Label Noise: Ignorance Is Bliss [6 pages]

Anonymous Author(s)
Affiliation
Address
email

Abstract

We establish a new theoretical framework for learning under multi-class, instance-dependent label noise. At the heart of our framework is the concept of *relative signal strength* (RSS), which is a point-wise measure of noisiness. We use relative signal strength to establish matching upper and lower bounds for excess risk . Our theoretical findings reveal a surprising result: the extremely simple *Noise Ignorant Empirical Risk Minimization (NI-ERM)* principle, which conducts empirical risk minimization as if no label noise exists, is minimax optimal. Finally, we translate these theoretical insights into practice: by using NI-ERM to fit a linear classifier on top of a frozen foundation model, we achieve state-of-the-art performance on the CIFAR-N data challenge.

1 Introduction

The problem of classification with label noise can be stated in terms of variables $(X, Y, \widetilde{Y})$, where X is the feature vector, $Y \in \{1, \ldots, K\}$ is the true label associated to X, and $\widetilde{Y} \in \{1, \ldots, K\}$ as a noisy version of Y. The learner has access to i.i.d. realizations of $(X, \widetilde{Y})$, and the objective is to learn a classifier that optimizes the risk associated with (X, Y).

In recent years, there has been a surge of interest in the challenging setting of instance (i.e., feature) dependent label noise, in which $\widetilde{Y}$ can depend on both Y and X. While several algorithms have been developed, there remains relatively little theory regarding algorithm performance and the fundamental limits of this learning paradigm.

This work develops a theoretical framework for learning under multi-class, instance-dependent label noise. Our framework hinges on the concept of *relative signal strength*, which is a point-wise measure of "noisiness" in a label noise problem. Using relative signal strength, we establish matching upper and lower bounds for excess risk. We further identify distributional assumptions that ensure that the lower bound tends to zero as the sample size n grows, implying that consistent learning is possible.

Suprisingly, our theoretical findings reveal that *Noise Ignorant Empirical Risk Minimization (NI-ERM)*, which conducts empirical risk minimization as if no label noise exists, is minimax optimal. To translate this insight into practice, we use NI-ERM to fit a linear classifier on top of a self-supervised feature extractor, achieving state-of-the-art performance on the CIFAR-N data challenge.

2 Literature review

Theory and algorithms for classification with label noise are often based on different probabilistic models. These can be categorized depending on how $\widetilde{Y}$ depends on Y and X. The simplest model is symmetric noise, where the distribution of $\widetilde{Y}$ is independent of Y and X [Angluin and Laird, 1988]. In this case, the probability that $\widetilde{Y} = k$ is the same for all $k \neq Y$, regardless of Y and X. In this

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

setting, it is easy to show that minimizing the noisy excess risk (associated to the 0/1 loss) implies minimizing the clean excess risk, a property known as *immunity*. When immunity holds, there is no need to modify the learning algorithm on account of noisy labels. In other words, the learner may be *ignorant* of the label noise and still learn consistently.

A more general model is classification with label dependent noise, in which the distribution of $\widetilde{Y}$ depends on Y, but not X. Many practical algorithms have been developed over the years, based on principles including data re-weighting [Liu and Tao, 2015], robust training [Han et al., 2018, Liu et al., 2020, Foret et al., 2021] and data cleaning [Brodley and Friedl, 1999, Northcutt et al., 2021]. Consistent learning algorithms still exist, such as those based on loss correction [Natarajan et al., 2013, Patrini et al., 2017, Van Rooyen and Williamson, 2018, Liu and Guo, 2020, Zhang et al., 2022]. However, the primary limitation of these methods is their reliance on the knowledge of noise transition probabilities, which are generally not identifiable [Zhang et al., 2021b].

In the most general setting, that of instance dependent label noise, the distribution of $\widetilde{Y}$ depends on both Y and X. While algorithms are emerging [Cheng et al., 2021, Zhu et al., 2021, Wang et al., 2022, Yang et al., 2023], the theoretical developments have primarily focused on the binary setting. Scott [2019] establishes immunity for a Neyman-Pearson-like performance criterion under a *posterior drift* model, discussed in more detail below. Cannings et al. [2020] establish an upper bound for excess risk under the strong assumption that the optimal classifiers for the clean and noisy distributions are the same. Closest to our work, Im and Grigas [2023] derive excess risk upper and lower bounds, and reach a similar conclusion, that noise-ignorant ERM attains the lower bound. Our results, based on the new concept of relative signal strength, provide a more refined analysis.

Additional connections between our contributions and prior work are made throughout the paper.

3 Problem statement and relative signal strength

Notation. $\mathcal{X}$ denotes a feature space and $\mathcal{Y} = \{1, 2, \ldots, K\}$ a label space, with $K \in \mathbb{N}$. The K-simplex is $\Delta^K := \{p \in \mathbb{R}^K : \forall i, p_i \geq 0, \sum p_i = 1\}$. A $K \times K$ matrix is *row stochastic* if all of its rows are in Δ^K. Denote the i-th element of a vector $\boldsymbol{v}$ as $[\boldsymbol{v}]_i$, and the (i, j)-th element of a matrix $\boldsymbol{M}$ as $[\boldsymbol{M}]_{i,j}$.

3.1 Learning from label noise

In conventional multiclass classification, we observe training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn i.i.d. from a joint distribution P_{XY}. The marginal distribution of X is denoted by P_X, and the *class posterior* probabilities $P_{Y|X=x}$ are captured by a K-simplex-valued vector $\boldsymbol{\eta} : \mathcal{X} \to \Delta^K$, where the j-th component of the vector is $[\boldsymbol{\eta}(x)]_j = \mathbb{P}(Y = j \mid X = x)$. A classifier $f : \mathcal{X} \to \mathcal{Y}$ maps an instance x to a class $f(x) \in \mathcal{Y}$. Denote the risk of a classifier f with respect to distribution P_{XY} as $R(f) = \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\mathbb{1}_{\{f(X)\neq Y\}}\right]$. The Bayes optimal classifier for P_{XY} is $f^*(x) \in \arg\max \boldsymbol{\eta}(x)$. The Bayes risk, which is the minimum achievable risk, is then denoted as $R^* = R(f^*) = \inf_f R(f)$.

We consider the setting where, instead of the true class label Y, a noisy label $\widetilde{Y}$ is observed. The training data $(X_1, \widetilde{Y}_1), \ldots, (X_n, \widetilde{Y}_n)$ can be viewed as an i.i.d. sample drawn from a "noisy" distribution $P_{X\widetilde{Y}}$. We define $P_{\widetilde{Y}|X=x}$, $\widetilde{\boldsymbol{\eta}}$, $\widetilde{R}$ and $\tilde{f}^*$ analogously to the "clean" distribution P_{XY}.

We view $(X, Y, \widetilde{Y})$ as a jointly distributed triple. Given $(X, Y) \sim P_{XY}$, the distribution of the noisy label $\widetilde{Y}$ is determined by the matrix-valued function $\boldsymbol{E} : \mathcal{X} \to \{\boldsymbol{M} \in \mathbb{R}^{K\times K} : \boldsymbol{M} \text{ is row stochastic}\}$, which is called the *noise transition matrix*, and whose (i, j)-th element is

$$[\boldsymbol{E}(x)]_{i,j} = \mathbb{P}\left(\widetilde{Y} = j \mid Y = i, X = x\right).$$

Note that the noisy and clean class posteriors satisfy $\widetilde{\boldsymbol{\eta}}(x) = \boldsymbol{E}(x)^\top \boldsymbol{\eta}(x)$, with $^\top$ denoting transpose.

The goal of learning from label noise is to find a classifier that is able to minimize the "clean test error," that is, the risk R defined w.r.t. P_{XY}, even though the learner's access is limited to corrupted training data $(X_i, \widetilde{Y}_i) \overset{\text{i.i.d.}}{\sim} P_{X\widetilde{Y}}$. In essence, label noise can be conceptualized as a type of domain adaptation problem, where P_{XY} is the source domain, $P_{X\widetilde{Y}}$ is the target domain, and the source

and target are linked by "posterior drift", which means that the source and target have the same X-marginal, but the "posteriors" (distribution of label given X) may change [Scott, 2019, Cai and Wei, 2021, Maity et al., 2023]. Thus, a label noise problem is captured by a triple $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$.

3.2 Relative signal strength

To study label noise, we introduce the concept of *relative signal strength* (RSS). This is a pointwise measure of how much "signal" (certainty about the label) is contained in the noisy distribution relative to the clean distribution. Previous work [Cannings et al., 2020, Cai and Wei, 2021] has examined a related concept within the context of binary classification, under the restriction that clean and noisy Bayes classifiers are identical. Our definition incorporates multi-class classification and relaxes the requirement that the clean and noisy Bayes classifiers agree.

Definition 1 (Relative Signal Strength) *For any class probability vectors $\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}$, define the* relative signal strength *(RSS) at $x \in \mathcal{X}$ as*

$$\mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = \min_{j \in \mathcal{Y}} \ \frac{\max_i [\widetilde{\boldsymbol{\eta}}(x)]_i - [\widetilde{\boldsymbol{\eta}}(x)]_j}{\max_i [\boldsymbol{\eta}(x)]_i - [\boldsymbol{\eta}(x)]_j}, \tag{1}$$

where $0/0 := +\infty$. Furthermore, for $\kappa \in [0, \infty)$, denote the set of points whose RSS exceeds κ as

$$\mathcal{A}_\kappa(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = \{x \in \mathcal{X} : \mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) > \kappa\}.$$

$\mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ is a point-wise measure of how much "signal" the noisy posterior contains about the clean posterior. To gain some intuition, first notice that if the noisy Bayes classifier predicts a different class than the clean Bayes classifier, the RSS is 0 by taking $j = \arg\max \widetilde{\boldsymbol{\eta}}$ (assuming for simplicity that the $\arg\max$ is a singleton set). Now suppose the clean and noisy Bayes classifiers *do* make the same prediction at x, say i^*, and consider a fixed j. If

$$\frac{[\widetilde{\boldsymbol{\eta}}(x)]_{i^*} - [\widetilde{\boldsymbol{\eta}}(x)]_j}{[\boldsymbol{\eta}(x)]_{i^*} - [\boldsymbol{\eta}(x)]_j}$$

is small, it means that the clean Bayes classifier is relatively certain that j is not the correct clean label, while the noisy Bayes classifier is less certain that j is not the correct noisy label. Taking the minimum over j gives the relative signal strength at x. As we formalize in the next section, a large RSS at x ensures that a small (pointwise) noisy excess risk at x implies a small (pointwise) clean excess risk. To gain more intuition, consider the following examples.

Example 1 *When $\boldsymbol{\eta}(x) = [0 \ \ 1 \ \ 0]^\top$ and $\widetilde{\boldsymbol{\eta}}(x) = [0.3 \ \ 0.6 \ \ 0.1]^\top$,*

$$\mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = \min_{j \in \mathcal{Y}} \ \frac{\max_i [\widetilde{\boldsymbol{\eta}}(x)]_i - [\widetilde{\boldsymbol{\eta}}(x)]_j}{\max_i [\boldsymbol{\eta}(x)]_i - [\boldsymbol{\eta}(x)]_j} = \frac{[\widetilde{\boldsymbol{\eta}}(x)]_2 - [\widetilde{\boldsymbol{\eta}}(x)]_1}{[\boldsymbol{\eta}(x)]_2 - [\boldsymbol{\eta}(x)]_1} = \frac{0.6 - 0.3}{1 - 0} = 0.3.$$

Here, the clean Bayes classifier is absolutely certain about its prediction, while the noisy Bayes classifier is much less certain.

