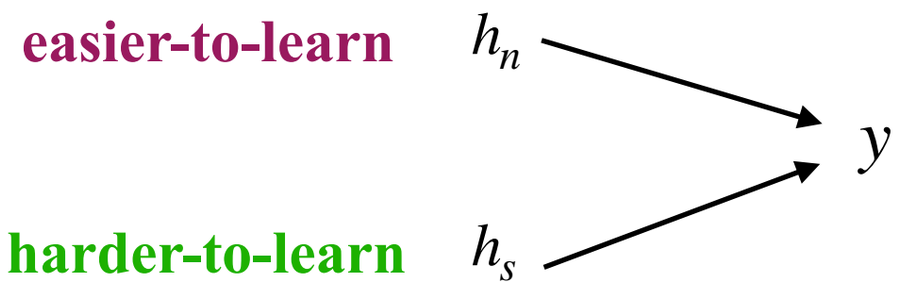
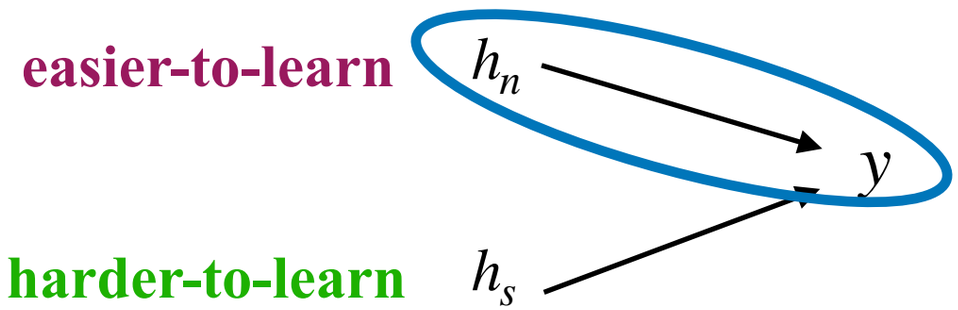
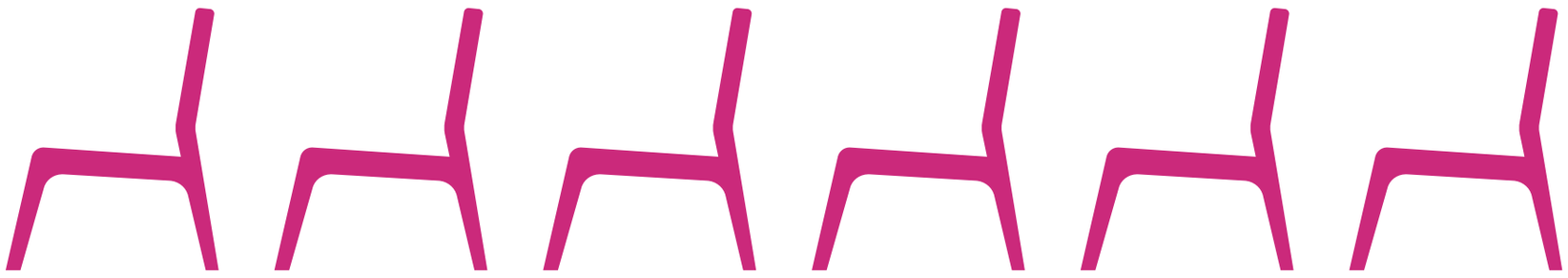
Systematic generalisation with group invariant predictions

Faruk Ahmed[1]*, **Yoshua Bengio**[1,2], **Harm van Seijen**[3], **Aaron Courville**[1,2]
[1] Université de Montréal, Mila, [2]CIFAR Fellow, [3]Microsoft Research
*Correspondence to faruk.ahmed@umontreal.ca

