NEURAL INFORMATION PROCESSING SYSTEMS

SCIENCES SORBONNE UNIVERSITÉ

Diverse Weight Averaging for Out-of-Distribution Generalization

Alexandre Ramé (Sorbonne & Meta AI)
Matthieu Kirchmeyer (Sorbonne & Criteo)
Thibaud Rahier (Criteo)
Alain Rakotomamonjy (LITIS & Criteo)
Patrick Gallinari (Sorbonne & Criteo)
Matthieu Cord (Sorbonne & Valeo.ai)

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

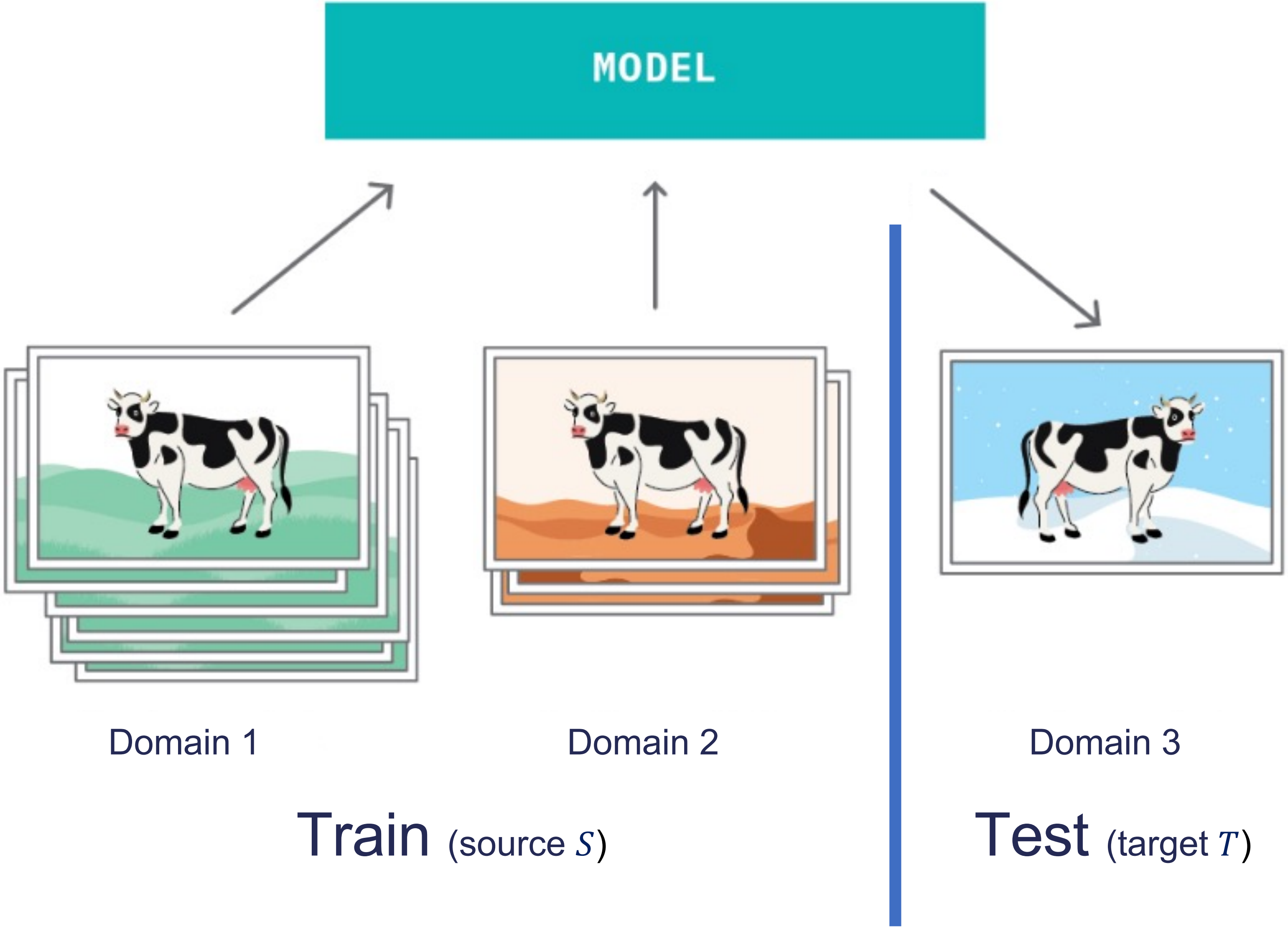
MLIA Machine Learning & Deep Learning for Information Access

