THE UNIVERSITY OF SYDNEY

Advanced Machine Learning

(COMP 5328)

Week 10 Tutorial:
Learning with Noisy Data II: Label Noise

Anjin Liu
anjin.liu@sydney.edu.au

https://github.com/Anjin-Liu/usyd_comp_5328_tutorial_S22025

Tutorial Contents

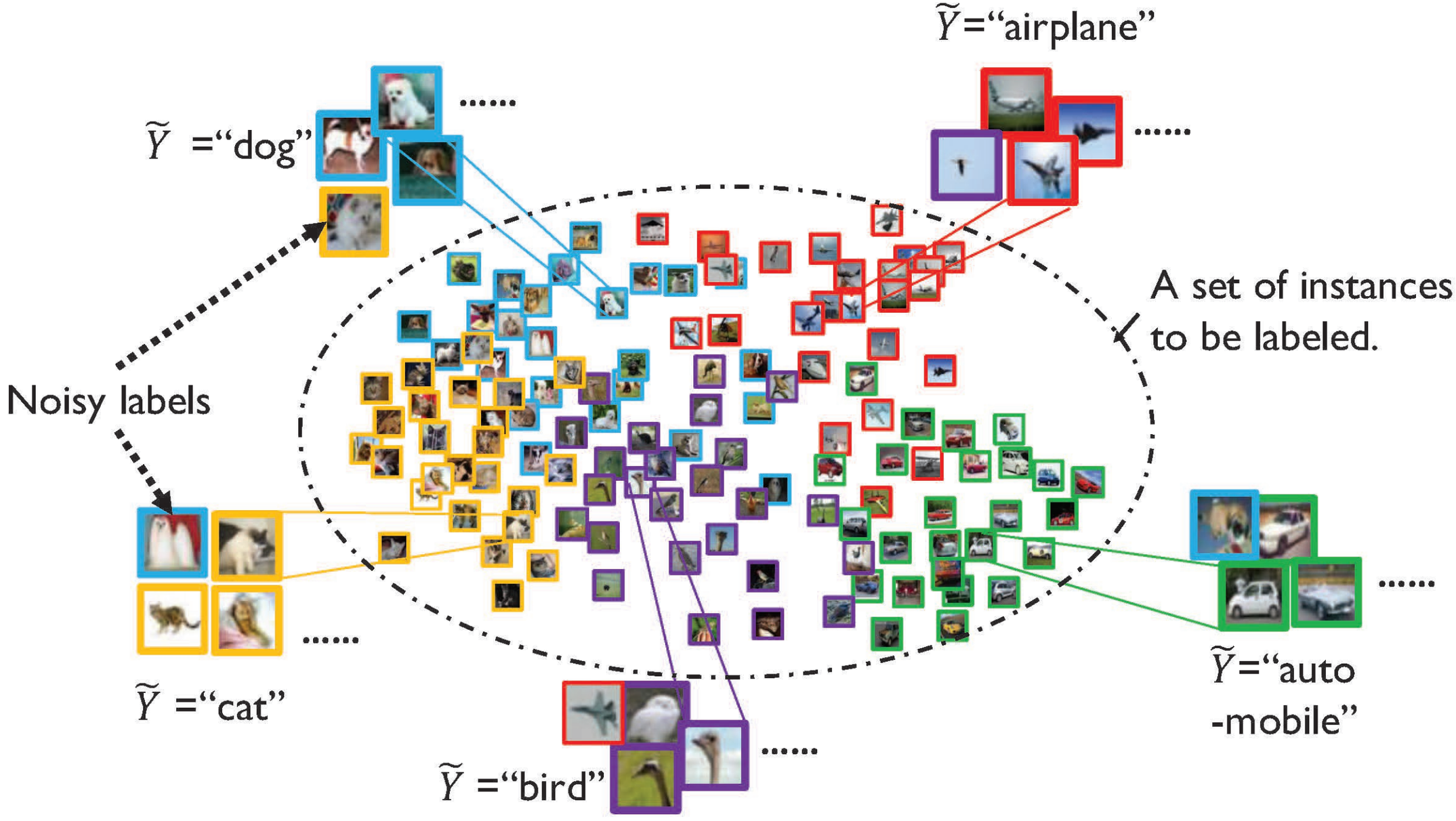
THE UNIVERSITY OF SYDNEY

- Review (20min):
  - Lecture 8: Learning with Noisy Data II: Label Noise
- Tutorial exercise & QA (40min):
- Announcements: Important Information for Assignment 2

Key points

- Learning with Noisy Labels – Problem Setup
- Random Classification Noise
