OSL2019, Les Houches, France March 25, 2019

Weakly Supervised Classification, Reliable Machine Learning, and More

---Overview of Our Recent Advances---

Masashi Sugiyama

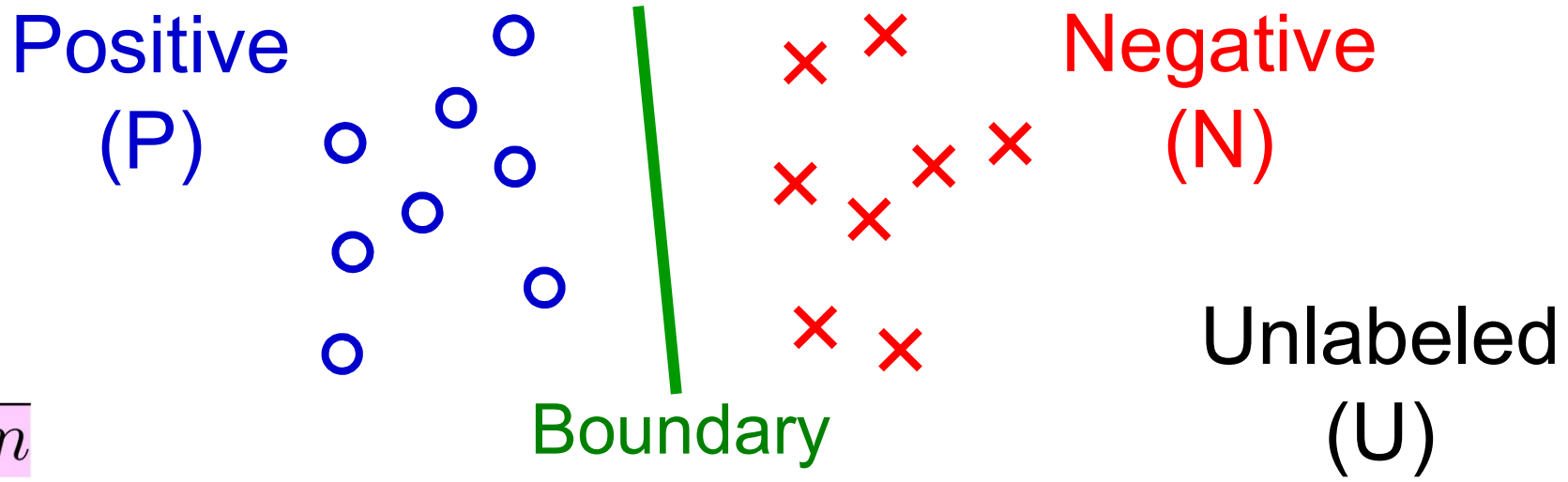
RIKEN Center for Advanced Intelligence Project/
The University of Tokyo

My Talk

1. Weakly supervised classification
2. Reliable machine learning
3. More

Weakly Supervised Learning

- Collecting labeled data is often expensive:
  - Improve data collection (e.g., crowdsourcing)
  - Use a simulator to generate pseudo data
  - Use domain knowledge (i.e., engineering)
  - Use cheap but weak data (e.g., unlabeled)
- Let's focus on binary classification:

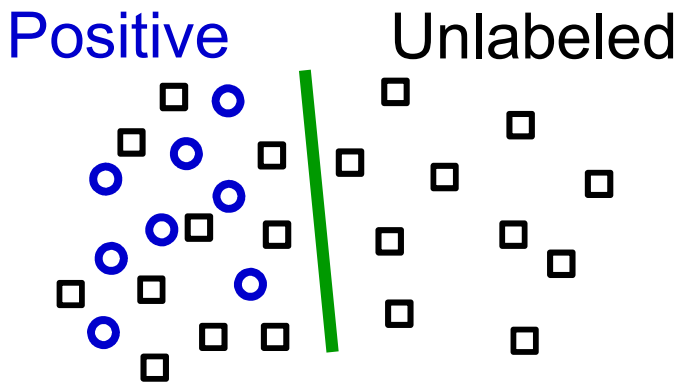


$1/\sqrt{n}$

(1) PU Classification

du Plessis, Niu & Sugiyama (NIPS2014, ICML2015)
Niu, du Plessis, Sakai, Ma & Sugiyama (NIPS2016)
Kiryo, Niu, du Plessis & Sugiyama (NIPS2017)

- Only PU data is available; N data is missing:
  - Click vs. non-click
  - Friend vs. non-friend



- We want to minimize the risk of classifier f:

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\Big[\ell\Big(yf(\boldsymbol{x})\Big)\Big]$$

$$\pi = p(y=+1)$$

$$= \pi\mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big] + (1-\pi)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]$$

Risk for P data

Risk for N data

- But N-risk cannot be estimated directly.

