Fishr:

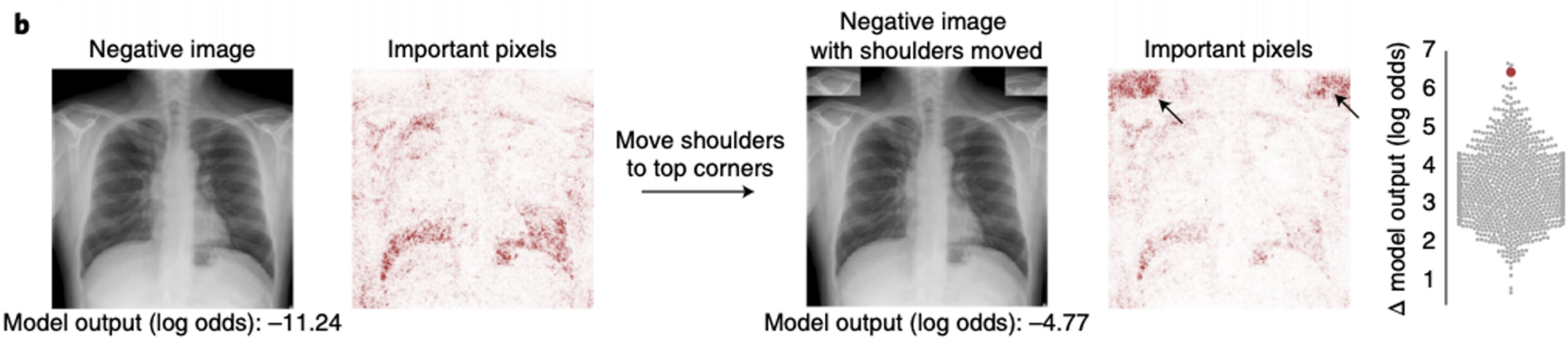
Invariant Gradients Variances for Out-of-distribution Generalization

Alexandre Ramé, Corentin Dancette and Matthieu Cord

Covid-19

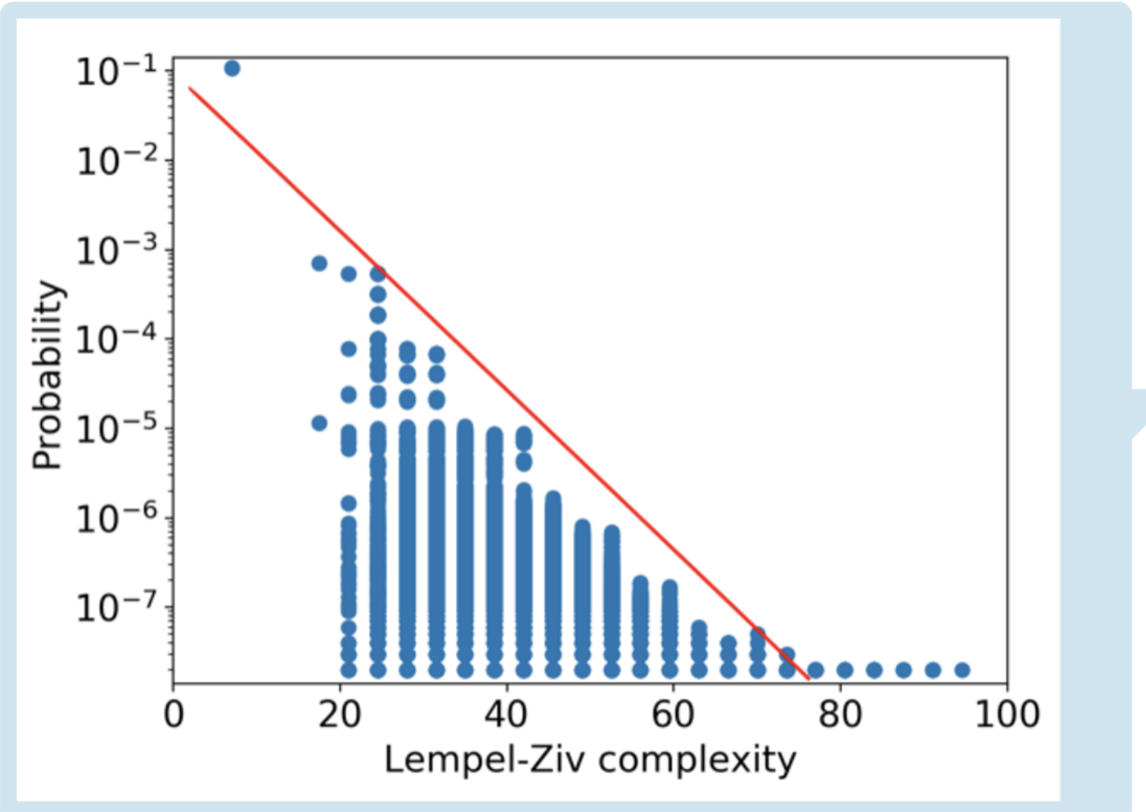
Deep networks to analyze chest scans. But bias such as:

- Detect position: shoulders, standing up etc
- Classify children vs. adults: data from external dataset
- Markers on the image

