IDENTIFIABILITY OF LABEL NOISE TRANSITION MATRIX

Yang Liu[1], **Hao Cheng**[1], **Kun Zhang**[2,3]
[1] University of California, Santa Cruz
[2] Carnegie Mellon University
[3] Mohamed bin Zayed University of Artificial Intelligence
{yangliu, haocheng}@ucsc.edu, kunz1@cmu.edu

ABSTRACT

The noise transition matrix plays a central role in the problem of learning with noisy labels. Among many other reasons, a large number of existing solutions rely on access to it. Identifying and estimating the transition matrix without ground truth labels is a critical and challenging task. When label noise transition depends on each instance, the problem of identifying the instance-dependent noise transition matrix becomes substantially more challenging. Despite recent works proposing solutions for learning from instance-dependent noisy labels, the field lacks a unified understanding of when such a problem remains identifiable. The goal of this paper is to characterize the identifiability of the label noise transition matrix. Building on Kruskal's identifiability results, we are able to show the necessity of multiple noisy labels in identifying the noise transition matrix for the generic case at the instance level. We further instantiate the results to explain the successes of the state-of-the-art solutions and how additional assumptions alleviated the requirement of multiple noisy labels. Our result also reveals that disentangled features are helpful in the above identification task and we provide empirical evidence.

1 Introduction

The literature of learning with noisy labels concerns the scenario when the observed labels $\tilde{Y}$ can differ from the true one Y. The noise transition matrix $T(X)$, defined as the transition probability from Y to $\tilde{Y}$ given X, plays a central role in this problem. Among many other benefits, the knowledge of $T(X)$ has demonstrated its use in performing either risk [1, 2], or label [2], or constraint corrections [3]. In beyond, it also finds applications in ranking small loss samples [4] and detecting corrupted samples [5]. On the other hand, applying the wrong transition matrix $T(X)$ can lead to a number of issues. The literature has well-documented evidence that a wrongly inferred transition matrix can lead to performance drops [1, 6, 7, 8], and false sense of fairness [3, 6]. Knowing whether a $T(X)$ is identifiable or not helps understand if the underlying noisy learning problem is indeed learnable.

The earlier results have focused on class- but not instance-dependent transition matrix $T(X) \equiv T := [\mathbb{P}(\tilde{Y} = j|Y = i)]_{i,j}, \forall X$. The literature has provided discussions of the identifiability of T under the mixture proportion estimation setup [9], and has identified a reducibility condition for inferring the inverse noise rate. Later works have developed a sequence of solutions to estimate T under a variety of assumptions, including irreducibility [9], anchor points [10, 7, 11], separability [12], rankability [13, 14], redundant labels [15], clusterability [8], among others [16, 17].

The question of identifying and estimating T becomes much trickier when the noise transition matrix is instance-dependent. The potentially complicated dependency between X and $T(X)$ renders it even less clear whether solving this problem is viable or not. We observe a recent surge of different solutions towards solving the instance-dependent label noise problem [12, 18, 19, 20]. Some of the results took on the problem of estimating $T(X)$, while the others proposed solutions to learn directly from instance-dependent noisy labels. We will survey these results in Section 1.1.

Despite the above successes, there lacks a unified understanding of when this learning from instance-dependent noisy label problem is indeed identifiable and therefore learnable. The mixture of different observations calls for the need for demystifying: (1) Under what conditions are the noise transition matrices $T(X)$ identifiable? (2) When and why do the existing solutions work when handling the instance-dependent label noise? (3) When $T(X)$ is not identifiable, what can we do to improve its identifiability? Providing answers to these questions will be the primary focus of this paper. The main contributions of this paper are to characterize the identifiability of instance-dependent label noise, use them to

explain the current existing results and point out possible directions to improve. Among other findings, some highlights of the paper are 1. We find many existing solutions have a deep connection to the celebrated Kruskal's identifiability results that date back to the 1970s [21, 22]. 2. Three separate independent and identically distributed (i.i.d.) noisy labels (random variables) are both necessary and sufficient for instance-level identifiability. 3. Disentangled features help with identifiability.

Our paper will proceed as follows. Section 2 and 3 will present our formulation and the highly relevant preliminaries. Section 4 provides characterizations of the identifiability at the instance level and lays the foundations for our discussions. Section 5 extends the discussion to different instantiations that help us explain the sucess of existing solutions. Section 6 provides some empirical observations and Section 7 concludes this paper.

1.1 Related works

In the literature of learning with label noise, a major set of works focus on designing *risk-consistent* methods, i.e., performing empirical risk minimization (ERM) with specially designed loss functions on noisy distributions leads to the same minimizer as if performing ERM over the corresponding unobservable clean distribution. The *noise transition matrix* is a crucial component for implementing risk-consistent methods, e.g., loss correction [23], loss reweighting [24], label correction [25] and unbiased loss [1]. A number of solutions were proposed to estimate this transition matrix for class-dependent label noise, which we have discussed in the introduction.

To handle instance-dependent noise, recent solutions include estimating local transition matrices for different groups of data [18], using confidence scores to revise transition matrices [26], and using clusterability of the data [8]. More recent works have used the causal knowledge to improve the estimation [20], and the deep neural network to estimate the transition matrix defined between the noisy label and the Bayes optimal label [27]. Other works chose to focus on the learning from instance-dependent label noise directly, without explicitly estimating the transition matrix [28, 19, 29, 30, 31].

The identifiability issue with label noise has been discussed in the literature, despite not being formally treated. Relevant to us is the identifiability results studied in the Mixture Proportion Estimation setting [9, 32, 33]. We'd like to note that the identifiability was defined for the inverse noise rate, which differs from our focus on the noise transition matrix T. To our best knowledge, we are not aware of other works that specifically address the identifiability of $T(X)$, particularly for an instance-dependent label noise setting. Highly relevant to us is the Kruskal's identifiability results [21, 22, 34, 35], which reveals a sufficient condition for identifying a parametric model that links a hidden variable to a set of observed ones. Some of this paper's efforts include translating the Kruskal's and its follow-up results to the problem of learning with noisy labels.

2 Formulation

We will use (X, Y) to denote a supervised data in the form of (feature, label) drawn from an unknown distribution over $X \times Y$. We consider a K-class classification problem where the label $Y \in \{1, 2, ..., K\}$ with $K \geq 2$. In our setup, we do not observe the clean true label Y, but rather a noisy one, denoting by $\tilde{Y}$. The generation of $\tilde{Y}$ follows the following transition matrix: $T(X) := [\mathbb{P}(\tilde{Y} = j|Y = i, X)]_{i,j=1}^{K}$. That is $T(X)$ is a $K \times K$ matrix with its (i, j) entry being $\mathbb{P}(\tilde{Y} = j|Y = i, X)$.

To define identifiability, we will denote by Ω an observation space. We first define identifiability for a general parametric space Θ. Denote the distribution induced by the parameter $\theta \in \Theta$ of a statistical model on the observation space Ω as $\mathbb{P}_\theta$ [21, 35]. To give an example, for a fixed X (when consider instance-level identifiability), and Ω is simply the outcome space for its associated noisy label $\tilde{Y}$, i.e., $\{1, 2, ..., K\}$. In this case, each θ is the combination of a possible transition matrix $T(X)$ and the hidden prior of $\mathbb{P}(Y|X)$, which we use to denote the conditional probability distribution of Y given X. $\mathbb{P}_\theta$ is then the distribution (probability density function) $\mathbb{P}(\tilde{Y}|X)$. Later in Section 4 when we introduce three noisy labels $\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3$ for each X, $\mathbb{P}_\theta$ is the joint distribution $\mathbb{P}(\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3|X)$. Identifiability defines as follows:

Definition 1 (Identifiability)**.** *The parameter θ (statistical model) is identifiable if $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}, \forall \theta \neq \theta'$.*

We now formally define identifiability for the task of learning with noisy labels for an instance X. Denote by $\theta(X) := \{T(X), \mathbb{P}(Y|X)\}$. $\mathbb{P}_{\theta(X)}$ is the distribution (probability density function) over Ω, defined by the noise transition matrix $T(X)$ and the prior $\mathbb{P}(Y|X)$. To emphasize, Ω is not necessarily the observation space of the noisy label $\tilde{Y}$ only. The exploration of an effective Ω will be one of the focuses of the paper.

