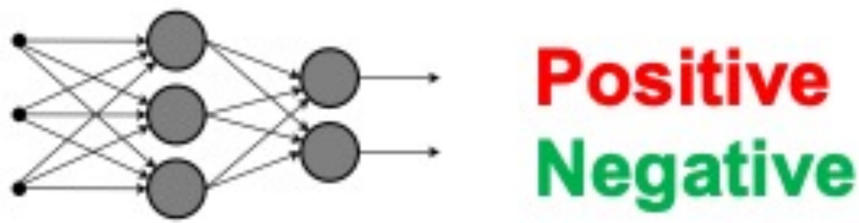
Fishr

Invariant Gradient Variances for Out-of-Distribution Generalization

Alexandre Ramé (PhD)
Corentin Dancette (PhD)
Matthieu Cord (Professor)

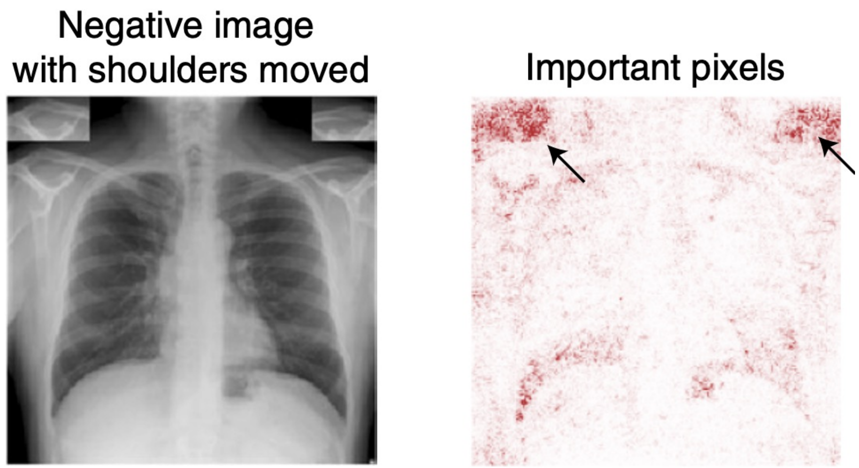
ICML
International Conference On Machine Learning

