# Identifiability of Label Noise Transition Matrix

**Yang Liu** [1 2] **Hao Cheng** [1] **Kun Zhang** [3 4]

## 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 the knowledge of 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 recently proposed solutions for learning from instance-dependent noisy labels, the literature 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 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 reveals that disentangled features improve identification. This discovery led us to an approach that improves the estimation of the transition matrix using properly disentangled features. Code is available at https://github.com/UCSC-REAL/Identifiability.

## 1. Introduction

The literature of learning with noisy labels concerns the scenario when the observed training 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 both defining and solving this problem. Among many other benefits, the knowledge of $T(X)$ has demonstrated its use in performing risk correction (Natarajan et al., 2013; Patrini et al., 2017), label correction (Patrini et al., 2017), and constraint corrections (Wang et al., 2021a; Wei et al., 2023). In beyond, it also finds applications in ranking small loss samples (Han et al., 2020) and detecting corrupted samples (Zhu et al., 2021a). 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 decline of model performance (Natarajan et al., 2013; Liu & Wang; Xia et al., 2019; Zhu et al., 2021c), and false sense of fairness (Wang et al., 2021a; Liu & Wang; Zhu et al., 2022b). Prior works have also documented challenges in estimating the noise transition matrices when the quality of available training information remains unclear. For instance, in (Zhu et al., 2022a) the authors show that when the quality of representations dropped, the estimation error in $T(X)$ increases significantly (Figure 1 therein). Other references have observed these challenges too (Xia et al., 2019).[1]

Knowing whether a $T(X)$ is identifiable or not is crucial and informs us if $T(X)$ and the underlying noisy learning problem are indeed learnable. The earlier results have focused on class-dependent transition matrix $T(X) \equiv T := [\mathbb{P}(\tilde{Y} = j|Y = i)]_{i,j}, \forall X$, that is different $X$s observe the same transition matrix. The literature has provided discussions of the identifiability of this class-dependent $T$ (Scott, 2015), and has identified a reducibility condition for inferring the inverse noise rate, which closely relates to $T$. Later works have developed a sequence of solutions to estimate $T$ under a variety of assumptions, including irreducibility (Scott, 2015), anchor points (**?**Xia et al., 2019; Yao et al., 2020a), separability (Cheng et al., 2020), rankability (Northcutt et al., 2017; 2021), redundant labels/tensor (Liu et al., 2020; Traganitis et al., 2018; Zhang et al., 2014), clusterability (Zhu et al., 2021c), among others (Zhang et al., 2021; Li et al., 2021).

Nonetheless, recent study (Wei et al., 2021) has shown that the above class-dependent model fails to capture the real-world noise patterns, but rather real human-level noise follows an instance-dependent model. Intuitively, the in-

[1]University of California, Santa Cruz [2]ByteDance Research [3]Carnegie Mellon University [4]Mohamed bin Zayed University of Artificial Intelligence. Correspondence to: Yang liu <yangliu@ucsc.edu>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

[1]We provide experiments to validate this in Appendix C.4.

stance $X$ encodes the difficulties in generating the label for it. At the same time, we observe a recent surge of different solutions towards solving the instance-dependent label noise problem (Cheng et al., 2020; Xia et al., 2020b; Cheng et al., 2021a; Yao et al., 2021). Some of the results took on the problem of estimating $T(X)$, while the others proposed solutions to intervene directly on the 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 potentially complicated dependency between $X$ and $T(X)$ renders it even less clear whether solving this problem is feasible or not. This observation 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 provide evidences to the success of existing solutions 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 (Kruskal, 1976; 1977). 2. Three separate independent and identically distributed (i.i.d.) noisy labels (random variables) are both necessary and sufficient for instance-level identifiability. This observation echoes the previous successes of developing tensor-based approaches for identifying the hidden models. 3. Disentangled features help improve identifiability and learnability.

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 provides evidence of the success of existing solutions. Section 6 provides some empirical observations.

### 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 (Patrini et al., 2017), loss reweighting (Liu & Tao, 2015), label correction (Xiao et al., 2015) and unbiased loss (Natarajan et al., 2013). 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 (Xia et al., 2020b), using confidence scores to revise transition matrices (Berthon et al., 2020), and using clusterability of the data (Zhu et al., 2021c). More recent works have used the causal knowledge to improve the estimation (Yao et al., 2021), and the deep neural network to estimate the transition matrix defined between the noisy label and the Bayes optimal label (Yang et al., 2021). Other works chose to focus on the learning from instance-dependent label noise directly, without explicitly estimating the transition matrix (Zhu et al., 2021b; Cheng et al., 2021a; Berthon et al., 2020; Xia et al., 2020a; Li et al., 2020).

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 (Scott, 2015; Yao et al., 2020b; Menon et al., 2015). 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 (Kruskal, 1976; 1977; Sidiropoulos & Bro, 2000; Allman et al., 2009), which reveals a sufficient condition for identifying a parametric model that links a hidden variable to a set of observed ones. Kruskal's early results were developed under the context of tensor, which later proves to be a powerful tool for learning latent variable models (Sidiropoulos et al., 2017; Zhang et al., 2014; Anandkumar et al., 2014).

## 2. Formulation

We 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:

$$T(X) := [\mathbb{P}(\tilde{Y} = j|Y = i, X)]_{i,j=1}^{K} .$$

$T(X)$ 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 $\Omega$ an observation space. We first define identifiability for a general parametric space $\Theta$. Denote the distribution induced by the parameter $\theta \in \Theta$ of a statistical model on the observation space $\Omega$ as

$\mathbb{P}_\theta$ (Kruskal, 1976; Allman et al., 2009). To give an example, for a fixed $X$ (when consider instance-level identifiability), and $\Omega$ is simply the outcome space for its associated noisy label $\tilde{Y}$, i.e., $\{1, 2, ..., K\}$. In this case, each $\theta$ 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 2.1** (Identifiability)**.** The parameter $\theta$ (statistical model) is identifiable if $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}, \forall \theta \neq \theta'$.

We define identifiability for the task of learning with noisy labels for an $X$. Denote by

$$\theta(X) := \{T(X), \mathbb{P}(Y|X)\} .$$

$\mathbb{P}_{\theta(X)}$ is the distribution (probability density function) over $\Omega$, defined by the noise transition matrix $T(X)$ and the prior $\mathbb{P}(Y|X)$. To emphasize, $\Omega$ is not necessarily the observation space of the noisy label $\tilde{Y}$ only. The exploration of an effective $\Omega$ will be one of the focuses.