Definition 2 (Identifiability of $T(X)$)**.** *For a given X, $T(X)$ is identifiable if $\mathbb{P}_{\theta(X)} \neq \mathbb{P}_{\theta'(X)}$ for $\theta(X) \neq \theta'(X)$, up to label permutation.*

Label permutation relabels the label space, e.g., $1 \to 2,\ 2 \to 1$, and the rows in $T(X)$ will swap.

3 Preliminary

In this section, we will introduce two highly relevant results on Mixture Proportion Estimation (MPE) [9] and Kruskal's identifiability result [21, 22].

3.1 Preliminary results using irreducibility and anchor points

The problem of learning from noisy labels ties closely to another problem called Mixture Proportion Estimation (MPE) [9], which concerns the following problem: let F, J, H be distributions defined over a Hilbert space $\mathcal{Z}$. The three relate to each other as follows: $F = (1-\kappa^*)J + \kappa^* H$. The identifiability problem concerns the ability to identify the mixture proportion κ^* from only observing F and H. The following identifiability result has been established:

Proposition 1. *[36] κ^* is identifiable if J is irreducible with respect to H, that J can not be written as $J = \gamma H + (1-\gamma)F'$, where $0 \leq \gamma \leq 1$, and F' is another distribution.*

Later, the anchor point condition [32], a stronger requirement of the irreducibility was established:

Proposition 2. *[32] κ^* is identifiable if there exists a subset $S \subseteq \mathcal{Z}$ such that $H(S) > 0$, but $\frac{J(S)}{H(S)} = 0$, where $J(S), H(S)$ denote the probabilities of S measured by J, H.*

The above set S is called an anchor set. A sequence of follow-up works have emphasized the necessity of anchor points in identifying a class-dependent noise transition matrix T [7, 17].

Prior work has established the connection between the MPE problem and the learning from noisy label one [32] for the identifiability of an inverse noise rate $\mathbb{P}(Y|\tilde{Y})$ but not the noise transition $T(X)$. We reproduce the discussion and fill in the gap. The discussion and results are for the class-dependent but not instance-dependent label noise, i.e., $T(X) \equiv T$, and for a binary classification problem. To follow the convention, we assume $Y \in \{-1, +1\}$. There are two things we need to do: (1) State the noisy label problem as an MPE one; and (2) show that the identifiability of κ^* is equivalent to the identifiability of T. We start with the first thing above. We want to acknowledge that this equivalence appeared before in [32, 33]. We reproduce it here to make our paper self-contained. We first present:

Lemma 1. *Denote by $\tilde{\pi}_- = \frac{\pi_-}{1-\pi_+}, \tilde{\pi}_+ = \frac{\pi_+}{1-\pi_-}$ and we have*

$$\mathbb{P}(X|\tilde{Y}=-1) = \tilde{\pi}_- \cdot \mathbb{P}(X|\tilde{Y}=+1) + (1-\tilde{\pi}_-) \cdot \mathbb{P}(X|Y=-1) \tag{1}$$

$$\mathbb{P}(X|\tilde{Y}=+1) = \tilde{\pi}_+ \cdot \mathbb{P}(X|\tilde{Y}=-1) + (1-\tilde{\pi}_+) \cdot \mathbb{P}(X|Y=+1)\,. \tag{2}$$

Now $\mathbb{P}(X|\tilde{Y}=+1), \mathbb{P}(X|\tilde{Y}=-1)$ correspond to the observed mixture distribution F, H, while $\mathbb{P}(X|Y=+1)$ and $\mathbb{P}(X|Y=-1)$ are the two unobserved Js, $\tilde{\pi}_-, \tilde{\pi}_+$ correspond to the mixture proportion κ^*. This has established the learning with noisy label problem as two MPE problems corresponding for the two associated distributions $\mathbb{P}(X|\tilde{Y}=-1), \mathbb{P}(X|\tilde{Y}=+1)$. Therefore to formally establish the equivalence between identifying κ^* and T, we will only need to establish the equivalence between identifying $\tilde{\pi}_-, \tilde{\pi}_+$ and identifying T. Denote by $e_+ := \mathbb{P}(\tilde{Y}=-1|Y=+1), e_- := \mathbb{P}(\tilde{Y}=+1|Y=-1)$ which determine the T for the binary case. We have the following equivalence theorem:

Theorem 3. *Identifying $\{\tilde{\pi}_-, \tilde{\pi}_+\}$ is equivalent with identifying $\{e_-, e_+\}$.*

The above theorem concludes the same irreducibility and anchor point conditions proposed under MPE also apply to identifying noise transition matrix T. This conclusion aligns with previous successes in estimating class-dependent noise transition matrix T when the anchor point conditions are satisfied [10, 7, 17]. The above result has **limitations**. Notably, the result focuses on two mixed distributions, leading to the binary classification setup in the noisy learning setting. The authors did not find an easy extension to the multi-class classification problem. Secondly, the translation to the noisy learning problem requires the noise transition matrix to stay the same for a distribution of X (e.g., $\mathbb{P}(X|\tilde{Y}=+1)$), instead of providing instance-level understanding for each X.

3.2 Kruskal's identifiability result

Our results build on the Kruskal's identifiability result [21, 22]. The setup is as follows: suppose that there is an unobserved variable Z that takes values in a K-sized discrete domain $\{1, 2, ..., K\}$. Z has a non-degenerate prior

$\mathbb{P}(Z=i)>0$. Instead of observing Z, we observe p variables $\{O_i\}_{i=1}^p$. Each O_i has a finite state space $\{1,2,...,\kappa_i\}$ with cardinality κ_i. Let M_i be a matrix of size $K \times \kappa_i$, which j-th row is simply $[\mathbb{P}(O_i=1|Z=j),...,\mathbb{P}(O_i=\kappa_i|Z=j)]$. In this case, $[M_1,M_2,...,M_p]$ and $\mathbb{P}(Z=i)$ are the hidden parameters that control the generation of observations - together, these form our θ. We now introduce the Kruskal rank of a matrix, which plays a central role in Kruskal's identifiability results.

Definition 3 (Kruskal rank). *[21, 22] For a matrix M, the Kruskal rank of M is the largest number I such that every set of I rows [1] of M are linearly independent.*

In this paper, we will use $\mathsf{Kr}(M)$ to denote the Kruskal rank of matrix M. To give an example, $M=\begin{bmatrix}1&0&0\\0&1&0\\2&0&0\end{bmatrix} \Rightarrow \mathsf{Kr}(M)=1$. This is because $[1,0,0]$ and $[2,0,0]$ are linearly dependent. We first reproduce the following theorem:

Theorem 4. *[21, 22, 34] The parameters $M_i, i=1,...,p$ are identifiable, up to label permutation, if*

$$\sum_{i=1}^{p}\mathsf{Kr}(M_i)\geq 2K+p-1 \tag{3}$$

The result for $p=3$ was first established in [22], and then it was shown in [34] that the proof extends to a general p. The proof builds on showing that different parameter θ leads to different stacking of Ms: $[M_1,...,M_p]$. For example, when $p=3$, $[M_1,M_2,M_3]$ forms the tensor of the observations.

4 Instance-Level Identifiability

This section will characterize the identifiability of $T(X)$ at the instance level.

4.1 Single noisy label might not be sufficient and setting up for multiple noisy labels

At a first sight, it is impossible to identify $\mathbb{P}(\tilde{Y}|Y,X)$ from only observing $\mathbb{P}(\tilde{Y}|X)$.[2] unless X satisfies the anchor point definition that $\mathbb{P}(Y=k|X)=1$ for a certain k: since $\mathbb{P}(\tilde{Y}|X)=\mathbb{P}(\tilde{Y}|Y,X)\cdot\mathbb{P}(Y|X)$, different combinations of $\mathbb{P}(\tilde{Y}|Y,X),\mathbb{P}(Y|X)$ can lead to the same $\mathbb{P}(\tilde{Y}|X)$. More specifically, consider the following example:

Example 1. *Suppose we have a binary classification problem with $T(X)=\begin{bmatrix}1-e_-(X)&e_-(X)\\e_+(X)&1-e_+(X)\end{bmatrix}$. Note that using chain rule we have*

$$\begin{aligned}\mathbb{P}(\tilde{Y}=+1|X)&=\mathbb{P}(\tilde{Y}=+1|Y=+1,X)\cdot\mathbb{P}(Y=+1|X)+\mathbb{P}(\tilde{Y}=+1|Y=-1,X)\cdot\mathbb{P}(Y=-1|X)\\&=(1-e_+(X))\cdot\mathbb{P}(Y=+1|X)+e_-(X)\cdot\mathbb{P}(Y=-1|X)\end{aligned}$$

Consider two cases: (1): $\mathbb{P}(Y=+1|X)=1,\ e_+(X)=e_-(X)=0.3$ and (2): $\mathbb{P}(Y=+1|X)=0.7,\ e_+(X)=0.1,\ e_-(X)=0.233$. Both cases will return the same $\mathbb{P}(\tilde{Y}=+1|X)=0.7$.