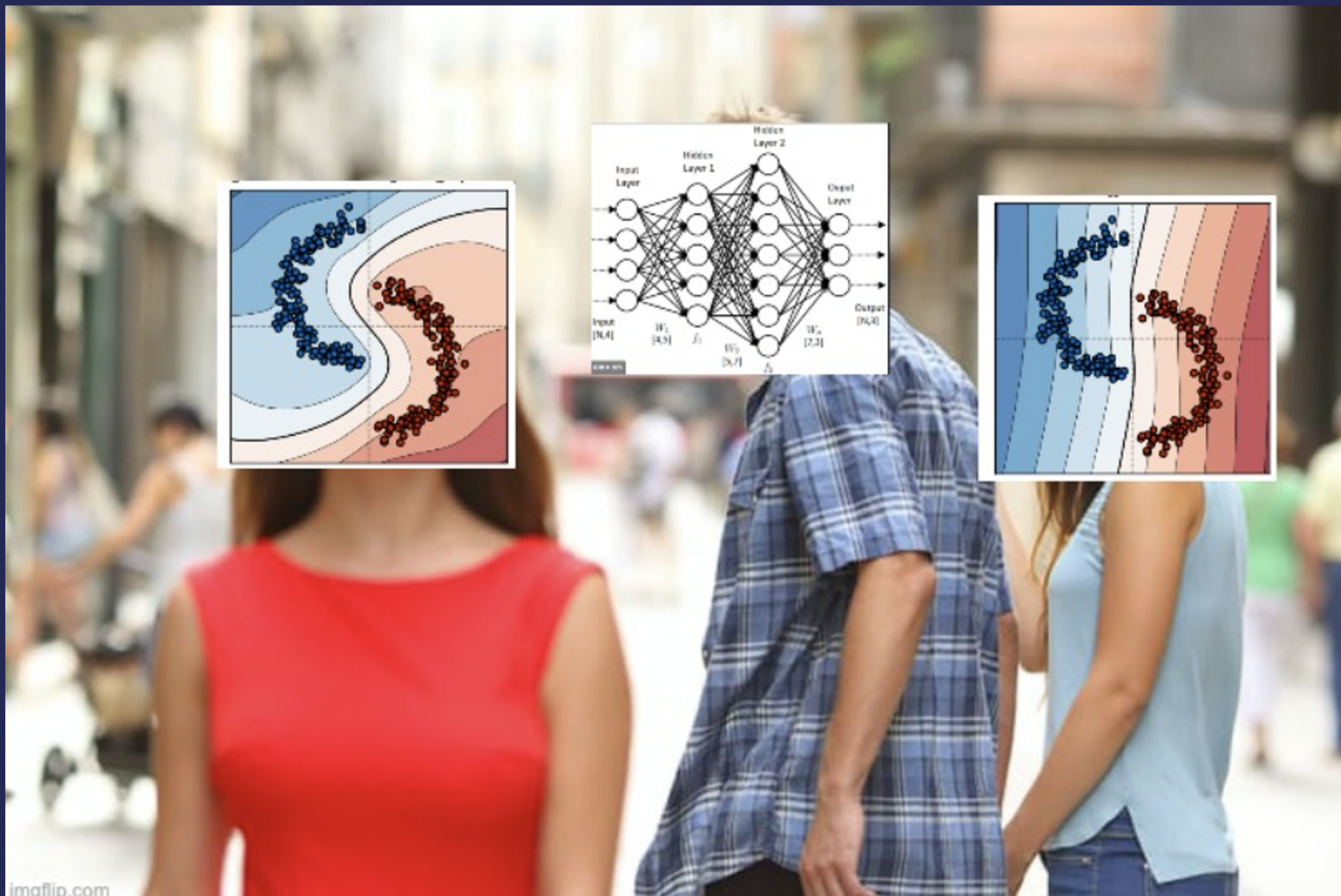


[1] : Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Roberts *et al.*, Nature 2021

NN biased towards simple functions



[1] Deep learning generalizes because the parameter-function map is biased towards simple functions. Valle-Perez *et al.*, ICLR 2019
[2] The Low-Rank Simplicity Bias in Deep Networks. Huh *et al.*, 2021


Input Layer
Hidden Layer 1
Hidden Layer 2
Output Layer
Input [N,4]
Output [N,3]
imgflip.com

Simplicity bias: pros & cons ?

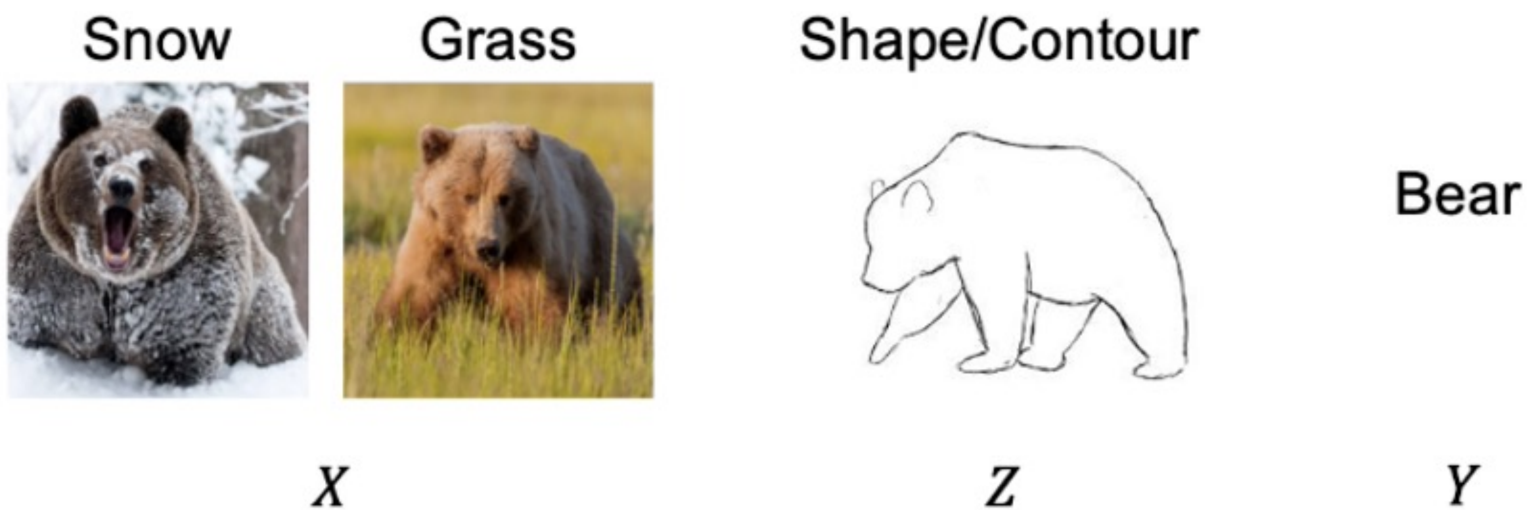
Pros: Occam's razor

1. Overfitting does not really hurt
2. Implicit regularization

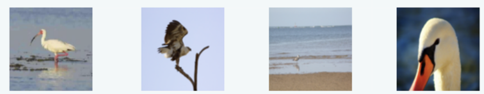
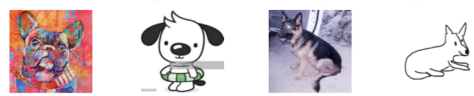
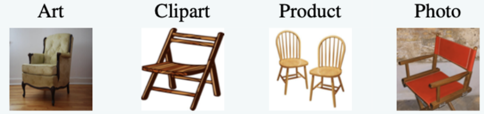
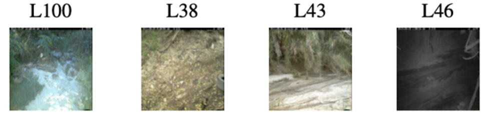
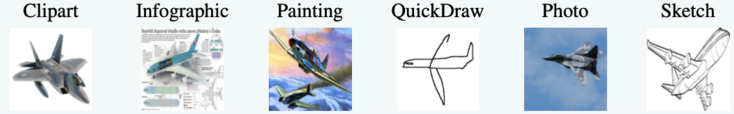
Cons: not robust

1. In distribution:
    - Can hurt accuracies ! rely on the simplest feature
    - Confidence estimates, calibration, ….
2. Out of Distribution generalization: correlation != causation

