

Label noise: Problems and Solutions

.....

Sandhya Tripathi & N. Hemachandra

Tutorial @IEEE DSAA'2020

October 8, 2020

Email: sandhyat@wustl.edu, nh@iitb.ac.in



Outline

1. Introduction to label noise problems
2. Learning in presence of label noise via Empirical Risk Minimization
3. Learning in the presence of label noise via deep neural networks
4. Learning is more than 0-1 loss based binary classification
5. Experiment protocols and other logistics while designing label noise experiments
6. What next from here?

Introduction to label noise problems

Binary Classification Problem



Image source: Internet

Binary Classification in the presence of label noise

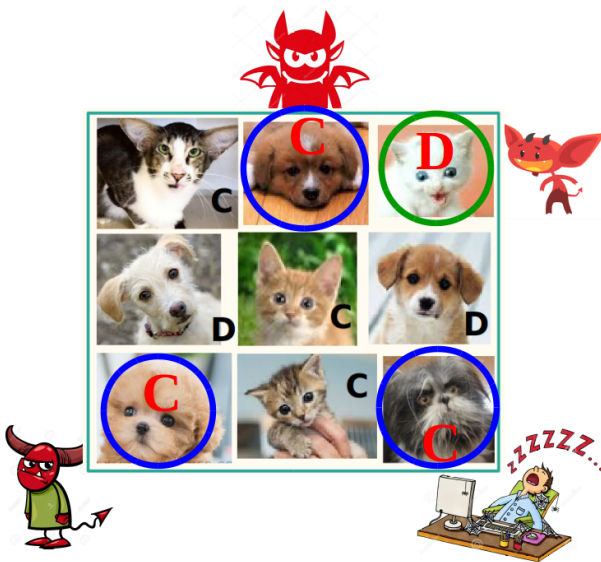
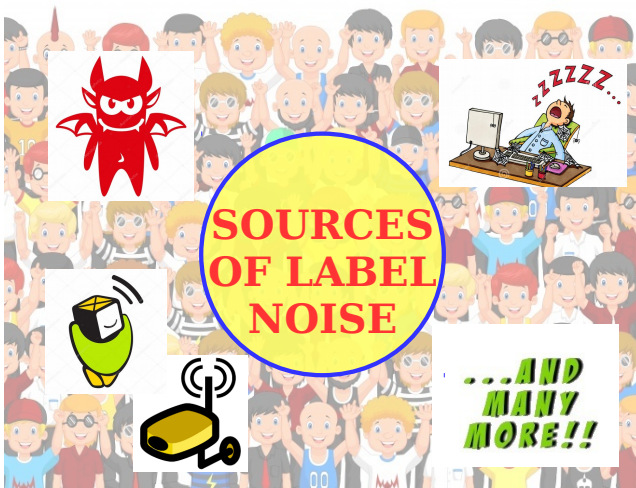


Image source: Internet

Motivation: Noisy labels in real life



Adversarial noise, Crowd sourcing, Sensor noise, Labelling is expensive especially for x-ray images etc.

Learning a classifier: ERM terminology

- \mathcal{D} : Joint distribution over $\mathbf{X} \times Y$ with $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^n$ and $Y \in \mathcal{Y} = \{-1, 1\}$
- $\eta(\mathbf{x}) := P(Y = 1|\mathbf{x})$, $\pi := P(Y = 1)$
- Decision function $f : \mathbf{X} \mapsto \mathbb{R}$
- \mathcal{H} : Hypothesis class of all measurable functions
- \mathcal{H}_{lin} : Class of linear hypothesis
- **Ideal situation**: Minimize $R_{\mathcal{D}}(f) := E_{\mathcal{D}}[\ell_{0-1}(f(\mathbf{x}), y)] = E_{\mathcal{D}}[\mathbf{1}_{\{y \neq \text{sign}(f(\mathbf{x}))\}}]$.
- **Solution** is Bayes classifier's prediction, i.e., $\text{sign}(2\eta(\mathbf{x}) - 1)$
- Training sample $S := \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \sim \mathcal{D}^m$
- Empirical risk minimization (ERM): $\hat{f}_l^* \in \arg \min_{f \in \mathcal{H}} \hat{R}_{\mathcal{D}, l}(f) := \frac{1}{m} \sum_{i=1}^m \ell(f(\mathbf{x}_i), y_i)$
- Solving ideal loss function (ℓ_{0-1}) based ERM is NP-hard.
- Use surrogate loss function ℓ_{sur} , a convex upper bound on ℓ_{0-1} .

Label noise: formal setup

- $\tilde{\mathcal{D}}$: Joint distribution on $\mathbf{X} \times \tilde{Y}$ obtained by inducing noise to \mathcal{D} , $\tilde{Y} \in \{-1, 1\}$
- $\tilde{S} = \{(\mathbf{x}_1, \tilde{y}_1), \dots, (\mathbf{x}_m, \tilde{y}_m)\} \sim \tilde{\mathcal{D}}^m$ (*corrupted data*)
- Noise rate $\rho_y(\mathbf{X}) := P(\tilde{Y} = -y | Y = y, \mathbf{X} = \mathbf{x})$
- $\tilde{\eta}(\mathbf{x}) = P(\tilde{Y} = 1 | Y = 1, \mathbf{X} = \mathbf{x})\eta(\mathbf{x}) + P(\tilde{Y} = 1 | Y = -1, \mathbf{X} = \mathbf{x})(1 - \eta(\mathbf{x}))$
- Noisy true ℓ -risk of classifier f be $R_{\tilde{\mathcal{D}}, \ell}(f)$ and its minimizer be \tilde{f}_ℓ^* .
- Empirical noisy ℓ -risk is $\hat{R}_{\tilde{\mathcal{D}}, \ell}(f) := \frac{1}{m} \sum_{i=1}^m l(f(\mathbf{x}_i), \tilde{y}_i)$.

Definition (Noise tolerance under risk minimization)

[Manwani and Sastry, 2013] For a loss function l and classifiers \tilde{f}_ℓ^* and f_ℓ^* , risk minimization is said to be noise tolerant if

$$R_{\mathcal{D}, 0-1}(\tilde{f}_\ell^*) = R_{\mathcal{D}, 0-1}(f_\ell^*). \quad (1)$$

Noise models

Symmetric label noise (SLN) model

- The flipping probability is a constant $\rho = P(\tilde{Y} = -y|Y = y)$.
- Corrupted in-class probability $\tilde{\eta}(\mathbf{x}) = (1 - 2\rho)\eta(\mathbf{x}) + \rho$.
- Corrupted class marginal $\tilde{\pi} = (1 - 2\rho)\pi + \rho$.

Class conditional noise (CCN) model

- The flipping probability is class dependent, i.e.,
 $\rho_+ = (Y = -1|Y = 1) \ \& \ \rho_- = (Y = 1|Y = -1)$.
- Corrupted in-class probability $\tilde{\eta}(\mathbf{x}) = (1 - \rho_+ - \rho_-)\eta(\mathbf{x}) + \rho_-$.
- Corrupted class marginal $\tilde{\pi} = (1 - \rho_+ - \rho_-)\pi + \rho_-$.

Instance dependent noise (IDN)

- Label-and instance-dependent (LIN): noise rate $\rho_y(\mathbf{x})$
- Purely instance-dependent noise (PIN): noise rate $\rho_+(\mathbf{x}) = \rho_-(\mathbf{x}) = \rho(\mathbf{x})$.

Goal: Given a sample \tilde{S} from noisy or corrupted distribution $\tilde{\mathcal{D}}$, obtain a classifier, which is trained on $\tilde{S} \sim \tilde{\mathcal{D}}^m$ but evaluated on data from \mathcal{D} .

