Max Planck Institute for Intelligent Systems Sep. 22, 2022

Robust Machine Learning from Weakly-Supervised, Noisy-Labeled, and Biased Data

Masashi Sugiyama

RIKEN Center for Advanced Intelligence Project/
The University of Tokyo

http://www.ms.k.u-tokyo.ac.jp/sugi/

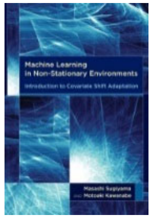
About Myself

- **Masashi Sugiyama:**
  - Director: RIKEN AIP, Japan
  - Professor: University of Tokyo, Japan
  - Consultant: several local startups

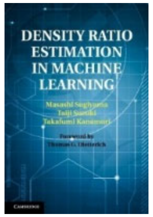
- **Interests: Machine learning (ML)**
  - ML theory & algorithm →
  - ML applications (signal, image, language, brain, robot, mobility, advertisement, biology, medicine, education…)

- **Academic activities:**
  - Program Chairs for NeurIPS2015, AISTATS2019, ACML2010/2020…

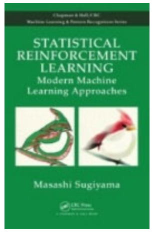
Sugiyama & Kawanabe, Machine Learning in Non-Stationary Environments, MIT Press, 2012



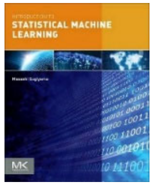
Sugiyama, Suzuki & Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012



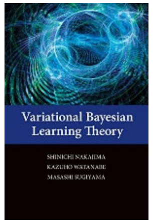
Sugiyama, Statistical Reinforcement Learning, Chapman and Hall/CRC, 2015



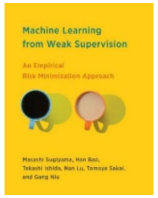
Sugiyama, Introduction to Statistical Machine Learning, Morgan Kaufmann, 2015



Nakajima, Watanabe & Sugiyama, Variational Bayesian Learning Theory, Cambridge University Press, 2019



Sugiyama, Bao, Ishida, Lu, Sakai & Niu. Machine Learning from Weak Supervision, MIT Press, 2022.



What is “RIKEN”?

- Name in Japanese: 理化学研究所
  - Pronounced as: rikagaku kenkyusho
  - Meaning: Physics and Chemistry Research Institute

- Acronym in Japanese: 理研 (RIKEN)

What is RIKEN-AIP?

MEXT Advanced Intelligence Project (2016-2025):

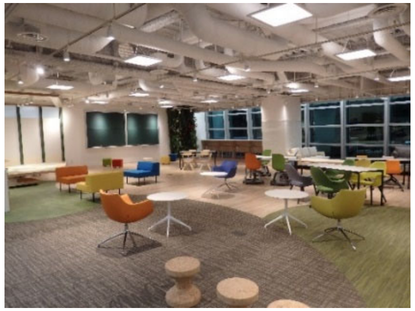
- 130 employed researchers (36% international, 23% female)
- 200 visiting researchers, 100 domestic students
- 140 international interns (total)

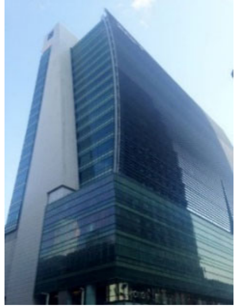
MEXT
MINISTRY OF EDUCATION, CULTURE, SPORTS, SCIENCE AND TECHNOLOGY-JAPAN

Missions:

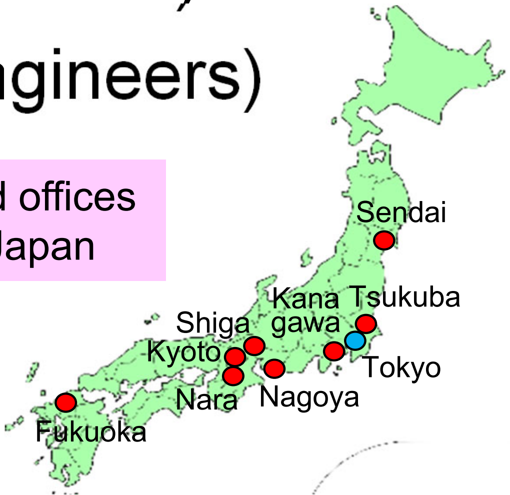
- Develop new AI technology (ML, Opt, math)
- Accelerate scientific research (cancer, material, genomics)
- Solve socially critical problems (disaster, elderly healthcare)
- Study of ELSI in AI (ethical guidelines, personal data)
- Human resource development (researchers, engineers)

Main office in the heart of Tokyo





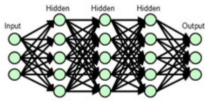
Distributed offices across Japan



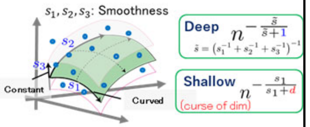
Selected Research

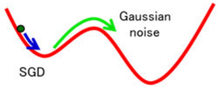
Developing New AI Technology

- **Theory of deep learning**:
  - Better prediction than shallow learning
  - No curse of dimensionality
  - Global optimization
- **Developing new methods**:
  - Weakly supervised learning
  - Noise robust learning
  - Causal inference



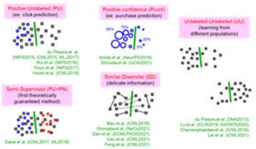
$$\mathbb{E}[\|f_T - f^*\|_{L_2}^2] \le \epsilon_M + O(T^{-\frac{2r\beta}{2r\beta+1}})$$


s_1, s_2, s_3: Smoothness
s_2
s_3
s_1
Constant
Curved
Deep $n^{-\frac{\tilde{s}}{\tilde{s}+1}}$
$\tilde{s} = (s_1^{-1} + s_2^{-1} + s_3^{-1})^{-1}$
Shallow $n^{-\frac{s_1}{s_1+d}}$
(curse of dim)



Weakly Supervised Classification

- Various weakly supervised classification problems can be solved by risk-rewriting systematically!


Positive-Unlabeled (PU)
(ex: click prediction)
du Plessis et al. (NIPS2014, ICML2015, MLJ2017)
Niu et al. (NIPS2016),
Kiryo et al. (NIPS2017)
Hsieh et al. (ICML2019)
Positive-confidence (Pconf)
(ex: purchase prediction)
Ishida et al. (NeurIPS2018)
Shinoda et al. (IJCAI2021)
Unlabeled-Unlabeled (UU)
(learning from different populations)
Semi-Supervised (PU+PN)
(first theoretically guaranteed method)
Sakai et al. (ICML2017, ML2018)
Similar-Dissimilar (SD)
(delicate information)
Bao et al. (ICML2018)
Shimada et al. (NeCo2021)
Dan et al. (ECMLPKDD2021)
Cao et al. (ICML2021)
Feng et al. (ICML2021)
du Plessis et al. (TAAI2013)
Lu et al. (ICLR2019, AISTATS2020)
Charoenphakdee et al. (ICML2019)
Lei et al. (ICML2021)

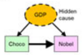
Noise Transition Correction

- Noise transition matrix T: $T^{\top} = \begin{bmatrix} 1 & 0.1 & 0.5 \\ 0 & 0.8 & 0.5 \\ 0 & 0.1 & 0 \end{bmatrix}$
  - Clean-to-noisy flipping probability.
- Major approaches: Patrini et al. (CVPR2017)
  - Loss correction by T^{-1} to eliminate noise.
  - Classifier adjustment by $T^{\top}$ to simulate noise.
- We want to estimate T only from noisy data:
  - Use human cognition as a "mask" for T. Han et al. (NeurIPS2018)
  - Learn T and a classifier dynamically. Xia et al. (NeurIPS2019)
  - Decompose T into simpler components. Yao et al. (NeurIPS2020)
  - Regularize T to be estimable. Zhang et al. (ICML2021), Li et al. (ICML2021)
  - Extension to input-dependent noise $T(\boldsymbol{x})$.

Xia et al. (NeurIPS2020), Berthon et al. (ICML2021)

Causal Inference in the Presence of Hidden Cause

- In causal inference, how to handle hidden cause is a big challenge!



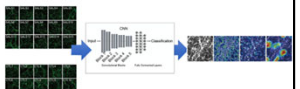
- We developed the first method to estimate the entire structure in the presence of hidden cause:
  - Speech separation technique is employed to separate hidden cause.

Maeda & Shimizu (AISTATS2020, UAI2021)

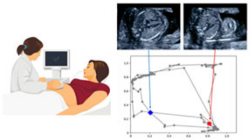
Accelerating Scientific Research

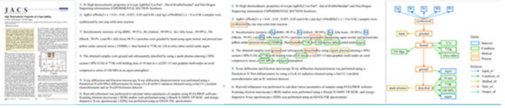
- **Medical science**:
  - Prostate/pancreatic cancer detection
  - ALS early diagnosis
  - Fetal heart screening
  - Colonoscopy
- **Material science**:
  - Database creation with text mining
- **Data-driven science**:
  - Selective inference for reliability evaluation

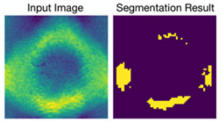
*Yoichiro Yamamoto, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. *Nature Communications* 10:5642, 2019.

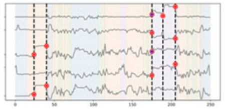


Imamura et al., Ann Neurology, 2021



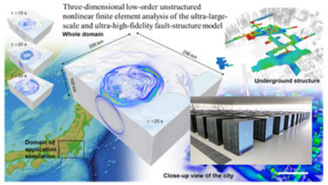


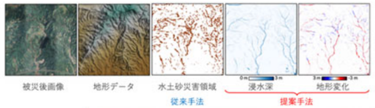


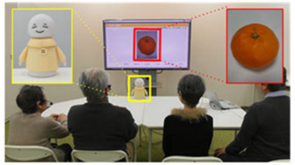


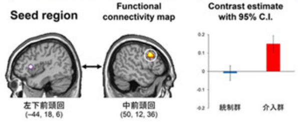
Solving Socially Critical Problems

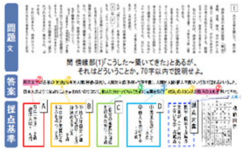
- **Natural disaster**:
  - Fugaku-based earthquake simulation
  - Remote sensing disaster analysis
- **Elderly healthcare**:
  - Chat-robot-guided cognitive function improvement
- **Education**:
  - Automatic essay evaluation
  - Interactive essay writing support


Three-dimensional low-order unstructured nonlinear finite element analysis of the ultra-large-scale and ultra-high-fidelity fault-structure model
Whole domain
205 km
256 km
100 km
Underground structure
Domain of application simulation
Close-up view of the city





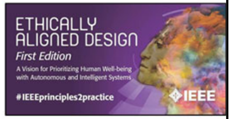


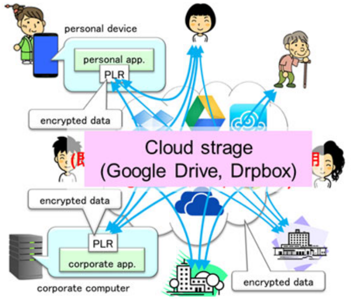

問題文
問 傍線部(1)「こうした～覆いてきた」とあるが、それはどういうことか。70字以内で説明せよ。
答案
採点基準
A
B
C
D

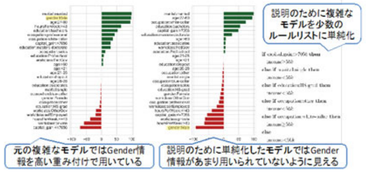
Smoking should be banned! It causes bad breath and makes your clothes stinky!
While your point is logically valid, it does not justify banning the substance. Can you provide a more relevant reason?
Oh..I know that many people die of second-hand smoke each day!
Indeed, people dying from second-hand smoke is much more convincing than causing bad breath. Where did you learn about this information?