Example 2 *When $\boldsymbol{\eta}(x) = [0 \ \ 1 \ \ 0]^\top$ and $\widetilde{\boldsymbol{\eta}}(x) = [0 \ \ 0 \ \ 1]^\top$,*

$$\mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = \min_{j \in \mathcal{Y}} \ \frac{\max_i [\widetilde{\boldsymbol{\eta}}(x)]_i - [\widetilde{\boldsymbol{\eta}}(x)]_j}{\max_i [\boldsymbol{\eta}(x)]_i - [\boldsymbol{\eta}(x)]_j} = \frac{[\widetilde{\boldsymbol{\eta}}(x)]_3 - [\widetilde{\boldsymbol{\eta}}(x)]_3}{[\boldsymbol{\eta}(x)]_2 - [\boldsymbol{\eta}(x)]_3} = \frac{1 - 1}{1 - 0} = 0.$$

The zero signal strength results from $\widetilde{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}$ leading to different predictions.

Example 3 (Comparison to KL divergence) *When $\boldsymbol{\eta}(x) = [0.05 \ \ 0.7 \ \ 0.25]^\top$, and $\widetilde{\boldsymbol{\eta}}^{(1)}(x) = [0.25 \ \ 0.7 \ \ 0.05]^\top$, $\widetilde{\boldsymbol{\eta}}^{(2)}(x) = [0.1 \ \ 0.6 \ \ 0.3]^\top$,*

$$\frac{1}{\mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{\eta} \,\middle\|\, \widetilde{\boldsymbol{\eta}}^{(1)}\right)} < \frac{1}{\mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{\eta} \,\middle\|\, \widetilde{\boldsymbol{\eta}}^{(2)}\right)} \quad \textit{while} \quad \mathcal{M}\left(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}^{(1)}\right) > \mathcal{M}\left(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}^{(2)}\right).$$

Here, $\widetilde{\boldsymbol{\eta}}^{(2)}$ is "closer" to $\boldsymbol{\eta}$ in terms of KL divergence, but $\widetilde{\boldsymbol{\eta}}^{(1)}$ provides more information in terms of predicting the $\arg\max$ *of $\boldsymbol{\eta}$. There is no conflict: KL divergence considers the similarity between two (whole) distributions, while the task of classification only focuses on predicting the* $\arg\max$.

This also illustrates why our notion of RSS is better suited for the label noise problem than other general-purpose distance measures between distributions.

A desirable learning scenario would be if $\mathcal{A}_\kappa(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = \mathcal{X}$ for some large κ, indicating that the signal strength is big across the entire space. Unfortunately, this ideal situation is generally not achievable. To understand this limitation, we begin by making a mild assumption about the label noise:

$$\left|\arg\max \boldsymbol{\eta}(x)\right| \leq \left|\arg\max \widetilde{\boldsymbol{\eta}}(x)\right|, \quad \text{almost surely } (P_X) \tag{C}$$

where $|\cdot|$ denotes set cardinality, that is, that $\arg\max \widetilde{\boldsymbol{\eta}}(x)$ contains at least as many entries as $\arg\max \boldsymbol{\eta}$. In other words, when there are ties for the clean Bayes prediction, the noisy Bayes prediction has at least as many ties. This assumption is reasonable because label noise typically introduces ambiguity by "confusing" the class probabilities, resulting in more ties.

Proposition 1 *If (**C**) holds, then* $\mathcal{A}_0(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = \left\{x \in \mathcal{X} : \arg\max \widetilde{\boldsymbol{\eta}}(x) = \arg\max \boldsymbol{\eta}(x)\right\}$.

This proposition, which is proved in Appendix A.1.1, indicates that $\mathcal{A}_0$, the region with positive RSS, precisely corresponds to the area where the true and noisy Bayes classifiers agree. Accordingly, $\mathcal{X} \setminus \mathcal{A}_0$, the zero signal region, is the region where Bayes decision rules differ. The "region of strong signal," $\mathcal{A}_\kappa$, is a subset of $\mathcal{A}_0$. Since the clean and noisy Bayes classifiers will typically disagree for at least some x, $\mathcal{A}_0 \neq \mathcal{X}$ in general. We note that the (somewhat strong) assumption that $\mathcal{A}_0 = \mathcal{X}$ has been made in prior studies [Cannings et al., 2020, Cai and Wei, 2021].

Posterior Drift Model Class. Now putting definitions together, we consider the posterior drift model Π defined over the triple $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$. Let $\epsilon \in [0, 1], \kappa \in (0, +\infty)$, and define

$$\Pi(\epsilon, \kappa) := \left\{ (P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) : P_X\Big(\mathcal{A}_\kappa\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right)\Big) \geq 1 - \epsilon \right\}.$$

This is a set of triples (label noise problems) such that $\mathcal{A}_\kappa$, the region with RSS at least κ, covers at least $1 - \epsilon$ of the probability mass. In the next section, we will demonstrate that the performance within $\mathcal{A}_\kappa$ can be guaranteed, whereas learning outside the region $\mathcal{A}_\kappa$ is provably challenging.

4 Upper and lower bounds

In this section, we establish both upper and lower bounds for excess risk under multi-class instance-dependent label noise.

4.1 Minimax lower bound

Our first theorem reveals a fundamental limit: no classifier trained using noisy data can surpass the constraints imposed by relative signal strength in a minimax sense. To state the theorem, we employ the following notation and terminology. Denote the noisy training data by $Z^n = \left\{(X_i, \widetilde{Y}_i)\right\}_{i=1}^n \overset{i.i.d.}{\sim} P_{X\widetilde{Y}}$. A *learning rule* $\hat{f}$ is an algorithm that takes Z^n and outputs a classifier. The risk $R(\hat{f})$ of a learning rule is a random variable, where the randomness is due to the draw Z^n.

Theorem 1 (Minimax Lower Bound) *Let* $\epsilon \in [0, 1], \kappa > 0$. *Then*

$$\inf_{\hat{f}} \sup_{(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \geq \frac{K-1}{K}\epsilon + \Omega\left(\frac{1}{\kappa}\sqrt{\frac{1}{n}}\right),$$

where the inf *is over all learning rules.*

Proof Sketch. The idea is to pick a jointly distributed triple $(X, Y, \widetilde{Y})$ such that: Y is independent of $\widetilde{Y}$ for $x \in \mathcal{X} \setminus \mathcal{A}_\kappa$ and Y and $\widetilde{Y}$ are supported on the same classes for $x \in \mathcal{A}_\kappa$. Consider the two regions separately: the excess risk on $\mathcal{X} \setminus \mathcal{A}_\kappa$ is no less than $\frac{K-1}{K}\epsilon$, while the excess risk on $\mathcal{A}_\kappa$ decreases with sample size n, and the rate is affected by the RSS κ. See Appendix A.1.2. ∎

The proof offers insights into how label noise impacts the learning process: if the signal is low, learning is hopeless, and the learner incurs an irreducible error of $(1 - 1/K)\epsilon$; in the high signal region, the signal strength κ determines the rate of convergence. These aspects determine fundamental limits that no classifier trained only on noisy data can overcome without additional assumptions.

4.2 Upper bound

This subsection establishes an upper bound for NI-ERM, the empirical risk minimizer trained on noisy data. This result implies that NI-ERM is minimax optimal, a potentially surprising result given that NI-ERM is arguably the simplest approach one might consider.

Theorem 2 (Excess Risk Upper Bound of NI-ERM) *Let $\epsilon \in [0,1], \kappa > 0$. Consider any $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa)$, assume function class $\mathcal{F}$ has Natarajan dimension V, and the noisy Bayes classifier $\tilde{f}^*$ belongs to $\mathcal{F}$. Let $\hat{f} \in \mathcal{F}$ be the ERM trained on $Z^n = \left\{(X_i, \widetilde{Y}_i)\right\}_{i=1}^n$. Then for $n > 2V$,*

$$\mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \leq \epsilon + \mathcal{O}\left(\frac{1}{\kappa}\sqrt{\frac{V}{n}}\right).$$

Natarajan dimension is a multiclass analogue of VC dimension. The upper bound (Theorem 2) matches the lower bound (which remains unchanged under the setting of Theorem 2 – see Appendix A.1.2 for a refined statement) in terms of both the irreducible error and the rate of convergence. This result is surprising as it indicates that the simplest possible approach, which ignores the presence of noise, is optimal. No learning rule could perform significantly better in this minimax sense.

5 Practical implication

The modern practice of machine learning often involves training a deep neural network. In complex tasks involving noisy labels, the naïve NI-ERM is often outperformed by state-of-the-art methods by a significant extent [Li et al., 2020, Xiao et al., 2023]. This is consistent with the finding that directly training a large neural network on noisy data frequently leads to overfitting [Zhang et al., 2021a].

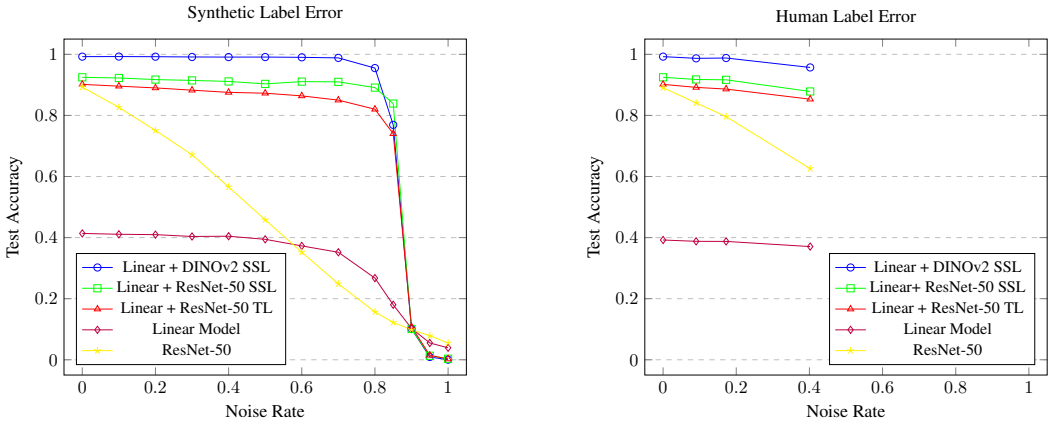
Yet this is not grounds for abandoning NI-ERM altogether as a practical strategy. Instead of using NI-ERM for end-to-end training of a deep neural network, we instead propose the following simple, two-step procedure, termed 'feature extraction + NI-ERM'.

1. Perform feature extraction using any method (e.g., transfer learning or self-supervised learning) that does not require labels.
2. Learn a simple classifier (e.g., a linear classifier) on top of these extracted features, using the noisily labelled data, in a noise-ignorant way.

We acknowledge the that practical idea of ignoring label noise is not new [Ghosh and Lan, 2021], but the full power of this approach has not been previously recognized. For example, prior work that has suggested ignoring the label noise usually augments this approach with additional heuristics [Zheltonozhskii et al., 2022, Xue et al., 2022].

