Supplementary Material:

Robust Inference via Generative Classifiers for Handling Noisy Labels

A. Preliminaries

Gaussian discriminant analysis. In this section, we describe the basic concepts of the discriminative and generative classifier (Ng & Jordan, 2002). Formally, denote the random variable of the input and label as $\mathbf{x}$ and $y = \{1, \cdots, C\}$, respectively. For the classification task, the discriminative classifier directly defines a posterior distribution $P(y|\mathbf{x})$, i.e., learning a direct mapping between input $\mathbf{x}$ and label y. A popular model for discriminative classifier is softmax classifier which defines the posterior distribution as follows: $P(y=c|\mathbf{x}) = \frac{\exp\left(\mathbf{w}_c^\top \mathbf{x} + b_c\right)}{\sum_{c'} \exp\left(\mathbf{w}_{c'}^\top \mathbf{x} + b_{c'}\right)}$, where $\mathbf{w}_c$ and b_c are weights and bias for a class c, respectively. In contrast to the discriminative classifier, the generative classifier defines the class conditional distribution $P(\mathbf{x}|y)$ and class prior $P(y)$ in order to indirectly define the posterior distribution by specifying the joint distribution $P(\mathbf{x}, y) = P(y) P(\mathbf{x}|y)$. Gaussian discriminant analysis (GDA) is a popular method to define the generative classifier by assuming that the class conditional distribution follows the multivariate Gaussian distribution and the class prior follows Bernoulli distribution: $P(\mathbf{x}|y=c) = \mathcal{N}(\mathbf{x}|\mu_c, \mathbf{\Sigma}_c), P(y=c) = \frac{\beta_c}{\sum_{c'} \beta_{c'}}$, where μ_c and $\mathbf{\Sigma}_c$ are the mean and covariance of multivariate Gaussian distribution, and β_c is the unnormalized prior for class c. This classifier has been studied in various machine learning areas (e.g., semi-supervised learning (Lasserre et al., 2006) and incremental learning (Lee et al., 2018)).

In this paper, we focus on the special case of GDA, also known as the linear discriminant analysis (LDA). In addition to Gaussian assumption, LDA further assumes that all classes share the same covariance matrix, i.e., $\mathbf{\Sigma}_c = \mathbf{\Sigma}$. Since the quadratic term is canceled out with this assumption, the posterior distribution of generative classifier can be represented as follows:

$$P(y=c|\mathbf{x}) = \frac{P(y=c) P(\mathbf{x}|y=c)}{\sum_{c'} P(y=c') P(\mathbf{x}|y=c')} = \frac{\exp\left(\mu_c^\top \mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mu_c^\top \mathbf{\Sigma}^{-1}\mu_c + \log \beta_c\right)}{\sum_{c'} \exp\left(\mu_{c'}^\top \mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mu_{c'}^\top \mathbf{\Sigma}^{-1}\mu_{c'} + \log \beta_{c'}\right)}.$$

One can note that the above form of posterior distribution is equivalent to the softmax classifier by considering $\mu_c^\top \mathbf{\Sigma}^{-1}$ and $-\frac{1}{2}\mu_c^\top \mathbf{\Sigma}^{-1}\mu_c + \log \beta_c$ as its weight and bias, respectively. This implies that $\mathbf{x}$ might be fitted in Gaussian distribution during training a softmax classifier.

Breakdown points. The robustness of MCD estimator can be explained by the fact that it has high breakdown points (Hampel, 1971). Specifically, the breakdown point of an estimator measures the smallest fraction of observations that need to be replaced by arbitrary values to carry the estimate beyond all bounds. Formally, denote $\mathcal{Y}_M$ as a set obtained by replacing M data points of set $\mathcal{Y}$ by some arbitrary values. Then, for a multivariate mean estimator $\mu = \mu(\mathcal{Y})$ from $\mathcal{Y}$, the breakdown point is defined as follows (see Appendix A for more detailed explanations including the breakdown point of covariance estimator):

$$\varepsilon^*(\mu, \mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \min\left\{ M \in \{1, \cdots, |\mathcal{Y}|\} : \sup_{\mathcal{Y}_M} \|\mu(\mathcal{Y}) - \mu(\mathcal{Y}_M)\| = \infty \right\}.$$