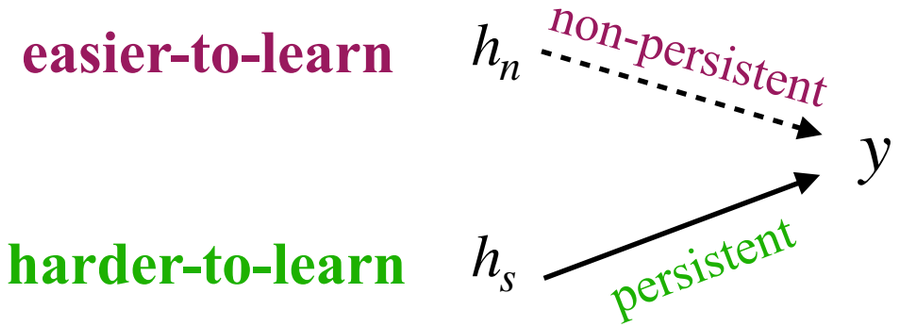

easier-to-learn
h_n
harder-to-learn
h_s
y

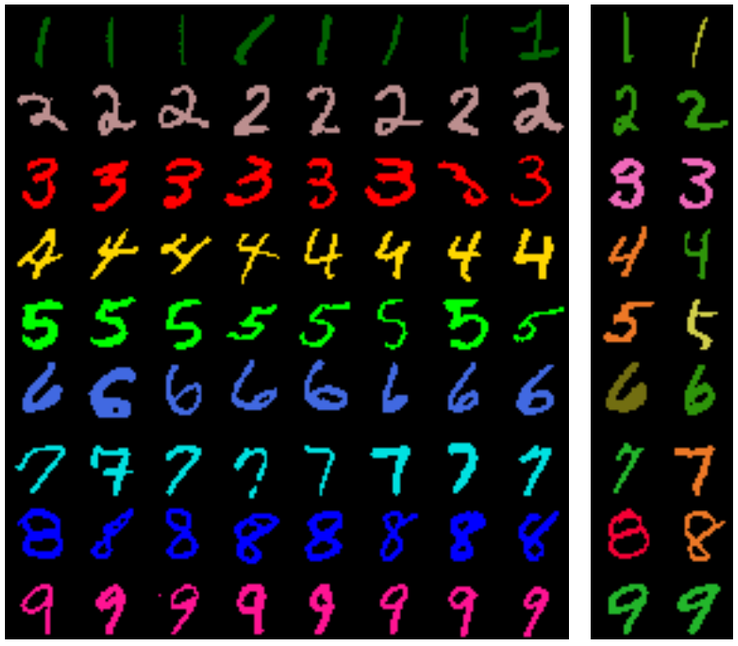


Occam's razor

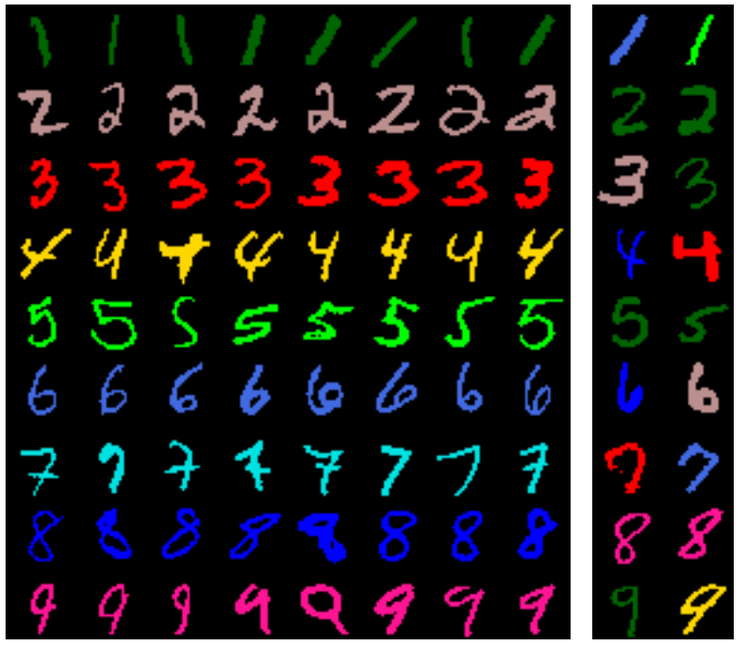


Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. CoRR, 2020.


easier-to-learn
h_n
non-persistent
y
harder-to-learn
h_s
persistent



Minority group colours are different



Minority group colours are recombinations

Systematically shifted test set

Minority colours	In-distribution	Systematic shift
Different	99.60 ± 0.02	38.72 ± 2.27
Recombinations	98.67 ± 0.39	97.56 ± 0.05

Assume data x is synthesised from *non-semantic factors* h_n and *semantic factors* h_s

