



ICML2019

# Breaking Inter-Layer Co-Adaptation by Classifier Anonymization

Ikuro Sato<sup>1</sup>

Kohta Ishikawa<sup>1</sup>

Guoqing Liu<sup>1</sup>

Masayuki Tanaka<sup>2</sup>

<sup>1</sup> Denso IT Laboratory. Inc., Japan

<sup>2</sup> National Institute of Advanced Industrial  
Science and Technology, Japan

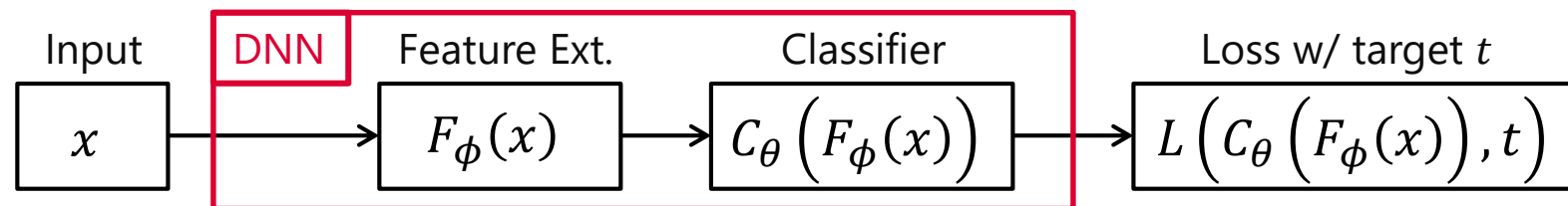
# Summary first

---

- About what?*     **Breaking co-adaptation** between feature extractor and classifier.
- How?*     By **classifier anonymization** technique.
- Theory?*     Proved: Features form simple **point-like distribution**.
- In reality?*     Point-like property largely confirmed on real datasets.

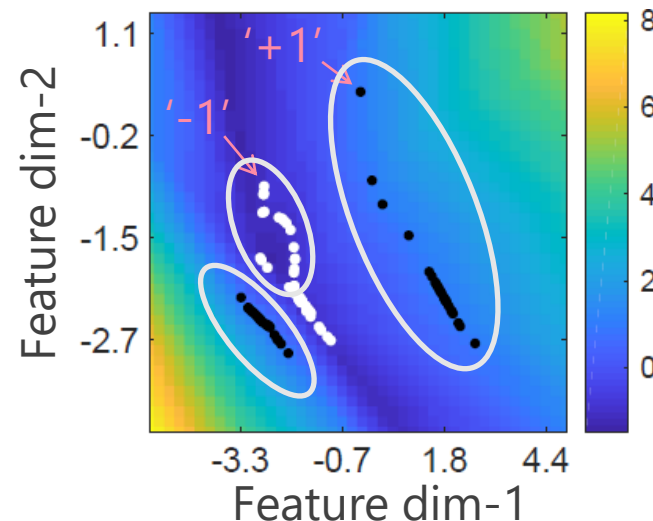
# E2E optimization scheme flourishes. Is it always good?

**E2E opt.**  $(\phi^*, \theta^*) = \arg \min_{\phi, \theta} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t) \in \mathcal{D}} L(C_\theta(F_\phi(x)), t)$



Feature extractor  $F_{\phi^*}$  adapts to a **particular classifier  $C_\theta$** .

Toy ex.)  
2-class regression



color:  $C_\theta$  value

Features may form  
**excessively complex distribution.**

- Disjointed
- Split



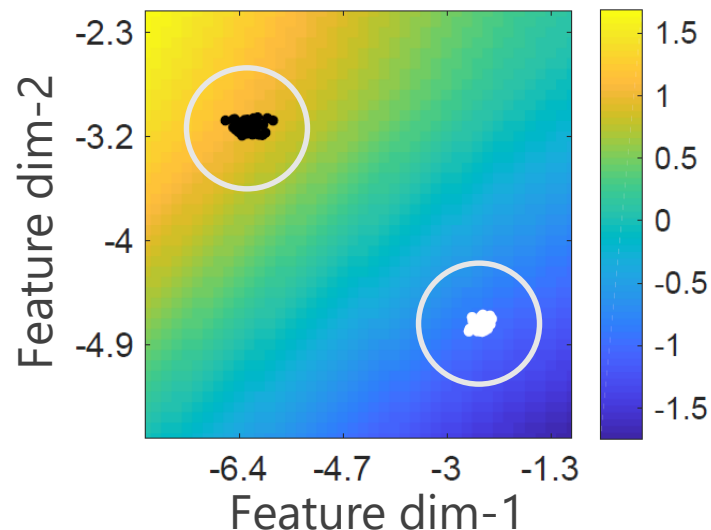
# FOCA: Feature-extractor Optimization through Classifier Anonymization

