Diverse Weight Averaging for Out-Of-Distribution Generalization

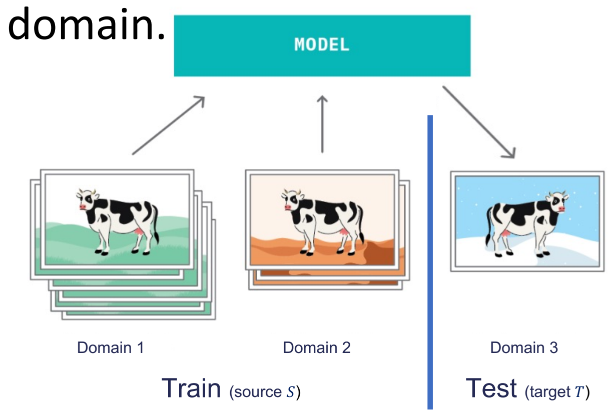
Alexandre Ramé* (Sorbonne U)
Matthieu Kirchmeyer* (Sorbonne U & Criteo)
Thibaud Rahier (Criteo)
* equal contribution
Alain Rakotomamonjy (LITIS & Criteo)
Patrick Gallinari (Sorbonne U & Criteo)
Matthieu Cord (Sorbonne U & Valeo.ai)

1 Setup and Challenge: OOD Generalization

Train on S source domain and test on T target domain.

Under domain shifts divided in [Ye2022] into:

- Diversity shift: $p_S(X) \neq p_T(X)$.
- Correlation shift: $p_S(Y|X) \neq p_T(Y|X)$.



2 Bias-Variance Analysis in OOD

Per [Kohavi1996]:

$$\mathbb{E}_\theta[\mathrm{err}_T(\theta)] = \mathbf{bias}_T^2 + \mathbf{var}_T.$$

- $\mathbf{bias}_T^2$: expected bias over T, $\mathbf{bias}(x,y) = y - \mathbb{E}_\theta\left[f_\theta(x)\right]$ 3
- $\mathbf{var}_T$: expected variance over T, $\mathbf{var}(x) = \mathbb{E}_\theta\left[\left(f_\theta(x) - \mathbb{E}_\theta\left[f_\theta(x)\right]\right)^2\right]$ 4

3 Bias and Correlation Shift

We show that for large NNs:

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

→ **bias** in OOD increases when the class posteriors mismatch.

4 Variance and Diversity Shift

We show that for NNs with diagonally dominant NTK:

$$\mathbf{var}_{d_T} \propto MMD^2_{NTK^2}(X_{d_S}, X_{d_T}) + \ \dots$$

d_S source dataset input support X_{d_S} resp. d_T target dataset support X_{d_T}.

→ **var** in OOD increases when the input marginals mismatch.

5 Controlling Diversity Shift with (Costly) Ensembling

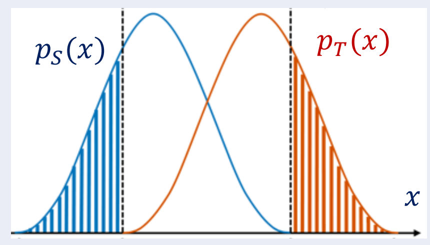
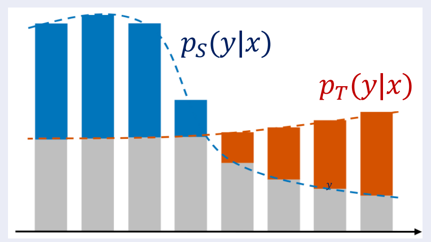
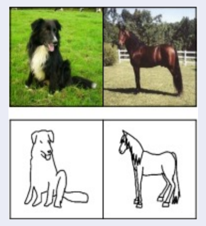
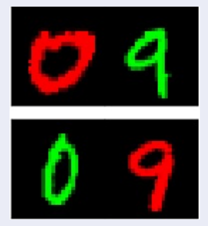
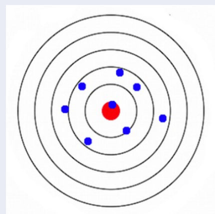
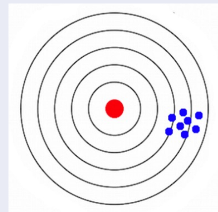
Ensembling averages predictions from different models: $f_{ENS} = \sum_{m=1}^{M} f_m$.

