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)

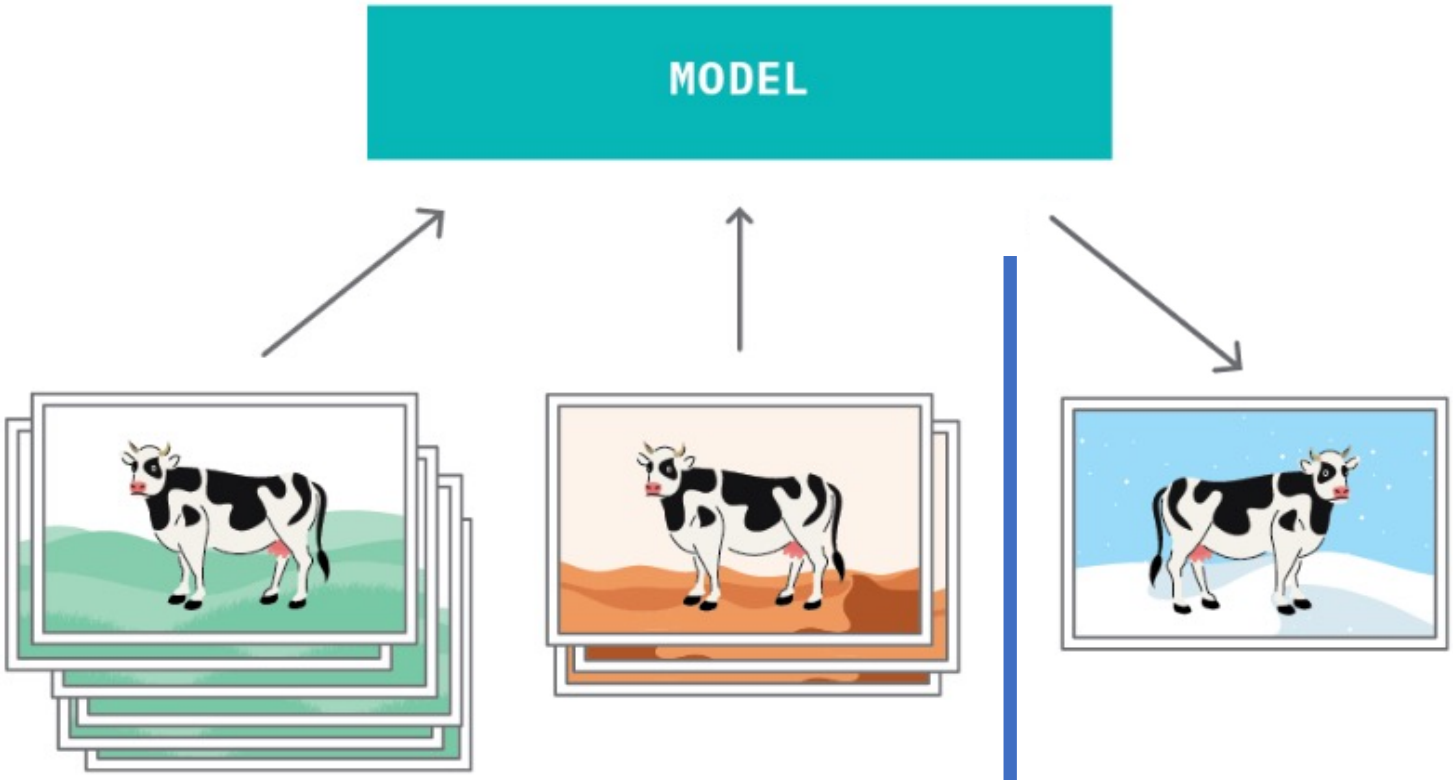
* Equal contribution

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access

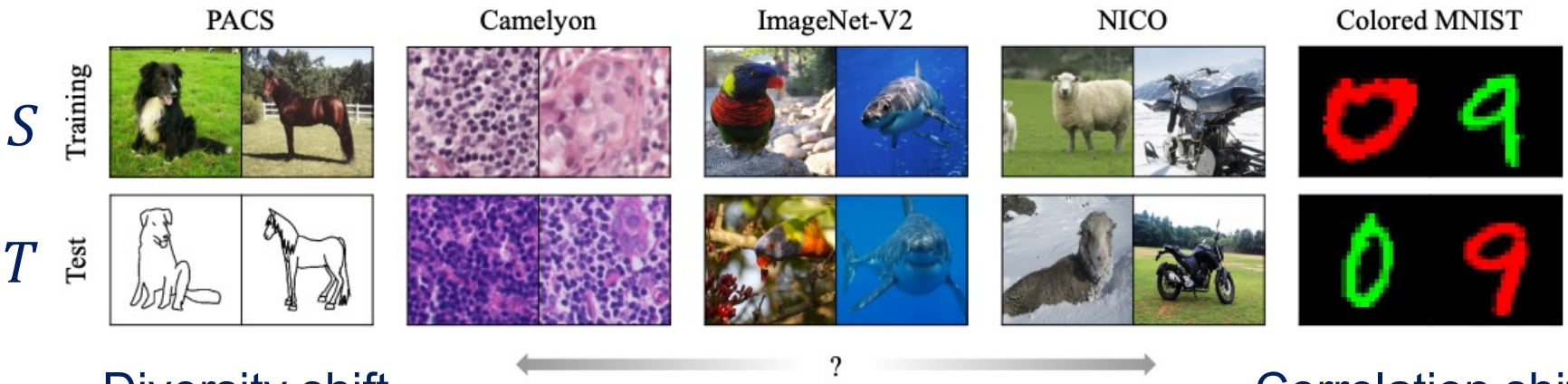
Meta CRITEO valeo.ai

Goal: generalization to unseen domains



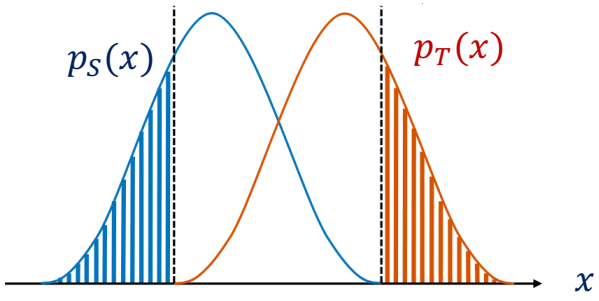
[1] Invariant risk minimization. Arjovsky *et al.*, 2019

