Fishr: Invariant Gradient Variances for Out-of-Distribution Generalization

Alexandre Ramé (PhD Sorbonne Univ.)
Corentin Dancette (PhD Sorbonne Univ.)
Matthieu Cord (Professor Sorbonne Univ.)

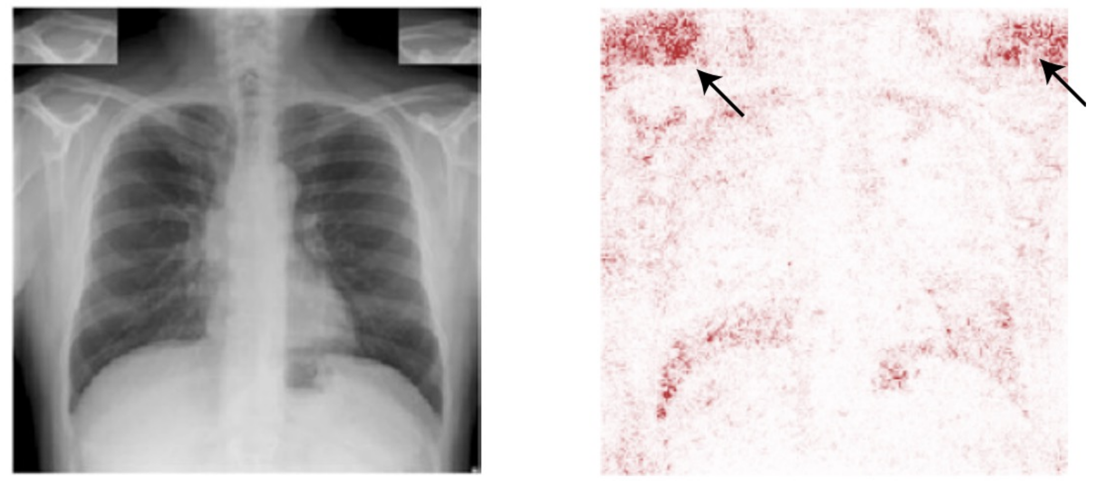
Github: alexrame/fishr

1 Simplicity Bias and Domain Shifts

DNNs learn simple/biased shortcuts rather than complex/stable features.

Ex: to detect Covid, DNNs analyze body/shoulder positions rather than lung fields.

Negative image with shoulders moved — Important pixels



$\Longrightarrow$ lack of robustness, failures under shifts.

Main challenge: we want to detect causation rather than correlation.

