

Advanced Machine Learning

(COMP 5328)

Learning with Noisy Data II: Label Noise

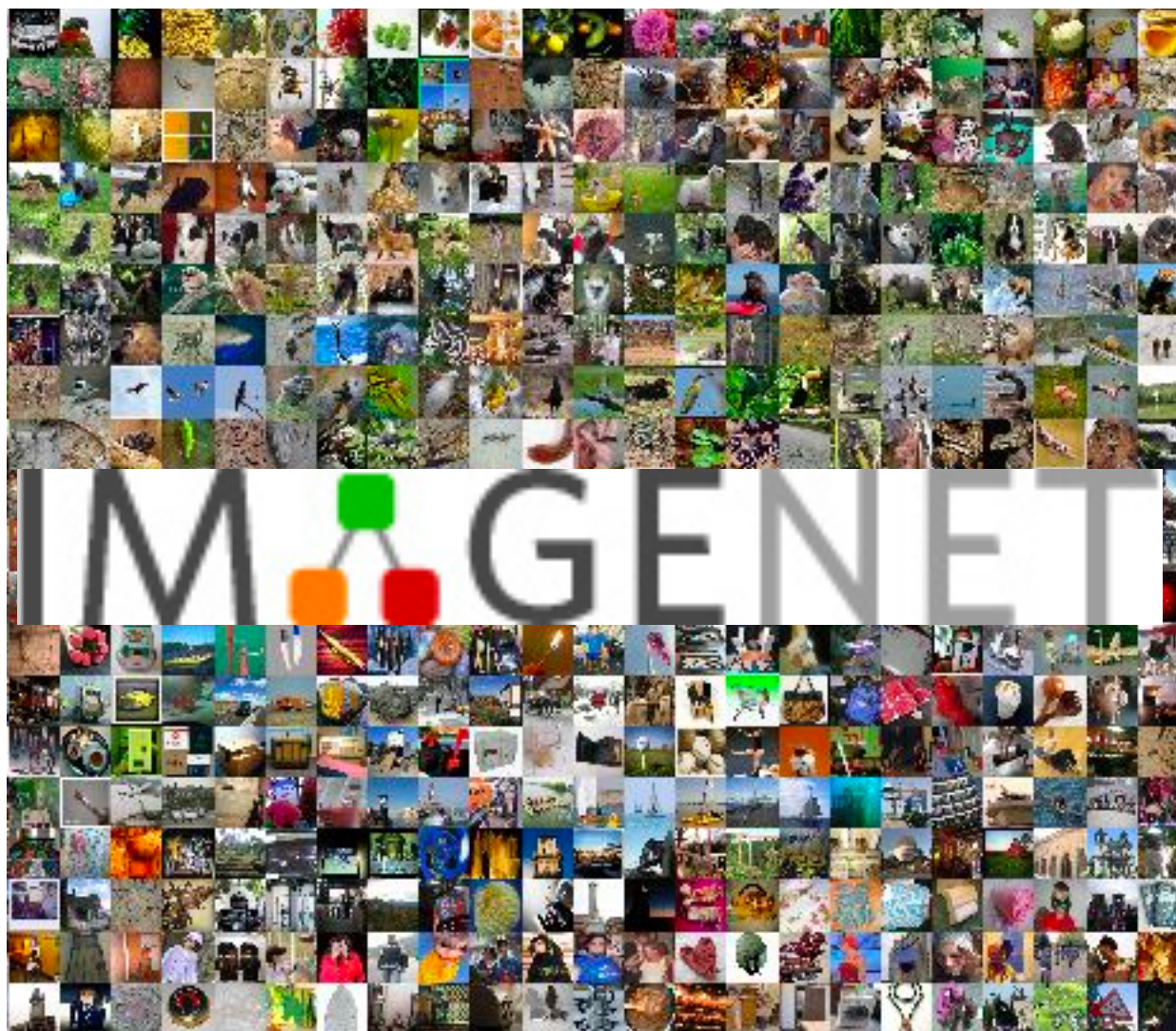
Tongliang Liu