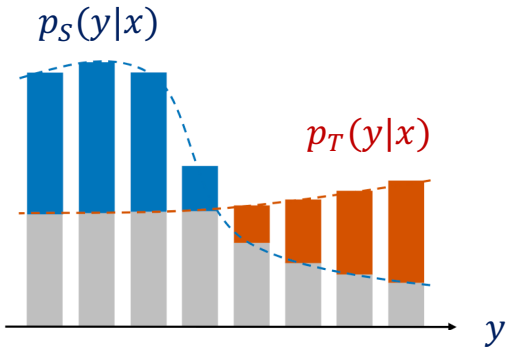
Two kind of source/target distribution shifts



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

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





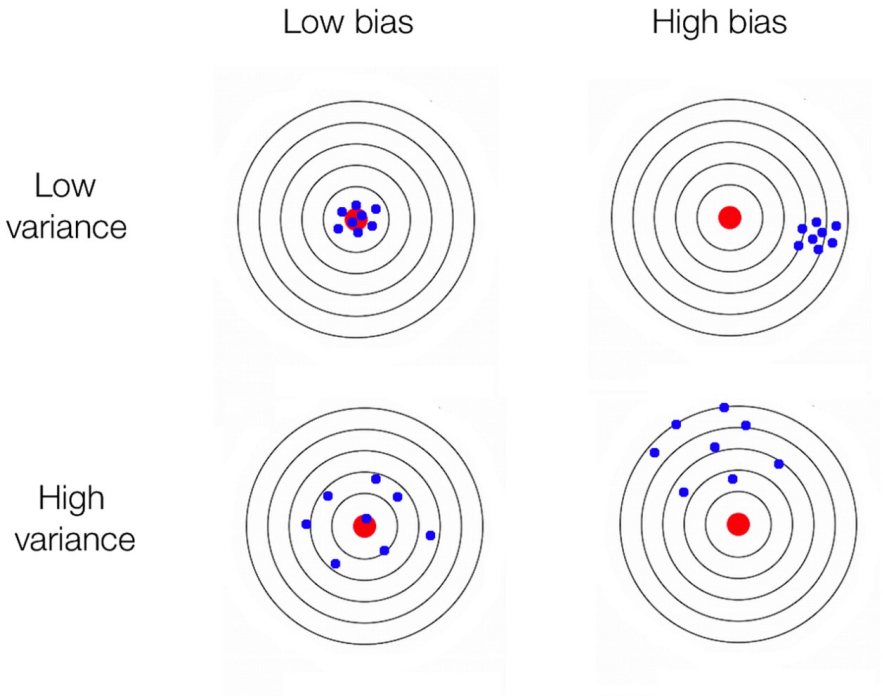
[1] OoD-Bench: Quantifying and Understanding Two Dimensions of OOD Generalization. Ye *et al.*, CVPR 2022

