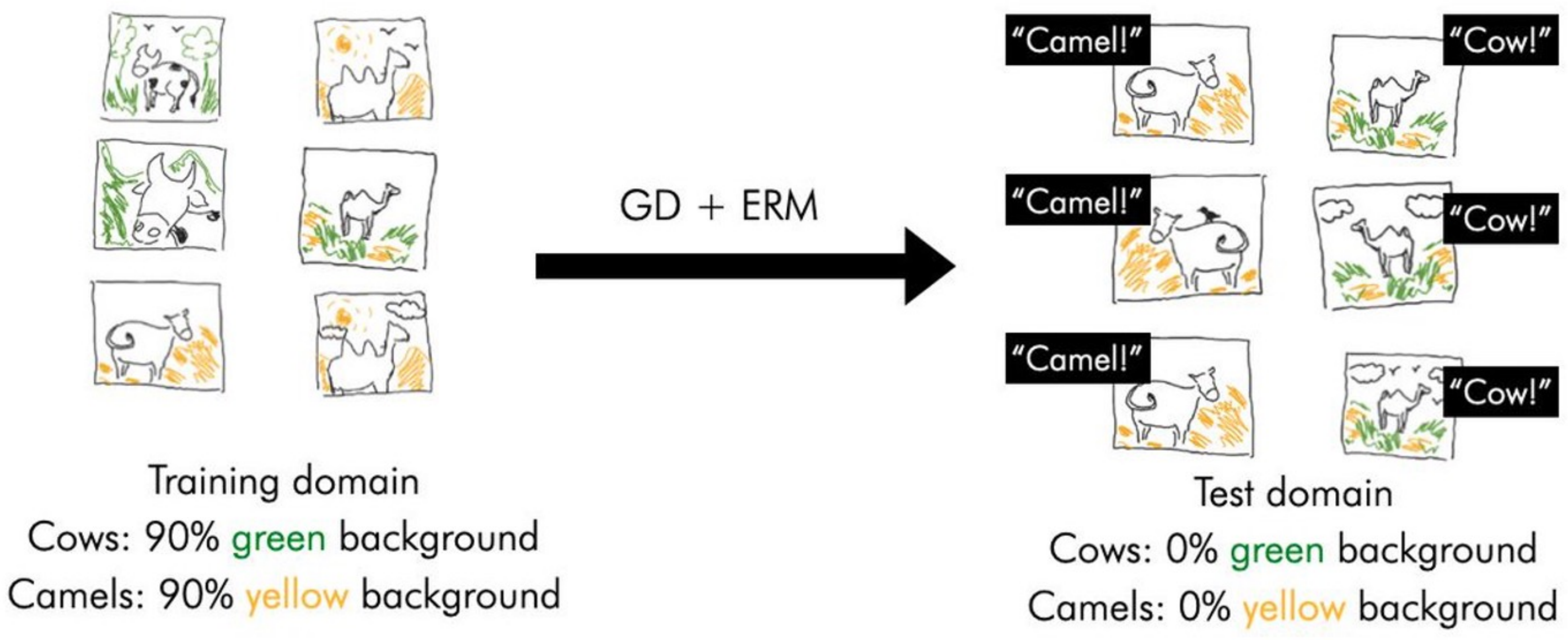
Pareto Invariant Risk Minimization: Towards Mitigating the Optimization Dilemma in OOD Generalization

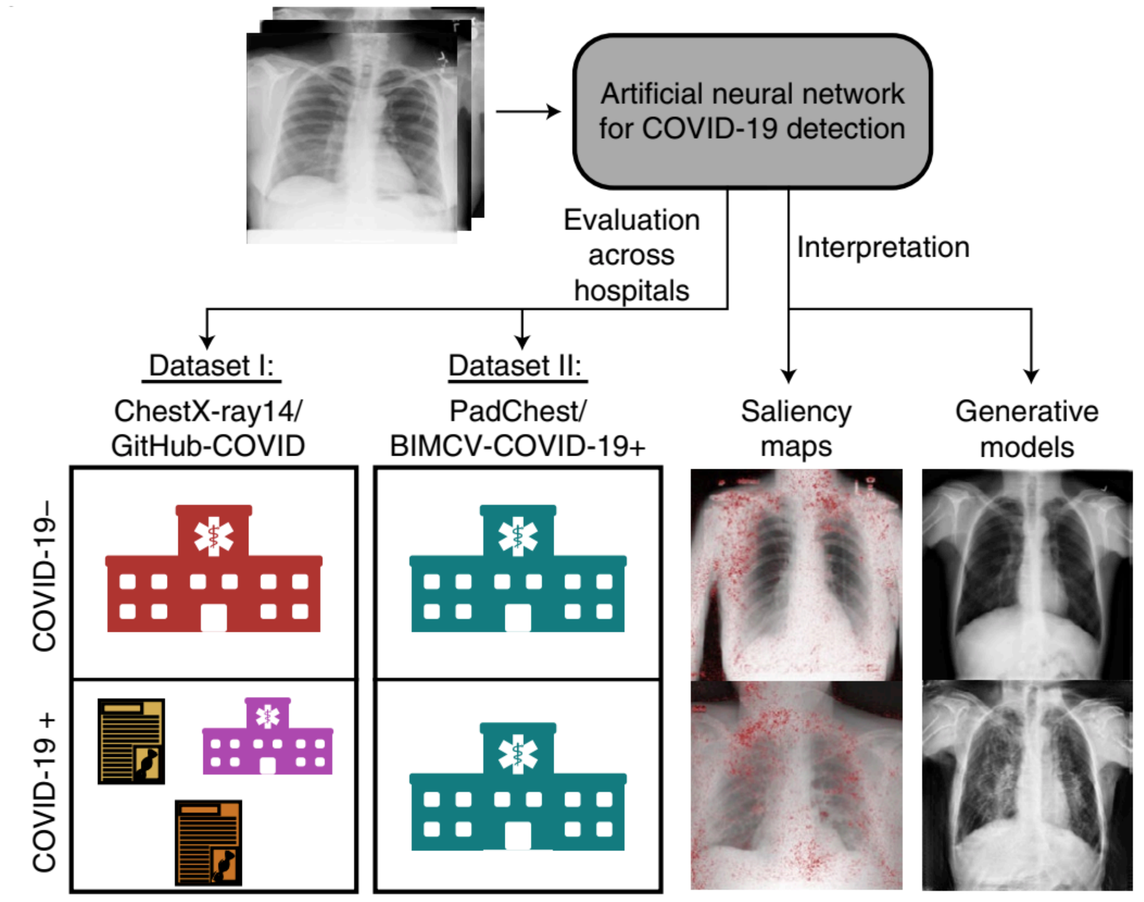
Yongqiang Chen

CUHK & Tencent AI Lab

*with Kaiwen Zhou, Yatao Bian, Binghui Xie,
Bingzhe Wu, Peilin Zhao, Bo Han, James Cheng and others.*

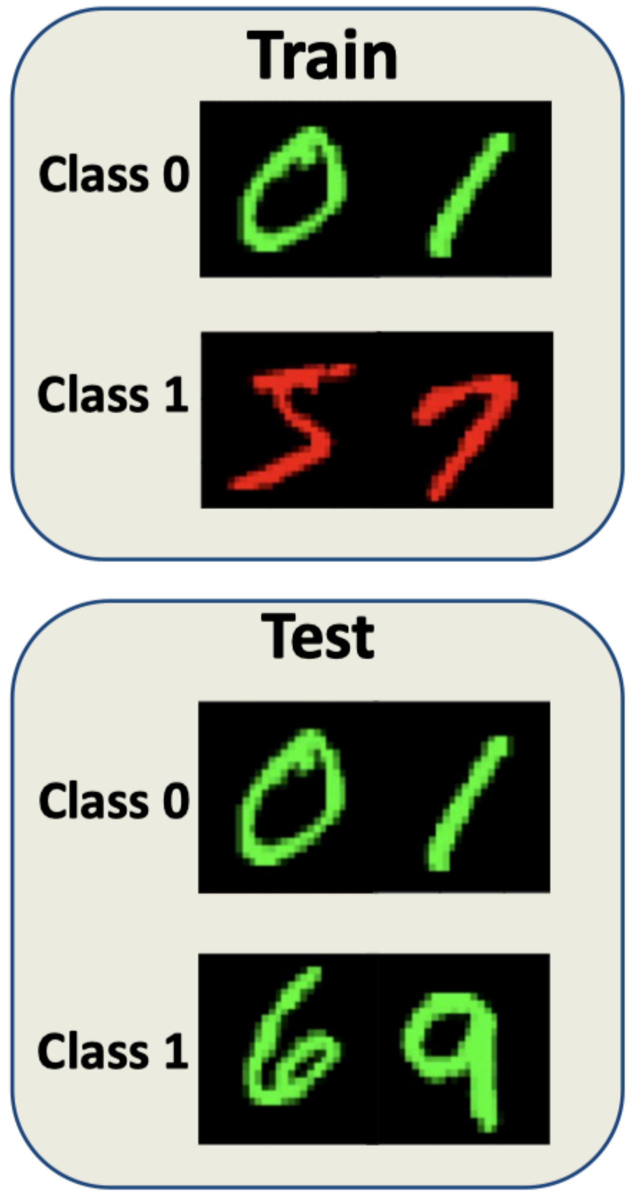
Out-of-Distribution generalization





b

	Dataset I			Dataset II		
	Combined	Chest-X-ray14	GitHub-COVID	Combined	PadChest	BIMCV-COVID-19+
No. radiographs	112,528	112,120	408	97,866	96,270	1,596
No. patients	31,067	30,805	262	64,954	63,939	1,105
% COVID-19+	0.2	0	76.5	1.6	0	100
% AP images	39.9	40	26	5.6	4.7	58.1