Is then the anchor point requirement necessary for identifying $T(X)$ at the instance level? The discussion in the rest of this section departs from the classical single noisy label setting. Instead, we assume for each instance X, we will have p conditionally independent (given X,Y) and identically distributed noisy labels $\tilde{Y}_1,...,\tilde{Y}_p$ generated according to $T(X)$. Let's assume for now we potentially have these labels. Later in this section, we discuss when having multiple redundant labels are possible, and connect to existing solutions in the literature in the next section.

Before we formally present the results for having multiple conditionally independent noisy labels, we offer intuitions. The reason behind this identifiability result ties close to latent class model [37] and tensor decomposition [38]. When the p noisy labels are conditionally independent given X and Y, we will have the joint distribution written as: $\mathbb{P}(\tilde{Y}_1,\tilde{Y}_2,...,\tilde{Y}_p|Y,X)=\prod_{i=1}^p\mathbb{P}(\tilde{Y}_i|Y,X)$ That is, the joint distribution of noisy labels can be encoded in a much smaller parameter space! In our setup, when we assume the i.i.d. $\tilde{Y}_i, i=1,2,...,p$ are generated according to the same transition matrix $T(X)$, the parameter space is fixed and determined by the size of $T(X)$. Yet, when we increase p, the observation space $\mathbb{P}(\tilde{Y}_1,\tilde{Y}_2,...,\tilde{Y}_p|Y,X)$ becomes richer to help us identify $T(X)$.

[1] There exists other definition that checks columns. Results would be symmetrical.

[2] We clarify that we will require knowing $\mathbb{P}(\tilde{Y}|X)$ - this requirement may appear weird when only one noisy label is sampled. But in practice, there are tools available to regress the posterior function $\mathbb{P}(\tilde{Y}|X)$ for each X.

4.2 The necessity of multiple noisy labels

We first define an *informative noisy label*.

Definition 4. *For a given (X, Y), we call their noisy label $\tilde{Y}$ informative if rank$(T(X)) = K$.*

Definition 4 simply requires the rank of $T(X)$ to be full, which is already assumed in the literature - e.g., loss correction [1, 2] would require the matrix has an inverse $T^{-1}(X)$, which is equivalent to $T(X)$ being full rank. In particular, it was required $e_+(X) + e_-(X) < 1$ in [1], which can be easily shown to imply $T(X)$ is full rank. Our first identifiability result states as follows:

Theorem 5. *With i.i.d. noisy labels, three informative noisy labels $\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3$ ($p = 3$) are both sufficient and necessary to identify $T(X)$.*

Proof sketch. We provide the key steps of the proof. The full proof can be found in the supplemental material. We first prove sufficiency. We first relate our problem setting to the setup of Kruskal's identifiability scenario: $Y \in \{1, 2, ..., K\}$ corresponds to the unobserved hidden variable Z. $\mathbb{P}(Y = i)$ corresponds to the prior of this hidden variable. Each $\tilde{Y}_i, i = 1, ..., p$ corresponds to the observation O_i. κ_i is then simply the cardinality of the noisy label space, K. In the context of this theorem, $p = 3$, corresponds to the three noisy labels we have.

Each $\tilde{Y}_i$ corresponds to an observation matrix M_i: $M_i[j, k] = \mathbb{P}(O_i = k|Z = j) = \mathbb{P}(\tilde{Y}_i = k|Y = j, X)$. Therefore, by definition of M_1, M_2, M_3 and $T(X)$, they all equal to $T(X)$: $M_i \equiv T(X), i = 1, 2, 3$. When $T(X)$ has full rank, we know immediately that all rows in M_1, M_2, M_3 are independent. Therefore, the Kruskal ranks satisfy $\mathsf{Kr}(M_1) = \mathsf{Kr}(M_2) = \mathsf{Kr}(M_3) = K$. Checking the condition in Theorem 4, we easily verify

$$\mathsf{Kr}(M_1) + \mathsf{Kr}(M_2) + \mathsf{Kr}(M_3) = 3K \geq 2K + 2 \, .$$

Calling Theorem 4 proves the sufficiency.

To prove necessity, we need to prove less than 3 informative labels will not suffice to guarantee identifiability. The idea is to show that the two different sets of parameters $T(X)$ can lead to the same joint distribution $\mathbb{P}(\tilde{Y}_1, \tilde{Y}_2|X)$. We leave the detailed constructions to the supplemental material.

□

The above result points out that to ensure identifiability of $T(X)$ at the instance level, we would need three conditionally independent and informative noisy labels. This result coincides with a couple of recent works that promote the use of three redundant labels [15, 8]. Per our theorem, these two proposed solutions have a more profound connection to the identifiability of hidden parametric models, and three labels are not only algorithmically sufficiently, but also necessary.

The crowdsourcing community has been largely focusing on soliciting more than one label from crowdsourced workers, yet the learning from noisy label literature has primarily focused on learning from a single one. One of the primary motivations of crowdsourcing multiple noisy labels is indeed to aggregate them into a cleaner one [39, 40, 41], which serves as a pre-processing step towards solving the noisy learning problem. Nonetheless, our result demonstrates the other significance of having multiple labels - they help the learner identify the underlying true noise transition parameters. There have been discussions about crowdsourcing additional label information being unnecessary when we have robust learning strategies in hand. This section presents a strong case against the above argument.

5 Instantiations and Practical Implications of Our Identifiability Results

Most of the learning with noisy label solutions focuses on the case of using a single label and have observed empirical successes. In this section, we provide extensions of our results to cover of state-of-the-art learning with noisy label methods, together with specific assumptions over X, $T(X) = [\mathbb{P}(\tilde{Y}|Y, X)]$ etc. We show that our results can easily extend to these specific instantiations that successfully avoided the requirements of having multiple noisy labels for each X. The high-level intuition for Section 5.1 is to leverage the smoothness and clusterability of the nearest neighbor Xs so that their noisy labels will jointly serve as the multiple noisy labels for the local group. Section 5.2 and 5.3 build on the notion that if $T(X)$ is the same for a group of Xs, each group can then be treated as one "instance" and a "disentangled" version of X will become observation variables that serve the similar role of the additionally required noisy labels.

5.1 Leveraging smoothness and clusterability of X

We start with a discussion using the smoothness and clusterability of X. Recent results have explored the clusterability of Xs [8, 42] to infer the noise transition matrix using the clusterability of X:

Definition 5. *The 2-NN clusterability requires each X and its two nearest neighbors X_1, X_2 share the same true label Y, that is $Y = Y_1 = Y_2$.*

This definition effectively helps us avoid the need for multiple noisy labels per each X: one can view it as for each X, borrowing the noisy labels from its 2-NN, all together we have three independent noisy labels $\tilde{Y}, \tilde{Y}_1, \tilde{Y}_2$, all given the same Y. This smoothness or clusterability condition allows us to apply our identifiability results when one believes the $T(X)$ stays the same for the 2-NN nearest neighborhood X, X_1, X_2.

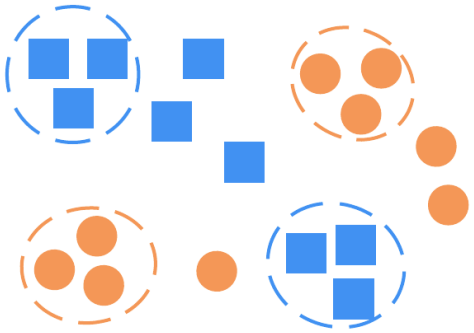

Figure 1: Data generation: label correlation among triplets.

