Compositional Risk Minimization

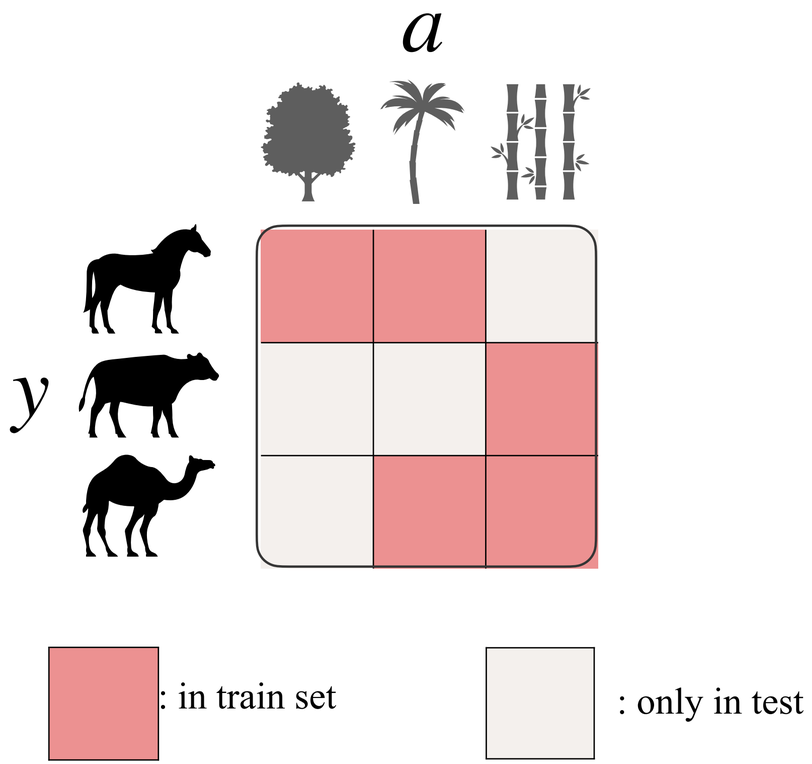
Divyat Mahajan[1,2⋆], **Mohammad Pezeshki**[1], **Charles Arnal**[1], **Ioannis Mitliagkas**[2], **Kartik Ahuja**[1,†], **Pascal Vincent**[1,2⋆,†]

[1]Meta FAIR, [2]Mila, Université de Montréal, DIRO
[⋆]Work done at Meta, [†]Joint last author

International Conference on Machine Learning (*ICML*) 2025

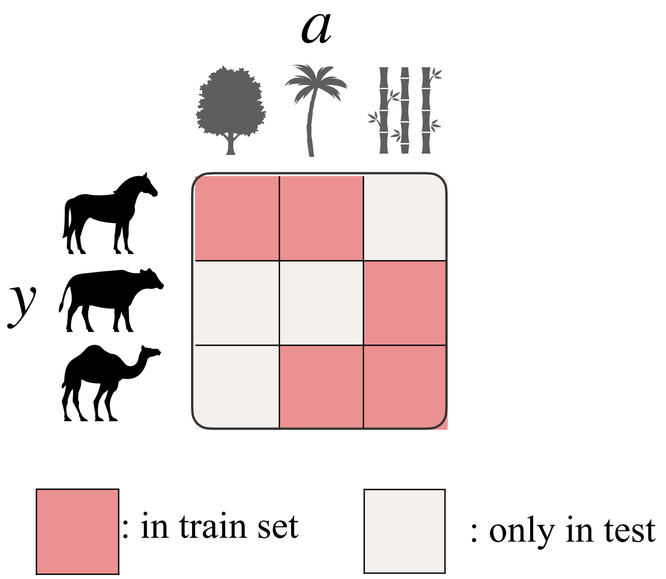
Meta Mila Université de Montréal

Compositional Shifts



- Some combinations of attributes are totally absent from the training distribution but present in the test distribution

Compositional Shifts



Compositional Distribution Shifts

- Assumption 1: $p(x \mid z) = q(x \mid z)\ \forall z \in \mathcal{Z}^{\times}$
- Assumption 2: $\mathcal{Z}^{\text{test}} \not\subseteq \mathcal{Z}^{\text{train}}$ but $\mathcal{Z}^{\text{test}} \subseteq \mathcal{Z}^{\times}$