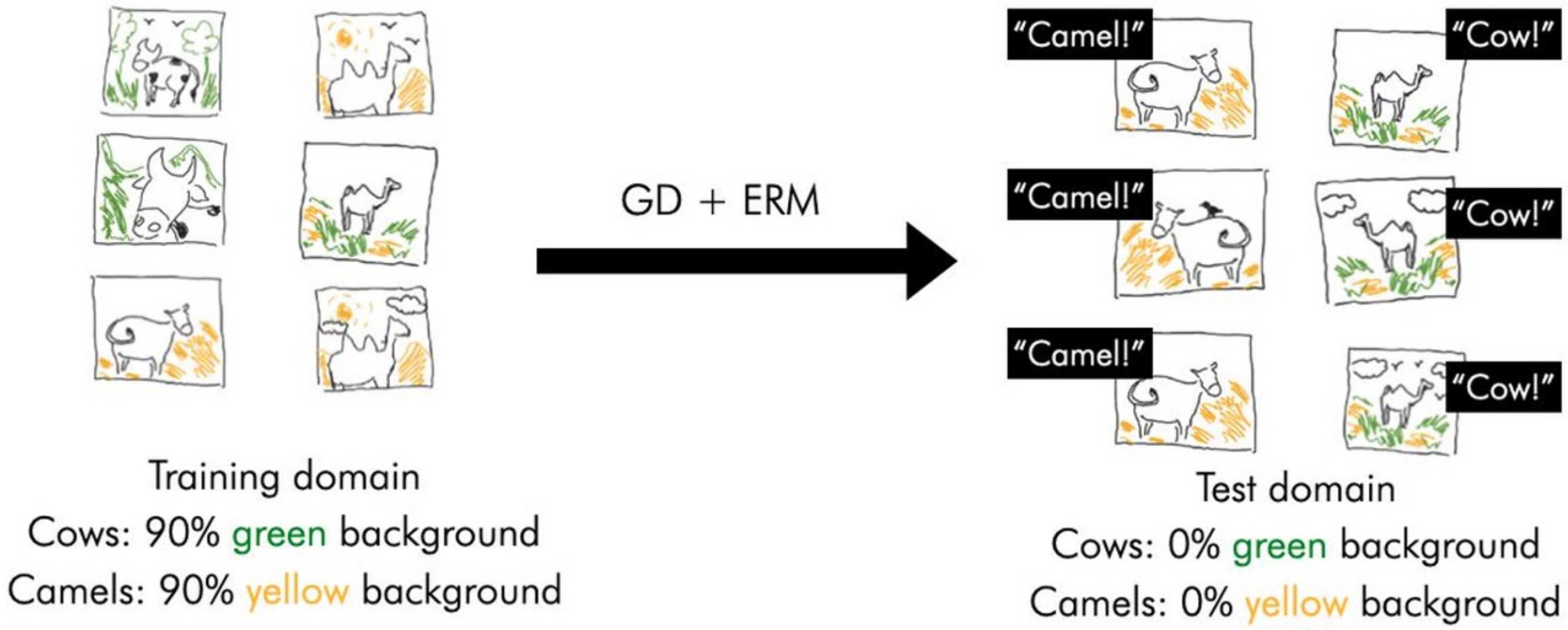


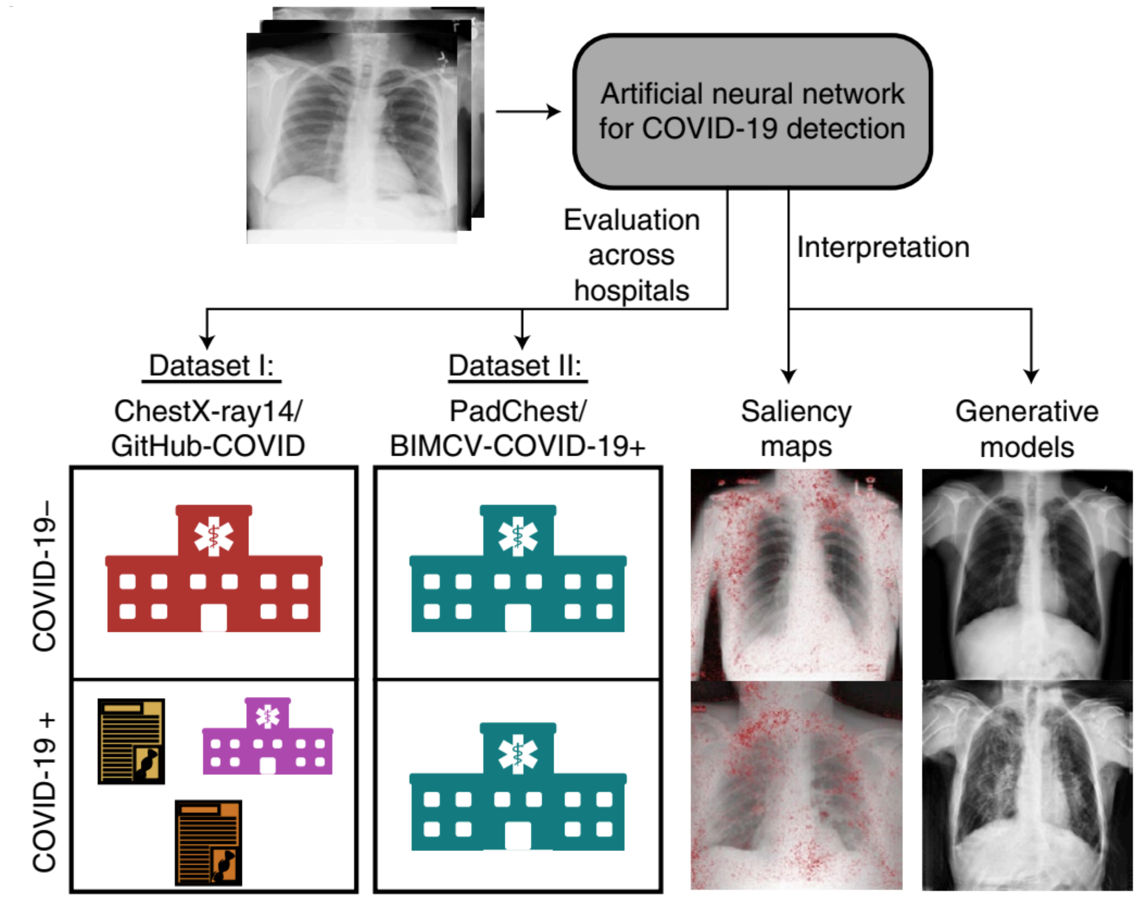
(Beery et al., 2018; Arjovsky et al., 2019; DeGrave et al. 2021; Ahuja et al., 2021; Zhang et al., 2022)

Models learned with Empirical Risk Minimization (ERM) are often:

- prone to **spurious correlations**
- can hardly generalize to **OOD** data

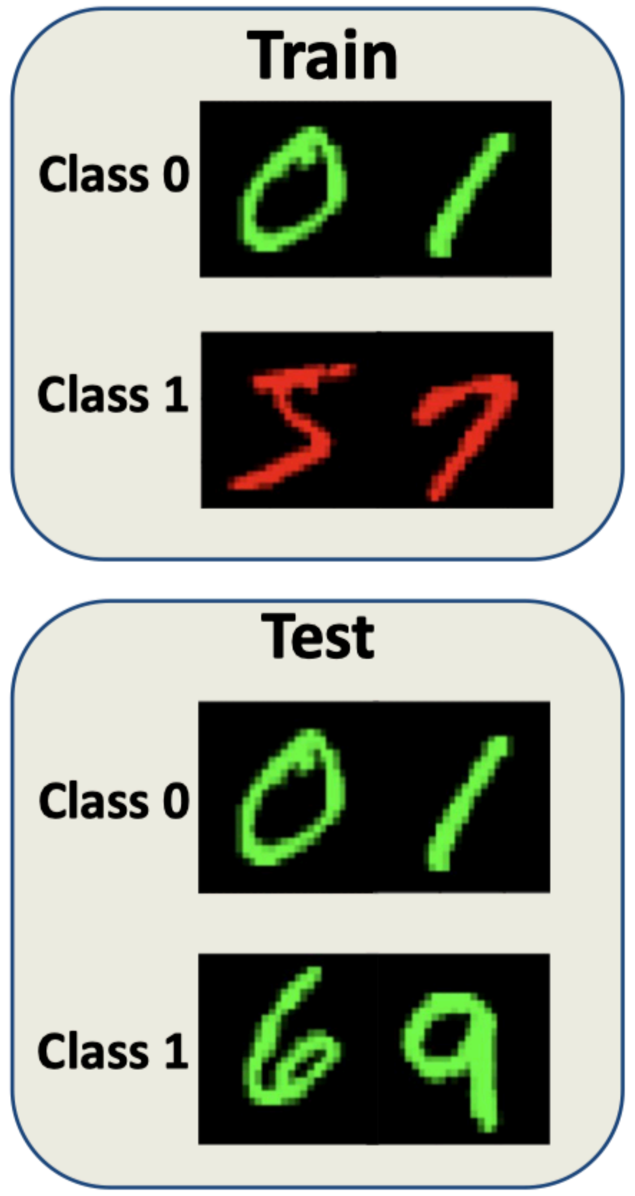
Out-of-Distribution generalization





b

	Dataset I			Dataset II		
	Combined	Chest-X-ray14	GitHub-COVID	Combined	PadChest	BIMCV-COVID-19+
No. radiographs	112,528	112,120	408	97,866	96,270	1,596
No. patients	31,067	30,805	262	64,954	63,939	1,105
% COVID-19+	0.2	0	76.5	1.6	0	100
% AP images	39.9	40	26	5.6	4.7	58.1



(Beery et al., 2018; Arjovsky et al., 2019; DeGrave et al. 2021; Ahuja et al., 2021; Zhang et al., 2022)

The goal of OOD generalization:

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \max_{e\in\mathcal{E}_{\text{all}}} \mathcal{L}_e(f)$$

given a subset of training **environments**/domains $\mathcal{E}_{\text{tr}} \subseteq \mathcal{E}_{\text{all}}$,

where each $e \in \mathcal{E}$ corresponds to a dataset $\mathcal{D}_e$ and a loss $\mathcal{L}_e$.

Previous works focus on OOD objectives

Previous works mostly focus on developing better ***optimization objectives***:

$$\min_f L_{\mathrm{ERM}} + \lambda \cdot \widehat{L}_{\mathrm{OOD}}$$

Regularization via some OOD objective

(Arjovsky et al., 2019; Krueger et al., 2021; Rame et al., 2021; Pezeshki et al., 2021; Ahuja et al., 2021; Zhang et al., 2022)

The Optimization Dilemma in OOD Generalization

Previous works mostly focus on developing better ***optimization objectives***:

$$\min_f L_{\text{ERM}} + \lambda \cdot \widehat{L}_{\text{OOD}}$$

Regularization via some ***relaxed*** OOD objective

IRM

$$\min_{f=w\circ\varphi} \sum_{e\in\mathscr{E}_{\text{tr}}} \mathscr{L}_e(w\circ\varphi),$$

$$\text{s.t. } w \in \arg\min_{\bar{w}} \mathscr{L}_e(\bar{w}\circ\varphi),\ \forall e \in \mathscr{E}_{\text{tr}}$$

Linearized IRM with $w \in \mathbb{R}^d$

IRM$_\mathcal{S}$

$$\min_{\varphi} \sum_{e\in\mathscr{E}_{\text{tr}}} \mathscr{L}_e(\varphi),$$

$$\text{s.t. } \nabla_{w|w=1}\mathscr{L}_e(w\cdot\varphi) = 0,\ \forall e \in \mathscr{E}_{\text{tr}}$$

Soften the constraints

IRMv1

$$\min_{\varphi} \sum_{e\in\mathscr{E}_{\text{tr}}} \mathscr{L}_e(\varphi) + \lambda \|\nabla_{w|w=1}\mathscr{L}_e(w\cdot\varphi)\|^2$$

(Arjovsky et al., 2019; Kamath et al., 2021)

The Optimization Dilemma in OOD Generalization

🥲 The practical variants of IRM can have very different behaviors from the original IRM.