Various models

Performance criteria: (for *designed* classifier using test set):

- 01-loss (or Accuracy), using $\ell_{01}(yf(\mathbf{x})) (= \ell_{01}(f))$ loss function.
- cost-sensitive criteria, $\ell_{\alpha}(f)$, $\alpha \in [0, 1]$ (i.e., α -weighted 01 loss fn.)
- Cross-entropy criteria
- Etc.

Classification nature:

- Binary class

Various types of noise:

- Symmetric Label Noise Model, SLN (also called uniform noise model); ρ
 - Class Conditional Noise Model, (CCN); ρ_+ and ρ_-
 - Instance dependent noise model; $\rho(\mathbf{x})$
- Multi-class noise model:
- Etc.

Methods for learning in the presence of label noise

[Sastry and Manwani, 2017]

- Noise cleaning: correct labels are restored
- Eliminating noisy points: after identifying the noisy points they are eliminated
- Designing schemes for dealing with label noise: goal is to minimize the effect of label noise
- Noise tolerant algorithms: designing algorithms that are unaffected by the label noise (identifying noise robust loss functions ℓ for minimization.)

Learning in presence of label noise via Empirical Risk Minimization

The resurrection of the label noise [Long and Servedio, 2010]

- Convex potential function: $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is convex, non-increasing, differentiable with $\phi'(0) < 0$.
- Boosting algorithms that minimize a convex potential function are susceptible to SLN. AdaBoost minimizes $\phi(z) = e^{-z}$, a convex potential.

The resurrection of the label noise [Long and Servedio, 2010]

- Convex potential function: $\phi : \mathbb{R} \mapsto \mathbb{R}$ is convex, non-increasing, differentiable with $\phi'(0) < 0$.
- Boosting algorithms that minimize a convex potential function are susceptible to SLN. AdaBoost minimizes $\phi(z) = e^{-z}$, a convex potential.

Which loss functions are not robust against SLN? (via counter examples)

1. Squared loss with $f \notin \mathcal{H}_{lin}$,
2. Hinge loss $\ell_{hinge}(f(\mathbf{x}), y) = \max(0, 1 - yf(\mathbf{x}))$,
3. Exponential loss $\ell_{exp}(f(\mathbf{x}), y) = \exp(-yf(\mathbf{x}))$
4. Log loss $\ell_{log-loss}(f(\mathbf{x}), y) = \ln(1 + \exp(-yf(\mathbf{x})))$

Which loss functions are robust to symmetric label noise (SLN)?

[Manwani and Sastry, 2013]

1. 0-1 loss $\ell_{0-1}(f(\mathbf{x}), y)$ with $f \in \mathcal{H}$
2. Squared loss $\ell_{sq}(f(\mathbf{x}), y) = (y - f(\mathbf{x}))^2$ when $f \in \mathcal{H}_{lin}$

Is there some magic that makes loss functions robust?

What property of 0-1 loss made it noise robust?

Sufficient condition for SLN robustness (Theorem 1, [Ghosh et al., 2015]): If

$\ell_1(f(\mathbf{x})) = \ell(f(\mathbf{x}), 1)$ and $\ell_{-1}(f(\mathbf{x})) = \ell(f(\mathbf{x}), -1)$ such that

$\ell_1(f(\mathbf{x})) + \ell_{-1}(f(\mathbf{x})) = K$, $K > 0$, then $\ell(f(\mathbf{x}), y)$ is SLN robust.

Any other loss function satisfying the sufficient condition?

- Novel unhinged loss [Van Rooyen et al., 2015] $\ell_{unhinged}(f(\mathbf{x}), y) = 1 - yf(\mathbf{x})$
- Sigmoid loss, $\ell_{sg}(\mathbf{x}) = \frac{1}{1 + \exp(\beta f(\mathbf{x})y)}$
- Ramp loss, $\ell_{ra}(\mathbf{x}) = (1 - \beta f(\mathbf{x})y)^+ - (-1 - \beta)f(\mathbf{x})y^+$
- Probit loss, $\ell_{pr}(\mathbf{x}) = 1 - \Phi(\beta f(\mathbf{x})y)$

Doesn't need any knowledge of the noise rates!

Is there some magic that makes loss functions robust?

What property of 0-1 loss made it noise robust?

Sufficient condition for SLN robustness (Theorem 1, [Ghosh et al., 2015]): If

$\ell_1(f(\mathbf{x})) = \ell(f(\mathbf{x}), 1)$ and $\ell_{-1}(f(\mathbf{x})) = \ell(f(\mathbf{x}), -1)$ such that

$\ell_1(f(\mathbf{x})) + \ell_{-1}(f(\mathbf{x})) = K$, $K > 0$, then $\ell(f(\mathbf{x}), y)$ is SLN robust.

Any other loss function satisfying the sufficient condition?

- Novel unhinged loss [Van Rooyen et al., 2015] $\ell_{unhinged}(f(\mathbf{x}), y) = 1 - yf(\mathbf{x})$
- Sigmoid loss, $\ell_{sg}(\mathbf{x}) = \frac{1}{1 + \exp(\beta f(\mathbf{x})y)}$
- Ramp loss, $\ell_{ra}(\mathbf{x}) = (1 - \beta f(\mathbf{x})y)^+ - (-1 - \beta)f(\mathbf{x})y^+$
- Probit loss, $\ell_{pr}(\mathbf{x}) = 1 - \Phi(\beta f(\mathbf{x})y)$

Doesn't need any knowledge of the noise rates!

But what happens when the noise is class conditional?

Class Conditional Noise (CCN) model, Performance assessment, etc

Recall, the CCN setup:

The flipping probability is class dependent, i.e.,

$$\rho_+ = (\tilde{Y} = -1|Y = 1) \ \& \ \rho_- = (\tilde{Y} = 1|Y = -1)$$

The SLN model is a special case of this.

More reasonable assumption

One of the first papers is by Nagarajan Natarajan et al

[\[Natarajan et al., 2013\]](#), [\[Natarajan et al., 2018\]](#)

Also, give performance guarantees

In terms of high probability bounds on the risk of the learned classifier (from noisy data) in terms of ‘best possible’ risk and some error terms.

Modifying the loss functions for noise robustness; Unbiased loss estimator [Natarajan et al., 2013]

Given a loss fn. $\ell(\cdot, \cdot)$, one defines a α -weighted version of it:

$$\ell_{\alpha}(t, y) := ((1 - \alpha)\mathbb{1}_{\{y=1\}} + \alpha\mathbb{1}_{\{y=-1\}})\ell(t, y)$$

Also, if $\ell(\cdot, \cdot)$ is bounded, one defines

$$\tilde{\ell}(t, y) := \frac{(1 - \rho_{-y})\ell(t, y) - \rho_y\ell(t, -y)}{1 - \rho_+ - \rho_{-1}}$$

Then,

$$E_{\tilde{y}}[\tilde{\ell}(t, \tilde{y})] = \ell(t, y)$$

That is, $\tilde{\ell}(t, y)$ is an unbiased loss estimator of loss $\ell(\cdot, \cdot)$

Unbiased loss estimator, cont.

In particular, we have,

$$E_{\tilde{y}}[\tilde{\ell}_{\alpha}(t, \tilde{y})] = \ell_{\alpha}(t, y)$$

with $\ell_{\alpha}(\cdot, \cdot)$ is the α -weighted version of bounded $\ell(\cdot, \cdot)$

Now, consider the following ERM scheme:

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_{\tilde{\ell}_{\alpha}}(f) := \sum_{i=1}^m \tilde{\ell}_{\alpha}(f(X_i, \tilde{Y}_i))$$

For a given classifier f , the above sample average converges to $R_{\mathcal{D}}(\ell_{\alpha}(f))$ even when it is computed in CCN set up.