For a multivariate estimator of covariance $\mathbf{\Sigma}$, we have

$$\varepsilon^*(\mathbf{\Sigma}, \mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \min\{M \in \{1, \cdots, |\mathcal{Y}|\} : \sup_{M} \max_{i} \{|\log \lambda_i(\mathbf{\Sigma}(\mathcal{Y})) - \log \lambda_i(\mathbf{\Sigma}(\mathcal{Y}_M))|\}\},$$

where the $k-$th largest eigenvalue of a general $n \times n$ matrix is denoted by $\lambda_k(\mathbf{\Sigma})$, $k = 1, \cdots, n$ such that $\lambda_1(\mathbf{\Sigma}) \leq \lambda_2(\mathbf{\Sigma}) \leq \cdots \leq \lambda_n(\Sigma)$. This implies that we consider a covariance estimator to be broken whenever any of the eigenvalues can become arbitrary large or arbitrary close to 0.

B. Experimental setup

We describe the detailed explanation about all the experiments in Section 4. The code is available at `[anonymized]`.

Detailed model architecture and datasets. We consider two state-of-the-art neural network architectures: DenseNet (Huang & Liu, 2017) and ResNet (He et al., 2016). For DenseNet, our model follows the same setup as in Huang & Liu

(2017): 100 layers, growth rate $k = 12$ and dropout rate 0. Also, we use ResNet with 34 and 44 layers, filters = 64 and dropout rate 0[5]. The softmax classifier is used, and each model is trained by minimizing the cross-entropy loss. We train DenseNet and ResNet for classifying CIFAR-10 (or 100) and SVHN datasets: the former consists of 50,000 training and 10,000 test images with 10 (or 100) image classes, and the latter consists of 73,257 training and 26,032 test images with 10 digits.[6] By following the experimental setup of Ma et al. (2018), All networks were trained using SGD with momentum 0.9, weight decay 10^{-4} and an initial learning rate of 0.1. The learning rate is divided by 10 after epochs 40 and 80 for CIFAR-10/SVHN (120 epochs in total), and after epochs 80, 120 and 160 for CIFAR-100 (200 epochs in total). For our method, we extract the hidden features at {79,89,99}-th layers and {27,29,31,33}-th layers for DenseNet and ResNet, respectively. We assume the uniform class prior distribution.

In the Table 5, we evaluate RoG for the NLP tasks on Twitter and Reuters dataset: the former has a task of part-of-speech (POS) tagging, and the latter has a task of text categorization. The Twitter dataset consists of 14,619 training data from 25 different classes, while some of classes only contain a small number of training data. We exclude such classes that are smaller than 100 in size, then we finally have 14,468 training data and 7,082 test data from 19 different classes. Similarly, we exclude the classes which have a size of less than 100 in the Reuters dataset. Then it consists of 5,444 training data and 2,179 test data from 7 different classes. For training, we use 2-layer FCNs with ReLU non-linearity and uniform noise on both Twitter and Reuters datasets and we extract the hidden features at both layers for 2-layer FCNs.

We also consider the open-set noisy scenarios (Wang et al., 2018). Figure 4 shows the examples of open-set noisy dataset which is built by replacing some training images in CIFAR-10 by external images in CIFAR-100 and Downsampled ImageNet (Chrabaszcz et al., 2017) which is equivalent to the ILSVRC 1,000-class ImageNet dataset (Deng et al., 2009), but with images downsampled to 32×32 resolution. In the Table 7, we maintain the original label of the CIFAR-10 dataset, while replacing 60% of the training data in CIFAR-10 with training data in CIFAR-100 and Downsampled ImageNet.

Validation. For our methods, the ensemble weights are chosen by optimizing the NLL loss over the validation set. We assume that the validation set consists of 1000 images with (same type and fraction of) noisy labels. However, one can expect that if we use all validation samples, the performance of our method can be affected by outliers. To relax this issue, we use only half of them, chosen by the MCD estimator. Specifically, we first compute the Mahalanobis distance for all validation samples using the parameters from MCD estimator, and select 500 samples with smallest distance. Then, one can expect that our ensemble method is more robust against the noisy labels in validation sets. In the case of Twitter and Reuters dataset, validation set consists of 570 and 210 samples with noisy labels, respectively.