**Definition 2.2** (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. Allowing for label permutation would mean that our results allow the high noise rate regime. For instance, for a binary classification problem, an $80\%$ noise rate would correspond to a counterfactual $20\%$ one. Finding either model would be regarded as being identifiable. In practice, further restriction such as noise rate should not exceed $50\%$ can help us remove one of the two cases.

## 3. Preliminary

In this section, we will introduce two highly relevant results on Mixture Proportion Estimation (MPE) (Scott, 2015) and Kruskal's identifiability result (Kruskal, 1976; 1977).

### 3.1. Preliminaries using irreducibility and anchor points

The problem of learning from noisy labels ties closely to another problem called Mixture Proportion Estimation (MPE) (Scott, 2015), 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 $\kappa^*$ from only observing $F$ and $H$. The following identifiability result has been established:

**Proposition 3.1.** *(Blanchard et al., 2010) $\kappa^*$ 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 (Yao et al., 2020b), a stronger requirement was established:

**Proposition 3.2.** *(Yao et al., 2020b) $\kappa^*$ 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 transition matrix $T$ (Xia et al., 2019; Li et al., 2021).

Prior work has established the connection between the MPE problem and the learning from noisy label one (Yao et al., 2020b) 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., $T(X) \equiv T$ ($\mathbb{P}(\tilde{Y}|Y, X) \equiv \mathbb{P}(\tilde{Y}|Y)$), 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 $\kappa^*$ is equivalent to the identifiability of $T$. We start with the first thing above. We want to acknowledge that this equivalence appeared before in (Yao et al., 2020b; Menon et al., 2015). We reproduce it here to make our paper self-contained. Denote by $\pi_+ := \mathbb{P}(Y = -1|\tilde{Y} = +1), \pi_- := \mathbb{P}(Y = +1|\tilde{Y} = -1)$ and $\tilde{\pi}_- = \frac{\pi_-}{1-\pi_+}, \tilde{\pi}_+ = \frac{\pi_+}{1-\pi_-}$.

**Lemma 3.3.** $\mathbb{P}(X|\tilde{Y} = -1), \mathbb{P}(X|\tilde{Y} = +1)$ *relate to* $\mathbb{P}(X|Y = -1), \mathbb{P}(X|Y = +1)$ *as follows:*

$$\begin{aligned}\mathbb{P}(X|\tilde{Y} = -1) &= \tilde{\pi}_- \cdot \mathbb{P}(X|\tilde{Y} = +1) \\ &\quad + (1 - \tilde{\pi}_-) \cdot \mathbb{P}(X|Y = -1) \\ \mathbb{P}(X|\tilde{Y} = +1) &= \tilde{\pi}_+ \cdot \mathbb{P}(X|\tilde{Y} = -1) \\ &\quad + (1 - \tilde{\pi}_+) \cdot \mathbb{P}(X|Y = +1) .\end{aligned}$$

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 $J$s, $\tilde{\pi}_-, \tilde{\pi}_+$ correspond to the mixture proportion $\kappa^*$. 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 $\kappa^*$ 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 then have:

**Theorem 3.4.** *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 (Liu & Tao, 2015; Xia et al., 2019; Li et al., 2021). 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 (Kruskal, 1976; 1977). The setup is as follows: suppose that there is an unobserved variable $Z$ that takes values in a $K$-sized discrete domain $\{1, 2, ..., r\}$. $Z$ 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 $\kappa_i$. Let $M_i$ be a matrix of size $r \times \kappa_i$, 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 $\theta$. We now introduce the Kruskal rank of a matrix, which plays a central role in Kruskal's identifiability results.

**Definition 3.5** (Kruskal rank)**.** (Kruskal, 1976; 1977) The Kruskal rank of a matrix $M$ is the largest number $I$ such that every set of $I$ rows [2] 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 3.6.** *(Kruskal, 1976; 1977; Sidiropoulos & Bro, 2000) The parameters $M_i, i = 1, ..., p$ are identifiable, up to label permutation, if*

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

The result for $p = 3$ was first established in (Kruskal, 1977) demonstrating the power of a three-way tensor, and then it was shown in (Sidiropoulos & Bro, 2000) that the proof extends to a general $p$. The proof builds on showing that different parameter $\theta$ leads to different stacking of $M$s: $[M_1, ..., M_p]$. When $p = 3$, $[M_1, M_2, M_3] := \sum_{k=1}^{K} \mathbf{m}_1^k \bigotimes \mathbf{m}_2^k \bigotimes \mathbf{m}_3^k$ forms the tensor of the observations, where $\mathbf{m}_i^k, i = 1, 2, 3$ is the $k$-th column of $M_i$.

## 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

At a first sight, it is impossible to identify $\mathbb{P}(\tilde{Y}|Y, X)$ from only observing $\mathbb{P}(\tilde{Y}|X)$,[3] 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 (probability) 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 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.

### 4.2. The necessity of multiple noisy labels

**Setups** 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. For each instance $X$, denote by $K_X \leq K$ the number of non-degenerated label classes

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

[3] 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 $X$.

$k$ such that $\mathbb{P}(Y = k|X) > 0$. W.l.o.g., let us assume the non-degenerate classes are simply $\{1, 2..., K_X\}$.

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 (Clogg, 1995) and tensor decomposition (Anandkumar et al., 2014). 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 $T(X)$. We now define an *informative noisy label*.

**Definition 4.1.** For a given $(X, Y)$, we call their noisy label $\tilde{Y}$ informative if $\text{rank}(T(X)) = K_X$.

Definition 4.1 requires the $K_X$ rows of $T(X)$ are linearly independent. When the observation space for $\tilde{Y}$ is the same as $Y$ (therefore $T(X)$ is a squared matrix), i.e., the true label $Y$ has a full support on the entire label space, then the requirement is stating that $T(X)$ is of full rank, which is already assumed in the literature - e.g., loss correction (Natarajan et al., 2013; Patrini et al., 2017; Traganitis et al., 2018) 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 (Natarajan et al., 2013), which can be easily shown to imply $T(X)$ is full rank.

But we do not remove the possibility that $T(X)$ is not a squared matrix and $K_X$ can be much smaller than the entire label space. The name for informativeness is inspired by the following observation: when the noisy label brings in useful information such that $\mathbb{P}(Y = i|\tilde{Y} = i) \neq \mathbb{P}(Y = i|\tilde{Y} = i)$, the rows in $T(X)$ have to be different. For binary classification

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

when $\mathbb{P}(Y = 1|\tilde{Y} = 1) \neq \mathbb{P}(Y = 1)$, we have $\frac{\mathbb{P}(\tilde{Y}=1|Y=1|)\cdot\mathbb{P}(Y=1)}{\mathbb{P}(\tilde{Y}=1)} \neq \mathbb{P}(Y = 1)$, which is equivalent to $\mathbb{P}(\tilde{Y} = 1|Y = 1|) \neq \mathbb{P}(\tilde{Y} = 1)$, i.e., $1 - e_1 \neq \mathbb{P}(Y = 1) \cdot (1 - e_1) + \mathbb{P}(Y = 0) \cdot e_0$, that is $1 - e_1 \neq e_0$.

