International Symposium on
Artificial Intelligence and Brain Science

July 4, 2022

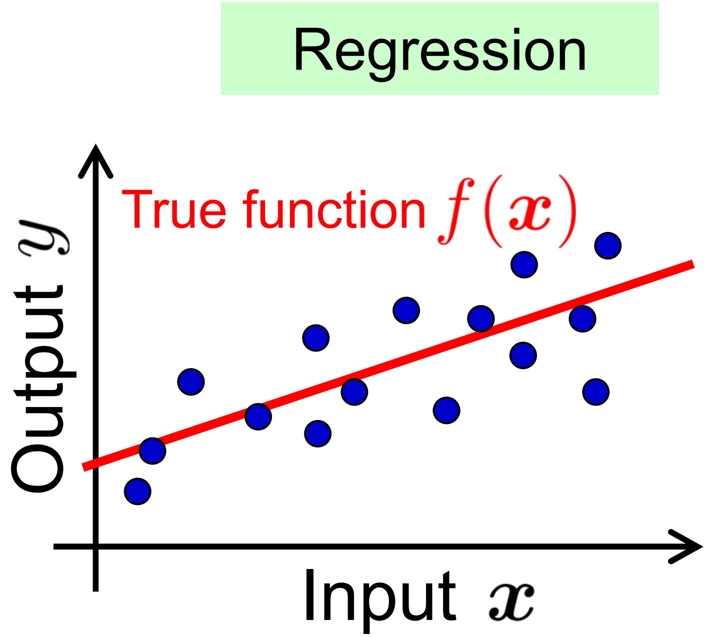
Recent Advances in Classification from Noisy Labels

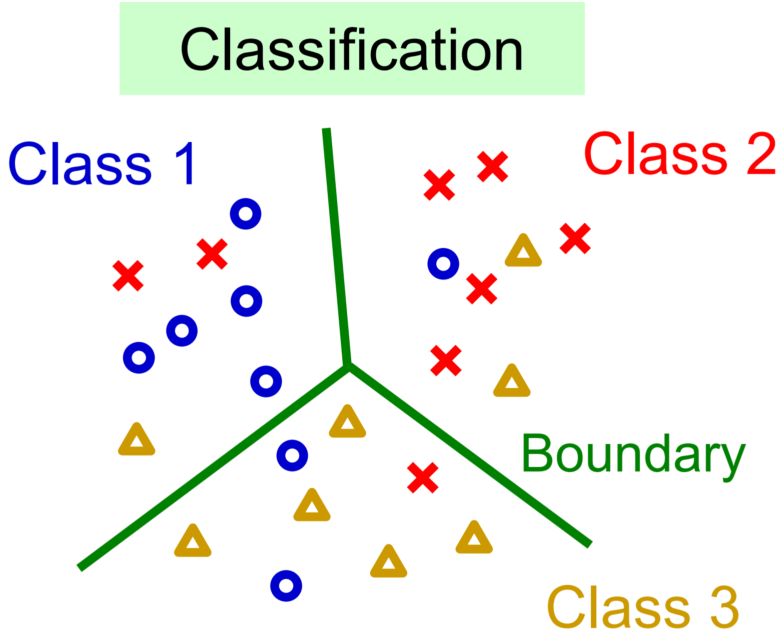
Masashi Sugiyama

RIKEN Center for Advanced Intelligence Project/
The University of Tokyo

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

Supervised Learning from Noisy Output Data





- Classical problem. Nothing to study further?
  - For regression, just using big data is fine.
  - For classification, big data doesn't necessarily help. Need further study to cope with label noise!

