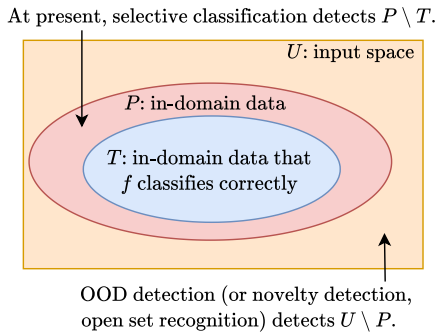

Figure 3: Relationship between selective classification and OOD detection (or open set recognition, novelty detection).

B PROOFS

We prove Proposition 1 in Section B.1 and prove Theorem 2 in Section B.2. The road map of the complete proof of the final result, i.e. Theorem 2, is shown in Figure 4.

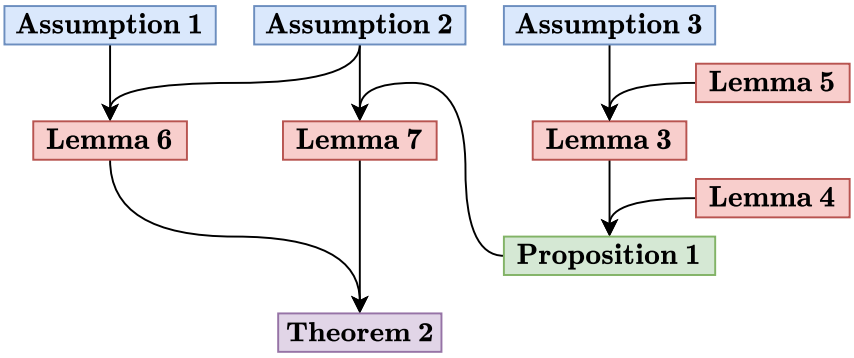

Figure 4: The road map of the proof of Theorem 2

B.1 PROOF OF PROPOSITION 1

To prove Proposition 1, we show that for ambiguous samples, the ensemble is conservative in providing the predictive probability of each class at first. By 'conservative', we mean that the predictive probability provided by the model is modest, i.e. away from the extreme values 0 and 1. Formally, we have the following lemma (proof is provided in Section B.1.1).

Lemma 3. *If Assumption 3 holds, then*

$$p_{\Pi^k_{\text{ens}}}(0|A) = 0, \tag{10}$$

$$p_{\Pi^k_{\text{ens}}}(1|A) = 0, \tag{11}$$

and

$$p_{\Pi^k_{\text{ens}}}(\pi^k_{\text{ens}}|A) = O((\pi^k_{\text{ens}})^{M-1})\ (\pi^k_{\text{ens}} \to 0^+), \tag{12}$$

$$p_{\Pi^k_{\text{ens}}}(\pi^k_{\text{ens}}|A) = O((1-\pi^k_{\text{ens}})^{M-1})\ (\pi^k_{\text{ens}} \to 1^-), \tag{13}$$

where the notation follows that of Assumption 3, and Π^k_{ens} *is the predictive probability of the ensemble for class* k.

Secondly, we show the relation of the PDF of the confidence to those of predictive probabilities. Note that for an SR model, its confidence is the maximum of predictive probabilities of all classes. With the following lemma (proof is provided in Section B.1.2), we can bound the PDF of the confidence by those of predictive probabilities.

Lemma 4. *Let* Π^k $(1 \leq k \leq K)$ *be* K *continuous random variables, and* $\mathrm{C} := \max_k \Pi^k$. *Then we have*

$$p_{\mathrm{C}}(\kappa) \leq \sum_{k=1}^{K} p_{\Pi^k}(\kappa). \tag{14}$$

Finally, since $\mathrm{C}_{\text{ens}} = \max_k \Pi^k_{\text{ens}}$, Lemma 3 and Lemma 4 derive that when $\kappa_{\text{ens}} \to 1^-$,

$$p_{\mathrm{C}_{\text{ens}}}(\kappa_{\text{ens}}|A) \leq \sum_{k=1}^{K} p_{\Pi^k_{\text{ens}}}(\kappa_{\text{ens}}|A) = O((1-\kappa_{\text{ens}})^{M-1}),$$

and thus Proposition 1 holds.

B.1.1 Proof of Lemma 3

We first derive the PDF of the average of multiple continuous random variables in terms of the PDFs of these random variables.

Lemma 5. *Let* $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_M$ *be* M *continuous random variables, and their average is* $\mathrm{X}_{\text{avg}} := \frac{1}{M}\sum_{i=1}^{M} \mathrm{X}_i$. *Then the PDF of* X_{avg} *is*

$$p_{\mathrm{X}_{\text{avg}}}(x_{\text{avg}}) = M \int_{\mathbb{R}^{M-1}} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_{M-1} \cdot p_{\vec{\mathrm{X}}}(x_1, x_2, \ldots, x_{M-1}, Mx_{\text{avg}} - \sum_{i=1}^{M-1} x_i), \tag{15}$$

where $p_{\vec{\mathrm{X}}}$ *is* $p_{\mathrm{X}_1,\mathrm{X}_2,\ldots,\mathrm{X}_M}$ *for short.*

Proof. The distribution function of X_{avg} is

$$\begin{aligned} F_{\mathrm{X}_{\text{avg}}}(x_{\text{avg}}) &= \int_{\sum_i x_i \leq M x_{\text{avg}}} \mathrm{d}x_1 \cdots \mathrm{d}x_{M-1} \mathrm{d}x_M \cdot p_{\vec{\mathrm{X}}}(x_1, \ldots, x_M) \\ &= \int_{\mathbb{R}^{M-1}} \mathrm{d}x_1 \cdots \mathrm{d}x_{M-1} \int_{-\infty}^{M x_{\text{avg}} - \sum_{i=1}^{M-1} x_i} \mathrm{d}x_M \cdot p_{\vec{\mathrm{X}}}(x_1, \cdots, x_M). \end{aligned}$$

