Positive-Unlabeled Learning with Non-Negative Risk Estimator

Ryuichi Kiryo[1,2]   Gang Niu[1,2]   Martins Christoffel du Plessis   Masashi Sugiyama[2,1]

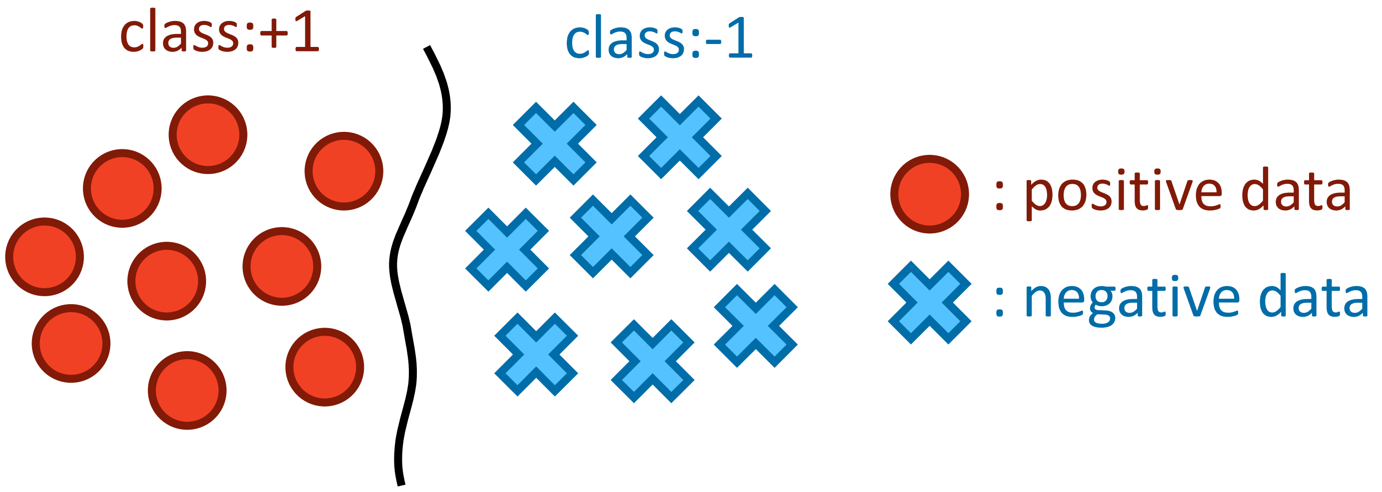
1 東京大学 THE UNIVERSITY OF TOKYO

2 RIKEN

Binary classification

- Classify input data $X \in \mathbb{R}^d$ to class $Y \in \{+1 \text{ (positive)}, -1 \text{(negative)}\}$

- Supervised learning:
  Classifier $g: \mathbb{R}^d \to \mathbb{R}$ is learned from positive data and negative data



Supervised binary classification (PN learning)

- **Goal**: minimize expected risk

$$R(g) = \mathbb{E}_{(X,Y)\sim p(x,y)}[l(g(X),Y)] = \underbrace{\pi_\mathrm{p}\mathbb{E}_\mathrm{p}[l(g(X),+1)]}_{\text{risk for positive class}} + \underbrace{\pi_\mathrm{n}\mathbb{E}_\mathrm{n}[l(g(X),-1)]}_{\text{risk for negative class}}$$

- Minimize empirical risk: approximation by data in hand

$$\hat{R}_\mathrm{pn}(g) = \frac{\pi_\mathrm{p}}{n_\mathrm{p}}\sum_{x\in\mathcal{X}_\mathrm{p}} l(g(x),+1) + \frac{\pi_\mathrm{n}}{n_\mathrm{n}}\sum_{x\in\mathcal{X}_\mathrm{n}} l(g(x),-1)$$

input data $X \in \mathbb{R}^d$
class label $Y \in \{\pm 1\}$
loss function $l: \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$

$\mathbb{E}_\mathrm{p}[\cdot] := \mathbb{E}_{X\sim p(x|Y=+1)}[\cdot]$
$\mathbb{E}_\mathrm{n}[\cdot] := \mathbb{E}_{X\sim p(x|Y=-1)}[\cdot]$

