A. Related works

Besides the works aforementioned, we survey other approaches to learning with noisy labels.

Robust losses. Various approaches propose to use a provably robust loss function in the learning process. In the case of class-dependent label noise, (Natarajan et al., 2013) constructed an unbiased estimator of any loss function under the noisy distribution. (Masnadi-shirazi & Vasconcelos, 2009) introduced a robust non-convex loss. Recently, works on symmetric losses showed that such loss offer theoretical robustness results to various types of noise (Ghosh et al., 2017; Charoenphakdee et al., 2019). Motivated by the robustness to noise of the mean absolute error loss (MAE) shown in (Ghosh et al., 2017), (Zhang & Sabuncu, 2018) introduced generalized cross entropy loss that allows for a trade-off between the efficient learning properties of the CCE loss and the noise-robustness of MAE. (Shen & Sanghavi, 2019) introduced a trimmed loss with an iterative minimization process that allows for theoretical guarantees in the simpler setting of generalized linear models.

Annotator-level modelling. Another recent line of related works attempts to model labels and worker's quality directly during the crowdsourcing annotation process, in order to produce more accurate labels efficiently. (Branson et al., 2017) modeled the annotators' skill and instances difficulty while incrementally training a computer vision model during the annotation process, effectively reducing the time burden of the annotation process as well as the error rate in the assigned labels. (Guan et al., 2018) modeled each annotator individually in order to better aggregate labels based on each worker's skill and area of expertise. (Khetan et al., 2018) introduced a method that allows to learn each workers' skill even when each example is only annotated once, by jointly modelling the assigned labels and the workers during the annotation process.

Learning with multiple noisy labels. A closely related setting is learning from multiple noisy labels, where the aim is to predict an unknown ground-truth label from $(X, (Y^j)_j)$, each Y^j referring to a noisy annotation. This setting can arise for example from crowdsourcing tasks; (Snow et al., 2008) showed that using multiple non-expert annotators to train a classifier can be as effective as using gold standard annotations from experts. In (Raykar et al., 2009), the authors derive a Bayesian approach to jointly learn the expertise of each annotator, the actual true label and the classifier. (Yan et al., 2010) extends this Bayesian approach by considering that each annotator's expertise varies across the input space. This setting differs from ours as it takes place before the aggregation of multiple annotations, which, for CSIDN, is only a way among others to obtain a confidence score for each noisy label.

Explicit/implicit regularizers. Recently, several other regularization techniques have shown good robustness in weakly-supervised settings. Temporal Ensembling (TE) (Laine & Aila, 2017) method labels some additional unlabeled instances using a consensus of predictions from models from previous epochs and with different regularizations and input augmentation conditions. Mean-teacher (MT) (Tarvainen & Valpola, 2017) instead uses predictions from a model obtained by averaging the weights of a set of models similar to TE, as using the prediction from a unique model is more efficient when a large amount of unlabeled data is available. Virtual Adversarial Training (Miyato et al., 2018) regularizes the network using a measure of local smoothness of the conditional label distribution given the input, defined as the robustness of the prediction to local adversarial perturbations in the input space. Introduced in (Zhang et al., 2018), *mixup* trains a neural network on convex combinations of instance pairs and their respective labels, and has been shown to reduce the memorization of corrupted labels.

Weak supervision. Recent approaches in weakly supervised learning provide alternatives to noisy label learning. Instead of considering a single imperfect labeller on the whole dataset, Data Programming (Ratner et al., 2016; 2020) considers a set of labelling functions providing approximate labels on subsets of the dataset and aggregates them by estimating their respective noise rates and modelling their dependencies. Meanwhile, Adversarial Data Learning (Arachie & Huang, 2019) considers a set of weak labellers providing soft labels of the data along with estimated error bounds, and trains a model minimizing the error rate on labels selected by an adversarial agent.