Let $x_M = u - \sum_{i=1}^{M-1} x_i$, then the integral above is equal to

$$\begin{aligned} &\int_{\mathbb{R}^{M-1}} \mathrm{d}x_1 \cdots \mathrm{d}x_{M-1} \int_{-\infty}^{M x_{\text{avg}}} \mathrm{d}u \cdot p_{\vec{\mathrm{X}}}(x_1, \ldots, x_{M-1}, u - \sum_{i=1}^{M-1} x_i) \\ =& \int_{-\infty}^{M x_{\text{avg}}} \mathrm{d}u \int_{\mathbb{R}^{M-1}} \mathrm{d}x_1 \cdots \mathrm{d}x_{M-1} \cdot p_{\vec{\mathrm{X}}}(x_1, \ldots, x_{M-1}, u - \sum_{i=1}^{M-1} x_i). \end{aligned} \tag{16}$$

The PDF of X_{avg} is the derivative of $F_{\mathrm{X}_{\text{avg}}}$, which combined with (16) derives

$$\begin{aligned} p_{\mathrm{X}_{\text{avg}}}(x_{\text{avg}}) =& F'_{\mathrm{X}_{\text{avg}}}(x_{\text{avg}}) \\ =& \frac{\mathrm{d}(M x_{\text{avg}})}{\mathrm{d}x_{\text{avg}}} \cdot \frac{\mathrm{d}F_{\mathrm{X}_{\text{avg}}}}{\mathrm{d}(M x_{\text{avg}})} \\ =& M \int_{\mathbb{R}^{M-1}} \mathrm{d}x_1 \cdots \mathrm{d}x_{M-1} \cdot p_{\vec{\mathrm{X}}}(x_1, \ldots, x_{M-1}, M x_{\text{avg}} - \sum_{i=1}^{M-1} x_i), \end{aligned}$$

which is exactly (15). □

Based on Lemma 5, we prove Lemma 3 as follows.

Proof. With Lemma 5 applied to Π_i^k, $1 \leq i \leq M$, we have

$$p_{\Pi_{\text{ens}}^k}(\pi_{\text{ens}}^k|A) = M \int_{\mathbb{R}^{M-1}} \mathrm{d}\pi_1^k \cdots \mathrm{d}\pi_{M-1}^k \cdot p_{\vec{\Pi}^k}(\pi_1^k, \ldots, \pi_{M-1}^k, M\pi_{\text{ens}}^k - \sum_{i=1}^{M-1} \pi_i^k|A), \tag{17}$$

where $p_{\vec{\Pi}^k}$ is $p_{\Pi_1^k,\ldots,\Pi_M^k}$ for short. The integrand in the right-hand side of (17) is non-zero requires

$$\begin{cases} 0 \leq \pi_i^k \leq 1, i = 1, 2, \ldots, M-1 \\ 0 \leq M\pi_{\text{ens}}^k - \sum_{i=1}^{M-1} \pi_i^k \leq 1 \end{cases}. \tag{18}$$

Firstly, we prove (10) and (12). When $0 \leq \pi_{\text{ens}}^k \leq \frac{1}{M}$, it easy to verify that (18) is equivalent to

$$\begin{cases} 0 \leq \pi_1^k \leq M\pi_{\text{ens}}^k \\ 0 \leq \pi_2^k \leq M\pi_{\text{ens}}^k - \pi_1^k \\ \cdots \\ 0 \leq \pi_i^k \leq M\pi_{\text{ens}}^k - \pi_1^k - \cdots - \pi_{i-1}^k \\ \cdots \\ 0 \leq \pi_{M-1}^k \leq M\pi_{\text{ens}}^k - \pi_1^k - \cdots - \pi_{M-2}^k \end{cases}. \tag{19}$$

Thus, (17) transforms into

$$\begin{aligned} p_{\Pi_{\text{ens}}^k}(\pi_{\text{ens}}^k|A) = & M \int_0^{M\pi_{\text{ens}}^k} \mathrm{d}\pi_1^k \cdots \int_0^{M\pi_{\text{ens}}^k - \sum_{j=1}^{i-1} \pi_j^k} \mathrm{d}\pi_i^k \cdots \int_0^{M\pi_{\text{ens}}^k - \sum_{j=1}^{M-2} \pi_j^k} \mathrm{d}\pi_{M-1}^k \\ & \cdot p_{\vec{\Pi}^k}(\pi_1^k, \ldots, \pi_{M-1}^k, M\pi_{\text{ens}}^k - \sum_{i=1}^{M-1} \pi_i^k|A). \end{aligned}$$

Considering that $p_{\vec{\Pi}^k}(\cdot|A)$ is bounded, as Assumption 3 claims, let B be one of its upper bounds. Then we have

$$\begin{aligned} p_{\Pi_{\text{ens}}^k}(\pi_{\text{ens}}^k|A) & \leq M \int_0^{M\pi_{\text{ens}}^k} \mathrm{d}\pi_1^k \cdots \int_0^{M\pi_{\text{ens}}^k - \sum_{j=1}^{i-1} \pi_j^k} \mathrm{d}\pi_i^k \cdots \int_0^{M\pi_{\text{ens}}^k - \sum_{j=1}^{M-2} \pi_j^k} \mathrm{d}\pi_{M-1}^k B \\ & \leq M \int_0^{M\pi_{\text{ens}}^k} \mathrm{d}\pi_1^k \cdots \int_0^{M\pi_{\text{ens}}^k} \mathrm{d}\pi_i^k \cdots \int_0^{M\pi_{\text{ens}}^k} \mathrm{d}\pi_{M-1}^k B \\ & = MB \cdot \int_0^{M\pi_{\text{ens}}^k} \mathrm{d}\pi_1^k \cdots \int_0^{M\pi_{\text{ens}}^k} \mathrm{d}\pi_{M-1}^k \\ & = MB \cdot (M\pi_{\text{ens}}^k)^{M-1} \\ & = M^M B \cdot (\pi_{\text{ens}}^k)^{M-1}, \end{aligned} \tag{20}$$

which directly derives (10) and (12) (note that the PDF is non-negative).