DNNs to detect Covid from medical scans …



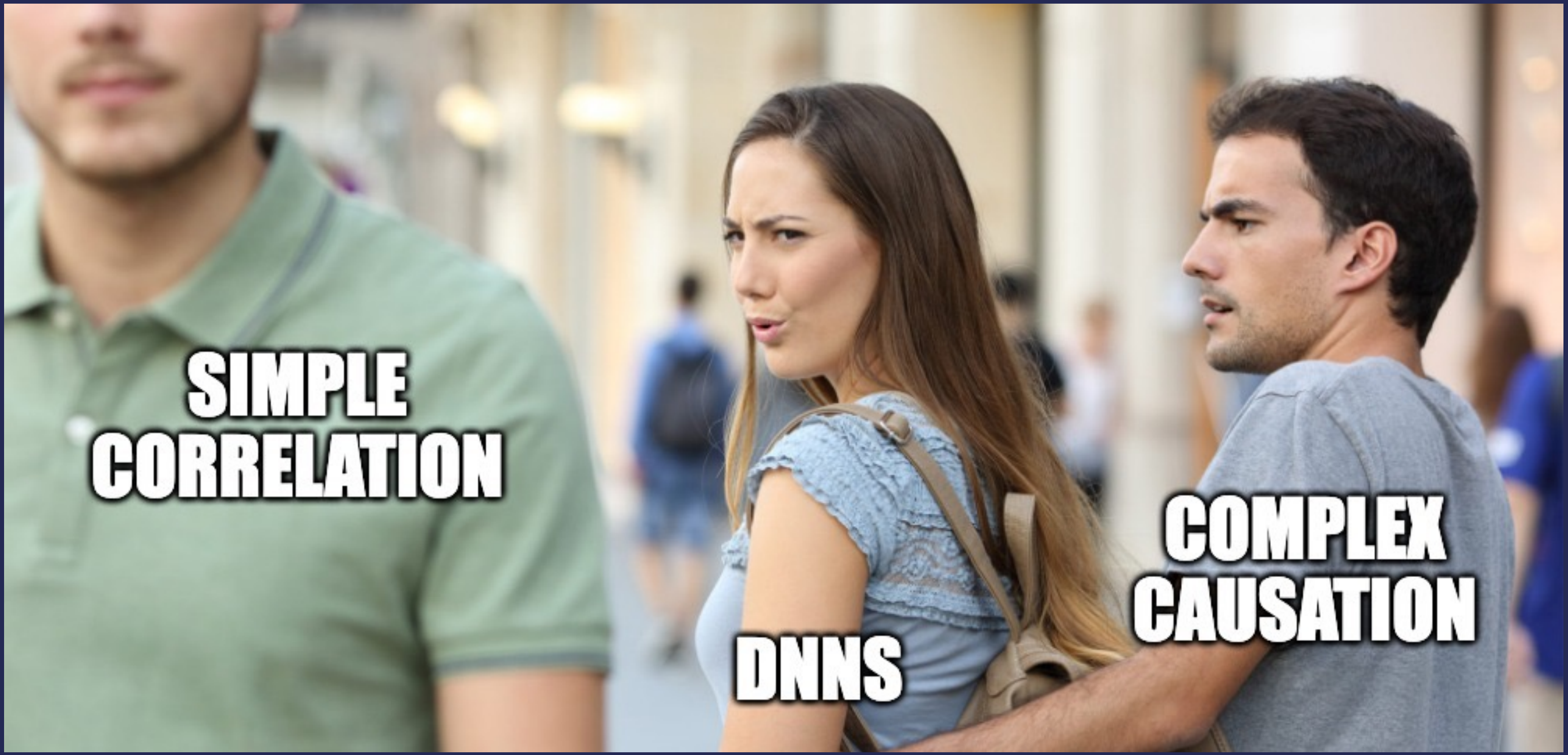
… but DNNs memorized biased shortcuts

- age: children *vs.* adults
- position: standing up *vs.* lying down



(rather than analyzing lung fields)

[1] Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Roberts *et al.*, Nature 2021



$\Longrightarrow$ Simplicity bias deteriorates out-of-distribution generalization

Framework: Training with Multiple Domains

Invariance paradigm: the causal mechanism is invariant across domains

[1] Causal inference by using invariant prediction: identification and confidence intervals. Peters *et al.*, 2016

ERM (empirical risk minimization) and invariance approaches

$$\mathcal{L}_{\mathrm{ERM}} = \mathrm{R_A} + \mathrm{R_B} \longleftarrow \text{Sum of domain-level risks}$$

As most works, we add an invariance regularization on top of ERM:

Hyperparameter

Domain-level statistics

$$\mathcal{L}_{invariance} = \mathcal{L}_{ERM} + \underbrace{\lambda \times distance(\phi_A, \phi_B)}_{\text{Invariance regularization}}$$

Invariance in features or losses

Features

$$f_\theta(x)$$

Ex: Dann [1],
CORAL [2]