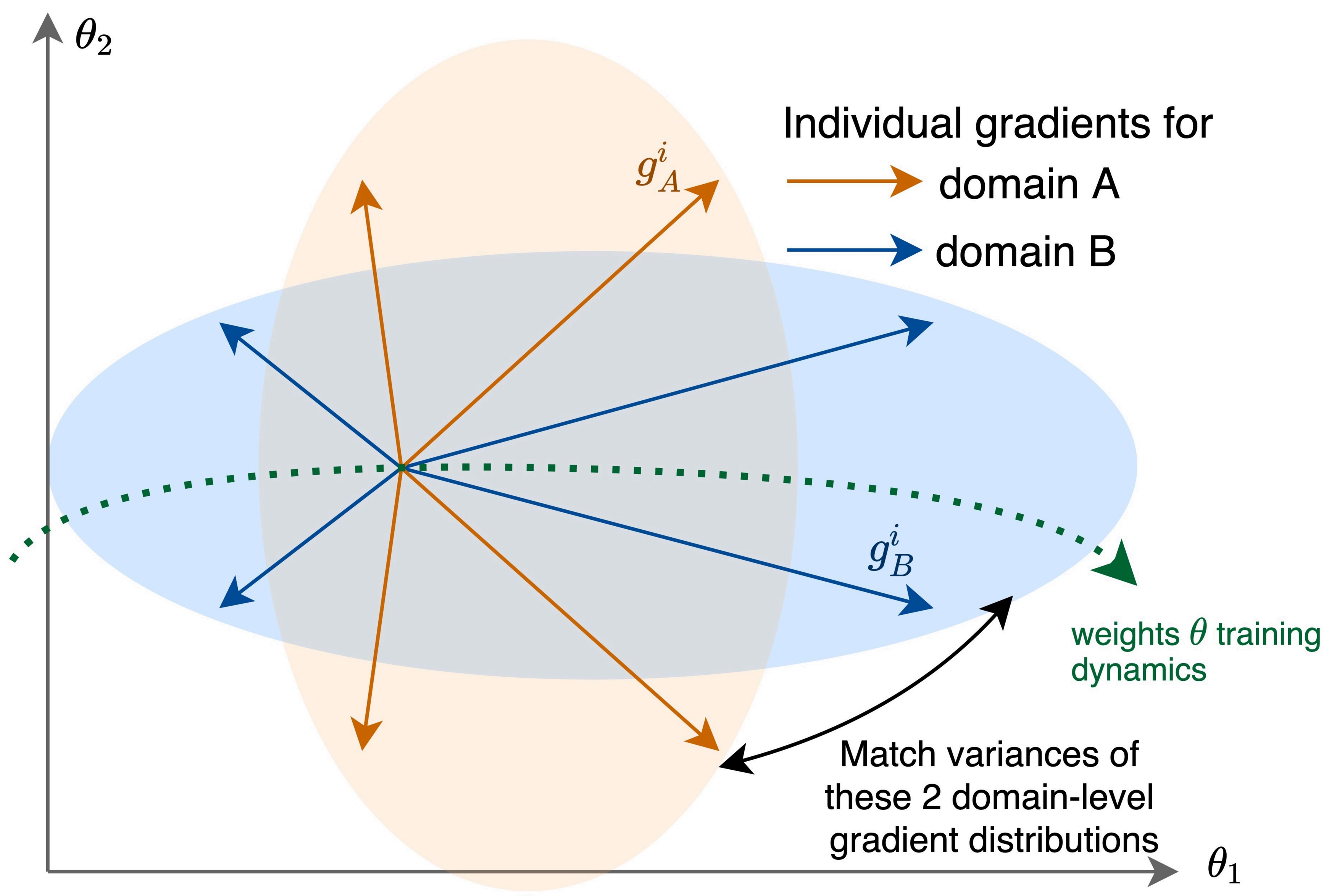
2 Invariance Paradigm

Assumption: the causal mechanism is invariant across the various training domains .

How to enforce *invariance* across domains?

- In *features* extracted by the encoder $f_\phi(x)$. Ex: CORAL(AAAI16), DANN(JMLR16).
- In *predictors* with invariant risks $l(f_\theta(x), y)$. Ex: IRM, V-Rex(ICML21).
- In *gradients* of the loss w.r.t. the weights of the network $\nabla_\theta l(f_\theta(x), y)$. Ex: Fish(ICLR22).

3 Fishr: Invariant Gradient Variances ... and also invariant risks and Hessians



$$\mathcal{L}_{Fishr} = \mathcal{L}_{ERM} + \lambda \parallel Var(G_A) - Var(G_B) \parallel_2^2,$$

λ controls the strength of our regularization matching the variance of gradients $G_e = \left[\nabla_\theta l\left(f_\theta(x_e^i), y_e^i\right)\right]_{i=1}^{n_e}$ across domains $e \in \{A, B\}$.

Theoretical analysis

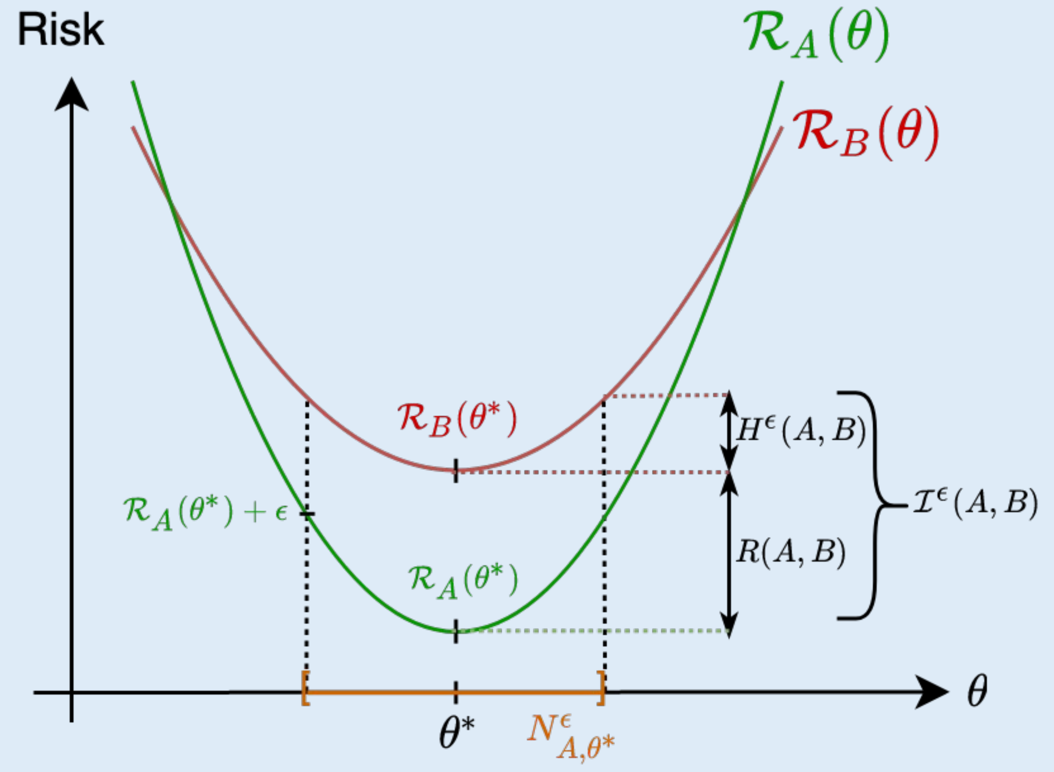
Gradient variances as a proxy to match:

- risks: $\mathcal{R}_e(\theta) = \sum_i l\left(f_\theta(x_e^i), y_e^i\right)/n_e$,
- Hessians: $\mathcal{H}_e(\theta) = \sum_i \nabla_\theta^2 l\left(f_\theta(x_e^i), y_e^i\right)/n_e$ (proof via the Fisher Information Matrix).