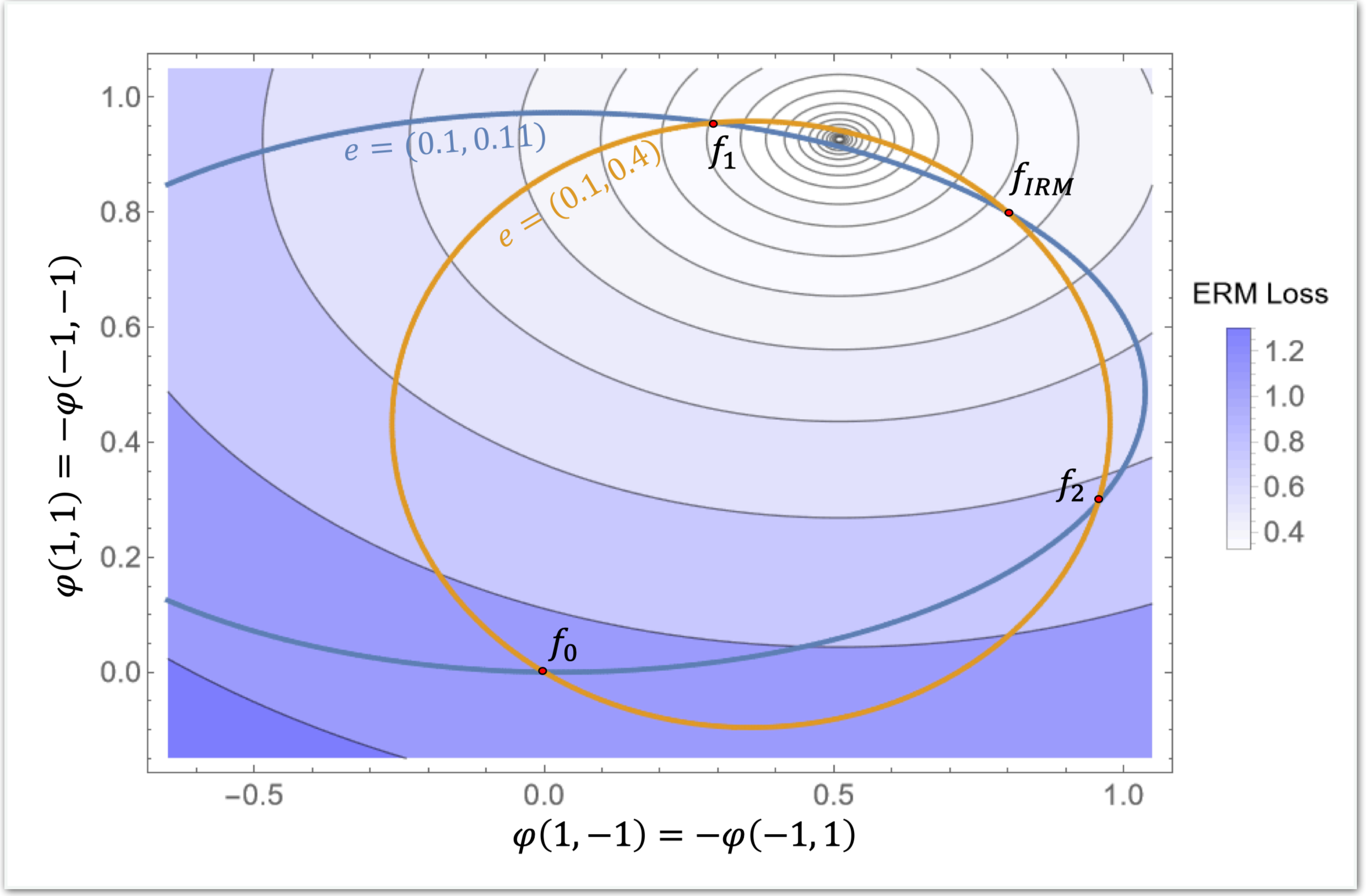


Illustration of IRMv1 failures

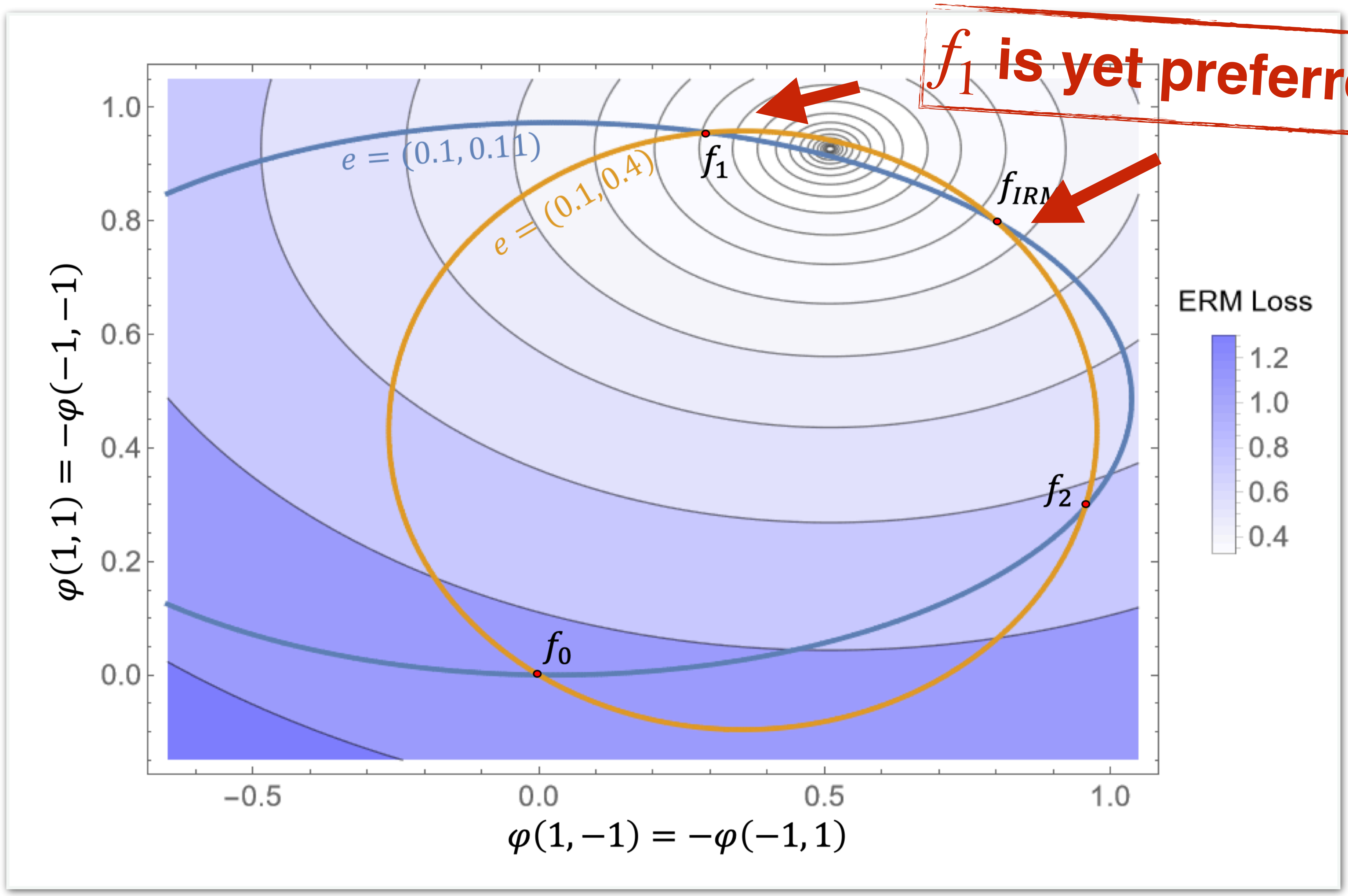
The ellipsoids are the solutions satisfying the **invariant constraints** in $\text{IRM}_{\mathcal{S}}$

$$\nabla_{w|w=1}\mathcal{L}_e(w \cdot \varphi) = 0, \forall e \in \mathcal{E}_{\text{tr}}$$

(Arjovsky et al., 2019; Kamath et al., 2021)

Invariant Risk Minimization in practice

The practical variants of IRM can have very different behaviors from the original IRM.



f_1 **is yet preferred than** f_{IRM}

Illustration of IRMv1 failures

The ellipsoids are the solutions satisfying the **invariant constraints** in $\text{IRM}_{\mathcal{S}}$

$$\nabla_{w|w=1}\mathcal{L}_e(w \cdot \varphi) = 0, \forall e \in \mathcal{E}_{\text{tr}}$$

(Arjovsky et al., 2019; Kamath et al., 2021)

The Optimization Dilemma in OOD Generalization

Previous works mostly focus on developing better ***optimization objectives***:

$$\min_f L_{\text{ERM}} + \lambda \cdot \widehat{L}_{\text{OOD}}$$

λ is ***hard to tune***

(Arjovsky et al., 2019; Krueger et al., 2021; Rame et al., 2021; Pezeshki et al., 2021; Ahuja et al., 2021; Zhang et al., 2022)

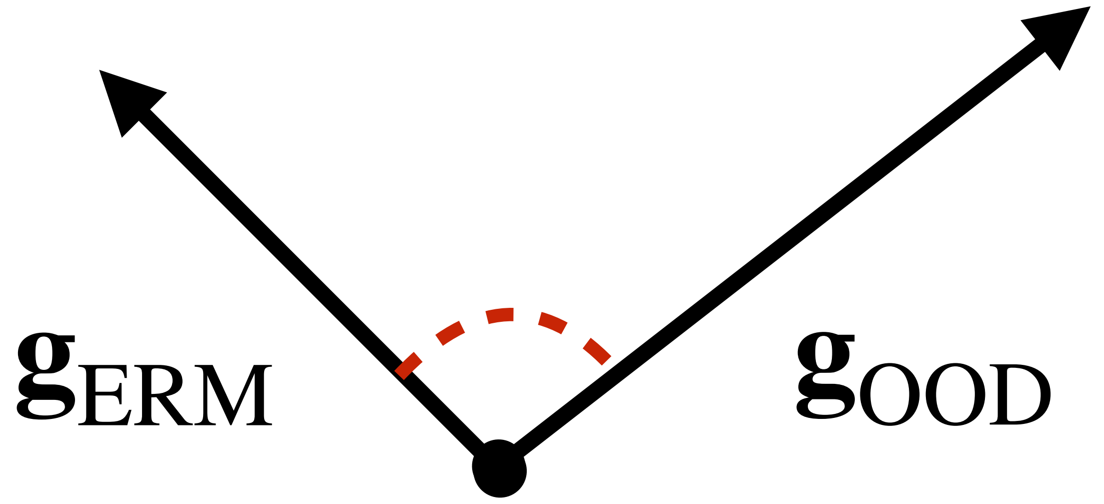
The Optimization Dilemma in OOD Generalization

