A Bias-variance Analysis of Weight Averaging for OOD Generalization

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

ICML International Conference On Machine Learning
Principles of Distribution Shift (PODS) Workshop

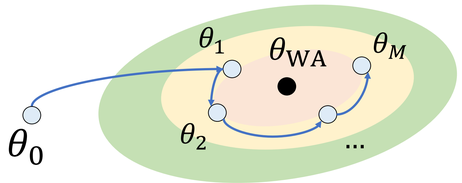
Check out DiWA extending this work

1 OOD Generalization and Weight Averaging (WA)

We consider OOD generalization:
- Train on S source domain and test on T target domain.
- Under domain shifts divided per [Ye2022] into:
  - Correlation shift (concept shift): $p_S(Y|X) \neq p_T(Y|X)$
  - Diversity shift (covariate shift): $p_S(X) \neq p_T(X)$

Various approaches:
- Domain-invariant OOD methods (IRM [Arjovsky2019], Fishr [Rame2022]) $\sim$ ERM [Gulrajani2021].
- Weight Averaging (WA) (SWAD [Cha2021], MA [Arpit2021]) $\gg$ ERM.

Dataset	Domains				
Colored MNIST $p_S(Y\|X) \neq p_T(Y\|X)$	+90%	+80%	-90%		
	(degree of correlation between color and label)				
Office-Home $p_S(X) \neq p_T(X)$	Art	Clipart	Product	Photo	



$\theta_{\text{WA}} = \frac{1}{M}\sum_{m=1}^{M}\theta_m$ averages all weights along optimization

2 Bias-Variance Analysis of WA

We extend the bias-variance-covariance decomposition for ensembling [Ueda1996] to WA.

$l_S = \{d_S, c\}$ a learning procedure: d_S source dataset and c randomness $\rightarrow \theta(l_S)$: trained weights.

In expectation over the M correlated learning procedures $L_S^M = \left\{l_S^{(m)}\right\}_{m=1}^{M}$ for the weight average $\theta_{\text{WA}}(L_S^M)$,

$$\mathbb{E}_{L_S^M}\mathcal{E}_T(\theta_{\text{WA}}(L_S^M)) = \mathbb{E}_{(x,y)\sim p_T}\left[\text{bias}^2(x,y) + \frac{1}{M}\text{var}(x) + \frac{M-1}{M}\text{cov}(x)\right] + O(\bar{\Delta}^2),$$

$\text{bias}(x,y) = y - \bar{f}_S(x)$, where $\bar{f}_S(x) = \mathbb{E}_{l_S}[f(x,\theta(l_S))]$

$\text{var}(x) = \mathbb{E}_{l_S}\left[\left(f(x,\theta(l_S)) - \bar{f}_S(x)\right)^2\right]$, $\Longrightarrow$ divided by M

$\text{cov}(x) = \mathbb{E}_{l_S,l'_S}\left[\left(f(x,\theta(l_S)) - \bar{f}_S(x)\right)\left(f(x,\theta(l'_S)) - \bar{f}_S(x)\right)\right]$, $\Longrightarrow$ functionally diverse members reduce covariance

$\bar{\Delta}^2 = \mathbb{E}_{L_S^M}\Delta_{L_S^M}^2$ with $\Delta_{L_S^M} = \max_{m=1}^{M}\|\theta_m - \theta_{\text{WA}}\|_2$. $\Longrightarrow$ ensures that $f_{\text{WA}} \sim \frac{1}{M}\sum_{m=1}^{M} f(\cdot,\theta_m)$ (ensembling)

3 Variance and Diversity Shift

We fix the source (resp. target) dataset d_S (d_T) with input support $X_{d_S}(X_{d_T})$

Assumption 1 (Kernel regime, constant norm and low similarity). *f is in the kernel regime. Its kernel K satisfies $\exists(\lambda_S,\epsilon)$ with $0 \leq \epsilon \ll \lambda_S$ s.t. $\forall x_S \in X_{d_S}, K(x_S,x_S) = \lambda_S$ and $\forall x'_S \neq x_S \in X_{d_S}, |K(x_S,x'_S)| \leq \epsilon$.*

Proposition 1. *With MMD, the empirical maximum mean discrepancy in the RKHS of K^2, under Assumption 2,*

$$\mathbb{E}_{x_T\in X_{d_T}}[\text{var}(x_T)] = \frac{n_S}{2\lambda_S}MMD^2(X_{d_S},X_{d_T}) + \lambda_T - \frac{n_S}{2\lambda_S}\beta_T + O(\epsilon)$$

WA divides variance by M explaining its success under diversity shift e.g. on real-world datasets from DomainBed [Gulrajani2021].