Bias-variance-covariance decomposition for ensembling [Ueda1996]:

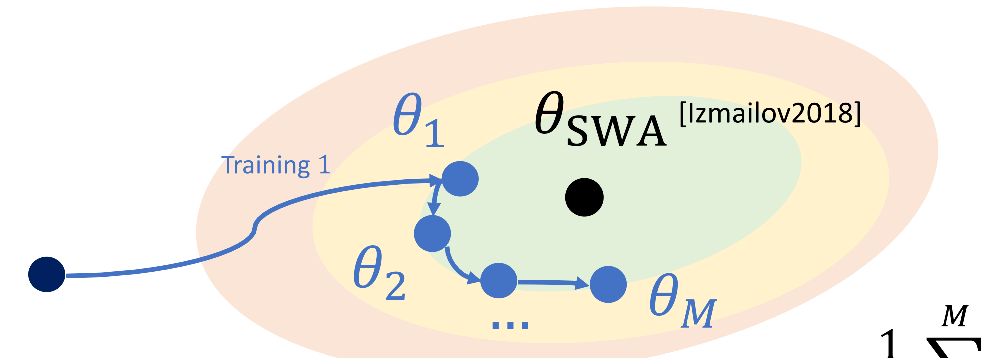
$$\mathbb{E}_{ens}[\mathrm{err}_T(ens)] = \mathbf{bias}_T^2 + \frac{1}{M}\mathbf{var}_T + \frac{M-1}{M}\mathbf{cov}_T,$$

- $\mathbf{bias}_T$: expected bias of a single model over T, 3
- $\mathbf{var}_T$: expected variance of a single model over T, 4
- $\mathbf{cov}_T$: expected covariance across models over T, 7
  $\mathbf{cov}(x) = \mathbb{E}_{\theta,\theta'}\left[\left(f_\theta(x) - \mathbb{E}_\theta\left[f_\theta(x)\right]\right)\left(f_{\theta'}(x) - \mathbb{E}_\theta\left[f_\theta(x)\right]\right)\right]$

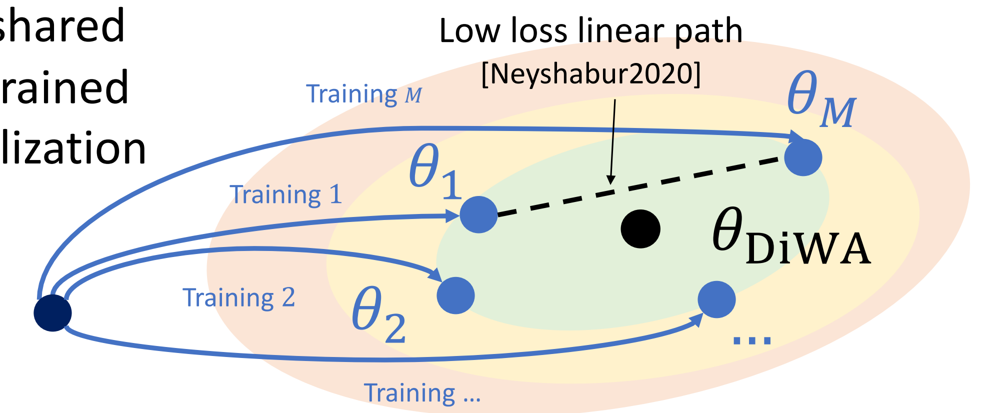
→ Factor $1/M$ reduces **var** *i.e.* ensembling handles diversity shift.
→ Ensembling cannot reduce **bias** *i.e.* correlation shift.
→ **cov** should be controlled.

Shift	Diversity	Correlation
Definition	$p_S(X) \neq p_T(X)$ 	$p_S(Y\|X) \neq p_T(Y\|X)$ 
Dataset	PACS, OfficeHome…	ColoredMNIST, CelebA…
Sample		
Bias-variance	Small bias Large variance 	Large bias Small variance 
Current SoTA	This paper: **DiWA**	Invariance: IRM, Coral Robust optim: gDRO