Studying AI-ELSI

- **AI Ethical guidelines**:
  - Japanese Society for AI, Ministry of Internal Affairs and Communications, Cabinet Office
  - IEEE, G20, OECD
- **Personal data management**:
  - Individual-based accessibility control system
- **AI security and reliability**:
  - Adversarial attack/defense
  - Fairness faking/guarantee


ETHICALLY ALIGNED DESIGN
First Edition
A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems
#IEEEprinciples2practice
IEEE





Today’s Topic: Robust Machine Learning

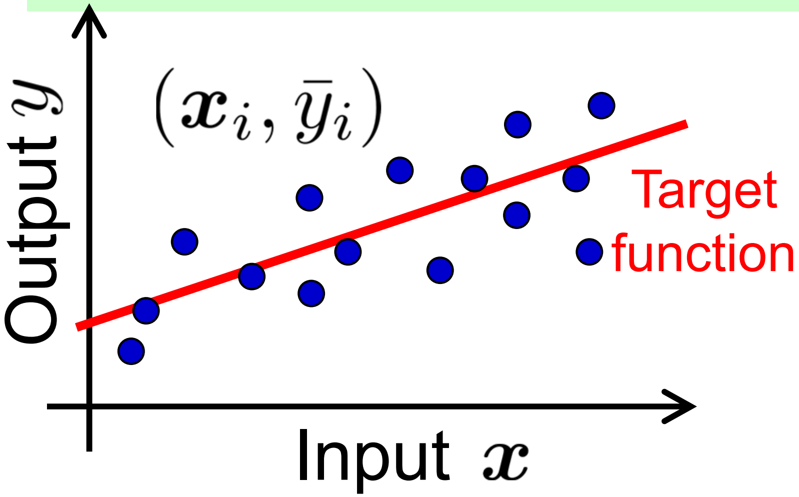
- Goal: Develop novel ML theories and algorithms that enable reliable learning from limited information.
  - Label noise: human error, sensor error.
  - Insufficient information: weak supervision.
  - Data bias: changing environments, privacy.
  - Attack: adversarial noise, distribution shift.

Contents

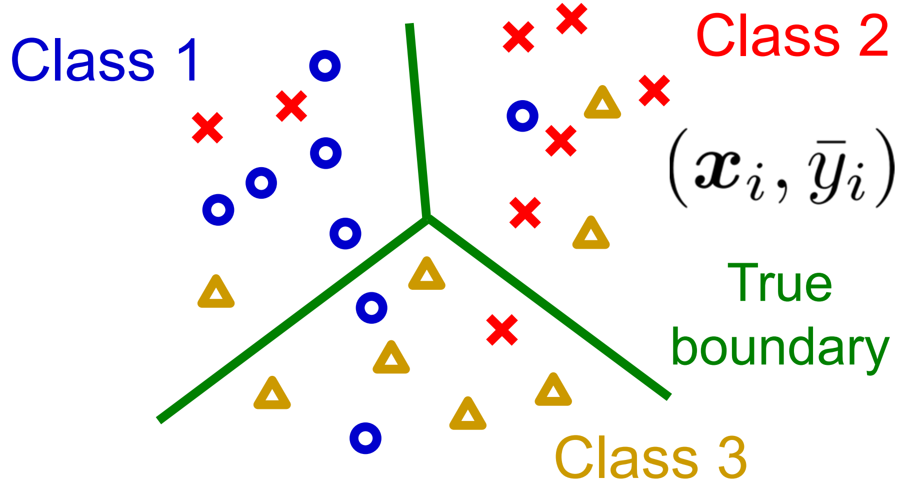
1. Noisy-Label Learning
   - A) Technical background
   - B) Single-step approach
   - C) Beyond anchor points
   - D) Further challenges
2. Weakly Supervised Learning
3. Transfer Learning
4. Summary

Supervised Learning with Noisy Output

Regression (additive noise)



Classification (label flipping noise)



$$\min_g \sum_{i=1}^{n} \ell\Big(\bar{y}_i, g(\boldsymbol{x}_i)\Big)$$

ℓ : loss

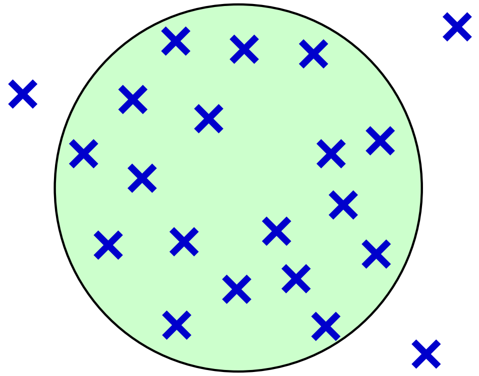
g : classifier

$\bar{y}$: noisy output

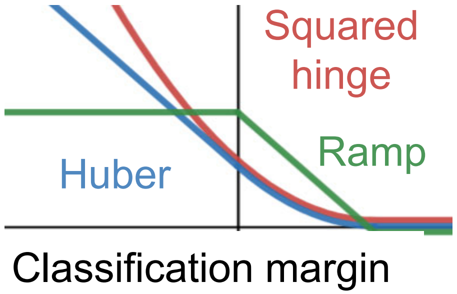
- Hasn't such a classic problem been solved?
  - Regression: Yes, big data yields consistency.
  - Classification: Specific noise reduction mechanism is needed to achieve consistency!

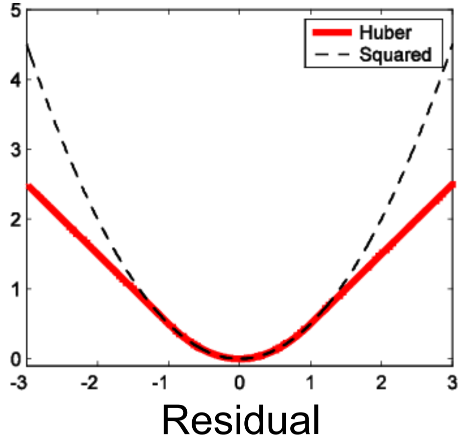
Classical Approaches

- **Unsupervised outlier removal**:
  - Substantially more difficult than classification.



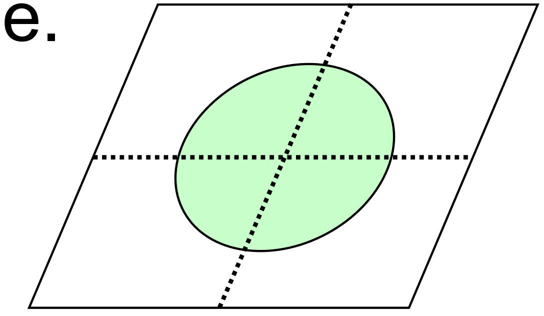
- **Robust loss:**
  - Works well for regression, but limited effectiveness for classification.



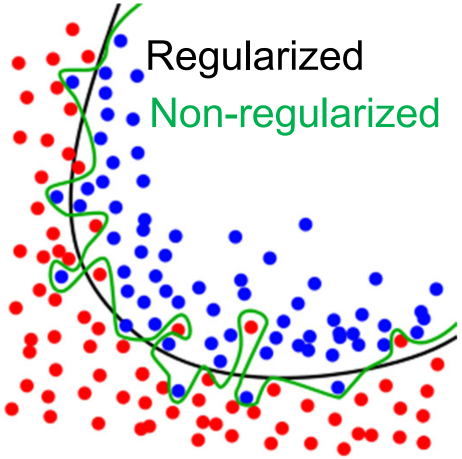


- **Regularization:**
  - Effective in suppressing overfitting, but too smooth for strong noise.

- **Need new approaches!**



ℓ_2-regularization



https://en.wikipedia.org/wiki/Overfitting

Contents

1. Noisy-Label Learning
   - A) Technical background
   - B) Single-step approach
   - C) Beyond anchor points
   - D) Further challenges
2. Weakly Supervised Learning
3. Transfer Learning
4. Summary

Formulation

- Clean training data: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y)$
- Noisy training data: $\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \bar{p}(\boldsymbol{x}, \bar{y})$

$\boldsymbol{x} \in \mathbb{R}^d$: Input instance

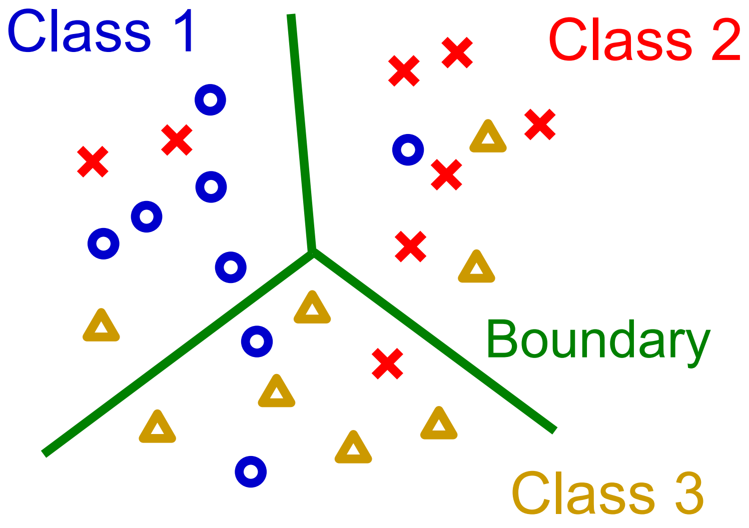
$y \in \{1, \dots, c\}$: Clean class label

$\bar{y} \in \{1, \dots, c\}$: Noisy class label

- Probabilistic classifier in simplex: $\boldsymbol{h}(\boldsymbol{x}) \in \Delta^{c-1}$
  - Each element approximates the class-posterior probability.

$$h_y(\boldsymbol{x}) \approx p(y|\boldsymbol{x})$$

- Loss: $\ell(y, \boldsymbol{h}(\boldsymbol{x})) \in \mathbb{R}$

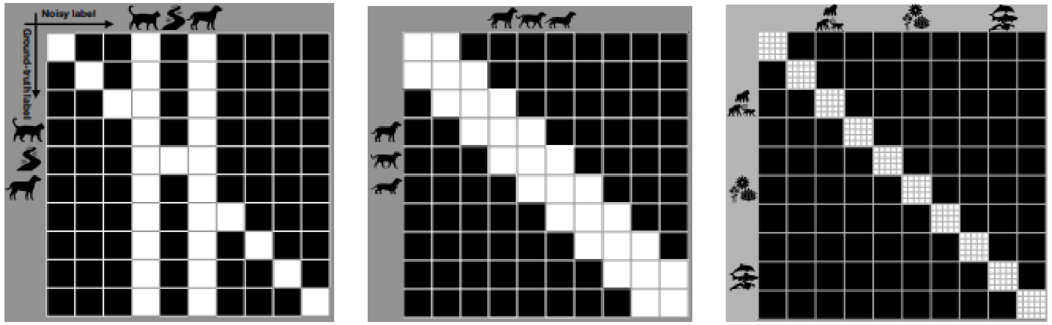


Modeling Class-Conditional Noise

- **Noise transition matrix:** $T_{y,\bar{y}} = \bar{p}(\bar{y}|y)$
  - Probability of flipping y to $\bar{y}$.

$$\begin{array}{c|ccc} & 1 & 0 & 0 \\ y & 0.1 & 0.8 & 0.1 \\ & 0.5 & 0.5 & 0 \\ \hline & & \bar{y} & \end{array}$$

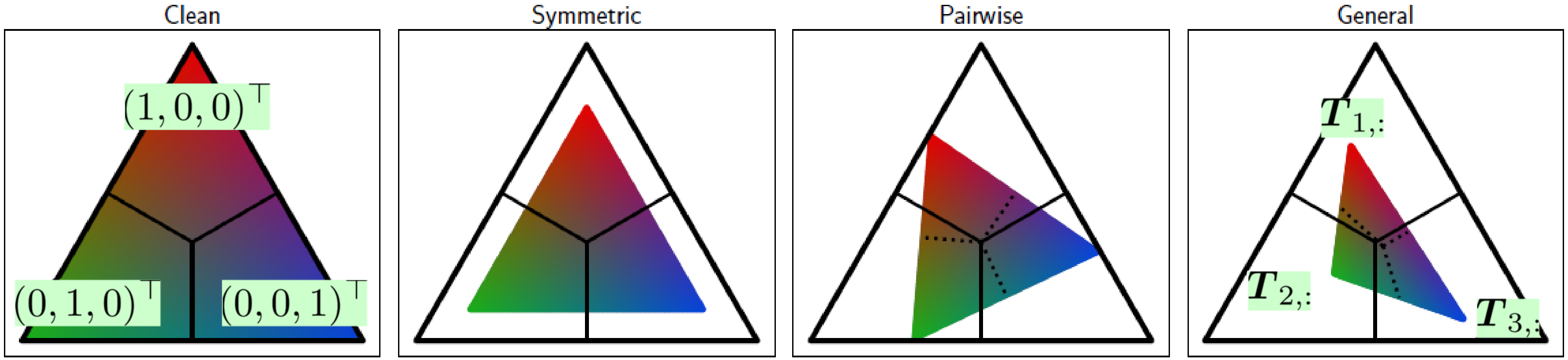
- We may encode human-cognitive bias:



(a) Column-diagonal (b) Tri-diagonal (c) Block-diagonal

Han, Yao, Niu, Zhou, Tsang, Zhang & Sugiyama (NeurIPS2018)

- Visualization as a simplex: Zhang, Niu & Sugiyama (ICML2021)



Loss Correction

Patrini, Rozza, Menon, Nock & Qu (CVPR2017)

- Forward correction: Add noise by $\boldsymbol{T}^\top$
  - $\boldsymbol{\ell}^{\rightarrow}(\boldsymbol{h}(\boldsymbol{x})) = \boldsymbol{\ell}(\boldsymbol{T}^\top \boldsymbol{h}(\boldsymbol{x}))$ $\quad \ell_y^*(\boldsymbol{h}(\boldsymbol{x})) = \ell^*(y, \boldsymbol{h}(\boldsymbol{x}))$

Classifier-consistency

$$\operatorname*{argmin}_{\boldsymbol{h}} \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell^{\rightarrow}(y, \boldsymbol{h}(\boldsymbol{x}))] = \operatorname*{argmin}_{\boldsymbol{h}} \mathbb{E}_{p(\boldsymbol{x},y)}[\ell(y, \boldsymbol{h}(\boldsymbol{x}))]$$

- Backward correction: Remove noise by $\boldsymbol{T}^{-1}$
  - $\boldsymbol{\ell}^{\leftarrow}(\boldsymbol{h}(\boldsymbol{x})) = \boldsymbol{T}^{-1}\boldsymbol{\ell}(\boldsymbol{h}(\boldsymbol{x}))$

Classifier-consistency

$$\operatorname*{argmin}_{\boldsymbol{h}} \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell^{\leftarrow}(y, \boldsymbol{h}(\boldsymbol{x}))] = \operatorname*{argmin}_{\boldsymbol{h}} \mathbb{E}_{p(\boldsymbol{x},y)}[\ell(y, \boldsymbol{h}(\boldsymbol{x}))]$$

Risk-consistency

$$\forall \boldsymbol{x},\ \mathbb{E}_{\bar{p}(\bar{y}|\boldsymbol{x})}[\ell^{\leftarrow}(y, \boldsymbol{h}(\boldsymbol{x}))] = \mathbb{E}_{p(y|\boldsymbol{x})}[\ell(y, \boldsymbol{h}(\boldsymbol{x}))]$$

- If $\boldsymbol{T}$ is given, consistency can be guaranteed!

Identifiability of Noise Transition

- In practice, we need to estimate $\boldsymbol{T}$ from noisy training data $\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n$.

- However, $\boldsymbol{T}$ is non-identifiable in general:
  - $\boldsymbol{T}$ can be decomposed as $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{V}$, where $\boldsymbol{U}, \boldsymbol{V}$ are some transition matrices.
  - Then $\bar{\boldsymbol{p}}_{\boldsymbol{x}} = \boldsymbol{T}^\top \boldsymbol{p}_{\boldsymbol{x}}$
    $= \boldsymbol{V}^\top (\boldsymbol{U}^\top \boldsymbol{p}_{\boldsymbol{x}})$

$$T_{y,\bar{y}} = \bar{p}(\bar{y}|y)$$

$$[\bar{\boldsymbol{p}}_{\boldsymbol{x}}]_{\bar{y}} = \bar{p}(\bar{y}|\boldsymbol{x})$$

$$[\boldsymbol{p}_{\boldsymbol{x}}]_y = p(y|\boldsymbol{x})$$

- Let's use anchor points (100%-certain samples):

$$\{\boldsymbol{x}^y \mid p(y|\boldsymbol{x}^y) = 1\}_{y=1}^c$$

Estimation of Noise Transition with Anchor Points

- Given anchor points $\{\boldsymbol{x}^y \mid p(y|\boldsymbol{x}^y) = 1\}_{y=1}^{c}$, $T_{y,\bar{y}} = \bar{p}(\bar{y}|y)$ can be naïvely estimated as

$$T_{y,\bar{y}} = \sum_{y'=1}^{c} p(\bar{y}|y')p(y'|\boldsymbol{x}^y) = \bar{p}(\bar{y}|\boldsymbol{x}^y) \approx \bar{h}_{\bar{y}}(\boldsymbol{x}^y)$$

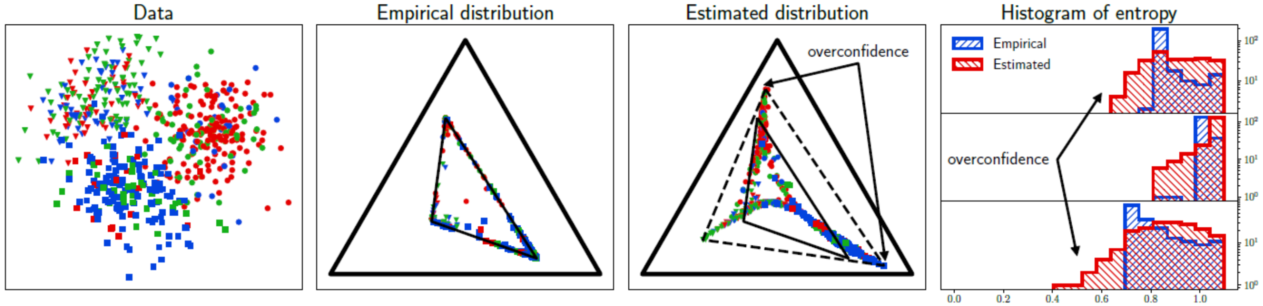
  - $\bar{\boldsymbol{h}}(\boldsymbol{x})$ is a probabilistic classifier learned from noisy training data $\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^{n}$.

- Even if anchor points are unknown, as long as they exist in noisy training data, we may find them as $\boldsymbol{x}^y \leftarrow \boldsymbol{x}_i \text{ s.t. } \bar{h}_y(\boldsymbol{x}_i) \approx 1$.

Further Improvements

$$\boldsymbol{x}^y \leftarrow \boldsymbol{x}_i \text{ s.t. } \bar{h}_y(\boldsymbol{x}_i) \approx 1$$

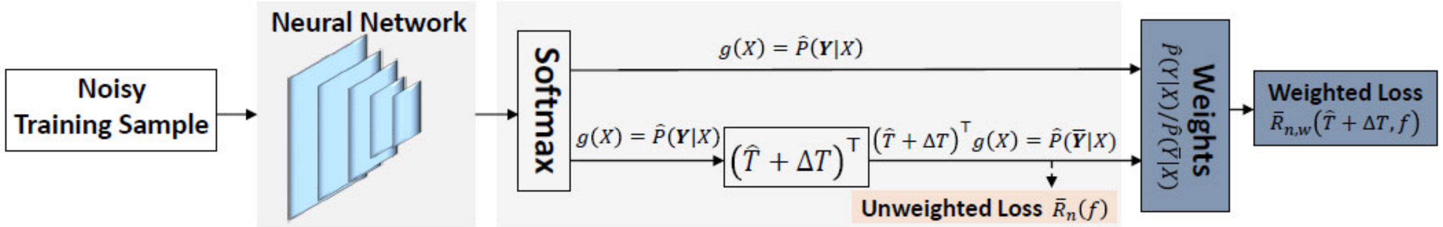
- We typically use deep learning to obtain $\bar{\boldsymbol{h}}(\boldsymbol{x})$:
  - Then it is often over-confident and unreliable.



Zhang, Niu & Sugiyama (ICML2021)

- Estimated $\boldsymbol{T}$ is revised during classifier training:

Xia, Liu, Wang, Han, Gong, Niu & Sugiyama (NeurIPS2019)


Noisy Training Sample
Neural Network
Softmax
$g(X) = \hat{P}(\boldsymbol{Y}|X)$
$g(X) = \hat{P}(\boldsymbol{Y}|X)$
$(\hat{T} + \Delta T)^{\top}$
$(\hat{T} + \Delta T)^{\top} g(X) = \hat{P}(\bar{\boldsymbol{Y}}|X)$
Unweighted Loss $\bar{R}_n(f)$
Weights
$\hat{P}(Y|X)/\hat{P}(\bar{Y}|X)$
Weighted Loss
$\bar{R}_{n,w}(\hat{T} + \Delta T, f)$

- Instead of explicitly finding anchor points, latent labels are utilized: $y_i' = \text{argmax}_{y'} \bar{h}_{y'}(\boldsymbol{x}_i)$

Yao, Liu, Han, Gong, Deng, Niu, Sugiyama & Tao (NeurIPS2020)

Contents

1. Noisy-Label Learning
   - A) Technical background
   - B) Single-step approach
   - C) Beyond anchor points
   - D) Further challenges
2. Weakly Supervised Learning
3. Transfer Learning
4. Summary

Challenge

- Current approaches are in two-step:
  1. Estimate transition matrix $\boldsymbol{T}$.
  2. Use estimated $\boldsymbol{T}$ to train a classifier $\boldsymbol{h}(\boldsymbol{x})$.

- Step 1 is done without regard to Step 2:
  - Estimation error of $\boldsymbol{T}$ in Step 1 can be magnified in Step 2.

- We want to estimate $\boldsymbol{T}$ and $\boldsymbol{h}(\boldsymbol{x})$ simultaneously in one-step.

Naïve Solution

- Naively, we may learn the noise transition and classifier at the same time as

$$\min_{\boldsymbol{U},\boldsymbol{h}} \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(\bar{y},\boldsymbol{U}^\top \boldsymbol{h}(\boldsymbol{x}))]$$

- However, the solution is not unique:
  - With any invertible transition matrix $\boldsymbol{Q}$, any $(\widehat{\boldsymbol{U}},\widehat{\boldsymbol{h}}) = (\boldsymbol{Q}^{-1}\boldsymbol{T},\boldsymbol{Q}^\top \boldsymbol{p}_{\boldsymbol{x}})$ are solutions.

$$T_{y,\bar{y}} = \bar{p}(\bar{y}|y) \qquad [\boldsymbol{p}_{\boldsymbol{x}}]_y = p(y|\boldsymbol{x})$$

- We need a certain constraint to obtain the right solution: $(\widehat{\boldsymbol{U}},\widehat{\boldsymbol{h}}) = (\boldsymbol{T},\boldsymbol{p}_{\boldsymbol{x}})$

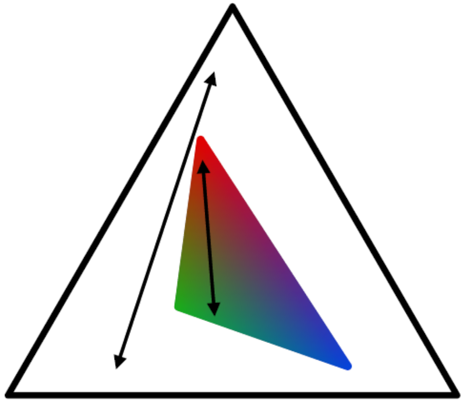
Total Variation Regularization

Zhang, Niu & Sugiyama (ICML2021)

- Noise transition $\boldsymbol{p}_{\boldsymbol{x}} \to \boldsymbol{U}^\top \boldsymbol{p}_{\boldsymbol{x}}$ is contraction in total variation distance:

$$\|\boldsymbol{U}^\top \boldsymbol{p}_{\boldsymbol{x}} - \boldsymbol{U}^\top \boldsymbol{p}_{\boldsymbol{x}'}\|_1 \leq \|\boldsymbol{p}_{\boldsymbol{x}} - \boldsymbol{p}_{\boldsymbol{x}'}\|_1$$

$$[\boldsymbol{p}_{\boldsymbol{x}}]_y = p(y|\boldsymbol{x})$$



  - Cleaner class-posteriors have larger total variation distances!

- Let's use this knowledge as a regularizer:

$$\min_{\boldsymbol{U},\boldsymbol{h}} \left[\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(\bar{y}, \boldsymbol{U}^\top \boldsymbol{h}(\boldsymbol{x}))] - \lambda \mathbb{E}_{p(\boldsymbol{x}),p(\boldsymbol{x}')}\|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{x}')\|_1\right]$$

  - Under the anchor point assumption, $\lambda > 0$ the empirical solution has statistical consistency.