Remarkably, this two-step approach attains extremely strong performance. We conducted experiments on the CIFAR image data under two scenarios: synthetic label flipping and realistic human label errors [Wei et al., 2022], as shown in Figure 1. We examine three different feature extractors: the DINOv2 foundation model [Oquab et al., 2023], ResNet-50 features extracted from training on ImageNet [He et al., 2016], and self-supervised ResNet-50 using contrastive loss [Chen et al., 2020]. We also compared to a simple linear model trained on the raw pixel intensities, and a ResNet-50 trained end-to-end. We observed that ResNet-50 exhibits a "tempered overfitting" behavior, consistent with previous findings [Zhang et al., 2021a, Mallinar et al., 2022]. The linear model demonstrates robustness to noise, but suffers from significant approximation error.

Conversely, the FE+NI-ERM approach enjoys the best of both worlds. Regardless of how the feature extraction is carried out, the resulting models exhibit robustness to label noise, while the overall accuracy depends entirely on the quality of the extracted features. This is illustrated in Figure 1, where the flatness of the accuracy curves as noise increases indicates the robustness, and the intercept at zero label noise is a measure of the feature quality. Importantly, this property holds true even under realistic label noise of CIFAR-N [Wei et al., 2022]. In fact, we find that using the DINOv2 [Oquab et al., 2023] extracted features in our FE+NI-ERM approach yields state of the art results on the CIFAR-10N and CIFAR-100N benchmarks, across the noise levels, as shown in Table 1.



(a) CIFAR-10 with synthetic label noise

(b) CIFAR-10 with realistic label noise

Figure 1: A linear model trained on features obtained from either transfer learning (pretrained ResNet-50 on ImageNet [He et al., 2016]), self-supervised learning (ResNet-50 trained on CIFAR-10 images with contrastive loss [Chen et al., 2020]), or a pretrained self-supervised foundation model DINOv2 [Oquab et al., 2023] significantly boosts the performance of the original linear model. In contrast, directly training a ResNet-50 leads to overfitting.

We reiterate that the only hyperparameters of our model are the hyperparameters of the linear classifier, which are tuned automatically using standard cross-validation on the noisy labels. This in contrast to the implementations of many methods on the CIFAR-N leaderboard (`http://noisylabels.com/`), where the hyperparameters are hard-coded. Altogether, this dominant performance, along with the simplicity of the approach and the lack of any untunable hyperparameters, suggest that the FE+NI-ERM is very powerful, and indicates a need for further investigation of its properties.

Table 1: Performance comparison with CIFAR-N leaderboard (`http://noisylabels.com/`) in terms of testing accuracy. "Aggre", "Rand1", . . . , "Noisy" denote various types of human label noise. We compare with four methods that covers the top three performance for all noise categories: ProMix [Xiao et al., 2023], ILL [Chen et al., 2023], PLS [Albert et al., 2023] and DivideMix [Li et al., 2020]. Our approach, a Noise Ignorant linear model trained on features extracted by the self-supervised foundation model DINOv2 [Oquab et al., 2023] achieves new state-of-the-art results, highlighted in bold. We employed Python's sklearn logistic regression and cross-validation functions without data augmentation; the results are deterministic and directly reproducible.

Leaderboard	CIFAR-10N					CIFAR-100N
Methods	Aggre	Rand1	Rand2	Rand3	Worst	Noisy
ProMix	97.65 ± 0.19	97.39 ± 0.16	97.55 ± 0.12	97.52 ± 0.09	**96.34 ± 0.23**	73.79 ± 0.28
ILL	96.40 ± 0.03	96.06 ± 0.07	95.98 ± 0.12	96.10 ± 0.05	93.55 ± 0.14	68.07 ± 0.33
PLS	96.09 ± 0.09	95.86 ± 0.26	95.96 ± 0.16	96.10 ± 0.07	93.78 ± 0.30	73.25 ± 0.12
DivideMix	95.01 ± 0.71	95.16 ± 0.19	95.23 ± 0.07	95.21 ± 0.14	92.56 ± 0.42	71.13 ± 0.48
FE + NI-ERM	**98.69 ± 0.00**	**98.80 ± 0.00**	**98.65 ± 0.00**	**98.67 ± 0.00**	95.71 ± 0.00	**83.17 ± 0.00**

6 Conclusions

This work presents a rigorous theory for learning under multi-class, instance-dependent label noise. We establish matching upper and lower bounds for excess risk . Our theory reveals the minimax optimality of Noise Ignorant Empirical Risk Minimizer (NI-ERM). To make this theory practical, we provide a simple modification leveraging a feature extractor with NI-ERM, demonstrating significant performance enhancements. A limitation of this work is that our methodology warrants more extensive experimental evaluation.

References

Paul Albert, Eric Arazo, Tarun Krishna, Noel E O'Connor, and Kevin McGuinness. Is your noise correction noisy? pls: Robustness to label noise with two stage detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 118–127, 2023.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine learning*, 2:343–370, 1988.

Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

Timothy I Cannings, Yingying Fan, and Richard J Samworth. Classification with imperfect training labels. *Biometrika*, 107(2):311–330, 2020.

Hao Chen, Ankit Shah, Jindong Wang, Ran Tao, Yidong Wang, Xing Xie, Masashi Sugiyama, Rita Singh, and Bhiksha Raj. Imprecise label learning: A unified framework for learning with various imprecise label configurations. *arXiv preprint arXiv:2305.12715*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

Luc Devroye, László Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer, 1996.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=6Tm1mposlrM`.

Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2703–2708, 2021.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Hyungki Im and Paul Grigas. Binary classification with instance and label dependent label noise. *arXiv preprint arXiv:2306.03402*, 2023.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International conference on machine learning*, pages 6226–6236. PMLR, 2020.

Subha Maity, Diptavo Dutta, Jonathan Terhorst, Yuekai Sun, and Moulinath Banerjee. A linear adjustment based approach to posterior drift in transfer learning. *Biometrika*, 2023. URL `https://doi.org/10.1093/biomet/asad029`.

Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. *Advances in Neural Information Processing Systems*, 35:1182–1195, 2022.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

Clayton Scott. A generalized Neyman-Pearson criterion for optimal domain adaptation. In *Algorithmic Learning Theory*, pages 738–761. PMLR, 2019.

Brendan Van Rooyen and Robert C Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL `https://doi.org/10.1137/1116025`.

Jialu Wang, Eric Xin Wang, and Yang Liu. Estimating instance-dependent label-noise transition matrix using a deep neural network. In *International Conference on Machine Learning*, 2022.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

Ruixuan Xiao, Yiwen Dong, Haobo Wang, Lei Feng, Runze Wu, Gang Chen, and Junbo Zhao. Promix: Combating label noise via maximizing clean sample utility. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4442–4450. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/494. URL `https://doi.org/10.24963/ijcai.2023/494`. Main Track.

Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pages 24851–24871. PMLR, 2022.

Shuo Yang, Songhua Wu, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. A parametrical model for instance-dependent label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Jianxin Zhang, Yutong Wang, and Clayton Scott. Learning from label proportions by learning with label noise. *Advances in Neural Information Processing Systems*, 35:26933–26942, 2022.

Mingyuan Zhang, Jane Lee, and Shivani Agarwal. Learning from noisy labels with no change to the training process. In *International Conference on Machine Learning*, pages 12468–12478. PMLR, 2021b.

Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1657–1667, 2022.

Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10113–10123, 2021.

A Appendix / supplemental material

A.1 Proofs

A.1.1 Proof of Proposition 1

Proposition *Assume (**C**) holds. Then for $\kappa \geq 0$,*

$$\mathcal{A}_0(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = \left\{ x \in \mathcal{X} : \arg\max \widetilde{\boldsymbol{\eta}}(x) = \arg\max \boldsymbol{\eta}(x) \right\} \quad \textit{almost surely}.$$

Proof. Notice that

$$\mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = 0 \quad \iff \quad \arg\max \widetilde{\boldsymbol{\eta}}(x) \not\subseteq \arg\max \boldsymbol{\eta}(x).$$

This is because $\mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = 0$ when the numerator is zero and the denominator is non-zero, which happens when $\arg\max \widetilde{\boldsymbol{\eta}}(x) \not\subseteq \arg\max \boldsymbol{\eta}(x)$. An equivalent statement of this is

$$\mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) > 0 \quad \iff \quad \arg\max \widetilde{\boldsymbol{\eta}}(x) \subseteq \arg\max \boldsymbol{\eta}(x).$$

Under assumption (**C**), the set $\arg\max \widetilde{\boldsymbol{\eta}}(x)$ cannot be a proper subset of $\arg\max \boldsymbol{\eta}(x)$, and therefore

$$\mathcal{M}(x; \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) > 0 \quad \iff \quad \arg\max \widetilde{\boldsymbol{\eta}}(x) = \arg\max \boldsymbol{\eta}(x).$$

Thus,

$$\mathcal{A}_0(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) = \left\{ x \in \mathcal{X} : \arg\max \widetilde{\boldsymbol{\eta}}(x) = \arg\max \boldsymbol{\eta}(x) \right\},$$

almost everywhere. ∎

A.1.2 Proof of Lower Bound: Theorem 1

Now we provide a more formal statement of the minimax lower bound and its proof. We begin with the scenario where the noisy distribution $P_{X\widetilde{Y}}$ has zero Bayes risk as an introductory example. The proof for the general case follows a similar strategy but involves more complex bounding techniques. We recommend that interested readers first review the proof of the zero-error version to build a solid understanding before tackling the general case.

Now consider a more restricted set of $\Pi(\epsilon, \kappa)$:

$$\Pi(\epsilon, \kappa, V, 0) := \left\{ (P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) : P_X\left(\mathcal{A}_\kappa\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right)\right) \geq 1 - \epsilon, P_X \text{ supported on } V+1 \text{ points}, \widetilde{R}^* = 0 \right\}.$$

Theorem (Minimax Lower Bound: when $\widetilde{R}^* = 0$) *Let $\epsilon \in [0,1], \kappa > 0, V > 1$. For any learning rule $\hat{f}$ based upon $Z^n = \left\{(X_i, \widetilde{Y}_i)\right\}_{i=1}^n$, and $n > \max(V-1, 2)$,*

$$\begin{aligned} \sup_{(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] &\geq \sup_{(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa, V, 0)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \\ &\geq \frac{K-1}{K}\epsilon + \frac{1}{\kappa}\frac{(V-1)(1-\epsilon)}{8en} \end{aligned}$$

Proof. Consider any $V+1$ distinct points $x_0, x_1, \ldots, x_V$. Choose

$$P_X(x) = \begin{cases} \epsilon & x = x_0 \\ (1-\epsilon) \cdot \frac{1}{n} & x = x_1, \ldots, x_{V-1} \\ (1-\epsilon) \cdot \left(1 - \frac{V-1}{n}\right) & x = x_V. \end{cases},$$

where $n > V-1$, and define the clean and noisy class posteriors by

$$\text{If } x = x_0, \text{ then } \boldsymbol{\eta}(x) = \boldsymbol{e}_j,\ \widetilde{\boldsymbol{\eta}}(x) = \boldsymbol{e}_1, \quad j \in \{1, 2, \dots K\} \tag{2}$$

$$\text{If } x = x_t, 1 \le t \le V-1, \text{ then } \boldsymbol{\eta}(x) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2(\kappa+\delta)} \cdot (-1)^{b_t+1} \\ \frac{1}{2} - \frac{1}{2(\kappa+\delta)} \cdot (-1)^{b_t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \widetilde{\boldsymbol{\eta}}(x) = \boldsymbol{e}_{b_t}, b_t \in \{1, 2\}, \delta > 0, \tag{3}$$

$$\text{If } x = x_V, \text{ then } \boldsymbol{\eta}(x) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2(\kappa+\delta)} \\ \frac{1}{2} - \frac{1}{2(\kappa+\delta)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \widetilde{\boldsymbol{\eta}}(x) = \boldsymbol{e}_1, \tag{4}$$

where $\boldsymbol{e}_i$ denotes the one-hot vector whose i-th element is one.