Why?

Unbiased loss estimator, cont.

In particular, we have,

$$E_{\tilde{y}}[\tilde{\ell}_{\alpha}(t, \tilde{y})] = \ell_{\alpha}(t, y)$$

with $\ell_{\alpha}(\cdot, \cdot)$ is the α -weighted version of bounded $\ell(\cdot, \cdot)$

Now, consider the following ERM scheme:

$$\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_{\tilde{\ell}_{\alpha}}(f) := \sum_{i=1}^m \tilde{\ell}_{\alpha}(f(X_i, \tilde{Y}_i))$$

For a given classifier f , the above sample average converges to $R_{\mathcal{D}}(\ell_{\alpha}(f))$ even when it is computed in CCN set up.

Why?

Unbiasedness of $\tilde{\ell}_{\alpha}$

CCN model, Unbiased loss estimator

Recall, that definition of unbiased loss estimator is in terms of CCN noise rates, ρ_+ and ρ_-

Noise rates are usually unknown

The last step in this ERM scheme is to *cross-validate* (CV) the above for various over a set/grid of CCN rates

Performance evaluation:

$$P[R_{\mathcal{D}}(\ell_{\alpha}(\hat{f})) \leq \min_{f \in \mathcal{F}} + 4L_{\rho}RC(\mathcal{F}) + 2\sqrt{\frac{\log(\frac{1}{\delta})}{2m}}] \geq 1 - \delta$$

Here, L_{ρ} is Lipschitz constant of ℓ and $RC(\mathcal{F})$ is Rademacher complexity of class of \mathcal{F}

CCN model, other performance measures

[Natarajan et al., 2018]

Noise robust classifiers are designed to optimize a given linear combination of the confusion matrix, called

Utility of margin based classifier $f(\cdot)$

$$U_{\mathcal{D}}(f) = a_{11}TP_{\mathcal{D}}(f) + a_{10}FP_{\mathcal{D}}(f) + a_{01}FN_{\mathcal{D}}(f) + a_{00}TN_{\mathcal{D}}(f)$$

for some given $\{a_{ij}\}$, $i, j \in \{0, 1\}$

These include AM, the arithmetic mean, apart from accuracy.

$AM(f(\cdot))$ is via $a_{11} = \frac{1}{2(1-\pi)}$, and $a_{00} = \frac{1}{2\pi}$, where $\pi = P(Y = 1)$ under \mathcal{D}

The Bayes classifier is a suitable threshold of the in-class probability $\eta(\mathbf{x})$ — for AM, it is π

A high probability bound on the *regret* of the algo – deviation from the best possible utility – is provided.

Modifying the loss fns. for noise robustness

[Patrini et al., 2016]

The (empirical) mean operator of a learning sample S is defined as

$$\mu_S := E_S[y\mathbf{x}]$$

A loss ℓ is a-linear-odd (a-LOL) when $\ell_o(x) = (\ell(x) - \ell(-x))/2 = ax$ for any $a \in \mathcal{R}$

Mean operator μ is a sufficient statistic among \mathcal{H}_{lin} and a -LOL loss $\ell(\cdot)$

Modifying the loss fns. for noise robustness

[Patrini et al., 2016]

Factorization theorem For a linear classifier $f \in \mathcal{H}_{lin}$

$$R_S(\ell(f)) = \frac{1}{2} R_{S_{2x}} + \ell_o(f(\mu_S))$$

where S_{2x} is ‘doubled-sample’ with features \mathbf{x} repeated with both labels and hence is label free.

This yields better high probability performance bound

Also, this bound has a term with a multiplier that *does not* blow with high noise rates ρ_+ and ρ_- as in Nagarajan Natarajan et al.

The multiplier is 0.6 irrespective of the noise rates.

Modifying the loss functions to account for noise robustness

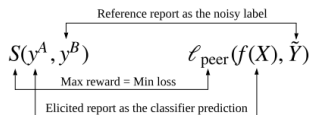
[Liu and Tao, 2016]

Importance weighted sampling

- Rely on the idea of rewriting the true clean risk $R_{\mathcal{D},\ell}(f)$ as $R_{\tilde{\mathcal{D}},\beta\ell}$ where $\beta = \frac{P_{\mathcal{D}}(\mathbf{X},Y)}{P_{\tilde{\mathcal{D}}}(\mathbf{X},\tilde{Y})}$
- As the noise is only in labels $\beta = \frac{P_{\mathcal{D}}(\mathbf{X},Y)}{P_{\tilde{\mathcal{D}}}(\mathbf{X},\tilde{Y})} = \frac{P_{\mathcal{D}}(Y|\mathbf{X})}{P_{\tilde{\mathcal{D}}}(\tilde{Y}|\mathbf{X})}$ which in turn becomes a function only of noise rates and corrupted in class probability $\tilde{\eta}$.
- Present three methods for $\tilde{\eta}$ estimation with the corresponding consistency analysis.
 - Probabilistic classifier approach
 - Kernel density estimation
 - Density ratio estimation method
- If perfect examples exist then $\hat{\rho}_{-\tilde{Y}} = \min_{\mathbf{x} \in \{\mathbf{x}_1, \dots, \mathbf{x}_m\}} \hat{\tilde{\eta}}(\mathbf{x})$

Peer loss functions [Liu and Guo, 2019]

- Inspired by peer prediction framework that studies how to elicit information from self-interested agents without verification.
- Reward function $S(y^A, y^B)$ where $y^{A/B}$ is the noisy observation of the ground truth by agent A/B.



- Correlation agreement (CA) builds Δ matrix that captures the stochastic correlation between y^A and y^B
- For $\rho_{+1} + \rho_{-1} < 1$, $\text{sign}(\Delta) = \mathbb{I}_{2 \times 2}$
- Peer samples: Randomly draw two distinct noisy samples say m_1 and m_2
- $\ell_{\text{peer}}(f(\mathbf{x}_m), \tilde{y}_m) := \ell(f(\mathbf{x}_m), \tilde{y}_m) - \ell(f(\mathbf{x}_{m_1}), \tilde{y}_{m_2})$
- If ℓ here is indicator function and $\pi = 0.5$, then the minimizer of corrupted true peer loss risk is equal to clean Bayes classifier.

Inherently class conditional noise robust!

Learning in the presence of label noise via deep neural networks

Multi-class noise robust loss functions

- [Ghosh et al., 2017] provided a sufficient condition for robustness of multi-class losses under uniform or symmetric noise
- $\sum_{i=1}^k \ell(f(\mathbf{x}), i) = C, \forall \mathbf{x}, \forall f.$
- $\ell_{MAE}(\hat{\mathbf{p}}, \mathbf{e}_j) = \|\mathbf{e}_j - \hat{\mathbf{p}}\|_1$ is robust
- Categorical cross entropy and Mean squared error do not satisfy it; MSE being bounded enjoys more robustness properties
- Issue with MAE : Difficult optimization as gradients saturate quickly.

Multi-class noise robust loss functions

- [Ghosh et al., 2017] provided a sufficient condition for robustness of multi-class losses under uniform or symmetric noise
- $\sum_{i=1}^k \ell(f(\mathbf{x}), i) = C, \forall \mathbf{x}, \forall f.$
- $\ell_{MAE}(\hat{\mathbf{p}}, \mathbf{e}_j) = \|\mathbf{e}_j - \hat{\mathbf{p}}\|_1$ is robust
- Categorical cross entropy and Mean squared error do not satisfy it; MSE being bounded enjoys more robustness properties
- Issue with MAE : Difficult optimization as gradients saturate quickly.