Our first identifiability result states as follows:

**Theorem 4.2.** *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 $T(X)$ when $K_X \geq 2$.*

Note that $K_X \geq 2$ is easily satisfied as long as there exists uncertainty in $\mathbb{P}(Y|X)$ - As long as there is uncertainty in $\mathbb{P}(Y|X)$, we have at least two non-degenerated label class $k$ with $\mathbb{P}(Y = k|X)$ and therefore $K_X \geq 2$. Further by Definition 4.1, we establish that $\text{rank}(T(X)) \geq K_X \geq 2$.

*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: $Y \in \{1, 2, ..., K_X\}$. 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 $O_i$. $\kappa_i$ 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 a rank $K_X$, we know immediately that all rows in $M_1, M_2, M_3$ are independent. Therefore, the Kruskal ranks satisfy $\mathsf{Kr}(M_1) = \mathsf{Kr}(M_2) = \mathsf{Kr}(M_3) = K_X$. Checking the condition in Theorem 3.6, we easily verify $\mathsf{Kr}(M_1) + \mathsf{Kr}(M_2) + \mathsf{Kr}(M_3) = 3K_X \geq 2K_X + 2$ . Calling Theorem 3.6 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 (Liu et al., 2020; Zhu et al., 2021c; Zhang et al., 2014). 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. This result also echoes the power of tensor (stacking third order information) in uncovering hidden models (Traganitis et al., 2018; Zhang et al., 2014). Particularly relevant to us is (Zhang et al., 2014) where it was shown a spectral EM approach that uses three noisy labels suffices to identify the noise transition matrix of labels. We want to highlight that our proof and results establish both the necessity and sufficiency for having three informative noisy labels, independent from the specific algorithms developed. Another note we want to add is that our main inquiry is on establishing the conditions for identifying $T(X)$, instead of proposing algorithms to estimate $T(X)$.

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 (Liu et al., 2012; Karger et al., 2011; Liu & Liu, 2015), which serves as a preprocessing 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.

## 5. Instantiations and Practical Implications of Our Identifiability Results

Most of the learning with noisy label solutions focus 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 $X$s 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 $X$s, 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. We carry the thoughts that each $X$'s noisy label $\tilde{Y}$ is informative as defined in Definition 4.1.

### 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 $X$s (Zhu et al., 2021c; Bahri et al., 2020) to infer the noise transition matrix:

**Definition 5.1.** The 2-NN clusterability requires each $X$ and its two nearest neighbors $X_1, X_2$ share the same true label $Y$, that is $Y = Y_1 = Y_2$, and $T(X) = T(X_1) = T(X_2)$.

This definition helps us remove the requirement for multiple noisy labels per each $X$: one can view it as for each $X$, borrowing the noisy labels from its 2-NN, we have three independent noisy labels $\tilde{Y}, \tilde{Y}_1, \tilde{Y}_2$, all from the same $Y$ (Figure 1). 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 $X, X_1, X_2$. 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),

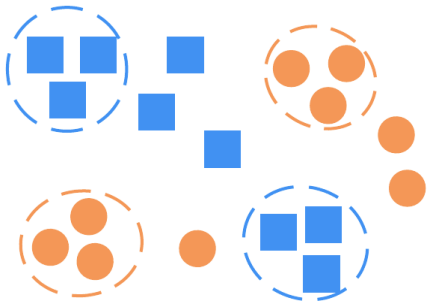

*Figure 1.* Data generation: label correlation among triplets. Orange circles and blue squares indicate different labels.

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, the 2-NN requirement seems to hold empirically: according to (Zhu et al., 2021c) (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 weaker feature extractor, the ratio is mostly always in or close to the 80% range.

The following data generation process for an unstructured discrete domain of classification problem (Feldman, 2020; Liu, 2021) justifies the 2-NN requirement. The intuition is that when $X$s are informative and sufficiently discriminative, the similar $X$s 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 $\lambda$.
- 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 $N$ $X$s 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 $\epsilon$ (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 $X$s that are really close. We prove the above data generation process satisfies the 2-NN clusterability requirement with high probability.

**Theorem 5.2.** *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 clusterability.*

**Smoothness conditions in semi-supervised learning** This above discussion also ties closely to the smoothness requirements in semi-supervised learning (Zhu et al., 2003; Zhu, 2005), where the neighborhood $X$s can provide and propagate label information in each local neighborhood of $X$s. Indeed, this idea echoes the co-teaching solution (Jiang et al., 2018; Han et al., 2018) 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 (Jiang et al., 2018; Han et al., 2018), counting the noisy label itself, and the "teacher" supervision, there are two such noisy supervision labels[4]. This observation raises an interesting question: does adding an additional teacher network for an additional supervision help? This question merits empirical verification.

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

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 $X$s 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:

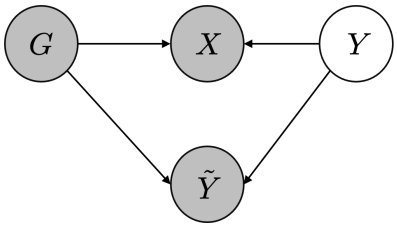


*Figure 2.* Graph for $(G, X, Y, \tilde{Y})$. Grey color indicates observable variables.

**Class-dependent** $T$ $\mathbb{P}(\tilde{Y}|Y, X) \equiv \mathbb{P}(\tilde{Y}|Y)$, a single group.

**Noise clusterability** The noise transition estimator proposed in (Zhu et al., 2021c) was primarily developed for class-dependent but not instance-dependent $T(X)$. Nonetheless, a noise clusterability definition is introduced therein to allow the approach to be applied to instance-dependent noise. Under noise clusterability, using 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 (Wang et al., 2021a; Liu & Wang; Wang et al., 2021b). For instance, (Wang et al., 2021a; Liu & Wang) 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 $X$s 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. 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 $|\mathcal{R}_i| = \kappa_i$. 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 5.3** (Disentangled $R$)**.** $R$ is disentangled if $\{R_i\}_{i=1}^{d^*}$ are conditional independent given $Y$.

