# All about OOD generalization I Invariant representation

Presenter: Yidong Ouyang

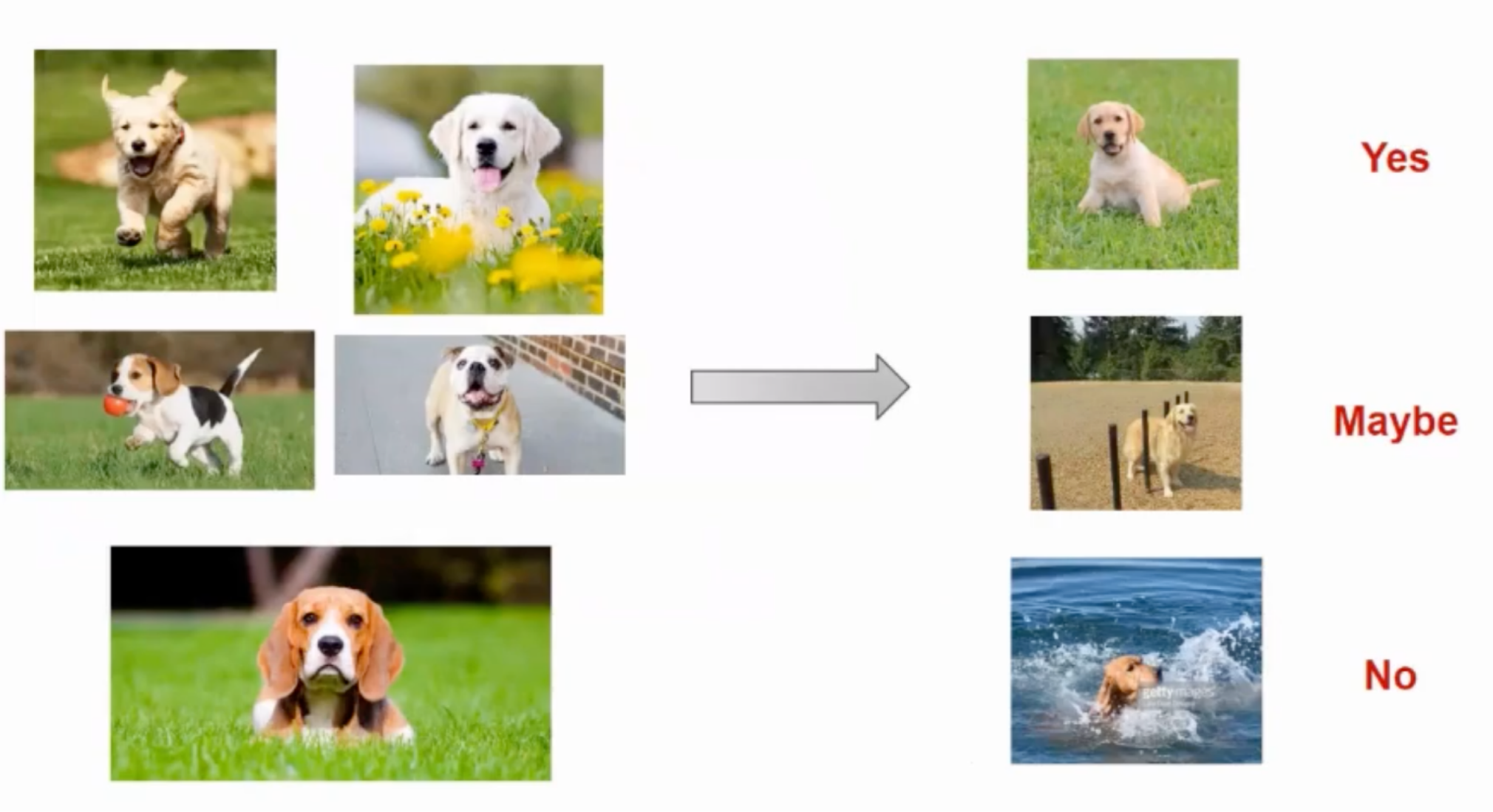
The Chinese University of Hong Kong, shenzhen

# Content

- 1.The big picture of out-of-distribution generalization (OOD)
- 2.The background of Invariant Risk Minimization (IRM) [Arjovsky2019InvariantRM]
- 3.An analysis on invariant principle [Ahuja2021InvariancePM]
- 4.The risk of IRM [Rosenfeld2021TheRO]
- 5.Provable out-of-distribution generalization [Chen2021IterativeFM]

# V

- 1.W
hav
  - (
  - |
  - V
  - |
- 2.M
fro




ay

e-ID

' suffer

Cite from Peng cui

# OOD has a very close relationship with several research area

- 1.transfer learning, domain adaptation, Multi-task learning
- 2.causal learning, counterfactuals
- 3.robust learning (worst-case generalization)
- 4.disentangle
- 5. meta learning, zero-shot learning
- 6. long-tail
- 7. lifelong learning, continue learning, online learning

# The big picture of OOD generalization

- Domain generalization bound
  - H-divergence bound [Albuquerque2019GeneralizingTU]
  - Kernel bound [Blanchard2021DomainGB]
  - Information theatrical lower bound [Zhao2019OnLI]
  - Recent progress [Ye2021TowardsAT, Federici2021AnIA]

- Invariant feature learning
  - Invariant risk minimization (IRM) and a dozen of its variants
  - The risk of IRM
  - IFM--Provable out-of-distribution generalization with Logarithmic Environments