- Class-dependent Label Noise: Binary
- Class-dependent Label Noise: Multi-class
- Instance- and Class-dependent Label Noise

What is label noise?



What is label noise?

Label noise widely exists even for small data because:

- Labels are provided by non-expert labellers, such as those in the Amazon Mechanical Turk.

- The labelling task is subjective, especially in tasks such as image captioning.

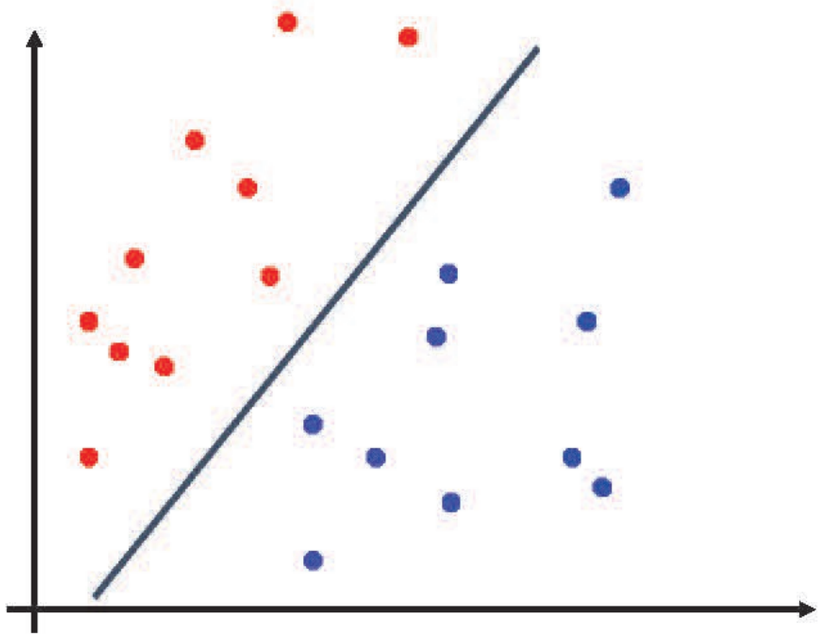
- Insufficient discriminative information for assigning reliable labels. Collecting reliable labels is time-consuming and costly.