But, when does an instance X and its 2-NN X_1, X_2 share the same true label? This requirement seems strange at the first sight: as long as $\mathbb{P}(Y|X), \mathbb{P}(Y_1|X_1)$ are not degenerate (being either 0 or 1 for different label classes), there always seems to be a positive probability that the realized $Y \neq Y_1$, no matter how close X and X_1 are. Nonetheless, empirically, the 2-NN requirement seems to hold well: according to [8] (Table 3 therein), when using a feature extractor built using the clean label, more than 99% of the instance satisfies the 2-NN condition. Even when using a feature extractor trained on noisy labels, the ratio is mostly always in or close to the 80% range.

The following data generation process for an unstructured discrete domain of classification problems [43, 44] helps us justify the 2-NN requirement. The intuition is that when Xs are informative and sufficiently discriminative, the similar Xs are going to enjoy the same true label.

- Let $\lambda = \{\lambda_1, ..., \lambda_n\}$ denote the priors for each $X \in \mathcal{X}$.
- For each $X \in \mathcal{X}$, sample a quantity q_X independently and uniformly from the set λ.
- The resulting probability mass function of X is given by $D(X) = \frac{q_X}{\sum_{X \in \mathcal{X}} q_X}$.
- A total of N Xs are observed. Denote by X_1, X_2 X's two nearest neighbors.
- Each (X, X_1, X_2) forms a triplet if $||X_1 - X||, ||X_2 - X||$ fall below a threshold ϵ (closeness).
- A single Y for the tuple (X, X_1, X_2) draws from $\mathbb{P}(Y|X, X_1, X_2)$.
- Based on Y, we further observe three $\tilde{Y}, \tilde{Y}_1, \tilde{Y}_2$ according to $\mathbb{P}(\tilde{Y}, \tilde{Y}_1, \tilde{Y}_2|Y)$.

The above data-generation process captures the correlation among Xs that are really close. We prove the above data generation process satisfies the 2-NN clusterability requirement with high probability.

Theorem 6. *When N is large enough such that $N > \frac{4\sum_{X \in \mathcal{X}} q_X}{\min_X q_X}$, w.p. at least $1 - Nexp(-2N)$, each X and its two nearest neighbor X_1, X_2 satisfy the 2-NN criterion.*

Smoothness conditions in semi-supervised learning This above discussion also ties closely to the smoothness requirements in semi-supervised learning [45, 46], where the neighborhood Xs can provide and propagate label information in each local neighborhood of Xs. Indeed, this idea echoes the co-teaching solution [47, 48] in the literature of learning with noisy labels, where a teacher/mentor network is trained to provide artificially generated noisy labels to supervise the training of the student network. Our identifiability result, to a certain degree, implies that the addition of the additional noisy supervision improves the chance for identifying $T(X)$. In [47, 48], counting the noisy label itself, and the "teacher" supervision, there are two such noisy supervision labels. This observation raises an interesting question: since our result emphasized three labels, does adding an additional teacher network for an additional supervision help? This question merits empirical verification. One caution we want to make is with more artificially inserted noisy labels, likely a subset of them will start to become dependent on each other, due to the similarities in training different models. This points out yet another interesting identifiability problem whose solution will help us detect the dependent labels and introduce additional and appropriate variables for modeling these dependencies.

5.2 Leveraging smoothness and clusterability of $T(X)$

In this section, we show that another "smoothness" assumption of $T(X)$ introduces new observation variables for us to identify $T(X)$. In Figure 2, we define variable $G = \{1, 2, ..., |G|\}$ to denote the group membership for each X. Consider a scenario that X can be grouped into $|G|$ groups such that each group of Xs share the same $T(X)$: $T(X_1) = T(X_2)$ if X_1, X_2 share the same group membership. We observe $G, X, \tilde{Y}$. This type of grouping has been observed in the literature:

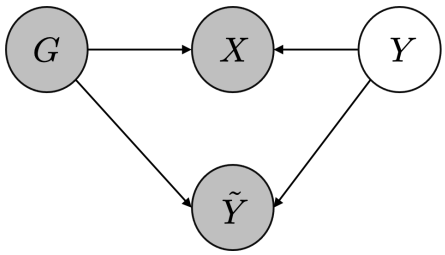

Figure 2: Causal graph for $(G, X, Y, \tilde{Y})$. Grey color indicates the variables are observable.

Class-dependent T Clearly, when $T(X) \equiv T$, the entire set of X can then be viewed as the group with $|G| = 1$.

Noise clusterability The noise transition estimator proposed in [8] was primarily developed for class-dependent but not instance-dependent $T(X)$. Nonetheless, a noise clusterability definition is introduced in [8] to allow the approach to be applied to the instance-dependent noise setting. Under the noise clusterability, using off-the-shelf clustering algorithms can help separate the dataset into local ones.

Group-dependent $T(X)$ Recent results have also studied the case that the data X can be grouped using additional information [3, 6, 49]. For instance, [3, 6] consider the setting where the data can be grouped by the associated "sensitive information", e.g., by age, gender, or race. Then the noise transition matrix remains the same for Xs that come from each group.

By this grouping, X becomes informative observations for each hidden Y and will fulfill the requirement of observing additional noisy labels. Below we formalize this discussion. We now define a disentangled feature and an informative feature: Denote by $R(X) \in \mathbb{R}^{d^*}$ a learned representation for X. Denote by R_i the random variable for $R_i(X), i = 1, 2, .., d^*$. For simplicity of the analysis, we assume each R_i has finite observation space $\mathcal{R}_i$ with cardinality $|\mathcal{R}_i| = \kappa_i$. We believe the result has implications for continuous feature space but the formalization of the results will be left for future technical developments. Define M_i for each R_i as $M_i[j, k] = \mathbb{P}(R_i = \mathcal{R}_i[k]|Y = j)$, where in above $\mathcal{R}_i[k]$ denotes the k-th element in $\mathcal{R}_i$.

Definition 6 (Disentangled R)**.** *R is disentangled if $\{R_i\}_{i=1}^{d^*}$ are conditional independent given Y.*

Definition 7 (Informative features)**.** *R_i is informative if its Kruskal rank is at least 2:* $\mathsf{Kr}(M_i) \geq 2$.

Assuming each X can be transformed into a set of disentangled features R, we prove:

Theorem 7. *For Xs in a given group $g \in G$, with a single informative noisy label, $T(X)$ is identifiable if the number of disentangled and informative features d^* satisfy that $d^* \geq K$.*

This result points out a new observation that even when we have a single noisy label, given a sufficient number of disentangled and informative features, the noise transition matrix T is indeed identifiable, without requiring either multiple noisy labels, or the anchor point condition.

The above result aligns with recent discussions of a neural network being able to disentangle features [50, 51] proves to be a helpful property. We establish that having disentangled feature helps identify $T(X)$. The required number of disentangled features grows linearly in K. When relaxing the unique identifiability to generic identifiability, i.e., the identifiability scenario has measure zero [35], the above theorem can be further extended to requiring $d^* \geq \lceil \log_2 \frac{K+2}{2} \rceil$. We leave the details in Appendix (Theorem 10). The above result is particularly informative when the number of classes K is large.

When the disentangled feature is not given, how do we disentangle X using only noisy labels to benefit from our results? In Section 6 we will test the effectiveness of a self-supervised representation learning approach that takes the side information relative to true label Y but operates independently from noisy labels. This result also implies when the noise rate is high such that $\tilde{Y}$ starts to become uninformative, dropping the noisy labels and focusing on obtaining the disentangled features helps with the identifiability of $T(X)$. This observation also helps explain recent successes in applying semi-supervised [19, 31, 52] and self-supervised [53, 54, 55] learning techniques to the problem of learning from noisy labels.

5.3 Smoothness and clusterability of $T(X)$ with unknown groupings

In practice, we often do not know the groupings of X that share the same $T(X)$, nor do we have a clear power (e.g., the noise clusterability condition) to separate the data into different groups. In reality, different from Figure 2, the group

membership can often remain hidden, if no additional knowledge of the data is solicited, leading to a situation in Figure 3.