# The background of Invariant Risk Minimization

- Data generating process:
  - Assume data are drawn from a set of $E$ training environments $\mathcal{E} = \{e_1, e_2, \dots, eE\}$
  - Label $y = \begin{cases} 1, w.p.\eta \\ -1, otherwise \end{cases}$
  - Invariant feature $z_c \sim N(y \cdot \mu_c, \sigma_c^2 I)$, spurious feature $z_e \sim N(y \cdot \mu_e, \sigma_e^2 I)$
  - $\mu_c \in \mathbb{R}^{d_c}$ , $\mu_e \in \mathbb{R}^{d_e}$
  - $x = f(z_c, z_e)$
- Model:
  - Feature extractor: $\Phi$, classifier $\beta$
  - $\hat{y} = \sigma(\beta^T \Phi(x))$
- Goal:
  - Make prediction relying on invariant feature $z_c$
  - $\min \mathcal{R}^{e_{E_+1}}(\Phi, \beta)$ where $\mathcal{R}^e(\Phi, \beta) = \mathbb{E}_{(x,y)\sim pe}[l(\sigma(\beta^T \Phi(x)))]$

# The proposal of IRM [Arjovsky2019InvariantRM]

- Review several methods: (They fail to discover the invariant feature)
  - 1. Empirical Risk Minimization (ERM)
    - Objective: $\min \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathcal{R}^e \ (\Phi, \beta)$
    - Comments: rely on $z_e$
  - 2. Minimization on the worst case
    - $\min max_{e \in \mathcal{E}} \ \mathcal{R}^e(\Phi, \beta) - r$

    **Proposition 2.** *Given KKT differentiability and qualification conditions,* $\exists \lambda_e \geq 0$ *such that the minimizer of* $R^{\text{rob}}$ *is a first-order stationary point of* $\sum_{e \in \mathcal{E}_{tr}} \lambda_e R^e(f)$.

    - Comments: also rely on $z_e$

# IRM

- Goal：
  - How to make feature extractor only rely on invariant feature $z_c$?
  - find a data representation such that the optimal classifier on top of that representation matches for all environments.
- Objective:
  - $\min_{\Phi,\beta} \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathcal{R}^e\ (\Phi,\beta)\ s.t.\beta \in \text{arg}min_{\hat{\beta}}\ \mathcal{R}^e\left(\Phi,\hat{\beta}\right) \forall \text{e} \in \mathcal{E}$ (4)
  - Practical version:
  - $\min \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}[\mathcal{R}^e\ (\Phi,\beta) + \lambda||\nabla_\beta \mathcal{R}^e(\Phi,\beta)||_2^2]$ ($\mathcal{R}^e(\Phi,\beta)$ is convex since $\beta$ is linear classifier)

# An analysis on invariant principle [Ahuja2021InvariancePM]

## Motivation:

**Despite the promising theory, why invariance principle-based approaches fail in common classification tasks**

## Key point:

**1.Revisit the fundamental assumptions in linear regression tasks and show that for linear classification tasks we need much stronger restrictions on the distribution shifts**

## Main results:

| Task | Invariant features capture label info | Support overlap invariant features | Support overlap spurious features | OOD generalization guarantee ($\mathcal{E}_{tr} \to \mathcal{E}_{all}$) ERM | IRM | IB-ERM | IB-IRM |
|---|---|---|---|---|---|---|---|
| Linear Classification | Full/Partial | No | Yes/No | Impossible for any algorithm to generalize OOD [Thm2] | | | |
| | Full | Yes | No | ✗ | ✗ | ✓ | ✓ [Thm3,4] |
| | Partial | Yes | No | ✗ | ✗ | ✗ | ✓ [Appendix] |
| | Full | Yes | Yes | ✓ | ✓ | ✓ | ✓ [Thm3,4] |
| | Partial | Yes | Yes | ✗ | ✓ | ✗ | ✓ |
| Linear Regression | Full | No | No | ✓ | ✓ | ✓ | ✓ |
| | Partial | No | No | ✗ | ✓ | ✗ | ✓ [Thm4] |

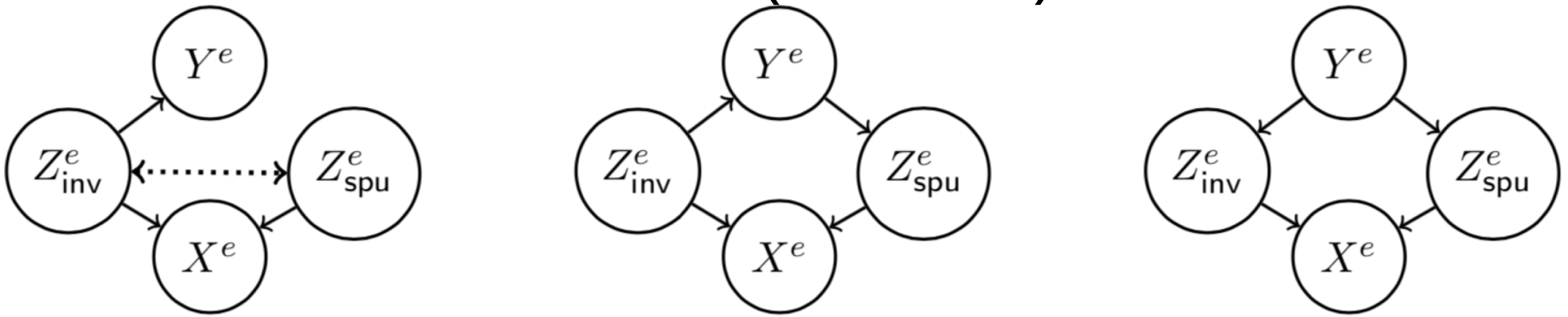
# Understanding the failures case for classification task

**Key idea:**

Whether the invariant feature fully capture the information about the label?

$$Y \perp X^e | \Phi^*(X^e) \quad \textbf{vs.} \quad Y \not\perp X^e | \Phi^*(X^e)$$