Secondly, we prove (11) and (13). These two equations could be derived like (10) and (12). However, here we take another way that uses a little trick to simplify the proof. Let Π_i^k and Π_{ens}^k be the corresponding random variables of π_i^k and π_{ens}^k respectively, and $\mathrm{U}_i = 1 - \Pi_i^k$, $\mathrm{U}_{\text{ens}} = \frac{1}{M}\sum_{i=1}^M \mathrm{U}_i = 1 - \Pi_{\text{ens}}^k$. Applying (20) to U_i and U_{ens}, we get that when $0 \leq u_{\text{ens}} \leq \frac{1}{M}$,

$$p_{\mathrm{U}_{\text{ens}}}(u_{\text{ens}}|A) \leq M^M B \cdot u_{\text{ens}}^{M-1}. \tag{21}$$

It is easy to see that $p_{\mathrm{U}_{\text{ens}}}(u_{\text{ens}}|A) = p_{\Pi_{\text{ens}}^k}(1 - u_{\text{ens}}|A)$. Sticking this into (21), we have

$$p_{\Pi_{\text{ens}}^k}(1 - u_{\text{ens}}|A) \leq M^M B \cdot u_{\text{ens}}^{M-1},$$

when $0 \leq u_{\text{ens}} \leq \frac{1}{M}$. With the $(1 - u_{\text{ens}})$ in the equation above replaced with π_{ens}^k, we have

$$p_{\Pi_{\text{ens}}^k}(\pi_{\text{ens}}^k|A) \leq M^M B \cdot (1 - \pi_{\text{ens}}^k)^{M-1},$$

when $1 - \frac{1}{M} \leq \pi_{\text{ens}}^k \leq 1$, which directly derives (11) and (13). □

B.1.2 PROOF OF LEMMA 4

Proof. First of all, we prove $\forall \kappa_1, \kappa_2, \kappa_1 < \kappa_2$,

$$F_{\mathrm{C}}(\kappa_2) - F_{\mathrm{C}}(\kappa_1) \leq \sum_{k=1}^{K} F_{\Pi^k}(\kappa_2) - F_{\Pi^k}(\kappa_1) \tag{22}$$

It is easy to see that

$$F_{\mathrm{C}}(\kappa) = F_{\Pi^1,\ldots,\Pi^K}(\kappa,\ldots,\kappa) = \int_{(-\infty,\kappa]^K} \mathrm{d}\pi^1 \cdots \mathrm{d}\pi^K \; p_{\Pi^1,\ldots,\Pi^K}(\pi^1,\cdots,\pi^K),$$

so the left-hand side of (22) is

$$\begin{aligned}
&\int_{(-\infty,\kappa_2]^K} \mathrm{d}\pi^1 \cdots \mathrm{d}\pi^K \; p_{\Pi^1,\ldots,\Pi^K}(\pi^1,\cdots,\pi^K) \\
-&\int_{(-\infty,\kappa_1]^K} \mathrm{d}\pi^1 \cdots \mathrm{d}\pi^K \; p_{\Pi^1,\ldots,\Pi^K}(\pi^1,\cdots,\pi^K) \\
=&\int_{(-\infty,\kappa_2]^K \setminus (-\infty,\kappa_1]^K} \mathrm{d}\pi^1 \cdots \mathrm{d}\pi^K \; p_{\Pi^1,\ldots,\Pi^K}(\pi^1,\cdots,\pi^K),
\end{aligned} \tag{23}$$

where the last equality is due to $(-\infty,\kappa_1] \subset (-\infty,\kappa_2]$, and the right-hand side of (22) is

$$\begin{aligned}
&\sum_{k=1}^{K} \int_{[\kappa_1,\kappa_2]} \mathrm{d}\pi^k \; p_{\Pi^k}(\pi^k) \\
=&\sum_{k=1}^{K} \int_{\mathbb{R}^{k-1}\times[\kappa_1,\kappa_2]\times\mathbb{R}^{K-k}} \mathrm{d}\pi^1 \cdots \mathrm{d}\pi^K \cdot p_{\Pi^1,\ldots,\Pi^K}(\pi^1,\cdots,\pi^K) \\
\geq&\int_{\bigcup_{k=1}^{K} \mathbb{R}^{k-1}\times[\kappa_1,\kappa_2]\times\mathbb{R}^{K-k}} \mathrm{d}\pi^1 \cdots \mathrm{d}\pi^K \cdot p_{\Pi^1,\ldots,\Pi^K}(\pi^1,\cdots,\pi^K),
\end{aligned} \tag{24}$$

where the last inequality is because $\mathbb{R}^{k-1} \times [\kappa_1,\kappa_2] \times \mathbb{R}^{K-k}$ for different k, $1 \leq k \leq K$, may have an intersection. To prove (22), we only need to prove that the right-hand side of (23) is less than or equal to the right-hand side of (24), which is equivalent to prove

$$(-\infty,\kappa_2]^K \setminus (-\infty,\kappa_1]^K \subset \bigcup_{k=1}^{K} \mathbb{R}^{k-1} \times [\kappa_1,\kappa_2] \times \mathbb{R}^{K-k}. \tag{25}$$