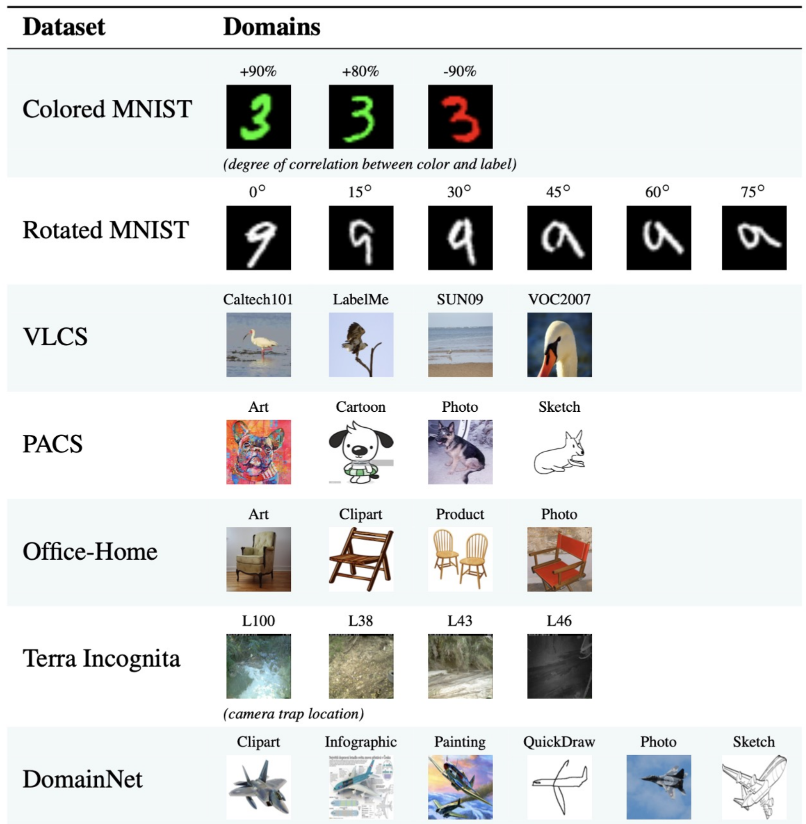
Predictor/Loss

$$l\left(f_\theta(x), y\right)$$

Ex: IRM [3],
V-Rex [4]

[1] Domain-Adversarial Training of Neural Networks. Ganin *et al.*, JMLR 2016
[2] Deep coral: Correlation alignment for deep domain adaptation. Sun and Saenko, ECCV 2016
[3] Invariant risk minimization. Arjovsky *et al.*, 2019
[4] Out-of-distribution generalization via risk extrapolation. Krueger *et al.*, ICML 2021

DomainBed



Dataset	Domains					
Colored MNIST	+90%	+80%	-90%			
	(degree of correlation between color and label)					
Rotated MNIST	0°	15°	30°	45°	60°	75°
VLCS	Caltech101	LabelMe	SUN09	VOC2007		
PACS	Art	Cartoon	Photo	Sketch		
Office-Home	Art	Clipart	Product	Photo		
Terra Incognita	L100	L38	L43	L46		
	(camera trap location)					
DomainNet	Clipart	Infographic	Painting	QuickDraw	Photo	Sketch

No traditional methods outperform ERM in DomainBed

[1] In search of lost domain generalization. Gulrajani and Lopez-Paz, ICLR 2021

Invariance in gradients !

