Diverse Weight Averaging for Out-Of-Distribution Generalization

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

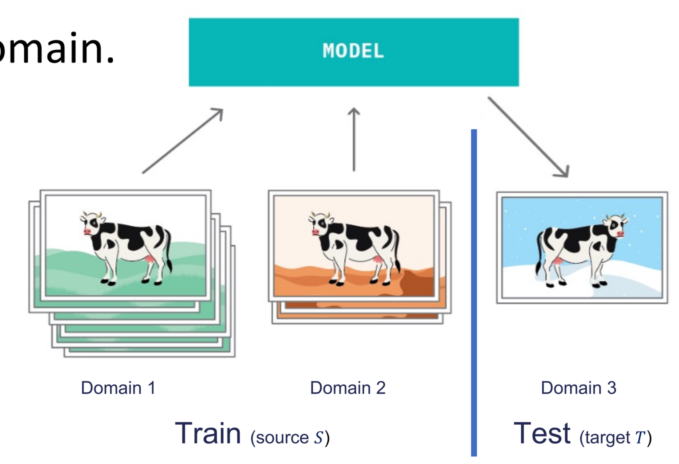
NEURAL INFORMATION PROCESSING SYSTEMS

1 OOD Generalization and Weight Averaging

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

Under domain shifts divided per [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]: $\mathbb{E}_\theta[\mathrm{err}_T(\theta)] = \mathrm{bias}_T^2 + \mathrm{var}_T$

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

3 Bias and Correlation Shift

For large NNs:

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

➜ bias in OOD increases when the posteriors mismatch.

4 Variance and Diversity Shift

For NNs with diagonally dominant NTK:

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

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

➜ var in OOD increases when the marginals mismatch.

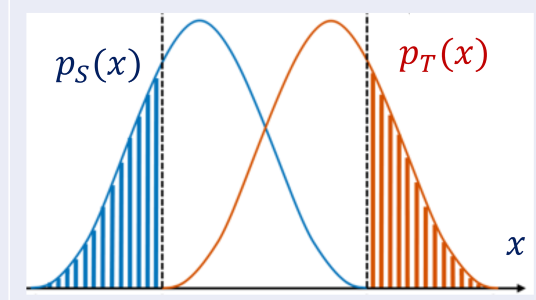
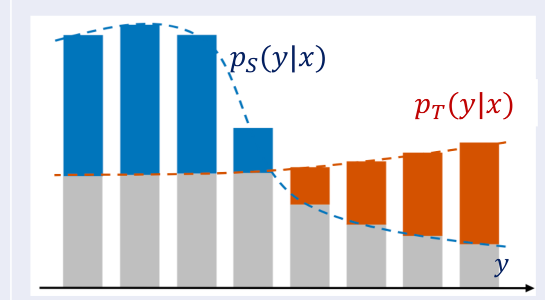
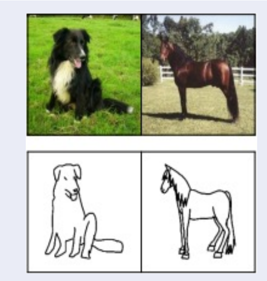
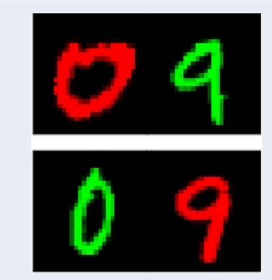
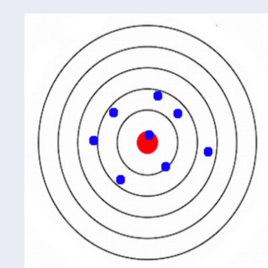
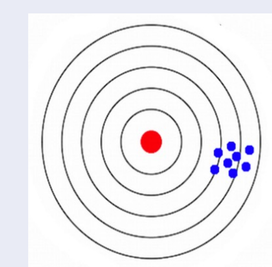
5 Controlling Diversity Shift with Ensembling