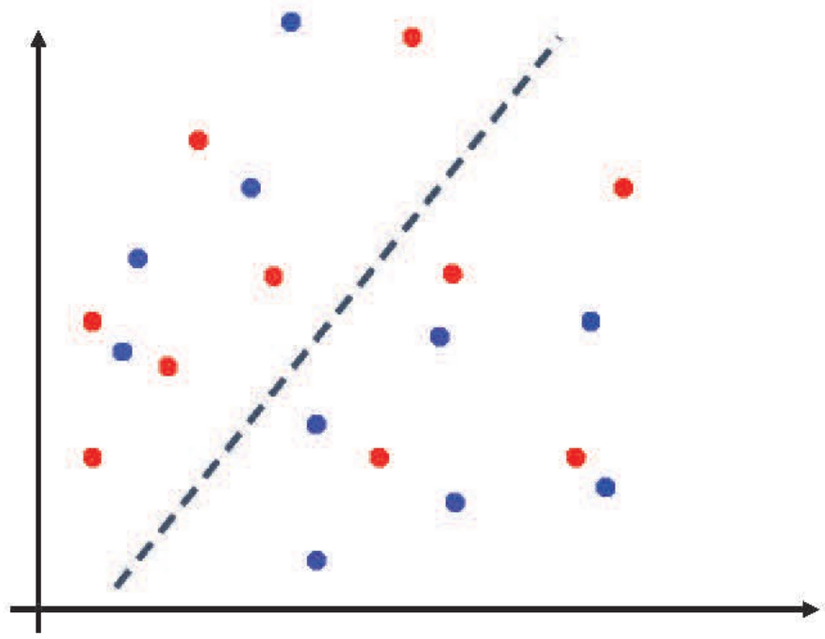
Learning with noisy labels

THE UNIVERSITY OF SYDNEY

$$Y \xrightarrow{\text{noise added}} \tilde{Y}$$

$$S=(X_1, Y_1), \ldots, (X_n, Y_n) \xrightarrow{\text{label noise}} (X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$$





Learning with noisy labels

THE UNIVERSITY OF SYDNEY

Problem Setup

- Observation: $X \in \mathcal{X} \subset \mathbb{R}^d$.
- Clean but unobservable label: $Y \in \mathcal{Y} = \{-1, +1\}$.

Ground Truth

- Observable but noisy label: $\tilde{Y} \in \mathcal{Y}$.
- Clean distribution: $D(X, Y)$; Noisy distribution: $D_\rho(X, \tilde{Y})$.

Learning with noisy labels

THE UNIVERSITY OF SYDNEY

Problem Setup

- Given the training examples $\{(X_i, \tilde{Y}_i)\}_{1 \le i \le n} \sim D_\rho(X, \tilde{Y})^n$.

- The target is to learn a discriminant function f_n: $\mathcal{X} \to \mathbb{R}$ such that the classifier predicts the correct label y given an observation $\boldsymbol{x}$.

Model Label Noise

THE UNIVERSITY OF SYDNEY

A probabilistic model:

$$\rho_Y(X) = P\left(\tilde{Y}\middle|Y, X\right),$$

where X is the feature, Y is the unobservable true label, and $\tilde{Y}$ is the observed noisy label.

$$\rho_{+1}(X) = P(\tilde{Y} = -1|Y = 1, X); \rho_{-1}(X) = P(\tilde{Y} = 1|Y = -1, X).$$

Note that if there is no label noise, we have

$$P\left(\tilde{Y} = 1\middle|Y = 1, X\right) = P\left(\tilde{Y} = -1\middle|Y = -1, X\right) = 1$$

otherwise

$$P\left(\tilde{Y} = -1|Y = 1, X\right) = P\left(\tilde{Y} = 1|Y = -1, X\right) = 0$$

$$P\left(\tilde{Y} = 1\middle|Y = -1, X\right), P\left(\tilde{Y} = -1\middle|Y = 1, X\right) \in (0,1).$$

Model Label Noise

THE UNIVERSITY OF SYDNEY

(1) Random Classification Noise (RCN):

$$\rho_Y(X) = P(\tilde{Y}|Y,X) = P(\tilde{Y}|Y);\ \rho_{+1}(X) = \rho_{-1}(X) = \rho.$$

$$\rho_{+1}(X) = P(\tilde{Y} = -1|Y = 1, X);\ \rho_{-1}(X) = P(\tilde{Y} = 1|Y = -1, X).$$

(2) Class-Dependent Noise (CCN):

$$\rho_Y(X) = P\left(\tilde{Y}\middle|Y,X\right) = P\left(\tilde{Y}\middle|Y\right);\ \rho_{+1}(X) = \rho_{+1},\ \rho_{-1}(X) = \rho_{-1}.$$

(3) Instance- and Label-Dependent Noise (ILN):

$$\rho_Y(X) = P\left(\tilde{Y}\middle|Y,X\right).$$

$\rho_{+1}(X)$ actual positive but observed negative
$\rho_{-1}(X)$ actual negative but observed positive

THE UNIVERSITY OF SYDNEY

Random Classification Noise

Random Classification Noise (RCN)

THE UNIVERSITY OF SYDNEY

The Impact of Label Noise

Under RCN, minimisation of *any* **convex potential** over a **linear function class** can result in classification performance equivalent to **random guessing**.