Even though CCE is not noise robust, a lot of research has been going on in making modifications that are either empirically or theoretically proven to be noise robust

Modified CCE losses

- Robust log loss [Kumar and Sastry, 2018]:
 - Design a loss function that has good properties of CCE like easy optimization and satisfies symmetry condition

$$\ell_{rll}(\hat{\mathbf{p}}, \mathbf{e}_j) = \log \frac{\alpha+1}{\alpha} - \log(\alpha + \hat{\mathbf{p}}) + \sum_{t=1, t \neq j} \frac{1}{k-1} \log(\alpha + \hat{\mathbf{p}}_t)$$

- Generalized cross entropy GCE [Zhang and Sabuncu, 2018]:
 - Inspired from the fact that CCE puts more emphasis on training difficult samples; overfitting problem in the presence of label noise

$$L_q(\hat{\mathbf{p}}, \mathbf{e}_j) = \frac{(1 - \hat{\mathbf{p}}_j^q)}{q}, \quad q \in [0, 1)$$

- $q \longrightarrow 0$ leads to CCE loss, $q \longrightarrow 1$ leads to MAE

Modified CE losses continued

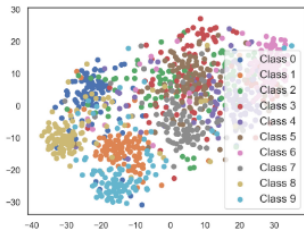
Symmetric cross entropy [Wang et al., 2019]:

- CCE over-fits on some easy losses but under-fits on some difficult classes;
Combine CCE and reverse CCE

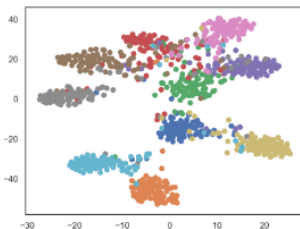
- Reverse CCE has been shown to be robust to uniform noise

$$\ell_{\text{symCCE}}(\hat{\mathbf{p}}, \mathbf{e}_j) = -\alpha \sum_{i=1}^k \mathbf{e}_{ji} \log(\hat{\mathbf{p}}_i) - \beta \sum_{i=1}^k \hat{\mathbf{p}}_i \log(\mathbf{e}_{ji})$$

- α takes care of over-fitting part of CCE and β provides flexibility in robustness
- Representation from second last layer projected to 2-D when noise is 60%



(a) CE



(b) SL

Multi-class noise robust loss functions

L_{DMI} [Xu et al., 2019]

- Relies on a generalized version of mutual information called Determinant based Mutual Information (DMI)
- L_{DMI} is claimed to be theoretically robust to any noise type and any noise level
- If $U := \frac{1}{N} \mathbf{O} \mathbf{L}$ then $L_{DMI} = -\log(\| \det(U) \|)$

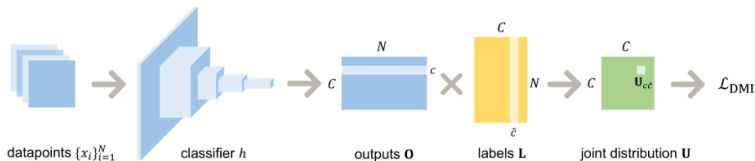


Figure 1: The computation of L_{DMI} in each step of iteration

Multi-class noise robust loss functions

- Active passive losses [Ma et al.,]: Can make any loss function robust by normalization but empirical performance is bad; propose active passive loss functions.
- Bi-tempered logistic loss [Amid et al., 2019]: Propose a theoretically sound bi-tempered loss function that is a non-convex generalization of logistic loss and requires tuning two temperature parameters. Robust to SLN only.
- Trimmed CCE [Rusiecki, 2019]: Decides a parameter h as threshold that decides the examples to include based on their CCE.
- Taylor CCE [Feng et al., 2020]: Uses Taylor expansion of the cross entropy loss and control the terms to include in the loss function.

Loss correction approach [Patrini et al., 2017]

Multi-class classification : $\mathbf{y} \in \{0, 1\}^c$ with c classes; network's prediction $\hat{p}(\mathbf{y}|\mathbf{x})$

Let $T \in [0, 1]^{c \times c}$ be the noise transition matrix s.t. $T_{ij} = p(\tilde{\mathbf{y}} = \mathbf{e}^i | \mathbf{y} = \mathbf{e}^j)$ and 'known' row stochastic

- Backward correction procedure: $\ell^{\leftarrow}(\hat{p}(\mathbf{y}|\mathbf{x})) = T^{-1}\ell(\hat{p}(\mathbf{y}|\mathbf{x}))$
 - Loss correction is unbiased
 - $R_{\tilde{\mathcal{D}}, \ell^{\leftarrow}}(\hat{p}(\mathbf{y}|\mathbf{x})) = R_{\mathcal{D}, \ell}(\hat{p}(\mathbf{y}|\mathbf{x})) \implies$ the minimizers are same
- Forward correction procedure: $\ell_{\psi}^{\rightarrow}(h(\mathbf{x})) = \ell(T^T \psi^{-1} h(\mathbf{x}))$ (better!)
 - l_{ψ} is a proper composite loss, i.e., its minimizer assumes the particular shape of the link function applied to the $p(\mathbf{y}|\mathbf{x})$
- Estimating T : find $\bar{\mathbf{x}}^i = \arg \max_{\mathbf{x} \in \mathcal{X}} \hat{p}(\tilde{\mathbf{y}} = \mathbf{e}^i | \mathbf{x})$ and then $\hat{T}_{ij} = \hat{p}(\tilde{\mathbf{y}} = \mathbf{e}^i | \bar{\mathbf{x}}^j)$

Loss correction approach [Patrini et al., 2017]

Multi-class classification : $\mathbf{y} \in \{0, 1\}^c$ with c classes; network's prediction $\hat{p}(\mathbf{y}|\mathbf{x})$

Let $T \in [0, 1]^{c \times c}$ be the noise transition matrix s.t. $T_{ij} = p(\tilde{\mathbf{y}} = \mathbf{e}^i | \mathbf{y} = \mathbf{e}^j)$ and 'known' row stochastic

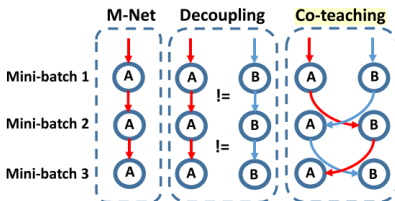
- Backward correction procedure: $\ell^{\leftarrow}(\hat{p}(\mathbf{y}|\mathbf{x})) = T^{-1}\ell(\hat{p}(\mathbf{y}|\mathbf{x}))$
 - Loss correction is unbiased
 - $R_{\tilde{\mathcal{D}}, \ell^{\leftarrow}}(\hat{p}(\mathbf{y}|\mathbf{x})) = R_{\mathcal{D}, \ell}(\hat{p}(\mathbf{y}|\mathbf{x})) \implies$ the minimizers are same
- Forward correction procedure: $\ell_{\psi}^{\rightarrow}(h(\mathbf{x})) = \ell(T^T \psi^{-1} h(\mathbf{x}))$ (better!)
 - l_{ψ} is a proper composite loss, i.e., its minimizer assumes the particular shape of the link function applied to the $p(\mathbf{y}|\mathbf{x})$
- Estimating T : find $\bar{\mathbf{x}}^i = \arg \max_{\mathbf{x} \in \mathcal{X}} \hat{p}(\tilde{\mathbf{y}} = \mathbf{e}^i | \mathbf{x})$ and then $\hat{T}_{ij} = \hat{p}(\tilde{\mathbf{y}} = \mathbf{e}^i | \bar{\mathbf{x}}^j)$

[Lukasik et al., 2020] show connection b/w label smoothing and loss correction

approaches; attribute its de-noising properties to ℓ_2 regularization.