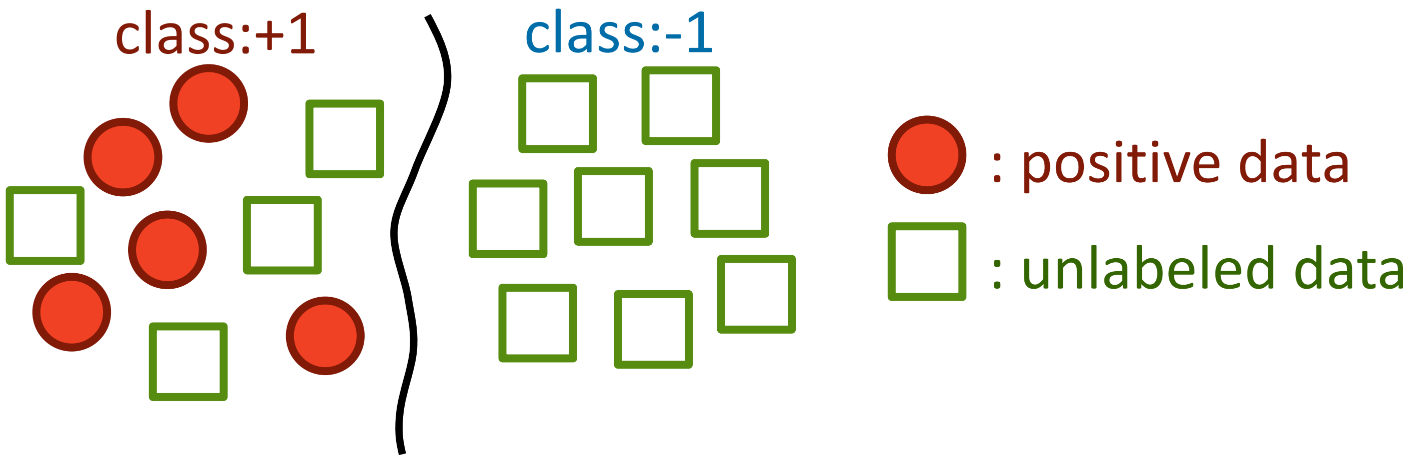
$\pi_\mathrm{p} := p(Y=+1)$
$\pi_\mathrm{n} := p(Y=-1)$

$\mathcal{X}_\mathrm{p} = \{x_i^\mathrm{p}\}_{i=1}^{n_\mathrm{p}} \overset{i.i.d.}{\sim} p(x|Y=+1)$
$\mathcal{X}_\mathrm{n} = \{x_i^\mathrm{n}\}_{i=1}^{n_\mathrm{n}} \overset{i.i.d.}{\sim} p(x|Y=-1)$

Classification when negative data is unavailable

- Example: click advertisement
  - Clicked : positive
  - Non-clicked : unlabeled (not interesting or unseen)

Learn a PN binary classifier from positive and unlabeled data (PU learning)



Unbiased PU learning [Natarajan+, NIPS 2013, du Plessis+, ICML 2015]

- **Goal**: minimize the same expected risk as PN learning

$$R(g) = \pi_{\mathrm{p}} \mathbb{E}_{\mathrm{p}}[l(g(X), +1)] + \underline{\pi_{\mathrm{n}} \mathbb{E}_{\mathrm{n}}[l(g(X), -1)]}$$

negative data unavailable

- **Idea**: unlabeled data = positive data + negative data

$$\mathbb{E}_{\mathrm{u}}[l(g(X), -1)] = \pi_{\mathrm{p}} \mathbb{E}_{\mathrm{p}}[l(g(X), -1)] + \pi_{\mathrm{n}} \mathbb{E}_{\mathrm{n}}[l(g(X), -1)]$$