- Attribute Vector: $z = (z_1, \ldots, z_m)$ that characterizes the group for the input x
  - Each attribute z_i is categorical and can take d possible values.
- Train Distribution: $p(x, z) = p(z)p(x \mid z)$ with support of z as $\mathcal{Z}^{\text{train}}$
- Test Distribution: $q(x, z) = q(z)q(x \mid z)$ with support of z as $\mathcal{Z}^{\text{test}}$
- Cartesian Product: $\mathcal{Z}^{\times} = \mathcal{Z}_1^{\text{train}} \times \mathcal{Z}_2^{\text{train}} \times \cdots \mathcal{Z}_m^{\text{train}}$

Contributions

Build classifiers that are robust to compositional distributions shifts!

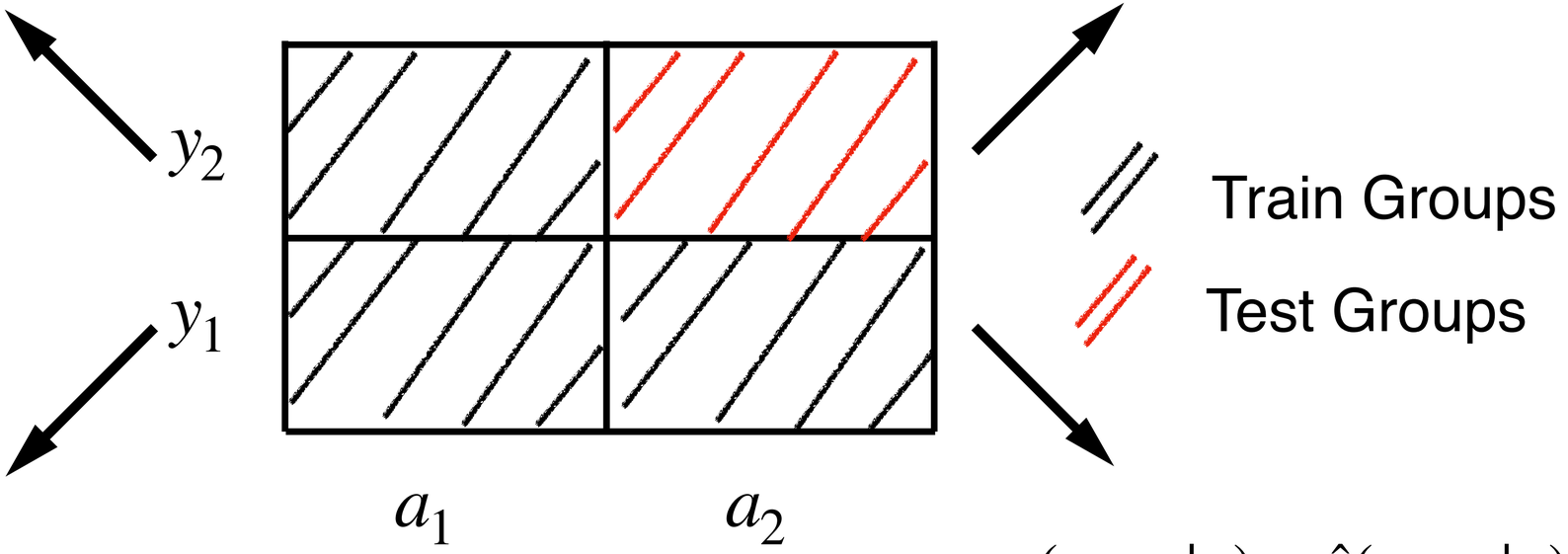
Theory of Compositional Shifts. For the family of additive energy distributions, we prove that additive energy classifiers generalize compositionally to novel combinations of attributes represented by a special mathematical object, which we call *discrete affine hull*.

A Practical Method. We propose simple algorithm Compositional Risk Minimization (CRM), which first trains an additive energy classifier and then adjusts the trained classifier for tackling compositional shifts.