Co-teaching [Han et al., 2018]

- Relies on training on selected/identified clean samples and the idea that DNNs learn easy instances first and then move on to difficult instances
- Correct labels would tend to have small loss instances
- Keep the batch size high initially and decrease in later stage of learning
- Two networks have different learning abilities so they can filter different types of error introduced by noisy labels and the error flows can be reduced by peer networks mutually.



Deep k-NN for noisy labels [Bahri et al., 2020]

- Fits in the class of data cleaning methods that detects and filters dirty data.
- Results are based on the notion of how spread out the noisy labelled points are. Also, doesn't assume any particular noise model
- The proposed example filters out the examples that disagree with the estimate of in-class probability $\eta(\mathbf{x})$ computed on noisy data
- Works both for the cases when some clean labelled examples are available too. And has been shown to be robust to the choice of neighbouring parameter in k-NN.
- Provide statistical guarantees like convergence rates and finite sample analysis.

Why can't I have 'some' clean labelled data?

[Hendrycks et al., 2018]

- Leverage the fact that a small set of clean labels is often procurable (Gold standard)
- Estimate the corruption matrix C (defined as T earlier)
 - Learn a classifier on corrupted data
 - Use the averaged probability predictions from this model on the clean data as estimates
- Experiments suggest major strength lies in the estimate of noise matrix

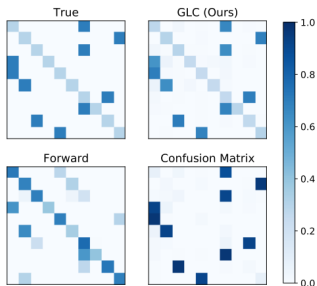


Figure 1: A label corruption matrix (top left) and three matrix estimates for a corrupted CIFAR-10 dataset. Entry C_{ij} is the probability that a label of class i is corrupted to class j , or symbolically $C_{ij} = p(\tilde{y} = j | y = i)$.

Using GANs for classification task from noisy labels

Generative adversarial networks [Goodfellow et al., 2014]

- Two player min-max game between discriminator D and generator G .

$$\min_G \max_D \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} [\log(D(\mathbf{x}))] + \mathbb{E}_{p(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]$$

- $G : \mathcal{Z} \mapsto \mathcal{X}$, synthesizes new samples by mimicking target distribution \mathcal{D} .
- $D : \mathcal{X} \mapsto \mathbb{R}$, decides whether the sample is real or fake (generated from G).
- $p(\mathbf{z})$ and $p_{\mathcal{D}}(\mathbf{x})$ are density functions of the random variables in \mathcal{Z} and \mathcal{X} .

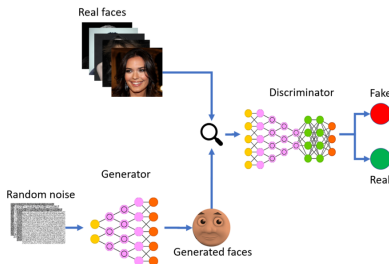


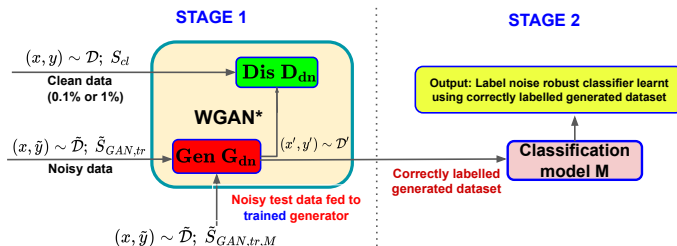
Figure: A schematic for GANs. Source: <https://medium.com/sigmoid/>

WasserteinGANY [Tripathi and Hemachandra, 2019b]

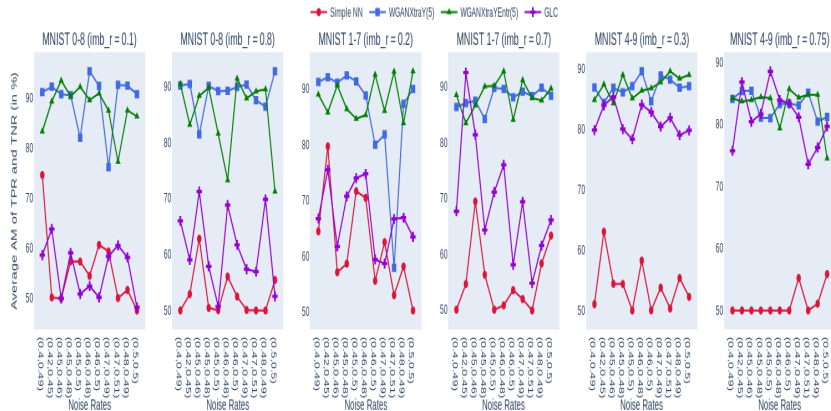
Objective function of WGANY

$$\min_{G_{dn}} \max_{D_{dn}} \mathbb{E}_{(x,y) \sim \mathcal{D}} [D_{dn}(x,y)] - \mathbb{E}_{(x',y') \sim \mathcal{D}'} [D_{dn}(x',y')]$$

- $G_{dn} : \mathcal{X} \times \tilde{\mathcal{Y}} \mapsto \mathcal{X} \times \mathcal{Y}, D_{dn} : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$
- $(x', y') := (G_{dn}^f(x, \tilde{y}), G_{dn}^l(x, \tilde{y})) = G_{dn}(x, \tilde{y})$, with $(x, \tilde{y}) \sim \tilde{\mathcal{D}}$
- Goal: Minimize the divergence between clean distribution \mathcal{D} and correctly labelled generated model distribution \mathcal{D}' .



Additional benefits for Imbalanced datasets (MNIST)



For high noise rates, WGAN based schemes have lead to higher average AM values (WGANXtraY(5) and WGANXtraYEntr(5)) in comparison to GLC [Hendrycks et al., 2018] and baseline SimpleNN with no modification for label noise robustness.

Other interesting recent studies

- [Chen et al., 2019]: Relate the noise rates to test set accuracy and propose an algorithm for learning
- SIGUA [Han et al.,]: avoids undesired memorization in over parameterized networks by reducing the learning rate for ascent on bad data (outliers, corrupted labels)
- Search strategies [Yang et al., 2019]: Deciding how many examples to discard that are supposedly noisy at every iteration by formulating it as a bi-level optimization problem
- Label noise information [Harutyunyan et al., 2020]: Memorized information of a network is quantified as the Shannon mutual information between weights and the vectors of all training labels
- Early stopping [Li et al., 2020]: prove that using first order methods with early stopping for training over parameterized neural networks make them robust to label noise
- ...

Instance dependent (aka non-uniform noise model), $\rho(\mathbf{x})$

- Loss fn $\ell(\cdot)$ is robust to *non-uniform* label noise if $\ell(\cdot)$ satisfies
 - $\ell_1(f(\mathbf{x})) + \ell_{-1}(f(\mathbf{x})) = K, \quad K > 0,$
 - If true risk is zero, i.e., $R_{\mathcal{D}}(f^*) = R^*(f^*) = R^* = 0$
- In fact, classifier from noisy data, $\tilde{f}_{\mathcal{D}}^*$, is same as f^*
- So, if true risk, R^* is zero, then 01-loss fn, $\ell_{01}(\cdot)$ is robust to non-uniform noise.
- Suppose true risk $R^* = 0$. In addition to 01-loss, $\ell_{01}(\cdot)$,
 - Sigmoid loss, $\ell_{sg}(\mathbf{x}) = \frac{1}{1 + \exp(\beta f(\mathbf{x})y)}$
 - Ramp loss, $\ell_{ra}(\mathbf{x}) = (1 - \beta f(\mathbf{x})y)^+ - (-1 - \beta)f(\mathbf{x})y^+$
 - Probit loss, $\ell_{pr}(\mathbf{x}) = 1 - \Phi(\beta f(\mathbf{x})y)$

are robust to non-uniform noise for sufficiently large (respective) β .