Meta

CRITEO

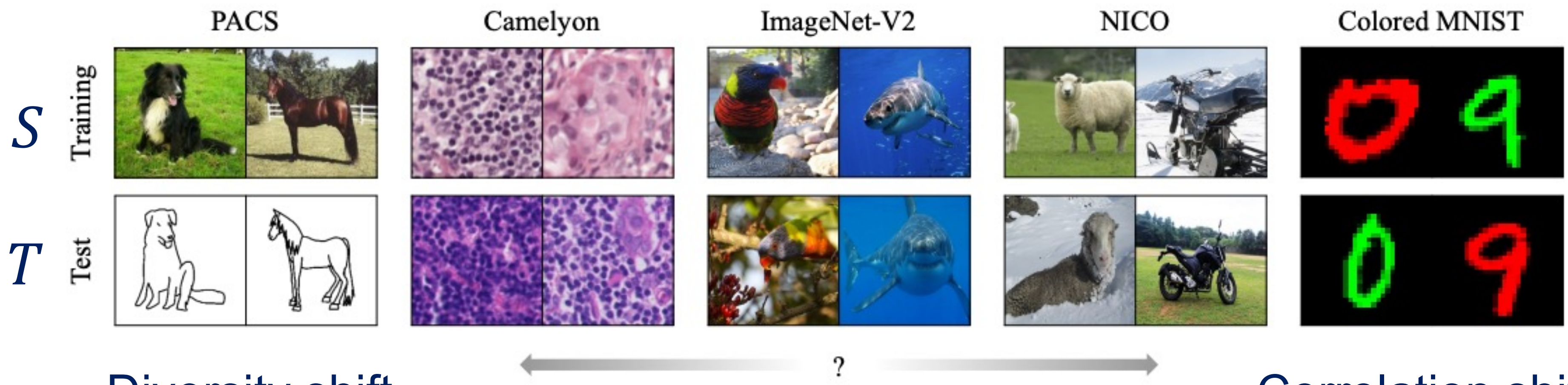
valeo.ai

Goal: generalization to unseen domains



[Arjovsky19] Invariant risk minimization.