B. Synthetic dataset

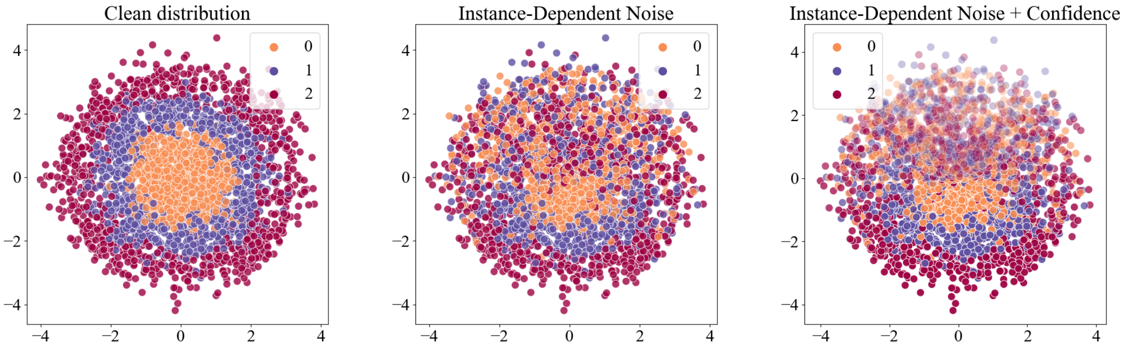
Figure 7 shows three synthetic datasets, which cover clean, IDN and CSIDN models.

C. Baselines

Here we detail the four baselines used in our experiments.

Forward correction. Introduced in (Patrini et al., 2017), *forward correction* estimates a fixed transition matrix T before training, and trains a classifier with the corrected loss $l_T : (y, \hat{y}) \mapsto l(y, T\hat{y})$.

Mean absolute error loss. Due to its symmetric property, the Mean Absolute Error (MAE) has been theoretically justified to be robust to label noise under assumptions (Ghosh et al., 2017). However, this loss is more difficult to train, especially on complex datasets.



(a) Clean data distribution.

(b) Data distribution with IDN.

(c) Data distribution with CSIDN.

Figure 7. Synthetic dataset. The clean distribution (a) consists in three classes of concentric circles. In the IDN setting (b), each point x has a probability $P(\bar{Y} \neq Y|x) = \rho\left(\frac{w \cdot x}{\|w\|\|x\|} + 1\right)/2$ with $w = (0, 1)$ of being corrupted, where ρ is a parameter controlling the mean noise rate. Therefore the noise is the strongest towards the direction $(0, 1)$ and the weakest in the direction $(0, -1)$. If corrupted, the label is flipped to another class uniformly. The CSIDN setting (c) is similar to the IDN setting, but each point is associated with measure of the confidence in the assigned label. A lower confidence is represented by a lower opacity in the figure.

L_q norm. Introduced in (Zhang & Sabuncu, 2018), L_q norm or Generalized Cross Entropy (GCE) Loss attempts to bring the best of both worlds between the CCE and the MAE loss: the CCE is easy to train, while the MAE is robust to label noise. The authors therefore define this loss using the negative box-cox transformation:

$$L_q\left(h(\boldsymbol{x}), \boldsymbol{e}_j\right) = \frac{(1 - h_j(\boldsymbol{x})^q)}{q},$$

so that the L_q tends to the CCE when $q \to 0$ and to the MAE when $q = 1$. In the following experiments, we set $q = 0.7$, suggested by authors.

Co-teaching (Han et al., 2018b). Co-teaching algorithm is a small-loss approach where two classifiers are trained in parallel. At each epoch, each classifier selects the instances with the smallest loss, and feed them to the other network as a training set for the next iteration. This work has proved to be a leading benchmark in the field of noisy labels.

D. Examples of real-world datasets

An example application of this work in building large real-world datasets with limited resources is constructing a dataset with images scraped from the web, and automatically labelling them from neighbouring text fields using a classifier such as a recurrent neural network. Then, a small subset of curated images can be used at the beginning of the process to calibrate the classifier, in order to make the predictions of the softmax output faithful to the confidence in each label. This way, one can build a very large dataset for a very low-cost that, while involving some instance-dependent noise, would be equipped with confidence information and therefore could be tackled with our proposed algorithm.