Key Trick

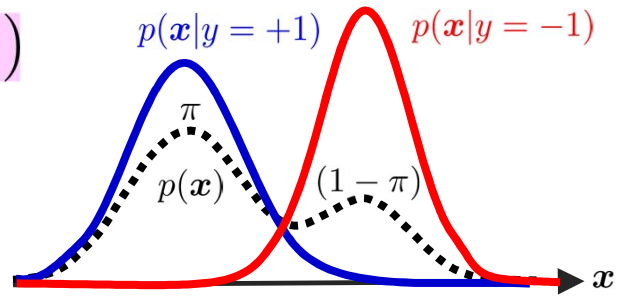
$$R(f) = \pi \mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big] + (1-\pi)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]$$

Risk for P data — Risk for N data

- Use "U-density is mixture of P- and N-densities":

$$p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=+1) + (1-\pi)p(\boldsymbol{x}|y=-1)$$

$$\pi = p(y=+1)$$



- Then

$$R(f) = \pi \mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big]$$

$$+\mathbb{E}_{p(\boldsymbol{x})}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big] - \pi \mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]$$

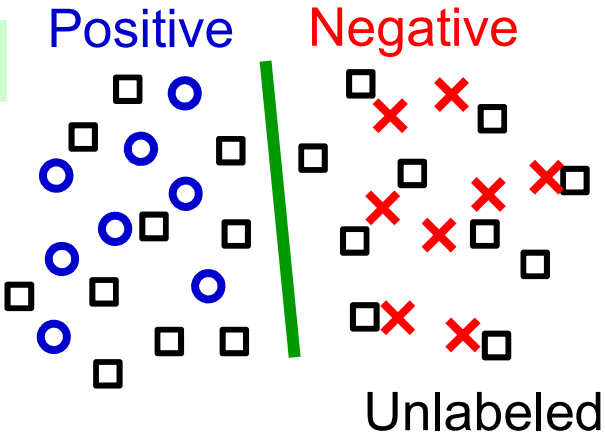
- Empirical risk minimization is possible from PU data, just by replacing expectations by sample averages!

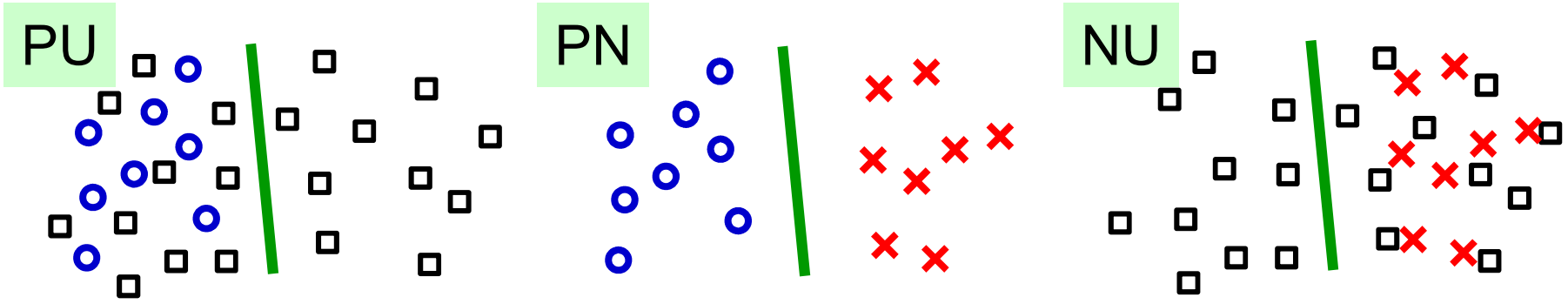
$$R(\widehat{f}_{\mathrm{PU}}) - R(f^*) \leq C(\delta)\left(\frac{2\pi}{\sqrt{n_{\mathrm{P}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right)$$

(2) PNU Classification

Sakai, du Plessis, Niu & Sugiyama (ICML2017), Sakai, Niu & Sugiyama (MLJ2018)

- PNU is semi-supervised classification.
- Let's decompose PNU into PU, PN, and NU:
  - Each is solvable.
  - Let's linearly combine them!
- Without cluster assumptions, PN classifiers are trainable!





$$R_{0/1}(f) \leq 2\widehat{R}^{\gamma}_{\mathrm{PN+PU}}(f) + \mathcal{O}(1/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{N}}} + 1/\sqrt{n_{\mathrm{U}}})$$

(3) Pconf Classification

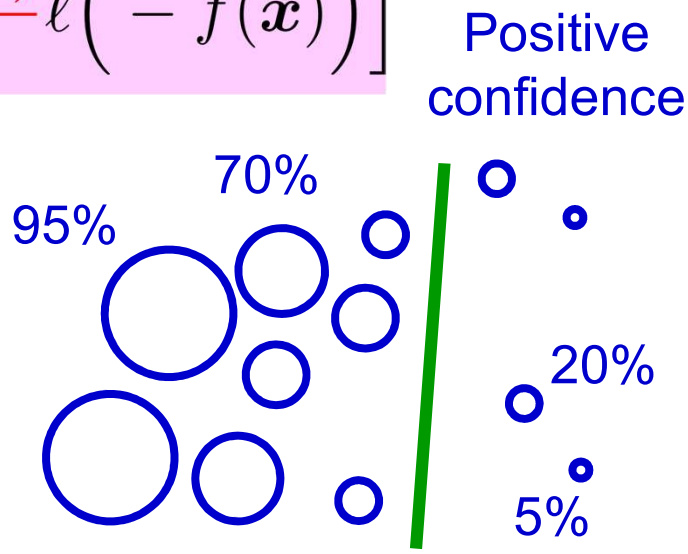
Ishida, Niu & Sugiyama (NeurIPS2018)

- Only P data is available, even not U data:
  - Data from rival companies cannot be obtained.
  - Only positive results are reported (publication bias).
- "Only-P learning" is unsupervised.
- From positive-confidence data, ERM is possible!