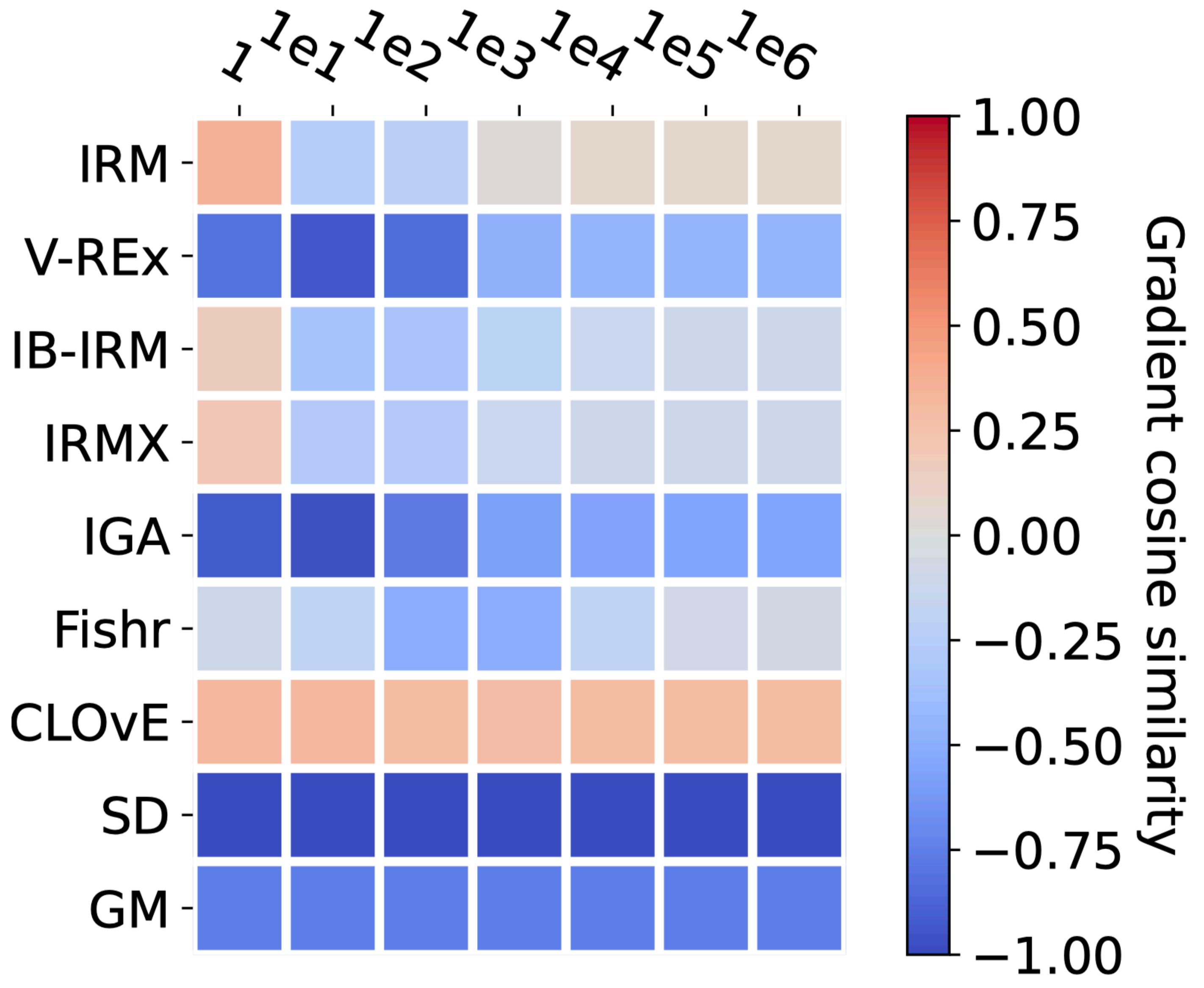
Previous works mostly focus on developing better ***optimization objectives***:

$$\min_f L_{\text{ERM}} + \lambda \cdot \widehat{L}_{\text{OOD}}$$

λ is ***hard to tune***

Gradient Conflicts generically exist between ERM and OOD objectives:





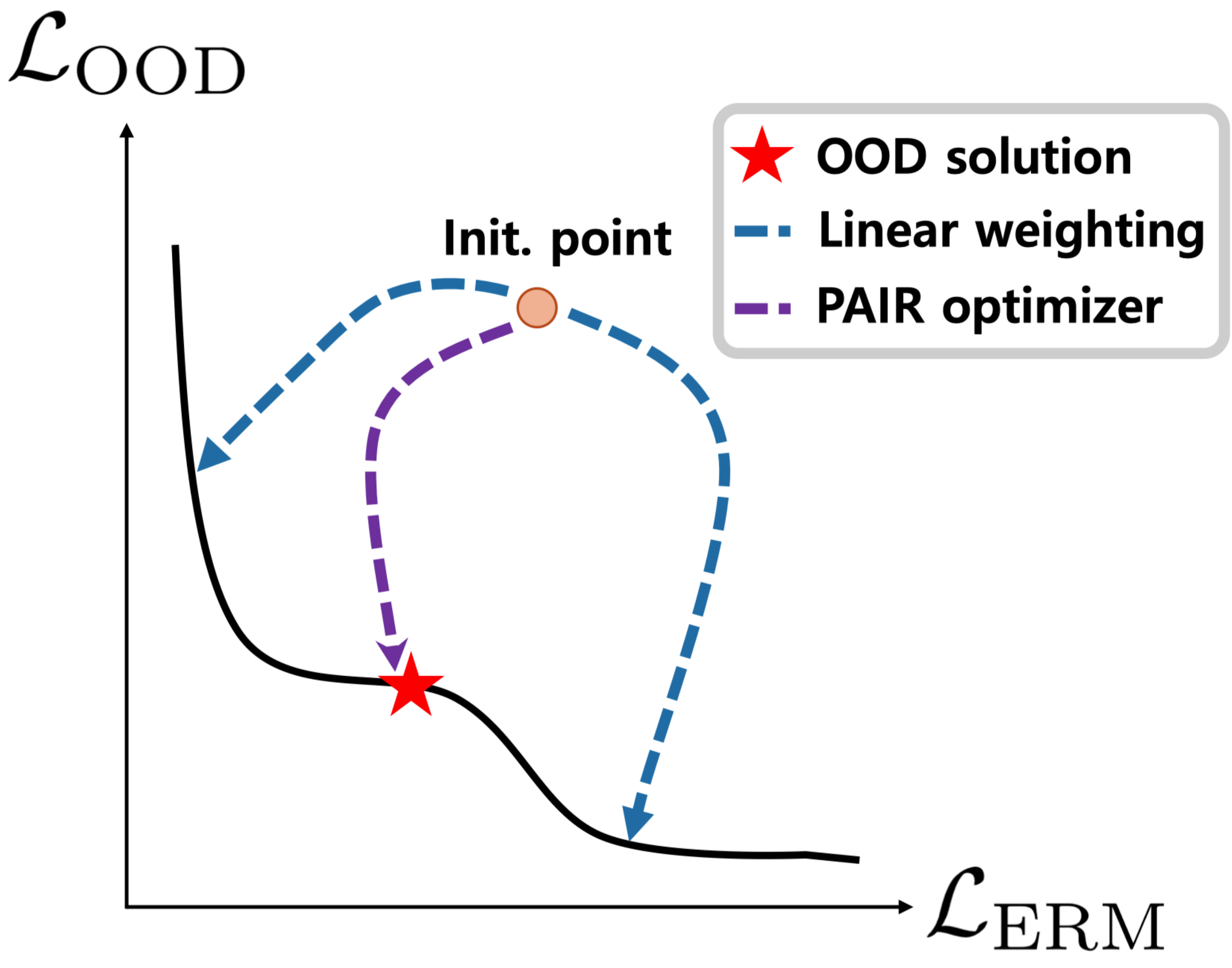
(Arjovsky et al., 2019; Krueger et al., 2021; Rame et al., 2021; Pezeshki et al., 2021; Ahuja et al., 2021; Zhang et al., 2022)

The Optimization Dilemma in OOD Generalization

The typically used linear weighting scheme cannot reach ***non-convex part of pareto front solutions***

$$\min_f L_{\mathrm{ERM}} + \lambda \cdot \widehat{L}_{\mathrm{OOD}}$$

The linear weight scheme

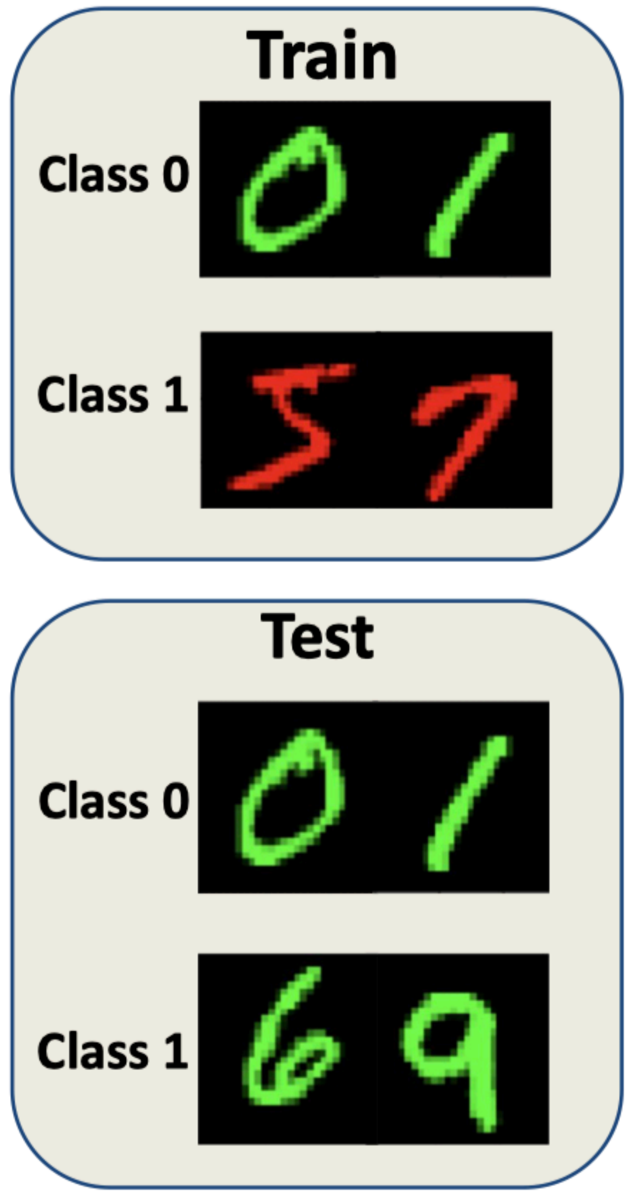


(Boyd & Vandenberghe, 2014)