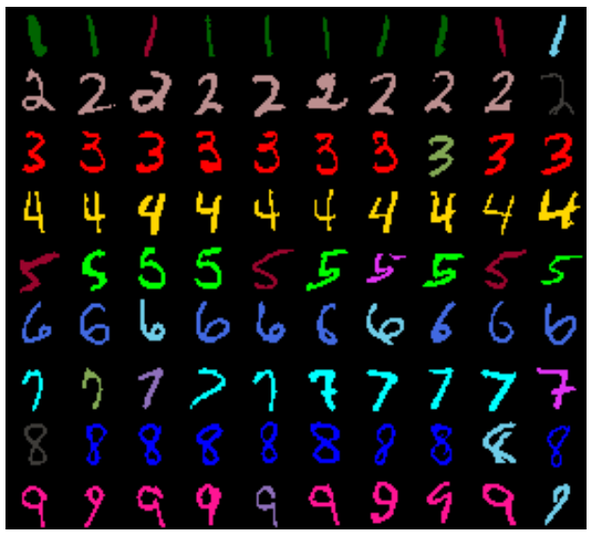
$$x = \mathscr{C}(h_n, h_s)\,.$$

If $\hat{y}(x)$ is our prediction, we can compute average accuracy, for synthesised x

$$\mathbb{E}\left[\mathbf{1}\{\hat{y}(\mathscr{C}(h_s, h_n)) = y\}\right]$$

with different sampling choices of h_s and h_n.

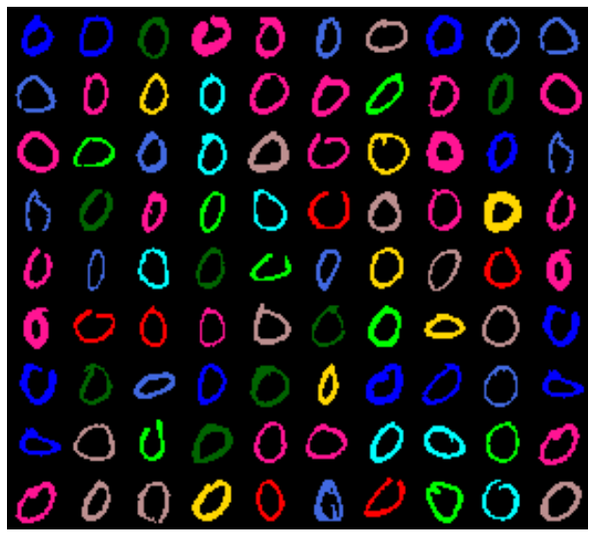
- *In-distribution generalisation:* $h_s \sim p(h_s \mid y), h_n \sim p(h_n \mid y)$
- *Generalisation under systematic-shift:* $h_s \sim p(h_s \mid y), h_n \sim p(h_n \mid y')$ where $y' \sim p(y)\ s.t.\ y' \neq y$
- *Generalisation under non-systematic-shift:* $h_s \sim p(h_s \mid y), h_n \nsim p(h_n)$
- *Semantic anomaly detection:* $h_s \nsim p(h_s), h_n \sim p(h_n)$