Learning invariant mechanism

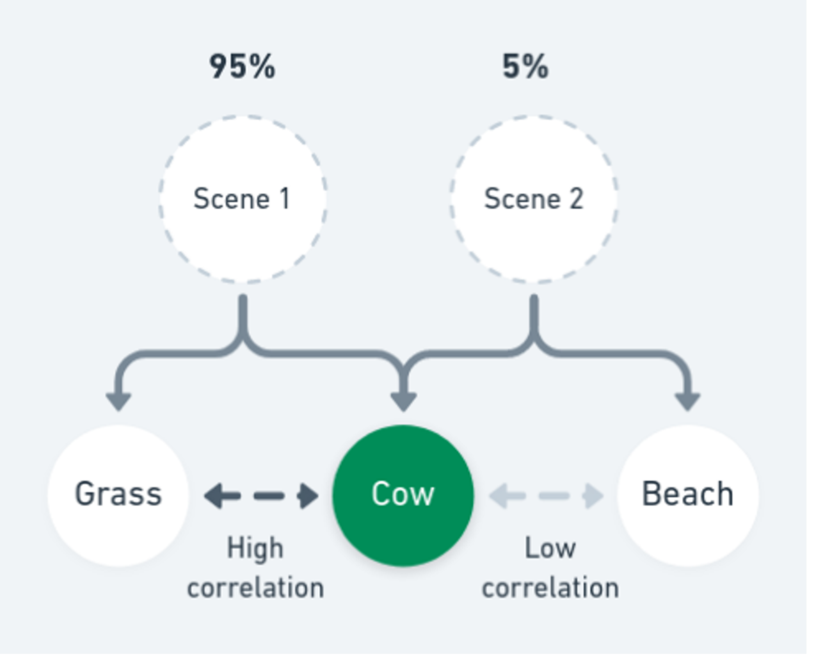


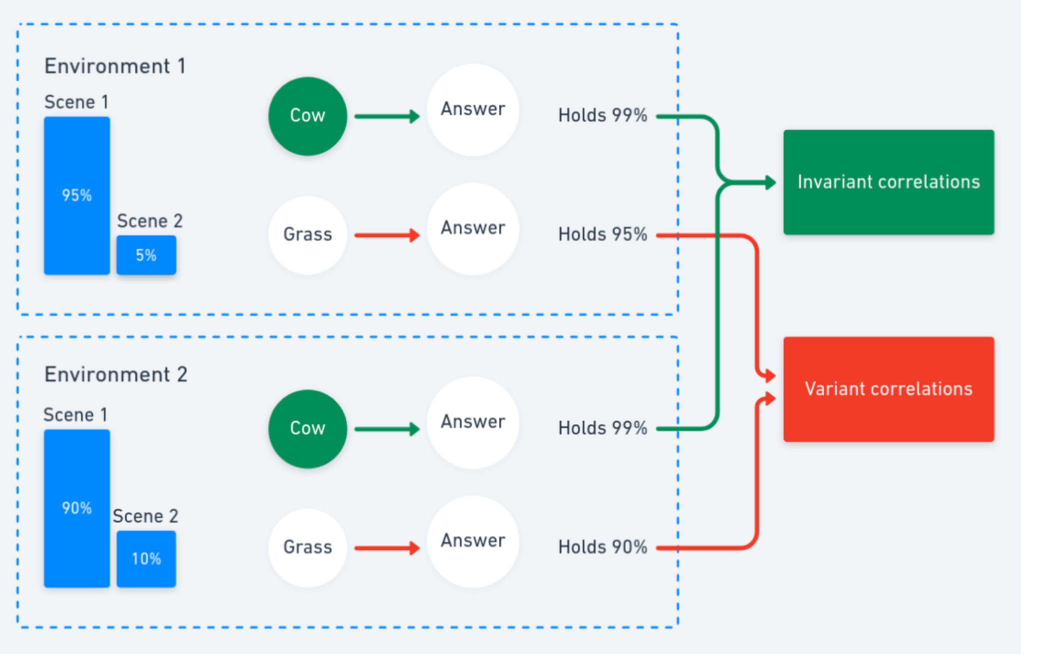
Assumption: multi domains

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 

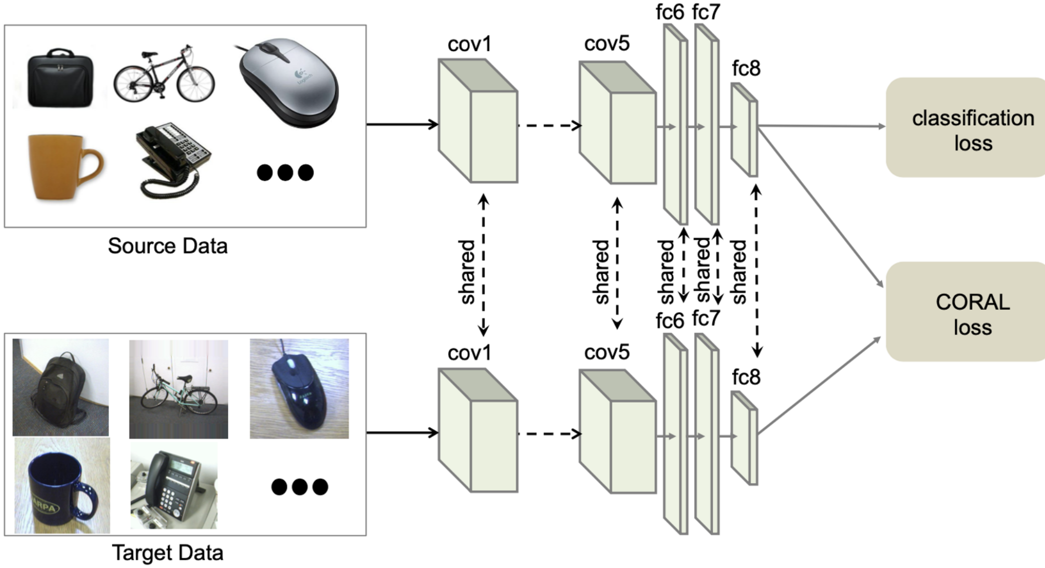
[1] In search of lost domain generalization. Gulrajani and Lopez-Paz, ICLR 2021

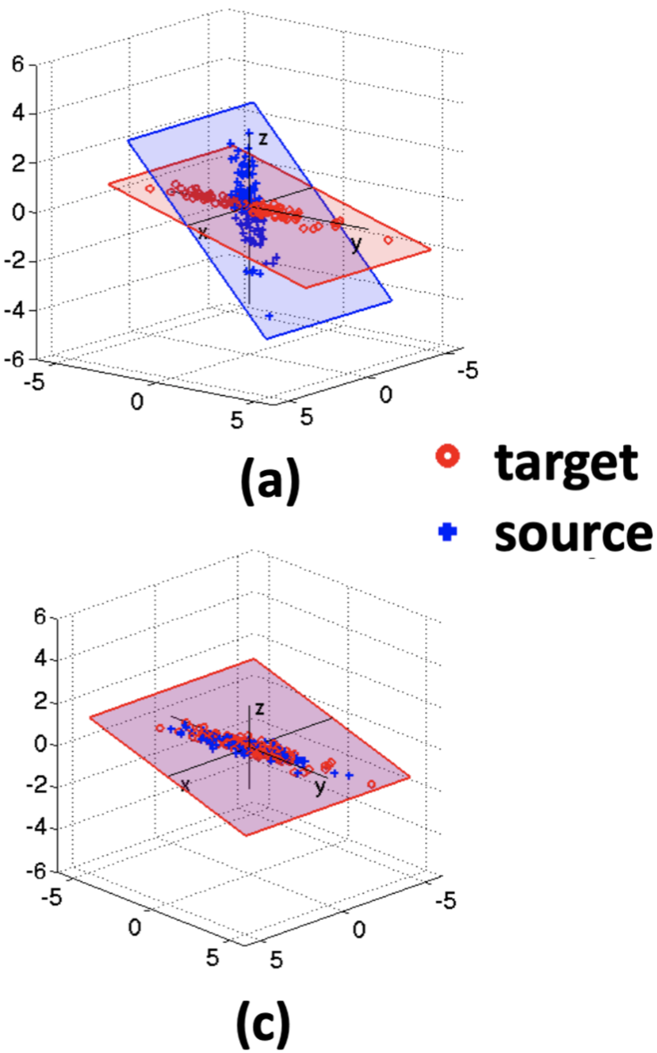
Invariant mechanism across domains





Invariant features





Coral regularizes:

$$\| Cov(Z_A) - Cov(Z_b) \|_F^2$$

[1] Deep coral: Correlation alignment for deep domain adaptation. Sun and Saenko, ECCV 2016