Note our definition of disentangled representation does have differences from conventional disentanglement. For example, a well accepted definition of disentangled representation is based on group theory which operates on $X$ (Higgins et al., 2018). A recent work also designs an algorithm to learn self-supervised disentangled representation based on this definition (Wang et al., 2021c). Our work is to study the identifiability of the label noise transition matrix. Thus our definition of disentangled representation is related to $Y$

**Definition 5.4** (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 5.5.** *For $X$s 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 (Higgins et al., 2018; Steenbrugge et al., 2018) 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 (Allman et al., 2009), the above theorem can be further extended to requiring $d^* \geq \lceil \log_2 \frac{2K_G^*+1}{2} \rceil$, where $K_G^* = \max_{X \in G} K_X$. Details are deferred to Appendix (Theorem B.1). Note that the existence of disentangled $X$

[4]Co-teaching is a sample selection approach. Part of the solution in it requires supervision datasets constructed according to another network's predictions. In a certain sense, the training process benefits from artificially generated noisy labels.

does not imply that we will be able to directly infer $\mathbb{P}(Y|X)$ which will help us complete the learning task directly. But rather, it is indeed possible to further identify the structure $\mathbb{P}(X|Y)$ (from unobserved to observed) but this is an identifiability problem defined on a much higher space.

When disentangled features are 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 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 successes in applying semi-supervised (Cheng et al., 2021a; Li et al., 2020; Nguyen et al., 2019) and self-supervised learning (Cheng et al., 2021b; Zheltonozhskii et al., 2022; Ghosh & Lan, 2021) to handle 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 $T(X)$.

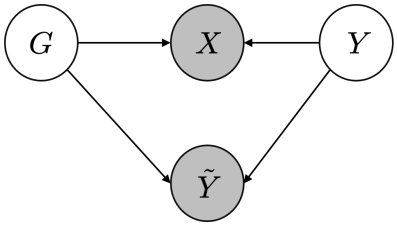


*Figure 3.* Graph with unobserved $G$. Grey color indicates observable variables.

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. Then:

**Theorem 5.6.** *Any estimator $T^*(X)$ will incur at least the following estimation error:*

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

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 $\{1, 2, ..., |G|K\}$[5]. 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 5.7.** *For $X$s 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, the requirement of the number of informative and disentangled features grows linearly in $|G|$. We now relate to the literature that implicitly groups $X$s. We will use $\mathcal{X}$ to denote the space of all possible $X$s.

**Part-dependent label noise** (Xia et al., 2020b) 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 subparts, 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 5.7 would be more likely to be satisfied.

**DNN approach** (Yang et al., 2021) 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 5.7, 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** (Yao et al., 2021) proposed improving the identifiability by exploring the causal structure. With causal inference, 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 $\tilde{\mathcal{X}} := \{\tilde{X} : \tilde{X} \to X \in \mathcal{X}\}$, and $|G| = |\tilde{\mathcal{X}}|$. Therefore in order to meet the identifiability condition in Theorem 5.7, the hidden variable $\tilde{\mathcal{X}}$ has to be sufficiently parameterized to induce a smaller $|G|$.

## 6. Empirical Evidence: Disentangled Features

Most of our results above verified the empirical success of existing approaches from the identifiability's perspective and we refer the interested reader to the detailed experi-

[5]By mapping $(G = 1, Y = 1) \to 1, (G = 1, Y = 2) \to 2, ..., (G = |G|, Y = K) \to |G|K$.

*Table 1.* Estimation error for different features on CIFAR-10. Each experiment is run 3 times and mean ± std is reported. *asymm.*: asymmetric label noise; *inst.*: instance-dependent label noise. (numbers) are noise rates. All the encoders are ResNet50 backbone.

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

ments in the corresponding references. We now empirically show the possibility of learning disentangled features to help identify the noise transition matrix. In this paper, we consider two types of label noise: asymmetric label noise (Han et al., 2018; Wei et al., 2020) and instance-dependent label noise (Cheng et al., 2021a; Zhu et al., 2021b). The label noise of each instance is characterized by $T_{ij}(X) = \mathbb{P}(\widetilde{Y} = j|X, Y = i)$. For asymmetric label noise, $T(X) \equiv T$, each clean label is randomly flipped to its adjacent label w.p. $\epsilon$, where $\epsilon$ 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 (Cheng et al., 2021a) to generate instance-dependent label noise. The generation process is detailed in Algorithm 2 in Appendix C.1. With these definitions, *asymm./inst.* $\epsilon$ in Table 1 denotes asymmetric/instance-dependent label noise with noise rate $\epsilon$.

**Experiment details** We consider three types of encoders that are used to generate features. The first encoder is pre-trained by cross-entropy (CE) loss under 0.1 symmetric label noise rate to simulate the weakly-supervised features, which is generally adopted in FW (Patrini et al., 2017) and HOC (Zhu et al., 2021c). 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 (Chen et al., 2020), a representative work on SSL literature which learns a good represention based on InfoNCE loss (Van den Oord et al., 2018). However, it is shown that the features learned by SimCLR are only *partly* disentangled on some simple augmentation features such as rotation and colorization (Wang et al., 2021c). The third encoder is trained by IPIRM (Wang et al., 2021c), which embeds InfoNCE loss into IRM (Invariant Risk Minimization) framework (Arjovsky et al., 2019) to learn *fully* disentangled features. We train SimCLR model and IPIRM model by referring official codebase of IPIRM [6]. After training these three encoders, we fix the encoder and generate features from raw samples to estimate the noise transition matrix using HOC estimator (Zhu et al., 2021c). The hyper-parameters for estimating transition matrix are consistent with official implementation of HOC [7]: optimizer: Adam, learning rate: 0.1, number of iterations: 1500. We defer the key steps of HOC to Algorithm 1 in Appendix C.1. Note all the three encoders are trained on CIFAR100 datasets and generate feature for CIFAR10 to estimate noise transition matrix.

**Evaluation** We evaluate the performance via absolute estimation error defined below: $\text{err} = \sum_{i=1}^{K}\sum_{j=1}^{K}|\hat{T}_{i,j} - T_{i,j}|/K^2 \cdot 100$, 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 observe that the estimation error decreases as features become more disentangled which supports our analyses. We defer the details, more experiments, as well as experiments on comparing training performances using disentangled features, to the Appendix.

## 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. We show the importance of disentangled features for identifying the noise transition matrix.

Our work has limitations. At multiple places of the work, we state such. For instance, we discussed the situation when we will have multiple noisy labels and our focus on discretized features. In Section 5.2, we clearly stated our requirement of the disentangled and informative features.

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.