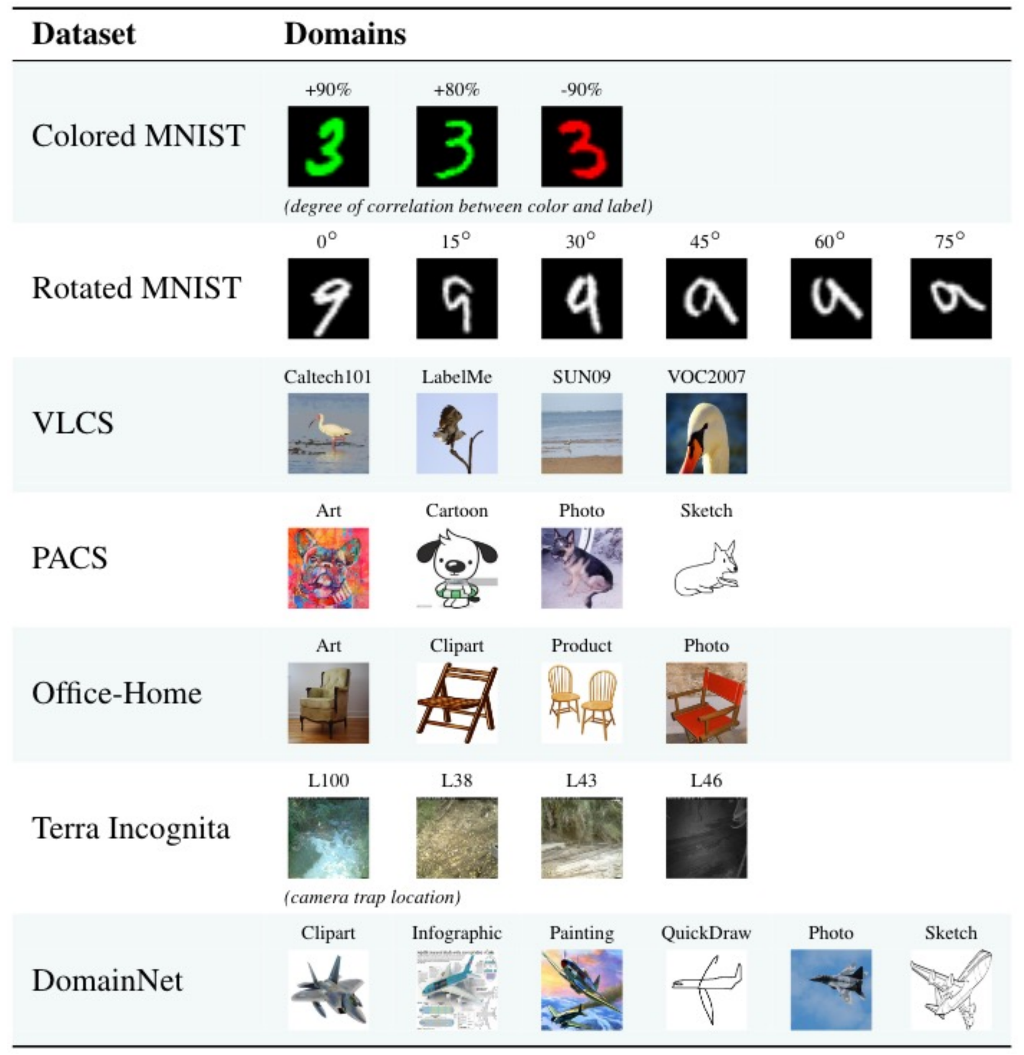
We prove Fishr reduces *domain inconsistencies*:

$$\mathcal{I}^\varepsilon(A,B) = max_{\theta \in N_{A,\theta^*}^\varepsilon} |R_B(\theta) - R_A(\theta^*)|$$

in a ε neighbourhood $N_{A,\theta^*}^\varepsilon$ around weights θ^*.



4 State-of-the-art Performances on DomainBed Benchmark (at almost no computational overhead)



Algo.	Invariance	Acc. ↑ cMNIST	rMNIST	VLCS	PACS	OHome	TerraI	DNet	Avg	Rank ↓ Avg
ERM	✗	57.8	97.8	77.6	86.7	66.4	53.0	41.3	68.7	9.1
CORAL	Features	58.6	**98.0**	77.7	87.1	**68.4**	52.8	41.8	69.2	4.6
DANN	Features	57.0	97.9	**79.7**	85.2	65.3	50.6	38.3	67.7	11.9
IRM	Predictors	67.7	97.5	76.9	84.5	63.0	50.5	28.0	66.9	14.7
V-REx	Predictors	67.0	97.9	78.1	**87.2**	65.7	51.4	30.1	68.2	7.7
Fish	Gradients	61.8	97.9	77.8	85.8	66.0	50.8	**43.4**	69.1	8.4
Fishr	Gradients	**68.8**	97.8	78.2	86.9	68.2	**53.6**	41.8	**70.8**	**3.9**