Cartesian Product Extrapolation (CPE)

$p(y_2, a_1 \mid x) = \hat{p}(y_2, a_1 \mid x)$

$q(y_2, a_2 \mid x) = \hat{q}(y_2, a_2 \mid x)$



$p(y_1, a_1 \mid x) = \hat{p}(y_1, a_1 \mid x)$

$p(y_1, a_2 \mid x) = \hat{p}(y_1, a_2 \mid x)$

Question: What assumptions should be place on $p(x \mid z)$ for CPE?

Additive Energy Distribution (AED)

$$p(x \mid z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-1^T E(x, z) \Big) \quad \text{where} \quad 1^T E(x, z) = \sum_{i=1}^{m} E_i(x, z_i)$$

Conditional distribution of data given factors

Partition Function

Energy Function

Energy Function for each component

- **Assumption**: The energy function can be decomposed as addition of energies with different components of z
  - Natural choice to model inputs that satisfy a conjunction of characteristics
- Partition function can model interaction between components of z

$$\mathbb{Z}(z) = \int \exp\Big(-1^T \, E(x, z) \Big) dx$$

Additive Energy Distribution (AED)

$$p(x \mid z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-1^T E(x, z) \Big) \quad \text{where} \quad 1^T E(x, z) = \sum_{i=1}^{m} E_i(x, z_i)$$

Conditional distribution of data given factors

Partition Function

Energy Function

Energy Function for each component

- AED expressed with inner product:

$$p(x \mid z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-\sigma(z)^T E(x) \Big)$$

where $\sigma(z) = [\text{onehot}(z_1), \ldots, \text{onehot}(z_m)]^\top$,

$$E(x) = [E_1(x,1), \ldots, E_1(x,d), \ldots, E_m(x,1), \ldots, E_m(x,d)]^\top$$

Provable Extrapolation with CRM: Step 1

True Model:

$$p(z\,|\,x) = Softmax(\log p(x\,|\,z) + \log p(z)) \quad \text{where} \quad p(x\,|\,z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-\sigma(z)^T E(x) \Big)$$

Learned Model (Train):

$$\hat{p}(z\,|\,x) = Softmax(\log \hat{p}(x\,|\,z) + \log p(z)) \quad \text{where} \quad \hat{p}(x\,|\,z) = \frac{1}{\hat{B}(z)} \exp\Big(-\sigma(z)^T \hat{E}(x) \Big)$$

Free parameter

- **CRM First Step:** $\hat{E}, \hat{B} \in argmin_{\tilde{E},\tilde{B}} R(\tilde{p})$ where $R(\tilde{p}) = \mathbb{E}_{(x,z)\sim p}\Big[-\log \tilde{p}(z\,|\,x) \Big]$

Provable Extrapolation with CRM: Step 2

True Model:

$$p(z\,|\,x) = Softmax(\log p(x\,|\,z) + \log p(z)) \quad \text{where} \quad p(x\,|\,z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-\sigma(z)^T E(x) \Big)$$

Learned Model (Train):

$$\hat{p}(z\,|\,x) = Softmax(\log \hat{p}(x\,|\,z) + \log p(z)) \quad \text{where} \quad \hat{p}(x\,|\,z) = \frac{1}{\hat{B}(z)} \exp\Big(-\sigma(z)^T \hat{E}(x) \Big)$$

Learned Model (Eval):

$$\hat{q}(z\,|\,x) = Softmax(\log \hat{q}(x\,|\,z) + \log \hat{q}(z)) \quad \text{where} \quad \hat{q}(x\,|\,z) = \frac{1}{B^\star(z)} \exp\Big(-\sigma(z)^T \hat{E}(x) \Big)$$

- **CRM Second Step:** Under *AED* assumption and test group as *affine combination* of train groups,

$$B^\star(z) = \log\left(\mathbb{E}_{x\sim p(x)} \left[\frac{\exp\Big(-\sigma(z)^T \hat{E}(x) \Big)}{\sum_{\tilde{z}\in\mathcal{Z}^{\text{train}}} \exp\Big(-\sigma(\tilde{z})^T \hat{E}(x) + \log p(\tilde{z}) - \hat{B}(\tilde{z}) \Big)} \right] \right)$$

Provable Extrapolation with CRM

True Model:

$$p(z\,|\,x) = Softmax(\log p(x\,|\,z) + \log p(z)) \quad \text{where} \quad p(x\,|\,z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-\sigma(z)^T E(x) \Big)$$