Now we prove (25). $\forall (\pi^1,\ldots,\pi^K) \in (-\infty,\kappa_2]^K \setminus (-\infty,\kappa_1]^K$, we have

$$\forall k, 1 \leq k \leq K, \pi^k \leq \kappa_2, \tag{26}$$

$$\exists k_0, 1 \leq k_0 \leq K, \pi^{k_0} > \kappa_1, \tag{27}$$

where (27) is because if all π^k is less than or equal to κ_1 instead, then $(\pi^1,\ldots,\pi^K) \in (-\infty,\kappa_1]^K$, which contradicts with $(\pi^1,\ldots,\pi^K) \in (-\infty,\kappa_2]^K \setminus (-\infty,\kappa_1]^K$. Thus, $\pi^{k_0} \in [\kappa_1,\kappa_2]$, so

$$(\pi^1,\ldots,\pi^K) \in \mathbb{R}^{k_0-1} \times [\kappa_1,\kappa_2] \times \mathbb{R}^{K-k_0} \subset \bigcup_{k=1}^{K} \mathbb{R}^{k-1} \times [\kappa_1,\kappa_2] \times \mathbb{R}^{K-k},$$

which is exactly (25), and therefore (22) is proved.

With (22) and the definition of derivatives, it is easy to see that $F'_{\mathrm{C}}(\kappa) \leq \sum_{k=1}^{K} F'_{\Pi^k}(\kappa)$, which is equivalent to $p_{\mathrm{C}}(\kappa) \leq \sum_{k=1}^{K} p_{\Pi^k}(\kappa)$. Thus, Lemma 4 is proved. □

B.2 PROOF OF THEOREM 2

To prove Theorem 2, we show the following lemmas first.

Lemma 6. *If Assumption 1-2 holds, then for any individual SR model,*

$$\lim_{\tau\to 1^-}\Pr(Err|\mathbf{C}\geq\tau) > \lim_{\tau\to 1^-}\Pr(Err|D,\mathbf{C}\geq\tau), \tag{28}$$

where the notation follows those of Assumption 1-2.

Proof. Using Bayes' rule, we have

$$\Pr(A|\mathbf{C}\geq\tau) = \frac{\Pr(\mathbf{C}\geq\tau|A)\Pr(A)}{\Pr(\mathbf{C}\geq\tau|A)\Pr(A)+\Pr(\mathbf{C}\geq\tau|D)\Pr(D)}. \tag{29}$$

Using L'Hospital's rule, we have

$$\begin{aligned}\lim_{\tau\to 1^-}\frac{\Pr(\mathbf{C}\geq\tau|D)}{\Pr(\mathbf{C}\geq\tau|A)} &= \lim_{\tau\to 1^-}\frac{\int_\tau^1 p_{\mathbf{C}}(\kappa|D)\mathrm{d}\kappa}{\int_\tau^1 p_{\mathbf{C}}(\kappa|A)\mathrm{d}\kappa}\\ &= \lim_{\tau\to 1^-}\frac{-p_{\mathbf{C}}(\tau|D)}{-p_{\mathbf{C}}(\tau|A)}\\ &= \lim_{\tau\to 1^-}\frac{p_{\mathbf{C}}(\tau|D)}{p_{\mathbf{C}}(\tau|A)}\\ &>0,\end{aligned} \tag{30}$$

where the last inequality is due to Assumption 2. Combining this with (29), we have

$$\begin{aligned}\lim_{\tau\to 1^-}\Pr(A|\mathbf{C}\geq\tau) =& \frac{\Pr(A)}{\Pr(A)+\Pr(D)\lim_{\tau\to 1^-}\frac{\Pr(\mathbf{C}\geq\tau|D)}{\Pr(\mathbf{C}\geq\tau|A)}},\\ =& \frac{\Pr(A)}{\Pr(A)+\Pr(D)\lim_{\tau\to 1^-}\frac{p_{\mathbf{C}}(\tau|D)}{p_{\mathbf{C}}(\tau|A)}},\\ >&0.\end{aligned} \tag{31}$$

Now we derive (28).

$$\begin{aligned}\Pr(Err|\mathbf{C}\geq\tau) =& \Pr(Err,A|\mathbf{C}\geq\tau)+\Pr(Err,D|\mathbf{C}\geq\tau)\\ =& \Pr(Err|A,\mathbf{C}\geq\tau)\Pr(A|\mathbf{C}\geq\tau)+\Pr(Err|D,\mathbf{C}\geq\tau)\Pr(D|\mathbf{C}\geq\tau)\\ =& \Pr(Err|A,\mathbf{C}\geq\tau)\Pr(A|\mathbf{C}\geq\tau)+\Pr(Err|D,\mathbf{C}\geq\tau)[1-\Pr(A|\mathbf{C}\geq\tau)]\\ =& [\Pr(Err|A,\mathbf{C}\geq\tau)-\Pr(Err|D,\mathbf{C}\geq\tau)]\Pr(A|\mathbf{C}\geq\tau)+\Pr(Err|D,\mathbf{C}\geq\tau).\end{aligned}$$

According to the equation above, we have

$$\begin{aligned}\lim_{\tau\to 1^-}\Pr(Err|\mathbf{C}\geq\tau) =& [\lim_{\tau\to 1^-}\Pr(Err|A,\mathbf{C}\geq\tau)-\lim_{\tau\to 1^-}\Pr(Err|D,\mathbf{C}\geq\tau)]\\ &\cdot\lim_{\tau\to 1^-}\Pr(A|\mathbf{C}\geq\tau)+\lim_{\tau\to 1^-}\Pr(Err|D,\mathbf{C}\geq\tau)\end{aligned} \tag{32}$$

Due to (31) and Assumption 1, the first term of the equation above is positive, so (28) is derived. □

Lemma 7. *If Assumption 2-3 hold, then*

$$\lim_{\tau_{\text{ens}}\to 1^-}\Pr(A|\mathbf{C}_{\text{ens}}\geq\tau_{\text{ens}})=0, \tag{33}$$

where the notation follows those of Assumption 2-3.