Instance dependent (aka non-uniform noise model), $\rho(\mathbf{x})$

- Loss fn $\ell(\cdot)$ is robust to *non-uniform* label noise if $\ell(\cdot)$ satisfies
 - $\ell_1(f(\mathbf{x})) + \ell_{-1}(f(\mathbf{x})) = K, \quad K > 0,$
 - If true risk is zero, i.e., $R_{\mathcal{D}}(f^*) = R^*(f^*) = R^* = 0$
- In fact, classifier from noisy data, $\tilde{f}_{\mathcal{D}}^*$, is same as f^*
- So, if true risk, R^* is zero, then 01-loss fn, $\ell_{01}(\cdot)$ is robust to non-uniform noise.
- Suppose true risk $R^* = 0$. In addition to 01-loss, $\ell_{01}(\cdot)$,
 - Sigmoid loss, $\ell_{sg}(\mathbf{x}) = \frac{1}{1 + \exp(\beta f(\mathbf{x})y)}$
 - Ramp loss, $\ell_{ra}(\mathbf{x}) = (1 - \beta f(\mathbf{x})y)^+ - (-1 - \beta)f(\mathbf{x})y^+$
 - Probit loss, $\ell_{pr}(\mathbf{x}) = 1 - \Phi(\beta f(\mathbf{x})y)$

are robust to non-uniform noise for sufficiently large (respective) β .

- For large β , these loss fns. are *sufficiently steep near 0* to apprx ℓ_{01} .

Non-uniform noise robustness, cont.

- The assumption $R_{\mathcal{D}}(f_{\mathcal{D}}^*) = R_{\mathcal{D}}(f^*) = 0$ is not very restrictive as \mathcal{D} is hypothetical, [\[\[Manwani and Sastry, 2013\]\]](#), [\[\[Ghosh et al., 2017\]\]](#), etc.
- This condition seems necessary – if you $R_{li}^* \neq 0$, then one can have a counter-example where ℓ_{01} is not noise robust among linear classifiers, [\[Manwani and Sastry, 2013\]](#).

Non-uniform noise robustness, An example

On going work: *Learning hedonic cooperative games with noisy preferences*

- Each player in a cooperative game has *preferences* on coalitions to join involving other players
- Prediction of core stable grand coalition formation is modelled as classification problem
- Say, preferences are *noisy*.
- A instance dependent noisy 'label' model.
- But, $R_{\mathcal{D}}(f^*) = 0$, as stable coalition prediction formation can be predicted when clean/original preferences are known.
- Thus, above loss fns with large enough tunable parameter β are robust to noisy preferences

Learning is more than 0-1 loss based binary classification

Cost sensitive learning with noisy labels

- *Differential costing of mis-classifications due to nature of decision systems like medicine where safety critical decisions are made.*
- *The availability of data could be skewed like the number of customers defaulting credit card is very low*

Some solutions

- *Balance out the training data by up-sampling (down-sampling) the minority (majority) class*
- *Use an α -weighted classification loss like*

$$\ell_{\alpha,0-1}(f(\mathbf{x}), y) = (1 - \alpha) \mathbb{1}_{[y=1, f(\mathbf{x}) < 0]} + \alpha \mathbb{1}_{[y=-1, f(\mathbf{x}) \geq 0]}$$

where α is either user given or cross-validated for.

Cost sensitive learning with noisy labels

- *Differential costing of mis-classifications due to nature of decision systems like medicine where safety critical decisions are made.*
- *The availability of data could be skewed like the number of customers defaulting credit card is very low*

Some solutions

- *Balance out the training data by up-sampling (down-sampling) the minority (majority) class*
- *Use an α -weighted classification loss like*

$$\ell_{\alpha,0-1}(f(\mathbf{x}), y) = (1 - \alpha) \mathbb{1}_{[y=1, f(\mathbf{x}) < 0]} + \alpha \mathbb{1}_{[y=-1, f(\mathbf{x}) \geq 0]}$$

where α is either user given or cross-validated for.

What if the labels are noisy?

Cost sensitive learning with noisy labels [Natarajan et al., 2018]

$\ell_\alpha = ((1 - \alpha)\mathbb{1}_{[y=1]} + \alpha\mathbb{1}_{[y=-1]})\ell(f(\mathbf{x}, y))$ loses its consistency property in the presence of label noise

- Approach of unbiased estimators as defined earlier has a version for weighted losses too.
- Cross validate the imbalance parameter α and the noise rates!
- Accuracy is not a good measure for imbalanced datasets; hence use Arithmetic mean.

What happens when α is given? Can we avoid estimating the noise rates?

Cost sensitive learning for SLN models

[Tripathi and Hemachandra, 2019a]

- *Is risk minimization under weighted 0-1 loss uniform noise tolerant?*
NO!! Look for surrogates! α -weighted uneven margin loss functions
[Scott, 2012].

Definition $((\alpha, \gamma, \rho)$ -robustness of risk minimization)

For a loss function $l_{\alpha,un}$ and classifiers $\tilde{f}_{l_{\alpha,un}}^*$ and $f_{l_{\alpha,un}}^*$, risk minimization is said to be $((\alpha, \gamma, \rho)$ -robust if

$$R_{D,\alpha}(\tilde{f}_{l_{\alpha,un}}^*) = R_{D,\alpha}(f_{l_{\alpha,un}}^*)$$

Further, if the classifiers in equation (2) are $f_{r,l_{\alpha,un}}^*$ and $\tilde{f}_{r,l_{\alpha,un}}^*$ then, we say that regularized risk minimization under $l_{\alpha,un}$ is $((\alpha, \gamma, \rho)$ -robust.

Cost sensitive learning for SLN models

[Tripathi and Hemachandra, 2019a]

- Is risk minimization under weighted 0-1 loss uniform noise tolerant?
NO!! Look for surrogates! α -weighted uneven margin loss functions
[Scott, 2012].





Definition $((\alpha, \gamma, \rho)$ -robustness of risk minimization)

For a loss function $l_{\alpha,un}$ and classifiers $\tilde{f}_{l_{\alpha,un}}^*$ and $f_{l_{\alpha,un}}^*$, risk minimization is said to be $((\alpha, \gamma, \rho)$ -robust if

$$R_{D,\alpha}(\tilde{f}_{l_{\alpha,un}}^*) = R_{D,\alpha}(f_{l_{\alpha,un}}^*)$$

Further, if the classifiers in equation (2) are $f_{r,l_{\alpha,un}}^*$ and $\tilde{f}_{r,l_{\alpha,un}}^*$ then, we say that regularized risk minimization under $l_{\alpha,un}$ is $((\alpha, \gamma, \rho)$ -robust.