The triple $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ is thus parameterized by $j, \boldsymbol{b} := [b_1\ b_2\ \cdots\ b_{V-1}]^\top$, and δ.

This construction ensures $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa, V, 0)$. In particular,

$$\begin{aligned} &\mathcal{A}_\kappa \supseteq \{x_1, x_2, \dots, x_V\}, && P_X(\mathcal{A}_\kappa) \ge 1-\epsilon, \\ &\mathcal{X} \setminus \mathcal{A}_\kappa \subseteq \{x_0\}, && P_X(\mathcal{X} \setminus \mathcal{A}_\kappa) \le \epsilon, \end{aligned}$$

and $\widetilde{R}^* = 0$ because $\widetilde{\boldsymbol{\eta}}(x)$ is one-hot for all x.

For classifier any f, by definition, its risk equals to

$$\begin{aligned} R(f) &= \mathbb{E}_{X,Y}\left[\mathbb{1}_{f(X)\neq Y}\right] \\ &= \mathbb{E}_X \mathbb{E}_{Y|X}[\mathbb{1}_{f(X)\neq Y}] \\ &= \mathbb{E}_X \mathbb{E}_{Y|X}[1 - \mathbb{1}_{f(X)=Y}] \\ &= \mathbb{E}_X\left[1 - [\boldsymbol{\eta}(X)]_{f(X)}\right] \\ &= \int_{\mathcal{X}} \left(1 - [\boldsymbol{\eta}(x)]_{f(x)}\right) dP_X(x), \end{aligned}$$

Under our construction of P_X, $R(f)$ can be decomposed into two parts

$$R(f) = \underbrace{\int_{\{x_0\}} \left(1 - [\boldsymbol{\eta}(x)]_{f(x)}\right) dP_X(x)}_{:=R_0(f)} + \underbrace{\int_{\{x_1,\dots,x_V\}} \left(1 - [\boldsymbol{\eta}(x)]_{f(x)}\right) dP_X(x)}_{:=R_V(f)},$$

so does the excess risk

$$R(f) - R(f^*) = \Big(R_0(f) - R_0(f^*)\Big) + \Big(R_V(f) - R_V(f^*)\Big).$$

Recall that in our construction, $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ is parameterized by $j, \boldsymbol{b}$, and δ. Therefore

$$\begin{aligned} \sup_{(P_X,\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}})\in\Pi(\epsilon,\kappa,V,0)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] &\ge \sup_{j,\boldsymbol{b},\delta} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \\ &= \sup_{j,\boldsymbol{b},\delta} \Big\{\mathbb{E}_{Z^n}\left[R_0\left(\hat{f}\right) - R_0(f^*)\right] \\ &\qquad + \mathbb{E}_{Z^n}\left[R_V\left(\hat{f}\right) - R_V(f^*)\right]\Big\} \\ &= \sup_{j} \mathbb{E}_{Z^n}\left[R_0\left(\hat{f}\right) - R_0(f^*)\right] \\ &\qquad + \sup_{\boldsymbol{b},\delta} \mathbb{E}_{Z^n}\left[R_V\left(\hat{f}\right) - R_V(f^*)\right] \end{aligned}$$

where the last equality holds because region $\{x_0\}$ only depends on j, while region $\{x_1, \ldots, x_V\}$ only depends on $\boldsymbol{b}, \delta$.

In the remaining part of the proof, we will examine

$$\sup_j \mathbb{E}_{Z^n}\left[R_0\left(\hat{f}\right) - R_0(f^*)\right] \tag{5}$$

and

$$\sup_{\boldsymbol{b},\delta} \mathbb{E}_{Z^n}\left[R_V\left(\hat{f}\right) - R_V(f^*)\right] \tag{6}$$

separately.

Let's start with the first term (5), which acts over the "low signal strength" region $\{x_0\}$. Since $\boldsymbol{\eta}$ is one-hot on $\{x_0\}$, its Bayes risk over that is zero

$$\begin{aligned}\sup_j \mathbb{E}_{Z^n}\left[R_0\left(\hat{f}\right) - R_0(f^*)\right] &= \sup_j \mathbb{E}_{Z^n}\left[R_0\left(\hat{f}\right)\right] \\ &= \sup_j \mathbb{E}_{Z^n}\left[\int_{\{x_0\}} \mathbb{1}_{\hat{f}(x)\neq j} dP_X(x)\right].\end{aligned}$$

To deal with $\sup_j$, we use a technique called "the probabilistic method": replace j with a random variable $J \sim \text{Uniform}\{1, 2, \ldots, K\}$:

$$\begin{aligned}\sup_j \mathbb{E}_{Z^n}\left[\int_{\{x_0\}} \mathbb{1}_{\hat{f}(x)\neq j} dP_X\right] &\geq \mathbb{E}_{J,\, Z^n}\left[\int_{\{x_0\}} \mathbb{1}_{\hat{f}(x)\neq J} dP_X(x)\right] \\ &= \mathbb{E}_{Z^n}\left[\mathbb{E}_{J|Z^n}\left[\int_{\{x_0\}} \mathbb{1}_{\hat{f}(x)\neq J} dP_X(x)\right]\right].\end{aligned}$$

Again, notice that J is an independent draw. Even if the point x_0 is observed in Z^n, the associated noisy label $\widetilde{Y} = 1$ does not give any information about the clean label $Y = J$. Thus

$$\begin{aligned}\mathbb{E}_{Z^n}\left[\mathbb{E}_{J|Z^n}\left[\int_{\{x_0\}} \mathbb{1}_{\hat{f}(x)\neq J} dP_X(x)\right]\right] &= \mathbb{E}_{Z^n}\left[\mathbb{E}_J\left[\int_{\{x_0\}} \mathbb{1}_{\hat{f}(x)\neq J} dP_X(x)\right]\right] \\ &= \mathbb{E}_{Z^n}\left[\int_{\{x_0\}} \mathbb{E}_J\left[\mathbb{1}_{\hat{f}(x)\neq J}\right] dP_X(x)\right] \\ &= \mathbb{E}_{Z^n}\left[\int_{\{x_0\}} \left(1 - \frac{1}{K}\right) dP_X(x)\right] \\ &= \left(1 - \frac{1}{K}\right)\epsilon.\end{aligned}$$

Now we have the minimax lower bound for the first part (5):

$$\sup_j \mathbb{E}_{Z^n}\left[R_{\{x_0\}}\left(\hat{f}\right) - R_{\{x_0\}}(f^*)\right] \geq \left(1 - \frac{1}{K}\right)\epsilon.$$

For the second part (6), which is over $\{x_1, \ldots, x_V\}$, due to the relative signal strength condition, and from our explicit construction in Eqn. (3) and (4), the excess risk w.r.t. true and noisy distribution are related by

$$\begin{aligned}R_V(f) - R_V(f^*) &= \int_{\{x_1,\ldots,x_V\}} \left(\max \boldsymbol{\eta}(x) - [\boldsymbol{\eta}(x)]_{f(x)}\right) dP_X(x) \\ &= \int_{\{x_1,\ldots,x_V\}} \frac{1}{\kappa + \delta}\left(\max \widetilde{\boldsymbol{\eta}}(x) - [\widetilde{\boldsymbol{\eta}}(x)]_{f(x)}\right) dP_X(x) \quad \because \text{by construction of } \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}} \\ &= \frac{1}{\kappa + \delta}\left(\widetilde{R}_V(f) - \widetilde{R}_V(\tilde{f}^*)\right),\end{aligned}$$

where $\widetilde{R}_V(f) := \int_{\{x_1,\ldots,x_V\}} \left(1 - [\widetilde{\boldsymbol{\eta}}(x)]_{f(x)}\right) dP_X(x)$. Also note that $f^*(x) = \tilde{f}^*(x)$ for $x \in \{x_1, \ldots, x_V\}$, which is a result of our construction of $\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}$.

Then

$$\sup_{\boldsymbol{b},\delta} \mathbb{E}_{Z^n}\left[R_V\left(\hat{f}\right) - R_V(f^*)\right] = \sup_{\boldsymbol{b},\delta} \mathbb{E}_{Z^n}\left[\frac{1}{\kappa+\delta}\left(\widetilde{R}_V(f) - \widetilde{R}_V(\tilde{f}^*)\right)\right].$$

This allows us to reduce the label noise problem into a standard learning problem: we have an iid sample Z^n from $P_{X\widetilde{Y}}$ and consider the risk evaluated on the same distribution $P_{X\widetilde{Y}}$. The remainder of the proof is similar to the proof of Theorem 14.1 in Devroye et al. [1996].

Notice that by our construction, $\widetilde{Y}$ is a deterministic function of X. To be specific, $\widetilde{Y} = \tilde{f}^*(X)$, where

$$\tilde{f}^*(x) = \begin{cases} 1 & x = x_0, \\ b_t & x = x_t,\ 1 \le t \le V-1 \\ 1 & x = x_V \end{cases}$$

is the noisy Bayes classifier.

We use the shorthand $f_{\boldsymbol{b}} := \tilde{f}^*$ to denote that the noisy Bayes classifier depends on $\boldsymbol{b}$, we also write the learning rule $\hat{f}(x; Z^n) := \hat{f}(x)$ to indicate the dependence of the learning rule $\hat{f}$ on the random sample Z^n.

Since the noisy Bayes risk is zero,

$$\sup_{\boldsymbol{b},\delta} \mathbb{E}_{Z^n}\left[\frac{1}{\kappa+\delta}\left(\widetilde{R}_V(f) - \widetilde{R}_V(\tilde{f}^*)\right)\right] = \sup_{\boldsymbol{b},\delta} \frac{1}{\kappa+\delta}\mathbb{E}_{Z^n}\left[\widetilde{R}_V(\hat{f})\right].$$

Again, use the probabilistic method, replace $\boldsymbol{b}$ with $\boldsymbol{B} \sim \text{Uniform}\{1,2\}^{V-1}$,

$$\begin{aligned}\sup_{\boldsymbol{b},\delta} \frac{1}{\kappa+\delta}\mathbb{E}_{Z^n}\left[\widetilde{R}_V(\hat{f})\right] &\ge \sup_{\delta} \frac{1}{\kappa+\delta}\mathbb{E}_{\boldsymbol{B},Z^n}\left[\widetilde{R}_V(\hat{f})\right] \\ &= \sup_{\delta} \frac{1}{\kappa+\delta}\mathbb{E}_{Z^n}\left[\mathbb{E}_{\boldsymbol{B}|Z^n}\left[\int_{\{x_1,\ldots,x_V\}} \mathbb{1}_{\hat{f}(x)\neq f_{\boldsymbol{B}}(x)} dP_X(x)\right]\right]\end{aligned}$$

Since we have $\boldsymbol{B} \sim \text{Uniform}\{1,2\}^{V-1}$ and also $Z|\boldsymbol{B} \sim P_{X\widetilde{Y}}$, then by Bayes rule (or eye-balling), we get the posterior distribution of $\boldsymbol{B}|Z^n$, to be specific:

$$\forall x \in \{x_1, \cdots, x_V\},\ \mathbb{P}\left(f_{\boldsymbol{B}}(x) = 1 | Z^n\right) = \begin{cases} \frac{1}{2} & x \neq X_1, \ldots, x \neq X_n, x \neq x_V \\ 0 \text{ or } 1 & \text{otherwise,} \end{cases}$$

where we overload the notation $\mathbb{P}$ to denote conditional probability of $\boldsymbol{B}|Z^n$.