- Risk can be expressed by only **positive** and **unlabeled** data

$$R_{\mathrm{pu}}(g) = \pi_{\mathrm{p}} \mathbb{E}_{\mathrm{p}}[l(g(X), +1)] + \mathbb{E}_{\mathrm{u}}[l(g(X), -1)] - \pi_{\mathrm{p}} \mathbb{E}_{\mathrm{p}}[l(g(X), -1)]$$

$$\mathbb{E}_{\mathrm{u}}[\cdot] = \mathbb{E}_{X \sim p(x)}[\cdot], \qquad \mathcal{X}_{\mathrm{u}} = \{x_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}} \overset{i.i.d.}{\sim} p(x) \coloneqq \pi_{\mathrm{p}} p(x|Y=+1) + \pi_{\mathrm{n}} p(x|Y=-1)$$

Theoretical properties of unbiased PU learning

[du Plessis+, NIPS 2014, ICML 2015; Niu+, NIPS 2016]

- Risk estimator is **unbiased**

$$\mathbb{E}[\hat{R}_{\mathrm{pu}}(g)] = R_{\mathrm{pu}}(g) = R(g)$$

- For linear-in-parameter models, estimation error vanishes in the **optimal** parametric rate

$$\mathcal{O}_p\left(\frac{\pi_{\mathrm{p}}}{\sqrt{n_{\mathrm{p}}}} + \frac{1}{\sqrt{n_{\mathrm{u}}}}\right)$$

cf. PN learning $\mathcal{O}_p\left(\frac{\pi_{\mathrm{p}}}{\sqrt{n_{\mathrm{p}}}} + \frac{\pi_{\mathrm{n}}}{\sqrt{n_{\mathrm{n}}}}\right)$

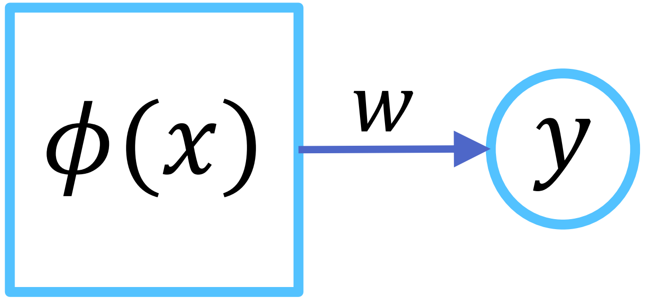
- PU learning can be **better** than PN learning if

$$\frac{\pi_{\mathrm{p}}}{\sqrt{n_{\mathrm{p}}}} + \frac{1}{\sqrt{n_{\mathrm{u}}}} < \frac{\pi_{\mathrm{n}}}{\sqrt{n_{\mathrm{n}}}}$$

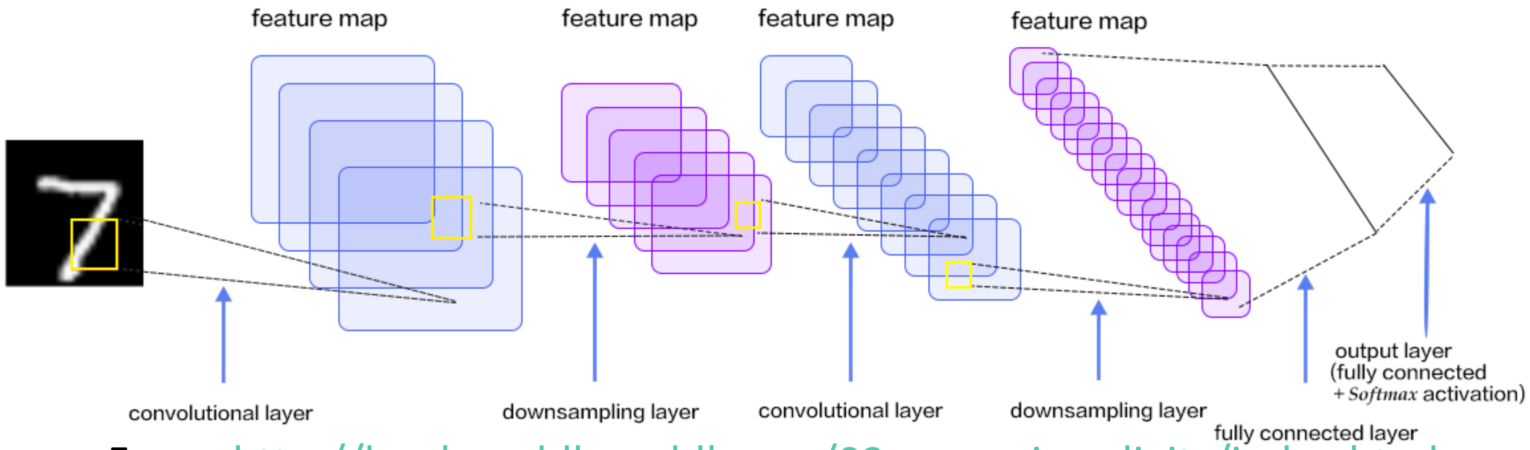
$\pi_{\mathrm{p}} := p(Y = +1), \pi_{\mathrm{n}} := p(Y = -1)$

