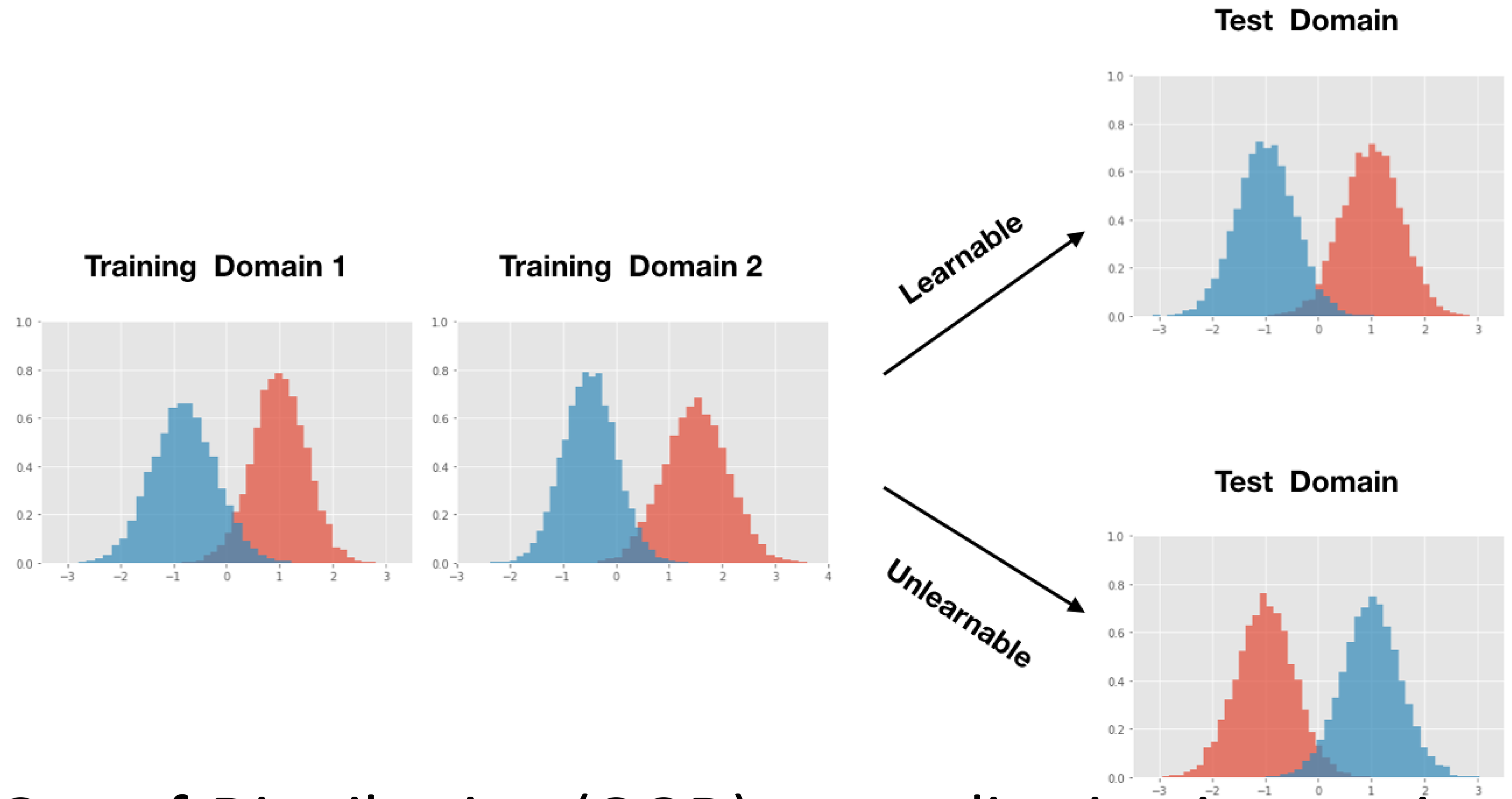
Towards a Theoretical Framework of Out-of-Distribution Generalization

Haotian Ye[1], Chuanlong Xie[2], Tianle Cai[1], Ruichen Li[1], Zhenguo Li[2], Liwei Wang[1]

[1]Peking University; [2]Huawei Noah's Ark Lab

Takeaways



- Out-of-Distribution (OOD) generalization is to train on sets of domains, and generalize to unseen domains.
- As the above figure shows, generalization to any domains is impossible. We lack theories of **when we can guarantee generalization** *(OOD – learnable)*.
- Intuition: the *extent* to which the *invariance* of *informative* feature is preserved, determines OOD learnability. **We build a framework to describe OOD learnability, bound OOD generalization error based on it, and propose an effective model selection algorithm.**

Our Framework (Informal)

- $\mathcal{E}_{avail} \subset \mathcal{E}_{all}$ are two sets of domains. We train on $\mathcal{E}_{avail}$, and (hope to) generalize to $\mathcal{E}_{all}$.
- Hypothesis Space: $f = \boldsymbol{g} \circ \boldsymbol{h}$, where $\boldsymbol{h}$ is d-dimensional feature extractors, and $\boldsymbol{g}$ is top classifier.
- For 1-dim feature ϕ, define $P_y^e = \mathrm{Prob}(\phi(X^e|Y = y))$.
- Goal: minimizes maximum loss of f in all domains, i.e.,

$$\mathcal{L}(f, \mathcal{E}) \triangleq \max_{e \in \mathcal{E}} \ell(f, e)\,.$$

❖ ***Variation*** of ϕ across domain set $\mathcal{E}$:

$$V(\phi, \mathcal{E}) = \max_{y \in Y} \max_{e, e' \in \mathcal{E}} \mathrm{dist}\left(P_y^e, P_y^{e'}\right)$$