**fully informative invariant features vs. partially informative invariant features (FIIF vs. PIIF)**




(a) FIIF (this work) (b) PIIF [Arjovsky et al., 2019] (c) PIIF [Rosenfeld et al., 2021b]

# Linear classification structural equation model

$$
\begin{aligned}
&Y^e \leftarrow \mathsf{I}\left(w_{\mathsf{inv}}^* \cdot Z_{\mathsf{inv}}^e\right) \oplus N^e, \qquad N^e \sim \text{Bernoulli}(q), q < \frac{1}{2}, \qquad N^e \perp (Z_{\mathsf{inv}}^e, Z_{\mathsf{spu}}^e), \\
&X^e \leftarrow S\left(Z_{\mathsf{inv}}^e, Z_{\mathsf{spu}}^e\right),
\end{aligned} \tag{5}
$$

*where* $w_{\mathsf{inv}}^* \in \mathbb{R}^m$ *with* $\|w_{\mathsf{inv}}^*\| = 1$ *is the labelling hyperplane,* $Z_{\mathsf{inv}}^e \in \mathbb{R}^m$, $Z_{\mathsf{spu}}^e \in \mathbb{R}^o$, $N^e$ *is binary noise with identical distribution across environments,* $\oplus$ *is the XOR operator,* $S$ *is invertible.*

**Contrast with Invariant feature $z_c \sim N\left(y \cdot \mu_c, \sigma_c^2 I\right)$, spurious feature $z_e \sim N(y \cdot \mu_e, \sigma_e^2 I)$ mentioned before.**

# Theorem 2—the importance of $z_{inv}^{e}$ overlap

**Theorem 2.** ***Impossibility of guaranteed OOD generalization for linear classification.*** *Suppose each $e \in \mathcal{E}_{all}$ follows Assumption 2. If for all the training environments $\mathcal{E}_{tr}$, the latent invariant features are bounded and strictly separable, i.e., Assumption 3 and 7 hold, then every deterministic algorithm fails to solve the OOD generalization (eq. (1)), i.e., for the output of every algorithm $\exists\, e \in \mathcal{E}_{all}$ in which the error exceeds the minimum required value $q$ (noise level).*

**Assumption 3.** ***Bounded invariant features.*** $\cup_{e\in\mathcal{E}_{tr}} \mathcal{Z}^e_{\text{inv}}$ *is a bounded set.*[4]

**Assumption 4.** ***Bounded spurious features.*** $\cup_{e\in\mathcal{E}_{tr}} \mathcal{Z}^e_{\text{spu}}$ *is a bounded set.*

[4] A set $\mathcal{Z}$ is bounded if $\exists M < \infty$ such that $\forall z \in \mathcal{Z}, \|z\| \leq M$.

**Assumption 5.** ***Invariant feature support overlap.*** $\forall e \in \mathcal{E}_{all}, \mathcal{Z}^e_{\text{inv}} \subseteq \cup_{e'\in\mathcal{E}_{tr}} \mathcal{Z}^{e'}_{\text{inv}}$

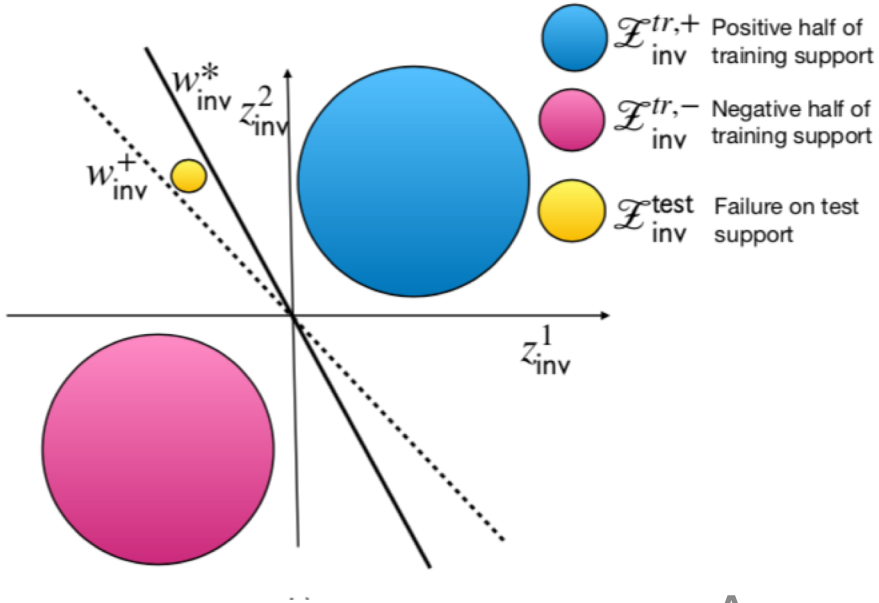
**Assumption 6.** ***Spurious feature support overlap.*** $\forall e \in \mathcal{E}_{all}, \mathcal{Z}^e_{\text{spu}} \subseteq \cup_{e'\in\mathcal{E}_{tr}} \mathcal{Z}^{e'}_{\text{spu}}$

**Assumption 7.** ***Strictly separable invariant features.*** Inv-Margin $> 0$.

# Theorem 3—whether $z^e_{spu}$ matters

**Theorem 3.** ***Sufficiency and Insufficiency of ERM and IRM.*** *Suppose each $e \in \mathcal{E}_{all}$ follows Assumption 2. Assume that a) the invariant features are strictly separable, bounded, and satisfy support overlap, b) the spurious features are bounded (Assumptions 3-5, 7 hold).*