Training method for noisy label learning. We consider the following training methods for noisy label learning:

(a) **Hard bootstrapping** (Reed et al., 2014): Training with new labels generated by a convex combination (the hard version) of the noisy labels and their predicted labels.

(b) **Soft bootstrapping** (Reed et al., 2014): Training with new labels generated by a convex combination (the soft version) of the noisy labels and their predictions.

(c) **Backward** (Patrini et al., 2017): Training via loss correction by multiplying the cross-entropy loss by a noise-aware correction matrix.

(d) **Forward** (Patrini et al., 2017): Training with label correction by multiplying the network prediction by a noise-aware correction matrix.

(e) **Forward (Gold)** (Hendrycks et al., 2018): An augmented version of Forward method which replaces its corruption matrix estimation with the identity on trusted samples.

(f) **GLC (Gold Loss Correction)** (Hendrycks et al., 2018): Training with the corruption matrix which is estimated by using the trusted dataset.

(g) **D2L** (Ma et al., 2018): Training with new labels generated by a convex combination of the noisy labels and their predictions, where its weights are chosen by utilizing the Local Intrinsic Dimensionality (LID).

(h) **Decoupling** (Malach & Shalev-Shwartz, 2017): Updating the parameters only using the samples which have different prediction from two classifier.

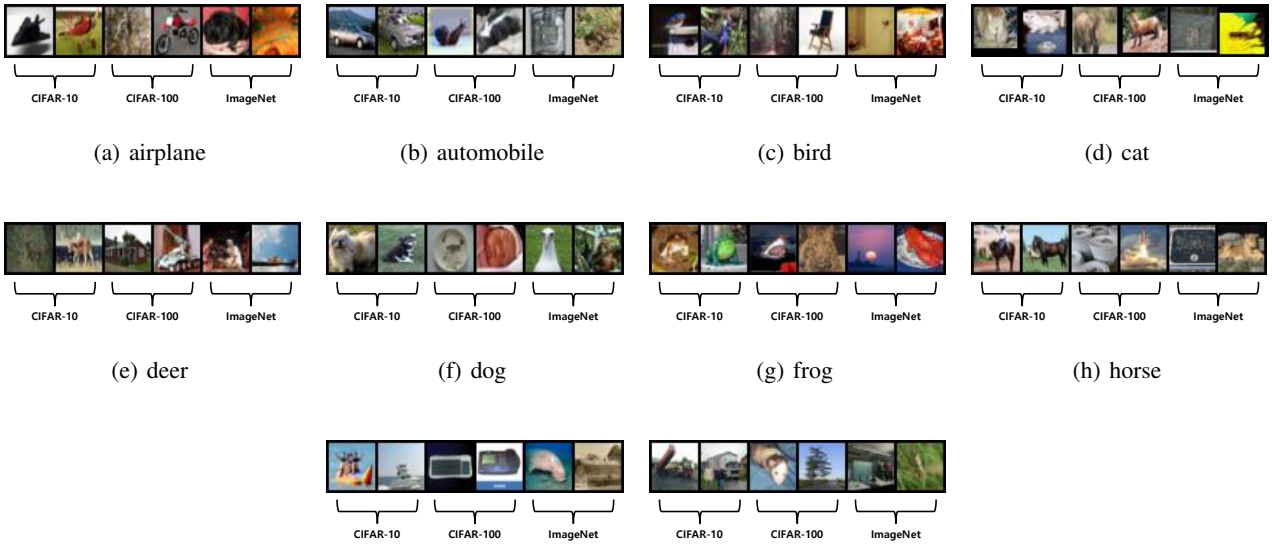
[5]ResNet architecture is available at `https://github.com/kuangliu/pytorch-cifar`.

[6]We do not use the extra SVHN dataset for training.

(i) **MentorNet** (Jiang et al., 2018): An extra teacher network is pre-trained and then used to select clean samples for its student network.

(j) **Co-teaching** (Han et al., 2018b): A simple ensemble method where each network selects its small-loss training data and back propagates the training data selected by its peer network.

(k) **Cross-entropy**: the conventional approach of training with cross-entropy loss.