Convex potential: any loss function $\ell: (h(X), Y) \mapsto \phi(Yh(X))$ where $\phi: \mathbb{R} \to \mathbb{R}$ is convex, non-increasing, differentiable with $\phi'(0) < 0$, and $\phi(+\infty) = 0$.

linear function class: $\mathcal{F}_{lin} = \{\boldsymbol{x} \mapsto \boldsymbol{\omega}^\top \boldsymbol{x} | \boldsymbol{\omega} \in \mathbb{R}^d\}$.

Long, Philip M., and Rocco A. Servedio. "Random classification noise defeats all convex potential boosters." Machine learning 78.3 (2010): 287-304.

Random Classification Noise (RCN)

THE UNIVERSITY OF SYDNEY

How to Reduce the Effects of RCN?

Symmetric loss function is robust to RCN when the function class $\mathcal{F}_{lin}$ is extended to the universal function space, which means the function in it can be of any form.

Theorem 1. The losses satisfying the following symmetric criterion are robust to RCN:

$$L(f(X), +1) + L(f(X), -1) = C,$$

where C is a constant. That is

$$\arg\min_f R_{D,L}(f) = \arg\min_f R_{D_\rho,L}(f).$$

Van Rooyen, Brendan, Aditya Menon, and Robert C. Williamson. "Learning with symmetric label noise: The importance of being unhinged." Advances in Neural Information Processing Systems. 2015.

Random Classification Noise (RCN)

THE UNIVERSITY OF SYDNEY

We will prove that $\arg\min_f R_{D,L}(f) = \arg\min_f R_{D_\rho,L}(f)$.

$$R_{D_\rho,L}(f) = \mathbb{E}_{(X,\tilde{Y})\sim D_\rho}[L(f(X),\tilde{Y})] = (1-2\rho)R_{D,L}(f) + \rho C.$$

Robust to noise labels means:
Its learned classification decision boundary remains the same under both noisy and noise-free conditions.

Random Classification Noise (RCN)

THE UNIVERSITY OF SYDNEY

Proof

$$P(\tilde{Y} = 1|X)$$

$$= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X)$$

$$= P(\tilde{Y} = 1|Y = 1, X)P(Y = 1|X) + P(\tilde{Y} = 1|Y = -1, X)P(Y = -1|X)$$

$$= (1 - \rho_{+1}(X))P(Y = 1|X) + \rho_{-1}(X)P(Y = -1|X)$$

$$= \big(1 - \rho_{+1}(X) - \rho_{-1}(X)\big)P(Y = 1|X) + \rho_{-1}(X).$$

Random Classification Noise (RCN)

THE UNIVERSITY OF SYDNEY

Proof Cont'd

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

Under RCN

$$P(\tilde{Y} = 1|X) = (1 - 2\rho)P(Y = 1|X) + \rho$$

$$P(\tilde{Y} = -1|X) = (1 - 2\rho)P(Y = -1|X) + \rho$$

Random Classification Noise (RCN)