$$R(f) = \pi \mathbb{E}_{p(\boldsymbol{x}|y=+1)}\left[\ell\Big(f(\boldsymbol{x})\Big) + \frac{1-r(\boldsymbol{x})}{r(\boldsymbol{x})}\ell\Big(-f(\boldsymbol{x})\Big)\right]$$

$$\pi = p(y=+1) \quad r(\boldsymbol{x}) = P(y=+1|\boldsymbol{x})$$

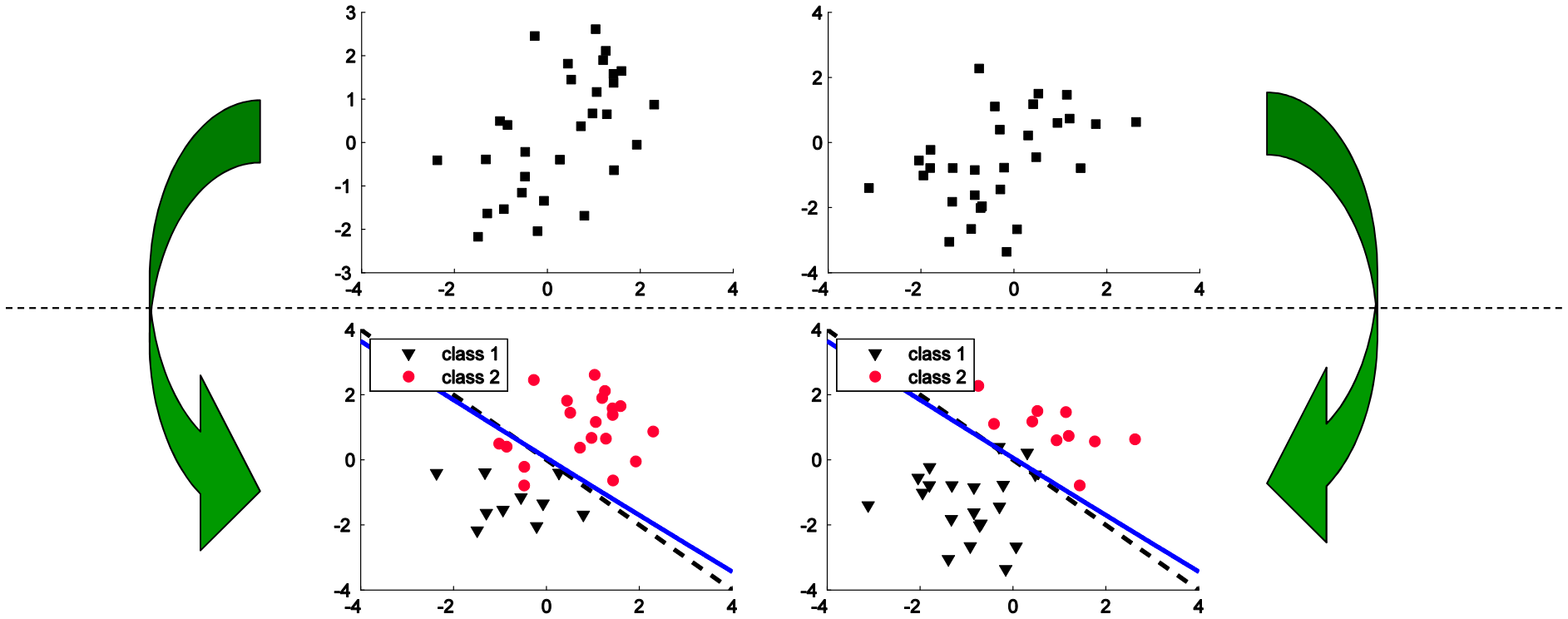
$$R\Big(f^*\Big) - R\Big(\widehat{f}\Big) = \mathcal{O}_p\Big(n^{-1/2}\Big)$$



(4) UU Classification

du Plessis, Niu & Sugiyama (TAAI2013), Nan, Niu, Menon & Sugiyama (ICLR2019)

- From two sets of unlabeled data with different class priors, PN classifiers are trainable!

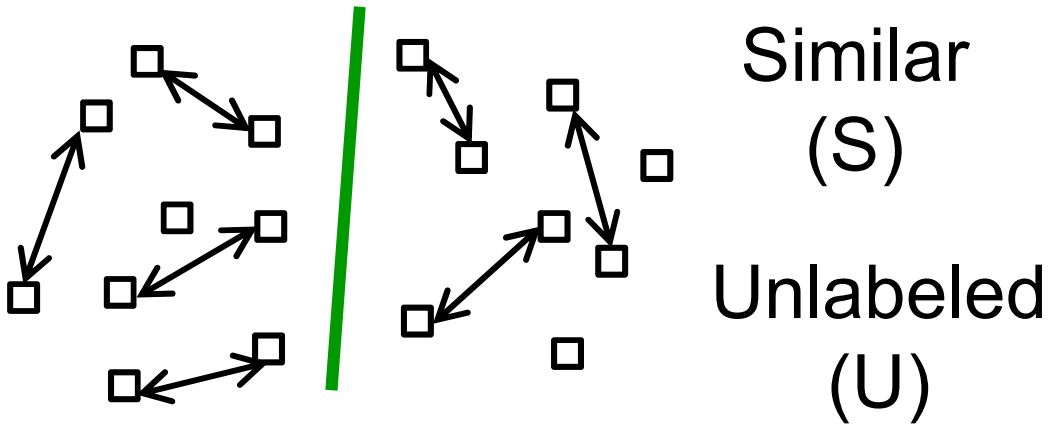


- In PU, we regarded U as noisy N.
- In UU, we use noisy P and noisy N!