Two kind of source/target distribution shifts

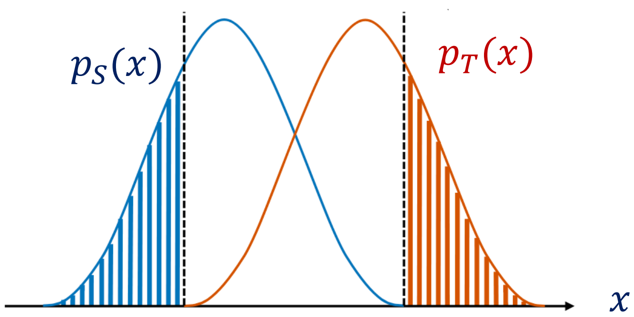


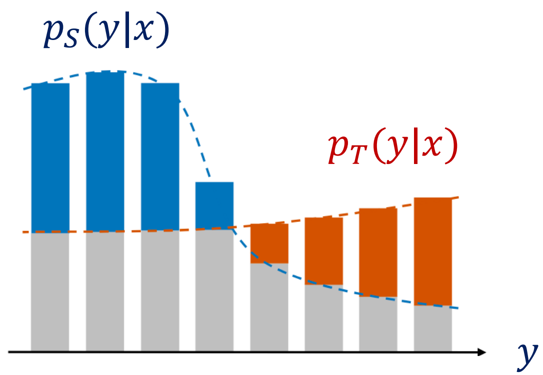
Diversity shift

$p_S(X) \neq p_T(X)$

Correlation shift

$p_S(Y|X) \neq p_T(Y|X)$





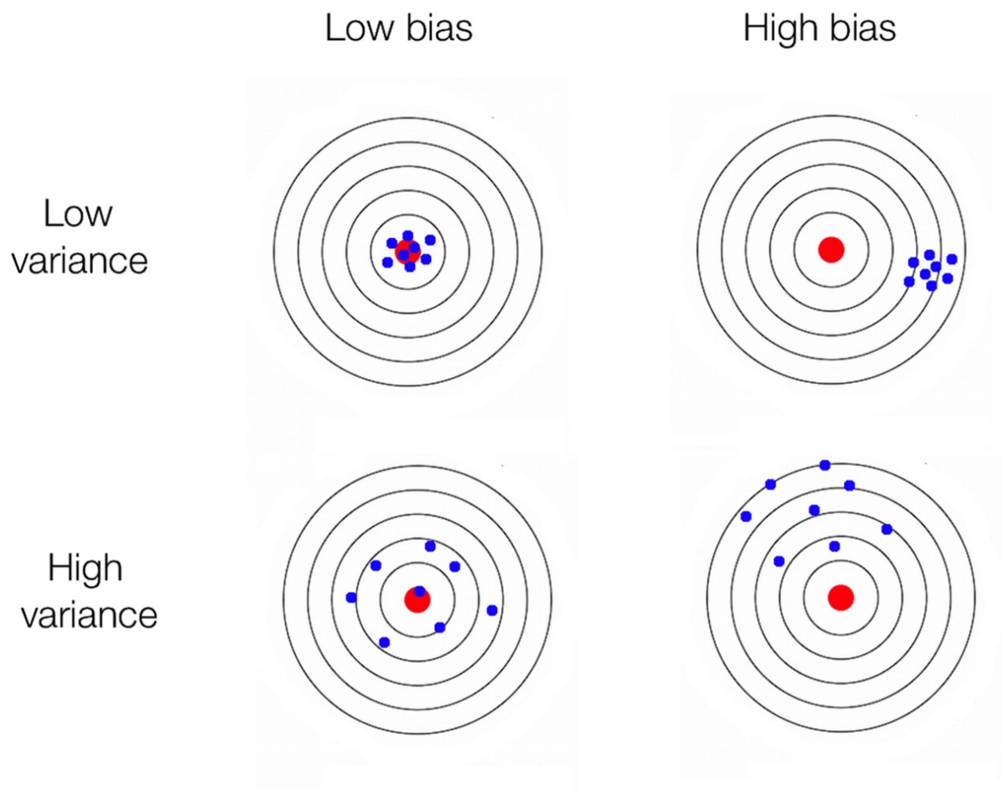
[Ye2022] OoD-Bench: Quantifying and Understanding Two Dimensions of OOD Generalization. CVPR

A bias-variance analysis in OOD

$$\mathbb{E}_\theta err_T(\theta) = bias^2 + var$$

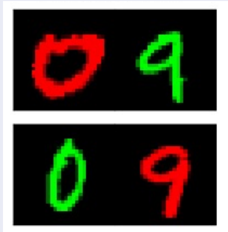
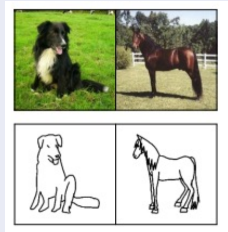
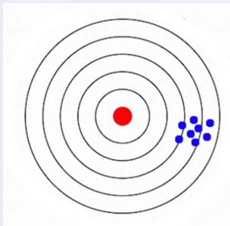
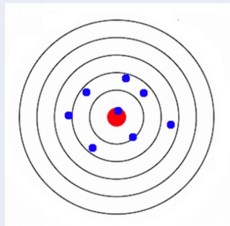
where, with $\bar{f}(x) = \mathbb{E}_\theta\, f_\theta(x)$:

- $bias(x, y) = y - \bar{f}(x)$,
- $var(x) = \mathbb{E}_\theta \left[\left(f_\theta(x) - \bar{f}(x)\right)^2\right]$.

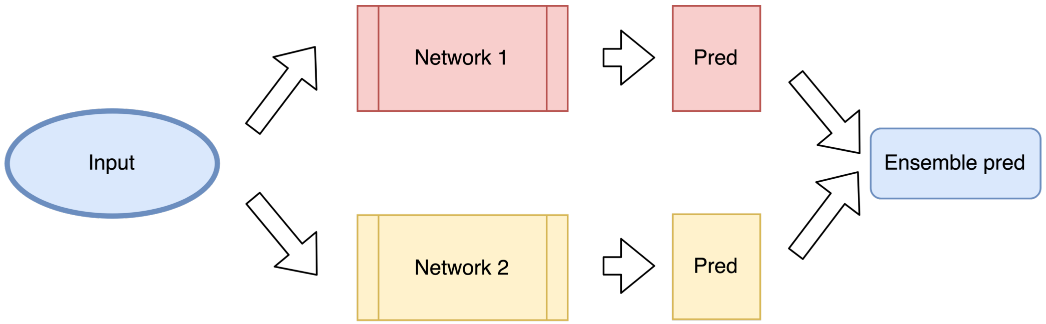


Question: how do $bias$ and var change with distribution shifts ?