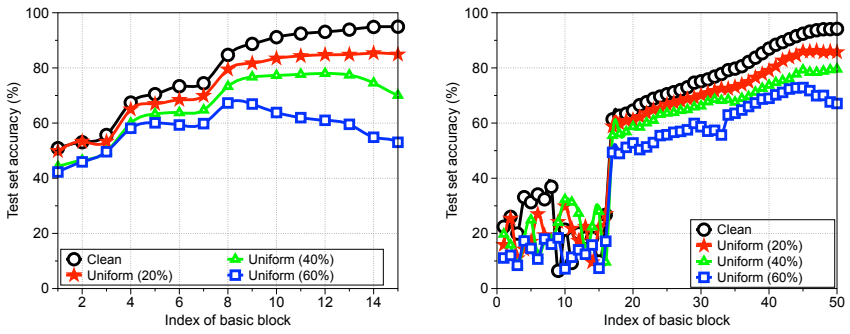
(a) airplane (b) automobile (c) bird (d) cat

(e) deer (f) dog (g) frog (h) horse

(i) ship (j) truck

Figure 4. Examples of open-set noise in CIFAR-10 dataset.

C. Layer-wise characteristics of generative classifiers



(a) Generalization from noisy labels (b) Generalization from noisy labels

Figure 5. Layer-wise characteristics of generative classifiers from (a) ResNet-34 and (b) DenseNet-100 trained on the CIFAR-10 dataset.

Figure 5 shows the classification accuracy of the generative classifiers from different basic blocks of ResNet-34 (He et al., 2016) and DenseNet-100 (Huang & Liu, 2017). One can note that the generative classifiers from DenseNet and ResNet have different patterns due to the architecture design. In the case of DenseNet, we found that it produces meaningful features after 20-th basic blocks.

D. Proof of Theorem 1

In this section, we present a proof of Theorem 1, which consists of two statements: the limit of estimation error (3) and estimated error ratio (4). We prove both statements one by one as stated in below. For convenience, we skip to mention the Continuous Mapping Theorem[7] and the number of training samples N goes to infinity for all convergences in the proof.

D.1. Proof of the limit of estimation error (3)

We start with a following lemma, which shows the convergences of sample and MCD estimators as the number of training samples N goes to infinity.

Lemma 1 *Suppose we have N number of d-dimensional training samples $\mathcal{X}_N = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ and $\mathcal{X}_N$ contains outlier samples with the fixed fraction $\delta_{\texttt{out}} < 1$. We assume the outlier samples are from an arbitrary distribution $P_{\texttt{out}}$ with zero mean and finite covariance matrix $\sigma^2_{\texttt{out}}\mathbf{I}$, and the clean samples are from a distribution of the hidden features of DNNs $P_{\texttt{clean}}$ with mean μ and covariance matrix $\sigma^2\mathbf{I}$. Let $\bar{\mu}$ and $\bar{\mathbf{\Sigma}}$ be the mean and covariance matrix of sample estimator, and let $\widehat{\mu}$ and $\widehat{\mathbf{\Sigma}}$ be the mean and covariance matrix of MCD estimator which selects samples from $\mathcal{X}_N$ with the fixed fraction $\frac{d}{N} < \delta_{\texttt{mcd}} < 1$ to optimize its objective (2). Then the mean and covariance matrix of sample estimator converge almost surely to below as $N \to \infty$:*

$$\bar{\mu} \overset{\text{a.s.}}{\to} (1-\delta_{\texttt{out}})\,\mu, \quad \bar{\mathbf{\Sigma}} \overset{\text{a.s.}}{\to} \left((1-\delta_{\texttt{out}})\,\sigma^2 + \delta_{\texttt{out}}\sigma^2_{\texttt{out}}\right)\mathbf{I} + \delta_{\texttt{out}}\,(1-\delta_{\texttt{out}})\,\mu\mu^T.$$

In addition, if $\delta_{\texttt{mcd}} \leq 1-\delta_{\texttt{out}}$ and $\sigma^2 < \sigma^2_{\texttt{out}}$, the mean and covariance matrix of MCD estimator converge almost surely to below as $N \to \infty$:

$$\widehat{\mu} \overset{\text{a.s.}}{\to} \mu, \qquad \widehat{\mathbf{\Sigma}} \overset{\text{a.s.}}{\to} \sigma^2\mathbf{I}.$$

A proof of the lemma is given in appendix D.3, where it is built upon the fact that the determinant of covariance matrix with some assumptions can be expressed as the d-th degree polynomial of outlier ratio.