THE UNIVERSITY OF SYDNEY

Proof Cont'd

$$R_{D_\rho,L}(f) = \mathbb{E}_{(X,\tilde{Y})\sim D_\rho}[L(f(X),\tilde{Y})]$$

$$= \int \Big(P\big(\tilde{Y}=1,X\big)L(f(X),1) + P\big(\tilde{Y}=-1,X\big)L(f(X),-1)\Big)dX$$

$$= \int \Big(P\big(\tilde{Y}=1|X\big)P(X)L(f(X),1) + P\big(\tilde{Y}=-1|X\big)P(X)L(f(X),-1)\Big)dX$$

$$= \int P(X)[(1-2\rho)P(Y=1|X)L(f(X),1) + \rho L(f(X),1)]dX$$

$$+\int P(X)[(1-2\rho)P(Y=-1|X)L(f(X),-1) + \rho L(f(X),-1)]dX$$

$$= (1-2\rho)\int \Big(P(Y=1,X)L(f(X),1) + P(Y=-1,X)L(f(X),-1)\Big)dX$$

$$+\rho\int P(X)[\ell(h(X),+1) + \ell(h(X),-1)]dX$$

Random Classification Noise (RCN)

THE UNIVERSITY OF SYDNEY

Proof Cont'd

$$R_{D_\rho,L}(f) = \mathbb{E}_{(X,\tilde{Y})\sim D_\rho}[L(f(X),\tilde{Y})]$$

$$= (1-2\rho)\int \Big(P(Y=1,X)L(f(X),1) + P(Y=-1,X)L(f(X),-1)\Big)dX$$

$$+\rho\int P(X)\big[\ell(h(X),+1) + \ell(h(X),-1)\big]dX$$

$$= (1-2\rho)\mathbb{E}_{(X,Y)\sim D}[L(f(X),Y)] + \rho C$$

Definition of Symmetric loss function

$$= (1-2\rho)R_{D,L}(f) + \rho C.$$

Completed!

Random Classification Noise (RCN)

THE UNIVERSITY OF SYDNEY

RCN-Robust Losses

The symmetric losses are robust to RCN:

(1) 0-1 Loss: $L(f(X), Y) = \mathbf{1}(\text{sign}(f(X)) \neq Y)$;

(2) Unhinged Loss: $L(f(X), Y) = 1 - Yf(X)$;

(3) Sigmoid Loss: $L(f(X), Y) = \frac{1}{1+e^{Yf(X)}}$;

(4) Ramp Loss: $L(f(X), Y) = \frac{1}{2}\max(0, \min(2, 1 - Yf(X)))$ ...

Ghosh, Aritra, Naresh Manwani, and P. S. Sastry. "Making risk minimization tolerant to label noise." Neurocomputing 160 (2015): 93-107.

Class-dependent Label Noise

(2) Class-Dependent Noise (CCN):

$$\rho_Y(X) = P(\tilde{Y}|Y,X) = P(\tilde{Y}|Y);\ \rho_{+1}(X) = \rho_{+1},\ \rho_{-1}(X) = \rho_{-1}.$$

Class-dependent Label Noise

THE UNIVERSITY OF SYDNEY

Modifying the loss function L to $\tilde{L}$ such that

$$\arg\min_{f\in\mathcal{F}} R_{D,L}(f) = \arg\min_{f\in\mathcal{F}} R_{D_\rho,\tilde{L}}(f).$$

Methods: **Importance reweighting**, unbiased estimator, cost-sensitive loss, rank pruning……

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.
Natarajan, Nagarajan, et al. "Learning with noisy labels." Advances in neural information processing systems. 2013.

Class-dependent Label Noise

THE UNIVERSITY OF SYDNEY

Viewing the noisy data and clean data are sampled from two domains, importance reweighting can be applied.

$$R_{D,L}(f) = \mathbb{E}_{(X,Y)\sim D}[L(f(X),Y)] = \int P_D(X,Y)\, L(f(X),Y)\mathrm{dXdY}$$

$$= \int P_{D_\rho}(X,Y)\frac{P_D(X,Y)}{P_{D_\rho}(X,Y)} L(f(X),Y)\mathrm{dXdY}$$

$$= \mathbb{E}_{(X,Y)\sim D_\rho}[\frac{P_D(X,Y)}{P_{D_\rho}(X,Y)} L(f(X),Y)]$$

$$= \mathbb{E}_{(X,Y)\sim D_\rho}[\beta(X,Y)L(f(X),Y)] \quad \text{where } \beta(\boldsymbol{x},y) = \frac{P_D(X=\boldsymbol{x},Y=y)}{P_{D_\rho}(X=\boldsymbol{x},\widetilde{Y}=y)}.$$

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.