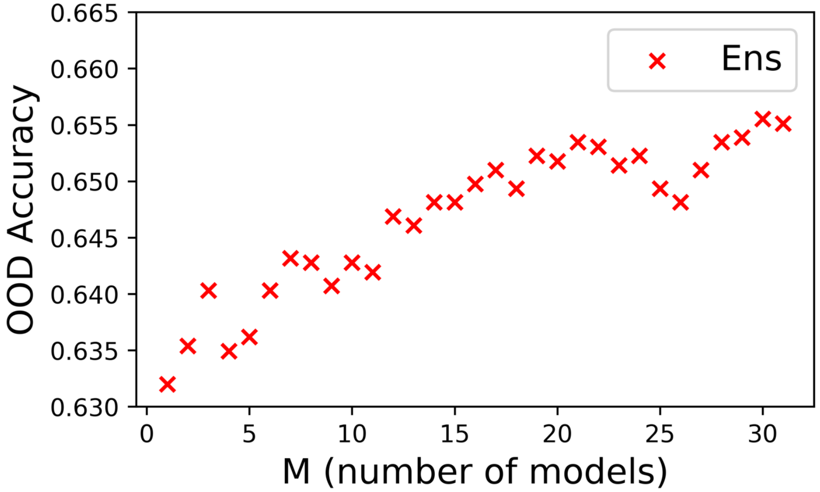
[Kohavi1996] Bias plus variance decomposition for zero-one loss functions. ICML

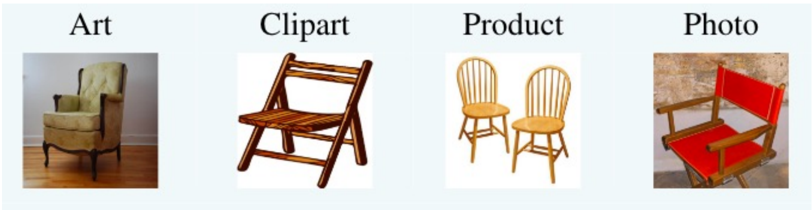
	Correlation shift	**Diversity shift**
Probabilistic perspective	$p_S(Y\|X) \neq p_T(Y\|X)$	$p_S(X) \neq p_T(X)$
Example		
Datasets	ColoredMNIST, CelebA…	OfficeHome, PACS …
Bias-variance	**Large bias** Small variance 	Small bias **Large variance** 
Approaches	• Invariance: IRM, Coral • Robust optimization: gDRO	• Variance reduction: ensembling, **DiWA**