It is a non-trivial task to jointly infer the group membership with $T(X)$. We first show that mixing the group membership can lead to non-negligible estimation errors. Suppose that there are two groups of X, each having a noise transition matrix $T_1(X), T_2(X)$. Suppose we ended up estimating one $T^*(X)$ for both groups mistakenly. We then have:

Theorem 8. *Any estimator $T^*(X)$ will incur at least the following estimation error:*

$$||T_1(X) - T^*(X)||_F + ||T_2(X) - T^*(X)||_F \geq \frac{1}{\sqrt{2}}||T_1(X) - T_2(X)||_F \tag{4}$$

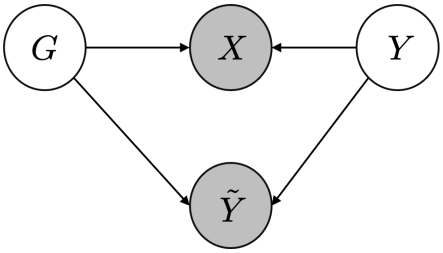

Figure 3: Causal graph with unobserved G. Grey indicates observable variables.

The above result shows the necessity of identifying G as well. Now we present our positive result on the identifiability when G is hidden too: Re-number the combined space of $G \times Y$ as $\{1, 2, ..., |G|K\}$. We are going to reuse the definition of M_i for each disentangled feature R_i: Define the "Kruskal matrix" for each R_i as $M_i[j, k] = \mathbb{P}(R_i = \mathcal{R}_i[k]|G \times Y = j)$.

Theorem 9. *For Xs in a given group $g \in G$, with a single informative noisy label, $T(X)$ is identifiable if the number of disentangled and informative features d^* satisfy that $d^* \geq 2|G|K - 1$.*

When we have unknown groups of noise, the requirement of the number of informative and disentangled features grows linearly in $|G|$. We now relate to the literature that implicitly groups Xs. We will use $\mathcal{X}$ to denote the space of all possible Xs.

Part-dependent label noise [18] discusses a part-dependent label noise model where each $T(X)$ can decompose into a linear combination of p parts: $T(X) = \sum_i^p \omega_i(X) \cdot T_i$. The motivation of the above model is each X can be viewed as a combination of multiple different sub-parts, and each of them has a certain difficulty being labeled. The hope is that the parameter space $\omega(X)$ can reduce the dependency between X and $T(X)$. Denote $\mathcal{W} := \{\omega(X) : X \in \mathcal{X}\}$. To put into our result, $|G| = |\mathcal{W}|$. If $\mathcal{W}$ has a much smaller space than $\mathcal{X}$, the condition specified in Theorem 9 would be more likely to be satisfied.

DNN approach [27] proposes using a deep neural network to encode the dependency between X and $T^*(X)$, with the only difference being that $T^*(X)$ is defined as the transition between $\tilde{Y}$ and the Bayes optimal label Y^*. Define: DNN $:= \{\text{DNN}(X) : X \in \mathcal{X}\}$. Similarly, in analogy to our results in Theorem 9, with replacing the hidden variable Y to Y^*, $|G|$ will be determined by |DNN|. So long as the DNN can identify the patterns in $T(X)$ and compress the space of DNN(X) as compared to $\mathcal{X}$, the identifiability becomes easier to achieve.

The causal approach [20] proposed improving the identifiability by exploring the causal structure. With causal inference approaches, one can identify a more representative and compressed $\tilde{X}$ for each X such that $\mathbb{P}(\tilde{Y}|Y, X, \tilde{X}) = \mathbb{P}(\tilde{Y}|Y, \tilde{X})$. Denote $\tilde{\mathcal{X}} := \{\tilde{X} : \tilde{X} \to X \in \mathcal{X}\}$. To plug in our results, we have $|G| = |\tilde{\mathcal{X}}|$.

6 Some Empirical Evidence: Disentangled Features

Most of our results above verified the empirical success of existing approaches and we refer the interested reader to the detailed experiments in the corresponding references. We now empirically show the possibility of learning disentangled features to help identify the noise transition matrix. We consider three types of encoders that are used to generate features. The first encoder is pre-trained by cross-entropy (CE) loss via a weakly supervised manner which is generally adopted in FW [23] and HOC [8]. However, since the training data is noisy, it is hard to guarantee that features are disentangled - this is our baseline. The second encoder is pre-trained by SimCLR [56] via a self-supervised manner. It is shown that the features trained by SimCLR are partly disentangled on some simple augmentation features such as rotation and colorization [57]. The third encoder is trained by IPIRM [57] via a self-supervised manner which can generate fully disentangled features. After training these three encoders, we fix the encoder and generate features from raw samples to estimate the noise transition matrix using HOC estimator [8]. We evaluate the performance via absolute estimation error defined below: err $= \frac{\sum_{i=1}^K \sum_{j=1}^K |\hat{T}_{i,j} - T_{i,j}|}{K^2} * 100$, where $\hat{T}$ is the estimated noise transition matrix, T is the real noise-transition matrix, K is the number of classes in the dataset, which is also the size of the transition matrix. The overall experiments are shown in Table 1. We can observe that the estimation error decreases as features become more disentangled which supports our analyses in the paper. We defer the details, more experiments, as well as experiments on comparing training performances using disentangled features, to the supplementary material.

Table 1: Comparison of transition matrix estimation error for different types of features on CIFAR-10. Each experiment is run 3 times and mean $\pm$ std is reported. *asymm.*: asymmetric label noise; *inst.*: instance-dependent label noise. Numbers are noise rates. All the encoders are from ResNet50 backbone.

Feature Type	*asymm. 0.3*	*asymm. 0.4*	*inst. 0.4*	*inst. 0.5*	*inst. 0.6*
Weakly-Supervised	14.51 ± 0.4	15.2 ± 0.02	8.39 ± 0.05	6.91 ± 0.06	6.18 ± 0.15
SimCLR	4.42 ± 0.01	4.41 ± 0.01	2.91 ± 0.02	2.55 ± 0.04	2.64 ± 0.03
IPIRM	$\mathbf{3.73 \pm 0.02}$	$\mathbf{3.74 \pm 0.01}$	$\mathbf{2.47 \pm 0.03}$	$\mathbf{2.20 \pm 0.02}$	$\mathbf{2.37 \pm 0.06}$

7 Concluding Remarks

This paper characterizes the identifiability of instance-level label noise transition matrix. We connect the problem to the celebrated Kruskal's identifiability result and present a necessary and sufficient condition for the instance-level identifiability. We extend and instantiate our results to practical settings to explain the successes of existing solutions. As a by-product, we show the importance of disentangled and informative features for identifying the noise transition matrix.

Future direction of work includes exploring the extension of our results to other weakly supervised learning settings (e.g., Positive Unlabeled learning, semi-supervised learning etc). Our results also encourage discussions on what assumptions are needed for the data in order to improve the identifiability of hidden factors.

References

[1] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[2] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 526–536, New York, NY, USA, 2021. Association for Computing Machinery.

[4] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pages 4006–4016. PMLR, 2020.

[5] Zhaowei Zhu, Zihao Dong, Hao Cheng, and Yang Liu. A good representation detects noisy labels. *arXiv preprint arXiv:2110.06283*, 2021.

[6] Yang Liu and Jialu Wang. Can less be more? when increasing-to-balancing label noise rates considered beneficial. NeurIPS'21.

[7] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32, 2019.

[8] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. *ICML*, 2021.

[9] Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*, 2015.

[10] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

[11] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *arXiv preprint arXiv:2006.07805*, 2020.

[12] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance-and label-dependent label noise. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, 2020.

[13] Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. *UAI*, 2017.

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

[15] Yang Liu, Juntao Wang, and Yiling Chen. Surrogate scoring rules and a dominant truth serum. *ACM EC*, 2020.

[16] Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. *arXiv preprint arXiv:2102.02414*, 2021.

[17] Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. *arXiv preprint arXiv:2102.02400*, 2021.

[18] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020.

[19] 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.

[20] Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34, 2021.

[21] Joseph B Kruskal. More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281–293, 1976.

[22] Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

[23] 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.

[24] 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.

[25] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015.

[26] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. *arXiv preprint arXiv:2001.03772*, 2020.

[27] Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent label-noise transition matrix using dnns. *arXiv preprint arXiv:2105.13001*, 2021.

[28] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. *CVPR*, 2021.

[29] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *International Conference on Machine Learning*, pages 825–836. PMLR, 2021.

[30] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*, 2020.

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

[32] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Gang Niu, Masashi Sugiyama, and Dacheng Tao. Towards mixture proportion estimation without irreducibility. *arXiv preprint arXiv:2002.03673*, 2020.

[33] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.

[34] Nicholas D Sidiropoulos and Rasmus Bro. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of chemometrics*, 14(3):229–239, 2000.

[35] Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

[36] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

[37] Clifford C Clogg. Latent class models. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 311–359. Springer, 1995.