- ***Sufficiency:*** *If the spurious features satisfy support overlap (Assumption 6 holds), then both ERM and IRM solve the OOD generalization problem (eq. (1)). Also, some of the ERM and IRM solutions rely on the spurious features and still achieve OOD generalization.*
- ***Insufficiency:*** *If spurious features do not satisfy support overlap, then both ERM and IRM fail at solving the OOD generalization problem (eq. (1)). Also, the classifiers that solve the OOD generalization problem do not rely on spurious features at all.*




**That is different from linear regression task.**

# The risk of IRM in linear case [Rosenfeld2021TheRO]

- **Theorem 5.1 (Linear case).** *Assume $f$ is linear. Suppose we observe $E$ training environments. Then the following hold:*

1. *Suppose $E > d_e$. Consider any linear featurizer $\Phi$ which is feasible under the IRM objective (4), with invariant optimal classifier $\hat{\beta} \neq 0$, and write $\Phi(f(z_c, z_e)) = Az_c + Bz_e$. Then under mild non-degeneracy conditions, it holds that $B = 0$. Consequently, $\hat{\beta}$ is the optimal classifier for all possible environments.*

2. *If $E \leq d_e$ and the environmental means $\mu_e$ are linearly independent, then there exists a linear $\Phi$—where $\Phi(f(z_c, z_e)) = Az_c + Bz_e$ with $rank(B) = d_e + 1 - E$—which is feasible under the IRM objective. Further, both the logistic and 0-1 risks of this $\Phi$ and its corresponding optimal $\hat{\beta}$ are strictly lower than those of the optimal invariant predictor.*

# Proof of Theorem 5.1

- Key idea:
  - Construct a example that $\Phi$ depends on the environment while the optimal classifier $\beta$ for each environment is constant.
  - Linear classifier to separate two gaussian
- 1.define $\Phi(x) = [z_c, p^T z_e]$ $where$ $\forall e \in \mathcal{E}$, $p^T \mu_e = \sigma_e^2 \ \tilde{\mu}$ $\quad \tilde{\mu}\ is\ a\ fixed\ scalar$
- $\Rightarrow p^T z_e | y \sim N(y \cdot p^T \mu_e, \|p\|_2^2 \sigma_e^2) = N(y \cdot \sigma_e^2 \tilde{\mu}, \sigma_e^2)$
- 2. For separating two Gaussians, the optimal linear classifier is $\Sigma^{-1}(\mu_1 - \mu_0)$ $\Rightarrow$the optimal classifier is $2\tilde{\mu}$

# Whole proof of Theorem 5.1 —lemma C.1

**Lemma C.1.** *Suppose we observe $E$ environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$, each with environmental mean of dimension $d_e \geq E$, such that all environmental means are linearly independent. Then there is a unique unit-norm vector $p$ such that*

$$p^T \mu_e = \sigma_e^2 \tilde{\mu} \ \ \forall e \in \mathcal{E}, \tag{11}$$

*where $\tilde{\mu}$ is the largest scalar which admits such a solution.*

*Proof.* Let $v_1, v_2, \ldots, v_E$ be a set of basis vectors for span$\{\mu_1, \mu_2, \ldots, \mu_E\}$. Each mean can then be expressed as a combination of these basis vectors: $u_i = \sum_{j=1}^{E} \alpha_{ij} v_j$. Since the means are linearly independent, we can combine these coefficients into a single invertible matrix

$$U = \begin{bmatrix} \alpha_{11} & \alpha_{21} & \ldots & \alpha_{E1} \\ \alpha_{12} & \alpha_{22} & \ldots & \alpha_{E2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1E} & \alpha_{2E} & \ldots & \alpha_{EE} \end{bmatrix}.$$

We can then combine the constraints (11) as

$$U^T p_\alpha = \boldsymbol{\sigma} \triangleq \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_E^2 \end{bmatrix},$$

where $p_\alpha$ denotes our solution expressed in terms of the basis vectors $\{v_i\}_{i=1}^{E}$. This then has the solution

$$p_\alpha = U^{-T} \boldsymbol{\sigma}.$$

the unique maximizing solution is

$$p = \sum_{i=1}^{E} p_{\alpha i} v_i.$$

$$p_\alpha = \frac{U^{-T}\boldsymbol{\sigma}}{\|U^{-T}\boldsymbol{\sigma}\|_2}, \quad \text{with} \quad \tilde{\mu} = \frac{1}{\|U^{-T}\boldsymbol{\sigma}\|_2}.$$

# Whole proof of Theorem 5.1—lemma C.2

**Lemma C.2.** *Assume $f$ is linear. Suppose we observe $E \leq d_e$ environments whose means are linearly independent. Then there exists a linear $\Phi$ with* $\mathrm{rank}(\Phi) = d_c + d_e + 1 - E$ *whose output depends on the environmental features, yet the optimal classifier on top of $\Phi$ is invariant.*

**High level idea:** Construct a feature extractor $\Phi$ such that it relies on spurious feature while we can find an optimal classifier which is the same for all environments.

When $E = d_e$

$$\Phi = \begin{bmatrix} I & 0 \\ 0 & M \end{bmatrix} \circ f^{-1}$$

$$M = \begin{bmatrix} — & p^T & — \\ — & 0 & — \\ & \vdots & \\ — & 0 & — \end{bmatrix}.$$