In-distribution

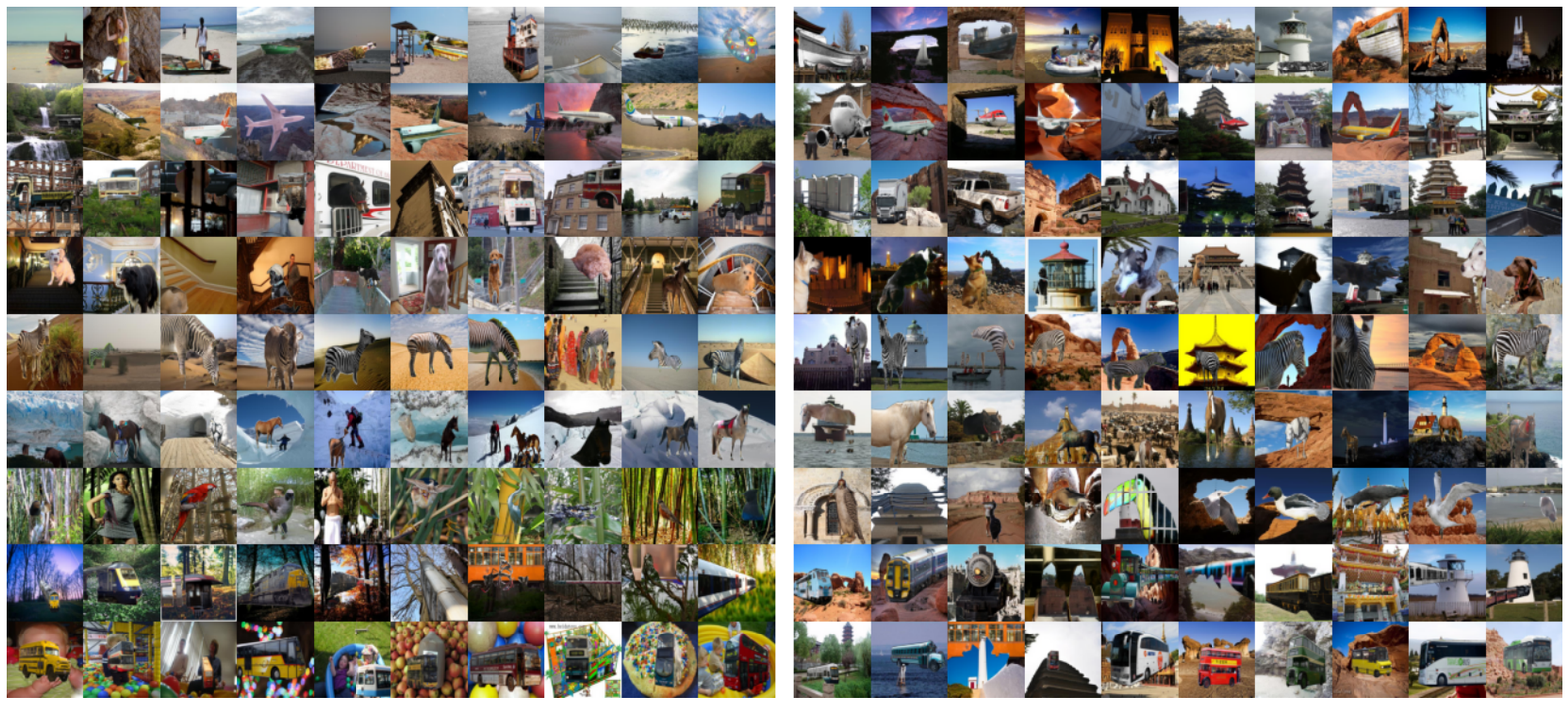
Systematic shift

Non-systematic shift



Semantic anomalies

COCO-on-Colours



COCO-on-Places

Do invariance methods/penalties for multi-group data help at such shifts?

Feature-distribution matching
Group-distributionally robust optimisation
Invariant Risk Minimisation
Others

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, 2019.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *CoRR*, 2020.

H. Li, S. J. Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. pp. 5400–5409, 2018.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ICLR*, 2020.

Predictive Group Invariance (PGI)

Feature extractor $f_\theta(x)$, predictive distribution:

$$p_w(y \mid x) = \sigma(w^\top f_\theta(x)) \,.$$

Split data into the biased majority ($\alpha = 0$) and unbiased minority group ($\alpha = 1$), for every class c, such that

$$x^{(i)} \sim \mathbb{P}^c \text{ if } \alpha^{(i)} = 0, y^{(i)} = c,$$
$$x^{(j)} \sim \mathbb{Q}^c \text{ if } \alpha^{(j)} = 1, y^{(j)} = c \,.$$

Loss function:

$$\ell(\theta, w \mid \mathcal{D}) + \lambda \left[\sum_c d\Big(\mathbb{E}_{x\sim\mathbb{Q}^c}[p_{\tilde{w}}(y \mid x)], \mathbb{E}_{x\sim\mathbb{P}^c}[p_{\tilde{w}}(y \mid x)] \Big) \right]_{\tilde{w}=w \text{ (fixed)}}$$

ℓ = standard (regularised) ERM term

d = KL-divergence

Predictive Group Invariance (PGI)

Feature extractor $f_\theta(x)$, predictive distribution:

$$p_w(y \mid x) = \sigma(w^\top f_\theta(x)) \,.$$

Split data :

"Environment Inference for Invariant Learning"

COLOURED MNIST	COCO-ON-COLOURS	COCO-ON-PLACES
97.26 ± 0.71	98.22 ± 1.05	80.43 ± 1.41

Partitioning accuracy

Loss function:

$$\ell(\theta, w \mid \mathscr{D}) + \lambda \left[\sum_c d\Big(\mathbb{E}_{x\sim\mathbb{Q}^c}[p_{\tilde{w}}(y \mid x)], \mathbb{E}_{x\sim\mathbb{P}^c}[p_{\tilde{w}}(y \mid x)] \Big) \right]_{\tilde{w}=w \text{ (fixed)}}$$

ℓ = standard (regularised) ERM term

d = KL-divergence

Elliot Creager, Jörn-Henrik Jacobson, and Richard Zemel. Environment inference for invariant learning. *ICML Workshop on Uncertainty and Robustness*, 2020.