[38] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

[39] Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. *Advances in neural information processing systems*, 25:692–700, 2012.

[40] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011.

[41] Yang Liu and Mingyan Liu. An online learning approach to improving the quality of crowd-sourcing. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '15, pages 217–230, New York, NY, USA, 2015. ACM.

[42] Dara Bahri, Heinrich Jiang, and Maya Gupta. Deep k-nn for noisy labels. In *International Conference on Machine Learning*, pages 540–550. PMLR, 2020.

[43] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[44] Yang Liu. Understanding instance-level label noise: Disparate impacts and treatments, 2021.

[45] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[46] Xiaojin Zhu. *Semi-supervised learning with graphs*. Carnegie Mellon University, 2005.

[47] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313. PMLR, 2018.

[48] 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. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

[49] Qizhou Wang, Jiangchao Yao, Chen Gong, Tongliang Liu, Mingming Gong, Hongxia Yang, and Bo Han. Learning with group noise. *arXiv preprint arXiv:2103.09468*, 2021.

[50] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[51] Xander Steenbrugge, Sam Leroux, Tim Verbelen, and Bart Dhoedt. Improving generalization for abstract reasoning tasks using disentangled feature representations. *arXiv preprint arXiv:1811.04784*, 2018.

[52] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. *arXiv preprint arXiv:1910.01842*, 2019.

[53] Hao Cheng, Zhaowei Zhu, Xing Sun, and Yang Liu. Demystifying how self-supervised features improve training from noisy labels. *arXiv preprint arXiv:2110.09022*, 2021.

[54] 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.

[55] 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.

[56] 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.

[57] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34, 2021.

[58] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13726–13735, 2020.

[59] 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.

[60] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

[61] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Appendix: Identifiability of Label Noise Transition Matrix

The Appendix is organized in the following way: Section A proves the Theorems in the main paper; Section B provides more discussions on generic identifiability; Section C provides more experiments on learning with noisy labels *w.r.t.* disentangled features and elaborates the detailed experimental settings in the paper.

A Omitted Proofs

Proof for Lemma 1

Proof. Using Bayes rule we easily obtain

$$\begin{aligned}\mathbb{P}(X|\tilde{Y}=+1) &= \mathbb{P}(X|Y=+1)\cdot\mathbb{P}(Y=+1|\tilde{Y}=+1)\\ &+\mathbb{P}(X|Y=-1)\cdot\mathbb{P}(Y=-1|\tilde{Y}=+1)\end{aligned} \tag{5}$$

The equality is due to the fact that $\tilde{Y}$ and X are assumed to be independent given Y. Similarly:

$$\begin{aligned}\mathbb{P}(X|\tilde{Y}=-1) &= \mathbb{P}(X|Y=+1)\cdot\mathbb{P}(Y=+1|\tilde{Y}=-1)\\ &+\mathbb{P}(X|Y=-1)\cdot\mathbb{P}(Y=-1|\tilde{Y}=-1)\end{aligned} \tag{6}$$

Denote by $\pi_+ := \mathbb{P}(Y=-1|\tilde{Y}=+1), \pi_- := \mathbb{P}(Y=+1|\tilde{Y}=-1)$. Since both $\mathbb{P}(X|Y=+1), \mathbb{P}(X|Y=-1)$ are unknown, solving Eqn. (5) and (6) we further have

$$\mathbb{P}(X|\tilde{Y}=-1) = \tilde{\pi}_-\cdot\mathbb{P}(X|\tilde{Y}=+1) + (1-\tilde{\pi}_-)\cdot\mathbb{P}(X|Y=-1) \tag{7}$$

$$\mathbb{P}(X|\tilde{Y}=+1) = \tilde{\pi}_+\cdot\mathbb{P}(X|\tilde{Y}=-1) + (1-\tilde{\pi}_+)\cdot\mathbb{P}(X|Y=+1). \tag{8}$$

□

Proof for Theorem 3

Proof. Further from $\tilde{\pi}_-, \tilde{\pi}_+$ we can solve and derive $\pi_- = \frac{\tilde{\pi}_-(1-\tilde{\pi}_+)}{1-\tilde{\pi}_-\tilde{\pi}_+}, \pi_+ = \frac{\tilde{\pi}_+(1-\tilde{\pi}_-)}{1-\tilde{\pi}_-\tilde{\pi}_+}$, establishing the equivalence between identifying $\tilde{\pi}_-, \tilde{\pi}_+$ with identifying π_-, π_+. Next we show that identifying π_-, π_+ is equivalent with identifying $\{e_+, e_-\}$.

We first show identifying $\{\pi_+, \pi_-\}$ suffices to identify $\{e_+, e_-\}$. To see this,

$$\mathbb{P}(\tilde{Y}=+1|Y=-1) = \frac{\mathbb{P}(Y=-1|\tilde{Y}=+1)\mathbb{P}(\tilde{Y}=+1)}{\mathbb{P}(Y=-1)}$$

And:

$$\mathbb{P}(Y=-1) = \mathbb{P}(Y=-1|\tilde{Y}=+1)\mathbb{P}(\tilde{Y}=+1) + \mathbb{P}(Y=-1|\tilde{Y}=-1)\mathbb{P}(\tilde{Y}=-1)$$

The derivation for $\mathbb{P}(\tilde{Y}=-1|Y=+1)$ is entirely symmetric. Since we directly observe $\mathbb{P}(\tilde{Y}=-1), \mathbb{P}(\tilde{Y}=+1)$, with identifying $\mathbb{P}(Y=+1|\tilde{Y}=-1), \mathbb{P}(Y=-1|\tilde{Y}=+1)$, we can identify $\mathbb{P}(\tilde{Y}=+1|Y=-1), \mathbb{P}(\tilde{Y}=-1|Y=+1)$.

Next we show that to identify $\{e_+, e_-\}$, it is necessary to identify $\{\pi_+, \pi_-\}$. Suppose not: we are unable to identify π_+, π_i but are able to identify $\{e_+, e_-\}$. This implies that there exists another pair $\{\pi'_+, \pi'_-\} \neq \{\pi_+, \pi_-\}$ such that (denote by $\tilde{p} := \mathbb{P}(\tilde{Y}=+1)$)

$$\mathbb{P}(\tilde{Y}=+1|Y=-1) = \frac{\pi_+\tilde{p}}{\pi_+\tilde{p}+(1-\pi_-)(1-\tilde{p})} \tag{9}$$

$$= \frac{\pi'_+\tilde{p}}{\pi'_+\tilde{p}+(1-\pi'_-)(1-\tilde{p})} \tag{10}$$

$$\mathbb{P}(\tilde{Y}=-1|Y=+1) = \frac{\pi_-(1-\tilde{p})}{(1-\pi_+)\tilde{p}+\pi_-(1-\tilde{p})} \tag{11}$$

$$= \frac{\pi'_-(1-\tilde{p})}{(1-\pi'_+)\tilde{p}+\pi'_-(1-\tilde{p})} \tag{12}$$

By dividing π_+, π'_+ in both the numerator and denominator in Eqn. (9) and (10), we conclude that

$$\frac{1-\pi_-}{\pi_+} = \frac{1-\pi'_-}{\pi'_+} \tag{13}$$

While from Eqn. (11) and (12) we conclude

$$\frac{1-\pi_+}{\pi_-} = \frac{1-\pi'_+}{\pi'_-} \tag{14}$$

From Eqn. (13) and (14) we have

$$(1-\pi_-)\pi'_+ = (1-\pi'_-)\pi_+ \tag{15}$$
$$(1-\pi'_+)\pi_- = (1-\pi_+)\pi'_- \tag{16}$$

Taking the difference and re-arrange terms we prove

$$\pi_+ + \pi_- = \pi'_+ + \pi'_-$$

From Eqn. (13) again, taking -1 on both side we have

$$\frac{1-\pi_- - \pi_+}{\pi_+} = \frac{1-\pi'_- - \pi'_+}{\pi'_+} \tag{17}$$

This proves $\pi_+ = \pi'_+$. Similarly we have $\pi_- = \pi'_-$ - but this contradicts the assumption that $\{\pi'_-, \pi'_+\}$ is a different pair. □

Proof for Theorem 5

Proof. We first prove sufficiency. We first relate our problem setting to the setup of Kruskal's identifiability scenario: $Y \in \{1, 2, ..., K\}$ corresponds to the unobserved hidden variable Z. $\mathbb{P}(Y = i)$ corresponds to the prior of this hidden variable. Each $\tilde{Y}_i, i = 1, ..., p$ corresponds to the observation O_i. κ_i is then simply the cardinality of the noisy label space, K. In the context of this theorem, $p = 3$, corresponding to the three noisy labels we have.