Learned Model (Train):

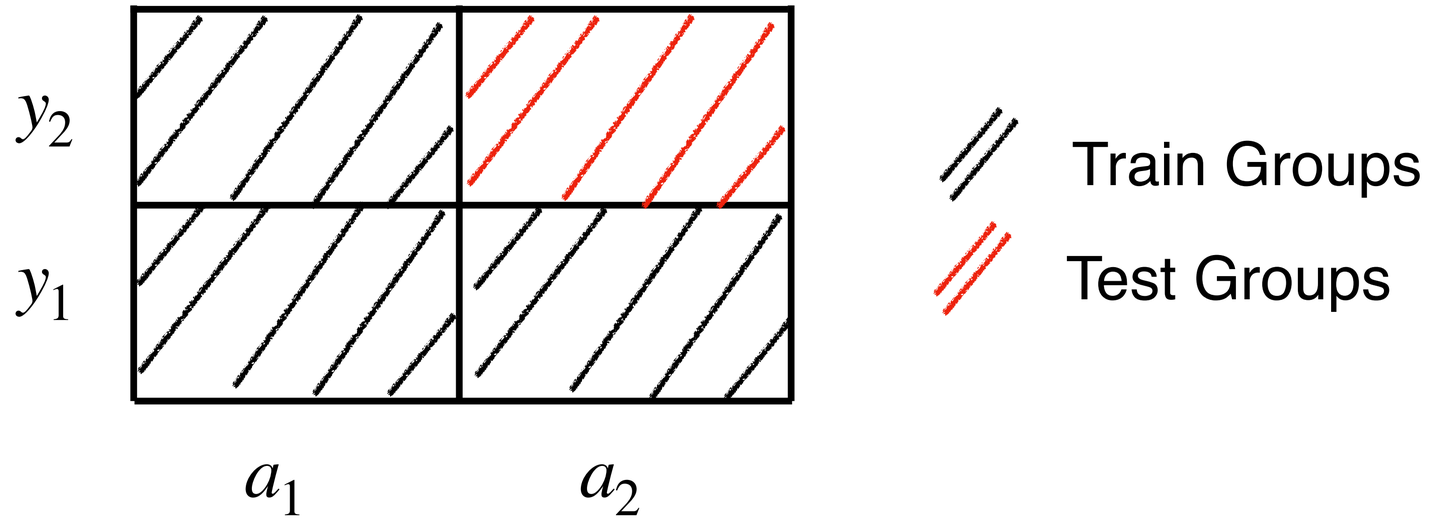
$$\hat{p}(z\,|\,x) = Softmax(\log \hat{p}(x\,|\,z) + \log p(z)) \quad \text{where} \quad \hat{p}(x\,|\,z) = \frac{1}{\hat{B}(z)} \exp\Big(-\sigma(z)^T \hat{E}(x) \Big)$$

Learned Model (Eval):

$$\hat{q}(z\,|\,x) = Softmax(\log \hat{q}(x\,|\,z) + \log \hat{q}(z)) \quad \text{where} \quad \hat{q}(x\,|\,z) = \frac{1}{B^\star(z)} \exp\Big(-\sigma(z)^T \hat{E}(x) \Big)$$

Theorem: If $\hat{p}(z\,|\,x) = p(z\,|\,x), \forall z \in \mathcal{Z}^{\text{train}}$, and $\hat{q}(z) = q(z)$ then $\hat{q}(z\,|\,x) = q(z\,|\,x), \forall z \in DAff(\mathcal{Z}^{\text{train}})$