**Acknowledgement** KZ was supported in part by the NSF-Convergence Accelerator Track-D award #2134901, by the National Institutes of Health (NIH) under Contract R01HL159805, by grants from Apple Inc., KDDI Research, Quris AI, and IBT, and by generous gifts from Amazon, Microsoft Research, and Salesforce.

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

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

## References


Allman, E. S., Matias, C., and Rhodes, J. A. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bahri, D., Jiang, H., and Gupta, M. Deep k-nn for noisy labels. In *International Conference on Machine Learning*, pp. 540–550. PMLR, 2020.

Berthon, A., Han, B., Niu, G., Liu, T., and Sugiyama, M. Confidence scores make instance-dependent label-noise learning possible. *arXiv preprint arXiv:2001.03772*, 2020.

Blanchard, G., Lee, G., and Scott, C. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., and Liu, Y. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2021a.

Cheng, H., Zhu, Z., Sun, X., and Liu, Y. Demystifying how self-supervised features improve training from noisy labels. *arXiv preprint arXiv:2110.09022*, 2021b.

Cheng, J., Liu, T., Ramamohanarao, K., and Tao, D. Learning with bounded instance-and label-dependent label noise. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, 2020.

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

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

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

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pp. 8527–8537, 2018.

Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I., and Sugiyama, M. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pp. 4006–4016. PMLR, 2020.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313. PMLR, 2018.

Karger, D. R., Oh, S., and Shah, D. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pp. 1953–1961, 2011.

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

Kruskal, J. B. 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.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgExaVtwr.

Li, X., Liu, T., Han, B., Niu, G., and Sugiyama, M. Provably end-to-end label-noise learning without anchor points. *arXiv preprint arXiv:2102.02400*, 2021.

Liu, Q., Peng, J., and Ihler, A. T. Variational inference for crowdsourcing. *Advances in neural information processing systems*, 25:692–700, 2012.

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

Liu, Y. Understanding instance-level label noise: Disparate impacts and treatments, 2021.

Liu, Y. and Liu, M. 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, pp. 217–230, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3486-0. doi: 10.1145/2745844.2745874. URL http://doi.acm.org/10.1145/2745844.2745874.

Liu, Y. and Wang, J. Can less be more? when increasing-to-balancing label noise rates considered beneficial. NeurIPS'21.

Liu, Y., Wang, J., and Chen, Y. Surrogate scoring rules and a dominant truth serum. *ACM EC*, 2020.

Menon, A., Van Rooyen, B., Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pp. 125–134, 2015.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *Advances in neural information processing systems*, pp. 1196–1204, 2013.

Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., and Brox, T. Self: Learning to filter noisy labels with self-ensembling. *arXiv preprint arXiv:1910.01842*, 2019.

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

Northcutt, C. G., Wu, T., and Chuang, I. L. Learning with confident examples: Rank pruning for robust classification with noisy labels. *UAI*, 2017.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. 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.

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

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

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., and Faloutsos, C. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

Steenbrugge, X., Leroux, S., Verbelen, T., and Dhoedt, B. Improving generalization for abstract reasoning tasks using disentangled feature representations. *arXiv preprint arXiv:1811.04784*, 2018.

Traganitis, P. A., Pages-Zamora, A., and Giannakis, G. B. Blind multiclass ensemble classification. *IEEE Transactions on Signal Processing*, 66(18):4737–4752, 2018.

Van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv e-prints*, pp. arXiv–1807, 2018.

Wang, J., Liu, Y., and Levy, C. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 526–536, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445915. URL https://doi.org/10.1145/3442188.3445915.

Wang, Q., Yao, J., Gong, C., Liu, T., Gong, M., Yang, H., and Han, B. Learning with group noise. *arXiv preprint arXiv:2103.09468*, 2021b.

Wang, T., Yue, Z., Huang, J., Sun, Q., and Zhang, H. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34, 2021c.

Wei, H., Feng, L., Chen, X., and An, B. 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*, pp. 13726–13735, 2020.

Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., and Liu, Y. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

Wei, J., Zhu, Z., Niu, G., Liu, T., Liu, S., Sugiyama, M., and Liu, Y. Fairness improves learning from noisily labeled long-tailed data. *arXiv preprint arXiv:2303.12291*, 2023.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32, 2019.

Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., and Chang, Y. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*, 2020a.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020b.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2691–2699, 2015.

Yang, S., Yang, E., Han, B., Liu, Y., Xu, M., Niu, G., and Liu, T. Estimating instance-dependent label-noise transition matrix using dnns. *arXiv preprint arXiv:2105.13001*, 2021.

Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., and Sugiyama, M. Dual t: Reducing estimation error for transition matrix in label-noise learning. *arXiv preprint arXiv:2006.07805*, 2020a.

Yao, Y., Liu, T., Han, B., Gong, M., Niu, G., Sugiyama, M., and Tao, D. Towards mixture proportion estimation without irreducibility. *arXiv preprint arXiv:2002.03673*, 2020b.

Yao, Y., Liu, T., Gong, M., Han, B., Niu, G., and Zhang, K. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Advances in neural information processing systems*, 27, 2014.

Zhang, Y., Niu, G., and Sugiyama, M. Learning noise transition matrix from only noisy labels via total variation regularization. *arXiv preprint arXiv:2102.02414*, 2021.

Zheltonozhskii, E., Baskin, C., Mendelson, A., Bronstein, A. M., and Litany, O. 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*, pp. 1657–1667, 2022.

Zhu, X. *Semi-supervised learning with graphs*. Carnegie Mellon University, 2005.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919, 2003.

Zhu, Z., Dong, Z., Cheng, H., and Liu, Y. A good representation detects noisy labels. *arXiv preprint arXiv:2110.06283*, 2021a.

Zhu, Z., Liu, T., and Liu, Y. A second-order approach to learning with instance-dependent label noise. *CVPR*, 2021b.

Zhu, Z., Song, Y., and Liu, Y. Clusterability as an alternative to anchor points when learning with noisy labels. *ICML*, 2021c.

Zhu, Z., Wang, J., and Liu, Y. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. *arXiv preprint arXiv:2202.01273*, 2022a.

Zhu, Z., Yao, Y., Sun, J., Liu, Y., and Li, H. Evaluating fairness without sensitive attributes: A framework using only auxiliary models. *arXiv preprint arXiv:2210.03175*, 2022b.