Each $\tilde{Y}_i$ corresponds to an observation matrix M_i:

$$M_i[j,k] = \mathbb{P}(O_i = k|Z = j) = \mathbb{P}(\tilde{Y}_i = k|Y = j, X)$$

Therefore, by definition of M_1, M_2, M_3 and $T(X)$, they all equal to $T(X)$: $M_i \equiv T(X), i = 1, 2, 3$. When $T(X)$ has full rank, we know immediately that all rows in M_1, M_2, M_3 are independent. Therefore, the Kruskal ranks satisfy

$$\mathsf{Kr}(M_1) = \mathsf{Kr}(M_2) = \mathsf{Kr}(M_3) = K$$

Checking the condition in Theorem 4, we easily verify

$$\mathsf{Kr}(M_1) + \mathsf{Kr}(M_2) + \mathsf{Kr}(M_3) = 3K \geq 2K + 2$$

Calling Theorem 4 proves the sufficiency.

Now we prove necessity. To prove so, we are allowed to focus on the binary case, where

$$T(X) = \begin{bmatrix} 1 - e_-(X) & e_-(X) \\ e_+(X) & 1 - e_+(X) \end{bmatrix}$$

Note in above, for simplicity we drop e_-, e_+'s dependency in X. We need to prove less than 3 informative labels will not suffice to guarantee identifiability. The idea is to show that the two different set of parameters e_-, e_+ can lead to the same joint distribution $\mathbb{P}(\tilde{Y}_1, \tilde{Y}_2|X)$.

The case with a single label is already proved by Example 1. Now consider two noisy labels $\tilde{Y}_1, \tilde{Y}_2$. We first claim the following three quantities fully capture the information provided by $\tilde{Y}_1, \tilde{Y}_2$:

- Posterior: $\mathbb{P}(\tilde{Y}_1 = +1|X)$
- Positive Consensus: $\mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2 = +1|X)$
- Negative Consensus: $\mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2 = -1|X)$

This is because other statistics in $\tilde{Y}_1, \tilde{Y}_2|X$ can be reproduced using combinations of the three quantities above:

$$\begin{aligned}
&\mathbb{P}(\tilde{Y}_1 = -1|X) = 1 - \mathbb{P}(\tilde{Y}_1 = +1|X)\,, \\
&\mathbb{P}(\tilde{Y}_1 = +1, \tilde{Y}_2 = -1|X) = \mathbb{P}(\tilde{Y}_1 = +1|X) - \mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2 = +1|X)\,, \\
&\mathbb{P}(\tilde{Y}_1 = -1, \tilde{Y}_2 = +1|X) = \mathbb{P}(\tilde{Y}_2 = +1|X) - \mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2 = +1|X)\,.
\end{aligned}$$

But $\mathbb{P}(\tilde{Y}_2 = +1|X) = \mathbb{P}(\tilde{Y}_1 = +1|X)$, since the two noisy labels are identically distributed. The above three quantities led to three equations that depend on e_+, e_-: denote by $\gamma := \mathbb{P}(Y = +1)$

Next we prove the following system of equations:

$$\begin{aligned}\mathbb{P}(\tilde{Y}=+1|X) &= \gamma\cdot(1-e_+)+(1-\gamma)\cdot e_-\\ \mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=+1|X) &= \gamma\cdot(1-e_+)^2+(1-\gamma)\cdot e_-^2\\ \mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=-1|X) &= \gamma\cdot e_+^2+(1-\gamma)\cdot(1-e_-)^2\end{aligned}$$

To see this:

$$\begin{aligned}&\mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=+1|X)\\ =&\mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=+1,Y=+1|X)\\ &+\mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=+1,Y=-1|X)\\ =&\mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=+1|Y=+1,X)\cdot\mathbb{P}(Y=+1|X)\\ &+\mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=+1|Y=-1,X)\cdot\mathbb{P}(Y=-1|X)\\ =&\gamma\cdot(1-e_+)^2+(1-\gamma)\cdot e_-^2\end{aligned}$$

The last equality uses the fact that $\tilde{Y}_1,\tilde{Y}_2$ are conditional independent given Y, so

$$\begin{aligned}&\mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=+1|Y=+1,X)=\\ &\mathbb{P}(\tilde{Y}_1=+1|Y=+1,X)\cdot\mathbb{P}(\tilde{Y}_2=+1|Y=+1,X)\\ &\mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=+1|Y=-1,X)=\\ &\mathbb{P}(\tilde{Y}_1=+1|Y=-1,X)\cdot\mathbb{P}(\tilde{Y}_2=+1|Y=-1,X)\end{aligned}$$

We can similarly derive for $\mathbb{P}(\tilde{Y}_1=\tilde{Y}_2=-1|X)$.

Now we show the above equations do not identify e_+,e_-. For instance, it is straightforward to verify that both of the solutions below satisfy the equations (up to numerical errors, exact solution exists but in complicated forms):

- $\gamma=0.7,\ e_+=0.2,\ e_-=0.2$
- $\gamma=0.8,\ e_+=0.242,\ e_-=0.07$

The above example proves that two informative noisy labels are insufficient to guarantee identifiability.

□

Proof for Theorem 6

Proof. In the unstructured model, we first show that, with a large N, with high probability, each X's will present at least 3 times. Denote by N_X the number of times X appears in the dataset. Then

$$N_X:=\sum_{i=1}^{N}1[X_i=X],\ \mathbb{E}[N_X]=\frac{q_X}{\sum_{X\in\mathcal{X}}q_X}N \tag{18}$$

When N is large enough such that $N>\frac{4\sum_{X\in\mathcal{X}}q_X}{\min_X q_X}$, we have $\mathbb{E}[N_X]>4$. Then using Hoeffding inequality we have

$$\mathbb{P}(N_X\le 3)\le exp(-2N).$$

Using union bound (across N samples), it implies that with probability at least $1-Nexp(-2N)$, $N_X\ge 3,\forall X$:

$$\mathbb{P}(N_X>3,\forall X)=1-\mathbb{P}(N_X\le 3,\exists X)\le 1-Nexp(-2N) \tag{19}$$

This further implies that with probability at least $1-Nexp(-2N)$, we have $X_1=X_2=X$ for each X: Their distance is 0, clearly falling below the closeness threshold ϵ. Therefore they will share the same true label. □

Proof for Theorem 7

Proof. The d^* features and the noisy label $\tilde{Y}$ jointly give us d^*+1 independent observations. Checking Kruskal's condition we have:

$$\mathsf{Kr}(T(X))+\sum_{i=1}^{d^*}\mathsf{Kr}(M_i)\ge K+2\cdot d^*\ge K+K+d^*=2K+d^*+1-1$$

Calling Theorem 4, we establish the identifiability. □

Proof for Theorem 8

Proof. By definition

$$||T_1(X) - T^*(X)||_F = \sqrt{\sum_i \sum_j (T_1[i,j] - T[i,j])^2} \tag{20}$$

Easy to show that

$$\begin{aligned}
&||T_1(X) - T^*(X)||_F + ||T_2(X) - T^*(X)||_F \\
=&\sqrt{\sum_i \sum_j (T_1[i,j] - T[i,j])^2} + \sqrt{\sum_i \sum_j (T_2[i,j] - T[i,j])^2} \\
=&\sqrt{\left(\sqrt{\sum_i \sum_j (T_1[i,j] - T[i,j])^2} + \sqrt{\sum_i \sum_j (T_2[i,j] - T[i,j])^2}\right)^2} \\
\geq&\sqrt{\sum_i \sum_j \left((T_1[i,j] - T[i,j])^2 + (T_2[i,j] - T[i,j])^2\right)} && \text{(Dropping the cross-product term which is positive)} \\
\geq&\sqrt{\sum_i \sum_j \left(T_1[i,j] - \frac{T_1[i,j] + T_2[i,j]}{2}\right)^2 + \left(T_2[i,j] - \frac{T_1[i,j] + T_2[i,j]}{2}\right)^2} && \text{(minimum distance is at half)} \\
=&\sqrt{\sum_i \sum_j 2\left(\frac{T_1[i,j] - T_2[i,j]}{2}\right)^2} \\
=&\frac{1}{\sqrt{2}}\sqrt{\sum_i \sum_j (T_1[i,j] - T_2[i,j])^2} \\
=&\frac{1}{\sqrt{2}}||T_1(X) - T_2(X)||_F
\end{aligned}$$

□

Proof for Theorem 9

Proof. The proof is straightforward by checking KrusKal's identifiability condition:

$$\mathsf{Kr}(T(X)) + \sum_{i=1}^{d^*} \mathsf{Kr}(M_i) \geq 1 + 2 \cdot d^* \geq 1 + 2|G|K - 1 + d^* = 2|G| \cdot K + d^* + 1 - 1$$

□

B Generic identifiability

We provide a bit more detail for the discussion on generic identifiability left in Section 5.2.