4 Covariance and Diversity

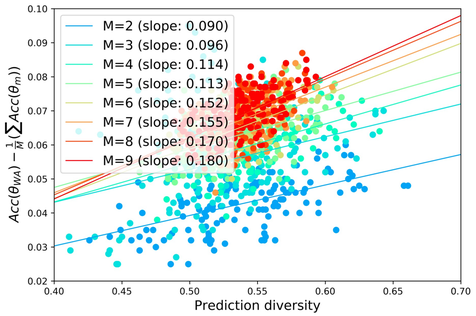


Figure 4: Each dot is the accuracy (↑) gain on "Art" (Office-Home) of WA over its members vs. prediction diversity (↑).

5 Bias and Correlation Shift

Assumption 2 (Small bias). *$\exists\epsilon > 0$ small s.t. $\forall x \in \mathcal{X}_S, |f_S(x) - \bar{f}_S(x)| \leq \epsilon$ where $\bar{f}_S(x) = \mathbb{E}_{l_S}[f(x,\theta(l_S))]$*

Proposition 2 *With a bounded difference between labeling functions $f_T - f_S$ on $\mathcal{X}_T \cap \mathcal{X}_S$, under Assumption 1,*

$\mathbb{E}_{(x,y)\sim p_T}[\text{bias}^2(x,y)] = \textit{Correlation Shift} + \textit{Support Mismatch} + O(\epsilon)$,

where Correlation Shift $= \int_{\mathcal{X}_T\cap\mathcal{X}_S}\left(f_T(x) - f_S(x)\right)^2 p_T(x)dx$ *and Support Mismatch* $= \int_{\mathcal{X}_T\setminus\mathcal{X}_S}\left(f_T(x) - \bar{f}_S(x)\right)^2 p_T(x)dx$.

WA cannot reduce bias ⇒ WA cannot tackle correlation shift (e.g. on Colored MNIST).

6 Limitations of Existing Analysis are Tackled by Our Analysis

[Cha2021] states that flatness of the loss landscape explains WA's success OOD. Yet,

1) <u>Uncontrolled OOD error:</u> flatness does not reduce domain shift in [Cha2021].
➢ We show that large M controls the variance thus the error under diversity shift.

2) <u>WA v.s. SAM:</u> SAM [Foret2021] also flattens minimas but fails OOD.
➢ We show that WA succeeds OOD thanks to similarity with ensembling (unlike SAM).

3) <u>WA+SAM:</u> flattens even more minimas but slightly worse OOD acc. than WA.
➢ Diversity across checkpoints for SAM is lower than ERM.

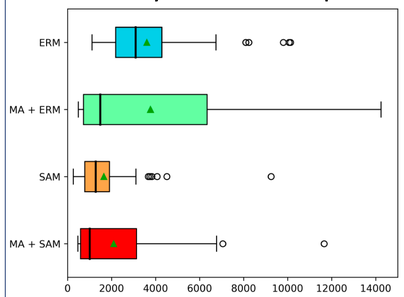


Figure 1: Train Hessian trace (↓) on "Clipart+Product+Photo" (OfficeHome)

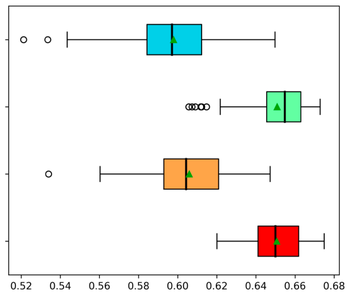


Figure 2: Accuracy in test OOD (↑) on "Art" (OfficeHome).

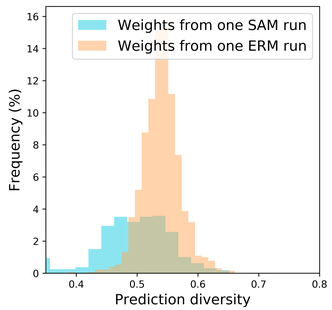


Figure 3: Diversity in ratio-error (↑) on "Art" (OfficeHome).

[Arjovsky2019]: Invariant Risk Minimization.
[Arpit2021]: Ensemble of averages: Improving model selection and boosting performance in domain generalization.
[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
[Rame2022]: Fishr: Invariant Gradient Variances for Out-of-Distribution Generalization. ICML
[Ueda1996]: Generalization error of ensemble estimators.
[Ye2022]: Ood-bench: Benchmarking and understanding OOD generalization datasets and algorithms. CVPR