$$\Phi(x) = \begin{bmatrix} z_c \\ p^T z_e \end{bmatrix},$$

$$\begin{aligned} p(y \mid z_c, \tilde{z}_e) &= \frac{p(z_c, \tilde{z}_e, y)}{p(z_c, \tilde{z}_e)} \\ &= \frac{\sigma(y \cdot \beta_c^T z_c) p(\tilde{z}_e \mid y \cdot \sigma_e^{e2} \tilde{\mu}, \sigma_e^2)}{[\sigma(y \cdot \beta_c^T z_c) p(\tilde{z}_e \mid y \cdot \sigma_e^{e2} \tilde{\mu}, \sigma_e^2) + \sigma(-y \cdot \beta_c^T z_c) p(\tilde{z}_e \mid -y \cdot \sigma_e^{e2} \tilde{\mu}, \sigma_e^2)]} \\ &= \frac{\sigma(y \cdot \beta_c^T z_c) \exp(y \cdot \tilde{z}_e \tilde{\mu})}{[\sigma(y \cdot \beta_c^T z_c) \exp(y \cdot \tilde{z}_e \tilde{\mu}) + \sigma(-y \cdot \beta_c^T z_c) \exp(-y \cdot \tilde{z}_e \tilde{\mu})]} \\ &= \frac{1}{1 + \exp(-y \cdot (\beta_c^T z_c + 2 \tilde{z}_e \tilde{\mu}))}. \end{aligned}$$

the optimal classifier is $\hat{\beta} = \begin{bmatrix} \beta_c \\ 2\tilde{\mu} \end{bmatrix}$

The risk of IRM

# Whole proof of Theorem 5.1—lemma C.2

- When $E < d_e$
- If we remove one of the environmental means, since they are linearly independent, we can simple redefine $M$ as

$$M = \begin{bmatrix} — & p^T & — \\ — & p'^T & — \\ — & 0 & — \\ & \vdots & \\ — & 0 & — \end{bmatrix}.$$

Thus, removing one environment increases the rank of $\Phi$ by 1.
Recursively, we constructed a feature extractor $\Phi$ with rank $d_c + 1 - (d_e - E)$ that relies on spurious feature.

# Whole proof of Theorem 5.1 —lemma C.3

**Lemma C.3.** *Suppose we observe $E$ environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$ whose parameters satisfy the non-degeneracy conditions (9, 10). Let $\Phi(x) = Az_c + Bz_e$ be any feature vector which is a linear function of the invariant and environmental features, and suppose the optimal $\hat{\beta}$ on top of $\Phi$ is invariant. If $E > d_e$, then $\hat{\beta} = 0$ or $B = 0$.*

**Non-degeneracy condition:**

1. $\mu_e = \sum_i \alpha_i^e \mu_i$. where $\sum_i \alpha_i^e \neq 1$,
2. $\exists \alpha^a, \alpha^b . \frac{\sigma_a^2 - \sum_i \alpha_i^a \sigma_i^2}{1 - \sum_i \alpha_i^a} \neq \frac{\sigma_b^2 - \sum_i \alpha_i^b \sigma_i^2}{1 - \sum_i \alpha_i^b}$.

*Proof.* Write $\Phi = [A|B]$ where $A \in \mathbb{R}^{d \times d_c}, B \in \mathbb{R}^{d \times d_e}$ and define

$$\bar{\mu}_e = \Phi \begin{bmatrix} \mu_c \\ \mu_e \end{bmatrix} = A\mu_c + B\mu_e,$$

$$\bar{\Sigma}_e = \Phi \begin{bmatrix} \sigma_c^2 I_{d_c} & 0 \\ 0 & \sigma_e^2 I_{d_e} \end{bmatrix} \Phi^T = \sigma_c^2 AA^T + \sigma_e^2 BB^T.$$

# Whole proof of Theorem 5.1 —lemma C.3

$$
\begin{aligned}
&\hat{\beta} = 2(\sigma_c^2 AA^T + \sigma_e^2 BB^T)^{-1}(A\mu_c + B\mu_e) \\
\iff &(\sigma_c^2 AA^T + \sigma_e^2 BB^T)\hat{\beta} = 2A\mu_c + 2B\mu_e \\
\iff &\sigma_e^2 BB^T\hat{\beta} - 2B\mu_e = 2A\mu_c - \sigma_c^2 AA^T\hat{\beta}.
\end{aligned}
$$

利用Non-degeneracy condition 1得到最终的结论$\mathbf{B} = \mathbf{0}$

# The risk of IRM in non-linear case

**Theorem D.3** (Non-linear case, full)**.** *Suppose we observe $E$ environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$. Then, for any $\epsilon > 1$, there exists a featurizer $\Phi_\epsilon$ which, combined with the ERM-optimal classifier $\hat{\beta} = [\beta_c, \beta_{e;ERM}, \beta_0]^T$, satisfies the following properties, where we define $p_{\epsilon,d_e} := \exp\{-d_e \min((\epsilon - 1), (\epsilon - 1)^2)/8\}$:*

1. *Define $\sigma^2_{\max} = \max_e \sigma^2_e$. Then the regularization term of $\Phi_\epsilon, \hat{\beta}$ is bounded as*

$$\frac{1}{E}\sum_{e\in\mathcal{E}} \|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi_\epsilon, \hat{\beta})\|_2^2 \in \mathcal{O}\left(p^2_{\epsilon,d_e}\left[\epsilon d_e \sigma^4_{\max}\exp\{2\epsilon\sigma^2_{\max}\} + \overline{\|\mu\|_2^2}\right]\right).$$

2. *$\Phi_\epsilon, \hat{\beta}$ exactly matches the optimal invariant predictor on at least a $1 - p_{\epsilon,d_e}$ fraction of the training set. On the remaining inputs, it matches the ERM-optimal solution.*

# The risk of IRM in non-linear case

*Further, for any test distribution with environmental parameters* $(\mu_{E+1}, \sigma^2_{E+1})$*, suppose the environmental mean* $\mu_{E+1}$ *is sufficiently far from the training means:*

$$\forall e \in \mathcal{E}, \min_{y \in \{\pm 1\}} \|\mu_{E+1} - y \cdot \mu_e\|_2 \geq (\sqrt{\epsilon} + \delta)\sigma_e \sqrt{d_e} \tag{15}$$