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 correlation and diversity shifts ?

[1] Bias plus variance decomposition for zero-one loss functions. Kohavi and Wolpert. ICML 1996

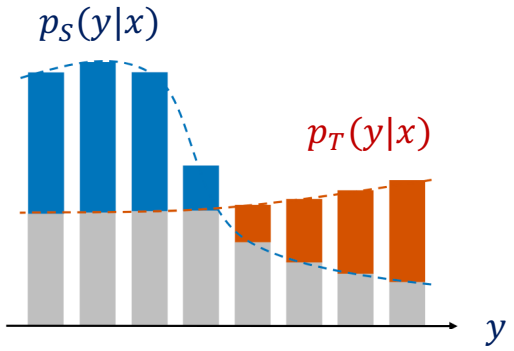
Bias as correlation shift

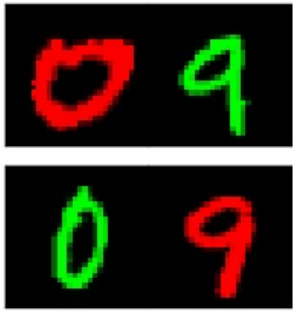
Definition: $bias^2 = \int_T \left(\mathbb{E}_T[Y|X=x] - \mathbb{E}_\theta f_\theta(x)\right)^2 p_T(x)dx.$

Intuition: bias in OOD increases when the posteriors vary across source and target.

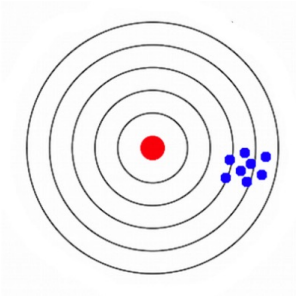
Proposition: for large networks,

$$bias^2 \approx \int_T (\mathbb{E}_T[Y|X=x] - \mathbb{E}_S[Y|X=x])^2 p_T(x)dx$$

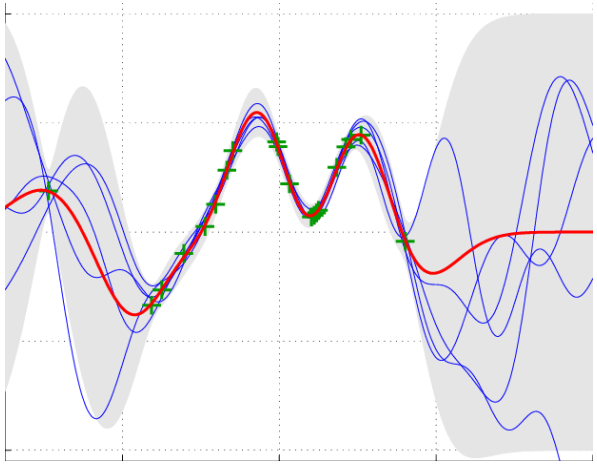




$\Rightarrow$



Variance as diversity shift

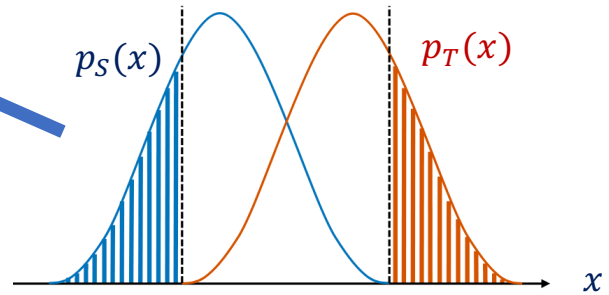


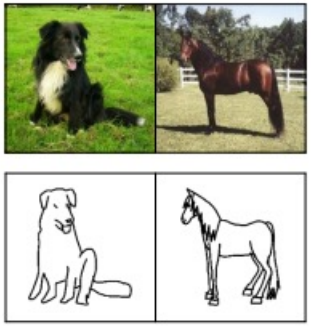
Definition: $var_{d_T} = \sum_{x \in d_T} \mathbb{E}_\theta \left[\left(f_\theta(x) - \bar{f}(x) \right)^2 \right]$.