Ensembling M models tackles variance



$$\mathbb{E}_{ens_M} err_T(ens_M) \approx bias^2 + \frac{1}{M} var$$





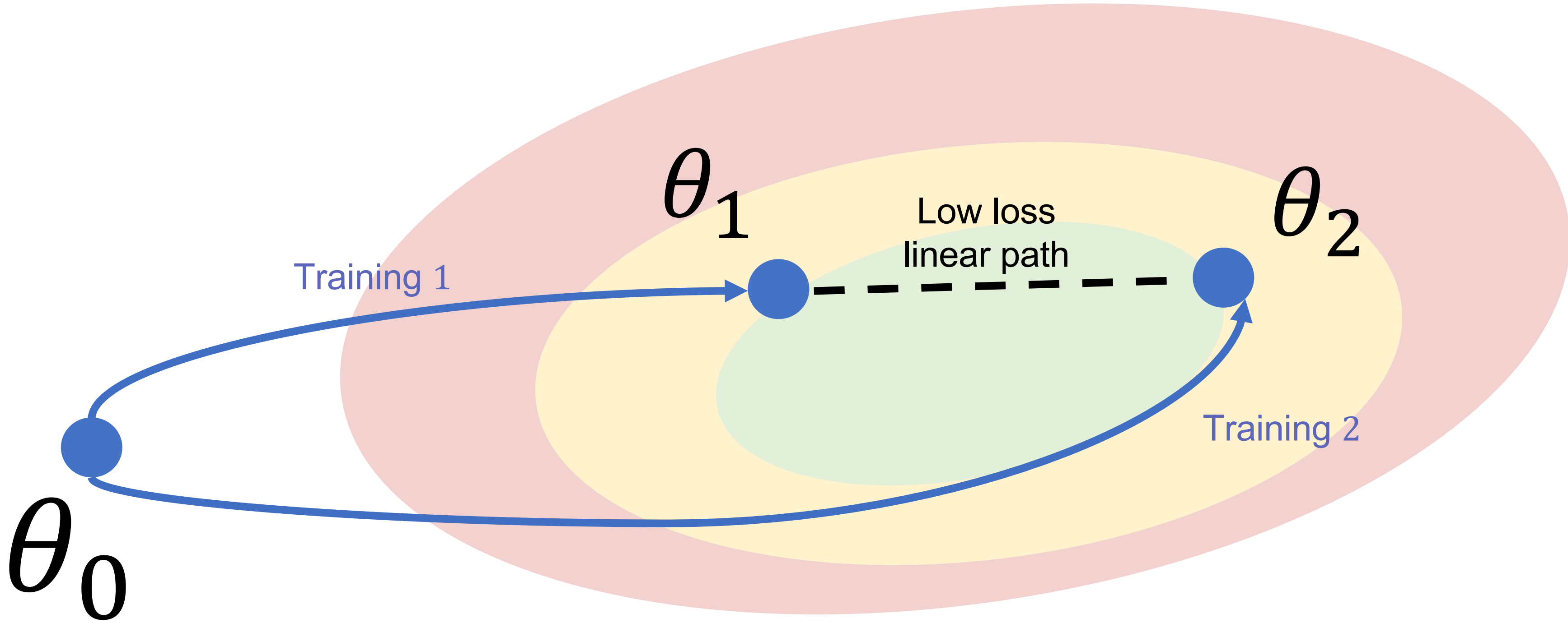
Setup: OfficeHome under diversity shift

- train on *Clipart, Product, Photo*
- test on OOD *Art*

[Ueda1996] Generalization error of ensemble estimators. ICNN

Yet traditional prediction ensembling is costly …

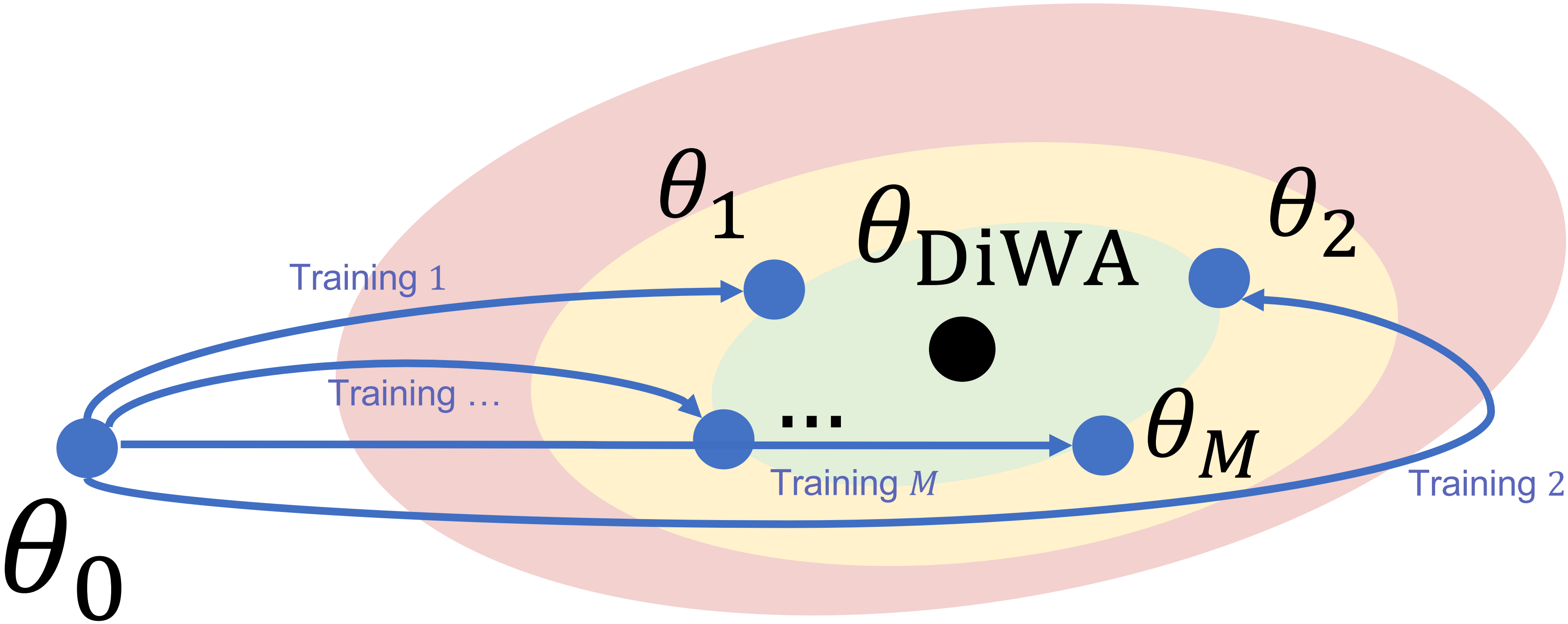
An empirical insight: linear mode connectivity



Low-loss linear path
when fine-tunings start from a shared pretrained initialization θ_0
(despite the architecture's non-linearities).