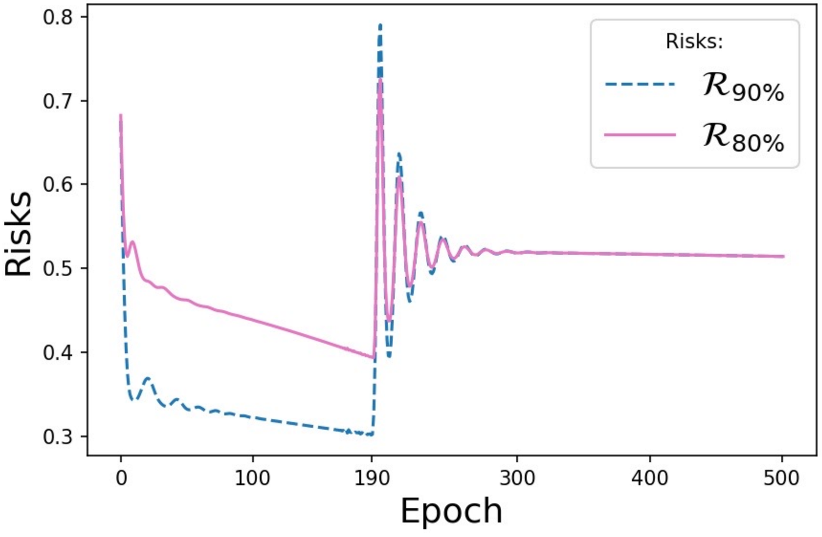
Invariant predictors

IRM regularizes:

$$\sum_{e \in \mathcal{E}} \| \nabla_{\omega|\omega=1.0} R_e(\omega \cdot \phi) \|^2$$

V-REx regularizes:

$$|R_A - R_B|^2$$



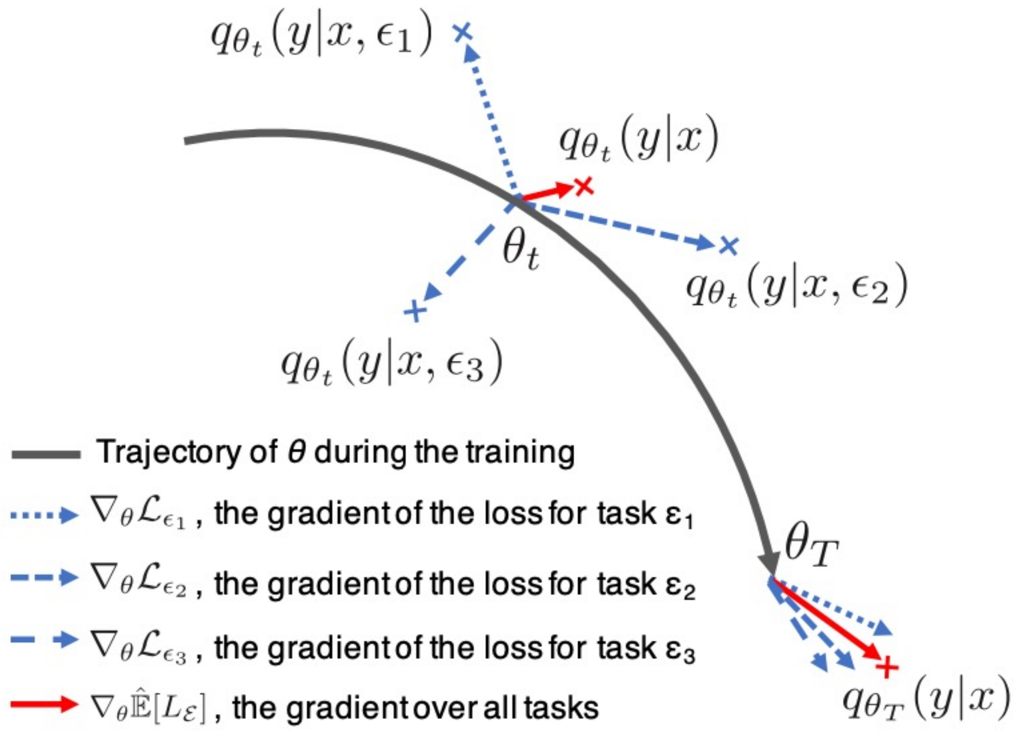
[1] Invariant risk minimization. Arjovsky *et al.*, 2019
[2] Out-of-distribution generalization via risk extrapolation. Krueger *et al.*, ICML 2021

Invariant gradients ?

Gradient mean matching

Individual gradients: $G_e = \left[\nabla_\theta l\left(f_\theta\left(x_e^i\right), y_e^i\right)\right]_{i=1}^{n_e}$

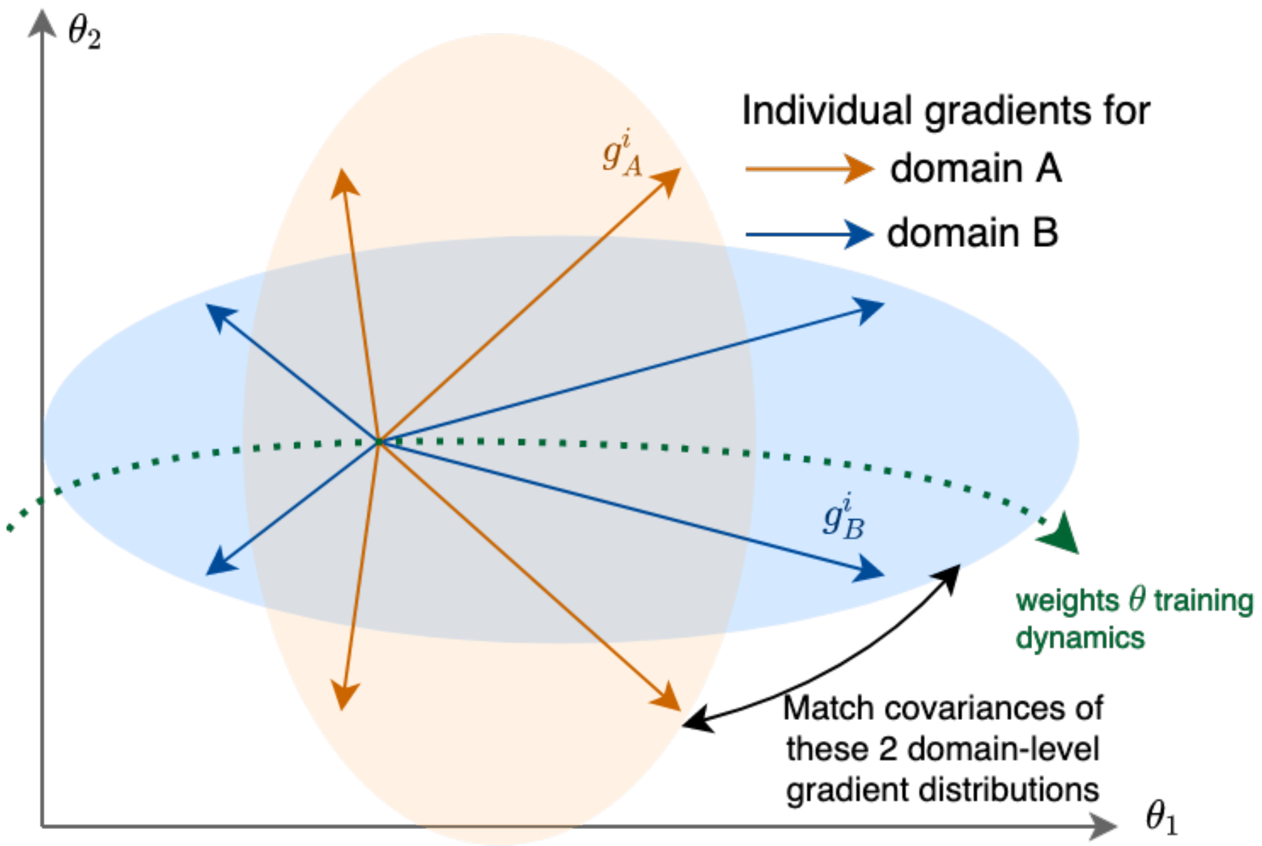
IGA regularizes: $\| \, Mean(G_A) - Mean(G_B) \, \|_2^2$