Intuition: variance increases away from the source training samples.

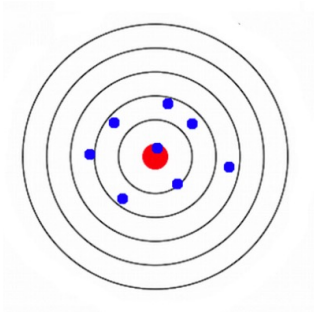
Proposition: for networks with diagonally dominant NTK [1]:

$$var_{d_T} \propto MMD^2_{NTK^2}(X_{d_S}, X_{d_T}) + \ldots$$

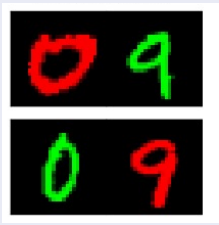
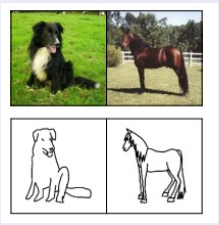
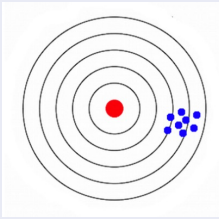
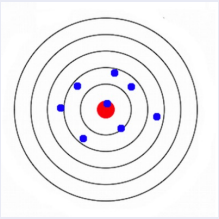




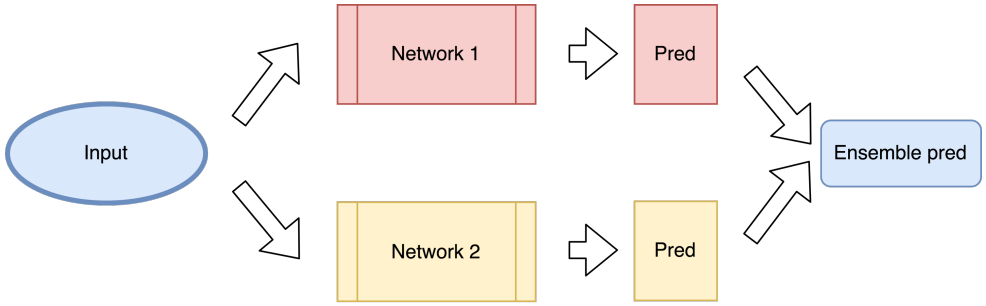
$\Rightarrow$



[1] Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. Seleznova and Kutyniok, ICML 2022

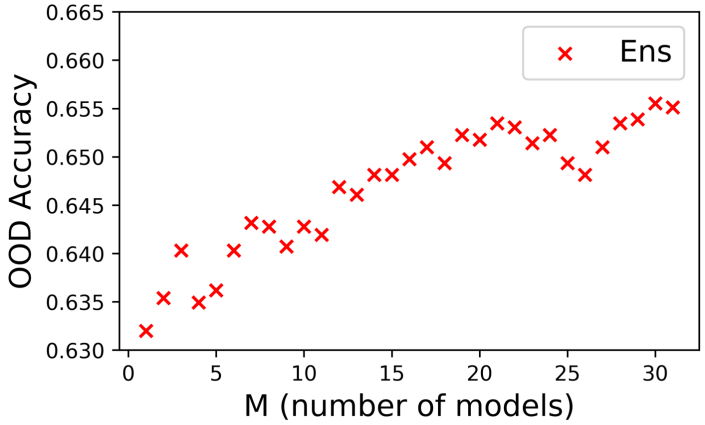
	Correlation shift	Diversity shift
Probabilistic perspective	$p_S(Y\|X) \neq p_T(Y\|X)$	$p_S(X) \neq p_T(X)$
Example		
Datasets	ColoredMNIST, CelebA…	OfficeHome, PACS …
Bias-variance	Large **bias** Small variance 	Small bias Large **variance** 
Approaches	Invariance: IRM, Coral Robust optimization: gDRO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