[Frankle2020] Linear Mode Connectivity and the Lottery Ticket Hypothesis. ICML
[Neyshabur2020] What is being transferred in transfer learning? NeurIPS

Diverse Weight Averaging (DiWA)



$$\theta_{\mathrm{DiWA}} = \frac{1}{M}\sum_{m=1}^{M}\theta_m$$

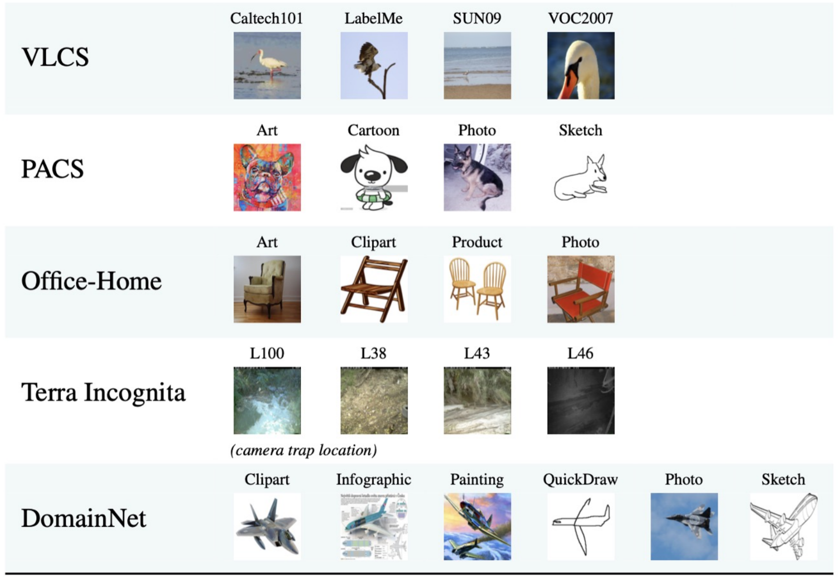
obtained from a shared pretrained initialization θ_0. Then:

$$f_{\theta_{\mathrm{DiWA}}} = f_{\frac{1}{M}\sum_{m=1}^{M}\theta_m} \approx \frac{1}{M}\sum_{m=1}^{M} f_{\theta_m}.$$

[Izmailov2018] Averaging Weights Leads to Wider Optima and Better Generalization. UAI
[Wortsman2022] Model soups: averaging weights of multiple fine-tuned models improves accuracy. ICML

SoTA on DomainBed [Gulrajani2021]

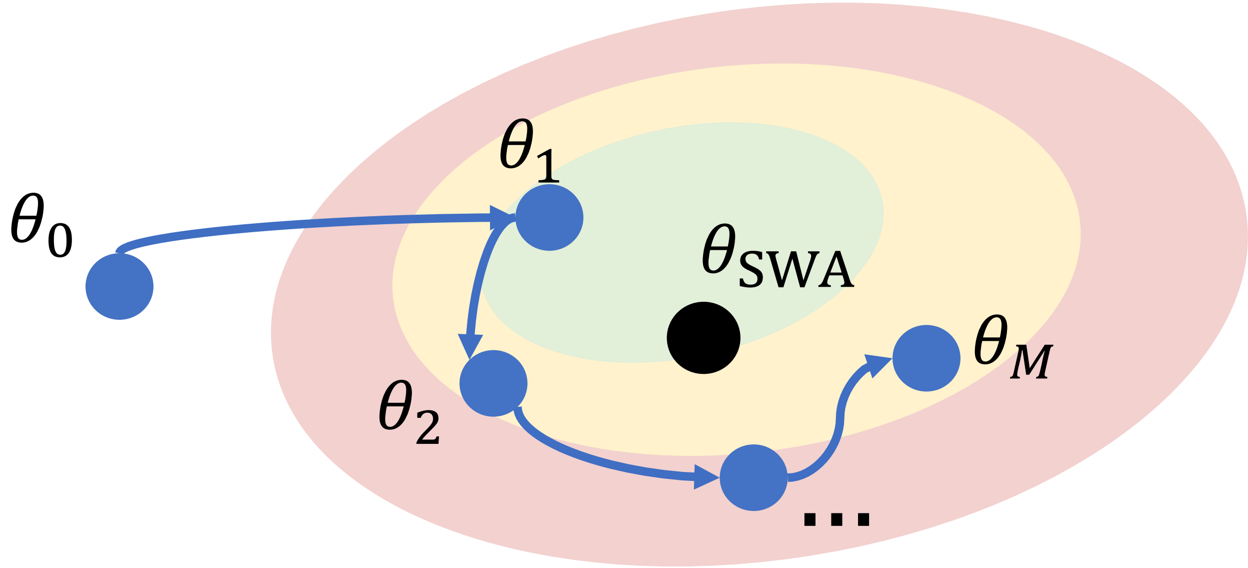
Reference benchmark for OOD generalization in computer vision, imposing the *code*, *datasets*, *training procedures*, *hyperparameter search* etc.