(5) SU Classification

Bao, Niu & Sugiyama (ICML2018)

- Delicate classification (income, religion…):
  - Highly hesitant to directly answer questions.
  - Less reluctant to just say "same as him/her".
- From similar and unlabeled data, PN classifiers are trainable!



  - Decoupling S-pairs results in UU classification!

(6) Complementary Classification

Ishida, Niu, Hu & Sugiyama (NIPS2017), Ishida, Niu, Menon & Sugiyama (arXiv2018)

- Labeling patterns in multi-class problems:
  - Selecting the collect class from a long class list is extremely painful.

https://www.bostondynamics.com/atla

1. Amazon Kiva
2. Aldebaran Nao
3. Softbank Pepper
4. Sony Aibo
5. iRobot Roomba

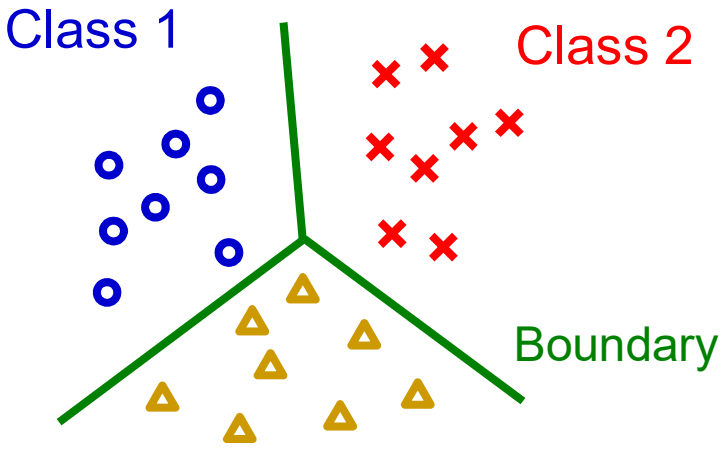
⋮

83. Boston Dynamics Atlas

⋮

100. Rethink Robotics Baxter

- Complementary labels:
  - Specify a class that a pattern does not belong to.
  - This is much easier and faster to collect!
- From complementary labels, classifiers are trainable!



$1/\sqrt{n}$

Incorporating Ordinary Labels

- Convert multiclass labeling into yes-no labeling:

http://www.softbank.jp/corp/group/sbr/news/press/2014/20141029_01/

https://www.bostondynamics.com/atlas

Is this Softbank Pepper?
Yes! (ordinary label)

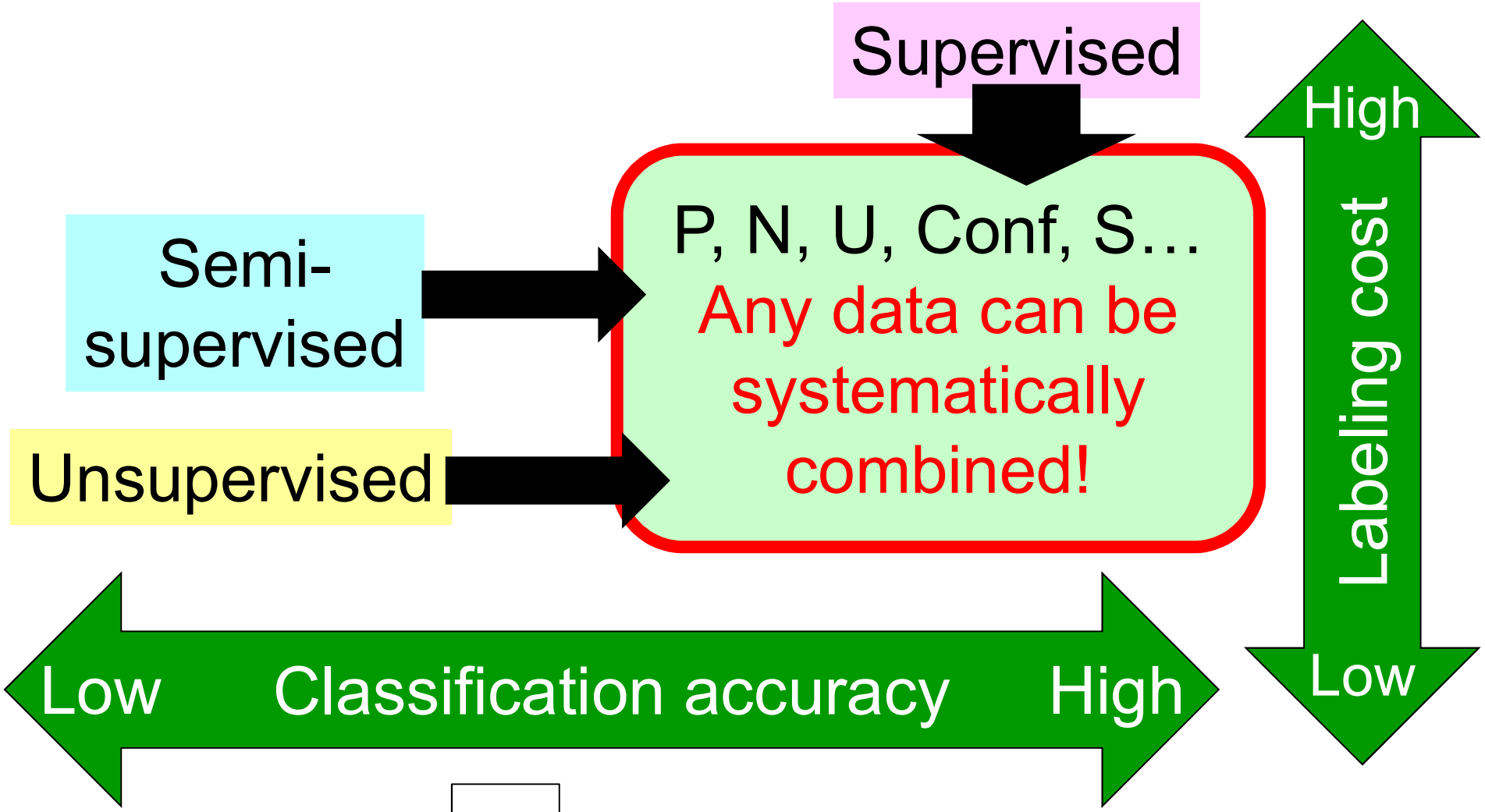
Is this iRobot Roomba?
No! (complementary label)