Theorem 10. *With a single informative noisy label, $T(X)$ is generically identifiable for each group $g \in G$ if the number of disentangled features d^* satisfies that $d^* \geq \lceil \log_2 \frac{K+2}{2} \rceil$, and $\tau_i \geq 2$.*

Proof. We first reproduce a relevant theorem in [35]:

Theorem 11. *[35] When $p = 3$ (3 independently observations), the model parameters are generically identifiable, up to label permutation, if*

$$\min(K, \kappa_1) + \min(K, \kappa_2) + \min(K, \kappa_3) \geq 2K + 2 \tag{21}$$

Based on the above theorem we have the following identifiability result:

Grouping d^* features evenly into two groups, each corresponding to a meta variable/feature:

$$R_1^* = \prod_{i=1}^{d_1^*} R_i,\ X_2^* = \prod_{j=d_1^*+1}^{d^*} R_j$$

Table 2: Comparison of test accuracy on CIFAR10 by using the estimated transition matrix.

Methods	*inst. 0.3*	*inst. 0.4*	*inst. 0.5*	*inst. 0.6*
FW (SimCLR)	66.61	65.82	64.51	62.81
FW (IPIRM)	73.24	72.54	71.33	69.42

Denote feature dimensions of each group as d_1^*, d_2^*:

$$\tau_1^* = \prod_{i=1}^{d_1^*} \geq 2^{d_1^*} \geq 2^{\lceil \log_2 \frac{K+2}{2} \rceil} \geq \frac{K+2}{2} \tag{22}$$

Similarly $\tau_2^* \geq \frac{K+2}{2}$. Denote by M_1^*, M_2^* the two observation matrices for the grouped variables

$$M_i^*[j,k] = \mathbb{P}(R_i^* = \mathcal{R}_i^*[k] | Y = j),\ i = 1, 2.$$

Then:

$$\mathsf{Kr}(T(X)) + \mathsf{Kr}(M_1^*) + \mathsf{Kr}(M_2^*) \geq K + 2\frac{K+2}{2} = 2K + 2,$$

which again satisfied the identifiability condition specified in Theorem 4. □

C More experiments

In this section, we elaborate the detailed experiment setting and perform more experiments *w.r.t.* disentangled features.

C.1 Experiment setting for Table 1

Label Noise Generation The label noise of each instance is characterized by $T_{ij}(X) = \mathbb{P}(\widetilde{Y} = j | X, Y = i)$. In this paper, we consider two types of label noise: asymmetric label noise [48, 58] and instance-dependent label noise [19, 28]. For asymmetric label noise, $T(X) \equiv T$, each clean label is randomly flipped to its adjacent label w.p. ϵ, where ϵ is the noise rate, i.e., $T_{ii} = 1 - \epsilon$, $T_{ii} + T_{i,(i+1)_K} = 1$, $(i+1)_K := i \mod K + 1$. For instance-dependent label noise, the generation of noisy labels also depends on the features. We follow CORES [19] to generate instance-dependent label noise. The generation process is detailed in Algorithm 1. With these definitions, *asymm./inst.* ϵ in Table 1 denotes asymmetric/instance-dependent label noise with noise rate ϵ.

Model pre-training. The network structures of all the three encoders in the Table 1 are ResNet50 [59]. Note that the encoders which generate features to estimate transition matrix can be pre-trained on different datasets. For example, HOC [8] utilizes ImageNet pre-trained encoders to generate features for CIFAR. Thus, following the pipeline of disentangled features generation [57], we pre-train all the three encoders on CIFAR100 dataset and generate feature for CIFAR10 to estimate transition matrix. The first encoder is trained under 0.1 symmetric label noise rate to simulate the weakly-supervised features while the second and third encoder is trained via self-supervised learning (SSL). Recall the goal of SSL is to learn a good representation without accessing labels. In this paper, we adopt SimCLR [56] and IPIRM [57] to perform SSL pre-training. SimCLR, as a representative work on SSL literature, learns a good represention based on InfoNCE loss [60]. However, it is shown that the features learned by SimCLR are only *partly* disentangled on some simple augmentation features such as rotation and colorization [57]. Thus IPIRM proposes a learning algorithm that embeds InfoNCE loss into IRM (Invariant Risk Minimization) framework [61] to learn *fully* disentangled features. We train SimCLR model and IPIRM model by referring official codebase of IPIRM [3]. The pre-trained models, as well as evaluation code are all released in the supplementary material. **Estimation error of Transition matrix**. After training these three encoders, we fix the encoder and generate features from raw samples to estimate the noise transition matrix using Global HOC estimator [8]. The hyper-parameters for estimating transition matrix are consistent with official implementation of HOC [4]: optimizer: Adam, learning rate: 0.1, number of iterations: 1500. After training, we evaluate the performance via absolute estimation error defined below:

$$\text{err} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} |\hat{T}_{i,j} - T_{i,j}|}{K^2} * 100,$$

where $\hat{T}$ is the estimated noise transition matrix, T is the real noise-transition matrix, K is the number of classes in the dataset.

C.2 Training performance using estimated transition matrix

We can further use the estimated transition matrix to perform forward loss correction (FW) [23]. Table 2 records the performance of FW by using the estimated transition matrix of SimCLR and IPIRM. The hyper-parameters for all the experiments in Table 2 are the same: optimizer: SGD, training epochs: 100, learning rate: 0.1 for first 50 epochs and 0.01 for last 50 epochs, batch-size: 256. From the results, we can observe that the test accuracy increases as features become more disentangled.

[3] https://github.com/Wangt-CN/IP-IRM

[4] https://github.com/UCSC-REAL/HOC

Algorithm 1 Instance-Dependent Label Noise Generation

Input:

1: Clean examples $(x_n, y_n)_{n=1}^N$; Noise rate: ε; Size of feature: $1 \times S$; Number of classes: K.

Iteration:

2: Sample instance flip rates q_n from the truncated normal distribution $\mathcal{N}(\varepsilon, 0.1^2, [0, 1])$;

3: Sample $W \in \mathcal{R}^{S \times K}$ from the standard normal distribution $\mathcal{N}(0, 1^2)$;

for $n = 1$ to N **do**

4: $p = x_n \cdot W$ // Generate instance dependent flip rates. The size of p is $1 \times K$.

5: $p_{y_n} = -\infty$ // Only consider entries different from the true label

6: $p = q_n \cdot \text{softmax}(p)$ // Let q_n be the probability of getting a wrong label

7: $p_{y_n} = 1 - q_n$ // Keep clean w.p. $1 - q_n$

8: Randomly choose a label from the label space as noisy label $\tilde{y}_n$ according to p;

end for

Output:

9: Noisy examples $(x_i, \tilde{y}_n)_{n=1}^N$.

Table 3: Comparison of test accuracy on CIFAR100 by using different DNN initialization.

Methods	*inst. 0.3*	*inst. 0.4*	*inst. 0.5*	*inst. 0.6*
CE (random init)	43.47	35.17	27.07	18.25
CE (SimCLR init)	58.95	49.7	36.87	25.07
CE (IPIRM init)	64.92	56.18	43.75	30.36

C.3 Initializing DNN using disentangled features

Except for estimating transition matrix, we can directly use disentangled features to perform training on noisy dataset. Table 3 shows the effect of using disentangled features as DNN initialization on CIFAR100. The hyper-parameters for all the experiments in Table 3 are consistent with Table 2. From the results, We can observe that even with vanilla Cross Entropy loss, the disentangled features are still beneficial to the performance.