Then the optimal decision rule for predicting $\boldsymbol{B}$ based on sample Z^n is:

$$f^*_{\boldsymbol{B}}(x; Z^n) := \begin{cases} \widetilde{Y}_i & x = X_i, i \in \{1, 2, \ldots, n\} \\ 1 & x = x_V \\ \text{random guess from } \{1,2\} & x \neq X_1, \ldots, x \neq X_n, x \neq x_V. \end{cases}$$

Therefore, roughly speaking, the error comes from the probability of $X \in \{x_1, \ldots, x_V\}$ not being one of observed X_i: for any $\hat{f}$,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{B}, Z^n}\left[\widetilde{R}_V(\hat{f})\right] &= \mathbb{E}_{Z^n}\left[\mathbb{E}_{\boldsymbol{B}|Z^n}\left[\int_{\{x_1,\ldots,x_V\}} \mathbb{1}_{\hat{f}(x)\neq f_{\boldsymbol{B}}(x)} dP_X(x)\right]\right] \\
&\geq \mathbb{P}\left(X \in \{x_1, \ldots, x_V\},\ f^*_{\boldsymbol{B}}(X; Z^n) \neq f_{\boldsymbol{B}}(X)\right) \qquad \because \text{error of } \hat{f} > \text{error of } f^*_{\boldsymbol{B}} \\
&= \left(1 - \frac{1}{2}\right)\mathbb{P}\left(X \neq X_1, \ldots, X \neq X_n, X \neq x_V, X \in \{x_1, \ldots, x_V\}\right) \\
&= \frac{1}{2}\sum_{t=1}^{V}\mathbb{P}\left(X \neq X_1, \ldots, X \neq X_n, X \neq x_V, X = x_t\right) \\
&= \frac{1}{2}\sum_{t=1}^{V}\mathbb{P}\left(X_1 \neq x_t, \ldots, X_n \neq x_t, x_V \neq x_t, X = x_t\right) \qquad \because \text{replace all } X \text{ with } x_t \\
&= \frac{1}{2}\sum_{t=1}^{V-1}\mathbb{P}\left(X_1 \neq x_t, \ldots, X_n \neq x_t, X = x_t\right) \\
&= \frac{1}{2}\sum_{t=1}^{V-1}\mathbb{P}\left(X_1 \neq x_t, \ldots, X_n \neq x_t | X = x_t\right)\mathbb{P}\left(X = x_t\right) \\
&= \frac{1}{2}\sum_{t=1}^{V-1}\left(1 - \mathbb{P}\left(X = x_t\right)\right)^n \mathbb{P}\left(X = x_t\right) \\
&= \frac{1}{2}(V-1)\left(1 - \frac{1-\epsilon}{n}\right)^n\left(\frac{1-\epsilon}{n}\right) \\
&= \frac{(V-1)(1-\epsilon)}{2n}\left(1 - \frac{1-\epsilon}{n}\right)^n \\
&= \frac{(V-1)(1-\epsilon)}{2n}\left(1 - \frac{1-\epsilon}{n}\right)^{1+\epsilon}\left(1 - \frac{1-\epsilon}{n}\right)^{n-1-\epsilon} \\
&\geq \frac{(V-1)(1-\epsilon)}{2n}\left(1 - \frac{1-\epsilon}{n}\right)^{1+\epsilon} e^{-1+\epsilon} \qquad \because \left(1 - \frac{1-\epsilon}{n}\right)^{n-1-\epsilon} \downarrow e^{-1+\epsilon} \\
&\geq \frac{(V-1)(1-\epsilon)}{2n}\left(1 - \frac{1}{n}\right)^2 e^{-1} \qquad \because \epsilon \in [0,1] \\
&\geq \frac{(V-1)(1-\epsilon)}{2n}\frac{e^{-1}}{4} = \frac{(V-1)(1-\epsilon)}{8en}. \qquad \text{take } n > 2
\end{aligned}
$$

Now we get the minimax risk for the second part (6)

$$
\begin{aligned}
\sup_{\boldsymbol{b},\delta} \mathbb{E}_{Z^n}\left[R_{\mathcal{A}_\kappa}\left(\hat{f}\right) - R_{\mathcal{A}_\kappa}(f^*)\right] &\geq \sup_{\delta} \frac{1}{\kappa + \delta}\frac{(V-1)(1-\epsilon)}{8en} \\
&\geq \frac{1}{\kappa}\frac{(V-1)(1-\epsilon)}{8en} \qquad \text{let } \delta \downarrow 0
\end{aligned}
$$

Combine the two parts together, we get the final result, for $n > \max(V-1, 2)$

$$
\sup_{(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa, V, 0)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \geq \frac{K-1}{K}\epsilon + \frac{1}{\kappa}\frac{(V-1)(1-\epsilon)}{8en}.
$$

■

As for the general version of the lower bound, now consider the set of triples:

$$\Pi(\epsilon,\kappa,V,L) := \Big\{ (P_X,\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}}) : P_X\Big(\mathcal{A}_\kappa\left(\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}}\right)\Big) \geq 1-\epsilon, \\ P_X \text{ supported on } V+1 \text{ points}, \frac{\widetilde{R}_{\mathcal{A}_\kappa}\left(\tilde{f}^*\right)}{P_X\Big(\mathcal{A}_\kappa\left(\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}}\right)\Big)} = L\Big\},$$

where $\widetilde{R}_C(f) = \int_C \left(1 - [\widetilde{\boldsymbol{\eta}}(x)]_{f(x)}\right) dP_X(x)$.

Theorem (Minimax Lower Bound (General Version)) *Let $\epsilon \in [0,1], \kappa > 0, V > 1, L \in (0,1/2)$. For any learning rule $\hat{f}$ based upon $Z^n = \left\{(X_i,\widetilde{Y}_i)\right\}_{i=1}^n$, for $n \geq \frac{V-1}{2L}\max\left\{9, \frac{1}{(1-2L)^2}\right\}$*

$$\sup_{(P_X,\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}})\in\Pi(\epsilon,\kappa)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \geq \sup_{(P_X,\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}})\in\Pi(\epsilon,\kappa,V,L)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \\ \geq \frac{K-1}{K}\epsilon + \frac{1-\epsilon}{\kappa}\sqrt{\frac{(V-1)L}{72n}}e^{-8}.$$

Proof.

Consider any $V+1$ distinct points $x_0, x_1, \ldots, x_V$. Choose

$$P_X(x) = \begin{cases} \epsilon & x = x_0 \\ (1-\epsilon)\cdot p & x = x_1,\ldots,x_{V-1} \\ (1-\epsilon)\cdot(1-(V-1)p) & x = x_V. \end{cases}$$

This imposes the constraint $(V-1)p \leq 1$, which will be satisfied in the end. Notice the difference compared to the previous zero-error proof: we place probability mass p, rather than $1/n$, on $x_1,\ldots,x_{V-1}$.

As for the clean and noisy class probabilities, choose

$$\text{If } x = x_0, \text{ then } \boldsymbol{\eta}(x) = \boldsymbol{e}_j,\ \widetilde{\boldsymbol{\eta}}(x) = \boldsymbol{e}_1, \quad j \in \{1,2,\ldots k\} \tag{7}$$

$$\text{If } x = x_t,\ 1 \leq t \leq V-1, \text{ then } \boldsymbol{\eta}(x) = \begin{bmatrix} \frac{1}{2} + \frac{c}{\kappa+\delta}\cdot(-1)^{b_t+1} \\ \frac{1}{2} - \frac{c}{\kappa+\delta}\cdot(-1)^{b_t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ \widetilde{\boldsymbol{\eta}}(x) = \begin{bmatrix} \frac{1}{2} + c\cdot(-1)^{b_t+1} \\ \frac{1}{2} - c\cdot(-1)^{b_t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$b_t \in \{1,2\},\ \delta > 0,\ c \in \left(0, \frac{1}{2}\right) \tag{8}$$

$$x = x_V, \quad \boldsymbol{\eta}(x) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2(\kappa+\delta)} \\ \frac{1}{2} - \frac{1}{2(\kappa+\delta)} \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ \widetilde{\boldsymbol{\eta}}(x) = \boldsymbol{e}_1, \tag{9}$$

where $\boldsymbol{e}_i$ denotes the one-hot vector whose i-th element is one.

The construction for class posterior is also similar to the previous proof, except that for $x = x_t, t \in \{1,\ldots,V-1\}$, $\widetilde{\boldsymbol{\eta}}$ is no longer a one-hot vector, rather has class probability separated by $2c$: $\left|\,[\widetilde{\boldsymbol{\eta}}(x)]_1 - [\widetilde{\boldsymbol{\eta}}(x)]_2\,\right| = 2c$.

The triple $(P_X,\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}})$ can be parameterized by $j, \boldsymbol{b} := [b_1\ b_2\ \cdots\ b_{V-1}]^\top$, δ, c and p.

Again, this construction ensures $(P_X,\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon,\kappa)$, to be specific:

$$\mathcal{A}_\kappa \supseteq \{x_1,x_2,\ldots,x_V\}, \qquad P_X(\mathcal{A}_\kappa) \geq 1-\epsilon, \\ \mathcal{X}\setminus\mathcal{A}_\kappa \subseteq \{x_0\}, \qquad P_X(\mathcal{X}\setminus\mathcal{A}_\kappa) \leq \epsilon,$$

For any classifier f, its risk can be decomposed into two parts

$$R\left(f\right) = \underbrace{\int_{\{x_0\}} \left(1 - [\boldsymbol{\eta}(x)]_{f(x)}\right) dP_X(x)}_{:=R_0(f)} + \underbrace{\int_{\{x_1,\ldots,x_V\}} \left(1 - [\boldsymbol{\eta}(x)]_{f(x)}\right) dP_X(x)}_{:=R_V(f)},$$

so does the excess risk

$$R\left(f\right) - R(f^*) = \Big(R_0\left(f\right) - R_0(f^*)\Big) + \Big(R_V\Big(f\Big) - R_V(f^*)\Big).$$

In our construction, $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ is parameterized by $j, \boldsymbol{b} := [b_1\ b_2\ \cdots\ b_{V-1}]^\top$, δ, c and p, therefore

$$\sup_{(P_X,\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}})\in\Pi(\epsilon,\kappa,V,L)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \geq \sup_j \mathbb{E}_{Z^n}\left[R_0\left(\hat{f}\right) - R_0(f^*)\right] \tag{10}$$

$$+ \sup_{\boldsymbol{b},\delta,c,p} \mathbb{E}_{Z^n}\left[\frac{1}{\kappa+\delta}\left(\widetilde{R}_V(f) - \widetilde{R}_V(\tilde{f}^*)\right)\right]. \tag{11}$$

Note that we have used the fact that

$$R_V(f) - R_V(f^*) = \frac{1}{\kappa+\delta}\left(\widetilde{R}_V(f) - \widetilde{R}_V(\tilde{f}^*)\right),$$

where $\widetilde{R}_V(f) := \int_{\{x_1,\ldots,x_V\}} \left(1 - [\widetilde{\boldsymbol{\eta}}(x)]_{f(x)}\right) dP_X(x)$.