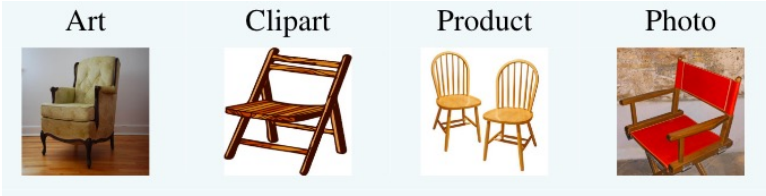
Bias-variance-covariance in ensembling (ENS)



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

where the covariance across models verify: $\text{cov} \leq var$.



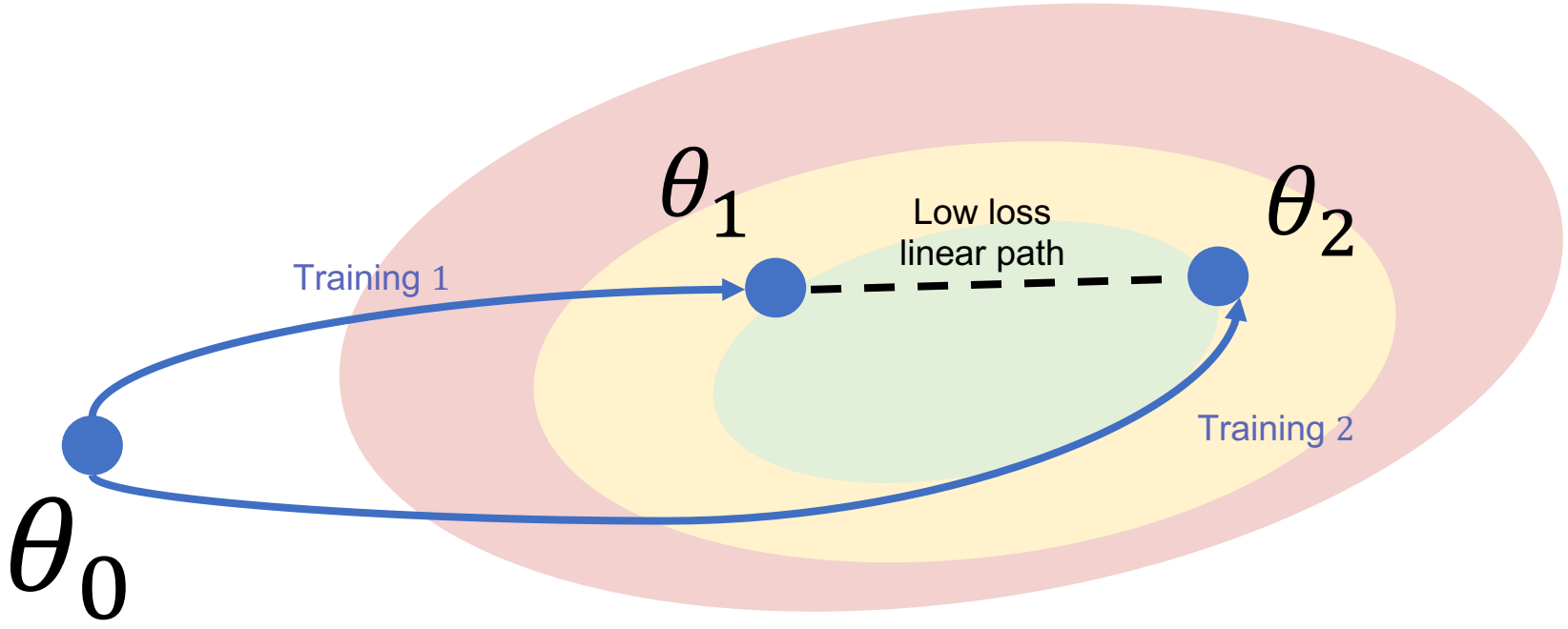


Setup: OfficeHome under diversity shift
- train on “Clipart,Product,Photo”,
- test on “Art” OOD.