Class-dependent Label Noise

THE UNIVERSITY OF SYDNEY

$$R_{D,L}(f) = \mathbb{E}_{(X,Y)\sim D_\rho}[\beta(X,Y)L(f(X),Y)]$$

where $\beta(\boldsymbol{x}, y) = \frac{P_D(X=\boldsymbol{x},Y=y)}{P_{D_\rho}(X=\boldsymbol{x},\tilde{Y}=y)} = \frac{P_{D_\rho}(\tilde{Y}=y|X=\boldsymbol{x})-\rho_{-y}}{(1-\rho_{+1}-\rho_{-1})P_{D_\rho}(\tilde{Y}=y|X=\boldsymbol{x})}.$

$$\arg\min_f R_{D,L}(f) = \arg\min_f R_{D_\rho,L}(f).$$

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.

THE UNIVERSITY OF SYDNEY

Class-dependent Label Noise: Multi-class

Transition matrix

We can obtain the following by using the product rule and the sum rule:

$$\begin{bmatrix} P(\tilde{Y}=1|\boldsymbol{x}) \\ \vdots \\ P(\tilde{Y}=C|\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} P(\tilde{Y}=1|Y=1,\boldsymbol{x}) & \cdots & P(\tilde{Y}=1|Y=C,\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ P(\tilde{Y}=C|Y=1,\boldsymbol{x}) & \cdots & P(\tilde{Y}=C|Y=C,\boldsymbol{x}) \end{bmatrix} \begin{bmatrix} P(Y=1|\boldsymbol{x}) \\ \vdots \\ P(Y=C|\boldsymbol{x}) \end{bmatrix}$$

Learning with noisy labels

Let T be the following flip matrix (also called transition matrix), e.g.,

$$T = \begin{bmatrix} P(\tilde{Y}=1|Y=1) & P(\tilde{Y}=1|Y=2) & \dots & P(\tilde{Y}=1|Y=C) \\ P(\tilde{Y}=2|Y=1) & P(\tilde{Y}=2|Y=2) & \dots & P(\tilde{Y}=2|Y=C) \\ \vdots & \vdots & \vdots & \vdots \\ P(\tilde{Y}=C|Y=1) & P(\tilde{Y}=C|Y=2) & \dots & P(\tilde{Y}=C|Y=C) \end{bmatrix}.$$

If we assume that given the clean label, the noisy label is independent with the instance, we have that $P(\tilde{Y}|Y) = P(\tilde{Y}|Y,X)$, and that

Forward

$$[P(\tilde{Y}=1|X),\dots,P(\tilde{Y}=C|X)]^\top = T\,[P(Y=1|X),\dots,P(Y=C|X)]^\top,$$

or $$[P(Y=1|X),\dots,P(Y=C|X)]^\top = T^{-1}[P(\tilde{Y}=1|X),\dots,P(\tilde{Y}=C|X)]^\top.$$