Contents

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

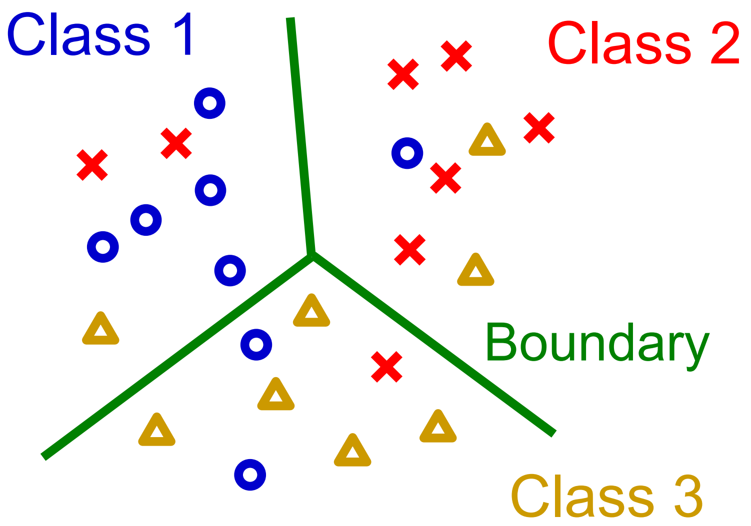
$\boldsymbol{x} \in \mathbb{R}^d$: Input instance

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

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

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

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



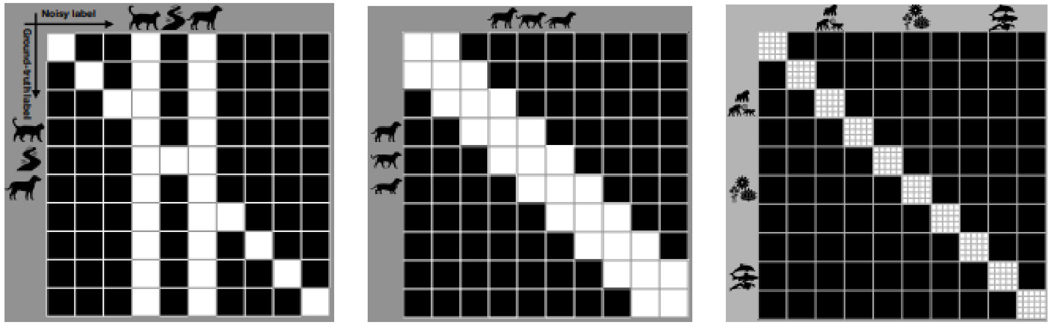
- Loss: $\ell(y, \boldsymbol{h}(\boldsymbol{x})) \in \mathbb{R}$

Modeling Class-Conditional Noise

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

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

- We may encode a human-cognitive bias:

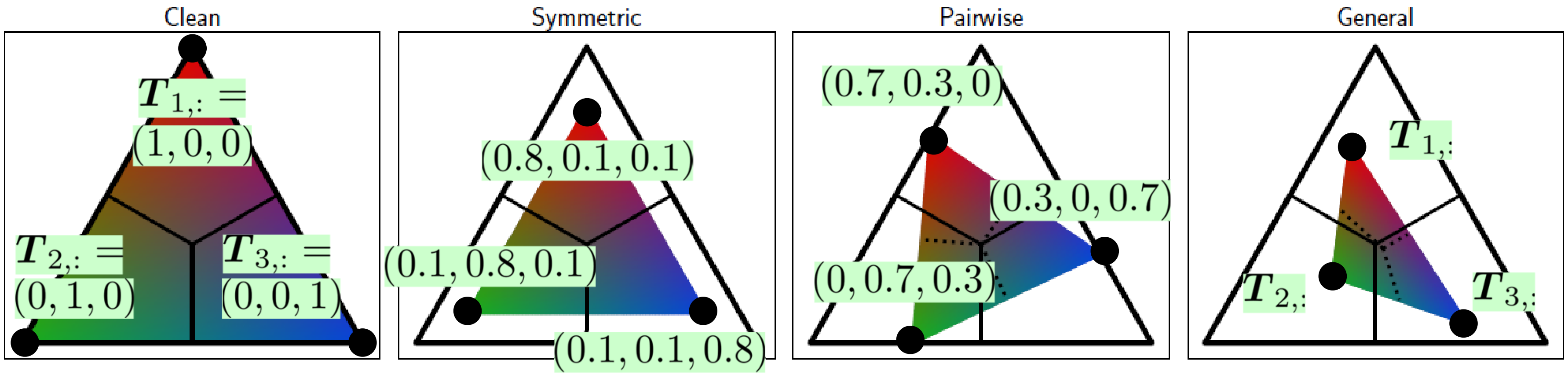


(a) Column-diagonal (b) Tri-diagonal (c) Block-diagonal

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 $\boldsymbol{T}^\top$
  - $\boldsymbol{\ell}^{\rightarrow}(\boldsymbol{h}(\boldsymbol{x})) = \boldsymbol{\ell}(\boldsymbol{T}^\top \boldsymbol{h}(\boldsymbol{x}))$ $\qquad \ell_y(\boldsymbol{h}(\boldsymbol{x})) = \ell(y, \boldsymbol{h}(\boldsymbol{x}))$