Discrete Affine Hull Extension



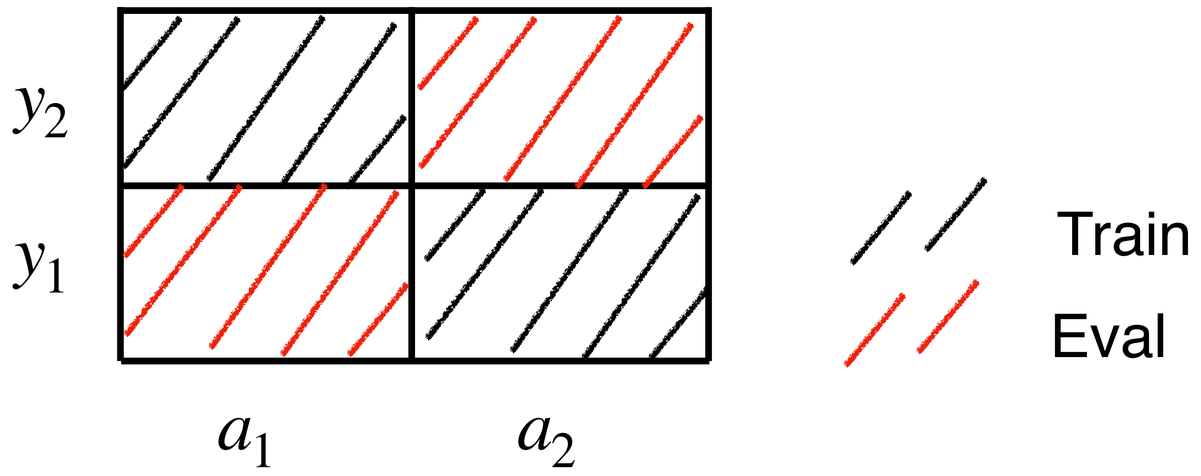
- Test group can be expressed as affine combination of train groups

$$\sigma(y_2, a_2) = \sigma(y_2, a_1) - \sigma(y_1, a_1) + \sigma(y_1, a_2)$$

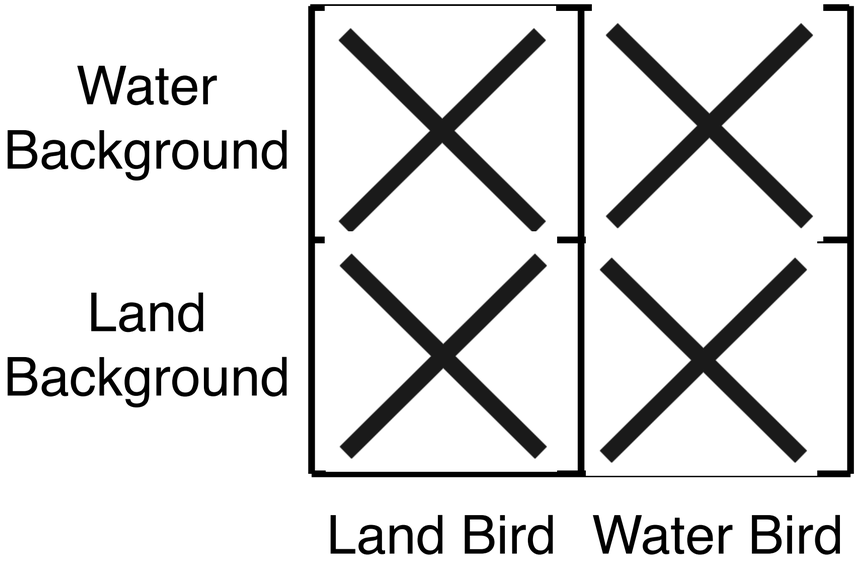
$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = (+1) \cdot \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + (-1) \cdot \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + (+1) \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

CPE is not always same as Discrete Affine Hull

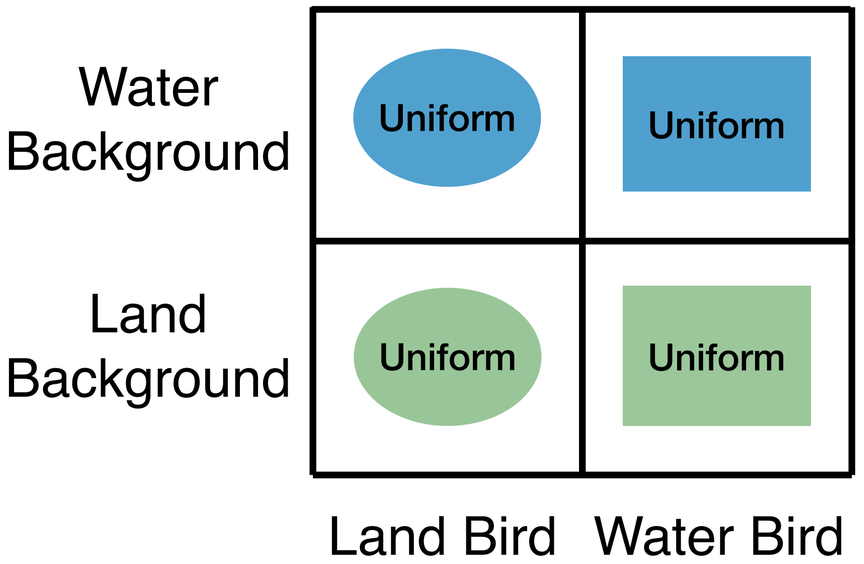
Note that extrapolation to novel groups depends on the support of train groups!