- α -weighted uneven margin squared loss

Hypothesis class Loss functions	Hypothesis class of all measurable functions	Hypothesis class of all linear functions
α -weighted 0-1 loss function		
α -weighted γ -uneven margin squared loss		

Cost sensitive learning in label noise

Approach 1: $l_{\alpha, usq}$ is (α, γ, ρ) robust with $f \in \mathcal{H}_{lin}$ **α -weighted γ -uneven margin squared loss:**

$$l_{\alpha, usq}(f(\mathbf{x}), y) = (1 - \alpha)\mathbf{1}_{\{y=1\}}(1 - f(\mathbf{x}))^2 + \alpha\mathbf{1}_{\{y=-1\}}\frac{1}{\gamma}(1 + \gamma f(\mathbf{x}))^2, \text{ for } \gamma > 0.$$

- Can be obtained by just solving a linear system of equations
- High probability bound available on the risk

Cost sensitive learning in label noise

Approach 1: $l_{\alpha, usq}$ is (α, γ, ρ) robust with $f \in \mathcal{H}_{lin}$ **α -weighted γ -uneven margin squared loss:**

$$l_{\alpha, usq}(f(\mathbf{x}), y) = (1 - \alpha) \mathbf{1}_{\{y=1\}} (1 - f(\mathbf{x}))^2 + \alpha \mathbf{1}_{\{y=-1\}} \frac{1}{\gamma} (1 + \gamma f(\mathbf{x}))^2, \text{ for } \gamma > 0.$$

- Can be obtained by just solving a linear system of equations
- High probability bound available on the risk

Approach 2: Re-sampling based scheme

- α -weighted uneven margin 0-1 loss function:

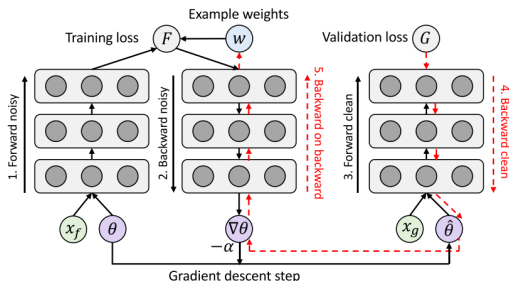
$$l_{0-1, \alpha, \gamma}(f(\mathbf{x}), y) = (1 - \alpha) \mathbf{1}_{\{y=1, f(\mathbf{x}) \leq 0\}} + \frac{\alpha}{\gamma} \mathbf{1}_{\{y=-1, \gamma f(\mathbf{x}) > 0\}}, \quad \forall \alpha \in (0, 1) \quad (2)$$

- **Noisy cost sensitive Bayes classifier:** $\tilde{f}_{0-1, \alpha, \gamma}^* = \text{sign} \left(\tilde{\eta}(\mathbf{x}) - \frac{\alpha}{\gamma + (1 - \alpha)\gamma} \right)$
- Issues: most algorithms use a threshold of 0.5 and can't use $\tilde{\eta}$ for prediction
- Solution:
 1. Re-balance noisy dataset with $r^* = \frac{\alpha}{\gamma(1 - \alpha)}$.
 2. Estimate $\tilde{\eta}_b$ from re-balanced noisy data.
 3. For test point \mathbf{x}_0 , prediction is $\hat{y}_0 = \text{sign}(\tilde{\eta}_b(\mathbf{x}_0) - 0.5)$.

Learning to re-weight examples for robust deep learning

[Ren et al., 2018]

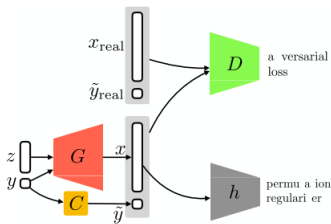
- Reduce the effect of training set biases like label noise and class imbalance on neural networks
- Propose a meta learning algorithm that learns to assign weights to training examples based on their gradient directions.
- Assumes availability of a clean and balanced validation set at the end of every iteration.



Generative Adversarial Networks and label noise I

Relevant problem for the ones that use label information like conditional GANs

- RCGAN [Thekumparampil et al., 2018]: Modifies generator's loss by adding a suitable noisy classifier (learnt before hand on noisy data on a certain hypothesis class using a certain loss function) based loss term as regularizer.



Generative Adversarial Networks and label noise II

- rcGAN and rACGAN [Kaneko et al., 2019]: Noise transition model is used while training the discriminator.

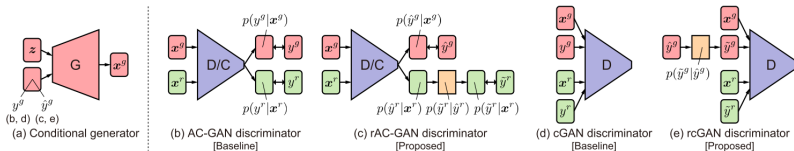


Figure 2. Comparison of naive and label-noise robust GANs. We denote the generator, discriminator, and auxiliary classifier by G , D , and C , respectively. Among all models, conditional generators (a) are similar. In our rAC-GAN (c) and rcGAN (e), we incorporate a noise transition model (viewed as an orange rectangle) into AC-GAN (b) and cGAN (d), respectively.

Positive-Unlabelled Learning

Exists when an observation of positive example is more reliable Eg.: a protein catalyzing a reaction or a social media user liking a product.

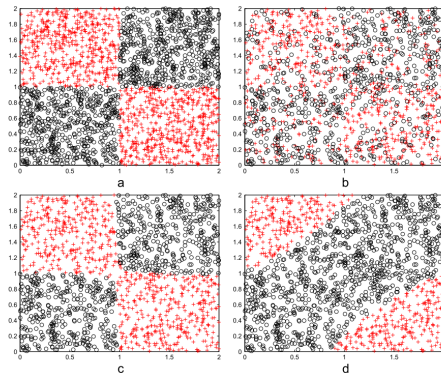
- [Jain et al., 2016]: label noise robust algorithm that estimates the class priors first, i.e., π and then posterior distribution η .
 - Works well for density estimation even in high dimension data as it transforms the full dimensional input space to a univariate space
- [Shi et al., 2018]: Treat a PU learning problem as a label noise problem by labelling all unlabelled examples as negative:
 - Decompose the empirical risk for positive and negative examples.
 - Upper bound the empirical risk on the negative examples using the results from [Patrini et al., 2016] for hinge loss.

Interaction of label noise with other problems

- Active Learning [Kremer et al., 2018]
- Domain Adaptation [Yu et al., 2020]
- Possibly in fairness literature with different interpretation of noise:
 - due to human biases [Jiang and Nachum, 2020]
 - noisy version of unobserved variable is only available [Fogliato et al., 2020]

Experiment designing

- Real datasets: Clothing1M [Xiao et al., 2015] is known to be noisy
- Synthetic data: Data can be generated in certain way and then noise is induced by flipping the labels. [Ghosh et al., 2015].



Categorical feature noise [Petety et al., 2020]

Possible existence examples:

- Room with many sensors connected in series (or with individual battery) measuring temperature, humidity, etc., as binary value, i.e., high or low.
- A power failure (or battery failures) will lead to all (or individual) sensors/attributes providing noisy observations with same (or different) probability.
- The feature vector \mathbf{x} is all categorical
- The flipping of these values can alter the classifier

Squared loss is robust if the flipping probability is same for all the features but there are counter examples for 0-1 loss

References I

- [Amid et al., 2019] Amid, E., Warmuth, M. K., Anil, R., and Koren, T. (2019).
Robust bi-tempered logistic loss based on bregman divergences.
In *Advances in Neural Information Processing Systems*, pages 14987–14996.
- [Bahri et al., 2020] Bahri, D., Jiang, H., and Gupta, M. (2020).
Deep k-nn for noisy labels.
arXiv preprint arXiv:2004.12289.
- [Chen et al., 2019] Chen, P., Liao, B. B., Chen, G., and Zhang, S. (2019).
Understanding and utilizing deep neural networks trained with noisy labels.
In *Proceedings of the International Conference on Machine Learning*, pages 1062–1070.
- [Dua and Graff, 2017] Dua, D. and Graff, C. (2017).
UCI machine learning repository.

References II

[Feng et al., 2020] Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. (2020).

Can cross entropy loss be robust to label noise?

IJCAI.

[Fogliato et al., 2020] Fogliato, R., G'Sell, M., and Chouldechova, A. (2020).