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

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

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

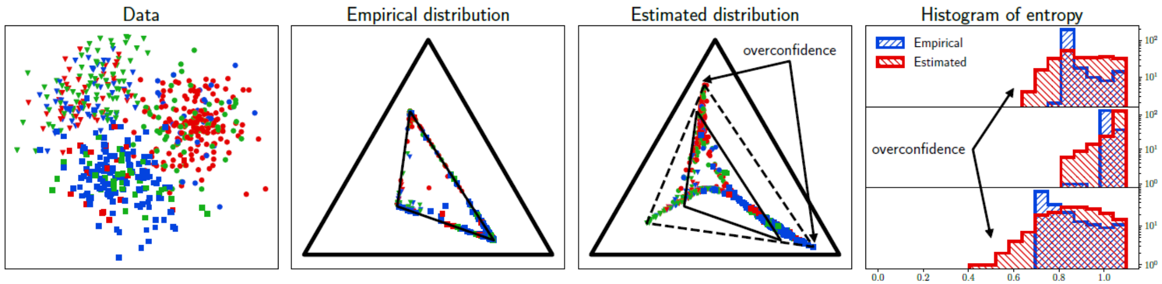
  - $\bar{\boldsymbol{h}}(\boldsymbol{x})$ is a probabilistic classifier learned from noisy training data $\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n$.

- 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

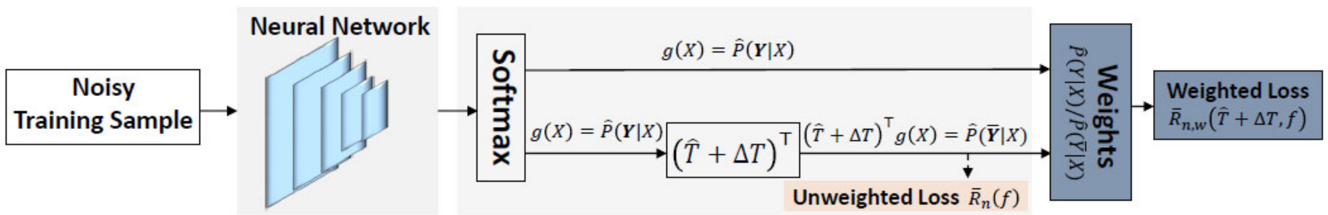
$$\boldsymbol{x}^y \leftarrow \boldsymbol{x}_i \text{ s.t. } \bar{h}_y(\boldsymbol{x}_i) \approx 1$$

- 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)

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

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

Contents

1. Formulation and existing results
2. One-step solution
3. Beyond anchor points
4. Further investigations

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

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

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

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

- 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.

- 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

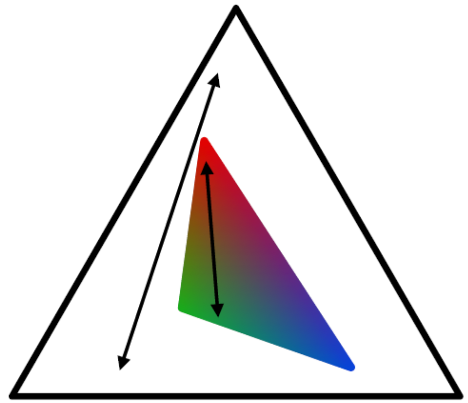
Zhang, Niu & Sugiyama (ICML2021)

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

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

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

  - Cleaner class-posteriors have a larger total variation distance!



- Let's use this knowledge as a regularizer:

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

$\lambda > 0$

  - Under the anchor point assumption, the empirical solution has statistical consistency.