Contents

1. Noisy-Label Learning
   - A) Technical background
   - B) Single-step approach
   - C) Beyond anchor points
   - D) Further challenges
2. Weakly Supervised Learning
3. Transfer Learning
4. Summary

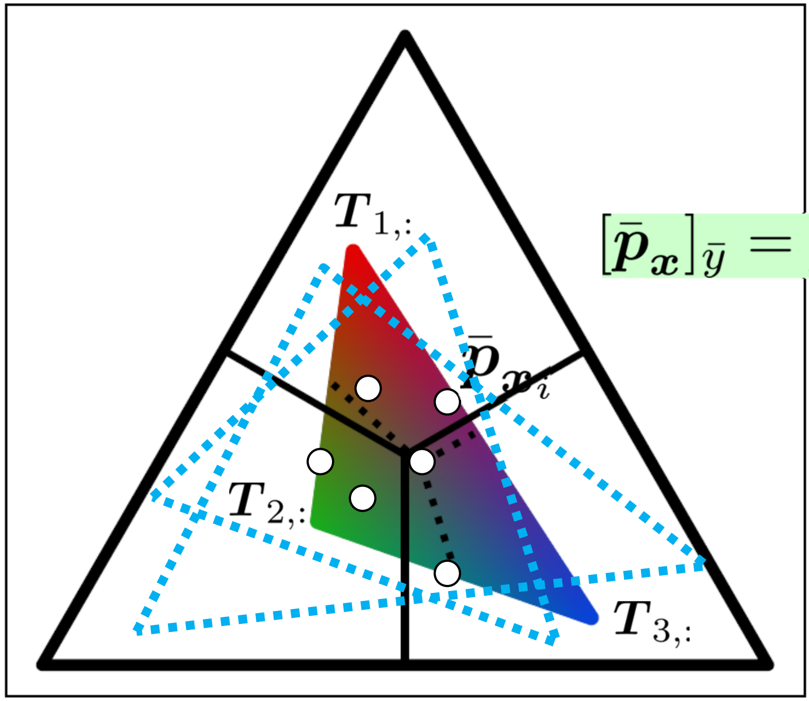
Challenges

$$\{\boldsymbol{x}^y \mid p(y|\boldsymbol{x}^y) = 1\}_{y=1}^{c}$$

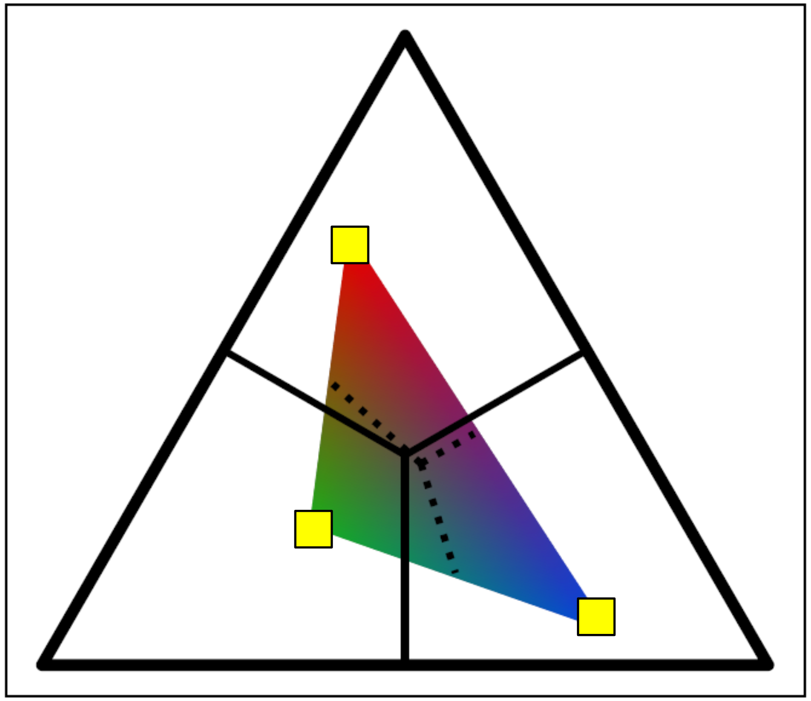
- To overcome the non-identifiability of $\boldsymbol{T}$:
  - Anchor points are explicitly used.
- This condition has been relaxed to:
  - Only the existence of anchor points is assumed.
- Can we further relax this assumption?

Non-identifiability of $\boldsymbol{T}$

- $\boldsymbol{T}$ can be visualized as a simplex, containing all training data.
- Generally, such a simplex is not unique.
- Anchor points are vertices of the true simplex.
  - Explicitly using anchor points naively recovers $\boldsymbol{T}$.

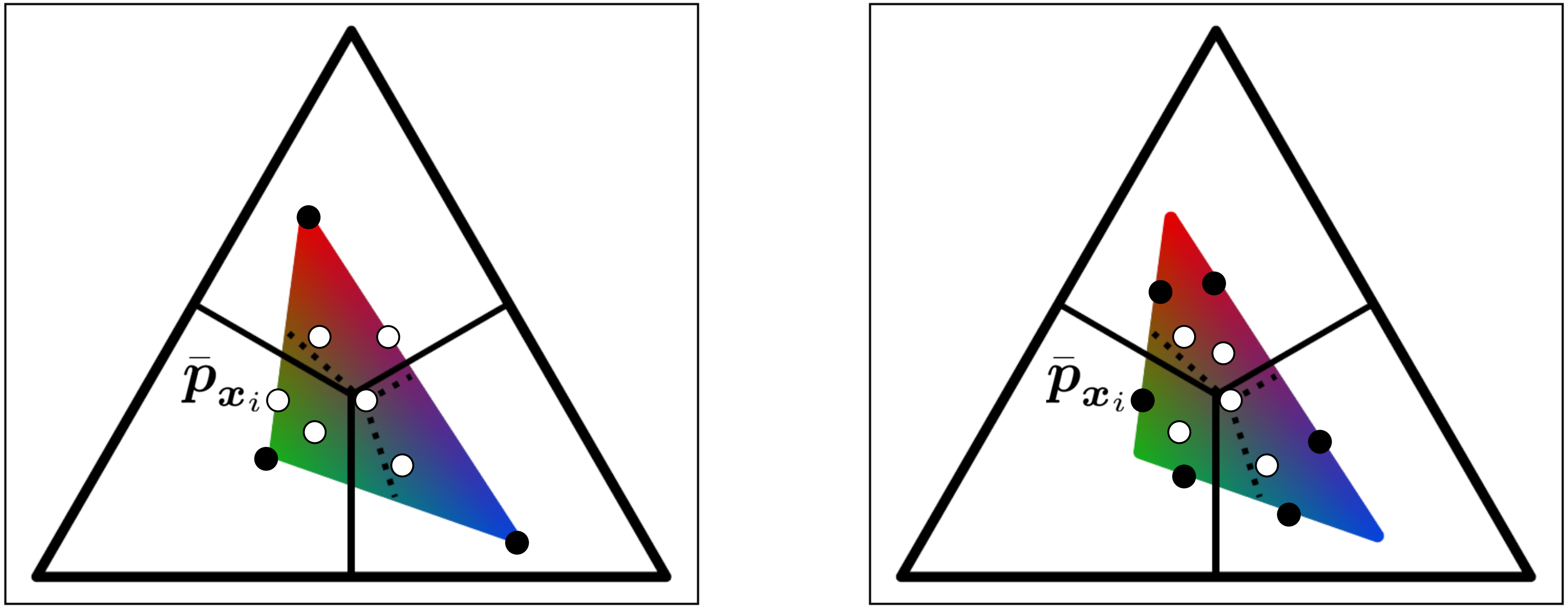


$[\bar{\boldsymbol{p}}_{\boldsymbol{x}}]_{\bar{y}} = \bar{p}(\bar{y}|\boldsymbol{x})$



Non-identifiability of T (cont.)

- Only the existence of anchor points still guarantees the identifiability of T.
- Even without anchor points, "sufficiently scattered" training data can guarantee the consistency (with the algorithm in the next page).

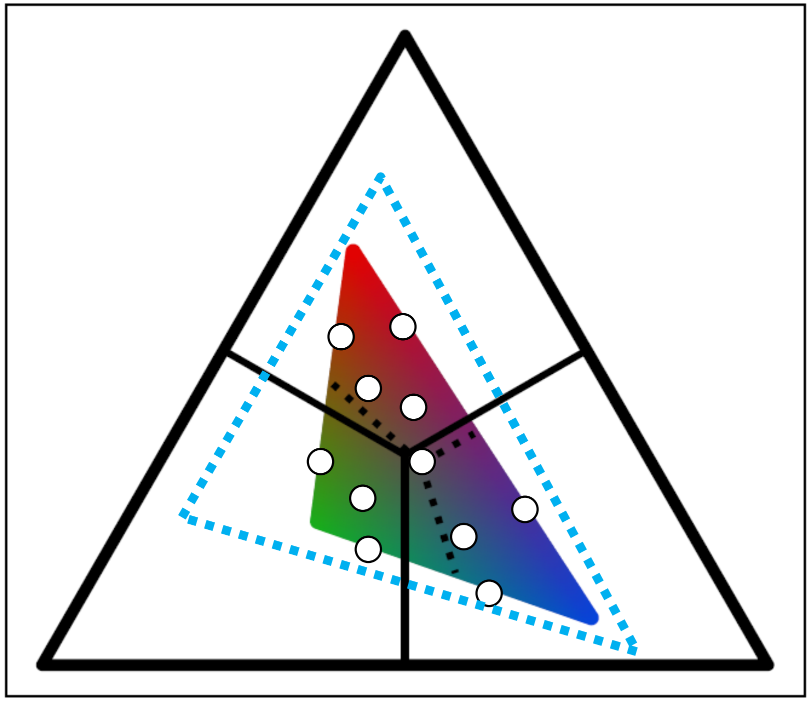


Volume Minimization

Li, Liu, Han, Niu & Sugiyama (ICML2021)

- Under the "sufficiently scattered" assumption, minimizing the volume of the transition matrix guarantees consistency!

$$\min_{\boldsymbol{U},\boldsymbol{h}} \left[\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(\bar{y},\boldsymbol{U}^\top \boldsymbol{h}(\boldsymbol{x}))] + \lambda \log\det(\boldsymbol{U})\right] \quad \lambda > 0$$

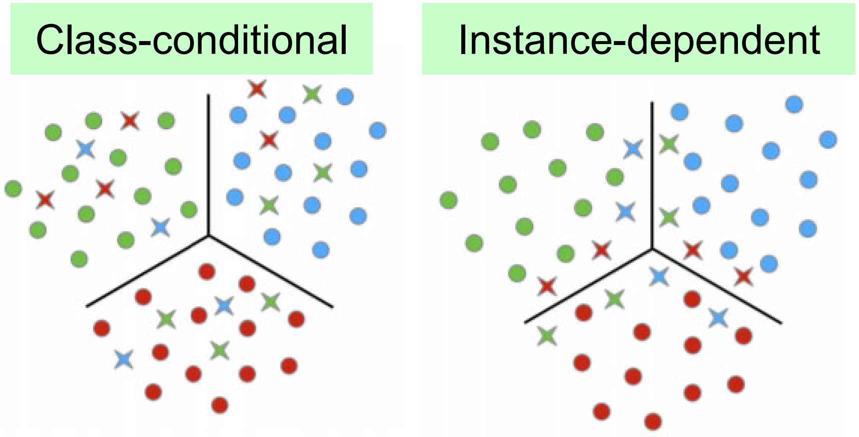


Contents

1. Noisy-Label Learning
   - A) Technical background
   - B) Single-step approach
   - C) Beyond anchor points
   - D) Further challenges
2. Weakly Supervised Learning
3. Transfer Learning
4. Summary

Beyond Class-Conditional Noise

- Instance-independence in class-conditional noise is restrictive.



- Instance-dependent noise: $T_{y,\bar{y}}(\boldsymbol{x}) = \bar{p}(\bar{y}|y,\boldsymbol{x})$
  - Extremely challenging problem!

- Various heuristic solutions:
  - Parts-based estimation

    Xia, Liu, Han, Wang, Gong, Liu, Niu, Tao & Sugiyama (NeurIPS2020)
  - Use of additional confidence scores

    Berthon, Han,Niu, Liu & Sugiyama (ICML2021)
  - Manifold regularization

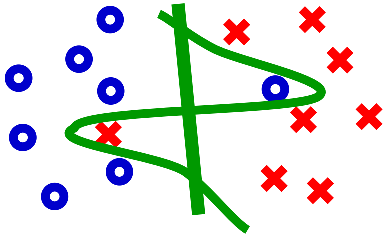
    Cheng, Liu, Ning, Wang, Han, Niu, Gao & Sugiyama (CVPR2022)

Co-teaching

Memorization of neural nets:

Arpit et al. (ICML2017)
Zhang et al. (ICLR2017)

- Stochastic gradient descent fits clean data faster.
- However, naïve early stopping does not work well.



"Co-teaching" between two neural nets:

- Teach small-loss data each other.