❖ ***Informativeness*** of ϕ across domain set $\mathcal{E}$:

$$I(\phi, \mathcal{E}) = \operatorname*{avg}_{y \neq y'} \min_{e \in \mathcal{E}} \mathrm{dist}\left(P_y^e, P_{y'}^e\right)$$

❖ ***OOD Learnability***: $(\mathcal{E}_{avail}, \mathcal{E}_{all})$ is $(s(\cdot), \delta)$-learnable, if for all $\phi \in \Phi$ satisfying $I(\phi, \mathcal{E}_{avail}) \geq \delta$, we have

$$s\big(V(\phi, \mathcal{E}_{avail})\big) \geq V(\phi, \mathcal{E}_{all}).$$

Here $s(\cdot)$ is a special type of monotonically increasing function, which we call ***expansion function***.

Theorems and Model Selections

- We prove that OOD generalization error can be bounded with our framework.
- **Main Theorem**: Under some conditions, if the problem is $(s(\cdot), I(\boldsymbol{h}, \mathcal{E}_{\text{avail}}))$-learnable, then

$$\mathcal{L}(f, \mathcal{E}_{all}) - \mathcal{L}(f, \mathcal{E}_{avail}) \leq O\left(s(V(\boldsymbol{h}, \mathcal{E}_{avail}))^{\frac{\alpha^2}{(\alpha+d)^2}}\right)$$

- **Lower Bound**: Exists a $(s(\cdot), \delta)$-learnable problem, where the optimal classifier f with $V(\boldsymbol{h}, \mathcal{E}_{avail}) = \varepsilon$ has

$$\mathcal{L}(f, \mathcal{E}_{all}) - \mathcal{L}(f, \mathcal{E}_{avail}) \geq \Omega\Big(s\big(V(\boldsymbol{h}, \mathcal{E}_{avail})\big)\Big)$$

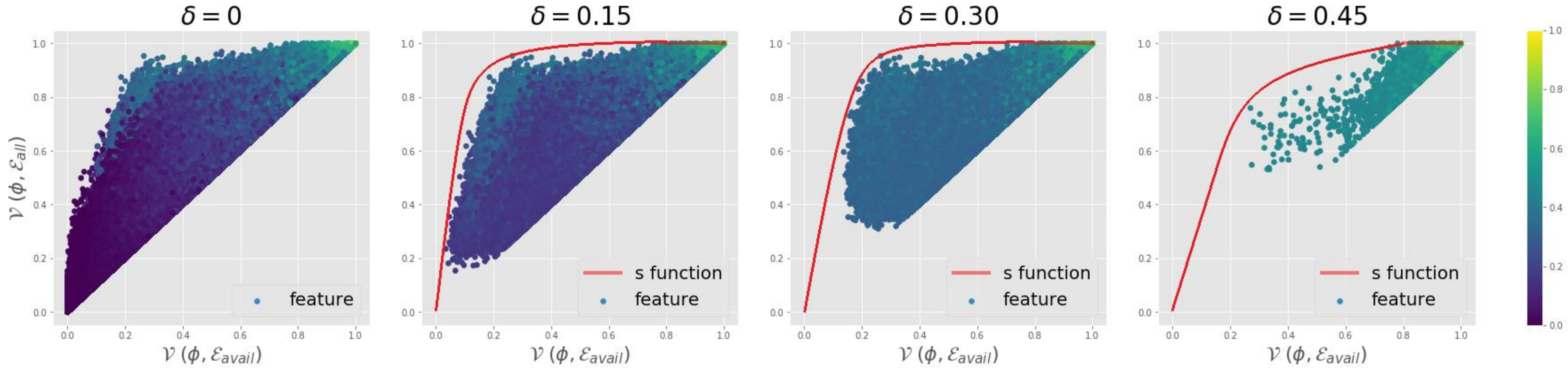
- We also propose a **model selection algorithm**, and it outperforms other selection methods:

Algorithm 1: Model Selection

```
Input: available dataset X_avail = (X_train, X_val), candidate models set M, var_acc_rate r_0.
for f = g ∘ h in M do
    for i in [d] do
        V̂_i ← max_{y∈Y, X^e≠X^e'∈X_avail} Total Variation(P(φ_i^e|y), P(φ_i^e'|y));   ▷Use GPU KDE
    end
    V_f ← mean_{i∈[d]} V̂_i
    Acc_f ← compute validation accuracy of f using X_val
end
Return argmax_{f∈M}(Acc_f − r_0 V_f)
```

Features' V and I (lightness) in Office-Home [1].

Larger δ corresponds to flatter $s(\cdot)$.



	Env	A	C	P	S	avg	acc inc
PACS	Val	85.20%	80.42%	96.17%	77.86%	84.91%	-
	Ours	**88.72%**	**81.74%**	**96.83%**	**79.00%**	**86.57%**	1.66%↑
	Env	A	C	P	R	avg	acc inc
OfficeHome	Val	61.85%	**55.56%**	74.72%	76.25%	67.09%	-
	Ours	**65.76%**	55.07%	**75.20%**	**76.31%**	**68.09%**	1.00%↑
	Env	C	L	S	V	avg	acc inc
VLCS	Val	97.46%	64.83%	**69.50%**[6]	**70.97%**	75.69%	-
	Ours	**97.81%**	**66.98%**	69.50%	**70.97%**	**76.32%**	0.63%↑

[1] H Venkateswara, et al. Deep hashing network for unsupervised domain adaptation.

NOAH'S ARK LAB