6 Cheap Approx. to Ensembling: Weight Averaging



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



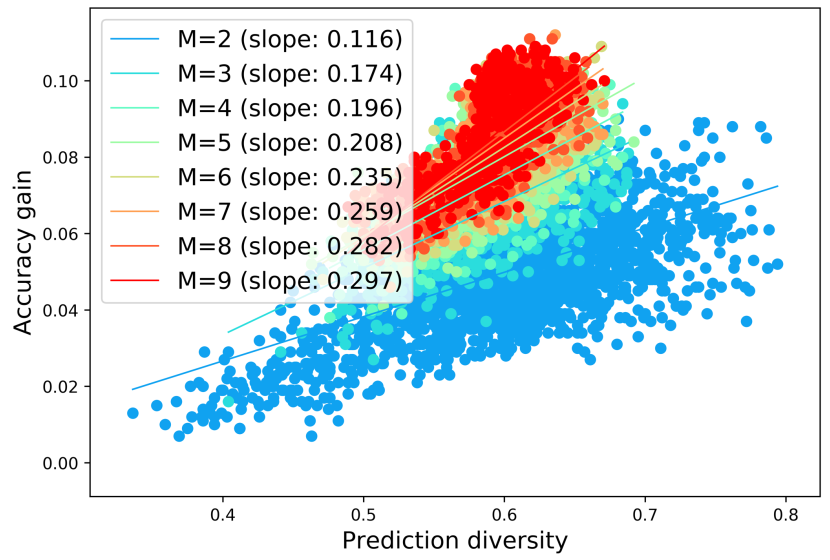
$$\mathbb{E}_{\theta_{WA}}[\mathrm{err}_T(\theta_{WA})] = \mathbb{E}_{ens}[\mathrm{err}_T(ens)] + \mathcal{O}(\bar{\Delta}^2),$$

- $\bar{\Delta}^2 = max_{m=1}^{M}\|\theta_m - \theta_{WA}\|^2$: locality constraint 8

→ Advantages of ensembling without inference cost.

7 Covariance and Diversity

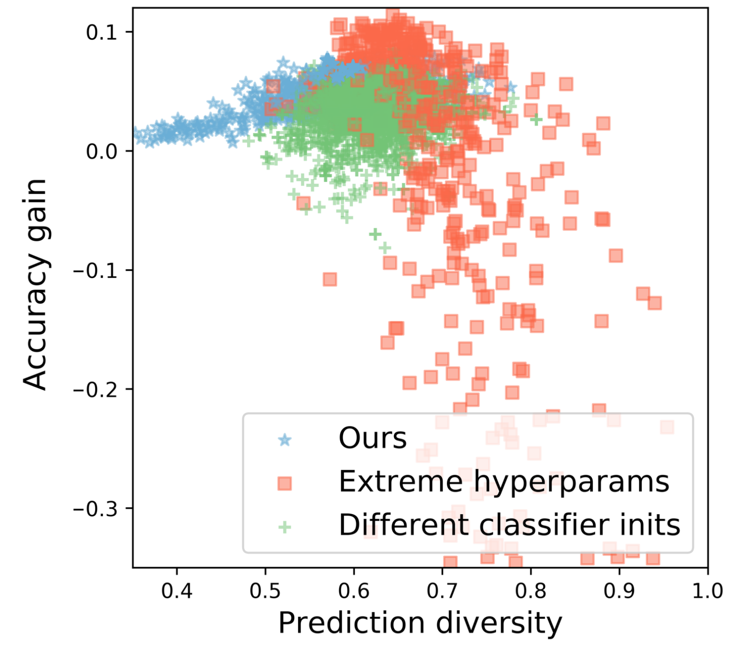
Legend: Each dot is the accuracy gain of combining M models over the average accuracy *w.r.t.* diversity.



→ **cov** reduced with diversity.
→ Gain of WA improves with diversity.
→ Regression's slope increases with M.

8 Diversity-Averageability Trade-off

Legend: Each dot is the accuracy gain of combining M models over the average accuracy *w.r.t.* diversity.



→ Increase diversity in data/learning procedure as long as linear mode connectivity is satisfied.

9 DiWA is state-of-the-art on DomainBed

Reference DomainBed benchmark [Gulrajani2021] and representative baselines.

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**

References

[Cha2021]: Swad: Domain generalization by seeking flat minima. NeurIPS.
[Gulrajani2021]: In search of lost domain generalization. ICLR.
[Izmailov2018]: Averaging Weights Leads to Wider Optima and Better Generalization. UAI.
[Kohavi1996]: Bias plus variance decomposition for zero-one loss functions. ICML.
[Neyshabur2020]: What is being transferred in transfer learning? NeurIPS.
[Sun2016]: Correlation Alignment for Unsupervised Domain Adaptation. AAAI.
[Ueda1996]: Generalization error of ensemble estimators.
[Ye2022]: Ood-bench benchmarking and understanding OOD generalization. CVPR.

ArXiv: https://arxiv.org/abs/2205.09739
Code: https://github.com/alexrame/diwa
Contact: first.last@sorbonne-universite.fr