[1] Generalization error of ensemble estimators. Ueda *et al.*, ICNN 1996

Yet ensembling is costly …

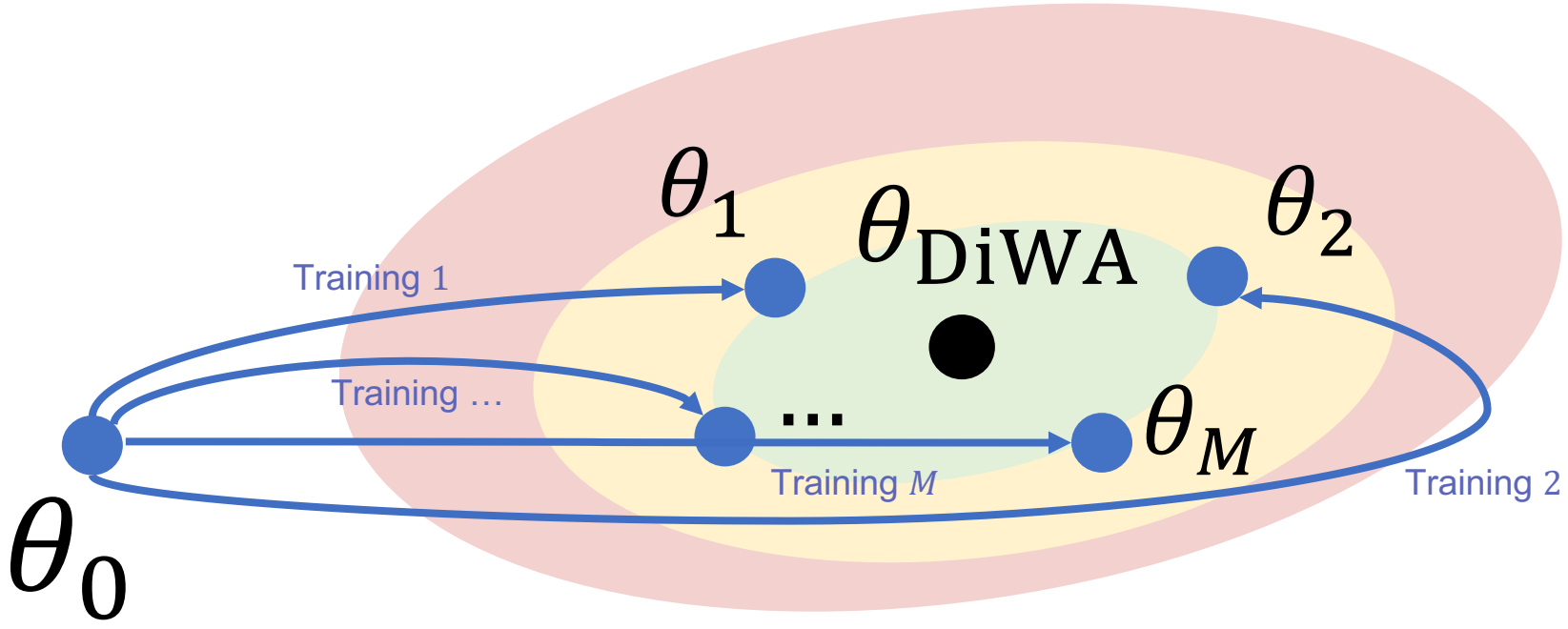
An empirical insight: linear mode connectivity



Possible when finetunings start from a shared pretrained initialization θ_0.

[1] Linear Mode Connectivity and the Lottery Ticket Hypothesis. Frankle *et al.,* ICML 2020
[2] What is being transferred in transfer learning? Neyshabur *et al.,* NeurIPS 2020

Diverse Weight Averaging (DiWA)