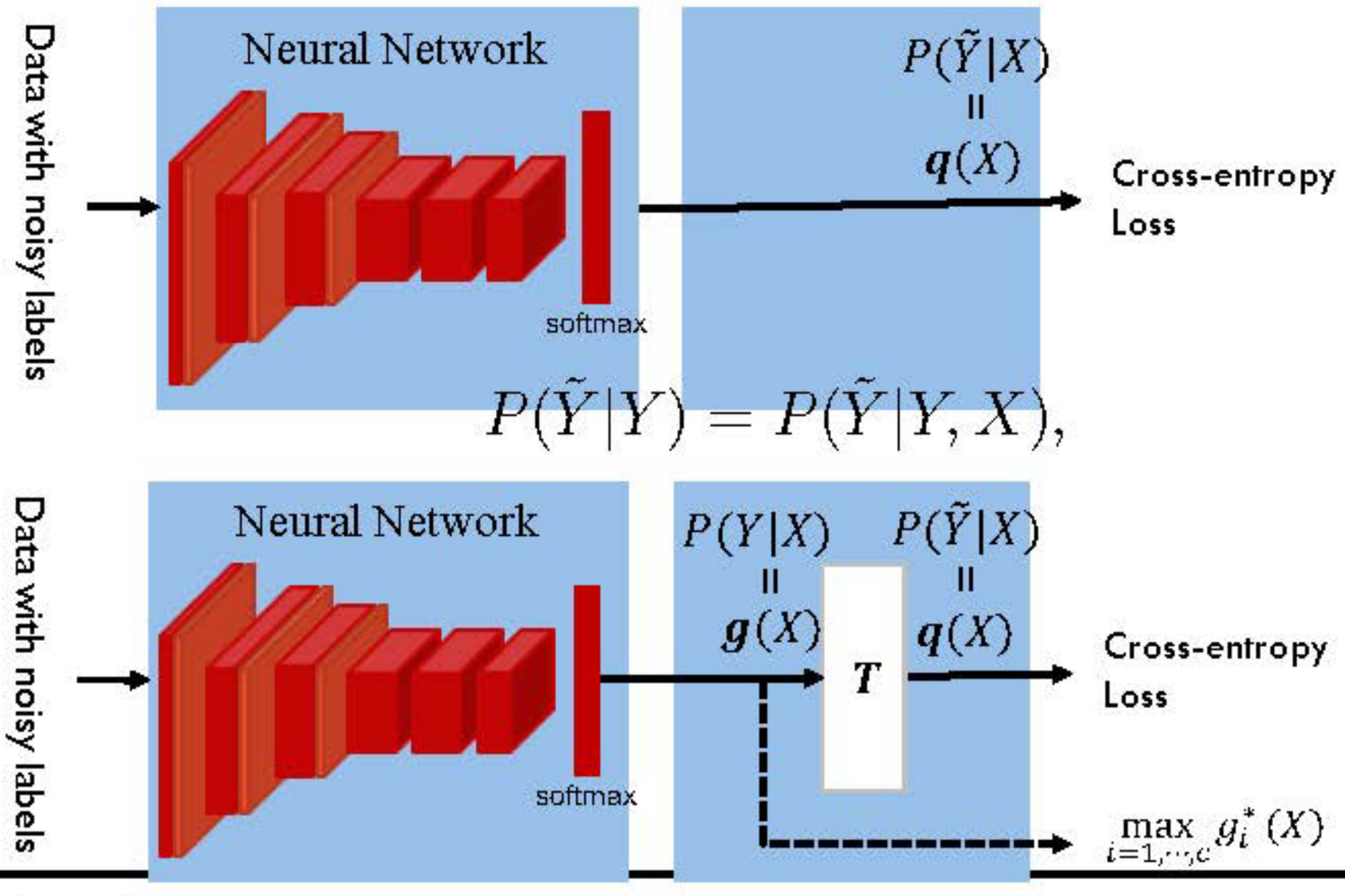
Backward

The above means that we can infer the clean class posterior by employing the noisy class posterior and the inverse transition matrix.