Contents

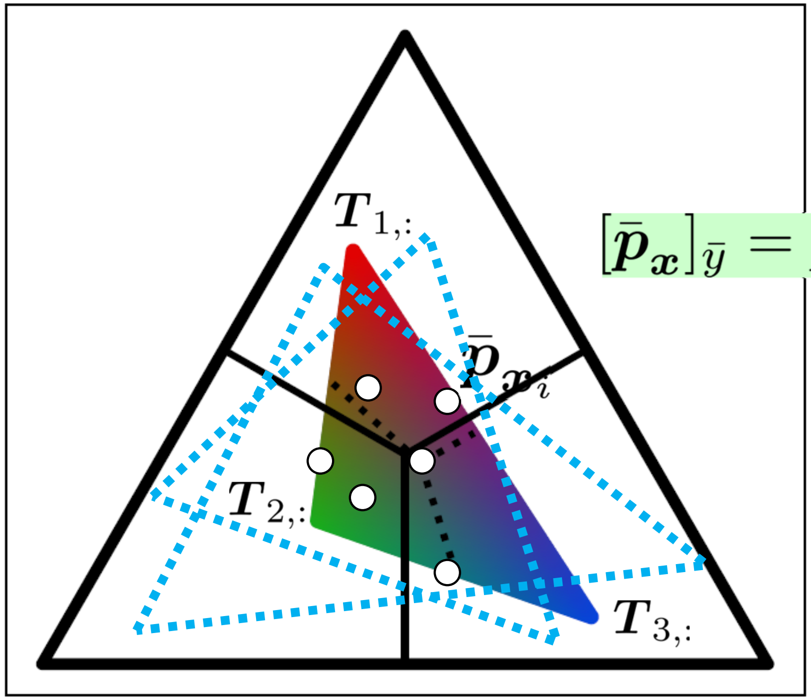
Beyond Anchor Points

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

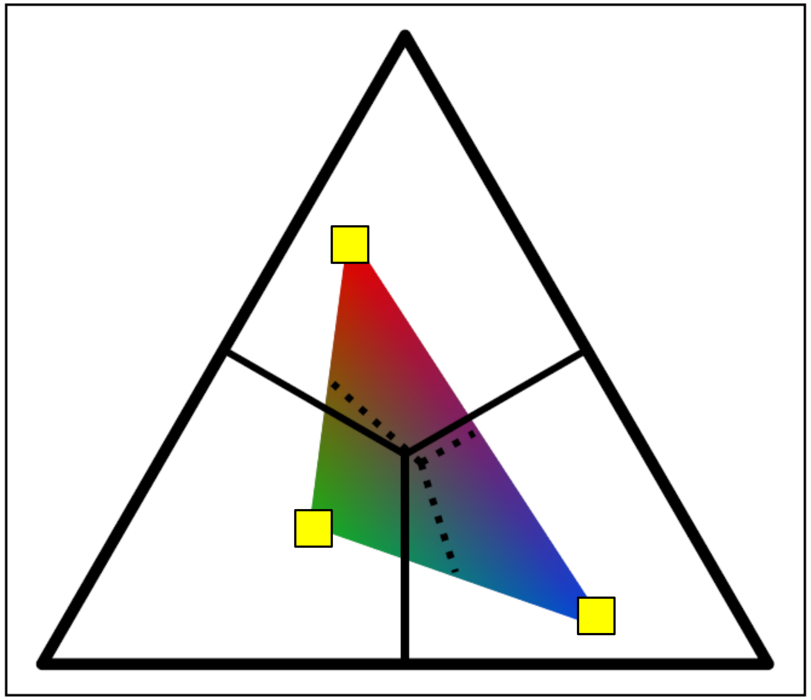
- To overcome the non-identifiability of $\boldsymbol{T}$:
  - Initial works used given anchor points explicitly.
- Later, it was relaxed to only assuming
  - Existence of anchor points in training data.
- Further, it was relaxed to assuming
  - Only existence of anchor regions (no noise regions) in the true distribution.
- Can we further relax this assumption?
  - Anchor regions rarely exist in reality.