Experiments: Setup



Train Distribution



Test Distribution

- Factors $z = (y, a)$ where y denotes the class label and a denotes the spurious attribute
- Compositional Shift: $\mathcal{Z}^{\text{train}} \neq \mathcal{Z}^{\text{test}}$ but $\mathcal{Z}^{\text{test}} = DAff(\mathcal{Z}^{\text{train}})$

Experiments: Results

Dataset	Method	Average Acc	WGA	WGA (No Groups Dropped)
Waterbirds	ERM	77.9 (0.1)	43.0 (0.1)	62.3 (1.2)
	G-DRO	77.9 (0.6)	42.3 (2.5)	87.3 (0.3
	LC	88.3 (0.7)	75.5 (0.8)	88.7 (0.3)
	sLA	89.3 (0.4)	77.3 (0.5)	89.7 (0.3)
	CRM	87.1 (0.7)	78.7 (1.6)	86.0 (0.6)
CelebA	ERM	85.8 (0.3)	39.0 (0.6)	52.0 (1.0)
	G-DRO	89.2 (0.5)	67.7 (1.3)	91.0 (0.6)
	LC	91.1 (0.2)	57.4 (0.6)	90.0 (0.6)
	sLA	90.9 (0.1)	57.4 (0.3)	86.7 (1.9)
	CRM	91.1 (0.2)	81.8 (1.2)	89.0 (0.6)
MetaShift	ERM	85.7 (0.4)	60.5 (0.6)	63.0 (0.0)
	G-DRO	86.0 (0.4)	63.8 (0.6)	80.7 (1.3)
	LC	88.5 (0.0)	68.2 (0.5)	80.0 (1.2)
	sLA	88.4 (0.1)	63.0 (0.5)	80.0 (1.2)
	CRM	87.6 (0.2)	73.4 (0.7)	74.7 (1.5)
MultiNLI	ERM	69.1 (0.7)	7.2 (0.6)	68.0 (1.7)
	G-DRO	70.4 (0.1)	34.3 (0.5)	57.0 (2.3)
	LC	75.9 (0.1)	54.3 (0.5)	74.3 (1.2)
	sLA	76.4 (0.5)	55.0 (1.8)	71.7 (0.3)
	CRM	74.6 (0.5)	57.7 (3.0)	74.7 (1.3)
CivilComments	ERM	80.4 (0.1)	55.8 (0.4)	61.0 (2.5)
	G-DRO	80.1 (0.2)	61.6 (0.4)	64.7 (1.5)
	LC	80.7 (0.1)	65.7 (0.5)	67.3 (0.3)
	sLA	80.6 (0.1)	65.6 (0.1)	66.3 (0.9)
	CRM	83.7 (0.1)	68.1 (0.5)	70.0 (0.6)
NICO++	ERM	85.0 (0.0)	35.3 (2.3)	35.3 (2.3)
	G-DRO	84.0 (0.0)	36.7 (0.7)	33.7 (1.2)
	LC	85.0 (0.0)	35.3 (2.3)	35.3 (2.3)
	sLA	85.0 (0.0)	33.0 (0.0)	35.3 (2.3)
	CRM	84.7 (0.3)	40.3 (4.3)	39.0 (3.2)

- We report test Average Accuracy and Worst Group Accuracy (WGA), averaged as a group is dropped from training and validation sets
- Last column is WGA under the dataset's standard subpopulation shift benchmark, i.e. with no group dropped
- All methods have a harder time to generalize when groups are absent from training, but CRM appears consistently more robust

Chat with us during the poster session!