[1] Out-of-distribution generalization with maximal invariant predictor. Koyama and Yamaguchi, 2020

Fishr: Gradient covariance matching

Fishr regularizes: $\| Cov(G_A) - Cov(G_B) \|_F^2$



Why gradients ?

Matching gradient covariances enables to match gradient distributions. But why ? Because gradients:

1. Dictate the learning process
2. More expressive than features
3. Takes into account the label: class-conditional!
4. Are weighted by the loss values: indirectly align risks

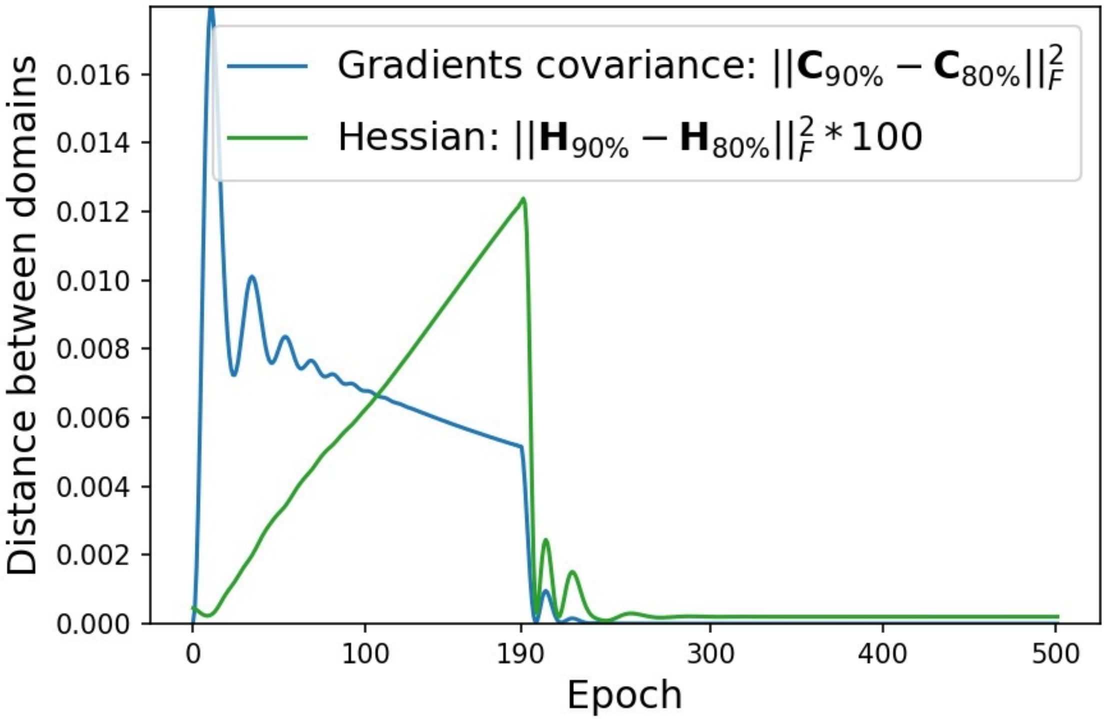
Hessian Motivations

Fishr matches domain-level Hessians

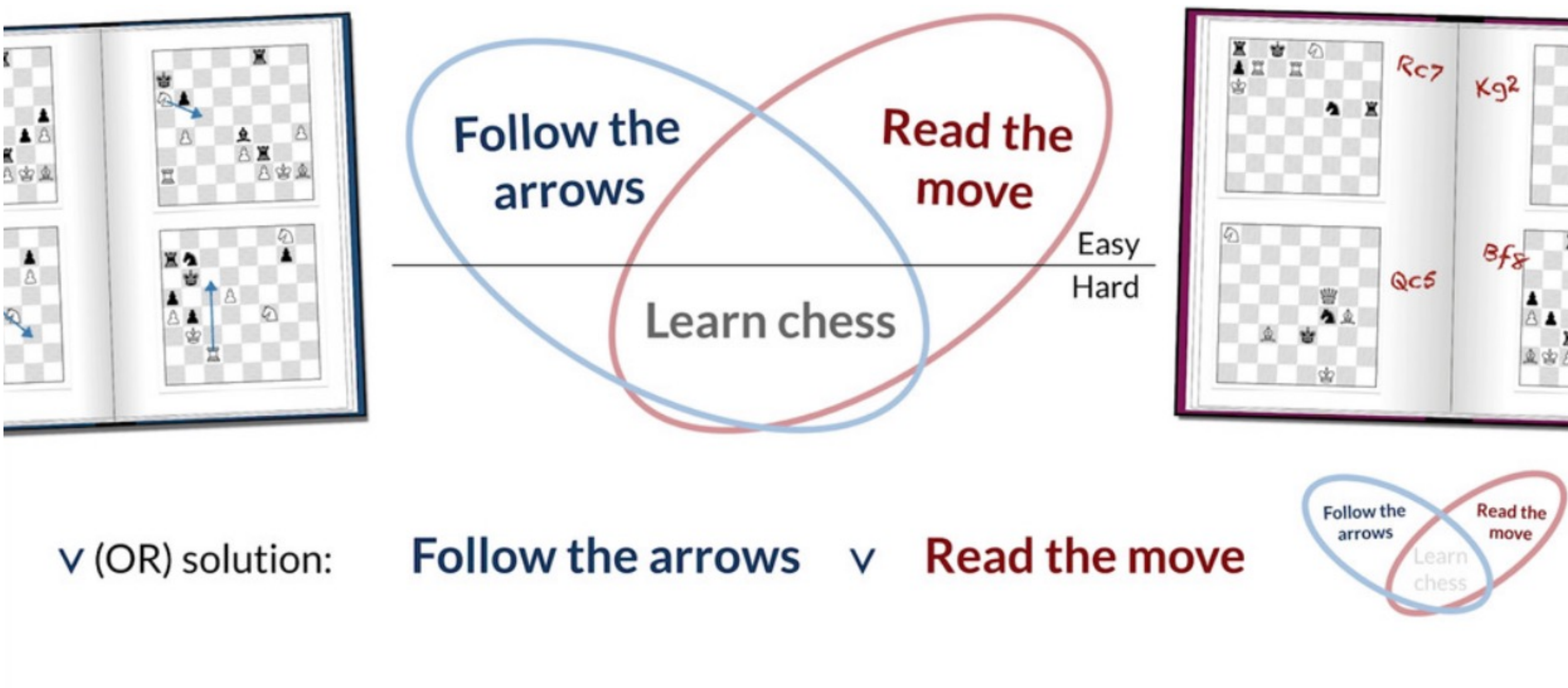
$$C \propto \tilde{F} \propto F \propto H$$

