# A. Related literature

**Statistical learning methods.** Statistical learning contributed a lot to the problem of learning with noisy labels, especially in theoretical aspects. Statistical learning approaches can be categorized into three strands: surrogate loss, noise rate estimation and probabilistic modeling. For example, in the surrogate losses category, Natarajan et al. (2013) proposed an unbiased estimator to provide the noise corrected loss approach. Masnadi-Shirazi & Vasconcelos (2009) presented a robust non-convex loss, which is the special case in a family of robust losses. In the noise rate estimation category, both Menon et al. (2015) and Liu & Tao (2016) proposed a class-probability estimator using order statistics on the range of scores. Sanderson & Scott (2014) presented the same estimator using the slope of the ROC curve. In the probabilistic modeling category, Raykar et al. (2010) proposed a two-coin model to handle noisy labels from multiple annotators. Yan et al. (2014) extended this two-coin model by setting the dynamic flipping probability associated with instances.

**Deep learning approaches.** Deep learning approaches are prevalent to handle noisy labels (Zhang & Sabuncu, 2018). Li et al. (2017) proposed a unified framework to distill the knowledge from clean labels and knowledge graph, which can be exploited to learn a better model from noisy labels. Veit et al. (2017) trained a label cleaning network by a small set of clean labels, and used this network to reduce the noise in large-scale noisy labels. Rodrigues & Pereira (2018) added a crowd layer after the output layer for noisy labels from multiple annotators. Tanaka et al. (2018) presented a joint optimization framework to learn parameters and estimate true labels simultaneously. Ren et al. (2018) leveraged an additional validation set to adaptively assign weights to training examples. Similarly, based on a small set of trusted data with clean labels, Hendrycks et al. (2018) proposed a loss correction approach to mitigate the effects of label noise on deep neural network classifiers. Ma et al. (2018) developed a new dimensionality-driven learning strategy, which monitors the dimensionality of deep representation subspaces during training and adapts the loss function accordingly. Wang et al. (2018) proposed an iterative learning framework for training CNNs on datasets with open-set noisy labels. Han et al. (2018a) proposed a human-assisted approach that conveys human cognition of invalid class transitions, and derived a structure-aware deep probabilistic model incorporating a speculated structure prior. Lee et al. (2019) proposed a novel inference method to obtain a robust decision boundary under any softmax neural classifier pre-trained on noisy datasets. Their idea is to induce a generative classifier on top of hidden feature spaces of the discriminative deep model.

# B. Training details

For *MNIST* and *NEWS*, we train Co-teaching+ by default at the beginning of training. For other datasets, we use a warm-up strategy to achieve a higher test accuracy. Specifically, for *CIFAR-10*, we warm-up Co-teaching+ with training Co-teaching for the first 20 epochs (i.e., only conducting cross-update for the first 20 epochs). For *CIFAR-100*, we warm-up Co-teaching+ with training Co-teaching for the first 5 epochs. For *T-ImageNet*, we start disagreement-update in the middle of training, i.e., we warm-up Co-teaching+ with training Co-teaching for the first 100 epochs. For *Open-sets*, we warm-up Co-teaching+ with training two networks in parallel for the first 55 epochs, where both networks leverage the small-loss trick. Inevitably, there is few chance that we cannot find enough small-loss instances for cross-update. In that case, we only conduct disagreement-update in a mini-batch data during training.