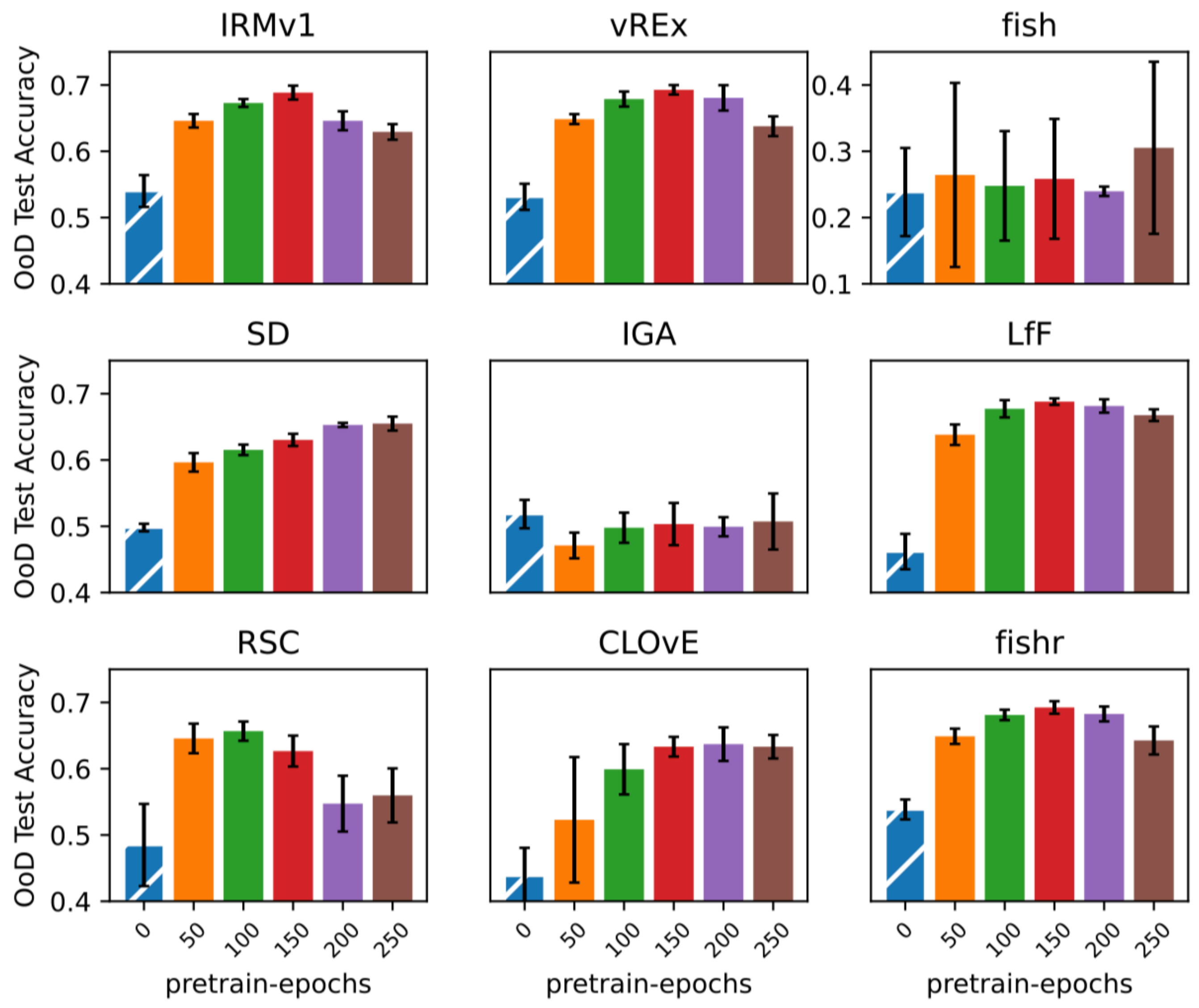
The Optimization Dilemma in OOD Generalization

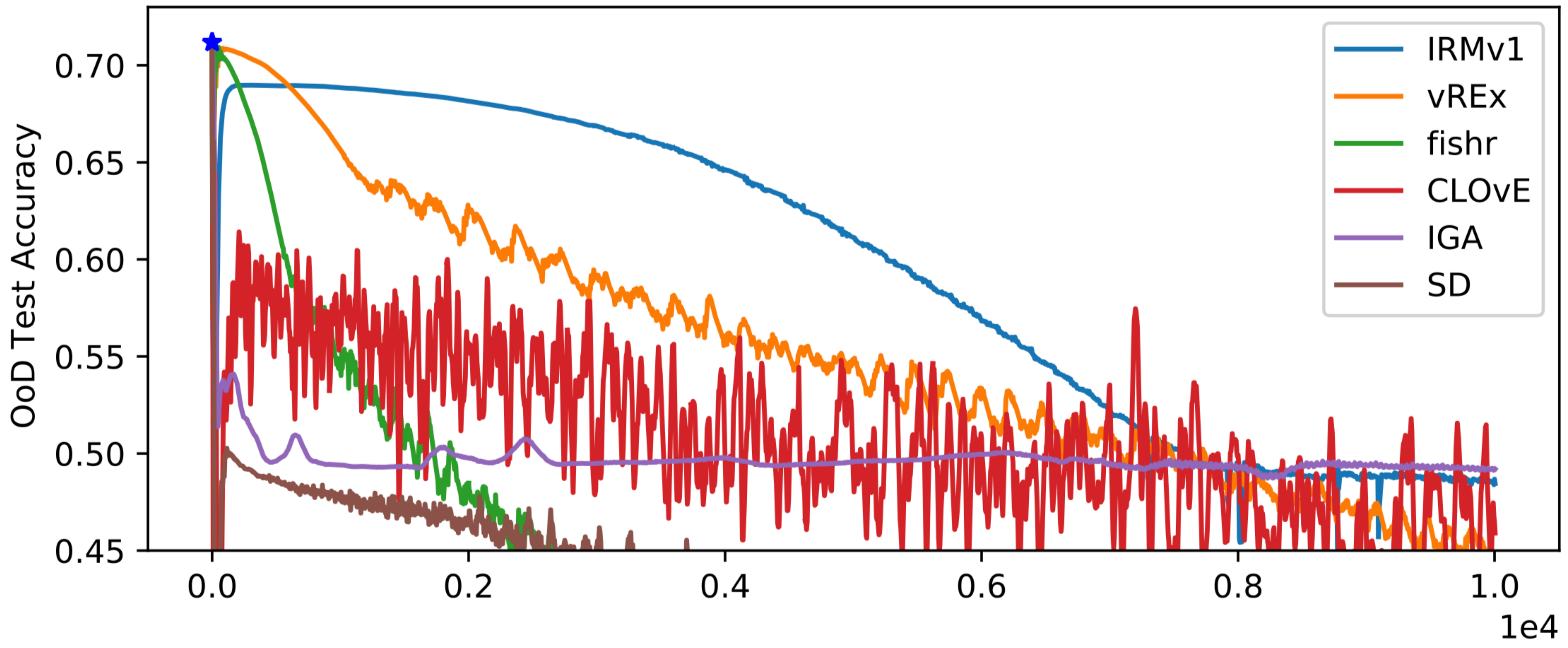
Even the desired solution is reachable, the scheme requires ***exhaustive hyperparemter tuning***:

$$\min_f L_{\text{ERM}} + \lambda \cdot \widehat{L}_{\text{OOD}}$$



λ is ***too strong*** to learn the correlation; λ is ***too weak*** to keep the invariance





(Arjovsky et al., 2019; Krueger et al., 2021; Rame et al., 2021; Pezeshki et al., 2021; Ahuja et al., 2021; Zhang et al., 2022)

The Optimization Dilemma in OOD Generalization

The usual optimization formula of OOD objectives in practice:

$$\min_f L_{\text{ERM}} + \lambda \cdot \widehat{L}_{\text{OOD}}$$

λ is ***hard to tune*** Regularization via some ***relaxed*** OOD objective

- $\widehat{L}_{\text{OOD}}$ usually has ***a large gap*** from the original one;
- λ is ***hard to tune***, i.e.,
  - Some solutions are unreachable with linear weight scheme;
  - Even reachable, it still requires exhaustive tuning efforts to find a proper λ;

(Arjovsky et al., 2019; Krueger et al., 2021; Rame et al., 2021; Pezeshki et al., 2021; Ahuja et al., 2021; Zhang et al., 2022)

As the traditional optimization scheme fails

How to obtain a desired OOD solution under the ERM and OOD conflicts?

From a Multi-Objective Optimization perspective…

The optimization of IRM essentially handles the ***trade-off*** between

$$\min_f L_{\text{ERM}} + \lambda \cdot \widehat{L}_{\text{OOD}}$$

Capturing the statistical correlations

Enforcing the invariance of learned correlations

From a Multi-Objective Optimization perspective…

The optimization of IRM essentially handles the ***trade-off*** between

$$\min_f L_{\text{ERM}} + \lambda \cdot \widehat{L}_{\text{OOD}}$$

⬇

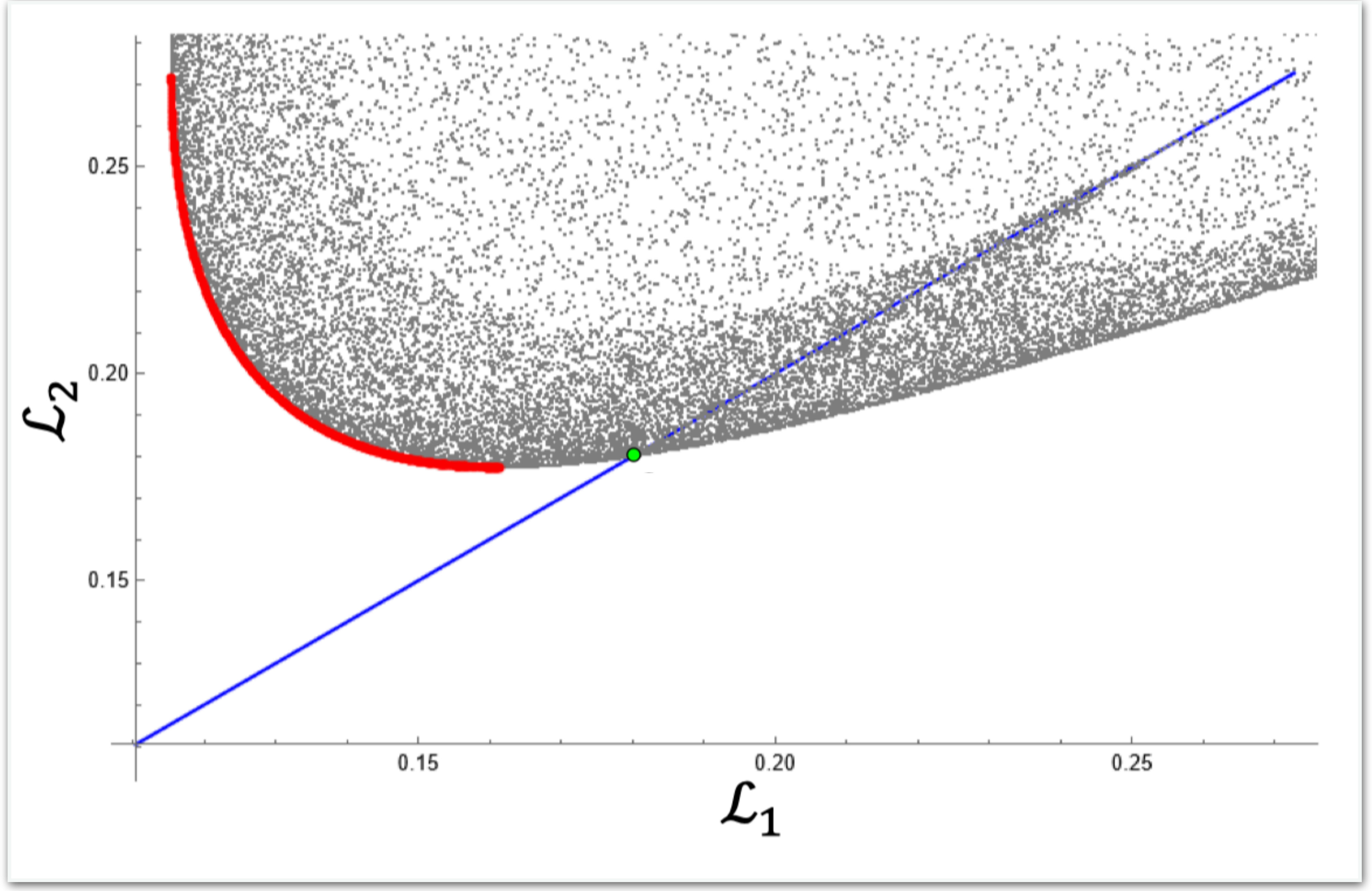
$$\min_f \{L_{\text{ERM}}, \quad \widehat{L}_{\text{OOD}}\}^T$$

Oh, it's a Multi-Objective Optimization (MOO)!