The first part (10) is exactly the same as the that of the zero-error proof, we have

$$\sup_j \mathbb{E}_{Z^n}\left[R_0\left(\hat{f}\right) - R_0(f^*)\right] \geq \left(1 - \frac{1}{K}\right)\epsilon.$$

From this point forward, the procedure is similar to the proof of Theorem 14.5 in Devroye et al. [1996]. For the second part (11), the noisy Bayes classifier is still

$$\tilde{f}^*(x) = \begin{cases} j & x = x_0, \\ b_t & x = x_t,\ 1 \leq t \leq V \\ 1 & x = x_V. \end{cases}$$

We also use the shorthand $f_{\boldsymbol{b}} := \tilde{f}^*$ to denote that the noisy Bayes classifier depends on $\boldsymbol{b}$. The noisy Bayes risk is no longer zero. In fact

$$\widetilde{R}_V(\tilde{f}^*) = (V-1)(1-\epsilon)p\left(\frac{1}{2} - c\right)$$

Notice that our construction $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa, V, L)$, by definition

$$\frac{\widetilde{R}_{\mathcal{A}_\kappa}\left(\tilde{f}^*\right)}{P_X\Big(\mathcal{A}_\kappa\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right)\Big)} = L,$$

therefore

$$L = \frac{\widetilde{R}_{\mathcal{A}_\kappa}\left(\tilde{f}^*\right)}{P_X\Big(\mathcal{A}_\kappa\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right)\Big)} \leq \frac{\widetilde{R}_V(\tilde{f}^*)}{P_X\Big(\left\{x_1, \ldots, x_V\right\}\Big)} = (V-1)p\left(\frac{1}{2} - c\right), \tag{12}$$

where the inequality holds from $\widetilde{R}_{\mathcal{A}_\kappa}(\tilde{f}^*) = \widetilde{R}_V(\tilde{f}^*)$ and $P_X\Big(\mathcal{A}_\kappa\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right)\Big) \geq P_X\Big(\left\{x_1, \ldots, x_V\right\}\Big)$.

It should be noted that since $(V-1)p \leq 1$ is required, and that $c > 0$, we have $L < 1 \cdot 1/2$. This is the origin of our condition $L < 1/2$ in the statement of the theorem. Naturally, the statement can be adjusted to $\min(L, 1/2)$ instead.

For fixed $\boldsymbol{b}$, the excess risk in region $\{x_1, \ldots, x_V\}$ becomes

$$\begin{aligned}\widetilde{R}_V(\hat{f}) - \widetilde{R}_V(\tilde{f}^*) &= \int_{\{x_1,\ldots,x_V\}} 2c\mathbb{1}_{\hat{f}(x)\neq f_{\boldsymbol{b}}(x)} dP_X(x) \\ &\geq 2c\sum_{t=1}^{V-1}(1-\epsilon)p\mathbb{1}_{\hat{f}(x_t)\neq f_{\boldsymbol{b}}(x_t)},\end{aligned}$$

where the inequality follows from the fact that we ignore the risk on point x_V.

Using the probabilistic method, replace $\boldsymbol{b}$ with $\boldsymbol{B} \sim \text{Uniform}\{1,2\}^{V-1}$,

$$\begin{aligned}\sup_{\boldsymbol{b},\delta,c,p} \mathbb{E}_{Z^n}\left[\frac{1}{\kappa+\delta}\left(\widetilde{R}_V(\hat{f}) - \widetilde{R}_V(\tilde{f}^*)\right)\right] &\geq \sup_{\delta,c,p} \mathbb{E}_{\boldsymbol{B},Z^n}\left[\frac{1}{\kappa+\delta}\left(\widetilde{R}_V(\hat{f}) - \widetilde{R}_V(\tilde{f}^*)\right)\right] \\ &= \sup_{\delta,c,p}\frac{1}{\kappa+\delta}\mathbb{E}_{Z^n}\left[\mathbb{E}_{\boldsymbol{B}|Z^n}\left[\left(\widetilde{R}_V(\hat{f}) - \widetilde{R}_V(\tilde{f}^*)\right)\right]\right]\end{aligned}$$

Now, we need to calculate $\boldsymbol{B}|Z^n$, which can be calculated using Bayes rule because we have $\boldsymbol{B} \sim \text{Uniform}\{1,2\}^{V-1}$ and also $Z|\boldsymbol{B} \sim P_{X\widetilde{Y}}$.

To be specific, for any $x \in \{x_0, x_1, \ldots, x_{V-1}\}$, assume point x_t is observed k times in training sample Z^n,

$$\mathbb{P}\left(f_{\boldsymbol{B}}(x) = 1|Z^n\right) = \begin{cases}\frac{1}{2} & x \neq X_1, \ldots, x \neq X_n, x \neq x_V \\ \mathbb{P}\left(B_t = 1|Y_{t_1}, \ldots, Y_{t_k}\right) & x = x_t = X_{t_1} = \cdots = X_{t_k},\ 1 \leq t \leq V-1,\end{cases}$$

where B_t denotes the t-th element of vector $\boldsymbol{B}$.

Next we compute $\mathbb{P}\left(B_t = 1|Y_{t_1} = y_1, \ldots, Y_{t_k} = y_k\right)$ for $y_1, \ldots, y_k \in \{1,2\}$. Denote the numbers of ones and twos by $k_1 = |\{j \leq k : y_j = 1\}|$ and $k_2 = |\{j \leq k : y_j = 2\}|$. Using Bayes rule, we get

$$\begin{aligned}\mathbb{P}\left(B_t = 1|Y_{t_1}, \ldots, Y_{t_k}\right) &= \frac{\mathbb{P}\left(B_t = 1 \cap Y_{t_1}, \ldots, Y_{t_k}\right)}{\mathbb{P}\left(Y_{t_1}, \ldots, Y_{t_k}\right)} \\ &= \frac{\mathbb{P}\left(Y_{t_1}, \ldots, Y_{t_k}|B_t = 1\right)\mathbb{P}\left(B_t = 1\right)}{\sum_{i=1}^{2}\mathbb{P}\left(Y_{t_1}, \ldots, Y_{t_k}|B_t = i\right)\mathbb{P}\left(B_t = i\right)} \\ &= \frac{(1/2+c)^{k_1}(1/2-c)^{k_2}(1/2)}{(1/2+c)^{k_1}(1/2-c)^{k_2}(1/2) + (1/2+c)^{k_2}(1/2-c)^{k_1}(1/2)}\end{aligned}$$

After some calculation, following Theorem 14.5 in Devroye et al. [1996], we get

$$\begin{aligned}\sup_{\boldsymbol{b},\delta,c,p} \mathbb{E}_{Z^n}\left[\frac{1}{\kappa+\delta}\left(\widetilde{R}_{\mathcal{A}_\kappa}(f) - \widetilde{R}_{\mathcal{A}_\kappa}(\tilde{f}^*)\right)\right] &\geq \sup_{\delta,c,p}\frac{1}{\kappa+\delta}c(V-1)(1-\epsilon)pe^{-\frac{8n(1-\epsilon)pc^2}{1-2c} - \frac{4c\sqrt{n(1-\epsilon)p}}{1-2c}} \\ &\geq \frac{1-\epsilon}{\kappa}\sup_{c,p} c(V-1)pe^{-\frac{8npc^2}{1-2c} - \frac{4c\sqrt{np}}{1-2c}} \quad \because \epsilon \geq 0, \text{ take } \delta \downarrow 0 \\ &= \frac{1-\epsilon}{\kappa}\sup_{c,p} c\,\frac{L}{1/2-c}e^{-\frac{8npc^2}{1-2c} - \frac{4c\sqrt{np}}{1-2c}}\end{aligned}$$

Take $c = \sqrt{\frac{(V-1)}{8nL}}$, and p is automatically specified by Eqn. (12). Then

$$\sup_{\boldsymbol{b},\delta,c,p} \mathbb{E}_{Z^n}\left[\frac{1}{\kappa+\delta}\left(\widetilde{R}_{\mathcal{A}_\kappa}(f) - \widetilde{R}_{\mathcal{A}_\kappa}(\tilde{f}^*)\right)\right] \geq \frac{1-\epsilon}{\kappa}\sqrt{\frac{V-1}{8nL}}\frac{L}{1/2-c}e^{-\frac{p(V-1)+\sqrt{\frac{4}{1-2c}}}{1-2c}}$$

Furthermore, take $n \geq \frac{9(V-1)}{2L}$, such that $c \leq 1/6$. Also use the fact that $(V-1)p \leq 1$, simplify the expression

$$\begin{aligned}\sup_{\boldsymbol{b},\delta,c,p} \mathbb{E}_{Z^n}\left[\frac{1}{\kappa+\delta}\left(\widetilde{R}_{\mathcal{A}_\kappa}(f) - \widetilde{R}_{\mathcal{A}_\kappa}(\tilde{f}^*)\right)\right] &\geq \frac{1-\epsilon}{\kappa}\sqrt{\frac{V-1}{8nL}}\frac{L}{1/2-1/6}e^{-\frac{1+\sqrt{6}}{1-1/3}} \\ &\geq \frac{1-\epsilon}{\kappa}\sqrt{\frac{3(V-1)L}{8n}}e^{-6}\end{aligned}$$

In order to satisfy the condition $L \leq (V-1)p(1/2-c)$ and $(V-1)p \leq 1$, plug in $c = \sqrt{\frac{(V-1)}{8nL}}$, we have $n \geq \frac{V-1}{2L(1-2L)^2}$.

Compare two parts together

$$\sup_{(P_X,\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}})\in\Pi(\epsilon,\kappa)} \mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] \geq \frac{K-1}{K}\epsilon + \frac{1-\epsilon}{\kappa}\sqrt{\frac{3(V-1)L}{8n}}e^{-6},$$

for $n \geq \frac{V-1}{2L}\max\left\{9, \frac{1}{(1-2L)^2}\right\}$.

■

A.1.3 Proof of Upper Bounds: Lemma 1 and Theorem 2

Lemma 1 (Oracle Inequality under Feature-dependent Label Noise) *For any $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ and any classifier f, we have*

$$R(f) - R(f^*) \leq \inf_{\kappa}\left\{P_X\Big(\mathcal{X}\setminus\mathcal{A}_\kappa\left(\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}}\right)\Big) + \frac{1}{\kappa}\left(\widetilde{R}(f) - \widetilde{R}\left(\tilde{f}^*\right)\right)\right\}.$$

Proof. For any $\kappa \geq 0$, the input space $\mathcal{X}$ can be divided into two regions: $\mathcal{X}\setminus\mathcal{A}_\kappa$ and $\mathcal{A}_\kappa$.