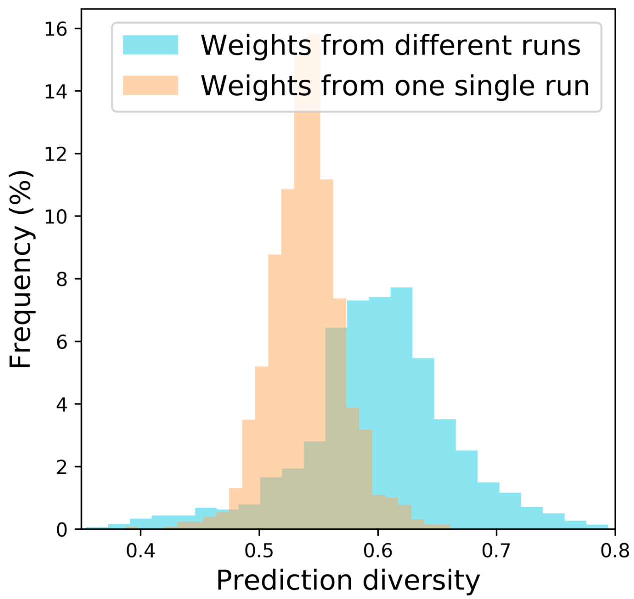


Algo	Cost	PACS	VLCS	OH	TI	DN	Avg
ERM	1	85.5	77.5	66.5	46.1	40.9	63.3
CORAL	1	86.2	78.8	68.7	47.6	41.5	64.6
SWAD	1	88.1	**79.1**	70.6	50.0	46.5	66.9
ENS	20	88.1	78.5	71.7	50.8	47.0	67.2
DiWA	1	**89.0**	78.6	**72.8**	**51.9**	**47.7**	**68.0**

[Gulrajani2021] In search of lost domain generalization. ICLR
[Cha2021] SWAD: Domain Generalization by Seeking Flat Minima. NeurIPS

Previous SoTA: single-run WA





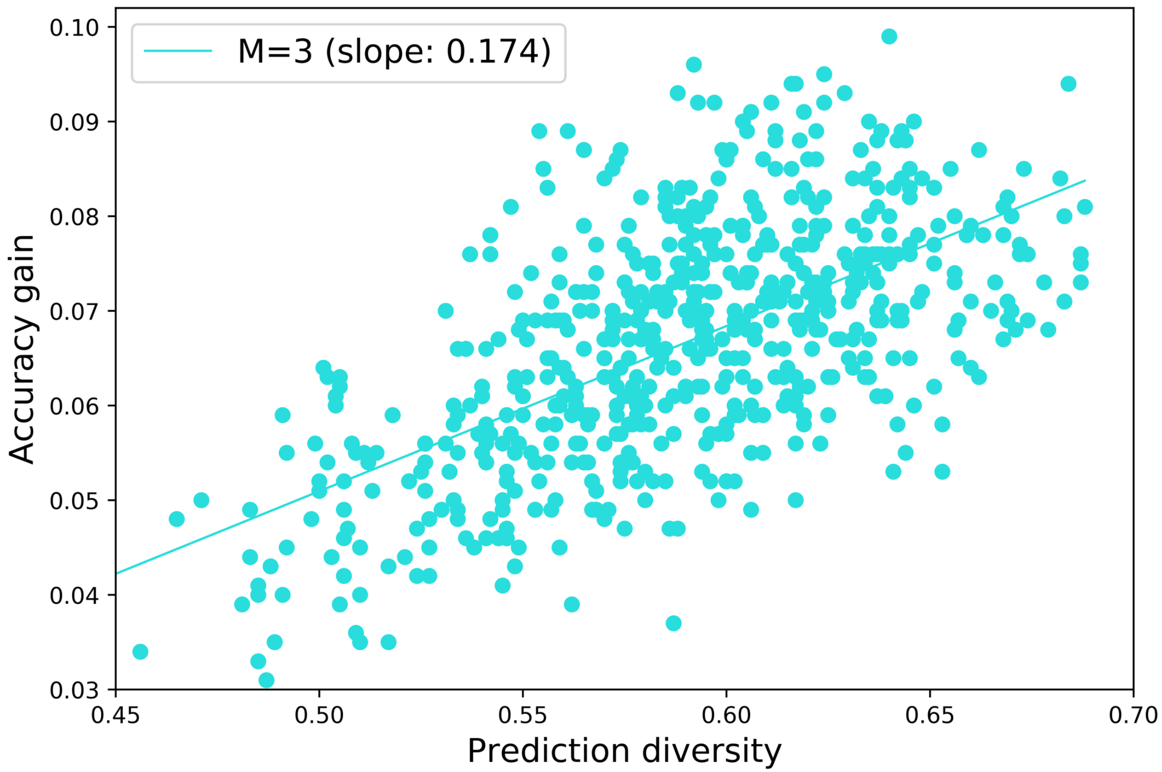
Weights from different runs are more diverse (left) thus their average is better (next slide).