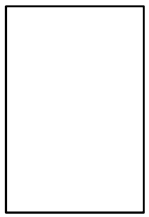
- Use both of ordinary and complementary labels!

$$R(f) = \alpha \mathbb{E}_{p(\boldsymbol{x},y)}\Big[\mathcal{L}\Big(f(\boldsymbol{x}),y\Big)\Big] + (1-\alpha)\Big\{(c-1)\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\Big[\bar{\mathcal{L}}\Big(f(\boldsymbol{x}),\bar{y}\Big)\Big] + \mathrm{Const.}\Big\}$$

$$\alpha \in [0,1]$$

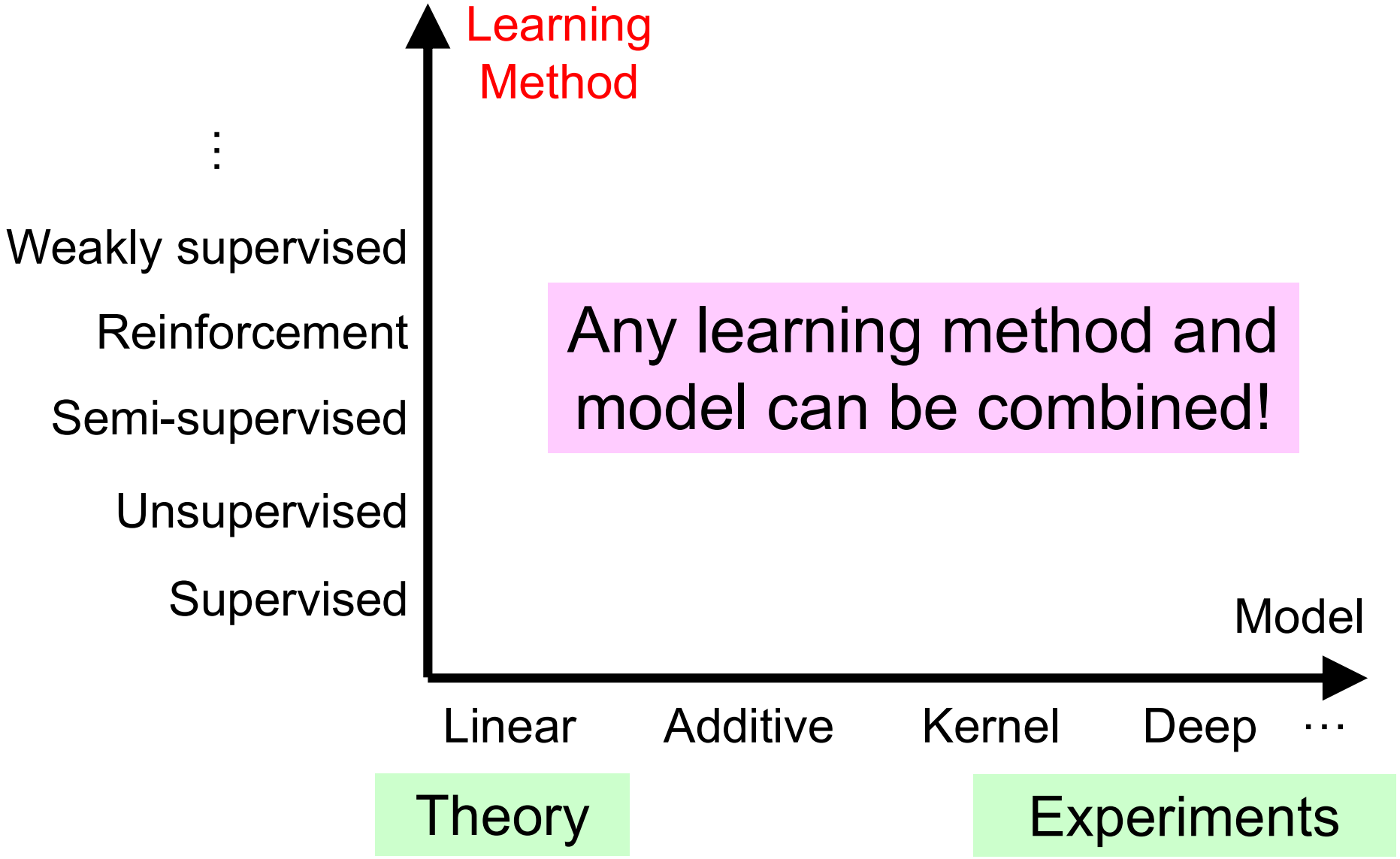
Learning from Weak Supervision





Sugiyama, Niu, Sakai & Ishida,
Machine Learning from Weak Supervision
MIT Press, 2020 (?)