Proof. Similar to (31), with Bayes' rule and L'Hospital's rule, we can derive that

$$\begin{aligned}\lim_{\tau_{\text{ens}}\to 1^-}\Pr(A|\mathbf{C}_{\text{ens}}\geq\tau_{\text{ens}}) =& \frac{\Pr(A)\lim_{\tau_{\text{ens}}\to 1^-}\frac{\Pr(\mathbf{C}_{\text{ens}}\geq\tau_{\text{ens}}|A)}{\Pr(\mathbf{C}_{\text{ens}}\geq\tau_{\text{ens}}|D)}}{\Pr(A)\lim_{\tau_{\text{ens}}\to 1^-}\frac{\Pr(\mathbf{C}_{\text{ens}}\geq\tau_{\text{ens}}|A)}{\Pr(\mathbf{C}_{\text{ens}}\geq\tau_{\text{ens}}|D)}+\Pr(D)},\\ =& \frac{\Pr(A)\lim_{\tau_{\text{ens}}\to 1^-}\frac{p_{\mathbf{C}_{\text{ens}}}(\tau_{\text{ens}}|A)}{p_{\mathbf{C}_{\text{ens}}}(\tau_{\text{ens}}|D)}}{\Pr(A)\lim_{\tau_{\text{ens}}\to 1^-}\frac{p_{\mathbf{C}_{\text{ens}}}(\tau_{\text{ens}}|A)}{p_{\mathbf{C}_{\text{ens}}}(\tau_{\text{ens}}|D)}+\Pr(D)},\end{aligned} \tag{34}$$

where $\mathrm{C}_{\mathrm{ens}}$ is the confidence score of the ensemble. Because for a definite sample, the confidence score of the ensemble is equal to that of the individual model, we have

$$\lim_{\tau_{\mathrm{ens}}\to 1^-} \frac{p_{\mathrm{C}_{\mathrm{ens}}}(\tau_{\mathrm{ens}}|A)}{p_{\mathrm{C}_{\mathrm{ens}}}(\tau_{\mathrm{ens}}|D)} = \lim_{\tau_{\mathrm{ens}}\to 1^-} \frac{p_{\mathrm{C}_{\mathrm{ens}}}(\tau_{\mathrm{ens}}|A)}{p_{\mathrm{C}}(\tau_{\mathrm{ens}}|D)} = 0,$$

where the last equality is due to Proposition 1 and Assumption 2. Substituting this to (34), we obtain (33). □

Now we provide the proof of the Theorem 2.

Proof. First, for the convenience of the proof, given an SR model (f, g), we define a threshold-to-coverage function $\rho_{(f,g)}$ of (f, g) that maps the confidence threshold to the corresponding coverage,

$$\rho_{(f,g)} : (0, 1) \to (0, 1), \tau \mapsto \phi(f, g; \tau).$$

Second, we prove that $\exists \delta \in (0, 1), \forall \tau_{\mathrm{ens}} \in (1 - \delta, 1)$,

$$\Pr(A|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) - \Pr(Err_{\mathrm{ind}}|\mathrm{C} \geq \tau) + \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) < 0, \tag{35}$$

$$\tau = \max \rho^{-1} \circ \rho_{\mathrm{ens}}(\tau_{\mathrm{ens}}) \tag{36}$$

where C is the confidence score of the individual SR model, Err_{ind} is the event that the individual model makes an error prediction, Err_{ens} is the event that the ensemble makes an error prediction, ρ and ρ_{ens} are the threshold-to-coverage functions of the individual model and the ensemble respectively. Note that the symbol ρ^{-1} denotes the preimage under ρ, rather than the inverse function of ρ. Because when $\tau_{ens} \to 1^-$, the coverage of the ensemble approaches 0, and due to the coverage of the individual model with the confidence threshold of τ is equal to the coverage of the ensemble ($\rho(\tau) = \rho_{\mathrm{ens}}(\tau_{\mathrm{ens}})$), we have $\tau \to 1^-$. Thus,

$$\lim_{\tau_{\mathrm{ens}}\to 1^-} \Pr(Err_{\mathrm{ind}}|\mathrm{C} \geq \tau) = \lim_{\tau\to 1^-} \Pr(Err_{\mathrm{ind}}|\mathrm{C} \geq \tau). \tag{37}$$

In addition, for definite samples, the confidence score of the ensemble and that of the individual model are the same, and the ensemble and the individual model make error predictions on the same set of samples, i.e. $Err_{\mathrm{ind}} = Err_{\mathrm{ens}}$, so

$$\begin{aligned}
\lim_{\tau_{\mathrm{ens}}\to 1^-} \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) &= \lim_{\tau_{\mathrm{ens}}\to 1^-} \Pr(Err_{\mathrm{ens}}|D, \mathrm{C} \geq \tau_{\mathrm{ens}}) \\
&= \lim_{\tau\to 1^-} \Pr(Err_{\mathrm{ens}}|D, \mathrm{C} \geq \tau) \\
&= \lim_{\tau\to 1^-} \Pr(Err_{\mathrm{ind}}|D, \mathrm{C} \geq \tau),
\end{aligned} \tag{38}$$

where the second equality is just a variable substitution. Finally, we have

$$\begin{aligned}
&\lim_{\tau_{\mathrm{ens}}\to 1^-} [\Pr(A|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) - \Pr(Err_{\mathrm{ind}}|\mathrm{C} \geq \tau) + \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}})] \\
=&\lim_{\tau_{\mathrm{ens}}\to 1^-} [0 - \Pr(Err_{\mathrm{ind}}|\mathrm{C} \geq \tau) + \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}})] \\
=&- \lim_{\tau\to 1^-} \Pr(Err_{\mathrm{ind}}|\mathrm{C} \geq \tau) + \lim_{\tau\to 1^-} \Pr(Err_{\mathrm{ind}}|D, \mathrm{C} \geq \tau) \\
<&0,
\end{aligned} \tag{39}$$

where the first equality is due to Lemma 7, the second equality is due to (37) and (38), and the last inequality is due to Lemma 6. Thus, with (39), it is easy to see that (35) holds.