Fairness evaluation in presence of biased noisy labels.

arXiv preprint arXiv:2003.13808.

[Ghosh et al., 2015] Ghosh, A., Manwani, N., and Sastry, P. (2015).

Making risk minimization tolerant to label noise.

Neurocomputing, 160:93–107.

[Ghosh et al., 2017] Ghosh, A., Manwani, N., and Sastry, P. (2017).

On the robustness of decision tree learning under label noise.

In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 685–697. Springer.

References III

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).

Generative adversarial nets.

In *Advances in Neural Information Processing Systems*, pages 2672–2680.

[Han et al.,] Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I. W., and Sugiyama, M. Sigua: Forgetting may make learning with noisy labels more robust.

[Han et al., 2018] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. (2018).

Co-teaching: Robust training of deep neural networks with extremely noisy labels.

In *Advances in Neural Information Processing Systems*, pages 8536–8546.

References IV

[Harutyunyan et al., 2020] Harutyunyan, H., Reing, K., Steeg, G. V., and Galstyan, A. (2020).

Improving generalization by controlling label-noise information in neural network weights.

arXiv preprint arXiv:2002.07933.

[Hendrycks et al., 2018] Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. (2018).

Using trusted data to train deep networks on labels corrupted by severe noise.

In Advances in Neural Information Processing Systems, pages 10456–10465.

[Jain et al., 2016] Jain, S., White, M., and Radivojac, P. (2016).

Estimating the class prior and posterior from noisy positives and unlabeled data.

In Advances in Neural Information Processing Systems, pages 2693–2701.

References V

[Jiang and Nachum, 2020] Jiang, H. and Nachum, O. (2020).

Identifying and correcting label bias in machine learning.

In *International Conference on Artificial Intelligence and Statistics*, pages 702–712.

[Kaneko et al., 2019] Kaneko, T., Ushiku, Y., and Harada, T. (2019).

Label-noise robust Generative Adversarial Networks.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2476.

[Kremer et al., 2018] Kremer, J., Sha, F., and Igel, C. (2018).

Robust active label correction.

In *International Conference on Artificial Intelligence and Statistics*, pages 308–316.

References VI

[Kumar and Sastry, 2018] Kumar, H. and Sastry, P. (2018).

Robust loss functions for learning multi-class classifiers.

In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 687–692. IEEE.

[Li et al., 2020] Li, M., Soltanolkotabi, M., and Oymak, S. (2020).

Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks.

In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR.

[Liu and Tao, 2016] Liu, T. and Tao, D. (2016).

Classification with noisy labels by importance reweighting.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(3):447–461.

References VII

[Liu and Guo, 2019] Liu, Y. and Guo, H. (2019).

Peer loss functions: Learning from noisy labels without knowing noise rates.
arXiv preprint arXiv:1910.03231.

[Long and Servedio, 2010] Long, P. M. and Servedio, R. A. (2010).

Random classification noise defeats all convex potential boosters.
Machine learning, 78(3):287–304.

[Lukasik et al., 2020] Lukasik, M., Bhojanapalli, S., Menon, A. K., and Kumar, S. (2020).

Does label smoothing mitigate label noise?
arXiv preprint arXiv:2003.02819.

[Ma et al.,] Ma, X., Huang, H., Wang, Y., Erfani, S. R. S., and Bailey, J.

Normalized loss functions for deep learning with noisy labels.

References VIII

[Manwani and Sastry, 2013] Manwani, N. and Sastry, P. (2013).

Noise tolerance under risk minimization.

IEEE transactions on Cybernetics, 43(3):1146–1151.

[Natarajan et al., 2018] Natarajan, N., Dhillon, I. S., Ravikumar, P., and Tewari, A. (2018).

Cost-sensitive learning with noisy labels.

Journal of Machine Learning Research, 18(155):1–33.

[Natarajan et al., 2013] Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. (2013).

Learning with noisy labels.

In *Advances in Neural Information Processing Systems*, pages 1196–1204.

References IX

- [Patrini et al., 2016] Patrini, G., Nielsen, F., Nock, R., and Carioni, M. (2016).
Loss factorization, weakly supervised learning and label noise robustness.
In *Proceedings of the International Conference on Machine Learning*, pages 708–717.
- [Patrini et al., 2017] Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. (2017).
Making deep neural networks robust to label noise: A loss correction approach.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2233–2241.
- [Petety et al., 2020] Petety, A., Tripathi, S., and Hemachandra, N. (2020).
Attribute noise robust binary classification (student abstract).
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13897–13898.

References X

- [Ren et al., 2018] Ren, M., Zeng, W., Yang, B., and Urtasun, R. (2018).
Learning to reweight examples for robust deep learning.
In *Proceedings of the International Conference on Machine Learning*, pages 4331–4340.
- [Rusiecki, 2019] Rusiecki, A. (2019).
Trimmed categorical cross-entropy for deep learning with label noise.
Electronics Letters, 55(6):319–320.
- [Sastry and Manwani, 2017] Sastry, P. and Manwani, N. (2017).
Robust learning of classifiers in the presence of label noise.
In *Pattern Recognition and Big Data*, pages 167–197. World Scientific.
- [Scott, 2012] Scott, C. (2012).
Calibrated asymmetric surrogate losses.
Electronic Journal of Statistics, 6:958–992.

References XI

- [Shi et al., 2018] Shi, H., Pan, S., Yang, J., and Gong, C. (2018).
Positive and unlabeled learning via loss decomposition and centroid estimation.
In Proceedings of the International Joint Conference on Artificial Intelligence,
pages 2689–2695.
- [Thekumparampil et al., 2018] Thekumparampil, K. K., Khetan, A., Lin, Z., and Oh, S.
(2018).
Robustness of conditional GANs to noisy labels.
In Advances in Neural Information Processing Systems, pages 10271–10282.
- [Tripathi and Hemachandra, 2019a] Tripathi, S. and Hemachandra, N. (2019a).
Cost sensitive learning in the presence of symmetric label noise.
In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 15–28. Springer.

References XII

- [Tripathi and Hemachandra, 2019b] Tripathi, S. and Hemachandra, N. (November, 2019b).
GANs for learning from very high class conditional noisy labels.
Working paper.
- [Van Rooyen et al., 2015] Van Rooyen, B., Menon, A., and Williamson, R. C. (2015).
Learning with symmetric label noise: The importance of being unhinged.
In *Advances in Neural Information Processing Systems*, pages 10–18.
- [Wang et al., 2019] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. (2019).
Symmetric cross entropy for robust learning with noisy labels.
In *Proceedings of the IEEE International Conference on Computer Vision*, pages 322–330.

References XIII

[Xiao et al., 2015] Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. (2015).

Learning from massive noisy labeled data for image classification.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699.

[Xu et al., 2019] Xu, Y., Cao, P., Kong, Y., and Wang, Y. (2019).

L_DMI: A novel information-theoretic loss function for training deep nets robust to label noise.

In *Advances in Neural Information Processing Systems*, pages 6222–6233.

[Yang et al., 2019] Yang, H., Yao, Q., Han, B., and Niu, G. (2019).

Searching to exploit memorization effect in learning from corrupted labels.

arXiv preprint arXiv:1911.02377.

References XIV

[Yu et al., 2020] Yu, X., Liu, T., Gong, M., Zhang, K., Batmanghelich, K., and Tao, D. (2020).

Label-noise robust domain adaptation.

ICML.

[Zhang and Sabuncu, 2018] Zhang, Z. and Sabuncu, M. (2018).

Generalized cross entropy loss for training deep neural networks with noisy labels.

In *Advances in Neural Information Processing Systems*, pages 8778–8788.



Thank you!
Questions? Comments?