Model vs. Learning Methods



My Talk

1. Weakly supervised classification
2. Reliable machine learning
3. More

Reliable Machine Learning

- For reliable deployment of machine learning systems in the real world, various types of robustness is needed:
  - Robustness to noisy training input
  - Robustness to noisy training output
  - Robustness to changing environments
  - Robustness to noisy test input

Noisy Training Output (1)

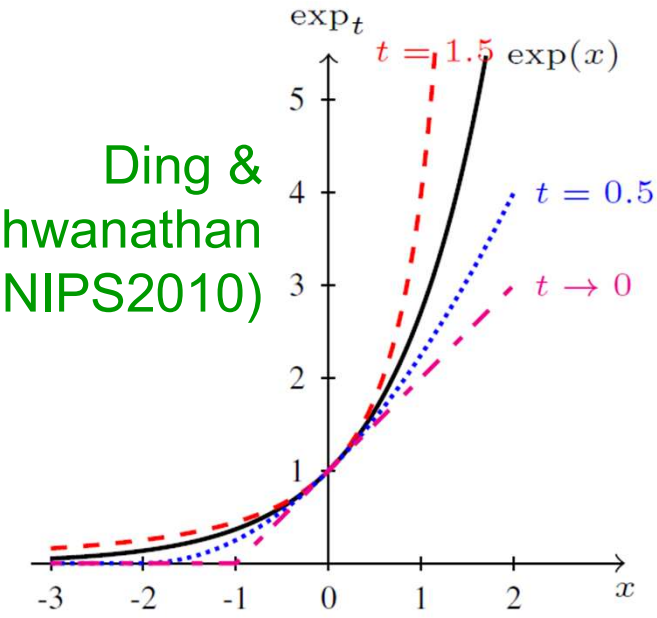
Futami, Sato & Sugiyama (NIPS2017)

- t-exponential family is useful:

$$p(x;\theta) = \exp_t(\langle \Phi(x), \theta \rangle - g_t(\theta))$$

$$\exp_t(x) = \begin{cases} \exp(x) & \text{if } t = 1 \\ [1 + (1-t)x]^{\frac{1}{1-t}} & \text{otherwise} \end{cases}$$

Ding & Vishwanathan (NIPS2010)



- However, it computationally challenging to use it in the Bayesian framework.
- Our proposal: Use t-algebra.

Nivanen, Le Mehaute & Wang (Reports on Mathematical Physics, 2003)

$$\exp_t(x)\exp_t(y) = \exp_t(x + y + (1-t)xy)$$

- We can develop an expectation-propagation algorithm for t-exponential family!

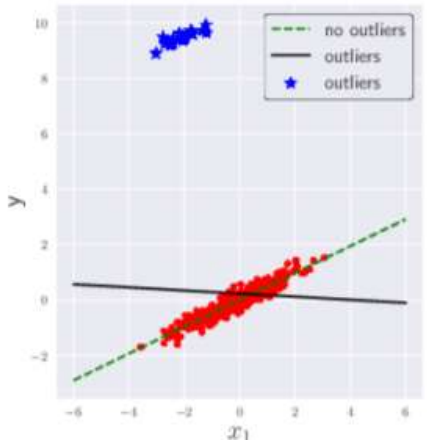
Noisy Training Output (2)

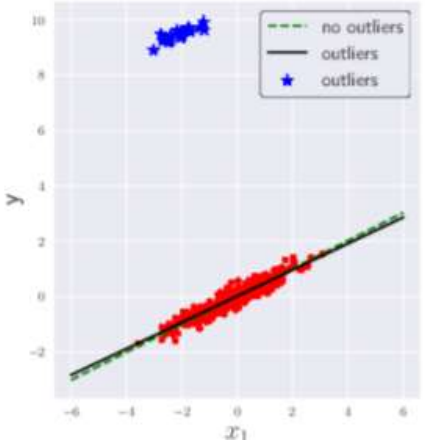
Futami, Sato & Sugiyama (AISTATS2018)

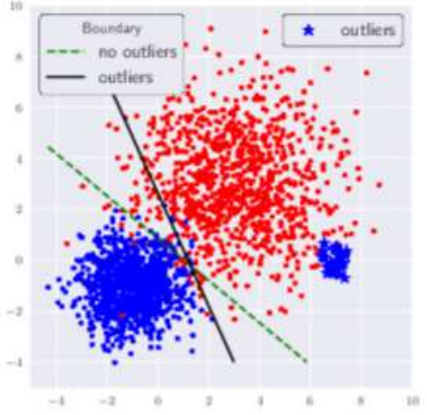
- We need more scalability when a more complex model such as a deep network is used.
- Our proposal: Change not the model, but KL-div to robust-div in variational inference.