[Izmailov2018] Averaging Weights Leads to Wider Optima and Better Generalization. UAI
[Cha2021] SWAD: Domain Generalization by Seeking Flat Minima. NeurIPS

Covariance as diversity

$$\mathbb{E}_{\theta_{WA}} err_T(\theta_{WA}) \approx bias^2 + \frac{1}{M} var + \frac{M-1}{M} cov,$$

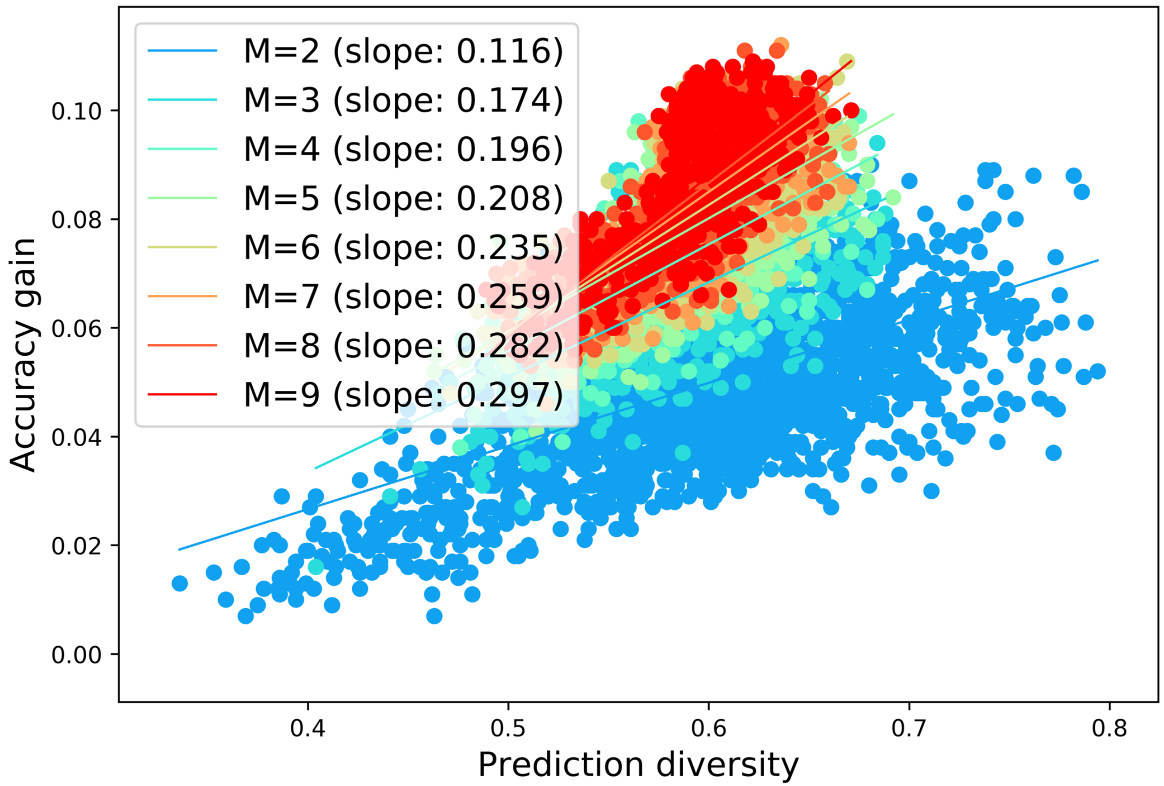
where cov is smaller when models are uncorrelated, *i.e.*, functionally diverse.



Legend: Each dot is the accuracy gain of averaging $M = 3$ models over the average accuracy *wrt* their diversity (normalized count of different errors).

Diversity even more important for larger M

$$\mathbb{E}_{\theta_{WA}} err_T(\theta_{WA}) = bias^2 + \frac{1}{M} var + \frac{M-1}{M} cov + \mathcal{O}(\bar{\Delta}^2).$$



Legend: Each dot is the accuracy gain of averaging M models over the average accuracy *wrt* their diversity (normalized count of different errors).

Conclusion

- Bias-variance analysis in OOD
  - ✓ Relate diversity shift to variance
  - ✓ Relate correlation shift to bias
- New weight averaging strategy
  - ✓ Average all weights obtained from the hyperparameter search
  - ✓ SoTA on DomainBed to tackle diversity shift

arXiv: https://arxiv.org/abs/2205.09739

code: https://github.com/alexrame/diwa