For any f, its risk can be decomposed into two parts

$$\begin{aligned}
R(f) &= \mathbb{E}_{X,Y}\left[\mathbb{1}_{f(X)\neq Y}\right] \\
&= \mathbb{E}_X\mathbb{E}_{Y|X}[\mathbb{1}_{f(X)\neq Y}] \\
&= \mathbb{E}_X\mathbb{E}_{Y|X}[1 - \mathbb{1}_{f(X)=Y}] \\
&= \mathbb{E}_X\left[1 - [\boldsymbol{\eta}(X)]_{f(X)}\right] \\
&= \int_{\mathcal{X}}\left(1 - [\boldsymbol{\eta}(x)]_{f(x)}\right)dP_X(x) \\
&= \int_{\mathcal{X}\setminus\mathcal{A}_\kappa}\left(1 - [\boldsymbol{\eta}(x)]_{f(x)}\right)dP_X(x) + \int_{\mathcal{A}_\kappa}\left(1 - [\boldsymbol{\eta}(x)]_{f(x)}\right)dP_X(x)
\end{aligned}$$

Thereofore, excess risk equals to

$$R(f) - R(f^*) \underbrace{\int_{\mathcal{X}\setminus\mathcal{A}_\kappa}\Big(\max\boldsymbol{\eta}(x) - [\boldsymbol{\eta}(x)]_{f(x)}\Big)\,dP_X(x)}_{ⓐ} + \underbrace{\int_{\mathcal{A}_\kappa}\Big(\max\boldsymbol{\eta}(x) - [\boldsymbol{\eta}(x)]_{f(x)}\Big)\,dP_X(x)}_{ⓑ}$$

Now examine the two terms separately,

$$ⓐ \leq \int_{\mathcal{X}\setminus\mathcal{A}_\kappa} 1\,dP_X(x) = P_X\Big(\mathcal{X}\setminus\mathcal{A}_\kappa\left(\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}}\right)\Big),$$

and

$$\begin{aligned}
ⓑ &< \int_{\mathcal{A}_\kappa}\frac{1}{\kappa}\Big(\max\widetilde{\boldsymbol{\eta}}(x) - [\widetilde{\boldsymbol{\eta}}(x)]_{f(x)}\Big)\,dP_X(x) \quad \because \text{by definition of relative signal strength} \\
&\leq \int_{\mathcal{X}}\frac{1}{\kappa}\Big(\max\widetilde{\boldsymbol{\eta}}(x) - [\widetilde{\boldsymbol{\eta}}(x)]_{f(x)}\Big)\,dP_X(x) \\
&= \frac{1}{\kappa}\left(\widetilde{R}(f) - \widetilde{R}(\tilde{f}^*)\right) \quad \because \text{by definition of } \widetilde{R}.
\end{aligned}$$

Since this works for any $\kappa > 0$, we then have

$$R(f) - R(f^*) \leq \inf_{\kappa>0}\left\{P_X\Big(\mathcal{X}\setminus\mathcal{A}_\kappa\left(\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}}\right)\Big) + \frac{1}{\kappa}\left(\widetilde{R}(f) - \widetilde{R}\left(\tilde{f}^*\right)\right)\right\}.$$

■

Now, to prepare for the rate of convergence proof, we first introduce the concept of shattering in multiclass and the Natarajan dimension [Natarajan, 1989], which can be viewed as a multi-class analogy of VC dimension [Vapnik and Chervonenkis, 1971].

Definition 2 (Shattering (Multiclass)) *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y} = \{1, 2, \ldots, K\}$. For any set containing n distinct elements $C_n = \{x_1, \ldots, x_n\} \subset \mathcal{X}$, denote*

$$\mathcal{H}_{C_n} = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\},$$

and therefore $|\mathcal{H}_{C_n}|$ is the number of distinct vectors of length n that can be realized by functions in $\mathcal{H}$.

The n^{th} shatter coefficient *is defined as*

$$S(\mathcal{H}, n) := \max_{C_n} |\mathcal{H}_{C_n}|.$$

We say that a set C_n is shattered by $\mathcal{H}$ if there exists $f, g : C_n \to \mathcal{Y}$ such that for every $x \in C_n$, $f(x) \neq g(x)$, and

$$\mathcal{H}_C \supseteq \{f(x_1), g(x_1)\} \times \{f(x_2), g(x_2)\} \times \cdots \times \{f(x_n), g(x_n)\}$$

If $\mathcal{Y} = \{1, 2\}$, this definition reduces the binary notion of shattering which says all labeling of points can be realized by some function in the hypothesis class $\mathcal{H}$, i.e., $\mathcal{H}_C = \{1, 2\}^{|C|}$. Contrary to the intuition that multiclass shattering means being able to label all K possible labels for each point $x \in C$: $\mathcal{H}_C = \{1, 2, \ldots, K\}^{|C|}$. This definition is more like "embed the binary cube into multiclass", where every $x \in C$ is allowed to pick from two labels rather than from K labels.

Definition 3 (Natarajan Dimension) *The Natarajan dimension of $\mathcal{H}$, denoted Ndim($\mathcal{H}$), is the maximal size of a shattered set $C \in \mathcal{X}$.*

Theorem (Excess Risk Upper Bound of NI-ERM) *Let $\epsilon \in [0, 1], \kappa \in (0, +\infty)$. Consider any $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa)$, assume function class $\mathcal{F}$ has Natarajan dimension V, and the noisy Bayes classifier $\tilde{f}^*$ belongs to $\mathcal{F}$. Let $\hat{f} \in \mathcal{F}$ be the ERM trained on $Z^n = \left\{(X_i, \widetilde{Y}_i)\right\}_{i=1}^n$, then*

$$\begin{aligned}\mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*)\right] &\leq \epsilon + \frac{1}{\kappa} \cdot 8\sqrt{\frac{V \log 2n + 2V \log K + 4}{2n}} \\ &\leq \epsilon + \mathcal{O}\left(\frac{1}{\kappa}\sqrt{\frac{V}{n}}\right) \quad \textit{up to log factor.}\end{aligned}$$

Proof. Following directly from Lemma 1, with $(P_X, \boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}) \in \Pi(\epsilon, \kappa)$, we already have

$$\begin{aligned}R(f) - R(f^*) &\leq P_X\left(\mathcal{X} \setminus \mathcal{A}_\kappa\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right)\right) + \frac{1}{\kappa}\left(\widetilde{R}(f) - \widetilde{R}\left(\tilde{f}^*\right)\right) \\ &\leq \epsilon + \frac{1}{\kappa}\left(\widetilde{R}(f) - \widetilde{R}\left(\tilde{f}^*\right)\right).\end{aligned}$$

Now replace f with NI-ERM $\hat{f}$, the following procedure is similar to the derivation of generalization error based on VC dimension in binary classification setup, except using a "Natarajan version" of Sauer's lemma.

Let's start from excess risk bound based on Shattering coefficient.

Lemma 2

$$\mathbb{E}_{Z^n}\left[\widetilde{R}\left(\hat{f}\right) - \widetilde{R}\left(\tilde{f}^*\right)\right] \leq 8\sqrt{\frac{\log(8eS(\mathcal{H}, 2n))}{2n}}$$

Proof Sketch. First convert excess risk $\widetilde{R}\left(\hat{f}\right) - \widetilde{R}\left(\tilde{f}^*\right)$ to generalization error with the inequality ([Devroye et al., 1996, Lemma 8.2])

$$\widetilde{R}\left(\hat{f}\right) - \widetilde{R}\left(\tilde{f}^*\right) \leq 2 \sup_{f \in \mathcal{F}} \left| \widetilde{R}(f) - \widetilde{R}_n\left(f\right) \right|,$$

where $\widetilde{R}_n$ denotes the empirical risk evaluated on training sample Z^n.

Then use the "symmetrization lemma" [Vapnik and Chervonenkis, 1971, Lemma 2], for $n\epsilon^2 \geq 2$,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \widetilde{R}(f) - \widetilde{R}_n\left(f\right) \right| \geq \epsilon \right) \leq 2\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \widetilde{R}'_n(f) - \widetilde{R}_n\left(f\right) \right| \geq \frac{\epsilon}{2} \right)$$

to convert the population risk $\widetilde{R}(f)$ into $\widetilde{R}'_n(f)$, the risk evaluated on the "ghost sample", a second, independent sample of n points.

Finally, with union bound and Hoeffding's inequality, we can bound the error probability with shattering coefficient on $2n$ points

$$\begin{aligned}
2\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \widetilde{R}'_n(f) - \widetilde{R}_n\left(f\right) \right| \geq \frac{\epsilon}{2} \right) &\leq 2S(\mathcal{H}, 2n)\mathbb{P}\left(\left| \widetilde{R}'_n(f) - \widetilde{R}_n\left(f\right) \right| \geq \frac{\epsilon}{2} \right) \\
&\leq 2S(\mathcal{H}, 2n)\left\{ \mathbb{P}\left(\left| \widetilde{R}'_n(f) - \widetilde{R}\left(f\right) \right| \geq \frac{\epsilon}{4} \right) + \mathbb{P}\left(\left| \widetilde{R}'_n(f) - \widetilde{R}\left(f\right) \right| \geq \frac{\epsilon}{4} \right) \right\} \\
&\leq 2S(\mathcal{H}, 2n)\left\{ 2 \cdot 2 \cdot e^{-2n\left(\frac{\epsilon}{4}\right)^2} \right\} \\
&\leq 8S(\mathcal{H}, 2n)e^{-\frac{n\epsilon^2}{8}},
\end{aligned}$$

notice that the inequality also holds for $n\epsilon^2 < 2$ because in this case the upper bound is bigger than 1. Therefore we get the convergence in probability bound for ERM $\hat{f}$,

$$\mathbb{P}\left(\widetilde{R}\left(\hat{f}\right) - \widetilde{R}\left(\tilde{f}^*\right) \geq \epsilon \right) \leq 8S(\mathcal{H}, 2n)e^{-\frac{n\epsilon^2}{32}}.$$

Convert it to expectation using the inequality from Devroye et al. [1996, Problem 12.1], we finally get

$$\mathbb{E}_{Z^n}\left[\widetilde{R}\left(\hat{f}\right) - \widetilde{R}\left(\tilde{f}^*\right) \right] \leq 8\sqrt{\frac{\log(8eS(\mathcal{H}, 2n))}{2n}}.$$

■

The last step is to bound multiclass shattering coefficient with Natarajan dimension. We leverage the following lemma.

Lemma 3 (Natarajan [1989]) *Let C and $\mathcal{Y}$ be two finite sets and let $\mathcal{H}$ be a set of functions from C to $\mathcal{Y}$.*

$$|\mathcal{H}| \leq |C|^{Ndim(\mathcal{H})} \cdot |\mathcal{Y}|^{2Ndim(\mathcal{H})}.$$

Replacing shattering coefficient with Natarajan dimension, we get

$$\begin{aligned}
\mathbb{E}_{Z^n}\left[\widetilde{R}\left(\hat{f}\right) - \widetilde{R}\left(\tilde{f}^*\right) \right] &\leq 8\sqrt{\frac{\log 8e + \log\left((2n)^V K^{2V}\right)}{2n}} \\
&\leq 8\sqrt{\frac{V\log 2n + 2V\log K + 4}{2n}}
\end{aligned}$$

Putting things together,

$$\mathbb{E}_{Z^n}\left[R\left(\hat{f}\right) - R(f^*) \right] \leq \epsilon + \frac{1}{\kappa} \cdot 8\sqrt{\frac{V\log 2n + 2V\log K + 4}{2n}}.$$

■

A.2 Experimental Details

A.2.1 2D Gaussian with synthetic label noise

For 2D Gaussian mixture data, we draw from two Gaussian centered at $[1\ 1]^\top$ and $[-1\ -1]^\top$, with covariance matrix being identity, 200 data points from each, with label $Y = 1, 2$ respectively. To generate noisy labels, we flip every label uniformly with some probability. We use Sklearn's logistic regression (with no ℓ_2 regularization). The experiment was conducted on AMD Ryzen 5 3600 CPU. The goal of the simulation is to experimentally verify noise immunity results in Section **??**. Notice that different trial corresponds to different draw of both instances and noisy labels.