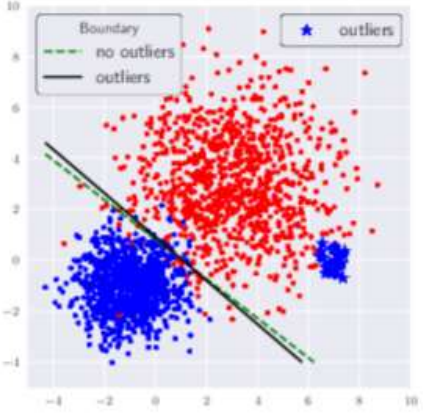
$$D_\beta(g\|f) = \frac{1}{\beta}\int g(x)^{1+\beta}dx + \frac{\beta+1}{\beta}\int g(x)f(x)^\beta dx + \int f(x)^{1+\beta}dx$$

Basu, Harris, Hjort & Jones (Biometrika1998)





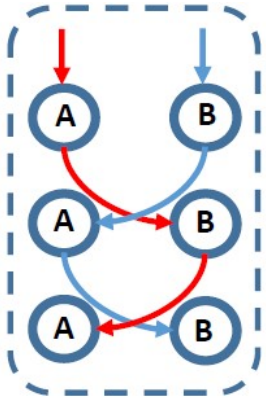


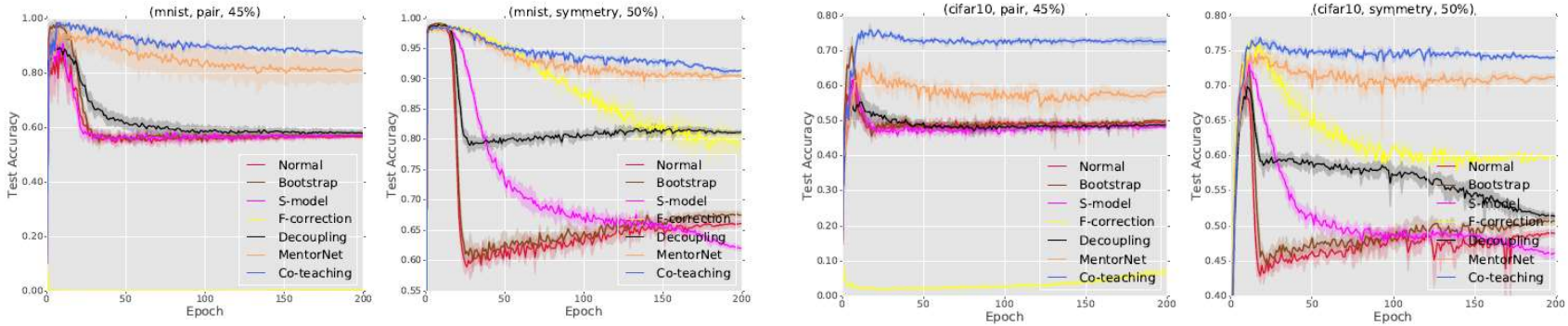


Noisy Training Output (3)

Han, Yao, Yu, Niu, Xu, Hu, Tsang & Sugiyama (NeurIPS2018)
Yu, Han, Yao, Niu, Tsang & Sugiyama (arXiv2019)

- Memorization of neural networks:
  - Empirically, clean data are fitted faster than noisy data.
- "Co-teaching" between two networks:
  - Select small-loss instances as clean data and teach them to another network.
- Experimentally works very well!
  - But no theory.





Noisy Training Output (4)

Kiryo, Niu, du Plessis & Sugiyama (NIPS2017), Han, Niu, Yao, Yu, Xu, Tsang & Sugiyama (arXiv2018)

- In PU learning, risk is estimated as

$$R(f) = \pi \mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big] + (1-\pi)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]$$

$$+\mathbb{E}_{p(\boldsymbol{x})}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big] - \pi\mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]$$

Risk for N data, which is non-negative by definition,
but its empirical approximation can be negative.

- Empirical N-risk going negative implies that something wrong is happening.
- Our proposal: Perform gradient ascent to step back to avoid poor local optima.
  - Can be generalized to general noisy data!
  - Empirically work very well, but no theory.

Changing Environments (1)

Hu, Niu, Sato & Sugiyama (ICML2018)

- Distributionally robust supervised learning:
  - Being robust to the worst test distribution.
  - Works well in regression.

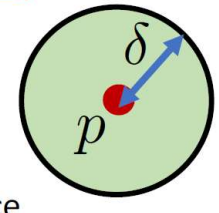
$$\min_{\theta} \sup_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(g_\theta(x), y)]$$

$$\mathcal{Q}_p = \{q \mid \mathrm{D}_f(q\|p) \leq \delta\}$$

"f-divergence ball"

[Bagnell 2005, Ben-Tal+ 2013, Namkoong+ 2016, 2017]

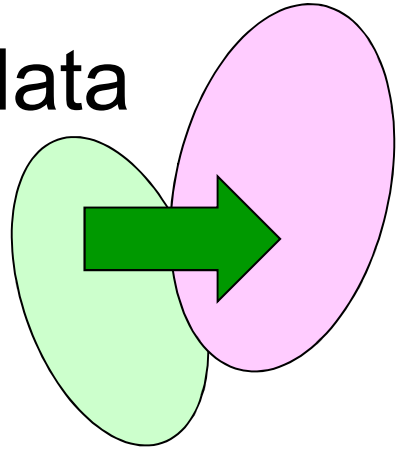
E.g. KL divergence, Chi-square divergence



- Our finding: In classification, this merely results in the same non-robust classifier.
  - Since the 0-1 loss is different from a surrogate loss.
- Additional distributional assumption can help:
  - E.g., latent prior change Storkey & Sugiyama (NIPS2007)

Changing Environments (2)

- Unsupervised domain adaptation:
  source labeled and target unlabeled data
- Concern: If source- and target-data distributions are completely different, domain adaptation does not work.