Table 2: Generalisation results on COLOURED MNIST.

Methods	In-distribution	Non-systematic shift	Systematic shift	Anomaly detection
Base (ERM)	**99.60 ± 0.02**	53.26 ± 1.89	38.72 ± 2.27	7.70 ± 0.23
IRMv1	99.47 ± 0.05	63.24 ± 3.04	55.19 ± 1.07	11.54 ± 1.18
REx	98.95 ± 0.11	72.12 ± 1.90	71.18 ± 3.27	15.54 ± 2.05
GroupDRO	89.47 ± 4.52	70.53 ± 1.79	79.17 ± 1.64	35.15 ± 10.83
Reweight	98.51 ± 0.12	75.01 ± 1.28	84.85 ± 0.61	28.60 ± 1.11
cIRMv1	99.36 ± 0.25	65.78 ± 3.53	61.09 ± 5.30	14.16 ± 2.12
cREx	98.56 ± 0.12	74.35 ± 2.09	80.01 ± 2.11	22.02 ± 2.52
cGroupDRO	95.65 ± 3.23	75.41 ± 3.45	81.14 ± 2.41	26.61 ± 6.61
cMMD	99.40 ± 0.03	97.17 ± 0.59	97.86 ± 0.16	78.32 ± 4.15
PGI	99.05 ± 0.08	**98.58 ± 0.06**	**98.48 ± 0.05**	**89.42 ± 1.95**

Table 3: Generalisation performance on COCO-ON-COLOURS.

Methods	In-distribution	Non-systematic shift	Systematic shift	Anomaly detection
Base (ERM)	90.57 ± 1.28	26.81 ± 4.93	1.10 ± 0.36	5.47 ± 0.08
IRMv1	91.61 ± 0.38	32.30 ± 4.52	2.11 ± 0.30	5.81 ± 0.17
REx	**91.69 ± 0.50**	36.57 ± 4.03	2.69 ± 0.81	5.73 ± 0.14
GroupDRO	43.06 ± 2.26	41.32 ± 4.39	43.24 ± 2.89	20.05 ± 3.08
Reweight	42.42 ± 3.47	47.56 ± 2.27	49.12 ± 1.63	18.15 ± 3.81
cIRMv1	91.53 ± 0.31	31.11 ± 4.51	1.74 ± 0.40	5.87 ± 0.16
cREx	74.75 ± 14.14	32.29 ± 7.71	29.75 ± 5.16	19.77 ± 14.98
cGroupDRO	41.10 ± 2.37	41.83 ± 2.96	42.10 ± 2.15	**21.81 ± 5.40**
cMMD	89.87 ± 1.13	55.02 ± 2.29	27.36 ± 1.57	8.82 ± 0.70
PGI	78.23 ± 2.01	**55.57 ± 4.60**	**51.62 ± 3.09**	18.84 ± 2.11

Table 4: Generalisation performance on COCO-ON-PLACES.

Methods	In-distribution	Non-systematic shift	Systematic shift	Anomaly detection
Base (ERM)	81.06 ± 1.01	45.25 ± 0.96	29.18 ± 1.24	9.21 ± 0.21
IRMv1	80.93 ± 0.71	45.17 ± 0.92	28.78 ± 0.73	9.39 ± 0.60
REx	**81.55 ± 0.70**	45.35 ± 0.92	29.56 ± 0.77	9.46 ± 0.51
GroupDRO	76.05 ± 0.87	43.72 ± 0.43	31.83 ± 0.54	9.61 ± 0.55
Reweight	81.14 ± 0.80	45.84 ± 0.70	30.37 ± 1.16	9.75 ± 0.69
cIRMv1	80.08 ± 1.90	44.96 ± 2.88	30.06 ± 2.07	9.64 ± 0.94
cREx	81.50 ± 0.76	45.44 ± 0.96	29.12 ± 0.97	9.17 ± 0.59
cGroupDRO	78.25 ± 0.31	41.69 ± 0.08	28.16 ± 0.91	9.45 ± 0.22
cMMD	79.64 ± 0.73	**49.44 ± 0.99**	35.86 ± 0.66	9.80 ± 0.45
PGI	75.00 ± 0.85	46.10 ± 0.79	**36.25 ± 0.42**	**11.12 ± 0.85**