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

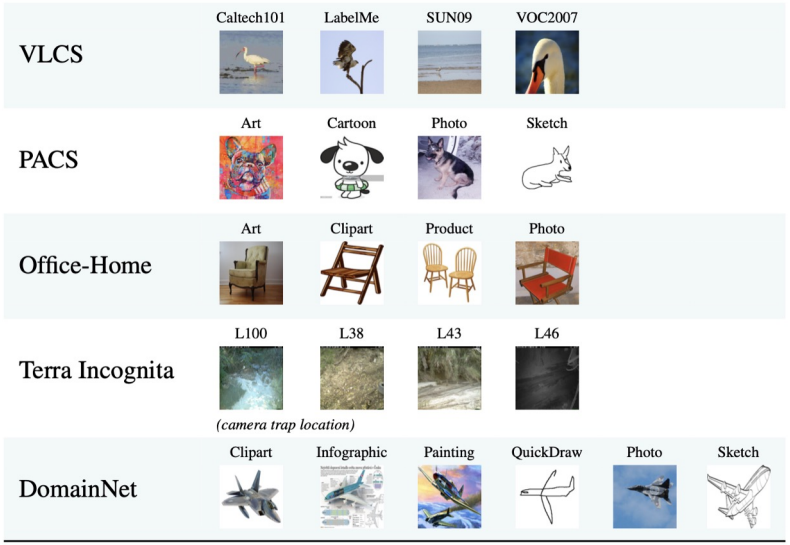
obtained from a shared pretrained initialization θ_0. By Taylor expansion around θ_{DiWA}:

$$f_{WA} \approx f_{ENS} + \mathcal{O}(max_{m=1}^{M}\|\theta_m - \theta_{WA}\|^2).$$

[1] Averaging Weights Leads to Wider Optima and Better Generalization. Izmailov *et al.*, UAI 2018
[2] Model soups: averaging weights of multiple fine-tuned models improves accuracy. Wortsman *et al.*, ICML 2022

SoTA on DomainBed [1]

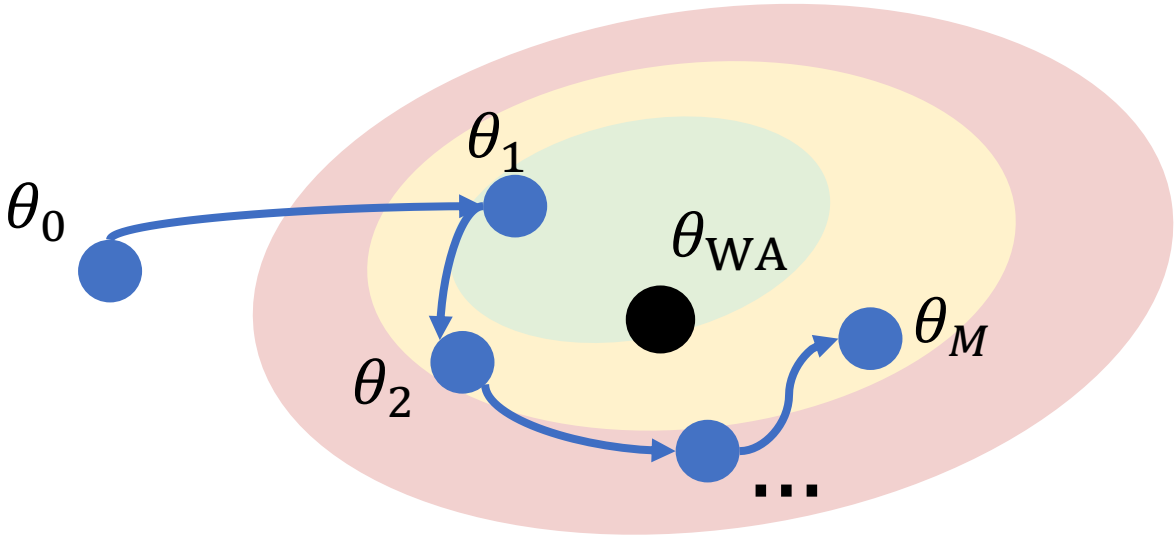
Reference benchmark for OOD generalization in computer vision, imposing the *code*, *datasets*, *training procedures*, *hyperparameter search* etc.