The big data era

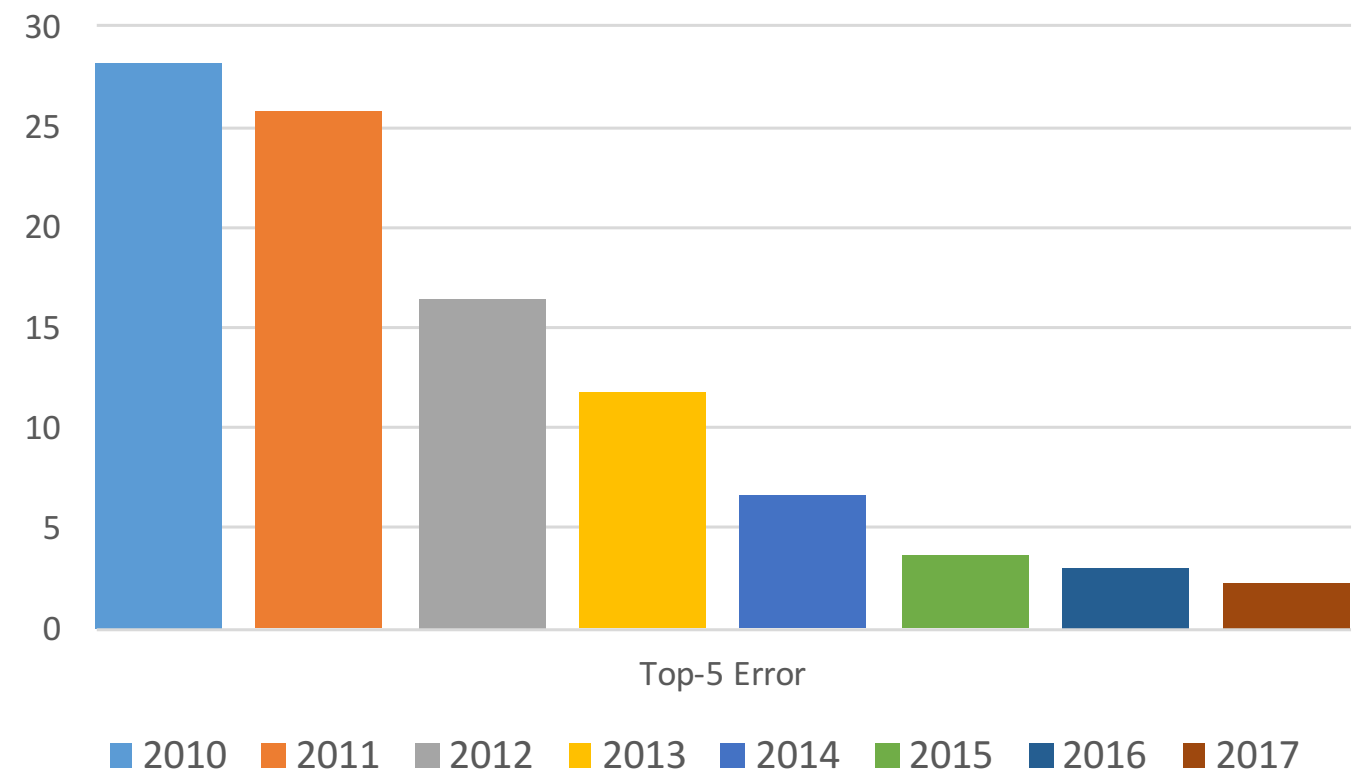


<https://www.zdnet.com/article/big-data-2018-cloud-storage-becomes-the-de-facto-data-lake/>