# 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.

## Notation Table

We provide the main notations below:

*Table 2.* Table of Notations for Frequently Used Variables

| | | |
|---|---|---|
| $X$ | $\triangleq$ | sample |
| $Y$ | $\triangleq$ | clean label |
| $\tilde{Y}$ | $\triangleq$ | noisy label |
| $K$ | $\triangleq$ | number of classes |
| $K_X$ | $\triangleq$ | number of non-degenerated label classes |
| $T(X)$ | $\triangleq$ | label noise transition matrix of $X$ |
| $\mathsf{Kr}(M)$ | $\triangleq$ | Kruskal rank of matrix $M$ |
| $e_+, e_-$ | $\triangleq$ | noise rate for label $+1$ and label $-1$, respectively. |
| $\pi_+, \pi_-$ | $\triangleq$ | inverse noise rate $\mathbb{P}(Y\|\tilde{Y})$ for label $+1$ and label $-1$, respectively. |
| $O_i, i = 1, ..., p$ | $\triangleq$ | observed variables. |
| $Z$ | $\triangleq$ | hidden variable |
| $\|G\|$ | $\triangleq$ | number of groups in $X$ |
| $R(X)$ | $\triangleq$ | learned representation of $X$ |
| $d^*$ | $\triangleq$ | dimension of $R(X)$ |

## A. Omitted Proofs

### Proof for Lemma 3.3

*Proof.* Using Bayes rule we easily obtain

$$\begin{aligned}\mathbb{P}(X|\tilde{Y}=+1) = \mathbb{P}(X|Y=+1)\cdot\mathbb{P}(Y=+1|\tilde{Y}=+1)\\ +\mathbb{P}(X|Y=-1)\cdot\mathbb{P}(Y=-1|\tilde{Y}=+1)\end{aligned} \tag{2}$$

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

$$\begin{aligned}\mathbb{P}(X|\tilde{Y}=-1) = \mathbb{P}(X|Y=+1)\cdot\mathbb{P}(Y=+1|\tilde{Y}=-1)\\ +\mathbb{P}(X|Y=-1)\cdot\mathbb{P}(Y=-1|\tilde{Y}=-1)\end{aligned} \tag{3}$$

Since both $\mathbb{P}(X|Y=+1), \mathbb{P}(X|Y=-1)$ are unknown, solving Eqn. (2) and (3) we further have

$$\mathbb{P}(X|\tilde{Y}=-1) = \tilde{\pi}_-\cdot\mathbb{P}(X|\tilde{Y}=+1) + (1-\tilde{\pi}_-)\cdot\mathbb{P}(X|Y=-1) \tag{4}$$

$$\mathbb{P}(X|\tilde{Y}=+1) = \tilde{\pi}_+\cdot\mathbb{P}(X|\tilde{Y}=-1) + (1-\tilde{\pi}_+)\cdot\mathbb{P}(X|Y=+1). \tag{5}$$

□

### Proof for Theorem 3.4

*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 $\pi_-, \pi_+$. Next we show that identifying $\pi_-, \pi_+$ 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 $\pi_+, \pi_i$ but are able to identify $\{e_+, e_-\}$. This implies that there exists another pair $\{\pi'_+, \pi'_-\} \neq \{\pi_+, \pi_-\}$ such that (denote by $\tilde{p} := \mathbb{P}(\tilde{Y} = +1)$)

$$\mathbb{P}(\tilde{Y} = +1|Y = -1) = \frac{\pi_+\tilde{p}}{\pi_+\tilde{p} + (1 - \pi_-)(1 - \tilde{p})} \tag{6}$$

$$= \frac{\pi'_+\tilde{p}}{\pi'_+\tilde{p} + (1 - \pi'_-)(1 - \tilde{p})} \tag{7}$$

$$\mathbb{P}(\tilde{Y} = -1|Y = +1) = \frac{\pi_-(1 - \tilde{p})}{(1 - \pi_+)\tilde{p} + \pi_-(1 - \tilde{p})} \tag{8}$$

$$= \frac{\pi'_-(1 - \tilde{p})}{(1 - \pi'_+)\tilde{p} + \pi'_-(1 - \tilde{p})} \tag{9}$$

By dividing $\pi_+, \pi'_+$ in both the numerator and denominator in Eqn. (6) and (7), we conclude that

$$\frac{1 - \pi_-}{\pi_+} = \frac{1 - \pi'_-}{\pi'_+} \tag{10}$$

While from Eqn (8) and (refeqn:e+:2) we conclude:

$$\frac{1 - \pi_+}{\pi_-} = \frac{1 - \pi'_+}{\pi'_-} \tag{11}$$

From Eqn. (10) and (11) we have

$$(1 - \pi_-)\pi'_+ = (1 - \pi'_-)\pi_+ \tag{12}$$

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

Taking the difference and re-arrange terms we prove

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

From Eqn. (10) again, taking $-1$ on both side we have

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

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

## Proof for Theorem 4.2

*Proof.* We first prove sufficiency. We first relate our problem setting to the setup of Kruskal's identifiability scenario: $Y \in \{1, 2, ..., K_X\}$ 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 $O_i$. $\kappa_i$ 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 a rank $K_X$, we know immediately that all rows in $M_1, M_2, M_3$ are linearly independent. Therefore, the Kruskal ranks satisfy

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

Checking the condition in Theorem 3.6, we easily verify

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

Calling Theorem 3.6 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.

For completeness we provide rationals for the multi-class case too. The idea is to show that the complete information returned by the single noisy label and two noisy labels do not always guarantee a unique solution.

For the first order information:

$$\begin{aligned}
\mathbb{P}(\tilde{Y} = i|X) &= \sum_{k \in [K]} \mathbb{P}(Y = k) \cdot \mathbb{P}(\tilde{Y} = i|Y = k, X) \\
&= \sum_{k \in [K]} \mathbb{P}(Y = k|X) \cdot T_{ki}(X)
\end{aligned}$$

Enumerating all $i$s, there are $K$ equations, written in a matrix form as:

$$\tilde{\mathbf{P}} = (T(X))^\top \cdot \mathbf{P}$$

where $\tilde{\mathbf{P}}$ is the vector form for $[\mathbb{P}(\tilde{Y} = 1|X); \mathbb{P}(\tilde{Y} = 2|X); ...; \mathbb{P}(\tilde{Y} = K|X)]$ and $\mathbf{P}$ is the one for $\mathbb{P}(Y = k|X)$.