*for some* $\delta > 0$*. Define the constants:*

$$k = \min_{e \in \mathcal{E}} \frac{\sigma_e^2}{\sigma_{E+1}^2}$$

$$q = \frac{2E}{\sqrt{k\pi}\delta} \exp\{-k\delta^2\}.$$

*Then the following holds:*

3. $\Phi_\epsilon, \hat{\beta}$ *is equivalent to the ERM-optimal predictor on at least a* $1 - q$ *fraction of the test distribution.*

4. *Under Assumption 1, suppose it holds that*

$$\mu_{E+1} = -\sum_{e \in \mathcal{E}} \alpha_e \mu_e \tag{16}$$

*for some set of coefficients* $\{\alpha_e\}_{e \in \mathcal{E}}$*. Then for any* $c \in \mathbb{R}$*, so long as*

$$\sum_{e \in \mathcal{E}} \alpha_e \frac{\|\mu_e\|_2^2}{\sigma_e^2} \geq \frac{\|\mu_c\|_2^2/\sigma_c^2 + |\beta_0|/2 + c\sigma_{ERM}}{1 - \gamma}, \tag{17}$$

*the 0-1 risk of* $\Phi_\epsilon, \hat{\beta}$ *is lower bounded by* $F(2c) - q$*.*

# Proof of Theorem 6.1(D.3)

- Key idea:
  - Construct $\Phi \text{ and } \beta$ which is almost identical to the optimal invariant predictor on the training data yet behaves like the ERM solution at test time.
  - Standard concentration results
- 1.Define $\mathcal{B}$ to be the union of balls centered at each $\mu_e$ which contains most of the samples from that environment. We can bound the measure of $\mathcal{B}_c = \mathbb{R}^{d_e} \backslash \mathrm{B}$ by at most p.
- 2. Define $\Phi(\mathrm{x}) = [z_c]$ in $\mathcal{B}$ and $\Phi(\mathrm{x}) = [z_c, z_e]^T$ in $\mathcal{B}_c$. When the new environment feature $\mu_{\mathrm{E}+1}$ is far away from $\mathcal{B}$, the proof is completed since this model differs from ERM solution at test time by at most p.

# Proof of Theorem 6.1 —lemma D.1

**Lemma D.1.** *Suppose we observe environments $\mathcal{E} = \{e_1, e_2, \ldots\}$. Given a set $\mathcal{B} \subseteq \mathbb{R}^{d_e}$, consider the predictor defined by Equation 19. Then for any environment $e$, the penalty term of this predictor in Equation 5 is bounded as*

$$\|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta})\|_2^2 \leq \left\| \mathbb{P}(z_e \in \mathcal{B}^c) \mathbb{E}[|z_e| \mid z_e \in \mathcal{B}^c] \right\|_2^2.$$

**The lemma shows that since only the environmental features contribute to the gradient penalty, the penalty can be bounded as a function of the measure and geometry of that set.**

*Proof.* We write out the precise form of the gradient for an environment $e$:

$$\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta}) = \int_{\mathcal{Z}_c \times \mathcal{Z}_e} p^e(z_c, z_e) \left[ \sigma(\hat{\beta}^T \Phi(f(z_c, z_e))) - p^e(y = 1 \mid z_c, z_e) \right] \Phi(f(z_c, z_e))\, d(z_c, z_e).$$

$\downarrow$ $z_c \perp\!\!\!\perp z_e \mid y$

$$\int_{\mathcal{Z}_c \times \mathcal{B}} p^e(z_c, z_e) \left[ \sigma(\beta_c^T z_c + \beta_0) - p^e(y = 1 \mid z_c, z_e) \right] [0]\, d(z_c, z_e)$$

$$+ \int_{\mathcal{Z}_c \times \mathcal{B}^c} p^e(z_c, z_e) \left[ \sigma(\beta_c^T z_c + \beta_{e;\mathrm{ERM}}^T z_e + \beta_0) - \sigma(\beta_c^T z_c + \beta_e^T z_e + \beta_0) \right] [z_e]\, d(z_c, z_e).$$

# Proof of Theorem 6.1 —lemma D.2

**Lemma D.2.** *For a set of $E$ environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$ and any $\epsilon > 1$, construct $\mathcal{B}_r$ as in Equation 18 and define $\Phi_\epsilon$ using $\mathcal{B}_r$ as in Equation 19. Suppose we now test on a new environment with parameters $(\mu_{E+1}, \sigma^2_{E+1})$, and assume Equation 15 holds with parameter $\delta$. Define $k = \min_{e\in\mathcal{E}} \frac{\sigma_e^2}{\sigma^2_{E+1}}$. Then with probability $\geq 1 - \frac{2E}{\sqrt{k\pi\delta}} \exp\{-k\delta^2\}$ over the draw of an observation from this new environment, we have*

$$\Phi_\epsilon(x) = f^{-1}(x) = \begin{bmatrix} z_c \\ z_e \end{bmatrix}.$$

**Assumption**

***1.*** *Any test distribution with environmental parameters* $\mu_{E+1}$ is sufficiently far from *the training means:*

$$\forall e \in \mathcal{E}, \min_{y\in\{\pm 1\}} \|\mu_{E+1} - y \cdot \mu_e\|_2 \geq (\sqrt{\epsilon} + \delta)\sigma_e\sqrt{d_e}$$