Unbiased PU learning works well in linear-in-parameter models experimentally.

[du Plessis+, NIPS 2014, ICML 2015; Niu+, NIPS 2016]

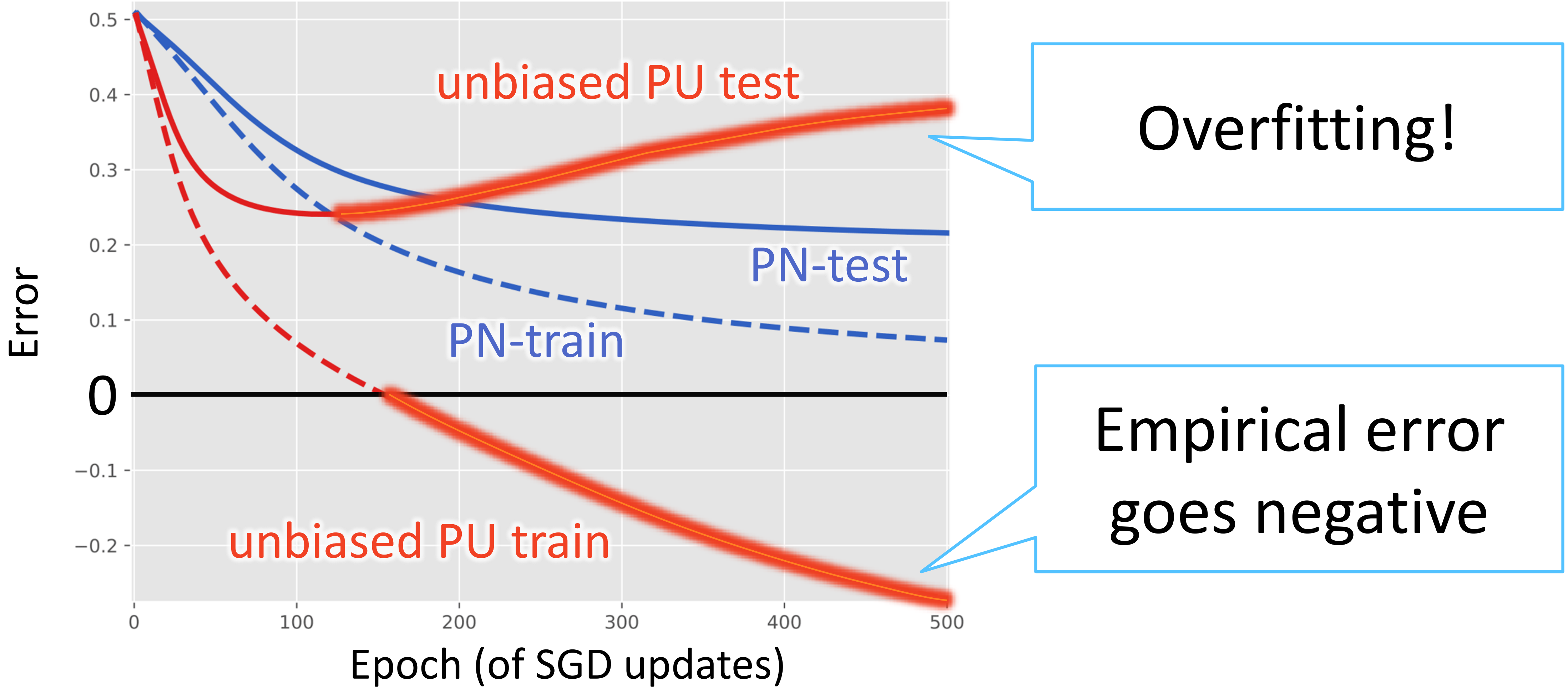


How about flexible models, like *deep nets*?