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

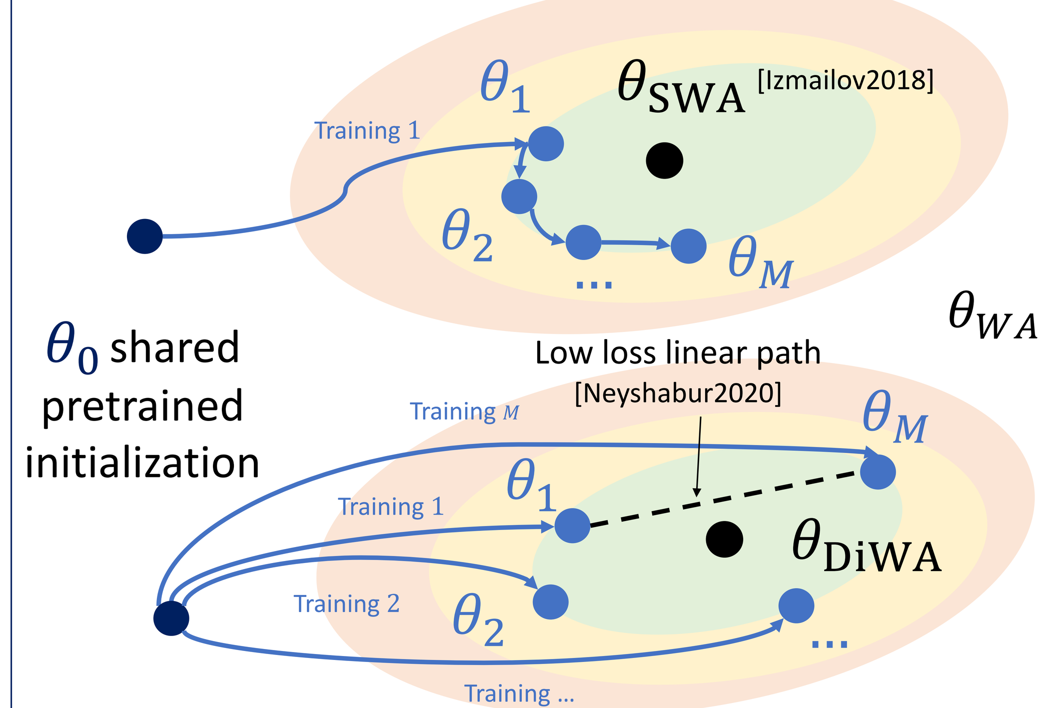
$$\mathbb{E}_{ens}[\mathrm{err}_T(ens)] = \mathrm{bias}_T^2 + \frac{1}{M}\mathrm{var}_T + \frac{M-1}{M}\mathrm{cov}_T,$$

- bias_T: bias of a single model averaged over T, 3
- var_T: variance of a single model averaged over T, 4
- cov_T: covariance, $\mathrm{cov}(x) = \mathbb{E}_{\theta,\theta'}[(f_\theta(x) - \mathbb{E}_\theta[f_\theta(x)])(f_{\theta'}(x) - \mathbb{E}_\theta[f_\theta(x)])]$ 7

➜ Factor $1/M$ reduces var *i.e.* ensembling handles diversity shift.
➜ Ensembling cannot reduce bias *i.e.* correlation shift.
➜ cov should be controlled to control the target error.

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 Weight Averaging and Ensembling



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

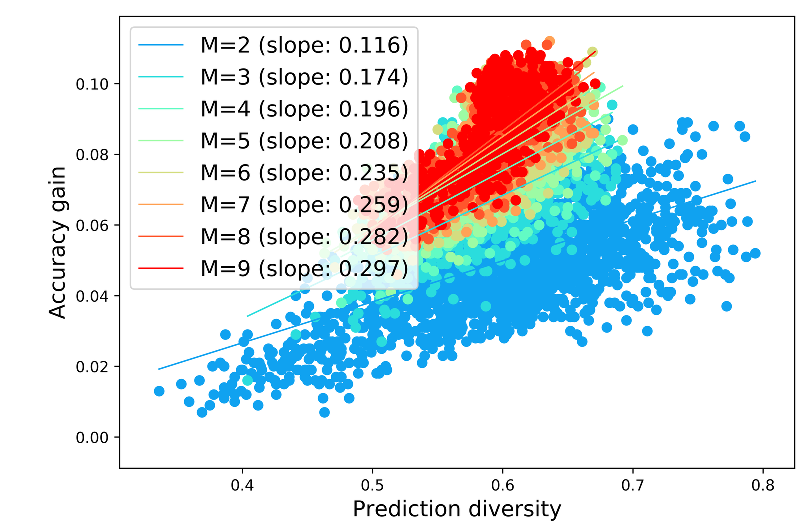
$$\mathbb{E}_{\theta_{WA}}[\mathrm{err}_T(\theta_{WA})] = \mathbb{E}_{ens}[\mathrm{err}_T(ens)] + \mathcal{O}(\bar{\Delta}^2)$$

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

➜ WA has the advantages of ensembling without inference cost.