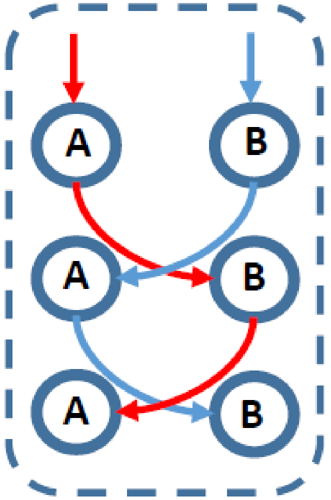
  Han, Yao, Yu, Niu, Xu, Hu, Tsang & Sugiyama (NeurIPS2018)

- Teach only disagreed data.

  Yu, Han, Yao, Niu, Tsang & Sugiyama (ICML2019)

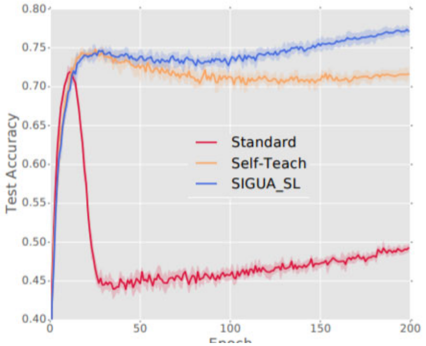
- Gradient ascent for large-loss data.

  Han, Niu, Yu, Yao, Xu, Tsang & Sugiyama (ICML2020)



No theory but very robust in experiments:

- Works well even if 50% random label flipping!

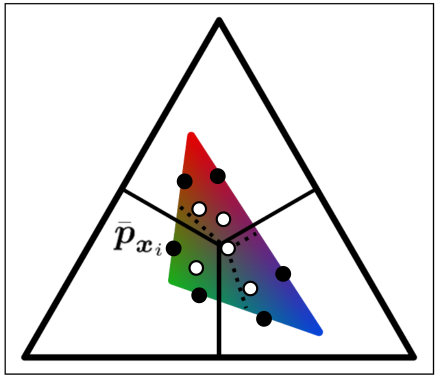


Summary: Noisy-Label Learning

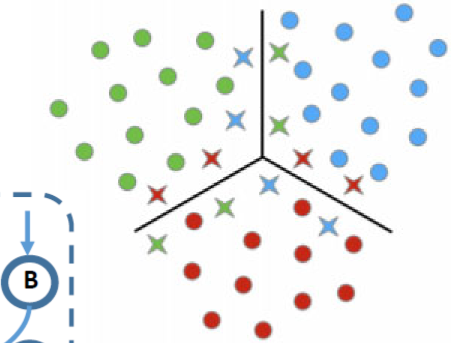
- Classification requires explicit treatment of label noise:
  - Loss correction by noise transition is promising.

$$T_{y,\bar{y}} = \bar{p}(\bar{y}|y)$$

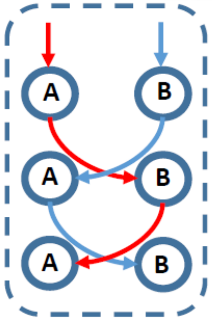
- However, noise transition is generally non-identifiable.
  - Recent development allows its consistent estimation under mild assumptions.



- Real-world noise is often instance-dependent:
  - Heuristic solutions have been developed.



- Super-robustness by co-teaching:
  - Heuristic solutions have been developed.

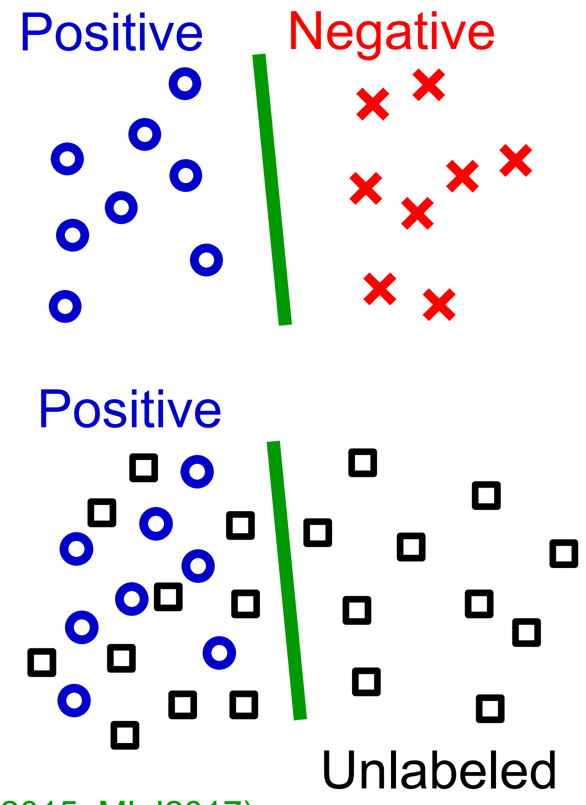


Contents

1. Noisy-Label Learning
2. Weakly Supervised Learning
3. Transfer Learning
4. Summary

Weakly Supervised Learning

- Fully supervised data is expensive to collect.
- Weakly supervised data can be collected easily:
  - Ex.) Click prediction in online ads: It is easy to automatically collect
    - Clicked ads (positive),
    - Unclicked ads (unlabeled).



- Learning only from P and U data is possible!

du Plessis et al. (NIPS2014, ICML2015, MLJ2017), Niu et al. (NIPS2016), Kiryo et al. (NIPS2017), Hsieh et al. (ICML2019)

  - Regard U data as noisy N data and correct the loss.
  - Statistically consistent. $\mathcal{O}_p\left(1/\sqrt{n}\right)$

Solution (Sketch)

du Plessis, Niu & Sugiyama
(NIPS2014, ICML2015)

- Given: Positive and unlabeled data

$$\{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\mathrm{i.i.d.}}{\sim} p(\boldsymbol{x}|y=+1) \qquad \{\boldsymbol{x}_j^{\mathrm{U}}\}_{j=1}^{n_{\mathrm{U}}} \overset{\mathrm{i.i.d.}}{\sim} p(\boldsymbol{x})$$

- Decomposition of the classification risk:

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

ℓ : loss

$\pi = p(y=+1)$:
Class prior (assumed known)

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

Risk for positive data

Risk for negative data

- Eliminate the expectation over negative data as

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

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

- Unbiased risk estimation:

$$\widehat{R}_{\mathrm{PU}}(f) = \frac{\pi}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\Big(f(\boldsymbol{x}_i^{\mathrm{P}})\Big) + \frac{1}{n_{\mathrm{U}}}\sum_{j=1}^{n_{\mathrm{U}}}\ell\Big(-f(\boldsymbol{x}_j^{\mathrm{U}})\Big) - \frac{\pi}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\Big(-f(\boldsymbol{x}_i^{\mathrm{P}})\Big)$$

Theoretical Properties (Sketch)

Niu, du Plessis, Sakai, Ma & Sugiyama (NIPS2016)

$$\widehat{R}_{\mathrm{PU}}(f) = \frac{\pi}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\Big(f(\boldsymbol{x}_i^{\mathrm{P}})\Big) + \frac{1}{n_{\mathrm{U}}}\sum_{j=1}^{n_{\mathrm{U}}}\ell\Big(-f(\boldsymbol{x}_j^{\mathrm{U}})\Big) - \frac{\pi}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\Big(-f(\boldsymbol{x}_i^{\mathrm{P}})\Big)$$

- Optimal parametric convergence rate:

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

with probability $1-\delta$

$$\widehat{f}_{\mathrm{PU}} = \mathrm{argmin}_f\, \widehat{R}_{\mathrm{PU}}(f)$$

$$f^* = \mathrm{argmin}_f\, R(f)$$

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

- Risk correction further improves the performance

Kiryo, Niu, du Plessis & Sugiyama (NIPS2017)

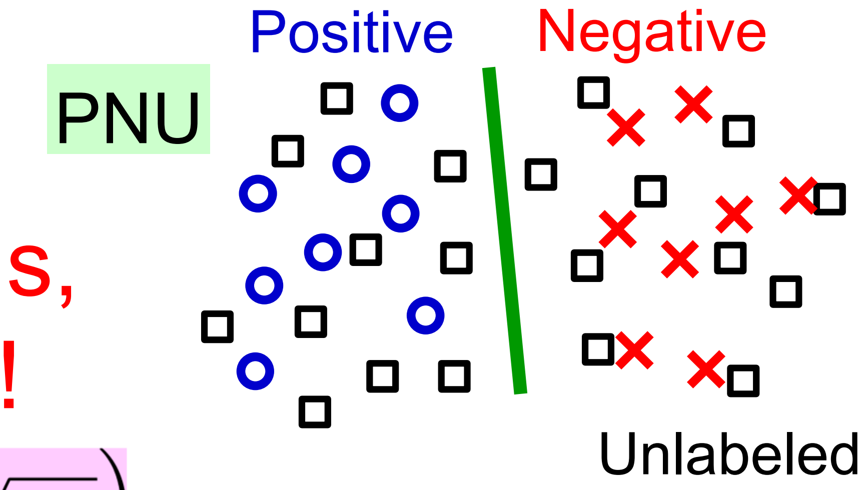
$$\widetilde{R}_{\mathrm{PU}}(f) = \frac{\pi}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\Big(f(\boldsymbol{x}_i^{\mathrm{P}})\Big) + \max\left\{0,\ \frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\ell\Big(-f(\boldsymbol{x}_i^{\mathrm{U}})\Big) - \frac{\pi}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\Big(-f(\boldsymbol{x}_i^{\mathrm{P}})\Big)\right\}$$

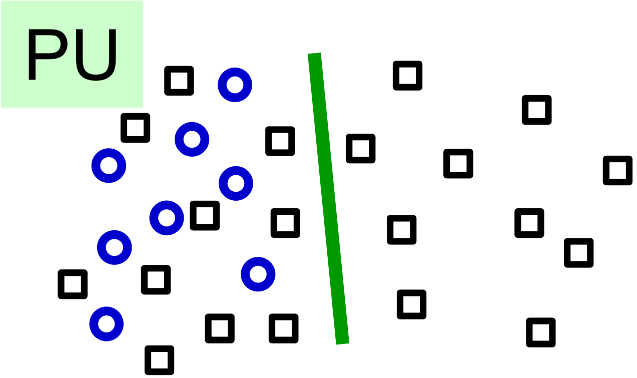
Semi-Supervised Classification (Positive-Negative-Unlabeled Classification)

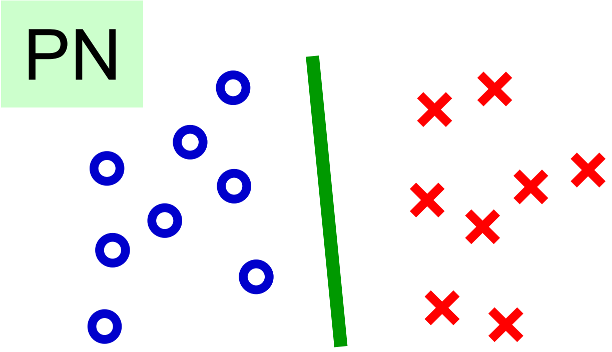
Sakai, du Plessis, Niu & Sugiyama (ICML2017)

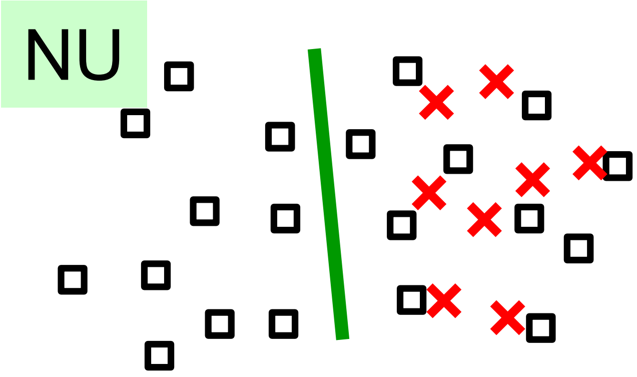
- Let’s decompose PNU into PU, PN, and NU:
  - Each is solvable.
  - Let’s combine them!
- Without cluster assumptions, PN classifiers are trainable!

$$\mathcal{O}_p\left(1/\sqrt{n_\mathrm{P}}+1/\sqrt{n_\mathrm{N}}+1/\sqrt{n_\mathrm{U}}\right)$$





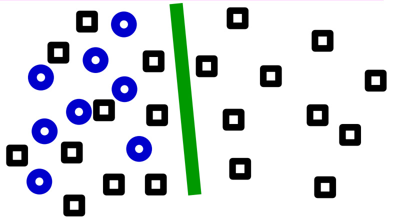




Various Extensions

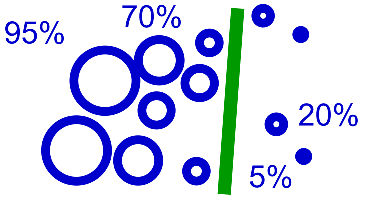
- Learning from weakly supervised data is possible in many different forms!