From: http://book.paddlepaddle.org/02.recognize_digits/index.html

Unbiased PU learning with flexible model



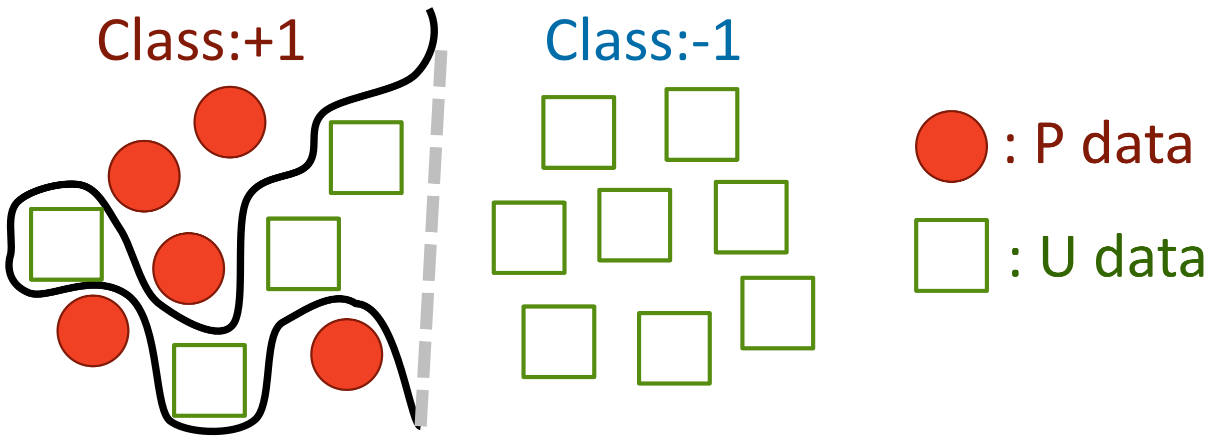
- Classify even digits and odd digits of MNIST by 3-layer MLP
- n_{p}=100, n_{n}=50, n_{u}=59900

Overfitting and negative risk of unbiased PU

- If classifier can perfectly separate P and U, training error w.r.t. 0-1 loss is:

$$\frac{\pi_{\mathrm{p}}}{n_{\mathrm{p}}}\sum_{x\in\mathcal{X}_{\mathrm{p}}} l(g(x),+1) + \frac{1}{n_{\mathrm{u}}}\sum_{x\in\mathcal{X}_{\mathrm{u}}} l(g(x),-1) - \frac{\pi_{\mathrm{p}}}{n_{\mathrm{p}}}\sum_{x\in\mathcal{X}_{\mathrm{p}}} l(g(x),-1)$$

$$\underbrace{\pi_p \cdot 0}_{\text{risk for positive class}} + \underbrace{0 - \pi_{\mathrm{p}} \cdot 1}_{\text{risk for negative class}} < 0$$



PU learning with non-negative risk estimator (non-negative PU learning)

- **Idea:**

Round-up risk for negative class to zero

- We propose **the new risk estimator** which is always **non-negative**

$$\tilde{R}_{\mathrm{pu}(g)} = \frac{\pi_{\mathrm{p}}}{n_{\mathrm{p}}} \sum_{x \in \mathcal{X}_{\mathrm{p}}} l(g(x), +1) + \mathbf{max}\left\{ \mathbf{0}, \frac{1}{n_{\mathrm{u}}} \sum_{x \in \mathcal{X}_{\mathrm{u}}} l(g(x), -1) - \frac{\pi_{\mathrm{p}}}{n_{\mathrm{p}}} \sum_{x \in \mathcal{X}_{\mathrm{p}}} l(g(x), -1) \right\}$$

classifier $g: \mathbb{R}^d \to \mathbb{R}$
loss function $l: \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$