For the second order information

$$\begin{aligned}
\mathbb{P}(\tilde{Y}_1 = i, \tilde{Y}_2 = j|X) &= \sum_{k \in [K]} \mathbb{P}(Y = k|X) \cdot \mathbb{P}(\tilde{Y}_1 = i|Y = k, X) \cdot \mathbb{P}(\tilde{Y}_2 = j|Y = k, X) \\
&= \sum_{k \in [K]} \mathbb{P}(Y = k|X) \cdot T_{ki}(X) \cdot T_{kj}(X)
\end{aligned}$$

Enumerating pairs of $(i, j)$ we have $K^2$ equations, written in matrix form as:

$$C = (T(X))^\top \cdot \Lambda \cdot T(X)$$

where in above $C$ is a $K \times K$ matrix with the $(i, j)$-th entry being $\mathbb{P}(\tilde{Y}_1 = i, \tilde{Y}_2 = j|X)$; $\Lambda$ is a diagonal matrix with $\Lambda_{ii} = \mathbb{P}(Y = k|X)$.

Notice that

$$\sum_j \mathbb{P}(\tilde{Y}_1 = i, \tilde{Y}_2 = j|X) = \mathbb{P}(\tilde{Y}_1 = i)$$

and

$$\sum_j \sum_{k\in[K]} \mathbb{P}(Y = k|X) \cdot T_{ki}(X) \cdot T_{kj}(X) = \sum_{k\in[K]} \mathbb{P}(Y = k|X) \cdot T_{ki}(X)$$

we know that for every $K$ equations from the second order information, there is at least one redundant equation. That is to conclude that we have at most $K + K^2 - K = K^2$ independent equations. Nonetheless, we have $K(\mathbb{P}(Y = k|X)) + K^2(T(X)) = K^2 + K$ unknown variables. So the equations are under-determined. Therefore we conclude for the general $K$, there exists cases two labels will not define a unique solution. For instance, for $K = 3$, we can easily find the following two sets of parameter settings will return us the same observed distribution for two labels:

Parameter setting 1:

$$[\mathbb{P}(Y = 1|X), \mathbb{P}(Y = 2|X), \mathbb{P}(Y = 3|X)] = [0.35, 0.35, 0.3], T(X) = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.175 & 0.65 & 0.175 \\ 0.15 & 0.15 & 0.7 \end{bmatrix}$$

Parameter setting 2:

$$[\mathbb{P}(Y = 1|X), \mathbb{P}(Y = 2|X), \mathbb{P}(Y = 3|X)] = [0.31, 0.34, 0.35], T(X) = \begin{bmatrix} 0.65 & 0.175 & 0.175 \\ 0.175 & 0.65 & 0.175 \\ 0.175 & 0.175 & 0.65 \end{bmatrix}$$

Similar examples can be obtained by searching through the solutions space of the equations.

□

## Proof for Theorem 5.2

*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

$$N_X := \sum_{i=1}^{N} 1[X_i = X],\ \mathbb{E}[N_X] = \frac{q_X}{\sum_{X\in\mathcal{X}} q_X} N \tag{15}$$

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

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

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

$$\mathbb{P}(N_X > 3, \forall X) = 1 - \mathbb{P}(N_X \leq 3, \exists X) \leq 1 - Nexp(-2N) \tag{16}$$

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 $\epsilon$. Therefore they will share the same true label.

Note that we are not imagining the exact same data appearing three times, but rather that three different data that happen to have the same pattern $X$ that appeared three times ([2]). For instance, these three $X$s can correspond to three independent users trying to apply for a credit card and ending up having the same application profiles (e.g., age, salary range, education level etc); it can also be three similar cat images ended up with the same encoding of the features. □

## Proof for Theorem 5.5

*Proof.* The $d^*$ features and the noisy label $\tilde{Y}$ jointly give us $d^*+1$ independent observations. Denote by $K_G^* = \max_{X\in G} K_X$. In Kruskal's setup, $Y \in \{1, 2, ..., K_G^*\}$ will then correspond. to the unobserved hidden variable $Z$. If the noisy label is

informative we know that $\mathsf{Kr}(T(X)) = K_G^* \leq K$. Then checking Kruskal's condition we have:

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

Calling Theorem 3.6, we establish the identifiability. □

## Proof for Theorem 5.6

*Proof.* By definition

$$||T_1(X) - T^*(X)||_F = \sqrt{\sum_i \sum_j (T_1[i,j] - T[i,j])^2} \tag{17}$$

Easy to show that

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

Then we prove that

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

□

## Proof for Theorem 5.7

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

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

Note $|G| \cdot K$ is the size of space for the unobserved variable ($G \times Y$ renumbered as $\{1, 2, ..., |G|K\}$). □

**Algorithm 1** Key Steps of HOC

**Input:** Noisy dataset: $\widetilde{D} = \{(x_n, \tilde{y}_n)\}_{n\in[N]}$, with disentangled features.
*//Find 2-NN using a similarity function* $\mathsf{Sim}(x, x')$.
With $1 - \mathsf{Sim}(x, x')$ as the distance metric:
$\{(\tilde{y}_n, \tilde{y}_{n_1}, \tilde{y}_{n_2}), \forall n\} \leftarrow \texttt{Get2NN}(\widetilde{D})$;
*//Count first-, second, and third-order consensus patterns:*
$(\hat{c}^{[1]}, \hat{c}^{[2]}, \hat{c}^{[3]}) \leftarrow \texttt{CountFreq}(\{(\tilde{y}_n, \tilde{y}_{n_1}, \tilde{y}_{n_2}), \forall n\})$
*//Solve equations:*
Find $T$ such that match the counts $(\hat{c}^{[1]}, \hat{c}^{[2]}, \hat{c}^{[3]})$.

## B. Generic identifiability

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

**Theorem B.1.** *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 $d^* \geq \lceil \log_2 \frac{2K_G^*+1}{2} \rceil$, and $\tau_i \geq 2$.*

*Proof.* We first reproduce a relevant theorem in (Allman et al., 2009):

**Theorem B.2.** *(Allman et al., 2009) When $p = 3$ (3 independently observations), the model parameters are generically identifiable, up to label permutation, if*

$$\min(K_G^*, \kappa_1) + \min(K_G^*, \kappa_2) + \min(K_G^*, \kappa_3) \geq 2K_G^* + 2 \tag{18}$$

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:

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

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

$$\tau_1^* = \prod_{i=1}^{d_1^*} \geq 2^{d_1^*} \geq 2^{\lceil \log_2 \frac{2K_G^*+1}{2} \rceil} \geq \frac{2K_G^*+1}{2} \tag{19}$$

Similarly $\tau_2^* \geq \frac{K_X+2}{2}$. 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:

$$\mathsf{Kr}(T(X)) + \mathsf{Kr}(M_1^*) + \mathsf{Kr}(M_2^*) \geq 1 + 2\frac{2K_G^*+1}{2} = 2K_G^* + 2,$$

which again satisfied the identifiability condition specified in Theorem 3.6. □

## C. More experiments

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

### C.1. More training details for Table 1

We present the key steps of HOC estimator in Algorithm 1 and the instance-dependent label noise generation in Algorithm 2.