**FOCA** 
$$\phi^* = \arg \min_{\phi} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t) \in \mathcal{D}} \mathbb{E}_{\theta \sim \Theta_{\phi}} L(C_{\theta}(F_{\phi}(x)), t)$$

Random weak classifier:  $\theta \sim \Theta_{\phi}$

*Want to know  
more about  $\Theta_{\phi}$ ?  
Please come to  
the poster!*

Feature extractor  $F_{\phi^*}$  adapts to a set of weak classifiers  $\{C_{\theta}\}$ .



Features form simple  
**point-like distribution**  
per class under some  
conditions.

# Proposition about the **point-like** property

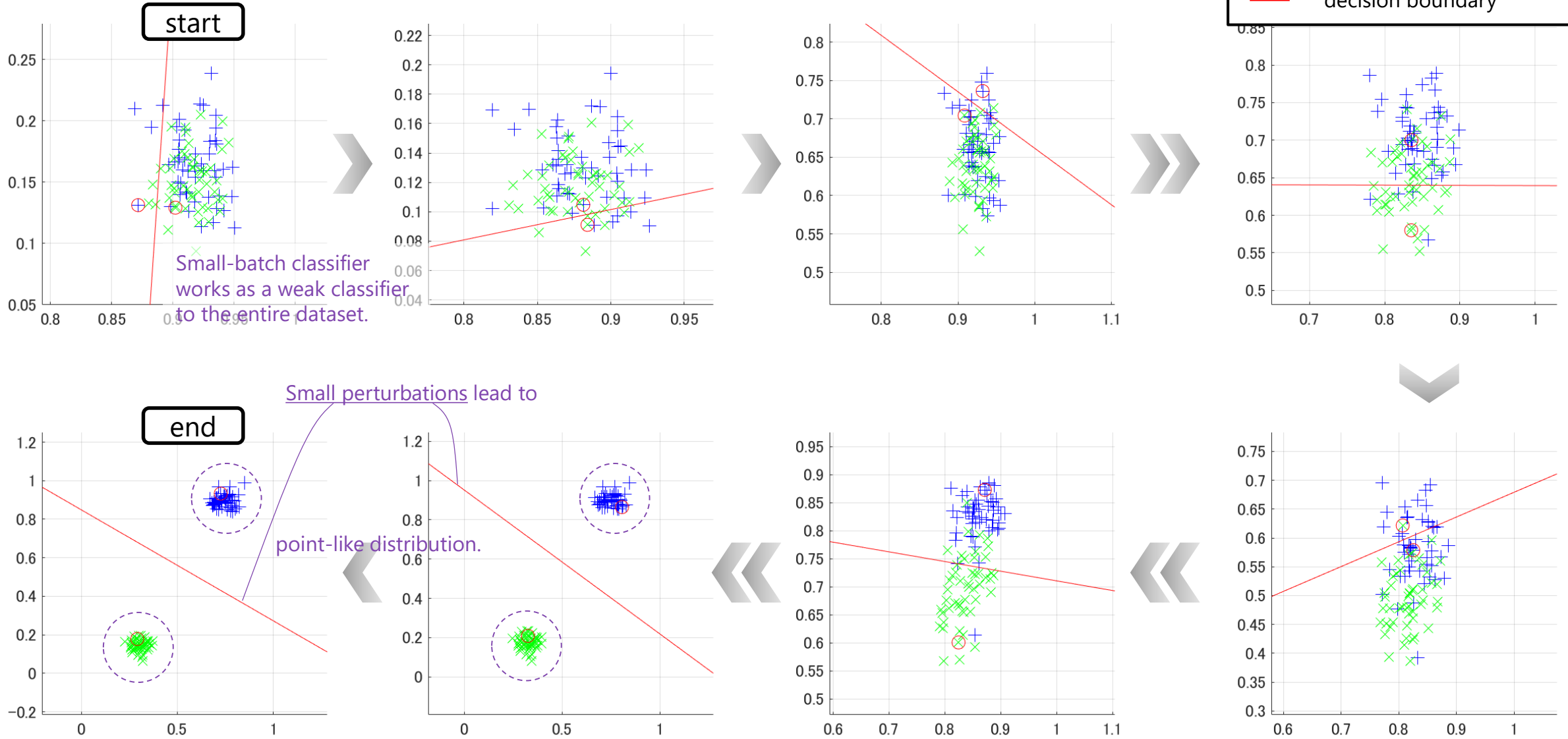
In words,

*If feature extractor has an enough representation ability,  
all input data of the same class are projected to a single point  
in the feature space in a class-separable way under  
certain conditions.*

*Please see  
the paper  
for the proof.*

# Toy problem demonstration

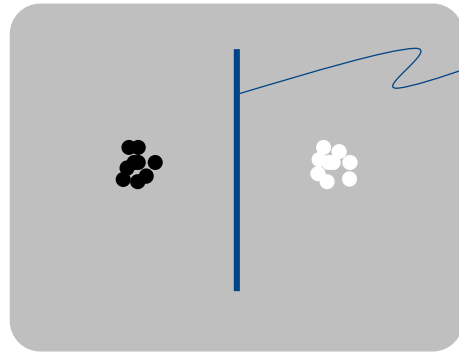
x-axis	Feature dim. #1
y-axis	Feature dim. #2
	data used to generate classifier
	decision boundary



# Experiment #1: partial-dataset training

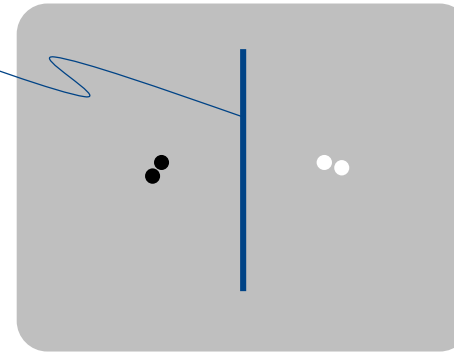
Thing we wish to confirm:

full-dataset classifier

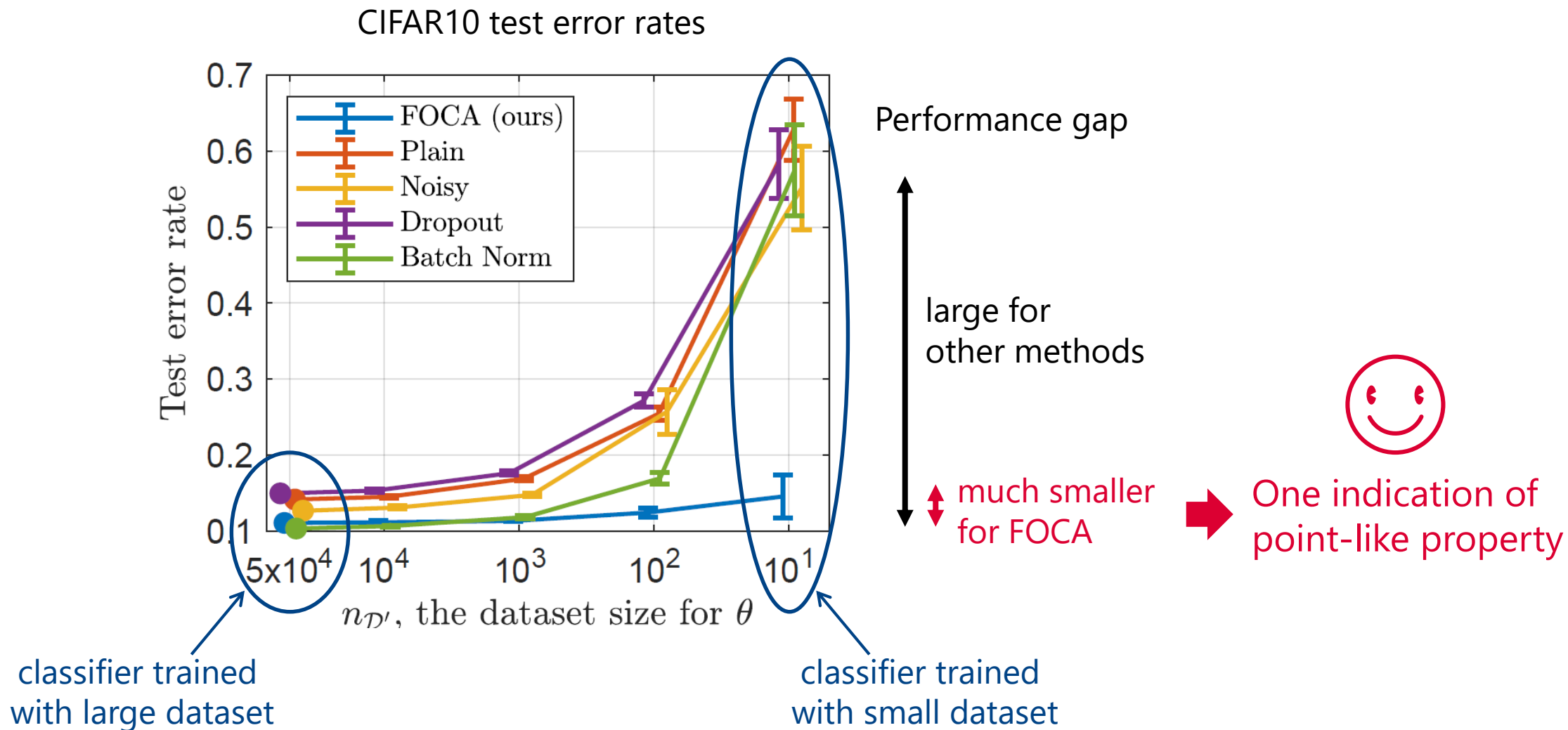


*Do they  
perform  
similarly  
for given  $F_{\phi^*}$   
??*

partial-dataset classifier



# Experiment #1: partial-dataset training



(The same, fixed feature extractor is used within each method.)



# More experiments ...

---

including:

- Approximate geodesic distance measurements between large- and small-dataset solutions
- Low-dimensional analyses

to further study the point-like property.

# Poster #28 tonight

#28  
Wed. June 12  
ICML 2019

## Breaking Inter-Layer Co-Adaptation by Classifier Anonymization

Ikuro Sato<sup>1</sup>, Kohta Ishikawa<sup>1</sup>, Guoqing Liu<sup>1</sup>, and Masayuki Tanaka<sup>2</sup>

<sup>1</sup>Denso IT Laboratory, Janan  
<sup>2</sup>AIST, Japan

### 1. Summary

**About what?** Breaking co-adaptation between feature extractor and classifier.

**How?** By **classifier anonymization** technique.

**Theory?** Proved: Features form simple **point-like distribution**.

**In reality?** Point-like property largely confirmed on real datasets.

### 2. Possible problem with co-adaptation

**E2E opt.**

$$(\phi^*, \theta^*) = \arg \min_{\phi, \theta} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t) \in \mathcal{D}} L(C_\theta(F_\phi(x)), t) \quad (1)$$

Feature extractor  $F_{\phi^*}$  adapts to a **particular classifier**  $C_\theta$ .

Features may form **excessively complex distribution**.

### 3. FOCA: Feature-extractor Optimization through Classifier Anonymization

**FOCA**  $\phi^* = \arg \min_{\phi} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t) \in \mathcal{D}} \mathbb{E}_{\theta \sim \Theta_\phi} L(C_\theta(F_\phi(x)), t) \quad (2)$

$\Theta_\phi = \mathcal{U}(\{\theta_{\phi,b}; b = b_1, b_2, \dots\})$ ,  $\mathcal{U}$ : discrete uniform dist.  $(3)$

$\theta_{\phi,b} = \arg \min_{\theta} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t) \in \mathcal{D}} L(C_\theta(F_\phi(x)), t) + \lambda \|\theta\|_2^2 \quad (4)$

Classifier of small data subset  $b$  (works as weak classifier to the entire dataset).

$F_{\phi^*}$  adapts to a set of random, weak classifiers  $\{C_\theta\}$ .

Features are expected to form **simple distribution**.