Lemma 1 states the convergences of sample and MCD estimators on a single distribution of hidden features of DNNs. Without loss of generality, one can assume the mean of outlier distribution is zero, i.e., $\mu_{\texttt{out}} = 0$ by an affine translation of hidden features. Furthermore, one can extend Lemma 1 to C number of distributions, which have the class mean μ_c and class covariance matrix $\mathbf{\Sigma}_c$ on each class label $c \in \{1, ..., C\}$ with the assumptions $\mathcal{A}1 \sim \mathcal{A}4$. Then the class mean of MCD and sample estimators converge almost surely as follows:

$$\widehat{\mu}_c \overset{\text{a.s.}}{\to} \mu_c, \qquad \bar{\mu}_c \overset{\text{a.s.}}{\to} (1-\delta_{\texttt{out}})\,\mu_c,$$

which implies that

$$\|\mu_c - \widehat{\mu}_c\|_1 \overset{\text{a.s.}}{\to} 0, \qquad \|\mu_c - \bar{\mu}_c\|_1 \overset{\text{a.s.}}{\to} \delta_{\texttt{out}}\|\mu_c\|_1.$$

This completes the proof of the limit of estimation error (3).

D.2. Proof of the limit of estimated error ratio (4)

Recall the class mean of MCD and sample estimators converge almost surely as follows:

$$\widehat{\mu}_c \overset{\text{a.s.}}{\to} \mu_c, \qquad \bar{\mu}_c \overset{\text{a.s.}}{\to} (1-\delta_{\texttt{out}})\,\mu_c.$$

Then one can induce that the limit of mean distance of sample and MCD estimators as follow:

$$\|\bar{\mu}_c - \bar{\mu}_{c'}\|_2 \overset{\text{a.s.}}{\to} (1-\delta_{\texttt{out}})^2\,\|\mu_i - \mu_c\|_2, \tag{5}$$

$$\|\widehat{\mu}_c - \widehat{\mu}_{c'}\|_2 \overset{\text{a.s.}}{\to} \|\mu_c - \mu_{c'}\|_2. \tag{6}$$

[7] P. Billingsley, Convergence of Probability Measures, John Wiley & Sons, 1999

On the other hand, the assumptions $\mathcal{A}1$ states that all class covariance matrices are the same, i.e., $\mathbf{\Sigma}_c = \sigma^2\mathbf{I}$. Then tied covariance matrices $\bar{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Sigma}}$ are given by gathering $\bar{\mathbf{\Sigma}}_c$ and $\widehat{\mathbf{\Sigma}}_c$ on each class c respectively:

$$\bar{\mathbf{\Sigma}} = \frac{\sum_c N_c \bar{\mathbf{\Sigma}}_c}{\sum_c N_c} = \frac{\sum_c \bar{\mathbf{\Sigma}}_c}{C}, \quad \widehat{\mathbf{\Sigma}} = \frac{\sum_c K_c \widehat{\mathbf{\Sigma}}_c}{\sum_c K_c} = \frac{\sum_c \widehat{\mathbf{\Sigma}}_c}{C}. \tag{7}$$

From the tied covariance matrices (7) and Lemma 1, one can induce their convergences and limits as follow:

$$\bar{\mathbf{\Sigma}} \overset{\text{a.s.}}{\to} \left((1-\delta_{\texttt{out}})\,\sigma^2 + \delta_{\texttt{out}}\sigma_{\texttt{out}}^2\right)\mathbf{I} + \delta_{\texttt{out}}\,(1-\delta_{\texttt{out}})\,\frac{1}{C}\sum_c \mu_c\mu_c^T, \tag{8}$$

$$\widehat{\mathbf{\Sigma}} \overset{\text{a.s.}}{\to} \sigma^2\mathbf{I}. \tag{9}$$

Next, we define a function of the tied covariance matrix as follow:

$$\phi(\widehat{\mathbf{\Sigma}}) = 4\|\widehat{\mathbf{\Sigma}}^{-1}\|_2\|\widehat{\mathbf{\Sigma}}\|_2\Big(1 + \|\widehat{\mathbf{\Sigma}}^{-1}\|_2\|\widehat{\mathbf{\Sigma}}\|_2\Big)^{-2}.$$