From a Multi-Objective Optimization perspective…

Assume we have the Multi-Objective Optimization (MOO) problem with 2 objectives:

$$\min_{f=w\cdot\varphi} \{L_1, L_2\}^T$$



Simulated Pareto front

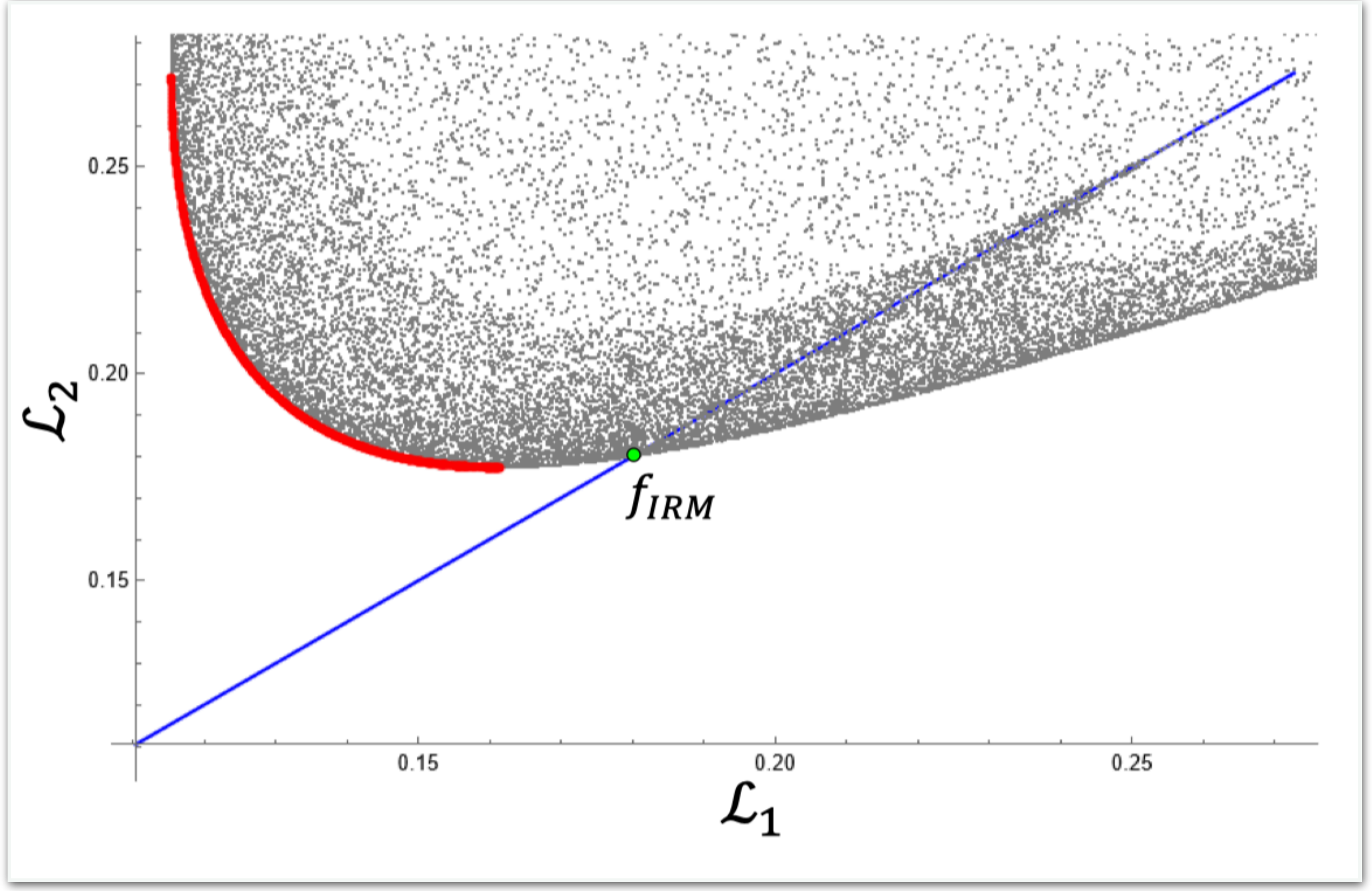
- A solution f (with $\{L_1, L_2\}^T$) **dominates** $\bar{f}$ (with $\{\bar{L}_1, \bar{L}_2\}^T$) if both $L_1 \leq \bar{L}_1$ and $L_2 \leq \bar{L}_2$;
- **Pareto optimal solutions** are the set of solutions dominated by none;
- Their images form the **Pareto front**;

(Kaisa, 1999)

From a Multi-Objective Optimization perspective…

Assume we have 2 training environments, a natural MOO formulation of IRMv1 is:

$$\min_{f=w\cdot\varphi} \{L_1, L_2, L_{\text{IRM}}\}^T$$



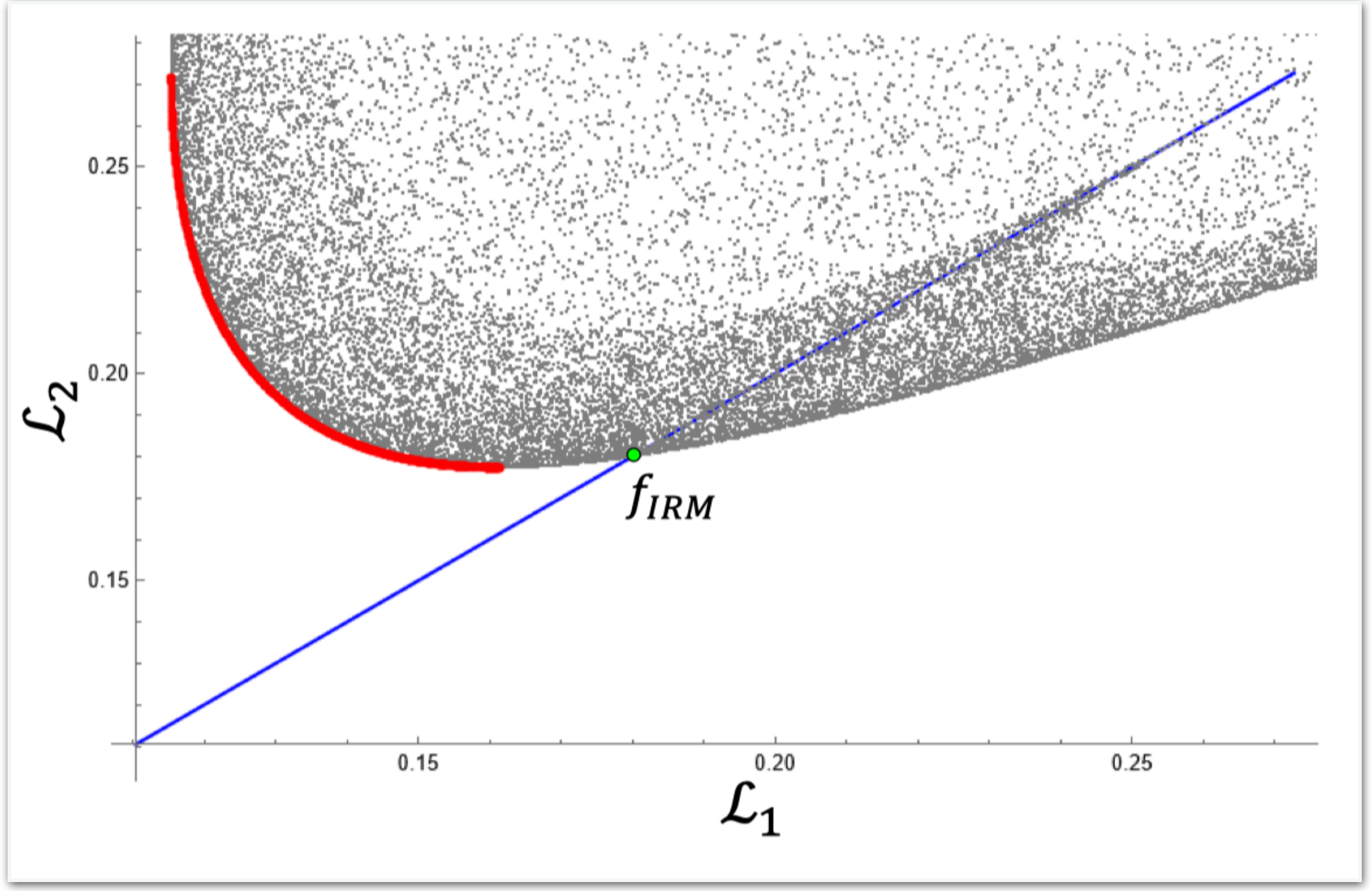
Simulated Pareto front

From a Multi-Objective Optimization perspective…

Observation I: Merely minimizing any environment-reweighted ERM cannot locate the f_{IRM};