Where:

- C is the gradient covariance
- $\tilde{F}$ is the empirical Fisher Information Matrix
- F is the true Fisher Information Matrix
- $H = \sum_{i=1}^{n} \nabla_\theta^2 l\left(f_\theta\left(x^i\right), y^i\right)$ is the Hessian



Good explanations are hard to vary



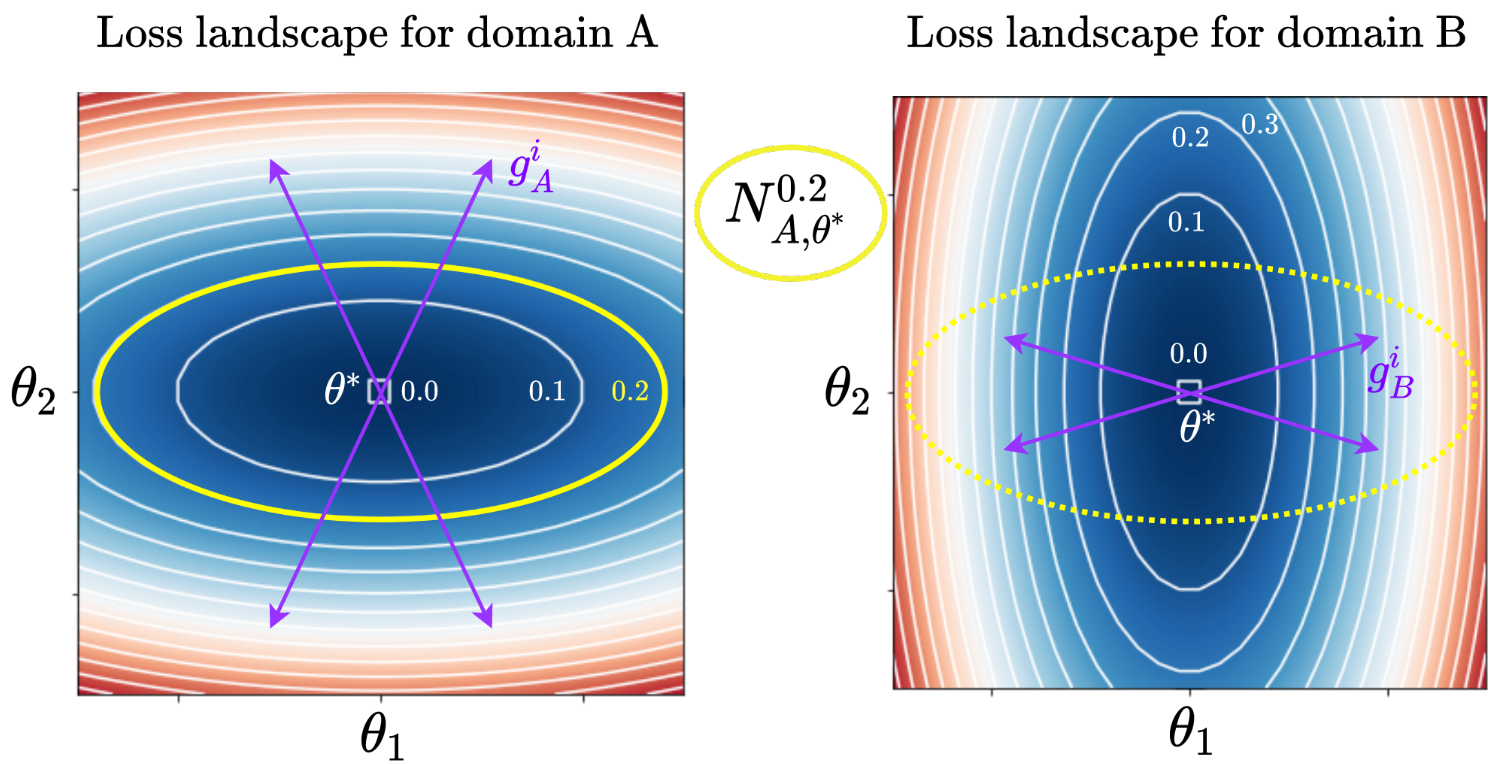
[1] Learning explanations that are hard to vary. Parascandolo *et al.*, ICLR 2021

Invariant Hessians for loss consistency

$$\mathcal{J}^{\epsilon} = max_{(A,B)\in\mathcal{E}^2} max_{\theta\in N^{\epsilon}_{A,\theta^*}} |R_B(\theta) - R_A(\theta^*)|$$

is minimal when:

$$H_A = H_B = \cdots$$



Neural Tangent Kernel intuition

$C \propto F$ sharing eigenvalues with K

Where:

- C is the gradient covariance
- F is the true Fisher Information Matrix
- K is the NTK matrix: $K[i,j] = \nabla_\theta f_\theta(x^i) \cdot \nabla_\theta f_\theta(x^j)$

Having similar spectral decompositions across $\{K\}_{e \in \mathcal{E}}$ would improve OOD generalization:

- Similar eigenvectors => same features across domains
- Similar eigenvalues => same learning speed

Scalable implementation

Approximations

1. Diagonal of the gradient covariance => Variance
2. Only in the classifier => ignore the features extractor weights

Thus the Fishr regularization ends up being:

$$\sum_{e \in \mathcal{E}} \sum_{\pi \in \omega} |v_e^{\pi} - v^{\pi}|^2$$

BackPACK package

BackPACK: Packing more into backprop

build passing | coverage 95% | python 3.6+

BackPACK is built on top of PyTorch. It efficiently computes quantities other than the gradient.

- **Website:** https://backpack.pt
- **Documentation:** https://docs.backpack.pt/en/master/
- **Bug reports & feature requests:** https://github.com/f-dangel/backpack/issues

Provided quantities include:

- Individual gradients from a mini-batch
- Estimates of the gradient variance or second moment
- Approximate second-order information (diagonal and Kronecker approximations)

Motivation: Computation of most quantities is not necessarily expensive (often just a small modification of the existing backward pass where backpropagated information can be reused). But it is difficult to do in the current software environment.