$\pi_{\mathrm{p}} \coloneqq p(Y = +1)$
$\pi_{\mathrm{n}} \coloneqq p(Y = -1)$
$\mathbb{E}_{\mathrm{n}}[\cdot] \coloneqq \mathbb{E}_{X \sim p(x|Y=-1)}[\cdot]$

positive dataset $\mathcal{X}_{\mathrm{p}} = \{x_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p}}} \overset{i.i.d.}{\sim} p(x|Y = +1)$
unlabeled dataset $\mathcal{X}_{\mathrm{u}} = \{x_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}} \overset{i.i.d.}{\sim} p(x)$

Theoretical analysis of non-negative PU learning

- Risk estimator is **consistent** and its bias decreases **exponentially**
  - Bias is negligible in practice

$$\mathcal{O}_p\left(\exp\left(-1/\left(\pi_{\mathrm{p}}^2/n_{\mathrm{p}} + 1/n_{\mathrm{u}}\right)\right)\right)$$

- Risk estimator may reduce mean squared error
  - Non-negative risk estimator is **more stable**

$$\mathbb{E}_{\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}}\left[\left(\tilde{R}_{\mathrm{pu}}(g) - R(g)\right)^2\right] \le \mathbb{E}_{\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}}\left[\left(\hat{R}_{\mathrm{pu}}(g) - R(g)\right)^2\right]$$

non-negative PU — unbiased PU

- For linear-in-parameter models, estimation error vanishes in the **optimal** parametric rate

$$\mathcal{O}_p\left(\frac{\pi_{\mathrm{p}}}{\sqrt{n_{\mathrm{p}}}} + \frac{1}{\sqrt{n_{\mathrm{u}}}}\right)$$

Large-scale algorithm

- Want to use mini-batch SGD

$(\mathcal{X}_{\mathrm{p}}^{i}, \mathcal{X}_{\mathrm{u}}^{i})$: i-th mini-batch ($i = 1, \dots, N$)

- Our objective function

$$\frac{\pi_{\mathrm{p}}}{n_{\mathrm{p}}} \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_{p}^{i}} l(g(x), +1) + \max\left\{0, \sum_{i}^{N} \left(\frac{1}{n_u} \sum_{x \in \mathcal{X}_{u}^{i}} l(g(x), -1) - \frac{\pi_p}{n_p} \sum_{x \in \mathcal{X}_{p}} l(g(x), -1) \right) \right\}$$

$\leq$ Still, upper bound can be minimized!