Then (9) implies that $\phi(\widehat{\mathbf{\Sigma}}) \overset{\text{a.s.}}{\to} 1$ clearly. Since the condition number t of $\widehat{\mathbf{\Sigma}}^{-1}$ is in $[1,\infty)$, i.e. $t = \|\widehat{\mathbf{\Sigma}}^{-1}\|_2\|\widehat{\mathbf{\Sigma}}\|_2 \in [1,\infty)$, one can induces that $\phi(\widehat{\mathbf{\Sigma}}) = \phi(t) = \frac{4t}{(1+t)^2}$ for $t \in [1,\infty)$ and it is a monotonic decreasing function by using the change of variables with t. Hence $\phi(t)$ has the maximum at $t = 1$ and it implies

$$1 = \phi(1) = \lim_{N\to\infty}\phi(\widehat{\mathbf{\Sigma}}) \geq \lim_{N\to\infty}\phi(\bar{\mathbf{\Sigma}}). \tag{10}$$

Therefore the limits of mean distance of estimators (5), (6) and ratio of the function ϕ (10) hold the statement (4),

$$\frac{\phi(\widehat{\mathbf{\Sigma}})\|\widehat{\mu}_c - \widehat{\mu}_{c'}\|_2}{\phi(\bar{\mathbf{\Sigma}})\|\bar{\mu}_c - \bar{\mu}_{c'}\|_2} \overset{\text{a.s.}}{\to} \lim_{N\to\infty}\frac{\phi(\widehat{\mathbf{\Sigma}})\|\widehat{\mu}_c - \widehat{\mu}_{c'}\|_2}{\phi(\bar{\mathbf{\Sigma}})\|\bar{\mu}_c - \bar{\mu}_{c'}\|_2} = \lim_{N\to\infty}\frac{1}{(1-\delta_{\texttt{out}})^2\,\phi(\bar{\mathbf{\Sigma}})} \geq 1.$$

This completes the proof of Theorem 1.

D.3. Proof of Lemma 1

In this part, we present a proof of Lemma 1. We show the almost surely convergences of sample and MCD estimators as the number of training samples N goes to infinity.

Proof of the convergence of sample estimator. First of all, the set of training samples $\mathcal{X}_N = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ contains outlier samples with the fixed fraction $\delta_{\texttt{out}}$. So, $\mathcal{X}_N$ is from a mixture distribution $P_{\texttt{mix}} = (1-\delta_{\texttt{out}})\,P_{\texttt{clean}} + \delta_{\texttt{out}}P_{\texttt{out}}$. Then mean and covWariance matrix of sample estimator, $\bar{\mu}$ and $\bar{\mathbf{\Sigma}}$, estimate mean $\mu_{\texttt{mix}}$ and covariance matrix $\mathbf{\Sigma}_{\texttt{mix}}$ of the mixture distribution $P_{\texttt{mix}}$, respectively. One can induce $\mu_{\texttt{mix}}$ and $\mathbf{\Sigma}_{\texttt{mix}}$ directly as follow:

$$\mu_{\texttt{mix}} = (1-\delta_{\texttt{out}})\,\mu, \quad \mathbf{\Sigma}_{\texttt{mix}} = (1-\delta_{\texttt{out}})\,\sigma^2\mathbf{I} + \delta_{\texttt{out}}\sigma_{\texttt{out}}^2\mathbf{I} + \delta_{\texttt{out}}\,(1-\delta_{\texttt{out}})\,\mu\mu^T. \tag{11}$$

Since $P_{\texttt{mix}}$ has the finite covariance matrix, i.e., $\mathbf{\Sigma}_{\texttt{mix}} < \infty$, one can apply the the Strong Law of Large Numbers[8] to the sample estimator of the mixture distribution $P_{\texttt{mix}}$. Then the mean and covariance matrix of sample estimator converge almost surely to the mean and covariance matrix of $P_{\texttt{mix}}$, respectively:

$$\bar{\mu} \overset{\text{a.s.}}{\to} \mu_{\texttt{mix}}, \qquad \bar{\mathbf{\Sigma}} \overset{\text{a.s.}}{\to} \mathbf{\Sigma}_{\texttt{mix}}.$$

This completes the proof of the convergence of sample estimator.