Third, we have

$$\begin{aligned}
\Pr(Err_{\mathrm{ens}}|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) =& \Pr(Err_{\mathrm{ens}}, A|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) + \Pr(Err_{\mathrm{ens}}, D|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) \\
=& \Pr(Err_{\mathrm{ens}}|A, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}})\Pr(A|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) \\
&+ \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}})\Pr(D|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) \\
=& \Pr(Err_{\mathrm{ens}}|A, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}})\Pr(A|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) \\
&+ \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}})[1 - \Pr(A|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}})] \\
=& [\Pr(Err_{\mathrm{ens}}|A, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) - \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}})] \\
&\cdot \Pr(A|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) + \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) \\
\leq& \Pr(A|\mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}) + \Pr(Err_{\mathrm{ens}}|D, \mathrm{C}_{\mathrm{ens}} \geq \tau_{\mathrm{ens}}),
\end{aligned} \tag{40}$$

where the last inequality is because any probability is in [0, 1]. Combining this with (35), we have $\exists\delta \in (0,1), \forall\tau_{\text{ens}} \in (1-\delta, 1)$,

$$\begin{aligned}\Pr(Err_{\text{ens}}|\mathbf{C}_{\text{ens}} \geq \tau_{\text{ens}}) \leq & \Pr(A|\mathbf{C}_{\text{ens}} \geq \tau_{\text{ens}}) + \Pr(Err_{\text{ens}}|D, \mathbf{C}_{\text{ens}} \geq \tau_{\text{ens}}) \\ < & \Pr(Err_{\text{ind}}|\mathbf{C} \geq \tau) - \Pr(Err_{\text{ens}}|D, \mathbf{C}_{\text{ens}} \geq \tau_{\text{ens}}) \\ & + \Pr(Err_{\text{ens}}|D, \mathbf{C}_{\text{ens}} \geq \tau_{\text{ens}}) \\ = & \Pr(Err_{\text{ind}}|\mathbf{C} \geq \tau). \end{aligned} \tag{41}$$

$\Pr(Err_{\text{ens}}|\mathbf{C}_{\text{ens}} \geq \tau_{\text{ens}})$ is the selective risk of the ensemble given the confidence threshold of τ_{ens} (or given coverage of $\rho_{\text{ens}}(\tau_{\text{ens}})$), and $\Pr(Err_{\text{ind}}|\mathbf{C} \geq \tau)$ is the selective risk of the individual model given the confidence threshold of τ (or given coverage of $\rho(\tau) = \rho_{\text{ens}}(\tau_{\text{ens}})$). These two selective risks are under the same coverage $\rho_{\text{ens}}(\tau_{\text{ens}})$. Thus, (41) is equivalent to that $\exists\delta \in (0,1)$, $\forall\tau_{\text{ens}} \in (1-\delta, 1)$, the ensemble has a lower selective risk than the individual model given the coverage of $\phi = \rho_{\text{ens}}(\tau_{\text{ens}})$. This can be simplified as $\exists\delta \in (0,1)$, $\forall\phi \in (0, \rho_{\text{ens}}(1-\delta))$, the ensemble has a lower selective risk than the individual model given the coverage of ϕ. □

C Details of Experiments

C.1 Datasets

The experiments were conducted on multiple data sets of image classification and text classification. The datasets of image classification are CIFAR-10, CIFAR-100, (Krizhevsky, 2009) and SVHN (Netzer et al., 2011), whose image sizes are all $32 \times 32 \times 3$ pixels. The datasets of text classification are MRPC (Dolan & Brockett, 2005), MNLI (Williams et al., 2018) and QNLI (Wang et al., 2018). The task of MRPC is to judge whether two paragraphs of text are semantically equivalent or not. MNLI's task is to judge the inferential relationship of sentences (three categories). The task of QNLI is to determine whether a paragraph has the answer to a given question. The sizes of the training set, development set, and test set of each data set used in experiments are shown in Table 2. MNLI's development set and test set are divided into *matched* and *mismatched* parts. In the table, (m) represents matched, and (mm) represents mismatched. The matched parts are sampled from the same source as the training set, while the mismatched parts are sampled from different sources. Current selective classification only considers test samples from the same distribution as the training set, so only the matched parts are used in experiments. In addition, test sets of MRPC, QNLI, and MNLI are not accessible, so we use their development sets as test sets. According to Liu et al. (2019); Huang et al. (2020), since CIFAR-10, CIFAR-100 and SVHN originally had no development set, their development sets were 2000 samples randomly divided from corresponding test sets.

Table 2: Sizes of training sets, development sets, and test sets for each dataset used in experiments

Datasets	Training Set	Development Set	Test Set	Number of Classes
CIFAR-10	50.0k		10.0k	10
CIFAR-100	50.0k		10.0k	100
SVHN	73.3k		26.0k	10
MRPC	3.7k	0.4k	1.7k	2
QNLI	104.7k	5.5k	5.5k	2
MNLI	392.7k	9.8k (m)/ 9.8k(mm)	9.8k(m)/9.8k(mm)	3

C.2 Model Implementations and Training Procedures

For image classification, the backbone model is VGG-16 (Simonyan & Zisserman, 2014) with Dropout (Srivastava et al., 2014), batch normalization (Ioffe & Szegedy, 2015). It is trained in the same way as Huang et al. (2020). The model is optimized using SGD with an initial learning rate of 0.1 (the learning rate decays by half in every 25 epochs), the momentum of 0.9, weight decay of 0.0005, batch size of 128, and a total training epoch of 300. Data preprocessing includes

data augmentation (random cropping and flip) and normalization. The implementations of the backbone model and data preprocessing are based on the official open-sourced implementation of SAT to ensure a fair comparison.