Learning with noisy labels

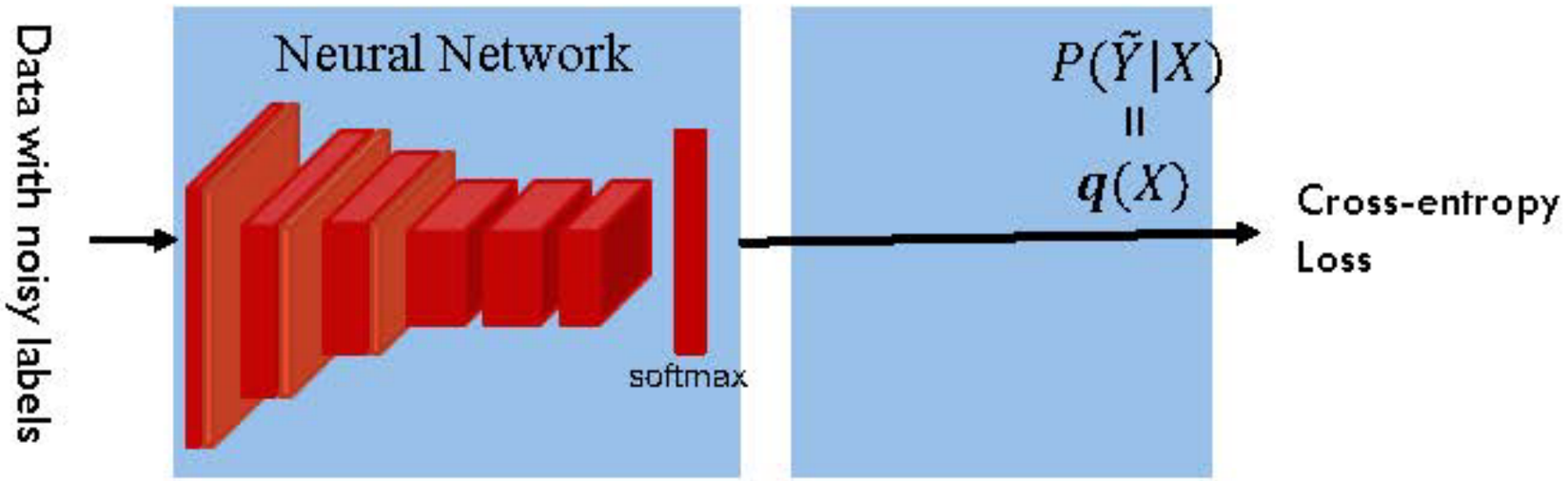
Forward learning:

$$[P(\tilde{Y}=1|X),\ldots,P(\tilde{Y}=C|X)]^\top = T\,[P(Y=1|X),\ldots,P(Y=C|X)]^\top.$$



Learning with noisy labels

Backward learning:



We have that $P(Y|X) = g(X) = T^{-1}q(X)$

because $[P(Y=1|X),\ldots,P(Y=C|X)]^\top = T^{-1}[P(\tilde{Y}=1|X),\ldots,P(\tilde{Y}=C|X)]^\top$.

Instance- and Class-dependent Label Noise

THE UNIVERSITY OF SYDNEY

(3) Instance- and Label-Dependent Noise (ILN):

$$\rho_Y(X) = P\left(\tilde{Y}\middle|Y, X\right).$$

Instance- and Class-dependent Label Noise

THE UNIVERSITY OF SYDNEY

Estimating flip rate for instance- and class-dependent label noise is ill-posed, e.g., if we assume

$$P(\tilde{Y} = 1|X) = 0.8,$$

We may have different possible solutions for $P(Y = 1|X)$ and $\rho(X)$, e.g.,

$$P(Y = 1|X) = 1; \rho_{-1}(X) = \rho_{+1}(X) = 0.2.$$

$$P(Y = 1|X) = 0.875; \rho_{-1}(X) = \rho_{+1}(X) = 0.1.$$

Instance- and Class-dependent Label Noise

THE UNIVERSITY OF SYDNEY

Estimating flip rate for instance- and class-dependent label noise is ill-posed.

Each sample may have its own label-flipping function, resulting in an infinitely large search space that cannot be uniquely determined.

Open problem: Can we make some reasonable assumptions such that the flip rate is identifiable?

Key points

- Learning with Noisy Labels – Problem Setup
- Random Classification Noise
- Class-dependent Label Noise: Binary
- Class-dependent Label Noise: Multi-class
- Instance- and Class-dependent Label Noise