Hyper-parameter selection

NS+S	Both non-systematic* and systematic
NS	Non-systematic* only
NS+ID	Non-systematic* + in-distribution
ID	In-distribution only

*Non-systematic validation sets use different colours than in training and test sets

Table 5: Hyper-parameters with different validation sets for COLOURED MNIST

Validation	In-distribution	Non-systematic shift	Systematic shift	Anomaly detection
NS+S (PGI)	99.05 ± 0.08	98.58 ± 0.06	98.48 ± 0.05	89.42 ± 1.95
NS (PGI)	99.31 ± 0.05	98.21 ± 0.26	97.54 ± 0.41	76.00 ± 4.06
NS+ID (PGI)	99.30 ± 0.07	98.31 ± 0.27	97.48 ± 0.45	76.07 ± 5.67
ID only (PGI)	99.69 ± 0.03	63.62 ± 2.05	58.18 ± 2.05	11.81 ± 1.89
Base (ERM)	99.60 ± 0.02	53.26 ± 1.89	38.72 ± 2.27	7.70 ± 0.23

Table 6: Hyper-parameters with different validation sets for COCO-ON-COLOURS

Validation	In-distribution	Non-systematic shift	Systematic shift	Anomaly detection
NS+S (PGI)	78.23 ± 2.01	55.57 ± 4.60	51.62 ± 3.09	18.84 ± 2.11
NS (PGI)	85.78 ± 1.45	51.02 ± 2.32	38.85 ± 2.29	15.71 ± 3.25
NS+ID (PGI)	85.78 ± 1.45	51.02 ± 2.32	38.85 ± 2.29	15.71 ± 3.25
ID only (cMMD)	92.51 ± 0.41	44.59 ± 3.28	10.48 ± 0.98	6.05 ± 0.23
Base (ERM)	90.57 ± 1.28	26.81 ± 4.93	1.10 ± 0.36	5.47 ± 0.08

Table 7: Hyper-parameters with different validation sets for COCO-ON-PLACES

Validation	In-distribution	Non-systematic shift	Systematic shift	Anomaly detection
NS+S (cMMD)	79.64 ± 0.73	49.44 ± 0.99	35.86 ± 0.66	9.80 ± 0.45
NS (cMMD)	79.64 ± 0.73	49.44 ± 0.99	35.86 ± 0.66	9.80 ± 0.45
NS+ID (cMMD)	79.64 ± 0.73	49.44 ± 0.99	35.86 ± 0.66	9.80 ± 0.45
ID only (PGI)	80.99 ± 0.52	47.63 ± 0.90	31.91 ± 0.89	9.59 ± 0.89
Base (ERM)	81.06 ± 1.01	45.25 ± 0.96	29.18 ± 1.24	9.21 ± 0.21

Best performing methods for a mix of non-systematic and systematic generalisation performance

Some takeaways

Invariance methods/penalties across inferred splits of a dataset appear to be useful at improving performance under distributional shift.

Such methods cannot be useful when spurious correlations are completely pervasive; ensuring diversity in data through curation could help.

The question of how to trade off in-distribution performance with that in unexpected situations is relevant. Real-world problems might call for more targeted invariance methods.

Some takeaways

Invariance methods/penalties across inferred splits of a dataset appear to be useful at improving performance under distributional shift.

Such methods cannot be useful when spurious correlations are completely pervasive; ensuring diversity in data through curation could help.

The question of how to trade off in-distribution performance with that in unexpected situations is relevant. Real-world problems might call for more targeted invariance methods.

Some takeaways

Invariance methods/penalties across inferred splits of a dataset appear to be useful at improving performance under distributional shift.

Such methods cannot be useful when spurious correlations are completely pervasive; ensuring diversity in data through curation could help.

The question of how to trade off in-distribution performance with that in unexpected situations is relevant. Real-world problems might call for more targeted invariance methods.

Thank you!