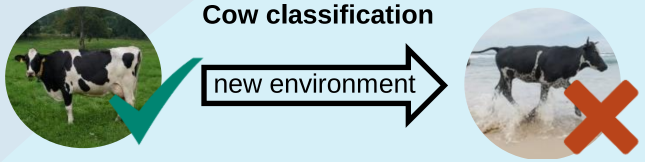
# Analysis of **Invariant Risk Minimization (IRM)** in **Out-of-domain Generalization**

## 1. Out-of-domain Generalization Problem

**Cow classification**




Learning algorithms can perform poorly in unseen environments when they learn **spurious correlations** (e.g. green pasture) [1].

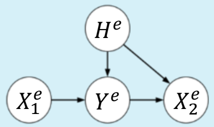
## 2. Invariant Risk Minimization (IRM)

- The IRM method attempts to **solve** this problem [2]
- By learning **invariant relationships** in the data (e.g. shape of cow)
- The simplified version IRMv1 is considered

## 3. Research Question

For which **data distribution shifts** is the IRMv1 method able to **capture invariance**?

## 4. Synthetic Data Model




$$H^e \leftarrow \mathcal{N}(0, \sigma_e^2)$$
$$X_1^e \leftarrow \mathcal{N}(0, \sigma_e^2)$$
$$Y^e \leftarrow \mathcal{N}(0, \sigma_y^2) + W_{1\to y}X_1^e + W_{h\to y}H^e$$
$$X_2^e \leftarrow \mathcal{N}(0, \sigma_2^2) + W_{y\to 2}Y^e + W_{h\to 2}H^e$$

*Figure 1: The synthetic data model used for the experiments, where $Y^e$ should be predicted from $X^e = [X_1^e, X_2^e]$. The symbol $\sigma_e^2$ is the variance in environment e.*

## 5. Data Distribution Shifts

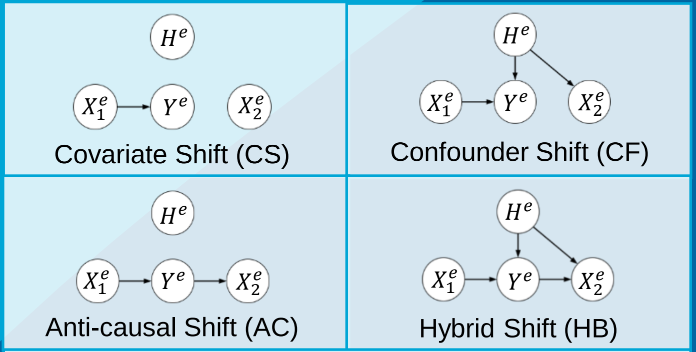



*Figure 2: The four data distribution shifts represented in the synthetic data model.*

## 6. Methodology

1. Train IRM on environments corresponding to the **data distribution shifts**.
2. IRM learns a prediction rule of the form $\hat{Y} = [\hat{W}_{1\to y}, \hat{W}_{y\to 2}]X$.
3. The **optimal invariant predictor** is $\hat{Y} = [W_{1\to y}, 0]X$.
4. The **model estimation error** is the distance between the learned prediction rule and the optimal invariant predictor: $||[\hat{W}_{1\to y}, \hat{W}_{y\to 2}] - [W_{1\to y}, 0]||^2$.
5. Compare IRM's error to that of the **non-invariant ERM**.

## 7. Experiment

- The variance of the noise of the underlying label (Y) can be **varying** (heteroskedastic) or **stable** (homoskedastic) [3].
- The experiment considers the following cases:
- **Heteroskedastic Y-noise** where $\sigma_y^2 = \sigma_e^2$ and $\sigma_2^2 = 1$
- **Homoskedastic Y-noise** where $\sigma_y^2 = 1$ and $\sigma_2^2 = \sigma_e^2$
- **Homoskedastic Y-noise with constant $X_2$** where $\sigma_y^2 = 1$ and $\sigma_2^2 = 1$

## 8. Results

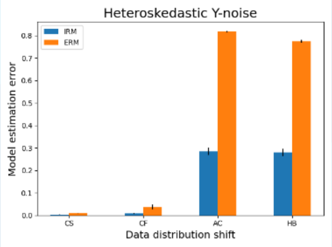



*Figure 3: Results under heteroskedastic Y-noise.*

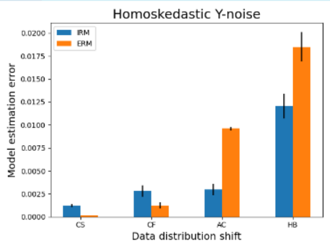



*Figure 4: Results under homoskedastic Y-noise.*

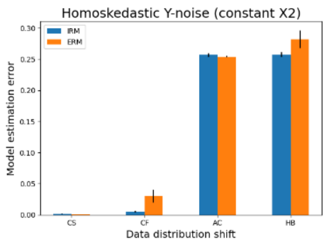



*Figure 5: Results under homoskedastic Y-noise with constant $X_2$.*

## 9. Discussion

- Figure 3:

IRM **outperforms** ERM in all shifts. It is **sub-optimal** in the AC and HB shift, because of the anti-causal link.

- Figure 4:

The errors are significantly smaller, because regression is **simpler**. ERM is better than IRM in the CS and CF shift, so the **regularizer** should have been smaller.

- Figure 5:

IRM recognizes the **confounder**. However, it yields large error in the presence of the anti-causal link. A **strong spurious correlation** is formed, because the $X_2$-noise follows the same distribution as the Y-noise.

## 10. Conclusion

For which **data distribution shifts** is the IRMv1 method able to **capture invariance**?

In the **CS** and **CF** shifts, IRM generally captures invariance. Except under homoskedastic Y-noise, when the regularizer is too large.
In the **AC** and **HB** shifts, IRM learns the invariant relationships when the spurious features do not follow the same distribution as the label.

## 11. Limitations

- The mentioned experiment is done on a fixed set of training environments (consult the paper for additional experiments).
- The weights related to the label do not reflect real-world randomness.
- Additional experiments with regards to the regularizer are needed to verify the discussion.

## References


[1] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. *CVPR*, 2011.
[2] Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ICLR*, 2019.
[3] Patrick J. Rosopa, Meline M. Schaffer, and Amber N. Schroeder. Managing heteroscedasticity in general linear models. *Psychological Methods*, 2013.



**Contact**
Jochem van Lith
j.a.e.vanlith@student.tudelft.nl

**Supervisors**
Rickard Karlsson, Stephan Bongers and Jesse Krijthe