The big data era

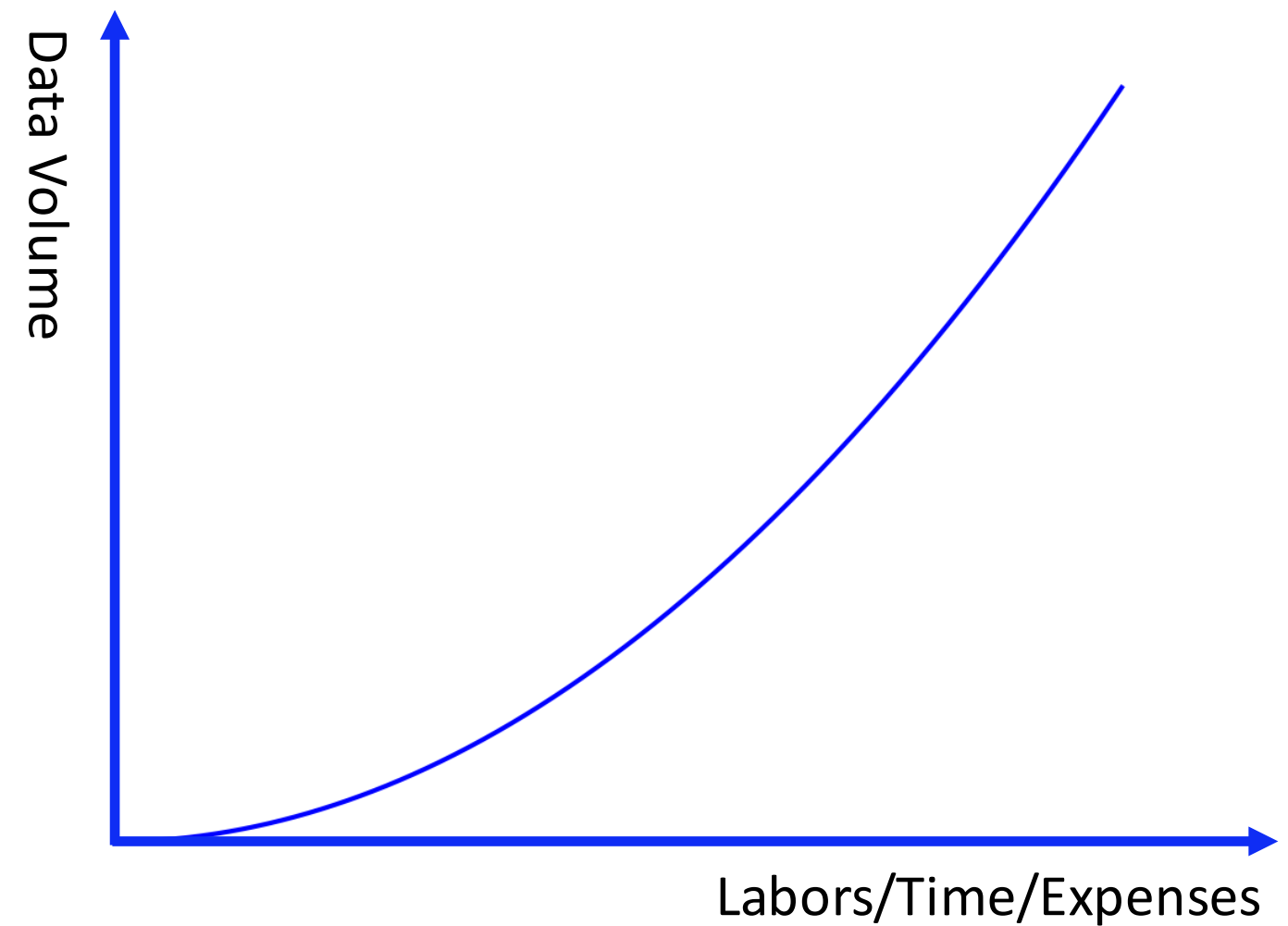


ImageNet Classification Error



Data credit: <http://www.image-net.org/>

The big data era

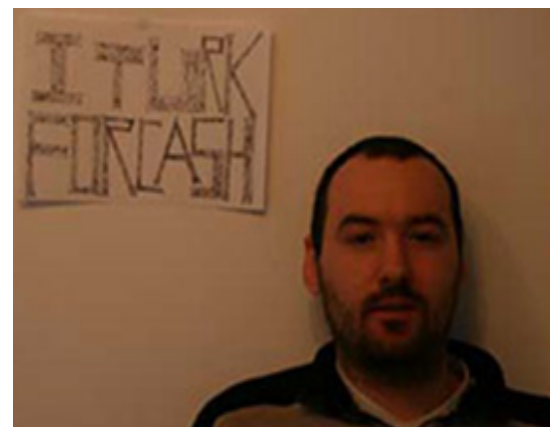
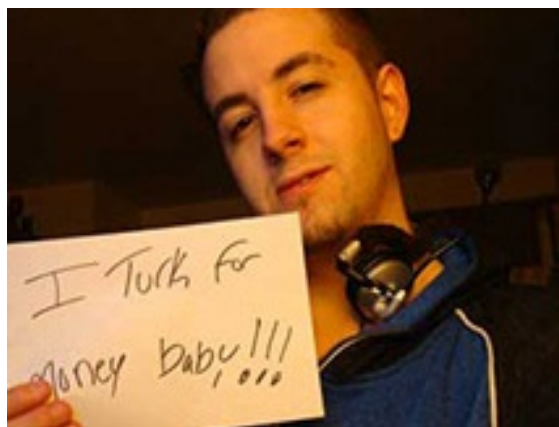