- Sum of risks in mini-batches

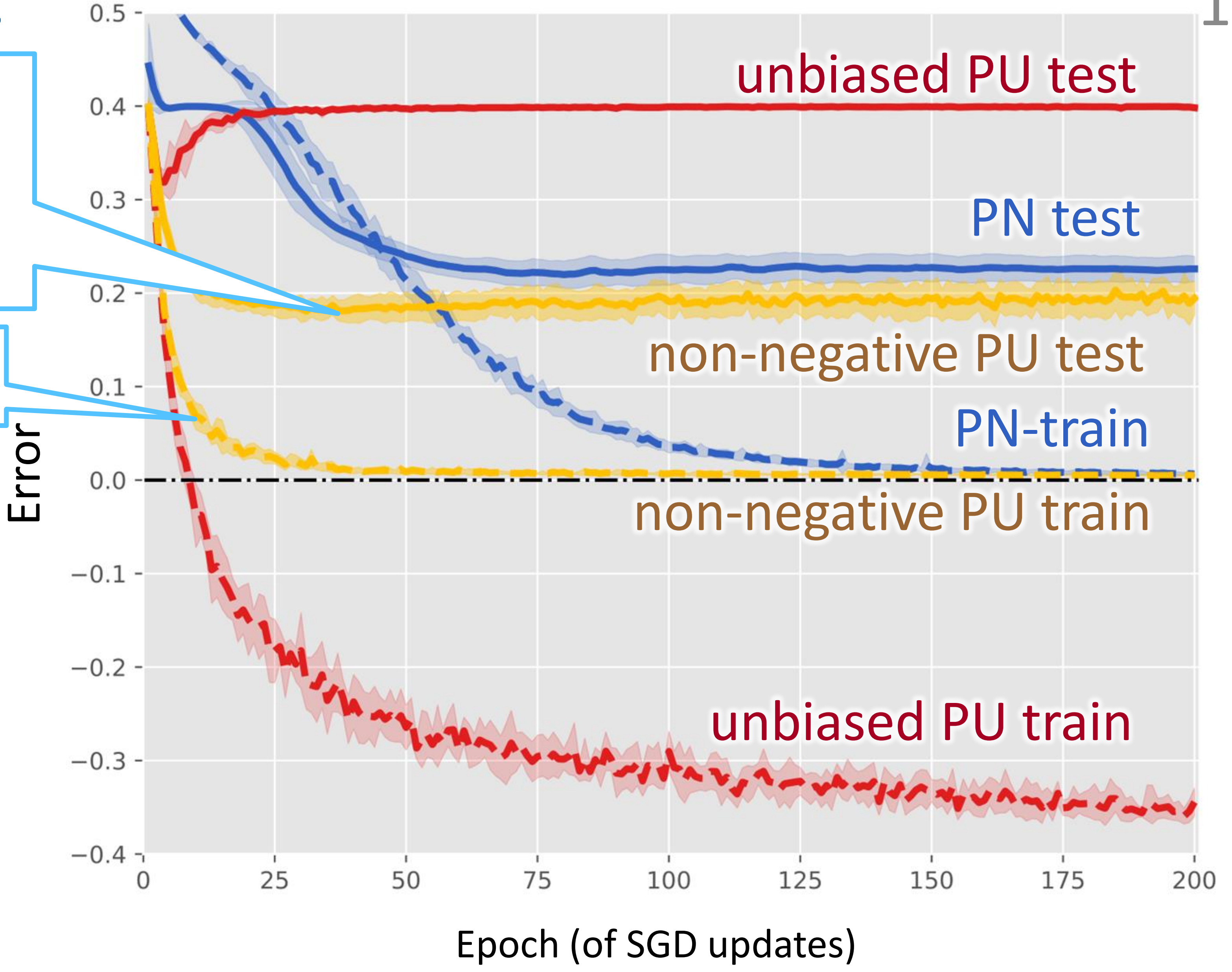
$$\sum_{i=1}^{N} \left[\frac{\pi_{\mathrm{p}}}{n_{\mathrm{p}}} \sum_{x \in \mathcal{X}_{\mathrm{p}}^{i}} l(g(x), +1) + \mathbf{max} \left\{ \mathbf{0}, \frac{1}{n_{\mathrm{u}}} \sum_{x \in \mathcal{X}_{\mathrm{u}}^{i}} l(g(x), -1) - \frac{\pi_{\mathrm{p}}}{n_{\mathrm{p}}} \sum_{x \in \mathcal{X}_{\mathrm{p}}} l(g(x), -1) \right\} \right]$$

CIFAR10 experiment

Achieved smaller test error than unbiased PU and even PN!

Training error → 0

- P=artifacts
- N=animals
- 13-layer CNN [Springenberg+, ICLR2015]
- $n_{\mathrm{p}} = 1000$
- $n_{\mathrm{n}} = 562$
- $n_{\mathrm{u}} = 50000$



Conclusions

- We proposed a non-negative risk estimator for PU learning which improves on the state-of-the-art unbiased risk estimators.
- The new risk estimator is more robust against overfitting, and training very flexible model given limited P data becomes possible.
- A large-scale PU learning algorithm was also developed.
- Extensive theoretical analyses were presented.
- Intensive experiments were carried out as well.

Reference

- Natarajan, Nagarajan, et al. "Learning with noisy labels." *Advances in neural information processing systems*. 2013.
- M. C. du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In *NIPS*, 2014.
- M. C. du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *ICML*, 2015.
- G. Niu, M. C. du Plessis, T. Sakai, Y. Ma, and M. Sugiyama. Theoretical comparisons of positive unlabeled learning against positive-negative learning. In *NIPS*, 2016.
- J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. 2015.