Features

$f_\theta(x)$

Ex: Dann, CORAL

Predictor/Loss

$l\left(f_\theta(x), y\right)$

Ex: IRM, V-REx

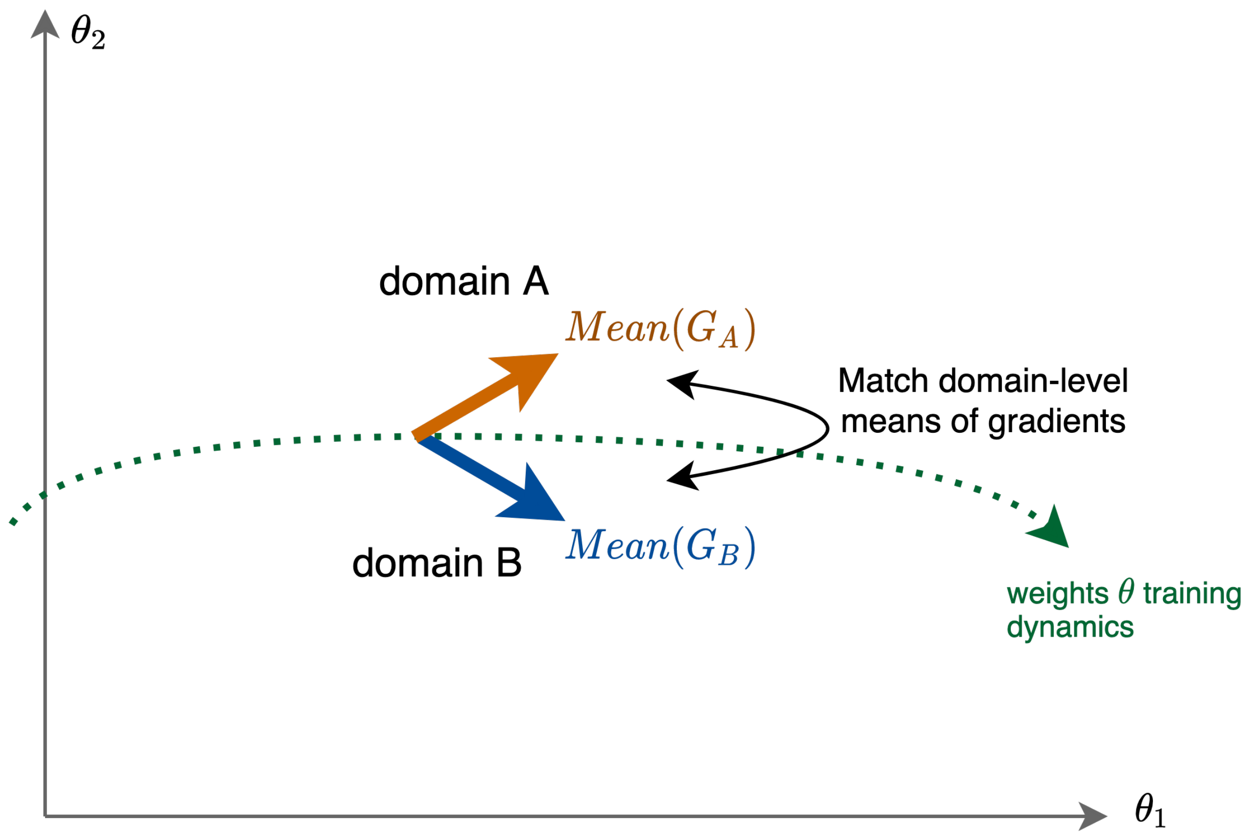
Gradients

$\nabla_\theta l\left(f_\theta(x), y\right)$

$$G_e = \left[\nabla_\theta l\left(f_\theta\left(x_e^i\right), y_e^i\right)\right]_{i=1}^{n_e} \text{ for domain } e \in \{A, B\}$$