Data credit: <http://www.image-net.org/>

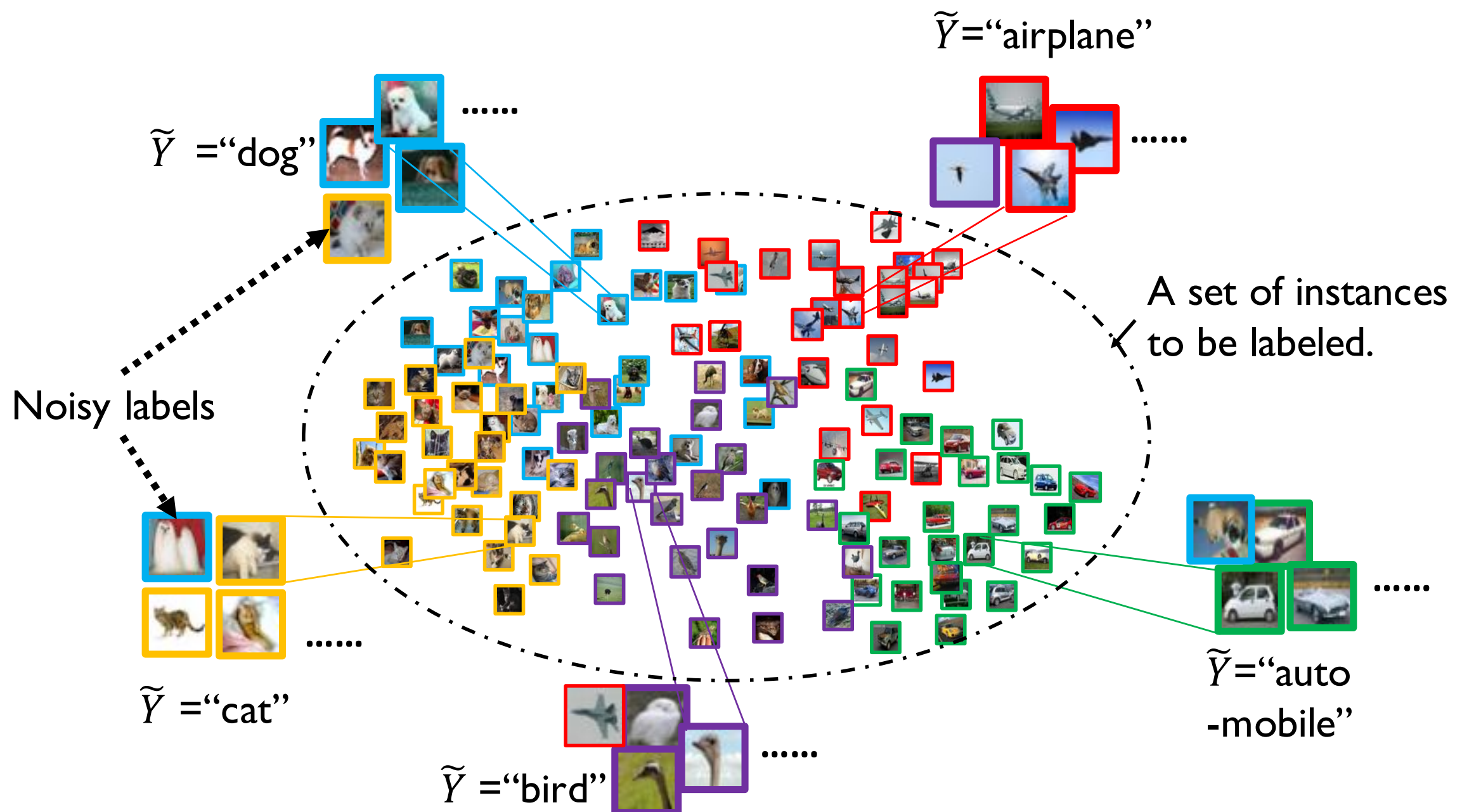
Labelling Costs



Labelling Costs



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.

Label noise widely exists









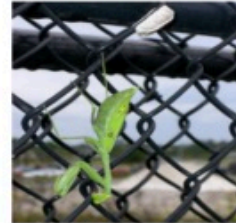
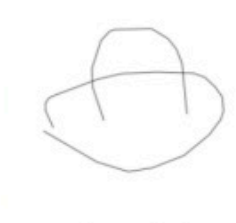
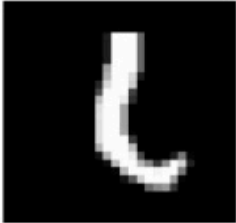



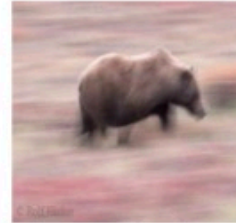





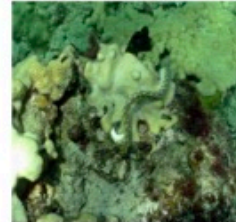

	MNIST	CIFAR-10	CIFAR-100	Caltech-256	ImageNet	QuickDraw
correctable	 given: 5 corrected: 3	 given: cat corrected: frog	 given: lobster corrected: crab	 given: ewer corrected: teapot	 given: white stork corrected: black stork	 given: tiger corrected: eye
multi-label	(N/A)	(N/A)	 given: hamster also: cup	 given: fried egg also: frying pan	 given: mantis also: fence	 given: hat also: flying saucer
neither	 given: 6 alt: 1	 given: deer alt: bird	 given: rose alt: apple	 given: porcupine alt: hot tub	 given: polar bear alt: elephant	 given: pineapple alt: raccoon
non-agreement	 given: 4 alt: 9	 given: deer alt: frog	 given: spider alt: cockroach	 given: minotaur alt: coin	 given: eel alt: flatworm	 given: bandage alt: roller coaster

Image credit: Northcutt et al. "Pervasive label errors in test sets destabilize machine learning benchmarks." arXiv preprint arXiv:2103.14749 (2021).

Real-world problems

- ImageNet dataset has noisy labels.
- The WebVision database and the extremely large scale JFT-300M database have about 20% of images which are believed to be incorrectly labelled.

[1] Li, W., Wang, L., Li, W., Agustsson, E., Berent, J., Gupta, A., Sukthankar, R. and Van Gool, L., 2017. WebVision challenge: visual learning and understanding with web data. arXiv preprint arXiv:1705.05640.

[2] Sun, C., Shrivastava, A., Singh, S. and Gupta, A., 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. arXiv preprint arXiv:1707.02968.