  - How to measure distribution discrepancy is key!
- Proposal: New discrepancy measures
  - Source labels are used for evaluation.
    Kuroki, Charoenphakdee, Bao, Honda, Sato & Sugiyama (AAAI2019)
  - Computable for complex models (e.g., DNNs).
    Lee, Charoenphakdee, Kuroki & Sugiyama (arXiv2019)

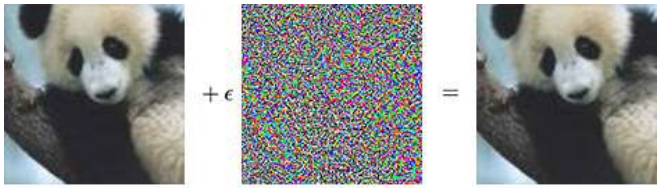
Noisy Test Input (1)

Tsuzuku, Sato & Sugiyama (NeurIPS2018)

- Adversarial attack can fool a classifier.
- Goal: Stabilize prediction

$$\forall \epsilon, \left(\|\epsilon\|_2 < c \;\Rightarrow\; t_X = \underset{i}{\operatorname{argmax}} \{F(X+\epsilon)_i\} \right)$$

https://blog.openai.com/adversarial-example-research/



- Our proposal: Lipschitz-margin training
  - Calculate the Lipschitz constant for each layer and derive the Lipschitz constant L_F for entire network.

$$\|F(X) - F(X+\epsilon)\|_2 \leq L_F \|\epsilon\|_2$$

  - Make the prediction margin large enough:

$$\forall i \neq t_X, (F_{t_X} \geq F_i + \sqrt{2} c L_F)$$

  - Provable guarded area for attacks.
  - Computationally efficient and empirically robust.

Noisy Test Input (2)

Ni, Charoenphakdee, Honda & Sugiyama (arXiv2019)

- In critical applications, it is better to reject difficult test inputs and ask humans' help.
- Approach 1: Reject low-confidence prediction
  - Existing methods have limitation in loss functions (e.g., logistic loss), resulting in weak performance.
  - We proposed new rejection criteria for general losses (e.g., cross-entropy loss).
- Approach 2: Train classifier and rejector
  - Existing methods only focus on binary problems.
  - We showed that this approach does not converge to the optimal solution in multi-class cases.

My Talk

1. Weakly supervised classification
2. Reliable machine learning
3. More

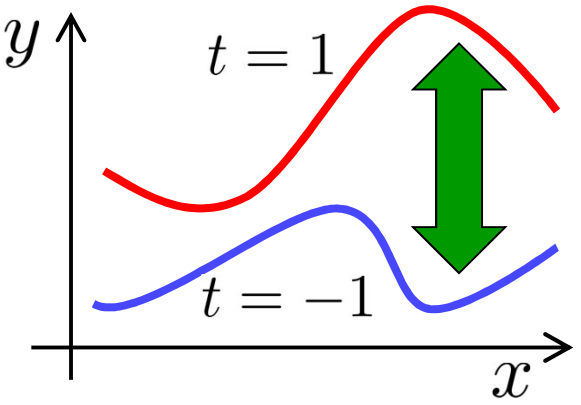
Individual Treatment Effect

Yamane, Yger, Atif & Sugiyama (NeurIPS2018)

$$\mathbb{E}[y|x,t=1]-\mathbb{E}[y|x,t=-1]$$

x: subject, y: outcome, t: treatment flag

- Setting: Due to privacy reasons, we cannot obtain (x,y,t)-triplets, but only (x,y)- and (x,t)-pairs without correspondence in x.

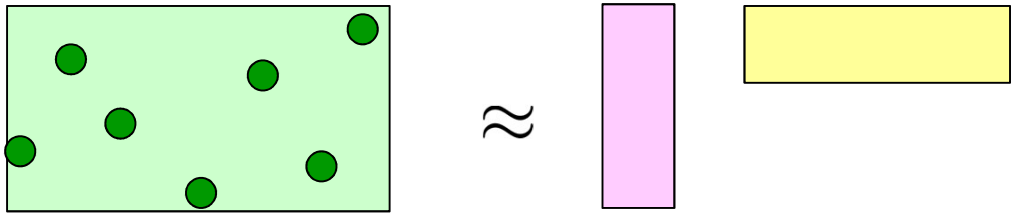


- Result: Solvable if we have (x,y)- and (x,t)-pairs from two different treatment policies.
  - Direct fitting without 4-time regression.

$$\widehat{u}(\boldsymbol{x})=2\cdot\frac{\widehat{\mathbf{E}}_{p_1}[y|\boldsymbol{x}]-\widehat{\mathbf{E}}_{p_2}[y|\boldsymbol{x}]}{\widehat{\mathbf{E}}_{p_1}[t|\boldsymbol{x}]-\widehat{\mathbf{E}}_{p_2}[t|\boldsymbol{x}]}$$

Matrix Completion

- Golden standard: Low-rank approximation of a matrix from its sparse observations.



- Matrix co-completion for multi-label classification with missing features and labels.

Features	Soft labels

Xu, Niu, Han, Tsang, Zhou & Sugiyama (arXiv2018)

- Clipped matrix factorization for ceiling effect.
  - Allowing entries taking beyond their upper-limits improves the recovery accuracy.

Teshima, Xu, Sato & Sugiyama (AAAI2019)

My Talk

1. Weakly supervised classification
2. Reliable machine learning
3. More

Summary

- ML from imperfect information:
  - Weakly supervised/noisy training data
  - Reinforcement/imitation learning, bandits
- Reliable deployment of ML systems:
  - Changing environments, adversarial test inputs
  - Bayesian inference
- Versatile ML:
  - Estimation of density ratio/difference/derivative