**Algorithm 1** Approximate minimization in Eq. (2)

**Input:** total number of iterations  $T$ ; number of classes  $C$ ; number of class samples  $n_c$  ( $c = 1, \dots, C$ ); number of samples per class for  $\theta$ -update  $k$ ; total number of samples  $n_{\mathcal{D}}$ ; minibatch size for co-update  $m$ ; learning rate  $\eta$

- 1: **Begin**
- 2: Initialize  $\phi$  by random variables.
- 3: **for**  $t = 1 : T$  **do**
- 4:  $L = \text{randi}(n_1, k), \dots, \text{randi}(n_C, k)$
- 5:  $\theta = \arg \min_{\theta} \sum_{c=1}^C L(C_\theta(F_\phi(x_c)), t) + \lambda \|\theta\|_2^2$
- 6:  $I_t = \text{randi}(n_{\mathcal{D}}, m)$
- 7:  $\phi \leftarrow \phi - \eta \sum_{c=1}^C \partial L(C_\theta(F_\phi(x_c)), t) / \partial \phi$
- 8: **end for**
- 9: **End**

**Output:** feature-extractor parameters  $\phi^* = \phi$

### 4. Proposition of point-like property

**Proposition 3.2.** Suppose that  $\phi^*$  simultaneously minimizes the classifier-anonymized, sample-wise losses  $\mathbb{E}_{\theta \sim \Theta_\phi} L_{\phi,\theta}(x, t)$  in a class-separable fashion for all  $(x, t) \in \mathcal{D}$ . Then, samples from the same class share the same features; i.e.,  $F_{\phi^*}(x) = F_{\phi^*}(x')$ ,  $\forall x, x' \in \mathcal{X}_c$ , but samples from different classes do not; i.e.,  $F_{\phi^*}(x) \neq F_{\phi^*}(x')$ ,  $\forall x \in \mathcal{X}_c, \forall x' \in \mathcal{X}_{c' \neq c}$ .

( $\mathcal{X}_c$  is the set of input var with class  $c$ .)

Assumption  $\begin{cases} L_{\phi,\theta}(x, t) = (C_\theta(F_\phi(x)) - t)^2 \\ C_\theta(F_\phi(x)) = \theta^T F_\phi(x) + \theta^0 \\ \Theta_\phi \text{ given by Eq.(3, 4)} \end{cases}$

(Proof is given in the main text.)

### 5. Exp. #1: Partial-dataset opt.

**Q:** If features form point-like distribution, the decision boundary should be robust against the dataset size. Is the classification performance robust?

**Result:** Yes, FOCA exhibits much smaller performance gap.

large-dataset classifier vs small-dataset classifier. (Each method uses the same, fixed feature extractor is used.)

### Exp. #2: Approx. geodesic dist.

**Q:** Let:  $\theta^{LD}$ : large-dataset classifier param.  $\theta^{SD}$ : small-dataset classifier param. Are they close?

**Procedure:**

- 1) Partitions straight line connecting  $\theta^{LD}$  and  $\theta^{SD}$  into  $P$  line segments of equal lengths.
- 2) Comp. approx. geodesic distance:  $d(\theta^{LD}, \theta^{SD}) = \left( \sum_{i=1}^{P-1} d(\theta^{LD}, \theta^{i+1})^2 \right)^{1/2} \quad (15)$

where  $d(\theta^{LD}, \theta^{i+1}) = \left( (\theta^{LD} - \theta^{i+1})^T \Sigma^{-1} (\theta^{LD} - \theta^{i+1}) \right)^{1/2}$   $(16)$

$\Sigma = \mathbb{E}_{x, t \in \mathcal{D}} \left( \frac{\partial L_{\phi,\theta}(x, t)}{\partial \theta} \frac{\partial L_{\phi,\theta}(x, t)}{\partial \theta} \right)^T$

**Result:** Yes, FOCA exhibits orders-of-magnitude smaller approx. geodesic distance (dashed lines).

### Exp. #3: Low-dim. analyses

**Q:** Qualitatively, is the low-dimensional structure point-like?

**Result:** Yes, FOCA (left) looks more point-like than BatchNorm (right) after LDA on normalized features.

Further, FOCA has the highest 2-class linear separability along principle axis:

Method	Eigenvalue	Test error rate
FOCA (ours)	247.28	2.01%
Plain	5.74	2.71%
Noisy	7.49	2.96%
Dropout	5.81	2.78%
Batch Norm	7.28	2.43%

*What?* **Breaking co-adaptation** between feature extractor and classifier.

*How?* By **classifier anonymization**.

*Theory?* Proved: Features form simple **point-like distribution**.

*Reality?* Point-like property largely confirmed on real datasets.