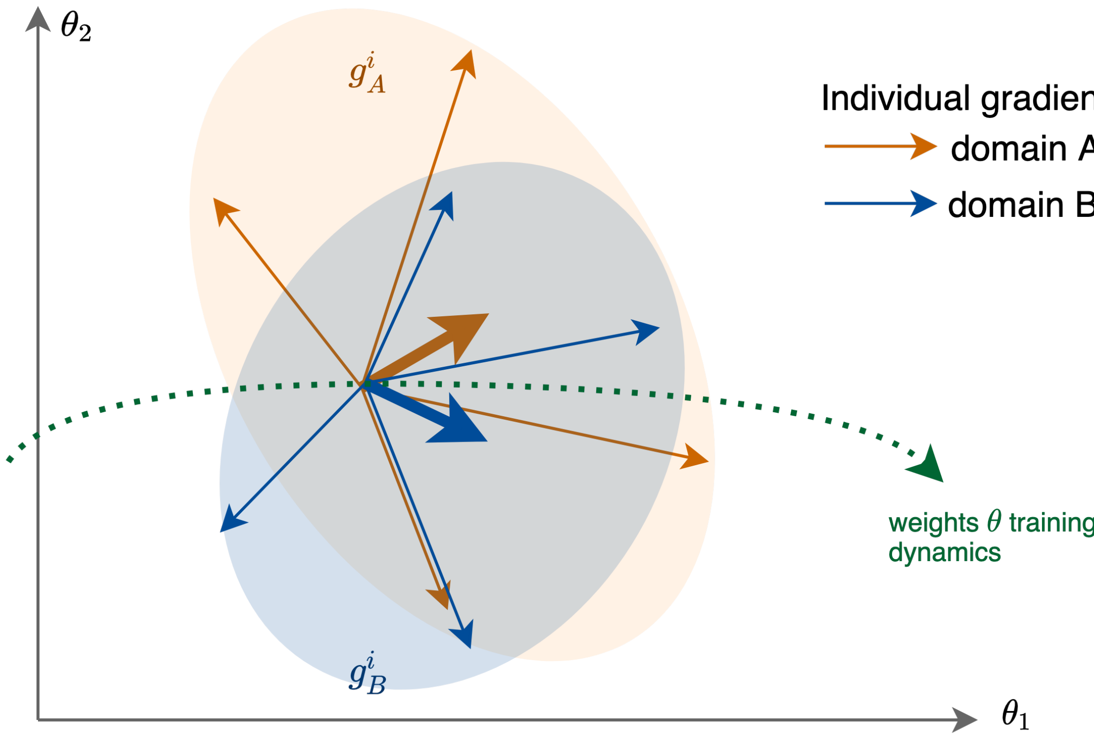
Matching domain-level gradient means

Regularization: $\| Mean(G_A) - Mean(G_B) \|_2^2$



[1] Out-of-distribution generalization with maximal invariant predictor. Koyama and Yamaguchi, 2020
[2] Fish: Gradient matching for domain generalization. Shi *et al.,* ICLR 2022