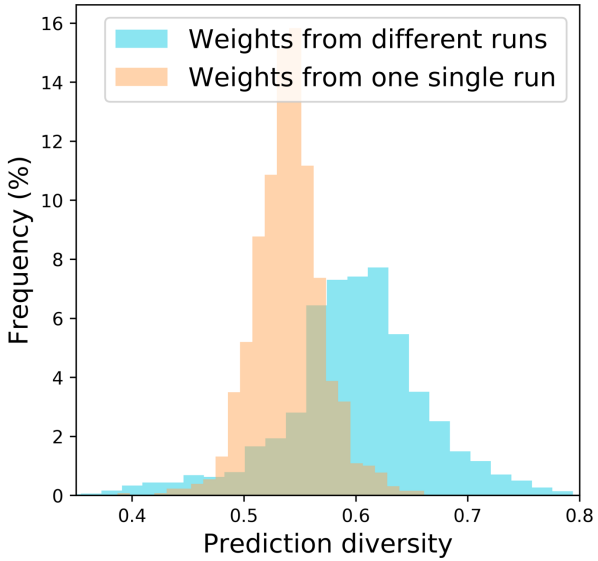


Algo	Cost	PACS	VLCS	OH	TI	DN	Avg
ERM	1	85.5	77.5	66.5	46.1	40.9	63.3
CORAL	1	86.2	78.8	68.7	47.6	41.5	64.6
SWAD [2]	1	88.1	**79.1**	70.6	50.0	46.5	66.9
ENS	20	88.1	78.5	71.7	50.8	47.0	67.2
DiWA	1	**89.0**	78.6	**72.8**	**51.9**	**47.7**	**68.0**

[1] In search of lost domain generalization. Gulrajani and Lopez-Paz, ICLR 2021
[2] SWAD: Domain Generalization by Seeking Flat Minima. Cha *et al.*, NeurIPS 2021

Previous SoTA: single-run WA





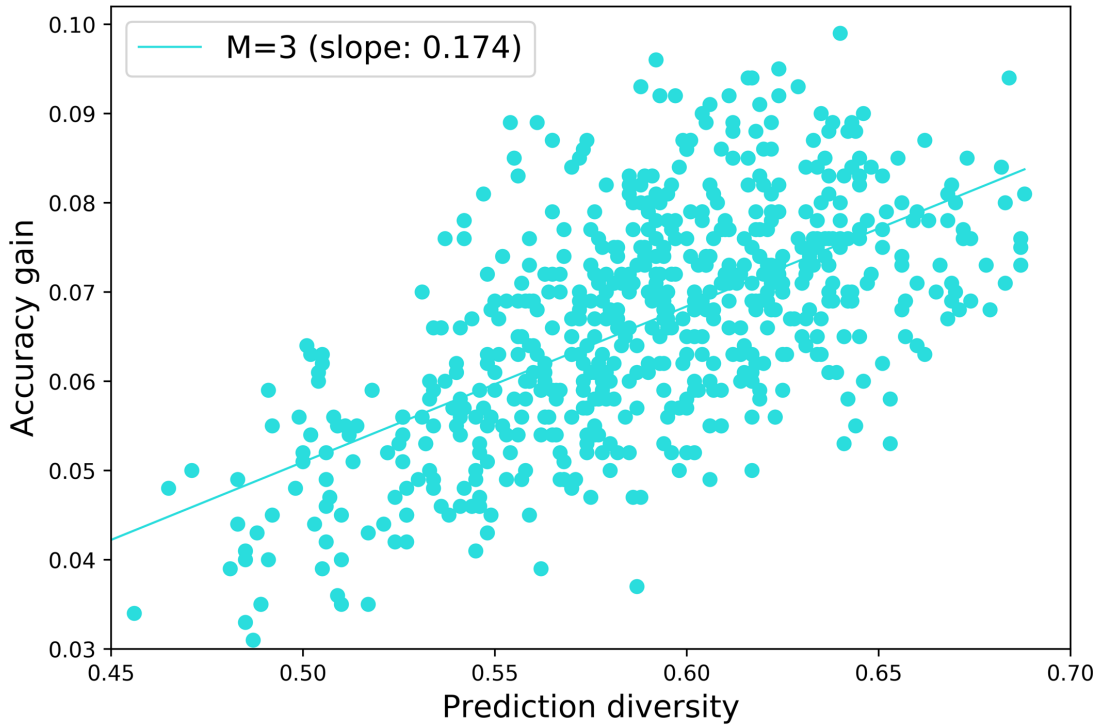
Weights from different runs are more diverse (left) thus their average is better (next slide).

[1] Averaging Weights Leads to Wider Optima and Better Generalization. Izmailov *et al.*, UAI 2018
[2] SWAD: Domain Generalization by Seeking Flat Minima. Cha *et al.*, NeurIPS 2021

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).

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