**Algorithm 2** Instance-Dependent Label Noise Generation

**Input:**

1: Clean examples $(x_n, y_n)_{n=1}^N$; Noise rate: $\varepsilon$; 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 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 |

*Table 4.* 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.2. Training performance using estimated transition matrix

We can further use the estimated transition matrix to perform forward loss correction (FW) (**?**). Table 3 records the performance of FW by using the estimated transition matrix of SimCLR and IPIRM. The hyper-parameters for all the experiments in Table 3 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.

### 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 4 shows the effect of using disentangled features as DNN initialization on CIFAR100. The hyper-parameters for all the experiments in Table 4 are consistent with Table 3. From the results, We can observe that even with vanilla Cross Entropy loss, the disentangled features are still beneficial to the performance.

### C.4. verifying the importance of characterizing the identifiability of the label noise transition matrix

#### C.4.1. CIFAR10 EXPERIMENT

Our first experiment is to show that when estimated transition matrices is far from the ground-truth matrix, it may make model perform worse even compared to the baseline (vanilla training with Cross Entropy).

**Experiment setting:** The training framework with transition matrix is followed from FW (Patrini et al., 2017). The dataset is CIFAR10 and the network structure is ResNet34. The hyper-parameters are as follows: batchsize (64), learning rate (0.1 for first 50 epochs and 0.01 for last 50 epochs), optimizer (SGD). For a randomly selected set of instances ($50\%$ of the

*Table 5.* Comparison of test accuracy on CIFAR10 by using different transition matrix.

| | CE | FW with $T_1$ | FW with $T_2$ | FW with $T_3$ |
|---|---|---|---|---|
| Test accuracy | 79.34 | 82.62 | 81.65 | 78.13 |

population), we generate noisy labels using the following transition matrix:

$$T = \mathbb{P}(\widetilde{Y}|Y,X) = \begin{bmatrix} \mathbf{0.9} & 0.011 & 0.011 & 0.011 & 0.011 & 0.011 & 0.011 & 0.011 & 0.011 & 0.011 \\ 0.019 & \mathbf{0.82} & 0.019 & 0.019 & 0.019 & 0.019 & 0.019 & 0.019 & 0.019 & 0.019 \\ 0.028 & 0.028 & \mathbf{0.74} & 0.028 & 0.028 & 0.028 & 0.028 & 0.028 & 0.028 & 0.028 \\ 0.037 & 0.037 & 0.037 & \mathbf{0.66} & 0.037 & 0.037 & 0.037 & 0.037 & 0.037 & 0.037 \\ 0.045 & 0.045 & 0.045 & 0.045 & \mathbf{0.58} & 0.045 & 0.045 & 0.045 & 0.045 & 0.045 \\ 0.054 & 0.054 & 0.054 & 0.054 & 0.054 & \mathbf{0.51} & 0.054 & 0.054 & 0.054 & 0.054 \\ 0.063 & 0.063 & 0.063 & 0.063 & 0.063 & 0.063 & \mathbf{0.43} & 0.063 & 0.063 & 0.063 \\ 0.071 & 0.071 & 0.071 & 0.071 & 0.071 & 0.071 & 0.071 & \mathbf{0.35} & 0.071 & 0.071 \\ 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & \mathbf{0.27} & 0.08 \\ 0.088 & 0.088 & 0.088 & 0.088 & 0.088 & 0.088 & 0.088 & 0.088 & 0.088 & \mathbf{0.2} \end{bmatrix}$$

The above transition matrix $T$ is uniform off-diagonal with diagonals evenly spaced over $[0.9, 0.2]$, which is the ground-truth transition matrix in our setting. The remaining unselected instances will enjoy a $T \equiv 0$.

We perform experiments using the following three uniform off-diagonal transition matrix with forward loss correction (**?**):

- $T_1$ with diagonals evenly spaced over $[0.9, 0.2]$
- $T_2$ with all diagonals $0.4$.
- $T_3$ with diagonals evenly spaced over $[0.2, 0.9]$

where $T_1$ is the ground-truth transition matrix while $T_3$ is far from the ground-truth. The results are listed in Table 5.

It can be observed that when using $T_3$, the performance is even worse than vanilla training with Cross Entropy, suggesting the importance of identifying and estimating the noise transition matrix.

### C.4.2. Gaussian experiment

Our second experiment is to show that in some settings, the transition matrix is hard to estimate correctly, which suggests the importance of identifiability. Consider a simple setting for binary classification and a set of instances generated according to the following setups:

- $X \sim \mathcal{N}(0, 3)$ where $\mathcal{N}$ denotes Gaussian distribution with mean 0 and variance 3.
- $\mathbb{P}(Y = 1|X) = \text{sigmoid}(X) = \frac{1}{1+e^{-X}}$

We generate $X$ and $Y$ following the above procedure and define the ground-truth transition matrix $T = \mathbb{P}(\widetilde{Y}|Y,X) = \begin{bmatrix} \mathbf{0.9} & 0.1 \\ 0.2 & \mathbf{0.8} \end{bmatrix}$ for generating $\widetilde{Y}$ from $Y$. Our goal is to examine whether we can estimate the correct transition matrix using $(X, \widetilde{Y})$.

**Experiment setting:** The training framework for estimating transition matrix is followed from FW (Patrini et al., 2017). We randomly sample 5000 $(x, y)$ pairs from the data generating procedure and using $T = \mathbb{P}(\widetilde{Y}|Y) = \begin{bmatrix} \mathbf{0.9} & 0.1 \\ 0.2 & \mathbf{0.8} \end{bmatrix}$ to generate $\widetilde{Y}$ from $Y$. The network structure is a simple FCN (fully connected network tructure) with one hidden layer (10 nodes) and ReLU activation. The hyper-parameters are as follows: learning rate (0.01 for 100 epochs ), optimizer (SGD). We perform the experiments with 30 runs and record the average performance in Table 6.

*Table 6.* Comparison of test accuracy for CE and FW.

| | CE | FW with estimated transition matrix |
|---|---|---|
| Test accuracy | 83.22 | 83.31 |

From Table 6, we can see that FW has very little gain compared to vanilla Cross Entropy training. We then calculate the average estimated transition matrix:

$$T_{estimated} = \begin{bmatrix} \mathbf{0.983} & 0.017 \\ 0.008 & \mathbf{0.992} \end{bmatrix}$$

We find that $T_{estimated}$ is nearly as the same as the identity matrix, suggesting that in this setting, FW is hard to estimate noise transition matrix correctly and contributes less to the performance.