Classification

- Expected 0-1 Risk:

$$R_D(f) = P(\text{sign}(f(X)) \neq Y) = \mathbb{E}_{(X,Y) \sim D} [\mathbf{1}(\text{sign}(f(X)) \neq Y)].$$

- Expected L -Risk:

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

- Empirical risk:

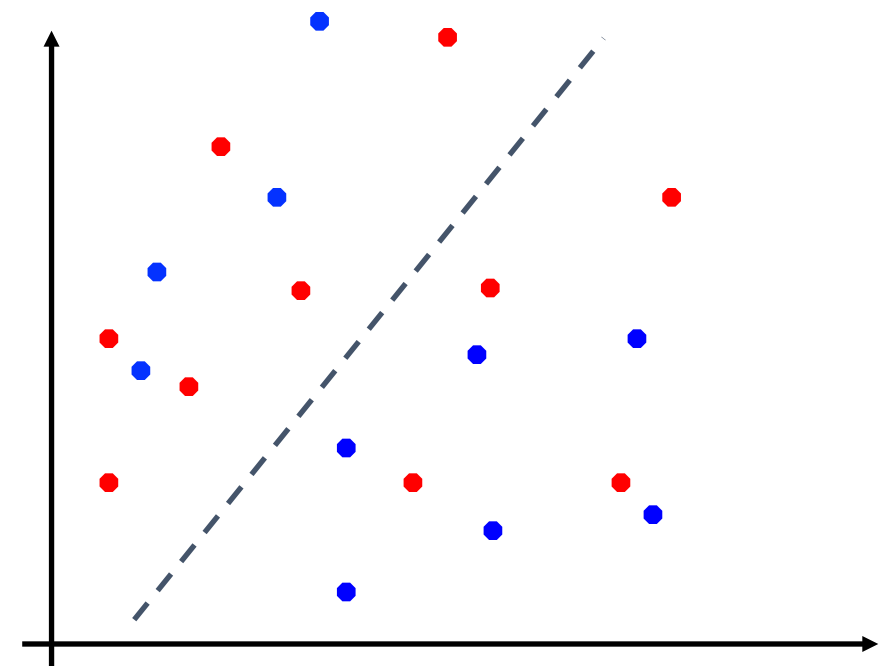
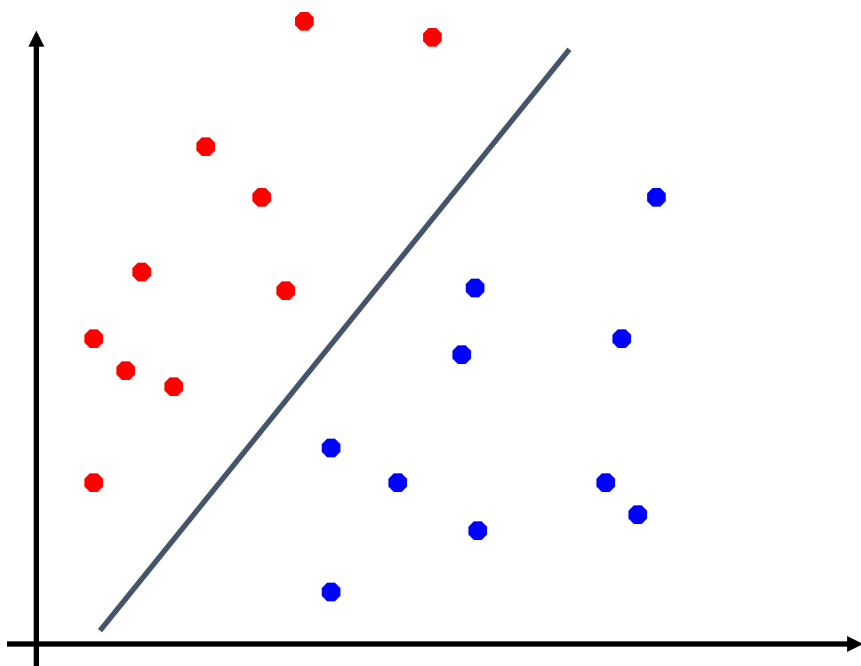
$$R_{D,L,n}(f) = \frac{1}{n} \sum_{i=1}^n L(f(X_i), Y_i).$$



Learning with noisy labels

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

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



Learning with noisy labels

Problem Setup

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

Problem Setup

- Given the training examples $\{(X_i, \tilde{Y}_i)\}_{1 \leq i \leq n} \sim D_\rho(X, \tilde{Y})^n$.
- The target is to learn a discriminant function $f_n: \mathcal{X} \rightarrow \mathbb{R}$ such that the classifier predicts the correct label y given an observation x .

Model Label Noise

A probabilistic model:

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

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

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

otherwise

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

Model Label Noise

(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):