Gradient distributions richer than gradient means

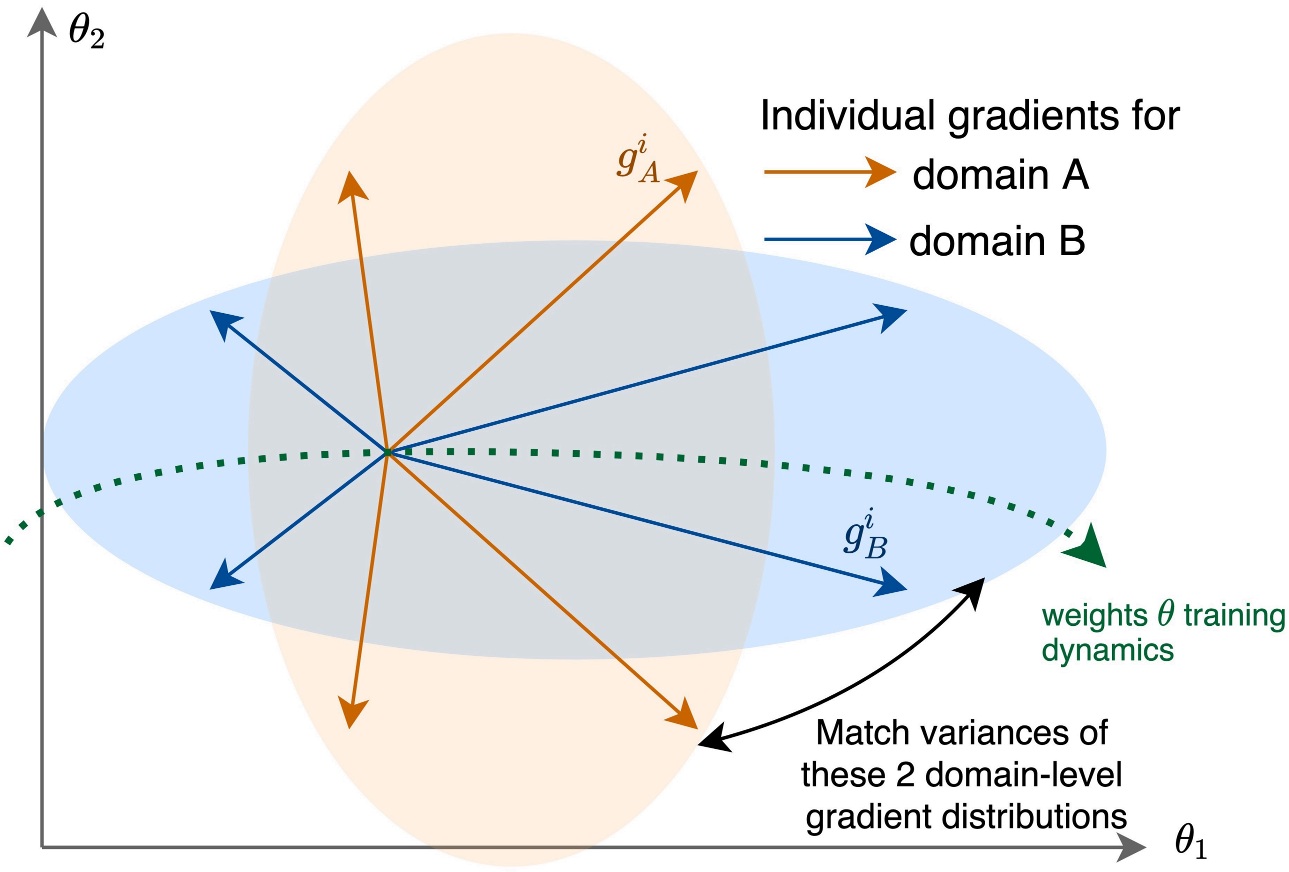


[1] Gradient diversity: a key ingredient for scalable distributed learning. Yin *et al.*, AISTATS 2018
[2] The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. Sankararaman *et al.*, ICML 2020

Fishr: invariant gradient variances

Regularization: $\| Var(G_A) - Var(G_B) \|_2^2$

where for $e \in \{A, B\}, G_e = \left[\nabla_\theta l\left(f_\theta\left(x_e^i\right), y_e^i\right)\right]_{i=1}^{n_e}$



Invariant gradients

Features

$f_\theta(x)$

Ex: Dann, CORAL

Predictors/ Losses

$l\left(f_\theta(x), y\right)$

Ex: IRM, V-REx

Gradients

$\nabla_\theta l\left(f_\theta(x), y\right)$

Direct objective

Fishr

Invariant gradients thus invariant losses ….