Observation II: …

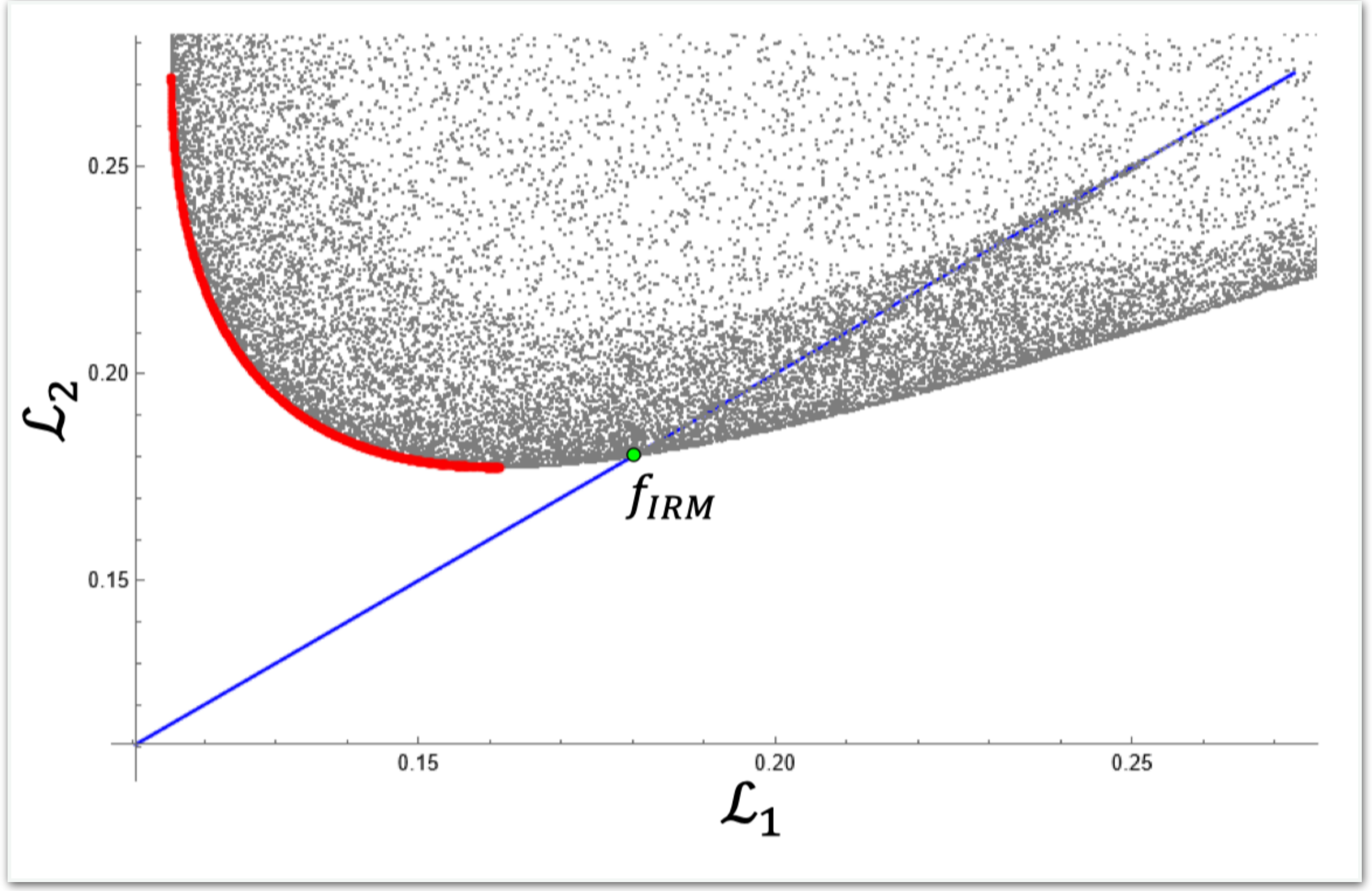
Observation III: …



Simulated Pareto front

From a Multi-Objective Optimization perspective…

👉 Observation I: Merely minimizing any environment-reweighted ERM cannot locate the f_{IRM};
Observation II: Incorporating the additional practical IRM penalty cannot locate the f_{IRM};
Observation III: …



Simulated Pareto front

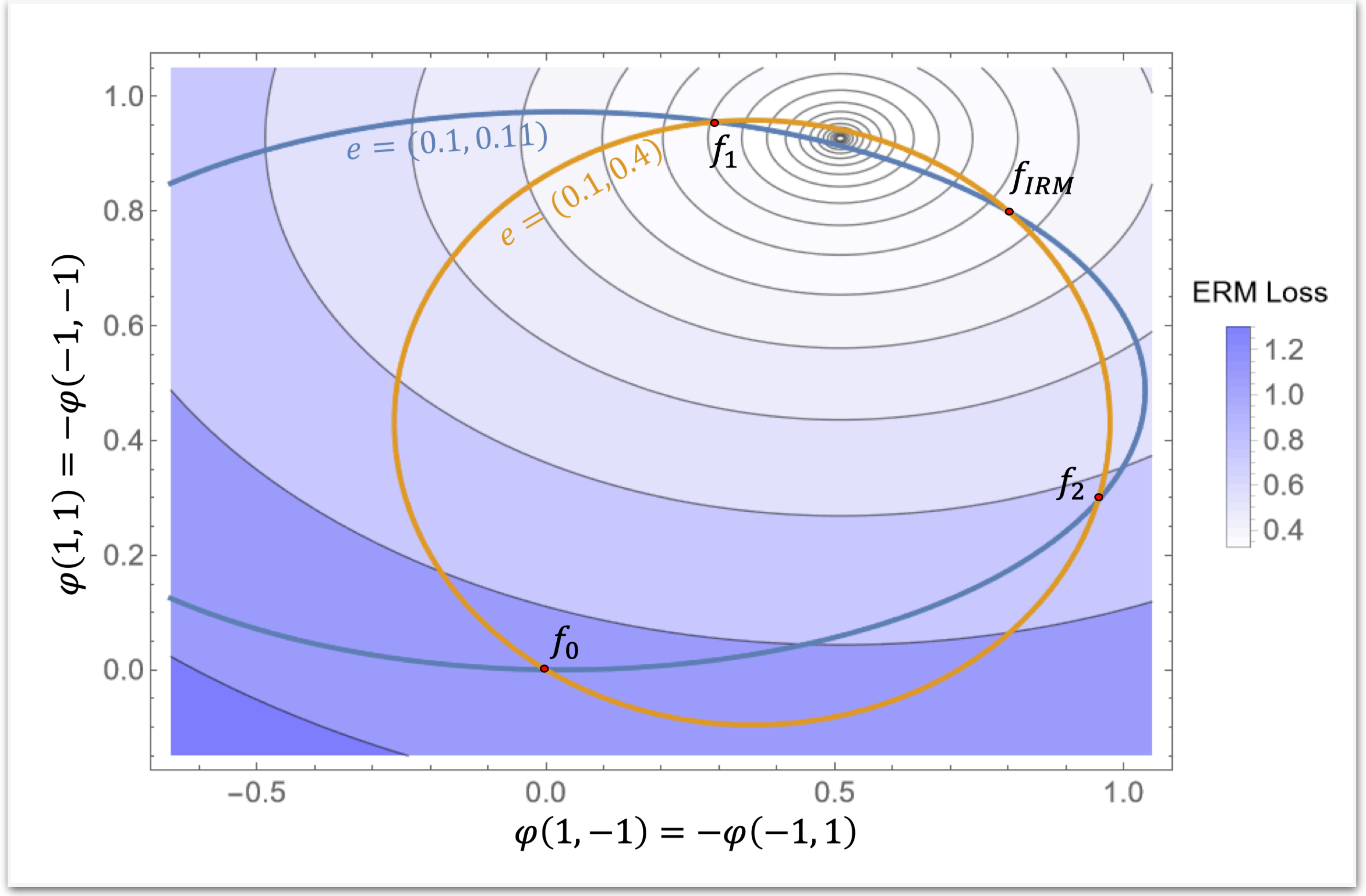
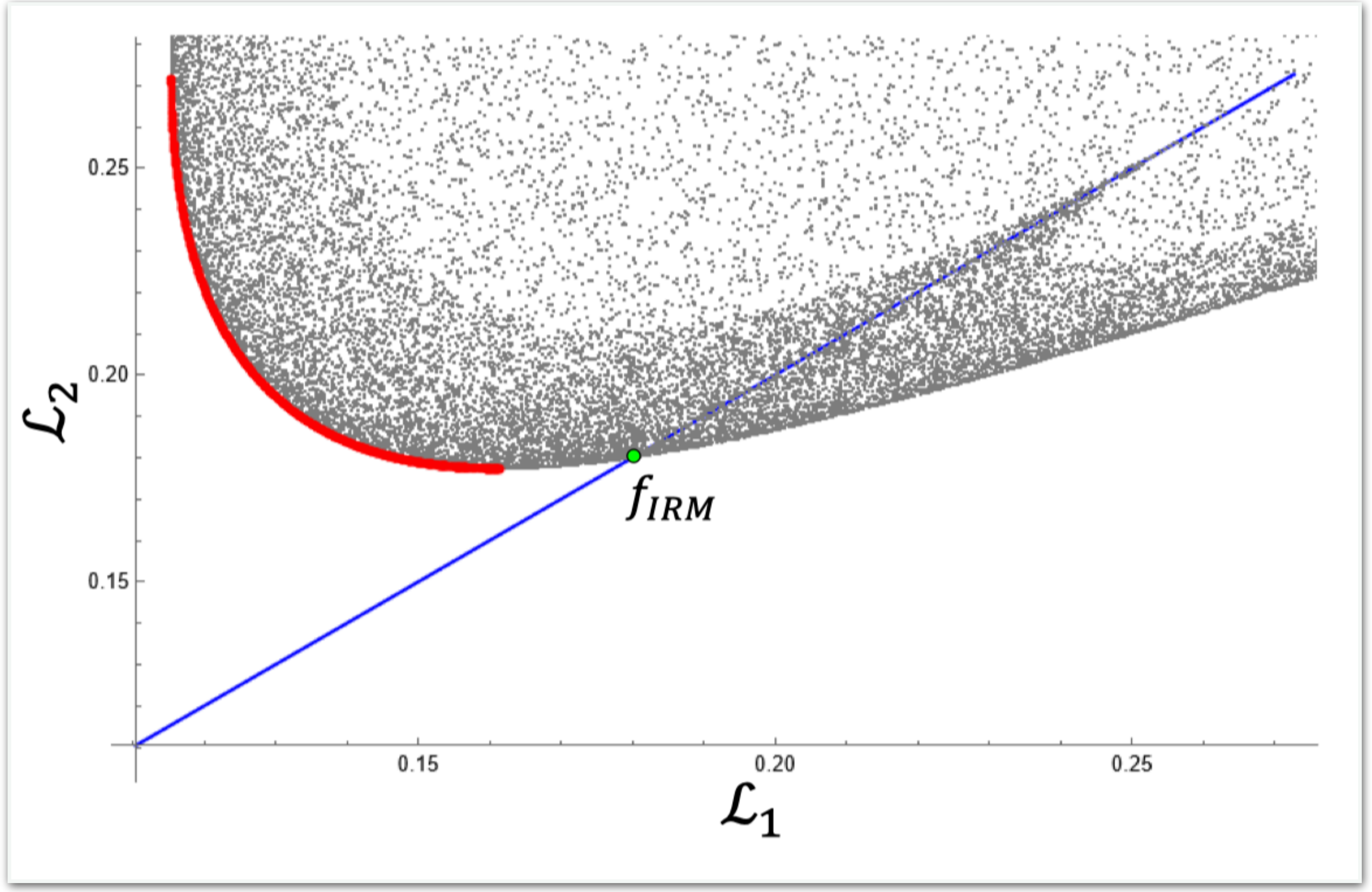


Illustration of IRMv1 failures

From a Multi-Objective Optimization perspective…

Observation I: Merely minimizing any environment-reweighted ERM cannot locate the f_{IRM};
Observation II: Incorporating the additional practical IRM penalty cannot locate the f_{IRM};
👉 Observation III: The failures of practical IRM variants is because of using bad objectives!