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

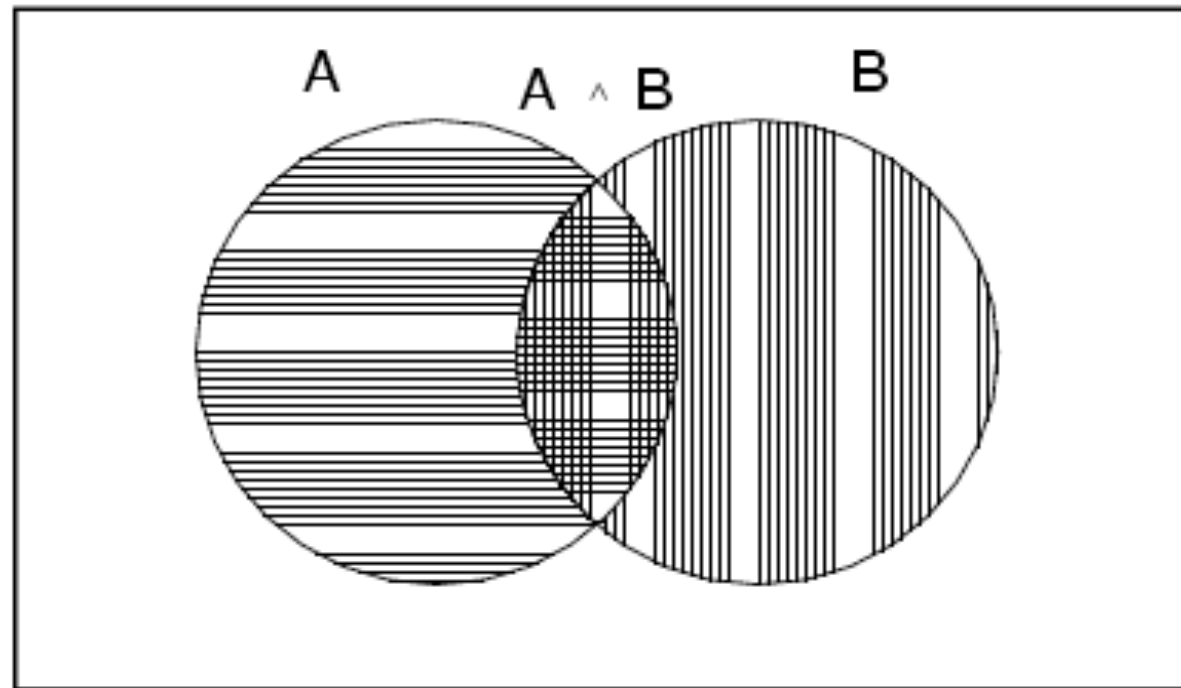
(3) Instance- and Label-Dependent Noise (ILN):

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

Basics in Probability Theory

Axioms of probability

True



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) \equiv P(AB) \equiv P(A, B)$$

$$P(A^c) = 1 - P(A)$$

Rules of Probability

- Sum Rule $p(X) = \sum_Y p(X, Y)$
- Product Rule $p(X, Y) = p(Y|X)p(X)$

Random Classification Noise

Random Classification Noise (RCN)

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 $\phi: \mathbb{R} \rightarrow \mathbb{R}$ is convex, non-increasing, differentiable with $\phi'(0) < 0$, and $\phi(+\infty) = 0$.

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

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)

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.

Random Classification Noise (RCN)

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 I. 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)

We will prove that

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

Random Classification Noise (RCN)

Proof

$$\begin{aligned} P(\tilde{Y} = 1|X) &= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X) \\ &= P(\tilde{Y} = 1|Y = 1, X)P(Y = 1|X) + P(\tilde{Y} = 1|Y = -1, X)P(Y = -1|X) \\ &= (1 - \rho_{+1}(X))P(Y = 1|X) + \rho_{-1}(X)P(Y = -1|X) \\ &= (1 - \rho_{+1}(X) - \rho_{-1}(X))P(Y = 1|X) + \rho_{-1}(X). \end{aligned}$$

Random Classification Noise (RCN)

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)

Proof Cont'd

$$\begin{aligned} R_{D_\rho, L}(f) &= \mathbb{E}_{(X, \tilde{Y}) \sim D_\rho} [L(f(X), \tilde{Y})] \\ &= \int \left(P(\tilde{Y} = 1, X) L(f(X), 1) + P(\tilde{Y} = -1, X) L(f(X), -1) \right) dX \\ &= \int \left(P(\tilde{Y} = 1|X) P(X) L(f(X), 1) + P(\tilde{Y} = -1|X) P(X) L(f(X), -1) \right) dX \\ &= \int P(X) [(1 - 2\rho) P(Y = 1|X) L(f(X), 1) + \rho L(f(X), 1)] dX \\ &\quad + \int P(X) [(1 - 2\rho) P(Y = -1|X) L(f(X), -1) + \rho L(f(X), -1)] dX \\ &= (1 - 2\rho) \int \left(P(Y = 1, X) L(f(X), 1) + P(Y = -1, X) L(f(X), -1) \right) dX \\ &\quad + \rho \int P(X) [\ell(h(X), +1) + \ell(h(X), -1)] dX \end{aligned}$$

Random Classification Noise (RCN)

Proof Cont'd

$$\begin{aligned} R_{D_\rho, L}(f) &= \mathbb{E}_{(X, \tilde{Y}) \sim D_\rho} [L(f(X), \tilde{Y})] \\ &= (1 - 2\rho) \int (P(Y = 1, X)L(f(X), 1) + P(Y = -1, X)L(f(X), -1))dX \\ &\quad + \rho \int P(X) [\ell(h(X), +1) + \ell(h(X), -1)]dX \\ &= (1 - 2\rho) \mathbb{E}_{(X, Y) \sim D} [L(f(X), Y)] + \rho C \\ &= (1 - 2\rho) R_{D, L}(f) + \rho C. \quad \text{Completed!} \end{aligned}$$