Features

$$f_\theta(x)$$

Ex: Dann, CORAL

Predictors/ Losses

$$l\,(f_\theta(x), y)$$

Ex: IRM, V-REx

Gradients

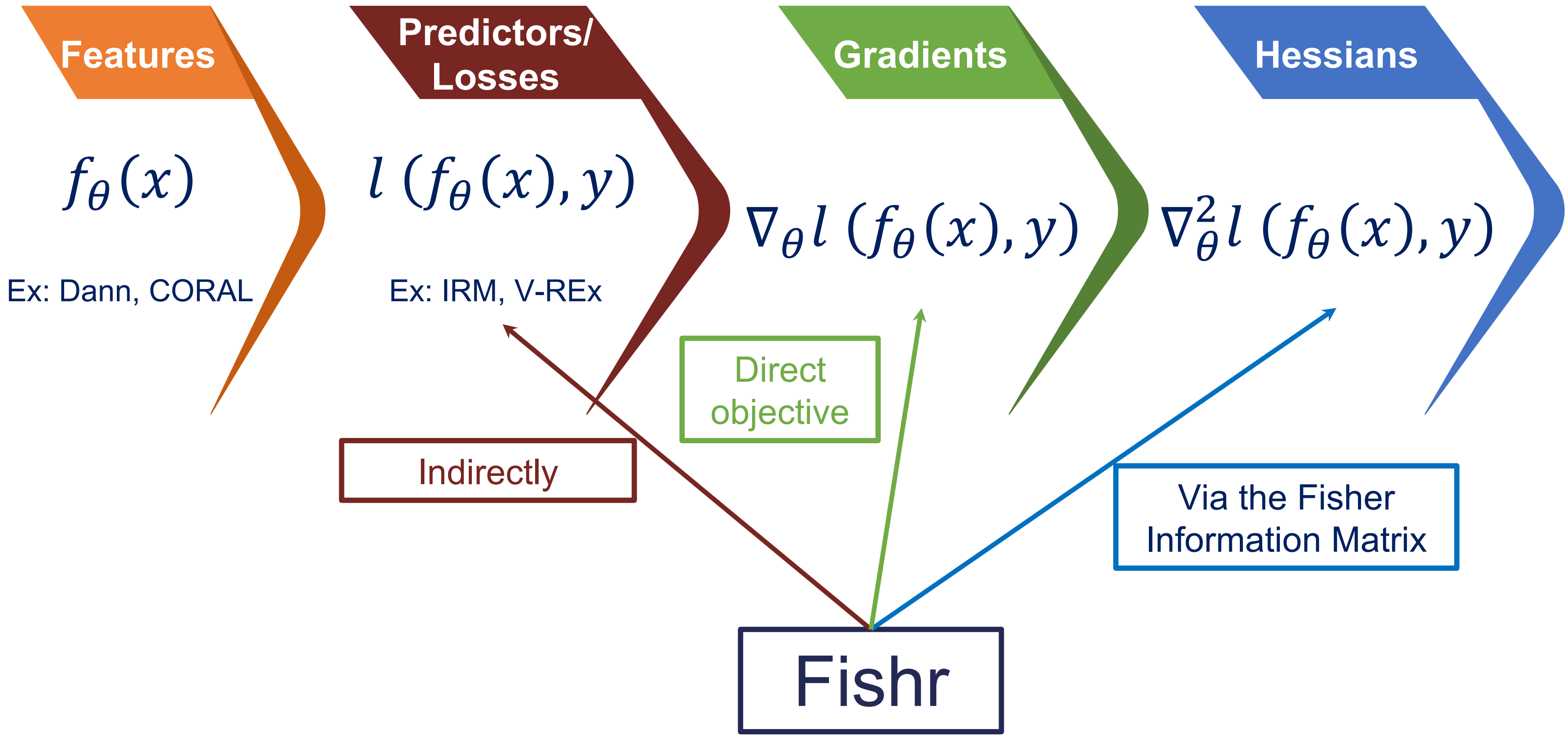
$$\nabla_\theta l\,(f_\theta(x), y)$$

Indirectly

Direct objective

Fishr

Invariant gradients thus invariant losses … and invariant Hessians!



Gradient covariance approximates the Hessian (via the Fisher Information Matrix)

Centered/uncentered

Empirically valid [1]

Theoretically valid [2]

$$C \underset{\approx}{\propto} \tilde{F} \underset{\approx}{\propto} F \underset{\approx}{\propto} H$$

Name	Statistics	Formula
C	Gradient Covariance	$Cov(G)$
$\tilde{F}$	Empirical Fisher Information Matrix	$\sum_{i=1}^{n} \nabla_\theta log\left(p_\theta\left(y^i\|x^i\right)\right) \nabla_\theta log\left(p_\theta\left(y^i\|x^i\right)\right)^T$
F	True Fisher Information Matrix	$\sum_{i=1}^{n} \mathbb{E}_{\hat{y}\sim P_\theta(\cdot\|x^i)}[\nabla_\theta log\left(p_\theta\left(\hat{y}\|x^i\right)\right) \nabla_\theta log\left(p_\theta\left(\hat{y}\|x^i\right)\right)^T]$
H	Hessian	$\sum_{i=1}^{n} \nabla_\theta^2 l\left(f_\theta\left(x^i\right), y^i\right)$

[1] On the interplay between noise and curvature and its effect on optimization and generalization. Thomas *et al.*, AISTATS 2020