Table 2: Testing accuracy of logistic regression on gaussian mixture data with uniform label noise. "Noise rate" refers to $\mathbb{P}(\widetilde{Y} \neq Y)$, the percentage of wrong labels in the training data. As theory in Section **??** predicts, when $\mathbb{P}(\widetilde{Y} \neq Y)$ reach 50%, there is a sharp decrease in performance.

Noise rate	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
Trial #1	93.00	92.83	92.38	92.08	91.78	91.93	92.25	92.90	91.83	92.58	74.68	25.12	9.70	7.73	7.52	7.25	7.38	7.15	7.18	7.10	7.00
Trial #2	91.73	91.60	92.05	91.63	91.78	91.78	91.68	91.63	91.55	91.48	80.40	21.10	9.93	8.55	8.38	8.22	8.20	8.35	8.33	8.40	8.28
Trial #3	92.73	92.75	92.78	92.78	92.58	92.45	91.68	88.15	82.58	59.83	49.53	35.80	21.28	14.35	9.33	8.53	8.12	7.70	7.13	7.23	7.28
Trial #4	91.55	91.58	91.60	91.63	91.68	91.60	91.25	90.98	89.98	86.38	60.53	9.95	8.75	10.00	10.45	9.08	9.00	9.53	9.20	9.03	8.45
Trial #5	91.55	91.58	91.60	91.63	91.68	91.60	91.25	90.98	89.98	86.38	60.53	9.95	8.75	10.00	10.45	9.08	9.00	9.53	9.20	9.03	8.45
Mean	92.11	92.07	92.08	91.95	91.90	91.87	91.62	90.93	89.18	83.33	65.13	20.40	11.68	10.10	9.23	8.43	8.34	8.45	8.21	8.16	7.89
Std	0.70	0.66	0.51	0.50	0.38	0.35	0.41	1.74	3.79	13.44	12.35	10.94	5.39	2.56	1.29	0.75	0.68	1.07	1.03	0.94	0.70

A.2.2 MNIST with synthetic label noise

We flip the clean training label of MNIST (`http://yann.lecun.com/exdb/mnist/`) uniformly (to any of the wrong classes). We use a shallow neural network with two convolution layers and two fully connected layers. We train with stochastic gradient descent with learning rate 0.01 for 10 epochs, batch size equals 64. We use the same hyperparamters for all tests. The experiments were conducted on a single NVIDIA GTX 1660S GPU. The goal of the simulation is to experimentally verify noise immunity results in Section **??**. Here randomness corresponds to different realization of noisy labels and stochastic gradient descent.

Table 3: Testing accuracy of a shallow CNN (2 conv layers with 2 fully connected layers) on MNIST with uniform label noise. "Noise rate" refers to $\mathbb{P}(\widetilde{Y} \neq Y)$, the percentage of wrong labels in the training data. As theory in Section **??** predicts, when $\mathbb{P}(\widetilde{Y} \neq Y)$ reach 90%, there is a sharp decrease in performance.

Noise rate	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
Trial #1	98.97	98.89	98.81	98.46	98.49	98.16	98.46	98.07	97.98	97.57	97.88	97.84	97.19	97.10	96.70	95.02	89.00	83.72	11.58	0.17	0.03
Trial #2	98.88	98.73	98.94	98.55	98.72	98.66	98.50	98.24	98.15	98.23	97.86	97.98	97.70	97.10	96.91	95.76	91.99	88.49	9.99	0.08	0.04
Trial #3	99.00	99.04	98.86	98.56	98.69	98.66	98.51	98.49	98.37	98.25	98.25	97.39	97.37	97.18	96.66	94.88	92.15	81.48	6.19	0.14	0.04
Trial #4	99.04	98.86	98.70	98.76	98.83	98.65	98.34	98.42	98.58	98.47	98.00	97.41	97.63	97.09	96.46	95.94	93.19	84.78	8.68	0.19	0.01
Trial #5	99.05	98.58	98.89	98.82	98.72	98.83	98.34	98.55	98.40	98.38	98.01	97.31	97.33	96.21	96.29	94.92	90.38	85.84	8.98	0.13	0.08
Mean	98.99	98.82	98.84	98.63	98.69	98.59	98.43	98.35	98.30	98.18	98.00	97.59	97.44	96.94	96.60	95.30	91.34	84.86	9.08	0.14	0.04
Std	0.07	0.17	0.09	0.15	0.12	0.25	0.08	0.20	0.23	0.36	0.16	0.30	0.21	0.41	0.24	0.51	1.65	2.59	1.98	0.04	0.03

A.2.3 CIFAR with synthetic label noise

We flip the clean training label of CIFAR-10 (`https://www.cs.toronto.edu/~kriz/cifar.html`) uniformly (to any of the wrong classes). To have a fair comparison between different methods, we fix the realization of noisy labels. Follow the 2-step procedure described in Section 5, we use different pre-trained neural networks as feature extractor: forward-passing the training image thought the network and record the feature. Then use sklearn's (`https://scikit-learn.org/stable/`) logistic regression function to fit the (feature, noisy label) pair in a full batch manner. We pre-specify a range of values for ℓ_2 regularization ($\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$) and number of iterations for lbfgs optimizer ($\{10, 20, 50, 100\}$), then do cross-validation on noisy data to pick the best hyper-parameters. We use the same range of hyper-parameters in all tests. The experiments were conducted on a single NVIDIA Tesla V100 GPU. The result is deterministic.

Table 4: Peformance on CIFAR-10 with synthetic label noise. We apply linear model on top of different feature extractors: "ResNet-50 TL" refers to using a pre-trained ResNet-50 on ImageNet [Deng et al., 2009] (available in Pytorch model library) in a transfer learning fashion, "ResNet-50 SSL" refers to using a pre-trained ResNet-50 on unlabeled CIFAR data with self-supervised loss [Chen et al., 2020] (publicly downloadable weights `https://github.com/ContrastToDivide/C2D?tab=readme-ov-file`) and "DINOv2 SSL" refers to using the self-supervised foundation model DINOv2 [Oquab et al., 2023] (available at `https://github.com/facebookresearch/dinov2`) as the feature extractor. "Noise rate" refers to $\mathbb{P}\big(\widetilde{Y} \neq Y\big)$, the percentage of wrong labels in the training data. As theory in Section **??** predicts, when $\mathbb{P}\big(\widetilde{Y} \neq Y\big)$ reach 90%, there is a sharp decrease in performance. We employed Python's sklearn logistic regression and cross-validation functions without data augmentation. The results are deterministic.

Noise rate	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.85	0.9	0.95	1
Linear	41.37	41.09	40.97	40.37	40.45	39.44	37.28	35.20	26.74	18.00	10.28	5.50	3.92
Linear + ResNet-50 TL	90.17	89.58	89.01	88.27	87.55	87.28	86.40	85.01	82.03	74.02	10.82	1.47	0.26
Linear + ResNet-50 SSL	92.48	92.26	91.74	91.46	91.13	90.33	91.07	90.99	89.11	83.89	10.08	1.31	0.34
Linear + DINOv2 SSL	99.25	99.27	99.23	99.14	99.10	99.11	99.02	98.84	95.50	76.91	10.13	0.92	0.03

A.2.4 CIFAR with human label error

We load the noisy human labels provided by `http://noisylabels.com/`, then follow exact the same procedure as above.

Table 5: Performance on CIFAR-N dataset (`http://noisylabels.com/`) in terms of testing accuracy. "Aggre", "Rand1", . . . , "Noisy" denote various types of human label noise. We apply linear model on top of different feature extractors: "ResNet-50 TL" refers to using a pre-trained ResNet-50 on ImageNet [Deng et al., 2009] in a transfer learning fashion, "ResNet-50 SSL" refers to using a pre-trained ResNet-50 on unlabeled CIFAR data with self-supervised loss [Chen et al., 2020] and "DINOv2 SSL" refers to using the self-supervised foundation model DINOv2 [Oquab et al., 2023] as the feature extractor. We employed Python's sklearn logistic regression and cross-validation functions without data augmentation; the results are deterministic and directly reproducible.

Methods	CIFAR-10N					CIFAR-100N
	Aggre	Rand1	Rand2	Rand3	Worst	Noisy
Linear	40.73	40.41	40.31	40.63	38.43	16.61
Linear + ResNet-50 TL	89.18	88.63	88.61	88.66	85.32	62.89
Linear + ResNet-50 SSL	91.78	91.66	91.39	91.28	87.84	57.95
Linear + DINOv2 SSL	98.69	98.80	98.65	98.67	95.71	83.17

A.2.5 Linear probing, then fine tuning (LP-FT)

We study whether LP-FT works in label noise learning scenario.

Table 6: Performance on CIFAR-N dataset (`http://noisylabels.com/`) in terms of testing accuracy. "Aggre", "Rand1", . . . , "Noisy" denote various types of human label noise. We apply linear model on top of different feature extractors: "ResNet-50 TL" refers to using a pre-trained ResNet-50 on ImageNet [Deng et al., 2009] in a transfer learning fashion, "ResNet-50 SSL" refers to using a pre-trained ResNet-50 on unlabeled CIFAR data with contrastive loss [Chen et al., 2020] and "DINOv2 SSL" refers to using the self-supervised foundation model DINOv2 [Oquab et al., 2023] as the feature extractor.

Feature	Method	CIFAR-10N						CIFAR-100N	
		Clean	Aggre	Rand1	Rand2	Rand3	Worst	Clean	Noisy
ResNet-50 TL	LP	90.17	89.18	88.63	88.61	88.66	85.32	71.79	62.89
	LP-FT	95.94	92.03	88.55	87.78	87.82	71.88	82.3	63.85
ResNet-50 SSL	LP	92.54	91.78	91.66	91.46	91.17	87.85	69.88	57.98
	LP-FT	94.11	89.11	84.49	83.75	84.15	65.00	74.41	54.49
DINOv2 (small) SSL	LP	96.09	94.8	94.39	94.42	94.35	91.14	83.82	72.46
	LP-FT	98.23	93.29	88.03	87.27	86.94	67.42	89.97	64.81

A.2.6 Robust learning strategy over DINOv2 feature

This section examines how different robust learning strategy works over DINOv2 feature, compared with only training with cross entropy.

Feature	Method	CIFAR-10N						CIFAR-100N	
		Clean	Aggre	Rand1	Rand2	Rand3	Worst	Clean	Noisy
DINOv2 SSL	CE	99.25	98.69	98.8	98.65	98.67	95.71	92.85	83.17
	MAE	99.27	99.04	99.01	99.09	99.11	95.55	90.68	82.55
	Sigmoid	99.26	98.86	98.91	98.87	98.96	96.66	92.82	82.03
	CE + SAM	99.09	97.66	98.47	98.53	98.47	95.47	89.97	82.85

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are either supported by theory statements or by reproducible experiment results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations about our practical method is described.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions are stated in the theorem statement. Full proofs are included in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Important information about the experiments are in main text. Details on the experimental setup is described in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is provided, common benchmark datase were used, instructions are given, the result is easily reproducible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See appendix and attached code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have done repeated experiments for data simulation, e.g., different iid draw. For the result on CIFAR-N data challenge, our result is deterministic and thus no error bar.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have read the NeurIPS Code of Ethics and confirm that this research follows the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theory-oriented paper towards better understanding of label noise problem.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Citations and urls are included.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve with this matter.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.