ICML International Conference On Machine Learning

Principles of Distribution Shift Workshop

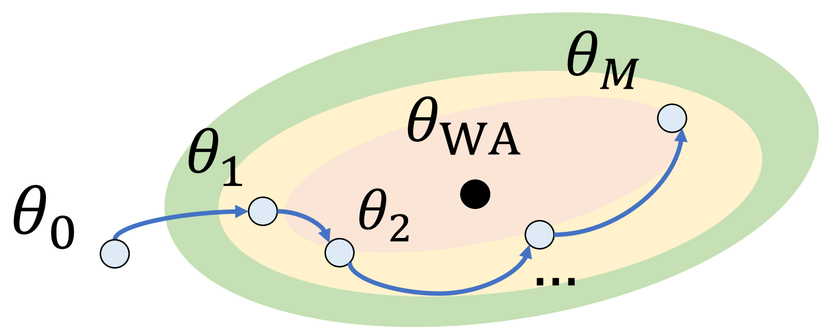
A Bias-variance Analysis of Weight Averaging for OOD Generalization

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

1 OOD Generalization and Weight Averaging

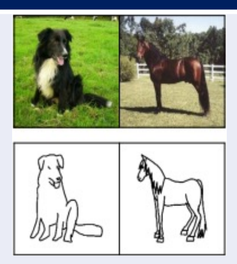
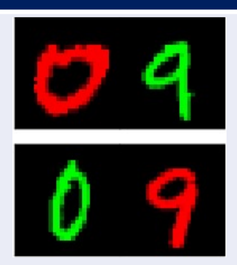
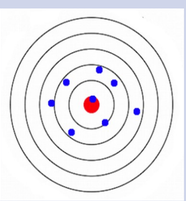
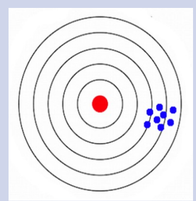
Context: OOD generalization
- Train on S source domain and test on T target domain
- Under domain shifts divided per [Ye2022] into:
  - Diversity shift: $p_S(X) \neq p_T(X)$
  - Correlation shift: $p_S(Y|X) \neq p_T(Y|X)$

Various methods; on real-world datasets [Gulrajani2021]:
- Domain-invariance [Arjovsky2019,Rame2022] ~ ERM
- Weight Averaging (WA) [Cha2021, Arpit2021] ≫ ERM



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

averages all weights along optimization

Shift	Diversity	Correlation
Sample		
Dataset	OfficeHome	ColoredMNIST
Bias-variance	Small bias Large variance 	Large bias Small variance 
Current SoTA	Multiple models: Ensembling / WA	Multiple domains: Invariance

2 Bias-Variance Analysis of Weight Averaging

Bias-variance-covariance decomposition for ensembling [Ueda1996] extended to WA:

$$\mathbb{E}_{\theta_{WA}}\text{err}_T(\theta_{WA}) = \frac{1}{M}\textbf{var} + \left(1-\frac{1}{M}\right)\textbf{cov} + \textbf{bias}^2 + \mathcal{O}(\bar{\Delta}^2)$$

(3) (4) (5)

- **var**: OOD variance of a single model averaged over T,
- **cov**: OOD covariance across models, averaged over T,
- **bias**: OOD bias of a single model averaged over T,
- $\bar{\Delta}^2$: locality constraint *s.t.* $f_{\text{WA}} \sim \frac{1}{M}\sum_{m=1}^{M} f(\cdot,\theta_m)$.

3 Variance and Diversity Shift

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

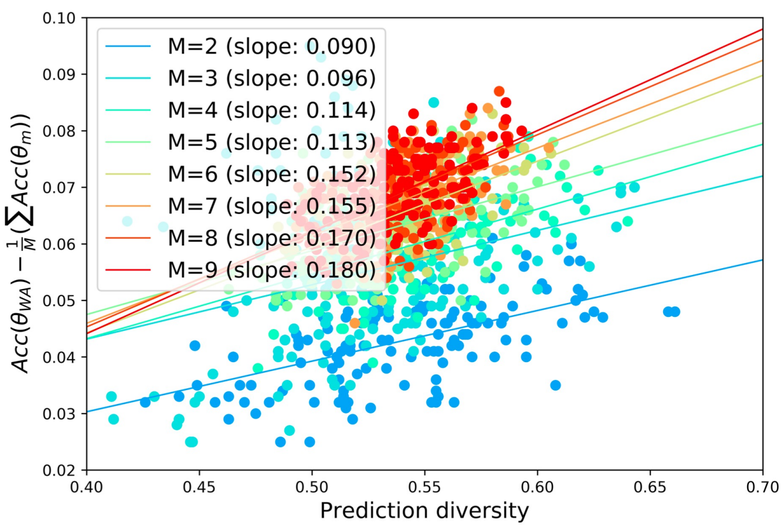
Proposition 1. *If f is in the kernel regime and its NTK K is diagonally dominant on d_S with off-diagonal terms $\leq \epsilon$,*