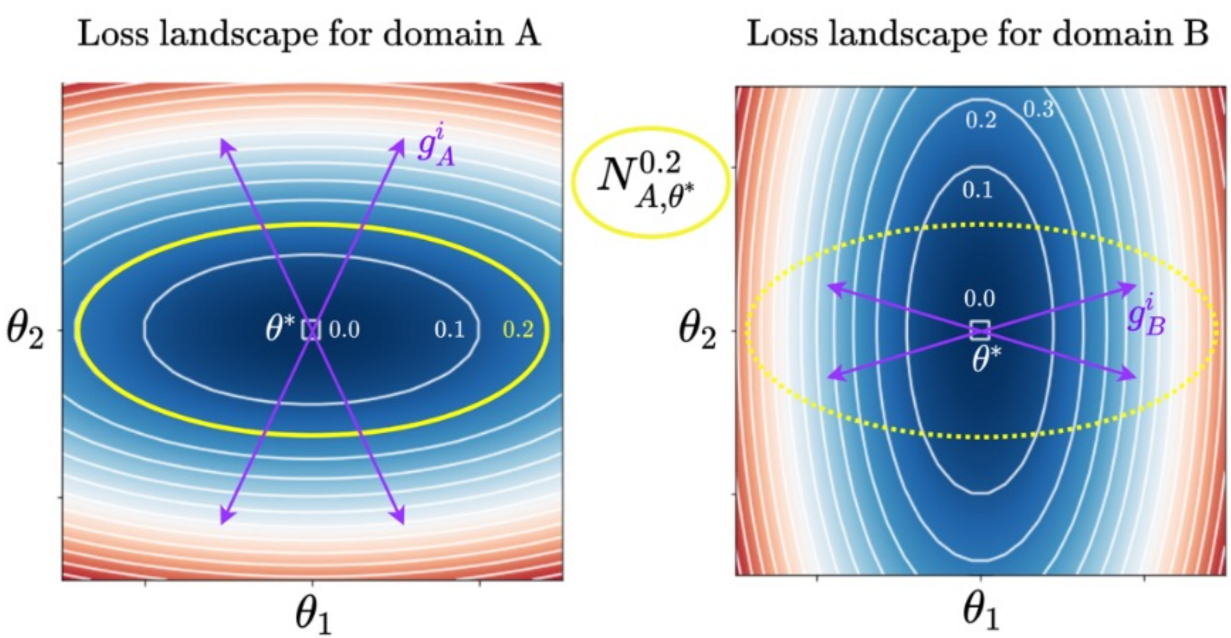
[2] New insights and perspectives on the natural gradient method. Martens, 2014

Fishr matches domain-level loss landscapes

With Fishr at convergence at θ^*, $R_A^{\theta^*} \approx R_B^{\theta^*}, G_A^{\theta^*} \approx G_B^{\theta^*}, H_A^{\theta^*} \approx H_B^{\theta^*}$

Via a 2nd-order Taylor expansion, ∀ weights θ close to θ^*:

$$R_A^{\theta} \approx R_A^{\theta^*} + (\theta - \theta^*)G_A^{\theta^*} + (\theta - \theta^*)H_A^{\theta^*}(\theta - \theta^*)$$
$$\approx R_B^{\theta^*} + (\theta - \theta^*)G_B^{\theta^*} + (\theta - \theta^*)H_B^{\theta^*}(\theta - \theta^*)$$
$$\approx R_B^{\theta}$$



[1] Learning explanations that are hard to vary. Parascandolo *et al.*, ICLR 2021

DomainBed

Reference benchmark for OOD generalization, imposing the *code*, *datasets*, *training procedures*, *hyperparameter search, model selection* etc.

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**

[1] In search of lost domain generalization. Gulrajani and Lopez-Paz, ICLR 2021

Fishr Contributions

- Theoretically
  - Invariant gradient variances …
  - but also invariant losses and Hessians to align landscapes
- Empirically
  - Simple and scalable
  - State of the Art on DomainBed for OOD generalization

Code available: https://github.com/alexrame/fishr

Merci !