Proof of the convergence of MCD estimator. Consider a collection $E_{\texttt{q}}$ of subsets $\mathcal{X}_{K,q} \subset \mathcal{X}_N$ with the size $K\,(= \lfloor\delta_{\texttt{mcd}}N\rfloor)$, and each subset $\mathcal{X}_{K,q} \in E_{\texttt{q}}$ contains the outlier samples with the fraction $q \in [0,1]$. Then $\mathcal{X}_{K,q} \in E_{\texttt{q}}$ is from a mixture distribution $P_{\texttt{q}} = (1-q)\,P_{\texttt{clean}} + qP_{\texttt{out}}$. One can induce that the mean $\mu_{\texttt{q}}$ and covariance matrix $\mathbf{\Sigma}_{\texttt{q}}$ of the mixture distribution $P_{\texttt{q}}$ as (11):

$$\mu_{\texttt{q}} = (1-q)\mu, \quad \mathbf{\Sigma}_{\texttt{q}} = (1-q)\sigma^2\mathbf{I} + q\sigma_{\texttt{out}}^2\mathbf{I} + q(1-q)\mu\mu^T. \tag{12}$$

[8] W. Feller, An Introduction to Probability Theory and Its Applications, John Wiley & Sons, 1968

Thus sample mean estimator $\bar{\mu}_{\mathcal{X}_{K,q}}$ and covariance estimator $\bar{\mathbf{\Sigma}}_{\mathcal{X}_{K,q}}$ of a subset $\mathcal{X}_{K,q}$ converge almost surely to $\mu_{\mathtt{q}}$ and $\mathbf{\Sigma}_{\mathtt{q}}$ respectively:

$$\bar{\mu}_{\mathcal{X}_{K,q}} \overset{\text{a.s.}}{\to} \mu_{\mathtt{q}}, \qquad \bar{\mathbf{\Sigma}}_{\mathcal{X}_{K,q}} \overset{\text{a.s.}}{\to} \mathbf{\Sigma}_{\mathtt{q}},$$

by the Strong Law of Large Numbers.

On the other hand, there is a subset $\mathcal{X}^*_{K,q^*} \subset \mathcal{X}_N$ in $E_{\mathtt{q}^*}$ which is selected by MCD estimator. Then the determinant of its covariance matrix is the minimum over all subset of size K in $\mathcal{X}_N$, and $\bar{\mu}_{\mathcal{X}^*_{K,q^*}} = \widehat{\mu} \overset{\text{a.s.}}{\to} \mu_{\mathtt{q}^*}$, $\bar{\mathbf{\Sigma}}_{\mathcal{X}^*_{K,q^*}} = \widehat{\mathbf{\Sigma}} \overset{\text{a.s.}}{\to} \mathbf{\Sigma}_{\mathtt{q}^*}$ as $N \to \infty$. Since the determinant is a continuous funciton, the Continuous Mapping Theorem[9] implies

$$\min_{\mathcal{X}_{K,q} \in E_{\mathtt{q}}, \forall q} \det(\bar{\mathbf{\Sigma}}_{\mathcal{X}_{K,q}}) \overset{\text{a.s.}}{\to} \min_{q} \det\left(\mathbf{\Sigma}_{\mathtt{q}}\right),$$

and

$$\min_{\mathcal{X}_{K,q} \in E_{\mathtt{q}}, \forall q} \det(\bar{\mathbf{\Sigma}}_{\mathcal{X}_{K,q}}) = \det(\bar{\mathbf{\Sigma}}_{\mathcal{X}^*_{K,q^*}}) = \det(\widehat{\mathbf{\Sigma}}) \overset{\text{a.s.}}{\to} \det\left(\mathbf{\Sigma}_{\mathtt{q}^*}\right).$$

Now, we'd like to show

$$\min_{q} \det\left(\mathbf{\Sigma}_{\mathtt{q}}\right) = \det\left(\mathbf{\Sigma}_{\mathtt{q}^*}\right) = \det\left(\mathbf{\Sigma}_{0}\right), \tag{13}$$

to complete the proof of Lemma 1.