7 Covariance and Diversity

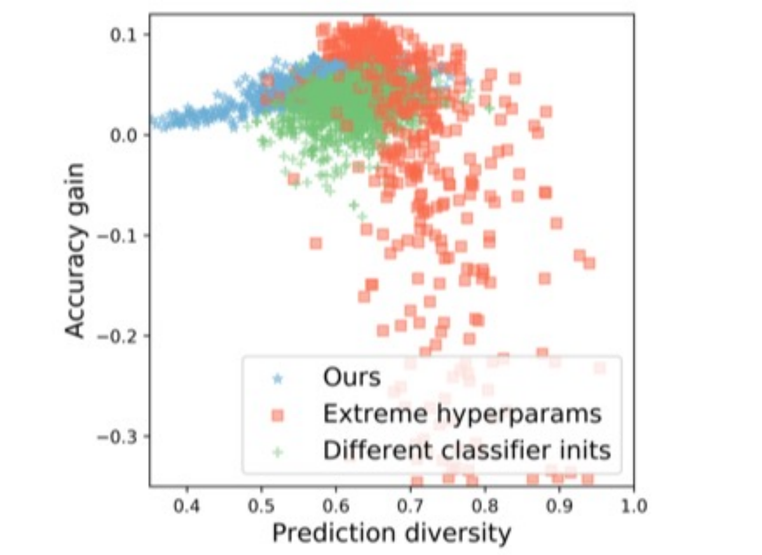
Legend: Each dot is the accuracy gain of combining M models over the average accuracy *w.r.t.* diversity.



➜ cov reduced with diversity
➜ Gain in accuracy of WA improves with diversity
➜ Linear regression's slope increases with M

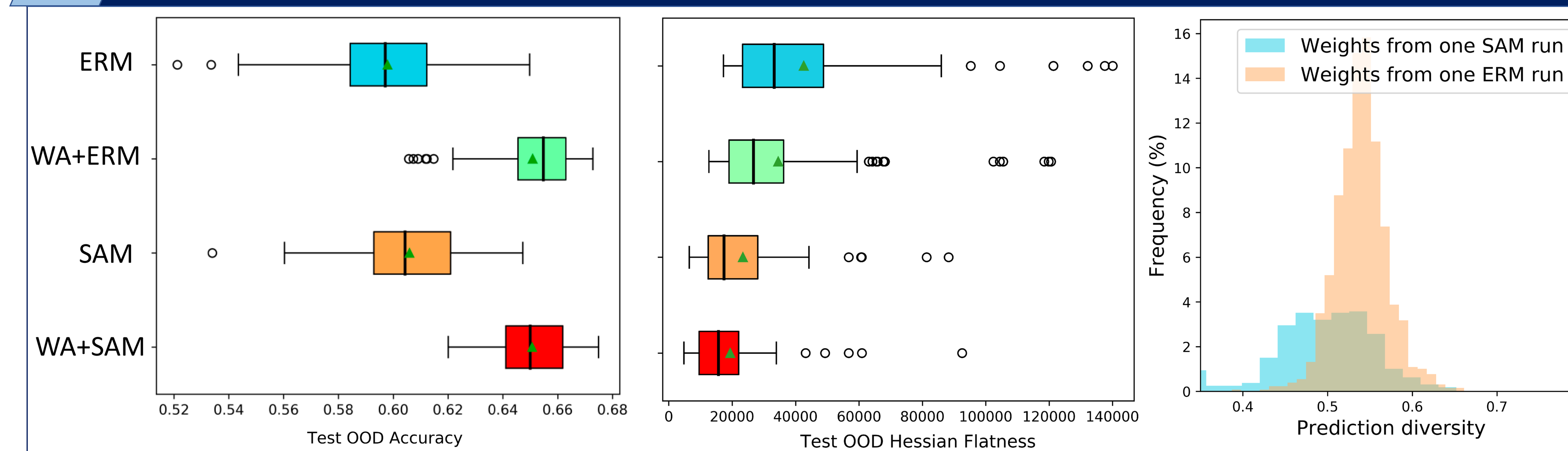
8 Diversity-Averageability trade-off

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 Prior Limitations Handled By Our Analysis



SAM [Foret2021]; WA+SAM [Kaddour2022] have worse OOD despite more flatness ➜ contradicts [Cha2021].