***2.***

Define $r = \sqrt{\epsilon\sigma_e^2 d_e}$ and construct $\mathcal{B}_r \subset \mathbb{R}^{d_e}$ as

$$\mathcal{B}_r = \left[\bigcup_{e\in\mathcal{E}} B_r(\mu_e)\right] \cup \left[\bigcup_{e\in\mathcal{E}} B_r(-\mu_e)\right], \qquad \Phi_\epsilon(x) = \begin{cases} \begin{bmatrix} z_c \\ 0 \end{bmatrix}, & z_e \in \mathcal{B}_r \\ \begin{bmatrix} z_c \\ z_e \end{bmatrix}, & z_e \in \mathcal{B}_r^c, \end{cases} \quad \text{and} \quad \hat{\beta} = \begin{bmatrix} \beta_c \\ \hat{\beta}_e \\ \beta_0 \end{bmatrix}.$$

# Proof of Theorem 6.1 —lemma D.1

- We construct a halfspace which is perpendicular to the line connecting $\mu_e$ and $\mu_{\mathrm{E+1}}$ and tangent to $\mathcal{B}_e$. This halfspace fully contains $\mathcal{B}_e$, and the measure of this halfspace is upper bounded by:

$$1-\Phi\left(\frac{\delta\sigma_e\sqrt{d_e}}{\sqrt{\sigma_{E+1}^2 d_e}}\right) \leq \Phi\left(-\sqrt{k}\delta\right)$$
$$\leq \frac{1}{\sqrt{k\pi}\delta}\exp\{-k\delta^2\},$$

# Proof of Theorem 6.1

**Theorem D.3** (Non-linear case, full)**.** *Suppose we observe $E$ environments $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$. Then, for any $\epsilon > 1$, there exists a featurizer $\Phi_\epsilon$ which, combined with the ERM-optimal classifier $\hat{\beta} = [\beta_c, \beta_{e;ERM}, \beta_0]^T$, satisfies the following properties, where we define $p_{\epsilon,d_e} := \exp\{-d_e \min((\epsilon - 1), (\epsilon - 1)^2)/8\}$:*

1. *Define $\sigma^2_{\max} = \max_e \sigma^2_e$. Then the regularization term of $\Phi_\epsilon, \hat{\beta}$ is bounded as*

$$\frac{1}{E}\sum_{e\in\mathcal{E}} \|\nabla_{\hat{\beta}} \mathcal{R}^e(\Phi_\epsilon, \hat{\beta})\|_2^2 \in \mathcal{O}\left(p^2_{\epsilon,d_e}\left[\epsilon d_e \sigma^4_{\max} \exp\{2\epsilon\sigma^2_{\max}\} + \overline{\|\mu\|_2^2}\right]\right).$$

2. *$\Phi_\epsilon, \hat{\beta}$ exactly matches the optimal invariant predictor on at least a $1 - p_{\epsilon,d_e}$ fraction of the training set. On the remaining inputs, it matches the ERM-optimal solution.*

# Iterative Feature Matching -- Toward Provable Domain Generalization with Logarithmic Environments [Chen2021IterativeFM]

- Strengthen the theorem 5.1
  - No algorithm can approximate the optimal invariant classifier in expectation with sub-linear environment.

**Theorem 3.2.** *Suppose $E \leq d_s$. Under Assumption 3.1, there exists a constant $c > 0$ such that, for any estimator $\widehat{w}$, there exists a hyper-distribution over parameters $\mathcal{P} = P(\mu_1, \Sigma_1, \{\mu_2, \sigma_2^e\}_{e=1}^E, S)$, such that if we draw those parameters from $\mathcal{P}$ and generate data $\mathcal{E}_{tr}$ from the $E$ environments parameterized by those parameters, then*

$$\mathbb{E}_{\mathcal{P}}[\|\widehat{w}(\mathcal{E}_{tr}) - w^*(\mathcal{E}_{tr})\|_2] \geq c.$$

# Proof of Theorem 3.2

- **Key idea**:
  - construct a hard instance to reduce the problem of finding the optimal invariant predictor to $r$-dimensional Gaussian mean estimation.

Setting:

$$
\begin{aligned}
Y &\sim \text{unif}\{\pm 1\} \\
Z_1|Y &\sim N(Y \cdot \mu_1, \Sigma_1) \in \mathbb{R}^r \\
Z_2|Y &\sim N(Y \cdot \mu_2^e, \Sigma_2^e) \in \mathbb{R}^{d_s} \\
Z &= [Z_1, Z_2] \in \mathbb{R}^d \\
X &= SZ.
\end{aligned}
$$

$w^* = S_1(S_1^\top S_1)^{-1}\Sigma_1^{-1}\mu_1 / \|S_1(S_1^\top S_1)^{-1}\Sigma_1^{-1}\mu_1\|_2$

where $S_1 \in \mathbb{R}^{d\times r}$

Proof:

hard instance:

$A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{d\times d_s}$ $\quad A = [I_r; \mathbf{0}_{d_s\times r}]$

B: top $r$ rows $\{u_i\}_{i=1}^r$ satisfy $u_i^\top u_j = 0$ for $i \neq j$, and $\|u_i\| = 1$

$\Sigma_1 = I_r$

$\mu_1$ has uniform distribution on the sphere of radius $\sqrt{r}$, $\sigma_2^e \sim N(0,1)$, and $\mu_2^e \sim \mathcal{N}(0, I_{d_s})$.

$\{(A\mu_1 + B\mu_2^e, A\Sigma_1 A^\top + (\sigma_2^e)^2 BB^\top)\}_{e=1}^E$

invariant classifier:

$w^* = A\mu_1/\sqrt{r}$

# Proof of Theorem 3.2

- Reduce "find a classifier $\omega$" to a Gaussian mean estimation problem where the mean parameter is $\theta^* = \omega$

$$\min_{\widehat{\theta}(\mathcal{E}_{tr})} \max_{\mathcal{P}} \|\theta^* - \widehat{\theta}\|_2 \geq c\sqrt{\frac{d_s r}{E}},$$

$$\min_{\widehat{\theta}(\mathcal{E}_{tr})} \max_{\mathcal{P}} \|w^* - \widehat{w}\|_2 \geq c \quad \text{when } E \leq d_s$$

# Modification of the covariance of spurious feature to overcome the impossible results

**Algorithm 1** Iterative Feature Matching (IFM) algorithm

**Require:** Invariant feature dimension $r$, target feature dimensions $r_0 = d > r_1 > \cdots > r_T = r$, number of training environments $E = |\mathcal{E}_{tr}|$, infinite samples $\{(X_i^e, Y_i^e)\}_{i=1}^{\infty} \sim P_e$ from each environment $e \in \mathcal{E}_{tr}$.