Random Classification Noise (RCN)

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

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



THE UNIVERSITY OF
SYDNEY

Class-dependent Label Noise: Binary

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

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

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

$$\begin{aligned} R_{D,L}(f) &= \mathbb{E}_{(X,Y) \sim D} [L(f(X), Y)] = \int P_D(X, Y) L(f(X), Y) dX dY \\ &= \int P_{D_\rho}(X, Y) \frac{P_D(X, Y)}{P_{D_\rho}(X, Y)} L(f(X), Y) dX dY \\ &= \mathbb{E}_{(X,Y) \sim D_\rho} \left[\frac{P_D(X, Y)}{P_{D_\rho}(X, Y)} L(f(X), Y) \right] \\ &= \mathbb{E}_{(X,Y) \sim D_\rho} [\beta(X, Y) L(f(X), Y)] \quad \text{where } \beta(\mathbf{x}, y) = \frac{P_D(X=\mathbf{x}, Y=y)}{P_{D_\rho}(X=\mathbf{x}, \tilde{Y}=y)}. \end{aligned}$$

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

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

Recall that

$$P_{D_\rho}(\tilde{Y} = y | X = \mathbf{x}) = (1 - \rho_{+1} - \rho_{-1})P_D(Y = y | X = \mathbf{x}) + \rho_{-y}$$

Then

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

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

Noise rate estimation

Recall that

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

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

We also assume that the flip rate is small such that

$$\rho_{+1} + \rho_{-1} \leq 1.$$

We have

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

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

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

Noise rate estimation

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

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

↓
0

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

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

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

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

We designed the following estimator: $\rho_{-y} = \min_{X \in \mathcal{X}} P(\tilde{Y} = y | X)$

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

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

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

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:

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

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

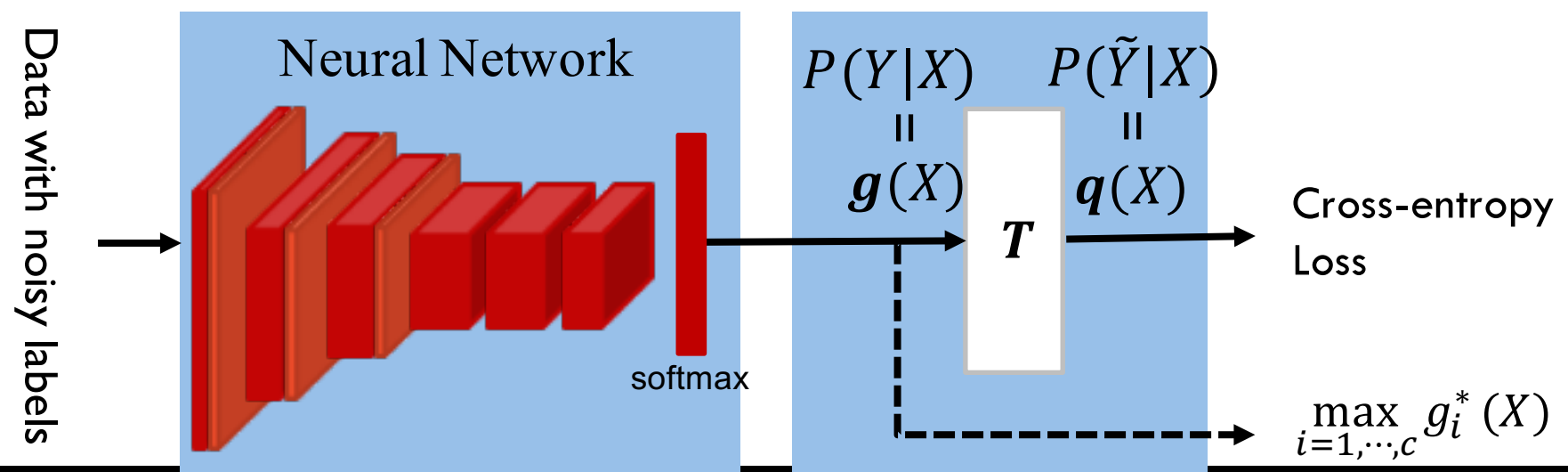
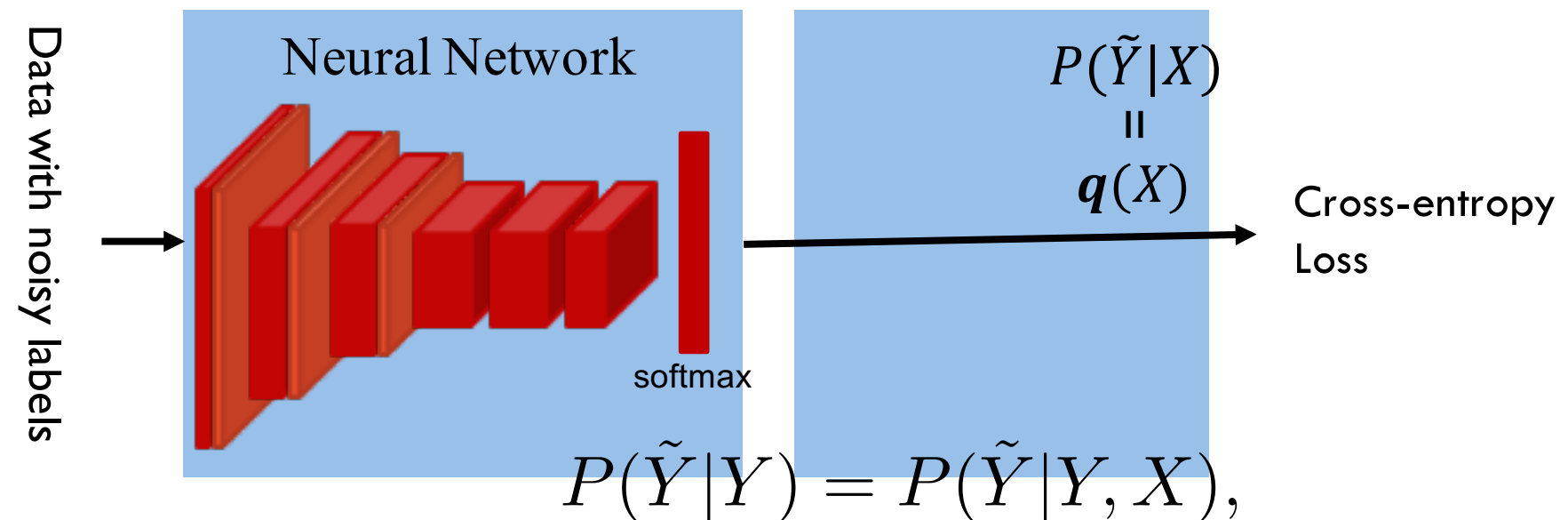
Backward