By the assumption $\delta_{\mathtt{mcd}} \leq 1 - \delta_{\mathtt{out}}$, E_0 is non-empty. It shows the existence of $\mathbf{\Sigma}_0$. From the covariance matrix $\mathbf{\Sigma}_{\mathtt{q}}$ (12), $\det(\mathbf{\Sigma}_{\mathtt{q}})$ is a d-th degree polynomial of q as follow:

$$\begin{aligned}
\det(\mathbf{\Sigma}_{\mathtt{q}}) &= \det\left((1-q)\sigma^2\mathbf{I} + q\sigma^2_{\mathtt{out}}\mathbf{I} + q(1-q)\mu\mu^T\right) \\
&= \left((1-q)\sigma^2 + q\sigma^2_{\mathtt{out}}\right)^{d-1}\left((1-q)\sigma^2 + q\sigma^2_{\mathtt{out}} + q(1-q)\mu^T\mu\right).
\end{aligned}$$

Since the assumption gives $\sigma^2_{\mathtt{out}} > \sigma^2$, $\det(\mathbf{\Sigma}_{\mathtt{q}})$ has the lower bound $\det(\mathbf{\Sigma}_0)$ as follow:

$$\begin{aligned}
\det(\mathbf{\Sigma}_{\mathtt{q}}) &= \left((1-q)\sigma^2 + q\sigma^2_{\mathtt{out}}\right)^{d-1}\left((1-q)\sigma^2 + q\sigma^2_{\mathtt{out}} + q(1-q)\mu^T\mu\right) \\
&\geq \left(\sigma^2\right)^{d-1}\left(\sigma^2 + q(1-q)\mu^T\mu\right) \\
&\geq \left(\sigma^2\right)^{d-1}\sigma^2 = \det(\mathbf{\Sigma}_0).
\end{aligned}$$

Then $\det(\mathbf{\Sigma}_{\mathtt{q}}) \geq \det(\mathbf{\Sigma}_0)$ for all $q \in [0,1]$ and the equality holds for only $q = 0$. It implies $q^* = 0$ and (13) is the shown. Therefore the mean and covariance matrix of MCD estimator converge almost surely to μ and $\sigma^2\mathbf{I}$, respectively:

$$\widehat{\mu} \overset{\text{a.s.}}{\to} \mu_0 = \mu, \quad \widehat{\mathbf{\Sigma}} \overset{\text{a.s.}}{\to} \mathbf{\Sigma}_0 = \sigma^2\mathbf{I}.$$

This completes the proof of Lemma 1.

E. Comparison with robust clustering methods

To remove the outliers (i.e., samples with noisy labels) in hidden feature, one can consider other robust clustering methods to estimate the parameters of generative classifiers. In this section, we test trimmed K-means (TKM) (Garcia-Escudero & Gordaliza, 1999) as a new baseline, and compare the performances with MCD estimator. Specifically, we use noisy labels to initialize clusters of TKM, and assign a "majority" label to each trained cluster. For pair comparison, a tied covariance across clusters is assumed, and we run the same number of iterations for both TKM and MCD. Table 8 reports the corresponding results under ResNet and CIFAR-10 with uniform noises when we induce a generative classifier only using a penultimate layer. First, we remark that TKM outperforms Softmax, which supports our claim that the clustering property of DNN features is useful to handle noisy labels. However, the generative classifier with MCD estimator still outperforms all baselines because it utilizes the information of noisy labels carefully to derive a new decision rule at all iterations, while TKM is essentially an unsupervised method and does not utilize the information, except for initialization and termination.

[9] P. Billingsley, Convergence of Probability Measures, John Wiley & Sons, 1999

Model	Inference method	CIFAR-10			
		Clean	Uniform (20%)	Uniform (40%)	Uniform (60%)
ResNet	Softmax	94.76%	80.88%	61.98%	39.96%
	Generative + TKM	94.35%	81.41%	63.27%	41.78%
	Generative + MCD (ours)	94.76%	**83.86%**	**68.03%**	**44.87%**

Table 8. Test accuracy (%) of ResNet on the CIFAR-10 dataset with uniform noise. The best results are highlighted in bold if the gain is bigger than 1%.