1: Let $\{\mathcal{E}_t\}_{t=1}^{T}$ be a partition of $\mathcal{E}_{tr}$ such that for $t < T$, $|\mathcal{E}_t| = \tilde{\Omega}((r_{t-1} - r_t)/(r_t - r - 1))$, and $|\mathcal{E}_T| = 3$.

2: **for** $t = 1$ to $T$ **do**

3: Find orthonormal $U_t \in \mathbb{R}^{r_t \times r_{t-1}}$ and $C_t \in \mathbb{R}^{r_t \times r_t}$ such that for all $e \in \mathcal{E}_t$,

$$\mathbb{E}_{(X,Y)\sim P_e}[U_t \dots U_1 X X^\top U_1^\top \dots U_t^\top | Y] = C_t. \tag{4.1}$$

4: Return classifier $\widehat{w} = \min_{w \in \mathbb{S}^{r-1}} \frac{1}{E} \sum_{e \in [E]} \mathbb{E}_{(X,Y)\sim P_e} l(U_t \dots U_1 X, Y)$.

**Main idea:**

In each round we shrink the dimension by a constant factor using a constant number of environments. The main theoretical challenge that remains is to show that in each iteration, with high probability, only spurious features are projected out.

# Proof of Theorem 4.1

- **Key idea**:
  - To prove IFM outputs a feature extractor $U_1 \dots UT$ that does not use the spurious features, we need to show that the right $d_s$ columns of matrix $U_T$ ... $U_1 S$ are all-zero.

**Lemma 5.1.** *If for all* $1 \leq t < T$, $|\mathcal{E}_t| = E_t = \Omega\left(\frac{r_{t-1}-r_t}{r_t-1} \max\left\{1, \log\left(\frac{D}{(r_t-1)d_s}\right), \log\left(\frac{d_s}{r_t-1}\right)\right\}\right)$, *and* $E_T \geq 3$, *and* $U_1, \dots, U_T$ *are the orthonormal matrices returned by IFM, then with probability* $1 - \exp(-\Omega(d_s))$, *if we write* $U_T \dots U_1 S = [A, B]$, *where* $B \in \mathbb{R}^{r \times d_s}$, *then* $B = \mathbf{0}_{r \times d_s}$.

- Since U is orthonormal, IFM outputs $\hat{\omega} = \omega^*$

# Key proof of Lemma 5.1

- Step1 with high probability, any one-layer feature extractor $U_1 \in \mathbb{R}^{k_1 \times d_s}$ that uses only spurious dimensions cannot match feature covariances from $\tilde{\Omega}(d_s/k_1)$ environments.
- $\Rightarrow$ feature extractor use at most $d_s/E_1$ spurious dimensions.
- Step2 We can recursively apply this argument until we have 0 spurious dimensions.

# Verify the advantages brought by IFM

- **ERM and IRM still suffer from linear environment complexity.**
- 1.They construct a hard instance where the ERM solution has worse-than-random performance on the test environments.
- 2.They prove there exists a feature extractor that only uses spurious dimensions while satisfies IRM penalty.

# Review

- 1.What is IRM
- 2.The difference between linear classification tasks and linear regression tasks under invariant principle
- 3.The failure case of IRM (Linear and non-linear)
- 4.What is IFM

# Takeaway message

- 1.How to find the invariant feature is still an open question.
  - IRM fails
  - IFM is a provable algorithm with Logarithmic Environments
- 2.There is some flaws with IFM
  - IFM still belongs to distribution matching(DM) methods
  - DM methods fails when the label distribution differs among environments

# Reference for IRM


- Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant Risk Minimization. *ArXiv, abs/1907.02893*.
- Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I., & Rish, I. (2021). Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. *ArXiv, abs/2106.06607*.
- Rosenfeld, E., Ravikumar, P., & Risteski, A. (2021). The Risks of Invariant Risk Minimization. *ICLR 2021*
- Chen, Y., Rosenfeld, E., Sellke, M., Ma, T., & Risteski, A. (2021). Iterative Feature Matching: Toward Provable Domain Generalization with Logarithmic Environments. *ArXiv, abs/2106.09913*.

# Reference for generalization bound


- Albuquerque, I., Monteiro, J.L., Bayazi, M.J., Falk, T., & Mitliagkas, I. (2019). Generalizing to unseen domains via distribution matching. *arXiv*
- Blanchard, G., Deshmukh, A., Dogan, Ü., Lee, G., & Scott, C. (2021). Domain Generalization by Marginal Transfer Learning. *J. Mach. Learn. Res., 22*, 2:1-2:55.
- Zhao, H., Combes, R.T., Zhang, K., & Gordon, G.J. (2019). On Learning Invariant Representations for Domain Adaptation. *ICML*.
- Ye, H., Xie, C., Cai, T., Li, R., Li, Z., & Wang, L. (2021). Towards a Theoretical Framework of Out-of-Distribution Generalization. *ArXiv, abs/2106.04496*.
- Federici, M., Tomioka, R., & Forr'e, P. (2021). An Information-theoretic Approach to Distribution Shifts. *ArXiv, abs/2106.03783*.