Positive-Unlabeled



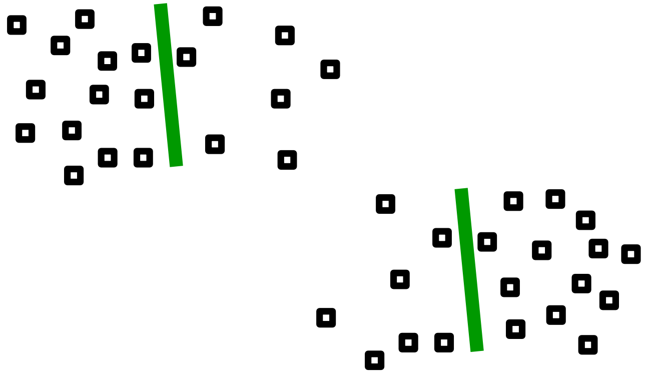
du Plessis et al. (NIPS2014, ICML2015, MLJ2017)
Niu et al. (NIPS2016),, Kiryo et al. (NIPS2017)
Hsieh et al. (ICML2019)

Positive-confidence



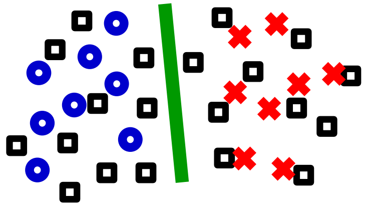
Ishida et al. (NeurIPS2018)
Shinoda et al. (IJCAI2021)

Unlabeled-Unlabeled



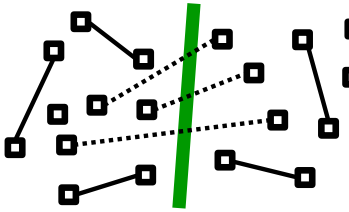
du Plessis et al.,(TAAI2013)
Lu et al. (ICLR2019, AISTATS2020)
Charoenphakdee et al. (ICML2019)
Lei et al. (ICML2021)

Semi-Supervised



Sakai et al. (ICML2017, ML2018)

Similar-Dissimilar

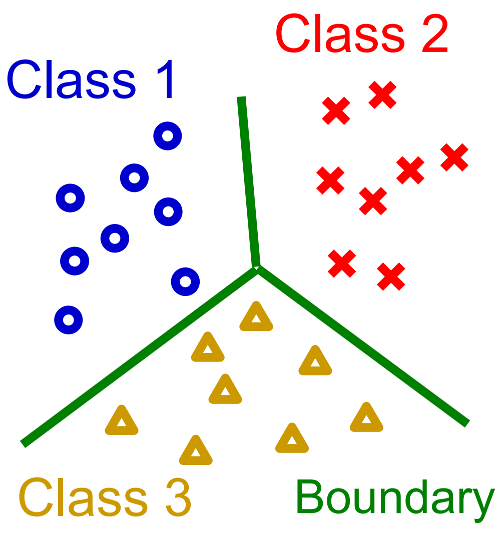


Bao et al. (ICML2018)
Shimada et al. (NeCo2021)
Dan et al. (ECMLPKDD2021)
Cao et al. (ICML2021)
Feng et al. (ICML2021)

$$\mathcal{O}_p\left(1/\sqrt{n}\right)$$

Multiclass Methods

- Labeling patterns in multi-class problems is extremely painful.



- Multi-class weak-labels:
  - Complementary labels: Specify a class that a pattern does not belong to ("not 1").

    Ishida et al. (NIPS2017, ICML2019)
    Chou et al. (ICML2020)
  - Partial labels: Specify a subset of classes that contains the correct one ("1 or 2").

    Feng et al. (ICML2020, NeurIPS2020)
    Lv et al. (ICML2020)
  - Single-class confidence: Cao et al. (arXiv2021)

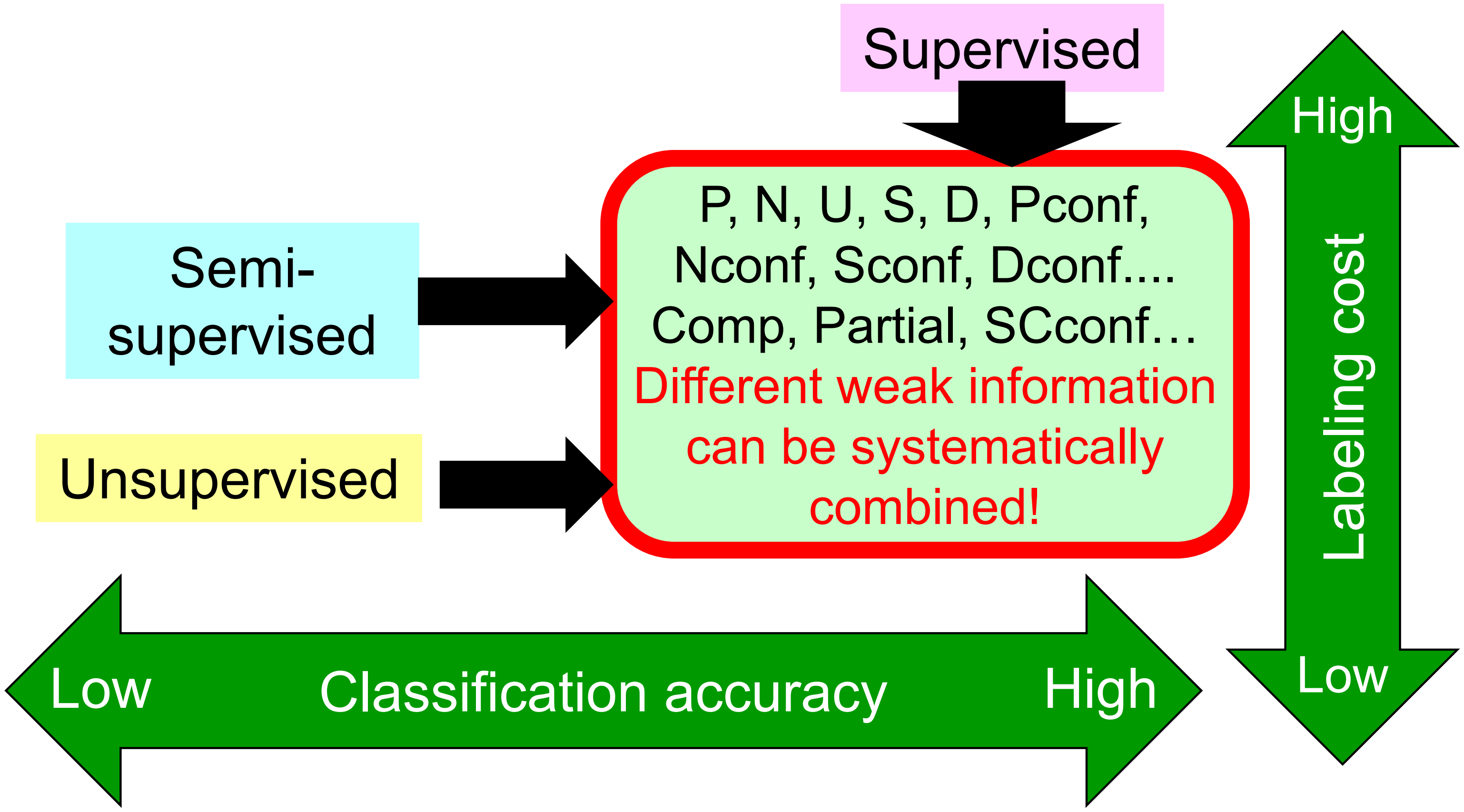
    One-class data with full confidence ("1 with 60%, 2 with 30%, and 3 with 10%")

- Similar loss correction is possible!

$\mathcal{O}_p\left(1/\sqrt{n}\right)$

Summary: Weakly Supervised Learning

- We developed an empirical risk minimization framework for weakly supervised learning:
  - Any loss, classifier, and optimizer can be used.
  - Statistical consistency with optimal convergence.



Supervised

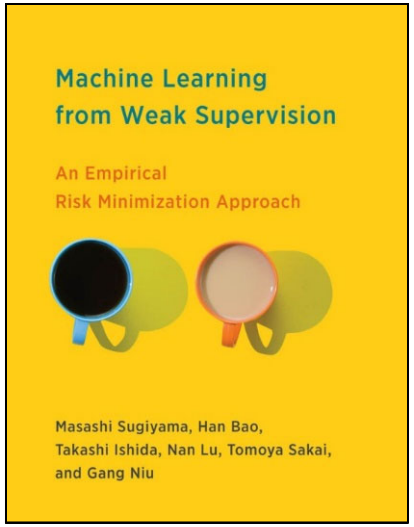
Semi-supervised

Unsupervised

P, N, U, S, D, Pconf, Nconf, Sconf, Dconf.... Comp, Partial, SCconf…

Different weak information can be systematically combined!

Sugiyama, Bao, Ishida, Lu, Sakai & Niu, Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach. MIT Press, August 2022.



Machine Learning from Weak Supervision

An Empirical Risk Minimization Approach

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, Tomoya Sakai, and Gang Niu

Contents

1. Noisy-Label Learning
2. Weakly Supervised Learning
3. Transfer Learning
4. Summary

Transfer Learning

$\boldsymbol{x}$: Input

y : Output

Given:

- Training data $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{tr}}(\boldsymbol{x}, y)$

Goal:

- Train a predictor $y = f(\boldsymbol{x})$
that works well in the test domain
(with some additional data from the test domain).

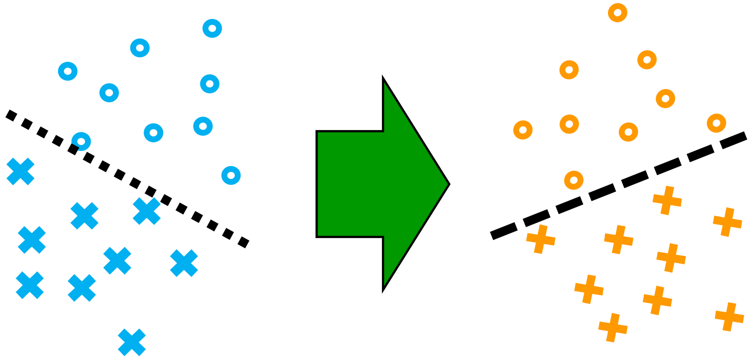
$$\min_f R(f) \qquad R(f) = \mathbb{E}_{p_{\mathrm{te}}(\boldsymbol{x},y)}[\ell(f(\boldsymbol{x}), y)]$$

ℓ : loss function

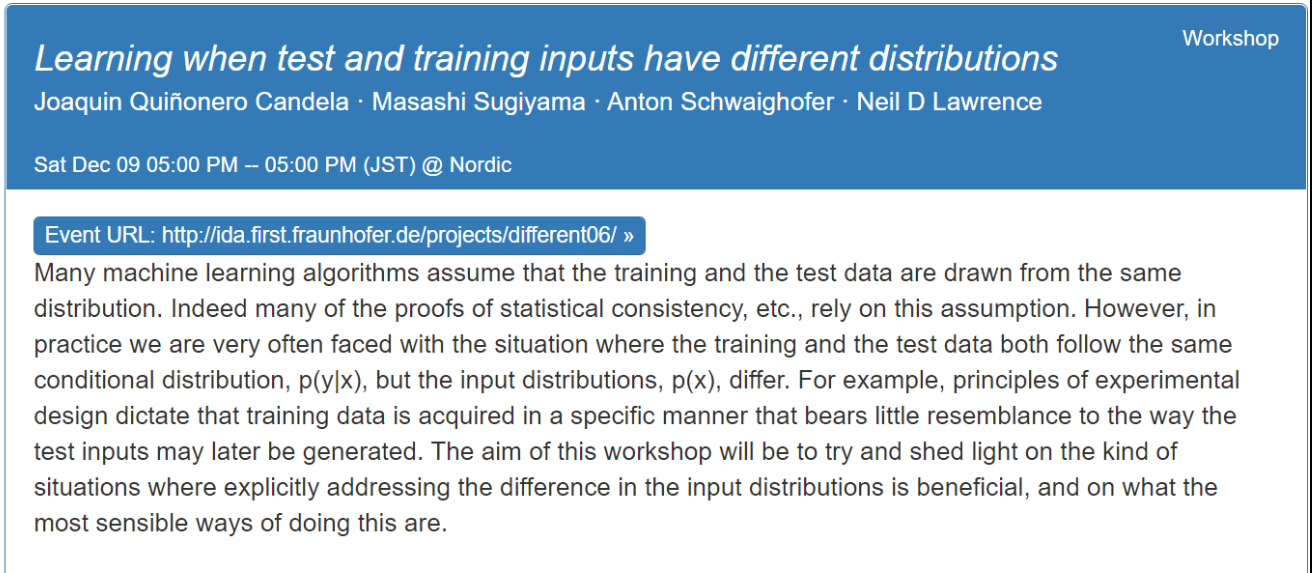
Challenge:

- Overcome changing distributions!

$$p_{\mathrm{tr}}(\boldsymbol{x}, y) \neq p_{\mathrm{te}}(\boldsymbol{x}, y)$$



NIPS Workshop on Learning when Test and Training Inputs Have Different Distributions, Whistler 2006



Workshop

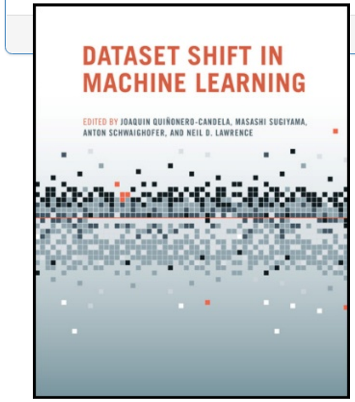
Learning when test and training inputs have different distributions

Joaquin Quiñonero Candela · Masashi Sugiyama · Anton Schwaighofer · Neil D Lawrence

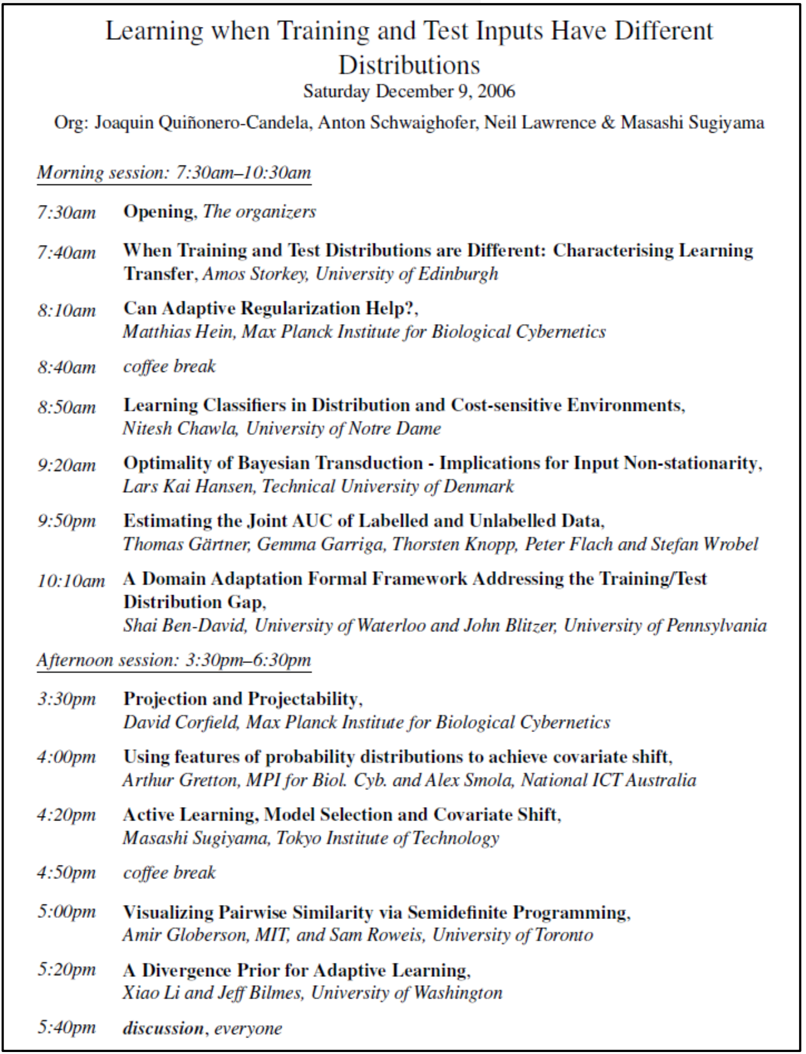
Sat Dec 09 05:00 PM -- 05:00 PM (JST) @ Nordic

Event URL: http://ida.first.fraunhofer.de/projects/different06/ »

Many machine learning algorithms assume that the training and the test data are drawn from the same distribution. Indeed many of the proofs of statistical consistency, etc., rely on this assumption. However, in practice we are very often faced with the situation where the training and the test data both follow the same conditional distribution, p(y|x), but the input distributions, p(x), differ. For example, principles of experimental design dictate that training data is acquired in a specific manner that bears little resemblance to the way the test inputs may later be generated. The aim of this workshop will be to try and shed light on the kind of situations where explicitly addressing the difference in the input distributions is beneficial, and on what the most sensible ways of doing this are.


DATASET SHIFT IN MACHINE LEARNING
EDITED BY JOAQUIN QUIÑONERO-CANDELA, MASASHI SUGIYAMA, ANTON SCHWAIGHOFER, AND NEIL D. LAWRENCE

Quiñonero-Candela, Sugiyama, Schwaighofer & Lawrence (Eds.), Dataset Shift in Machine Learning, MIT Press, 2009.



Learning when Training and Test Inputs Have Different Distributions

Saturday December 9, 2006

Org: Joaquin Quiñonero-Candela, Anton Schwaighofer, Neil Lawrence & Masashi Sugiyama

Morning session: 7:30am–10:30am

7:30am	**Opening**, *The organizers*
7:40am	**When Training and Test Distributions are Different: Characterising Learning Transfer**, *Amos Storkey, University of Edinburgh*
8:10am	**Can Adaptive Regularization Help?**, *Matthias Hein, Max Planck Institute for Biological Cybernetics*
8:40am	*coffee break*
8:50am	**Learning Classifiers in Distribution and Cost-sensitive Environments**, *Nitesh Chawla, University of Notre Dame*
9:20am	**Optimality of Bayesian Transduction - Implications for Input Non-stationarity**, *Lars Kai Hansen, Technical University of Denmark*
9:50pm	**Estimating the Joint AUC of Labelled and Unlabelled Data**, *Thomas Gärtner, Gemma Garriga, Thorsten Knopp, Peter Flach and Stefan Wrobel*
10:10am	**A Domain Adaptation Formal Framework Addressing the Training/Test Distribution Gap**, *Shai Ben-David, University of Waterloo and John Blitzer, University of Pennsylvania*

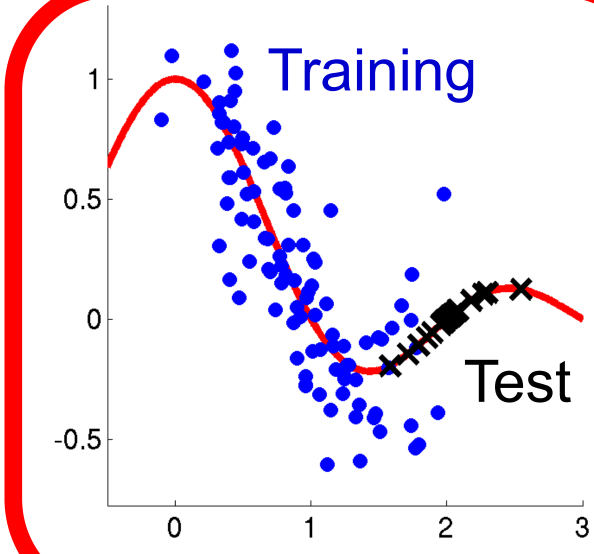
Afternoon session: 3:30pm–6:30pm

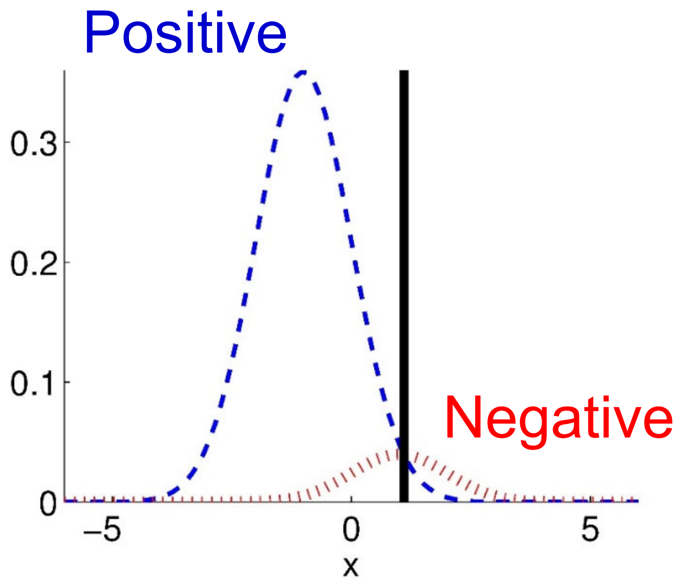
3:30pm	**Projection and Projectability**, *David Corfield, Max Planck Institute for Biological Cybernetics*
4:00pm	**Using features of probability distributions to achieve covariate shift**, *Arthur Gretton, MPI for Biol. Cyb. and Alex Smola, National ICT Australia*
4:20pm	**Active Learning, Model Selection and Covariate Shift**, *Masashi Sugiyama, Tokyo Institute of Technology*
4:50pm	*coffee break*
5:00pm	**Visualizing Pairwise Similarity via Semidefinite Programming**, *Amir Globerson, MIT, and Sam Roweis, University of Toronto*
5:20pm	**A Divergence Prior for Adaptive Learning**, *Xiao Li and Jeff Bilmes, University of Washington*
5:40pm	**discussion**, *everyone*

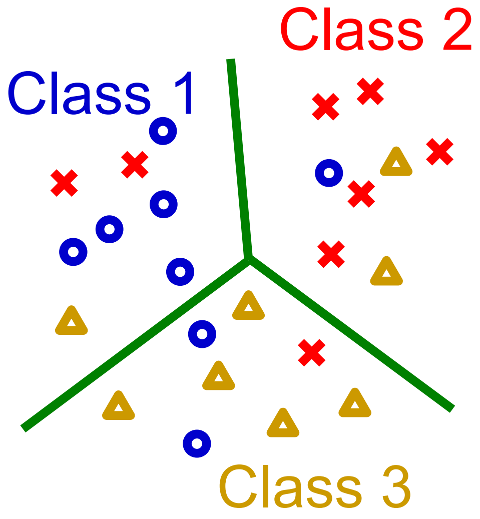
Various Scenarios

$\boldsymbol{x}$: Input y : Output

- Full-distribution shift: $p_{\mathrm{tr}}(\boldsymbol{x}, y) \neq p_{\mathrm{te}}(\boldsymbol{x}, y)$
- Covariate shift: $p_{\mathrm{tr}}(\boldsymbol{x}) \neq p_{\mathrm{te}}(\boldsymbol{x})$
- Class-prior shift: $p_{\mathrm{tr}}(y) \neq p_{\mathrm{te}}(y)$
- Output noise: $p_{\mathrm{tr}}(y|\boldsymbol{x}) \neq p_{\mathrm{te}}(y|\boldsymbol{x})$
- Class-conditional shift: $p_{\mathrm{tr}}(\boldsymbol{x}|y) \neq p_{\mathrm{te}}(\boldsymbol{x}|y)$







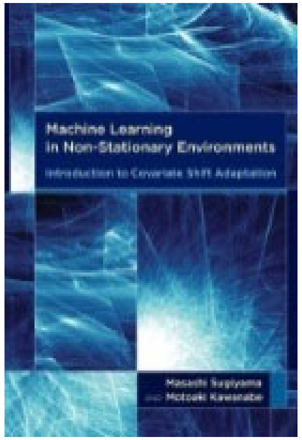
Classical Approach for Transfer Learning

- Two-step adaptation:
  1. Importance weight estimation:

$$\widehat{w} = \mathop{\mathrm{argmin}}_{w} \widehat{\mathbb{E}}_{p_{\mathrm{tr}}(\boldsymbol{x},y)}\left[D\left(w(\boldsymbol{x},y), \frac{p_{\mathrm{te}}(\boldsymbol{x},y)}{p_{\mathrm{tr}}(\boldsymbol{x},y)}\right)\right]$$

  2. Weighted predictor training:

$$\widehat{f} = \mathop{\mathrm{argmin}}_{f} \widehat{\mathbb{E}}_{p_{\mathrm{tr}}(\boldsymbol{x},y)}\left[\widehat{w}(\boldsymbol{x},y)\ell(f(\boldsymbol{x}),y)\right]$$


Machine Learning
in Non-Stationary Environments
Introduction to Covariate Shift Adaptation
Masashi Sugiyama
and Motoaki Kawanabe

Sugiyama & Kawanabe,
Machine Learning
in Non-Stationary
Environments,
MIT Press, 2012

- However, estimation error in Step 1 is not taken into account in Step 2.
  - We want to integrate these two steps!