Non-identifiability of $\boldsymbol{T}$

- $\boldsymbol{T}$ can be visualized as a simplex (triangle), 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 $\boldsymbol{T}$.

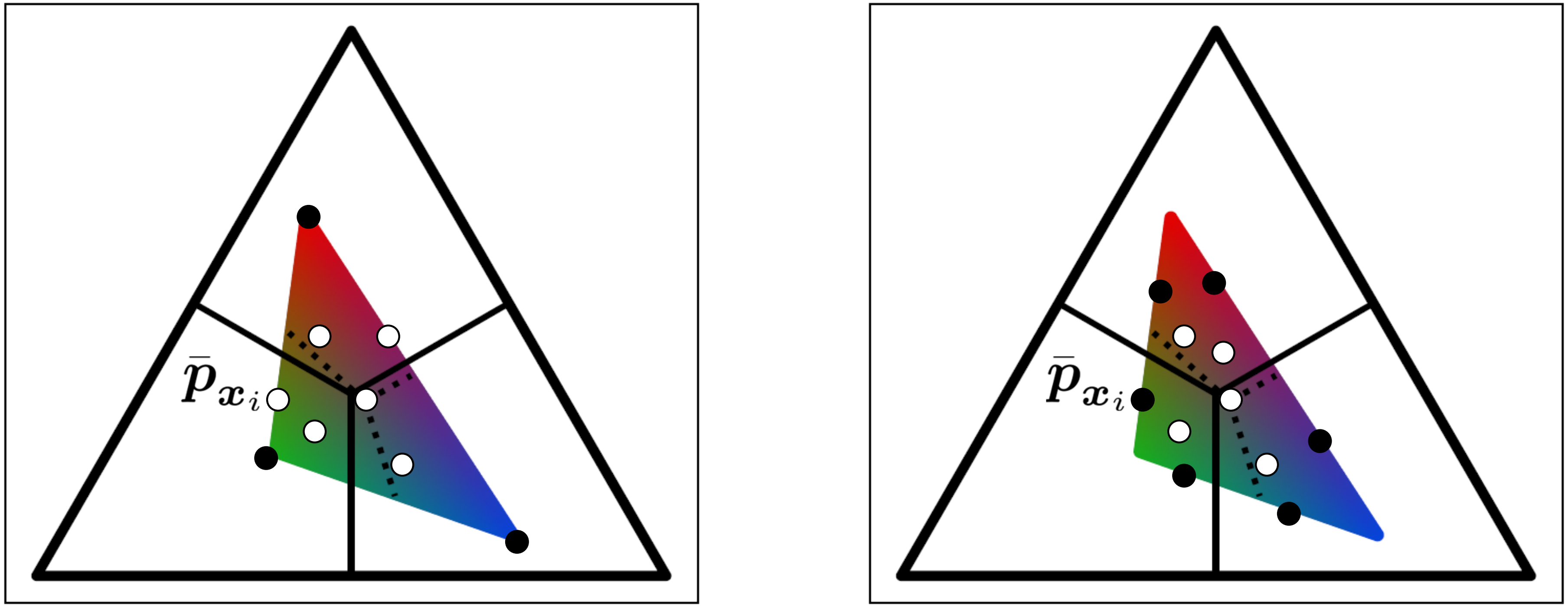


$[\bar{\boldsymbol{p}}_{\boldsymbol{x}}]_{\bar{y}} = \bar{p}(\bar{y}|\boldsymbol{x})$



Non-identifiability of $\boldsymbol{T}$ (cont.)

- Only the existence of anchor points still guarantees the identifiability of $\boldsymbol{T}$.
- Even without anchor points, "sufficiently scattered" training data can guarantee the identifiability.