For text classification, the backbone model of selective classifiers is BERT-base (Devlin et al., 2019). Pretrained BERT-base is provided by the Huggingface Transformer Library (Wolf et al., 2020). It is trained/fine-tuned in the same way as Xin et al. (2021). The model is trained/fine-tuned using AdamW (Loshchilov & Hutter, 2017) for 3 epochs, with a learning rate of 2×10^{-5}, batch size of 32, and the maximum input sequence length of 128.

C.3 Hyperparameters of Selective Classifiers

For the hyperparameter c of SN, we choose $c = 0.9$ for evaluating its selective risk given the coverage of 90%. The results of SN are reported in Appendix D. For the hyperparameter o of Gambler, we tune o on validation sets in the same way as Liu et al. (2019). For the hyperparameter α of SAT, we set $\alpha = 0.99$, the same as Huang et al. (2020). For the hyperparameter λ of Reg-curr, we set $\lambda = 0.05$.

D Empirical results of SN

With the hyperparameter c of 0.9, we report the selective risks given the coverage of 90% of SN ensembles and the individual SN in Figure 5. The coverage is set to 90% because the target coverage of the SN is c (Geifman & El-Yaniv, 2019)), and c is 0.9 in our experiments. The results show that each ensemble of SN have a lower selective risk than the individual SN.

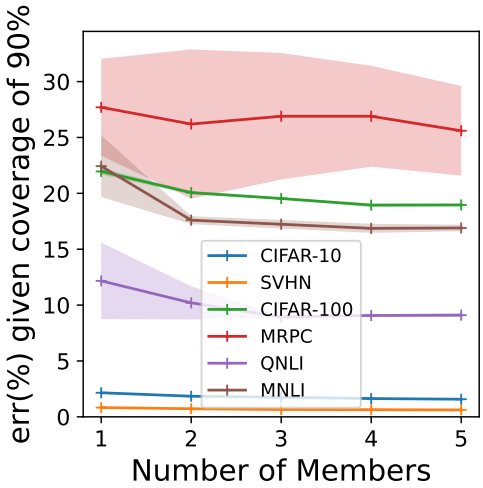

Figure 5: Selective risks of SN ensembles (with 2 to 5 members) and the individual SN (with only 1 member) given the coverage of 90%

E Further Properties of Selective Classifier Ensemble

E.1 The Effect of Number of Members on Selective Classifier Ensemble

We evaluate AURCs of SR ensemble, Gambler ensemble, and SAT ensemble of different numbers of members on CIFAR10, and find that an ensemble with more members has better performance, but is less efficient. The results are shown in Figure 6. In most cases, the AURC on the test set of CIFAR-10 decreases as the number of members in the ensemble increases. In addition, as the number of members in the ensemble grows, the effect of adding one member drops. On the one

hand, the result shows that an ensemble with a small number of members has good selective classification performance. On the other hand, it indicates that when the number of member models is large, increasing the number of members to improve the performance of the selective classification ensemble is not efficient.

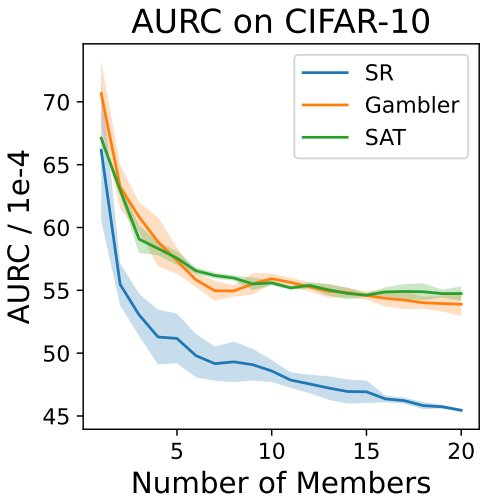

Figure 6: The AURCs on the test set of CIFAR-10 of SR ensemble, Gambler ensemble, and SAT ensemble of different numbers of members

E.2 Good Classification Performance Does Not Imply Good Selective Classification Performance

It is well known that the ensemble has better classification performance than an individual model, but this does not guarantee a better selective classification performance of the ensemble. To demonstrate this, we design an SR model with a big backbone, and show that it has as good classification performance as an SR ensemble with a normal backbone but worse selective classification performance than an SR model with a normal backbone. The big backbone is designed to have twice as many filters in every convolutional layer and neurons in every fully connected hidden layer as those of the standard VGG-16, which is therefore called *Big VGG-16*. It is easy to see that its number of parameters is approximately $2^2 = 4$ times as many as that of standard VGG-16. We train an SR ensemble of 4 VGG-16s and an SR model with a backbone of Big VGG-16 on CIFAR-10 and show the evaluation results in Figure 7 and Table 3. Figure 7 shows that when coverage is high, the ensemble and the big individual model have similar selective risks, and especially, the classification error rates (i.e. selective risk of 100% coverage) of the ensemble and the big individual model are similar. However, when coverage is low, the big individual model has significantly higher selective risk than the ensemble. Table 3 shows that the AURC of Big VGG-16 is much higher than the ensemble of 4 VGG-16s and even higher than SR. In summary, we show that a selective classifier with a good classification performance is not guaranteed to have good selective classification performance, so the good selective classification performance of the ensemble is not a trivial result of its good classification performance.

Table 3: The AURCs$(/10^{-4})$ of Big VGG-16, a vanilla VGG-16, and the ensemble of 4 VGG-16s on CIFAR-10. The best entries are marked in bold.