Our analysis explains this result:
➜ WA benefits from ensembling (unlike SAM).
➜ ERM has more diversity than SAM.

References

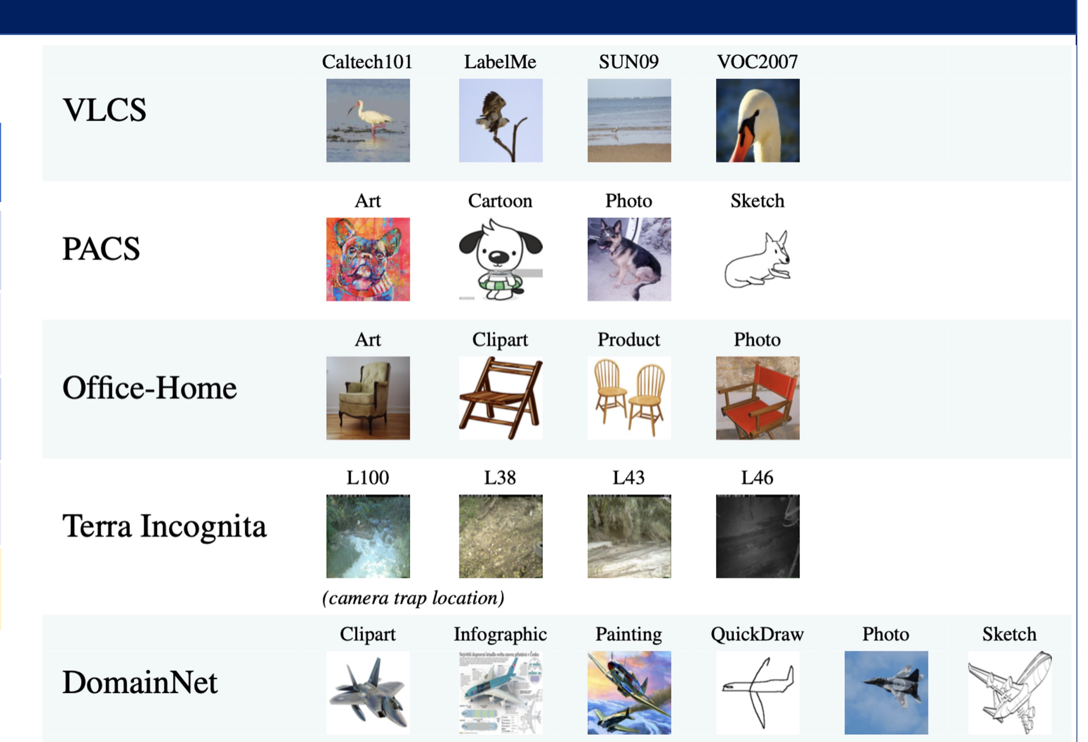
[Cha2021]: Swad: Domain generalization by seeking flat minima. NeurIPS.
[Foret2021]: Sharpness-aware minimization for efficiently improving generalization. ICLR.
[Gulrajani2021]: In search of lost domain generalization. ICLR.
[Izmailov2018]: Averaging Weights Leads to Wider Optima and Better Generalization. UAI.
[Kaddour2022]: A Fair Comparison of Two Popular Flat Minima Optimizers: SWA vs. SAM. NeurIPS.
[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 datasets and algorithms. CVPR.

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

Various methods on DomainBed [Gulrajani2021]:

- Invariance: CORAL [Sun2016] ~ ERM
- WA: SWAD [Cha2021] ≫ ERM
- Ensembling: ENS ≫ ERM - high inference cost
- DiWA is SoTA - low inference cost

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**



Contact

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