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

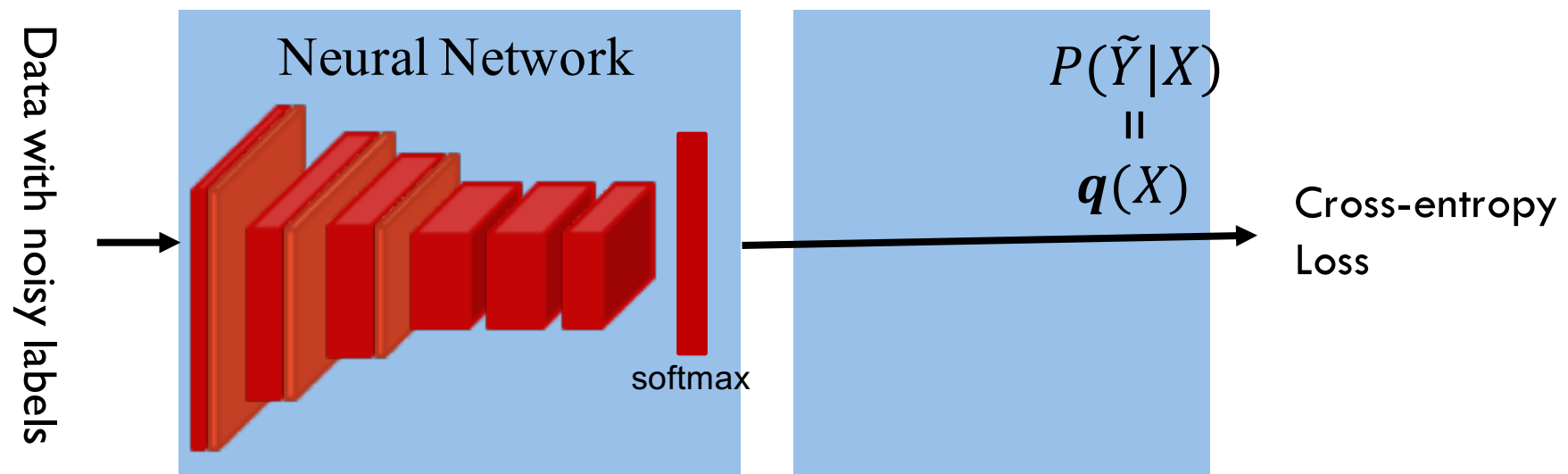
Forward learning:

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



Learning with noisy labels

Backward learning:



We have that

$$P(Y|X) = \mathbf{g}(X) = T^{-1}\mathbf{q}(X)$$

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

Instance- and Class- dependent Label Noise

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

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

Relationship: noisy data, flip rates, and clean data

$$\begin{aligned} P(\tilde{Y} = 1|X) &= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X) \\ &= P(\tilde{Y} = 1|Y = 1, X)P(Y = 1|X) + P(\tilde{Y} = 1|Y = -1, X)P(Y = -1|X) \\ &= (1 - \rho_{+1}(X))P(Y = 1|X) + \rho_{-1}(X)P(Y = -1|X) \\ &= (1 - \rho_{+1}(X) - \rho_{-1}(X))P(Y = 1|X) + \rho_{-1}(X). \end{aligned}$$

Instance- and Class- dependent Label Noise

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

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

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