Dataset	Big VGG-16	VGG-16	Ensemble
CIFAR-10	89.2	69.6	**49.3**

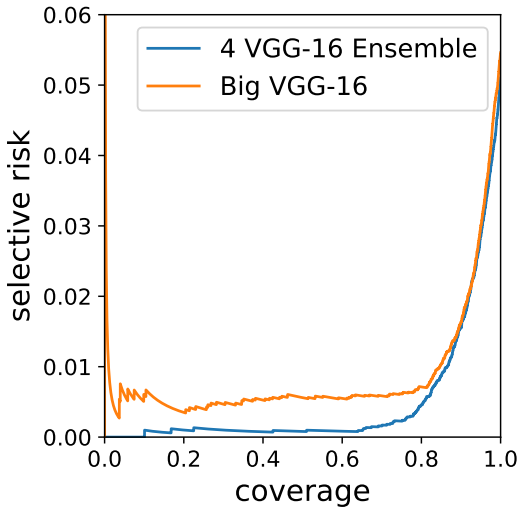

Figure 7: Risk-coverage curves of the ensemble of 4 VGG-16s and the Big VGG-16 on CIFAR-10

E.3 The Effects of Label Noise of SVHN on Selective Classifier Ensembles

In this section, we compare the effect of label noise of SVHN on SR ensemble with that on SAT ensemble, whose result might explain the abnormal experimental results (compared to results on other datasets) on SVHN in Section 6.2. SVHN is not a clean dataset, and there is much more label noise that can be detected in SVHN than in CIFAR-10 and CIFAR-100. Using the soft label of SAT (Huang et al., 2020), we detect label noise in SVHN, CIFAR-10, and CIFAR-100, and find that SVHN has significantly more label noise than CIFAR-10 and CIFAR-100. The result is presented in the following. In addition, it is known that SAT is robust to label noise (Huang et al., 2020), while SR is not so, so we conjecture that the label noise of SVHN is the reason why the SR ensemble is inferior to SAT on SVHN.

We detect label noise with the help of the soft label of SAT. For a sample x_i, the soft label of SAT (Huang et al., 2020), t_{i,y_i}, is used to measure x_i's learning difficulty. The soft label of SAT is initialized as 1 and updated at every training epoch as below

$$t_{i,y_i} \leftarrow \alpha \times t_{i,y_i} + (1 - \alpha) \times p_\theta(y_i|x_i),$$

where $p_\theta(Y|x)$ is the predictive probability distribution of the classifier, y_i is the label of x_i, α is a hyperparameter. The smaller the t_{i,y_i} is, the lower the true class predictive probability of the classifier on x_i during training time, indicating that x_i is more difficult to learn. By selecting a percentage of samples with the lowest t_{i,y_i}, we get the most difficult samples to learn for the classifier, from which we can easily detect label noise manually.

Table 4: Numbers of mislabeled samples in the top-0.1% difficult training samples of SVHN, CIFAR-10, and CIFAR-100.

Dataset	#Mislabeled	#Top-0.1%	Proportion
SVHN	73	73	100%
CIFAR-10	1	50	2%
CIFAR-100	1	50	2%

In training sets of SVHN, CIFAR-10, and CIFAR-100, we detect label noise manually among the top-0.1% difficult (measured by the soft label of SAT) samples. The numbers of mislabeled samples detected in SVHN, CIFAR-10, and CIFAR100 are shown in Table 4. The result shows that SVHN has significantly more mislabeled samples detected than CIFAR-10 and CIFAR-100, indicating much more label noise in SVHN than CIFAR-10 and CIFAR-100.

To verify the effect of label noise, the following experiments are designed. Firstly, we detect label noise manually among the 1% of the hardest-to-learn samples of SVHN training set and test set, using the soft label of SAT. Secondly, we remove the detected mislabeled samples from the original dataset. The remaining SVHN dataset is called the *clean SVHN*. Accordingly, the original dataset is called the *original SVHN*. Finally, we retrain and test the SR ensemble and SAT ensemble and compare their test results. In the second step, the reason for removing mislabeled samples rather than modifying them is that some samples cannot be classified even by humans, and some samples are not in the range of categories of SVHN. Thus, the label noise cannot be eliminated by modifying the labels but by removing mislabeled samples.

The test results of the SR ensemble and SAT ensemble on clean SVHN are shown in Table 5. It is not surprising that the AURCs of the SR ensemble and SAT ensemble are significantly lower on the clean SVHN than the original SVHN. Furthermore, on the clean SVHN, when the number of members is 5, the AURC of the SR ensemble is lower than that of SAT ensemble. Combined with results on the original SVHN, where the AURC of SR ensemble is higher than that of SAT ensemble, we conclude that label noise in SVHN is the reason why the SR ensemble has a higher AURC than SAT ensemble. In other words, label noise is the reason why the SR ensemble performs worse in selective classification than SAT ensemble on SVHN.

Table 5: AURC$/10^{-4}$ of SR ensemble and SAT ensemble on the clean SVHN

Dataset	#Member	SR	SAT
clean SVHN	1	12.3	**7.8**
	2	8.2	**7.1**
	3	7.3	**6.8**
	4	**6.8**	**6.8**
	5	**6.4**	6.8

In summary, by experiments, we show that the SR ensemble is not as robust to label noise as SAT ensemble, and label noise in SVHN is the reason why the SR ensemble is not as good as SAT ensemble on SVHN. We construct the *clean SVHN*, which is SVHN without some mislabeled samples. On the clean SVHN, we compare the SR ensemble with SAT ensemble and find that the SR ensemble is superior to SAT ensemble in selective classification performance. Combined with former experimental results, we conclude that label noise in SVHN is the reason why SR ensemble is inferior to SAT on SVHN.

Considering the experimental results on the clean SVHN and previous experimental results on CIFAR-10 and CIFAR-100 (see Table 5 and Table 1), it can be seen that the SR ensemble is superior to SAT ensemble in selective classification on clean image classification datasets, Thus, SR ensemble is the state-of-the-art selective classification method on clean image classification datasets, but is not as robust to label noise as SAT ensemble.