$$\textbf{var} \propto MMD^2_{K^2}(X_{d_S}, X_{d_T}) + cst_T + O(\epsilon)$$

→ Factor $1/M$ reduces **var** and handles diversity shift, *c.f.* WA's success on real-world datasets DomainBed [Gulrajani2021]

4 Covariance and Diversity

Figure 1: Each dot is the accuracy (↑) gain on "Art" (Office-Home) of WA over its members vs. prediction diversity (↑).



→ **Covariance** reduced with diversity
→ Gain in accuracy of WA improves with diversity
→ Linear regression's slope increases with M

5 Bias and Correlation Shift

$\mathcal{X}_S, \mathcal{X}_T$: input domains; f_S, f_T labelling functions

$$\forall x \in \mathcal{X}_S, f_S(x) = \mathbb{E}_{p_S}[Y|x],\ \forall x \in \mathcal{X}_T, f_T(x) = \mathbb{E}_{p_T}[Y|x]$$

Proposition 2. *If there is small bias on S ($\leq \epsilon$),*

$$\textbf{bias}^2 = \textit{Correlation Shift} + \textit{Support Mismatch} + O(\epsilon),$$

$$\textit{where Correlation Shift} = \int_{\mathcal{X}_T \cap \mathcal{X}_S} \left(f_T(x) - f_S(x)\right)^2 p_T(x)dx$$

$$\textit{and Support Mismatch} = \int_{\mathcal{X}_T \setminus \mathcal{X}_S} \left(f_T(x) - \mathbb{E}_\theta\left[f\left(x,\theta\right)\right]\right)^2 p_T(x)dx.$$

→ WA cannot reduce **bias** which dominates when posteriors mismatch, *i.e.*, under correlation shift

6 Prior Limitations Handled By Our Analysis

Flatness-based analysis [Cha2021] cannot explain why SAM [Foret2021] and WA+SAM [Kaddour2022] have flatter minimas than WA but worse OOD accuracy. Explained by our analysis:
→ WA benefits from ensembling (unlike SAM)
→ ERM has more diversity than SAM

Figure 2: Accuracy in test OOD (↑)

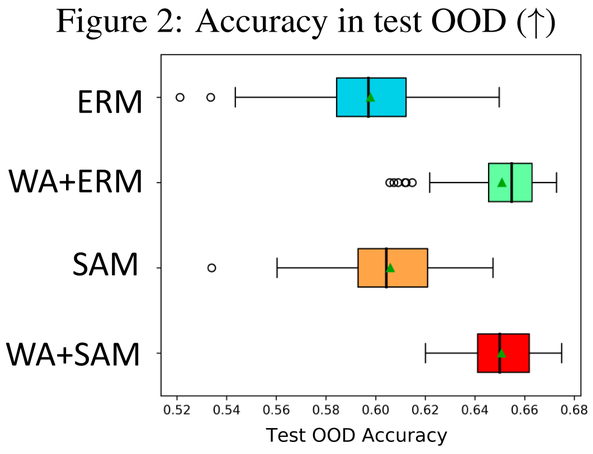
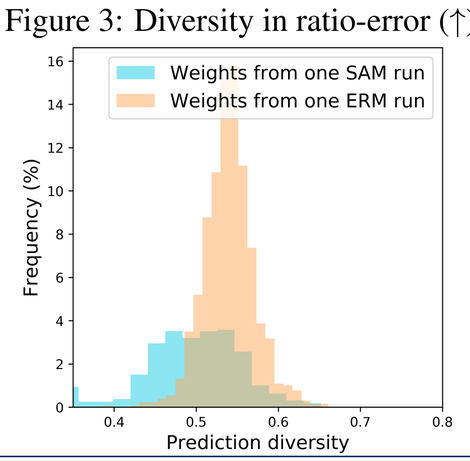


Figure 3: Diversity in ratio-error (↑)



References

[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.
[Kaddour2022]: A Fair Comparison of Two Popular Flat Minima Optimizers: Stochastic Weight Averaging vs. Sharpness-Aware Minimization.
[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.

Check out our follow-up, DiWA, SoTA on DomainBed !