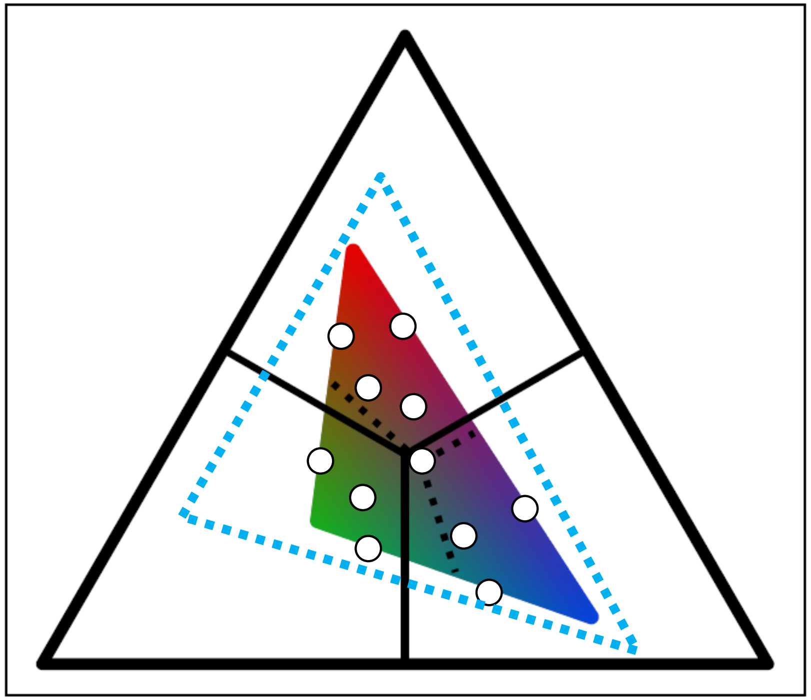


Volume Minimization

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

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

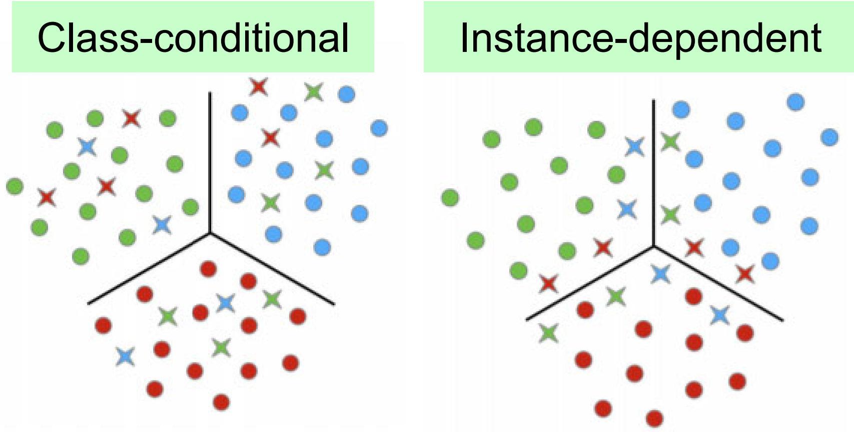
$$\min_{\boldsymbol{U},\boldsymbol{h}} \left[\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(\bar{y}, \boldsymbol{U}^\top \boldsymbol{h}(\boldsymbol{x}))] + \lambda \log\det(\boldsymbol{U}) \right] \quad \lambda > 0$$



Contents

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 new solutions emerge:
  - 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)

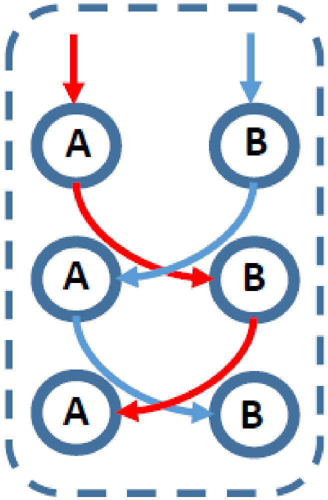
  - Stochastic gradient descent fits clean data faster.
  - However, naïve early stopping does not work well.

- "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% labels are randomly flipped.

Machine Learning, Neuroscience, and Society

- So far, various neuroscientific findings were brought to machine learning with great success:
  - Learning rule, model architecture, adversarial attack, etc.
  - How humans handle noisy observations?
- Beyond performance improvement, next-generation AI should take into account various constraints in society:
  - Culture, common sense, ethics, curiosity, friendliness, etc.
- Combining
  - neuroscientific findings (internal learning mechanism)
  - social demands (external constraints)

would be a promising direction.