Experiments

Proof-of-concept on Colored MNIST

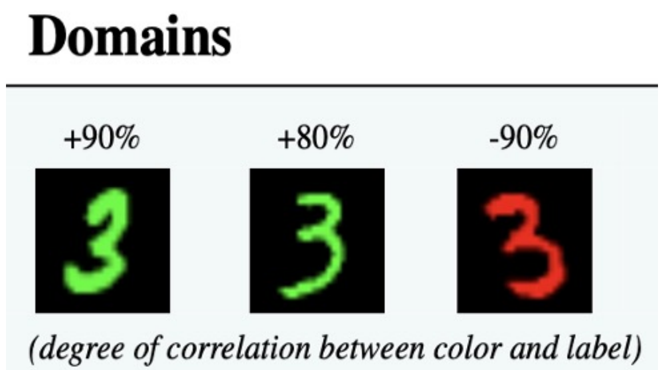


Table 2: **Colored MNIST** results. All methods use hyperparameters optimized for IRM.

Method	Train acc.	Test acc.	Gray test acc.
ERM	86.4 ± 0.2	14.0 ± 0.7	71.0 ± 0.7
IRM	71.0 ± 0.5	65.6 ± 1.8	66.1 ± 0.2
V-REx	71.7 ± 1.5	67.2 ± 1.5	68.6 ± 2.2
Fishr	71.0 ± 0.9	69.5 ± 1.0	70.2 ± 1.1

DomainBed 'Oracle'

Table 3: **Model selection: test-domain validation set (oracle).**

Algorithm	CMNIST	RMNIST	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Avg
ERM	57.8 ± 0.2	97.8 ± 0.1	77.6 ± 0.3	86.7 ± 0.3	66.4 ± 0.5	53.0 ± 0.3	41.3 ± 0.1	68.7
IRM	67.7 ± 1.2	97.5 ± 0.2	76.9 ± 0.6	84.5 ± 1.1	63.0 ± 2.7	50.5 ± 0.7	28.0 ± 5.1	66.9
GroupDRO	61.1 ± 0.9	97.9 ± 0.1	77.4 ± 0.5	87.1 ± 0.1	66.2 ± 0.6	52.4 ± 0.1	33.4 ± 0.3	67.9
Mixup	58.4 ± 0.2	98.0 ± 0.1	78.1 ± 0.3	86.8 ± 0.3	68.0 ± 0.2	**54.4** ± 0.3	39.6 ± 0.1	69.0
MLDG	58.2 ± 0.4	97.8 ± 0.1	77.5 ± 0.1	86.8 ± 0.4	66.6 ± 0.3	52.0 ± 0.1	41.6 ± 0.1	68.7
CORAL	58.6 ± 0.5	98.0 ± 0.0	77.7 ± 0.2	87.1 ± 0.5	**68.4** ± 0.2	52.8 ± 0.2	41.8 ± 0.1	69.2
MMD	63.3 ± 1.3	98.0 ± 0.1	77.9 ± 0.1	**87.2** ± 0.1	66.2 ± 0.3	52.0 ± 0.4	23.5 ± 9.4	66.9
DANN	57.0 ± 1.0	97.9 ± 0.1	79.7 ± 0.5	85.2 ± 0.2	65.3 ± 0.8	50.6 ± 0.4	38.3 ± 0.1	67.7
CDANN	59.5 ± 2.0	97.9 ± 0.0	**79.9** ± 0.2	85.8 ± 0.8	65.3 ± 0.5	50.8 ± 0.6	38.5 ± 0.2	68.2
MTL	57.6 ± 0.3	97.9 ± 0.1	77.7 ± 0.5	86.7 ± 0.2	66.5 ± 0.4	52.2 ± 0.4	40.8 ± 0.1	68.5
SagNet	58.2 ± 0.3	97.9 ± 0.0	77.6 ± 0.1	86.4 ± 0.4	67.5 ± 0.2	52.5 ± 0.4	40.8 ± 0.2	68.7
ARM	63.2 ± 0.7	**98.1** ± 0.1	77.8 ± 0.3	85.8 ± 0.2	64.8 ± 0.4	51.2 ± 0.5	36.0 ± 0.2	68.1
V-REx	67.0 ± 1.3	97.9 ± 0.1	78.1 ± 0.2	**87.2** ± 0.6	65.7 ± 0.3	51.4 ± 0.5	30.1 ± 3.7	68.2
RSC	58.5 ± 0.5	97.6 ± 0.1	77.8 ± 0.6	86.2 ± 0.5	66.5 ± 0.6	52.1 ± 0.2	38.9 ± 0.6	68.2
AND-mask	58.6 ± 0.4	97.5 ± 0.0	76.4 ± 0.4	86.4 ± 0.4	66.1 ± 0.2	49.8 ± 0.4	37.9 ± 0.6	67.5
SAND-mask	62.3 ± 1.0	97.4 ± 0.1	76.2 ± 0.5	85.9 ± 0.4	65.9 ± 0.5	50.2 ± 0.1	32.2 ± 0.6	67.2
Fish	61.8 ± 0.8	97.9 ± 0.1	77.8 ± 0.6	85.8 ± 0.6	66.0 ± 2.9	50.8 ± 0.4	**43.4** ± 0.3	69.1
Fishr	**68.8** ± 1.4	97.8 ± 0.1	78.2 ± 0.2	86.9 ± 0.2	68.2 ± 0.2	53.6 ± 0.4	41.8 ± 0.2	**70.8**

DomainBed ‘Training’

Table 4: **Model selection: training-domain validation set.**

Algorithm	CMNIST	RMNIST	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Avg
ERM	51.5 ± 0.1	98.0 ± 0.0	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.1	66.6
IRM	52.0 ± 0.1	97.7 ± 0.1	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 2.8	65.4
GroupDRO	52.1 ± 0.0	98.0 ± 0.0	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	43.2 ± 1.1	33.3 ± 0.2	64.8
Mixup	52.1 ± 0.2	98.0 ± 0.1	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	39.2 ± 0.1	66.7
MLDG	51.5 ± 0.1	97.9 ± 0.0	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	47.7 ± 0.9	41.2 ± 0.1	66.7
CORAL	51.5 ± 0.1	98.0 ± 0.1	$\mathbf{78.8} \pm 0.6$	86.2 ± 0.3	$\mathbf{68.7} \pm 0.3$	47.6 ± 1.0	41.5 ± 0.1	**67.5**
MMD	51.5 ± 0.2	97.9 ± 0.0	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	42.2 ± 1.6	23.4 ± 9.5	63.3
DANN	51.5 ± 0.3	97.8 ± 0.1	78.6 ± 0.4	83.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	38.3 ± 0.1	66.1
CDANN	51.7 ± 0.1	97.9 ± 0.1	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	45.8 ± 1.6	38.3 ± 0.3	65.6
MTL	51.4 ± 0.1	97.9 ± 0.0	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	45.6 ± 1.2	40.6 ± 0.1	66.2
SagNet	51.7 ± 0.0	98.0 ± 0.0	77.8 ± 0.5	$\mathbf{86.3} \pm 0.2$	68.1 ± 0.1	$\mathbf{48.6} \pm 1.0$	40.3 ± 0.1	67.2
ARM	$\mathbf{56.2} \pm 0.2$	$\mathbf{98.2} \pm 0.1$	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	45.5 ± 0.3	35.5 ± 0.2	66.1
V-REx	51.8 ± 0.1	97.9 ± 0.1	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	46.4 ± 0.6	33.6 ± 2.9	65.6
RSC	51.7 ± 0.2	97.6 ± 0.1	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	46.6 ± 1.0	38.9 ± 0.5	66.1
AND-mask	51.3 ± 0.2	97.6 ± 0.1	78.1 ± 0.9	84.4 ± 0.9	65.6 ± 0.4	44.6 ± 0.3	37.2 ± 0.6	65.5
SAND-mask	51.8 ± 0.2	97.4 ± 0.1	77.4 ± 0.2	84.6 ± 0.9	65.8 ± 0.4	42.9 ± 1.7	32.1 ± 0.6	64.6
Fish	51.6 ± 0.1	98.0 ± 0.0	77.8 ± 0.3	85.5 ± 0.3	68.6 ± 0.4	45.1 ± 1.3	$\mathbf{42.7} \pm 0.2$	67.1
Fishr	52.0 ± 0.2	97.8 ± 0.0	77.8 ± 0.1	85.5 ± 0.4	67.8 ± 0.1	47.4 ± 1.6	41.7 ± 0.0	67.1