Joint Weight-Predictor Optimization

- Covariate shift: Only input distributions change.

$p_{\mathrm{tr}}(\boldsymbol{x}) \neq p_{\mathrm{te}}(\boldsymbol{x})$ $p_{\mathrm{tr}}(y|\boldsymbol{x}) = p_{\mathrm{te}}(y|\boldsymbol{x})$ Shimodaira (JSPI2000)

- Suppose we are given
  - Labeled training data: $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{tr}}(\boldsymbol{x}, y)$
  - Unlabeled test data: $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{te}}(\boldsymbol{x})$
- Minimize a risk upper bound jointly w.r.t. weight w and predictor f: $J_{\ell_{\mathrm{tr}}}(f, w) \geq R_{\ell_{\mathrm{te}}}(f)^2$

Zhang et al. (ACML2020, SNCS2021)

$$\widehat{f} = \underset{f}{\mathrm{argmin}} \min_{w \geq 0} \widehat{J}_{\ell_{\mathrm{tr}}}(f, w)$$

$\widehat{J}_\ell$: Empirical approximation of J_ℓ

$$R_\ell(f) = \mathbb{E}_{p_{\mathrm{te}}(\boldsymbol{x}, y)}[\ell(f(\boldsymbol{x}), y)]$$

$$\ell_{\mathrm{te}} \leq 1, \ell_{\mathrm{tr}} \geq \ell_{\mathrm{te}}$$

  - Theoretical guarantee:

$$R_{\ell_{\mathrm{te}}}(\widehat{f}) \leq \sqrt{2} \min_f R_{\ell_{\mathrm{te}}}(f) + \mathcal{O}_p(n_{\mathrm{tr}}^{-1/4} + n_{\mathrm{te}}^{-1/4})$$

Dynamic Importance Weighting

- General changing distributions: $p_{\mathrm{tr}}(\boldsymbol{x}, y) \neq p_{\mathrm{te}}(\boldsymbol{x}, y)$
- Suppose we are given
  - Labeled training data: $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{tr}}(\boldsymbol{x}, y)$
  - Labeled test data: $\{(\boldsymbol{x}_i^{\mathrm{te}}, y_i^{\mathrm{te}})\}_{i=1}^{n_{\mathrm{te}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{te}}(\boldsymbol{x}, y)$
- For each mini-batch $\{(\bar{\boldsymbol{x}}_i^{\mathrm{tr}}, \bar{y}_i^{\mathrm{tr}})\}_{i=1}^{\bar{n}_{\mathrm{tr}}}, \{(\bar{\boldsymbol{x}}_i^{\mathrm{te}}, \bar{y}_i^{\mathrm{te}})\}_{i=1}^{\bar{n}_{\mathrm{te}}}$, importance weights are estimated by matching losses by kernel mean matching:

Fang et al. (NeurIPS2020)

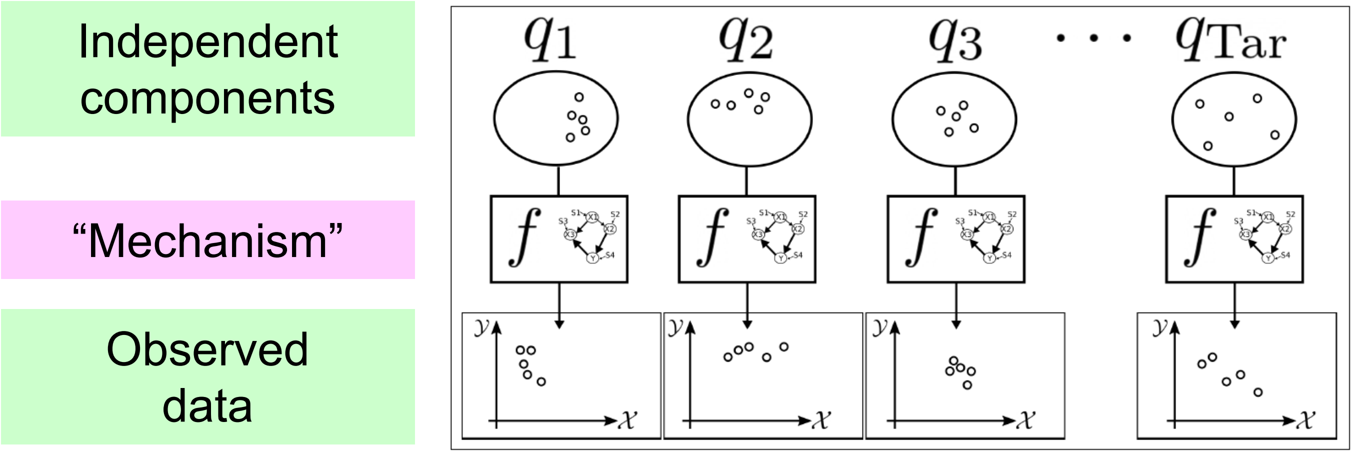
Huang et al. (NeurIPS2007)

$$\frac{1}{\bar{n}_{\mathrm{tr}}} \sum_{i=1}^{\bar{n}_{\mathrm{tr}}} r_i \ell(f(\bar{\boldsymbol{x}}_i^{\mathrm{tr}}), \bar{y}_i^{\mathrm{tr}}) \approx \frac{1}{\bar{n}_{\mathrm{te}}} \sum_{j=1}^{\bar{n}_{\mathrm{te}}} \ell(f(\bar{\boldsymbol{x}}_j^{\mathrm{te}}), \bar{y}_j^{\mathrm{te}})$$

- Extremely simple, but highly powerful!

Summary: Transfer Learning

- In transfer learning with importance weighting, simultaneously estimation of importance and predictor is promising.
- What should we do if training and test distributions look very different?
  - Mechanism transfer! Teshima, Sato & Sugiyama (ICML2020)



Independent components

"Mechanism"

Observed data

- New work: Continuous distribution change. Bai, Zhang, Zhao, Sugiyama & Zhou (NeurIPS2022)

Contents

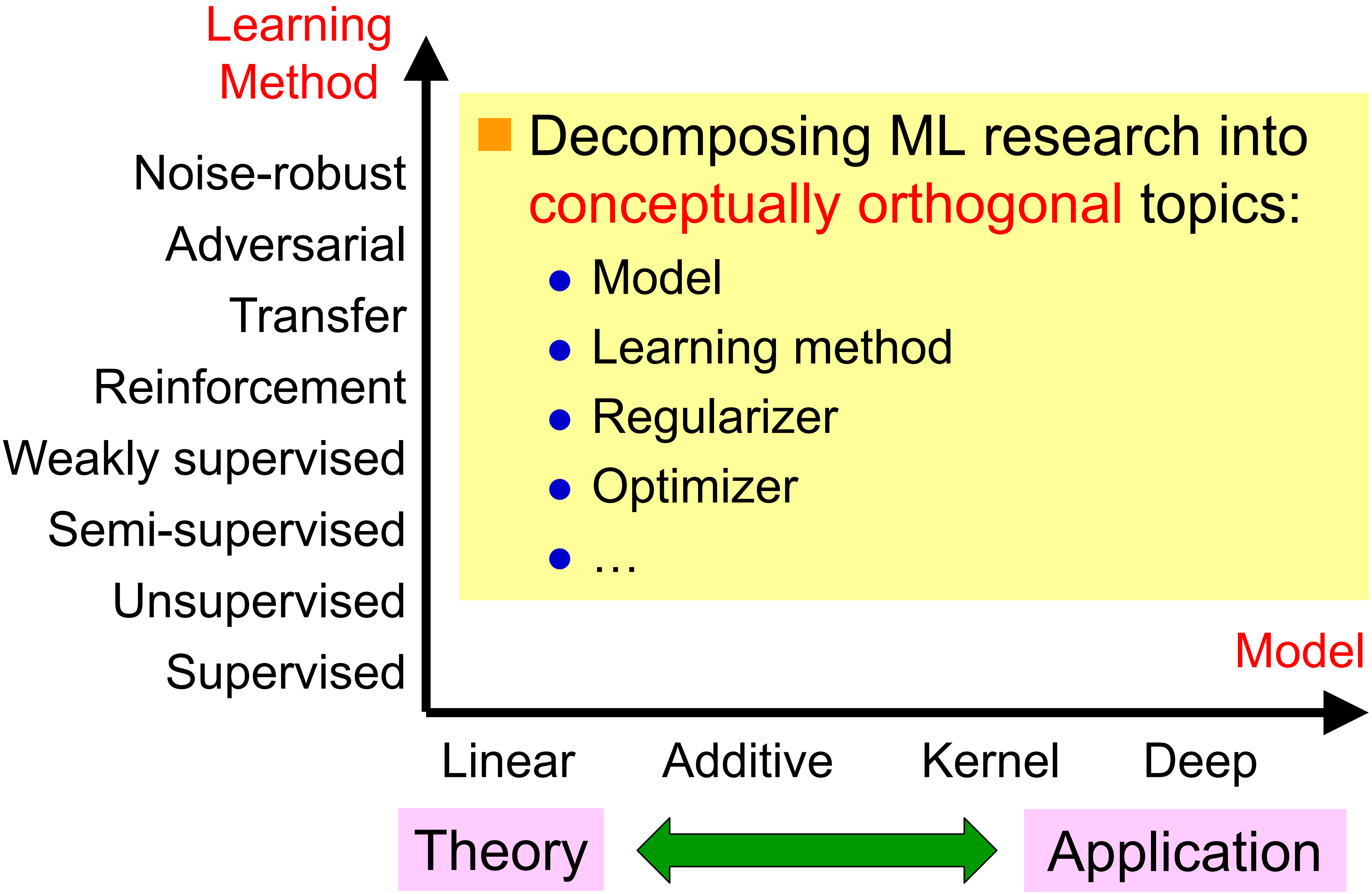
1. Noisy-Label Learning
2. Weakly Supervised Learning
3. Transfer Learning
4. Summary

More Challenges in Reliable Machine Learning

- Reliability for expectable situations:
  - Model the corruption process explicitly and correct the solution.
    - How to handle modeling error?
- Reliability for unexpected situations:
  - Consider worst-case robustness (“min-max”).
    - How to make it less conservative?
  - Include human support (“rejection”).
    - How to handle real-time applications?
- Exploring somewhere in the middle would be practically more useful:
  - Use partial knowledge of the corruption process.

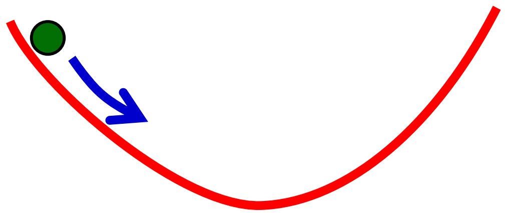
Axes of ML Research

- Decomposing ML research into conceptually orthogonal topics:
  - Model
  - Learning method
  - Regularizer
  - Optimizer
  - …

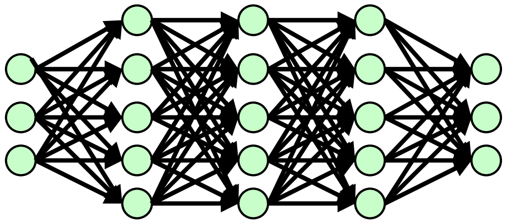


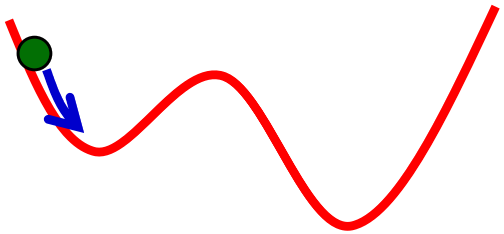
Further Investigations Needed

- Classical convex learning methods allow us to analyze the global solution.



- Since optimization in deep learning is complex, stochastic gradient descent is used.





- Thanks to the “gradual learning” nature, we can utilized intermediate learning results:
  - Strengthening supervision for weakly supervised learning.
  - Dynamic importance weighting for transfer learning.
  - Dynamic noise transition estimation for noise-robust learning.
  - Co-teaching for noise-robust learning.