Hyperparameters

regularization strength λ	1000	$10^{\text{Uniform}(1,4)}$
ema γ	0.95	Uniform(0.9, 0.99)
warmup iterations	1500	Uniform(0, 5000)

Table 13: **Impact of the λ distribution** from Table 7.

Model selection	λ distribution	CMNIST	RMNIST	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Avg
Oracle	Constant(0) (= ERM)	57.8 ± 0.2	97.8 ± 0.1	77.6 ± 0.3	86.7 ± 0.3	66.4 ± 0.5	53.0 ± 0.3	41.3 ± 0.1	68.7
	$10^{\text{Uniform}(1,4)}$	**68.8** ± 1.4	97.8 ± 0.1	78.2 ± 0.2	86.9 ± 0.2	**68.2** ± 0.2	**53.6** ± 0.4	41.8 ± 0.1	**70.8**
	$10^{\text{Uniform}(1,5)}$	68.7 ± 1.3	97.8 ± 0.0	**78.7** ± 0.3	**87.5** ± 0.1	68.0 ± 0.4	52.2 ± 0.5	**42.0** ± 0.1	70.7
Training	Constant(0) (= ERM)	51.5 ± 0.1	**98.0** ± 0.0	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.1	66.6
	$10^{\text{Uniform}(1,4)}$	**52.0** ± 0.2	97.8 ± 0.0	77.8 ± 0.1	85.5 ± 0.4	**67.8** ± 0.1	**47.4** ± 1.6	41.7 ± 0.0	**67.1**
	$10^{\text{Uniform}(1,5)}$	51.8 ± 0.3	97.9 ± 0.0	**77.9** ± 0.1	85.5 ± 0.6	67.4 ± 0.3	47.2 ± 1.0	**41.8** ± 0.1	**67.1**

Contributions

❖ Theoretically

Invariant gradients criterion

Gradient covariance, Fisher information matrix, Hessian, loss landscapes etc...

❖ Empirically

State of the art on DomainBed

Simple and scalable strategy

https://github.com/alexrame/fishr

Merci !

Finding lost DG: explaining domain generalization via model complexity

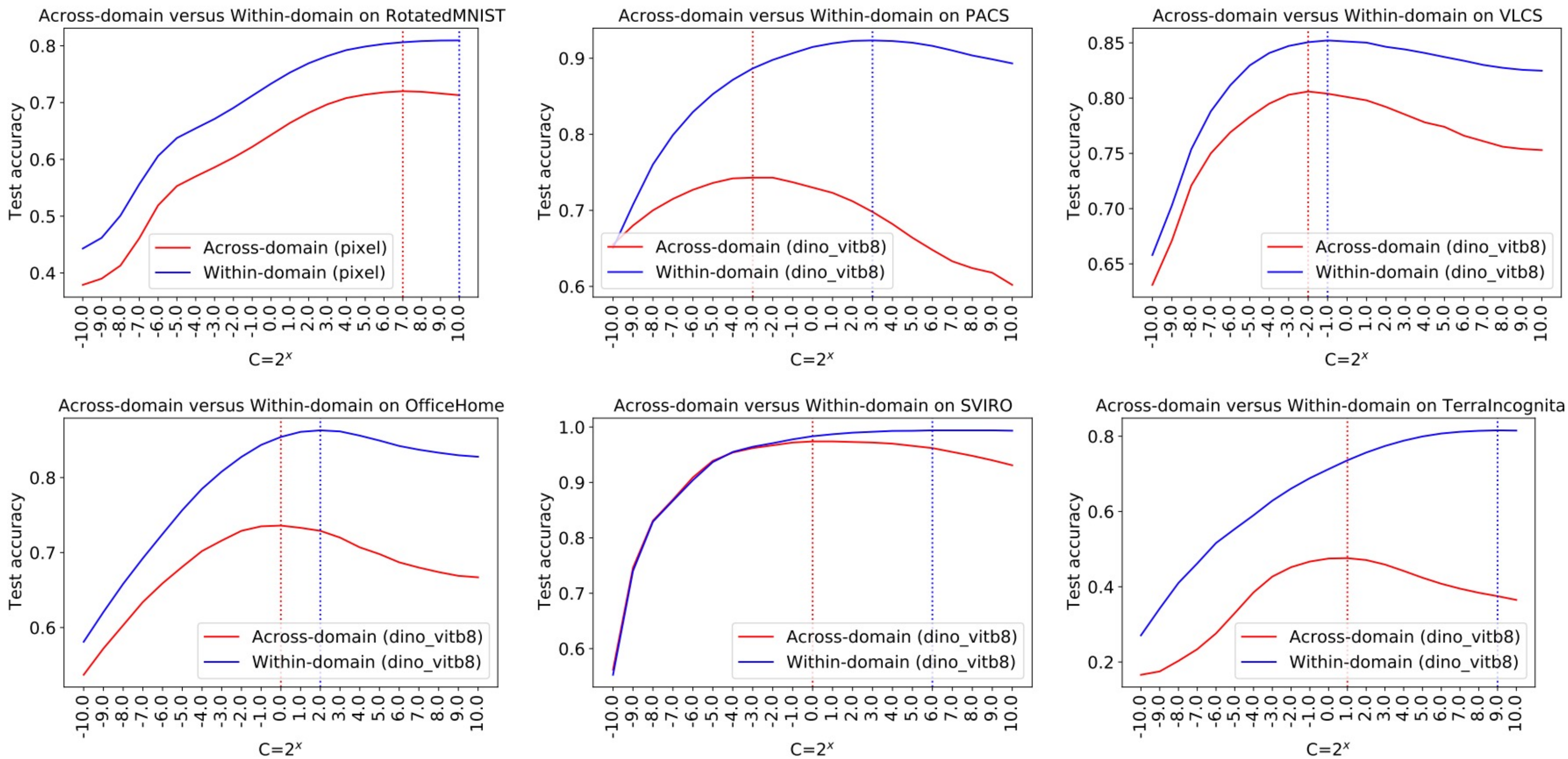


Figure 1: Linear SVC performance on DomainBed benchmark datasets is governed by model complexity parameter C. Optimal tuning for performance on novel target domains (DG condition, red) always requires stronger regularization (lower C) than for performance on seen domains (blue).