Simulated Pareto front

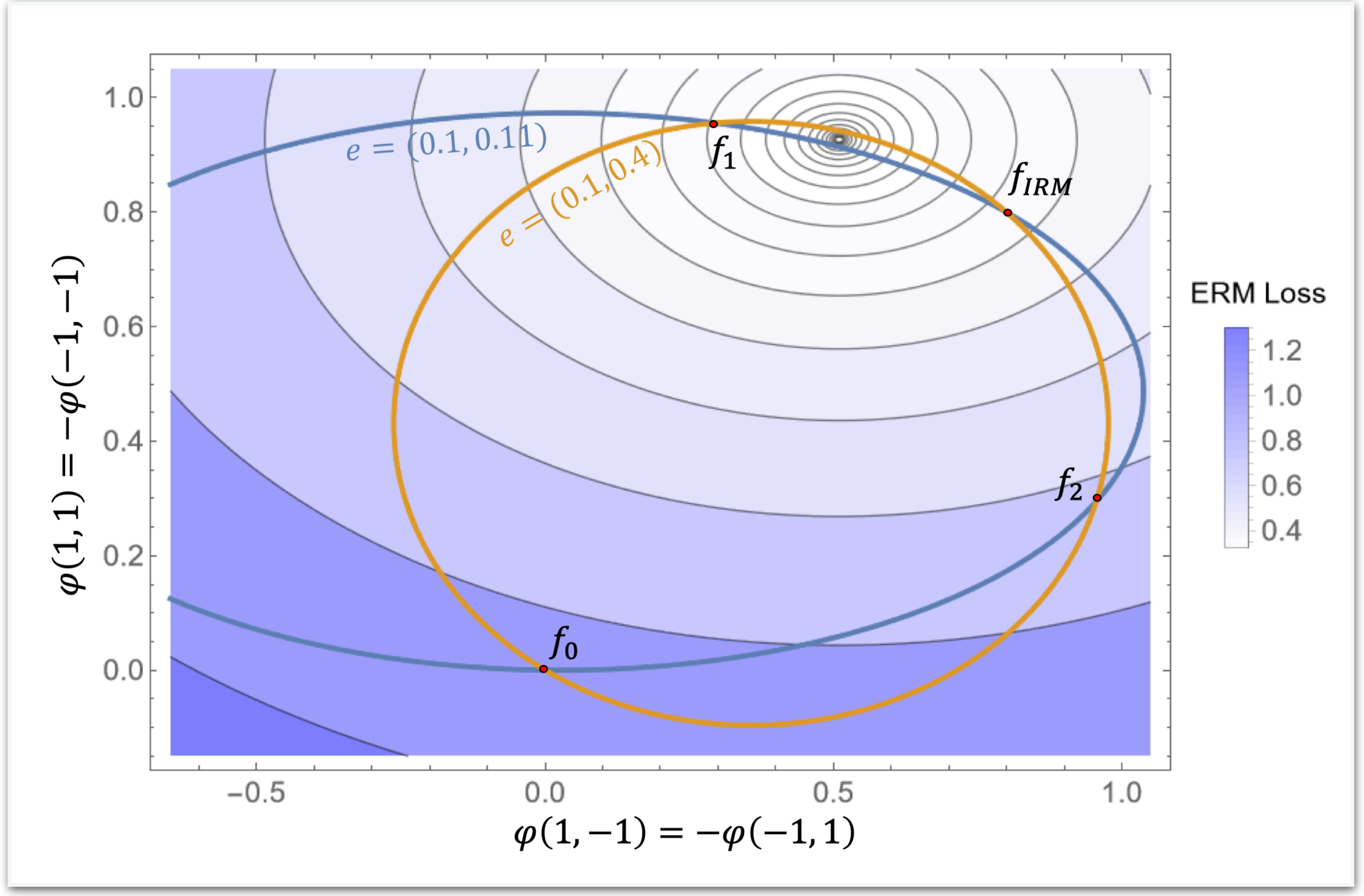
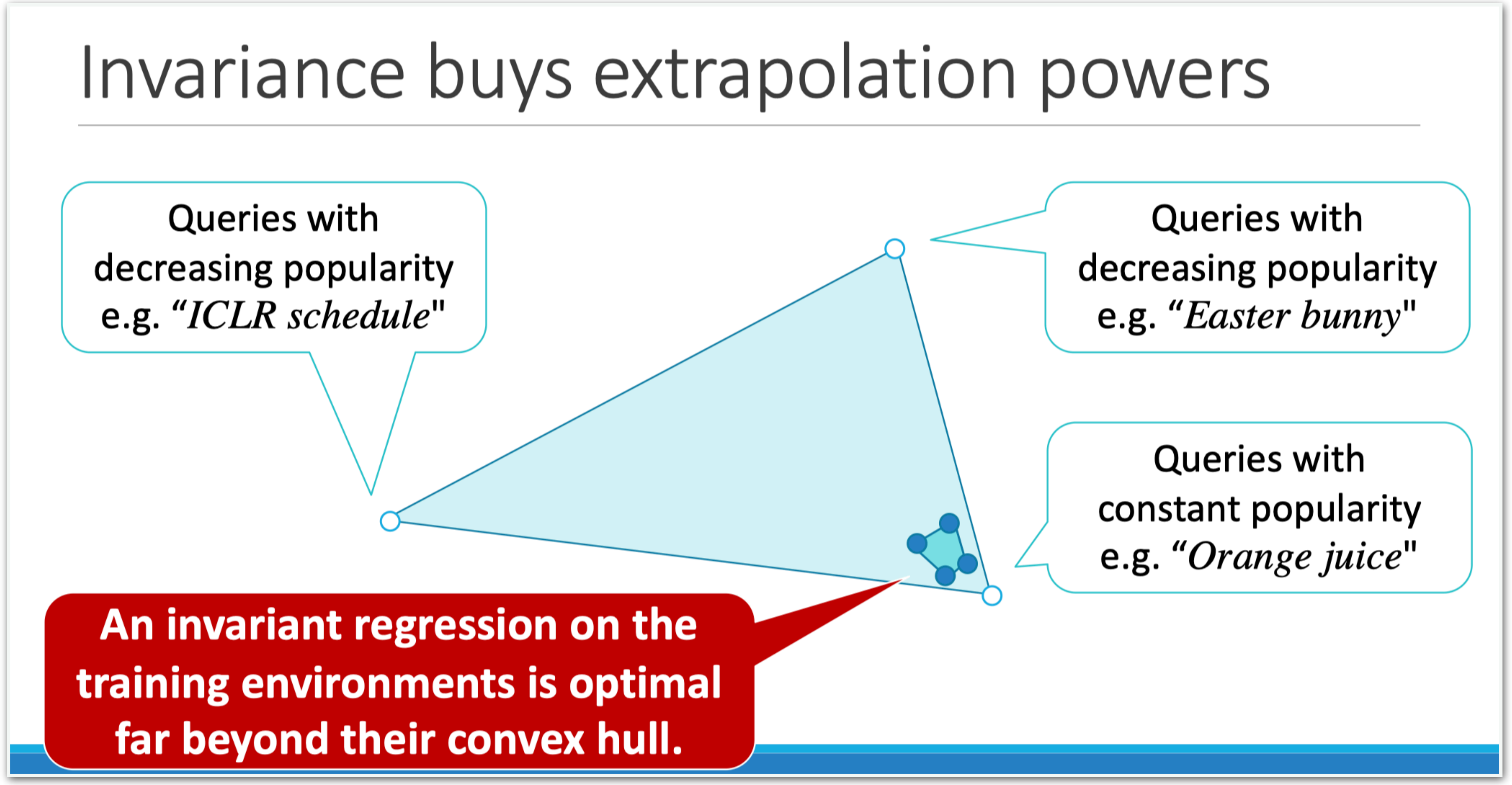


Illustration of IRMv1 failures

Robustify MOO objectives

IRM can extrapolate **stationary points** of **negative** combinations of training environments:

$$\{\sum_{e\in\mathscr{E}_{\mathrm{tr}}} \lambda_e \mathscr{D}_e \mid \sum_{e\in\mathscr{E}_{\mathrm{tr}}} \lambda_e = 1, \lambda_e \geq 0, \forall e\} \Rightarrow \{\sum_{e\in\mathscr{E}_{\mathrm{tr}}} \lambda_e \mathscr{D}_e \mid \sum_{e\in\mathscr{E}_{\mathrm{tr}}} \lambda_e = 1, \lambda_e \leq 0, \forall e\}$$

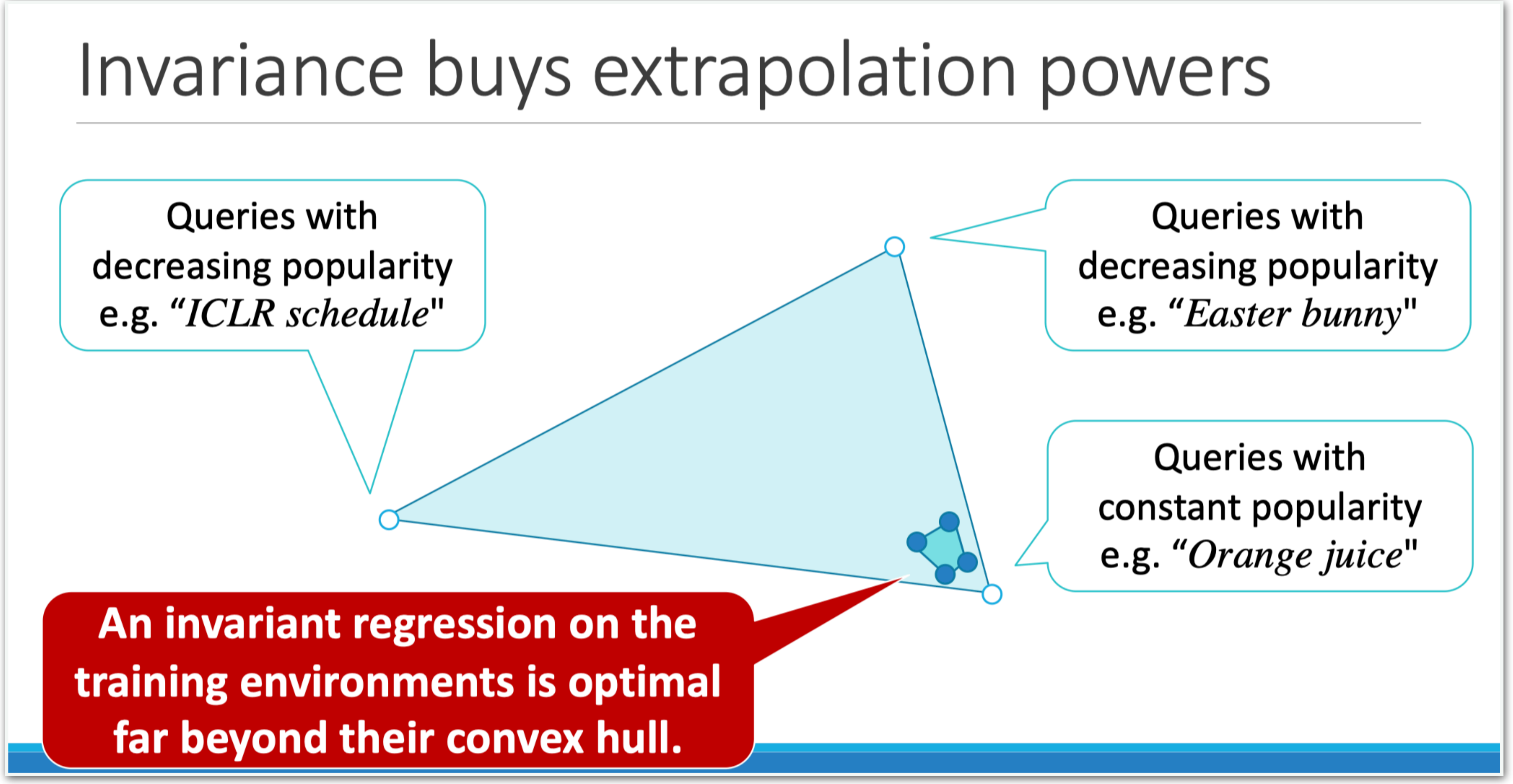


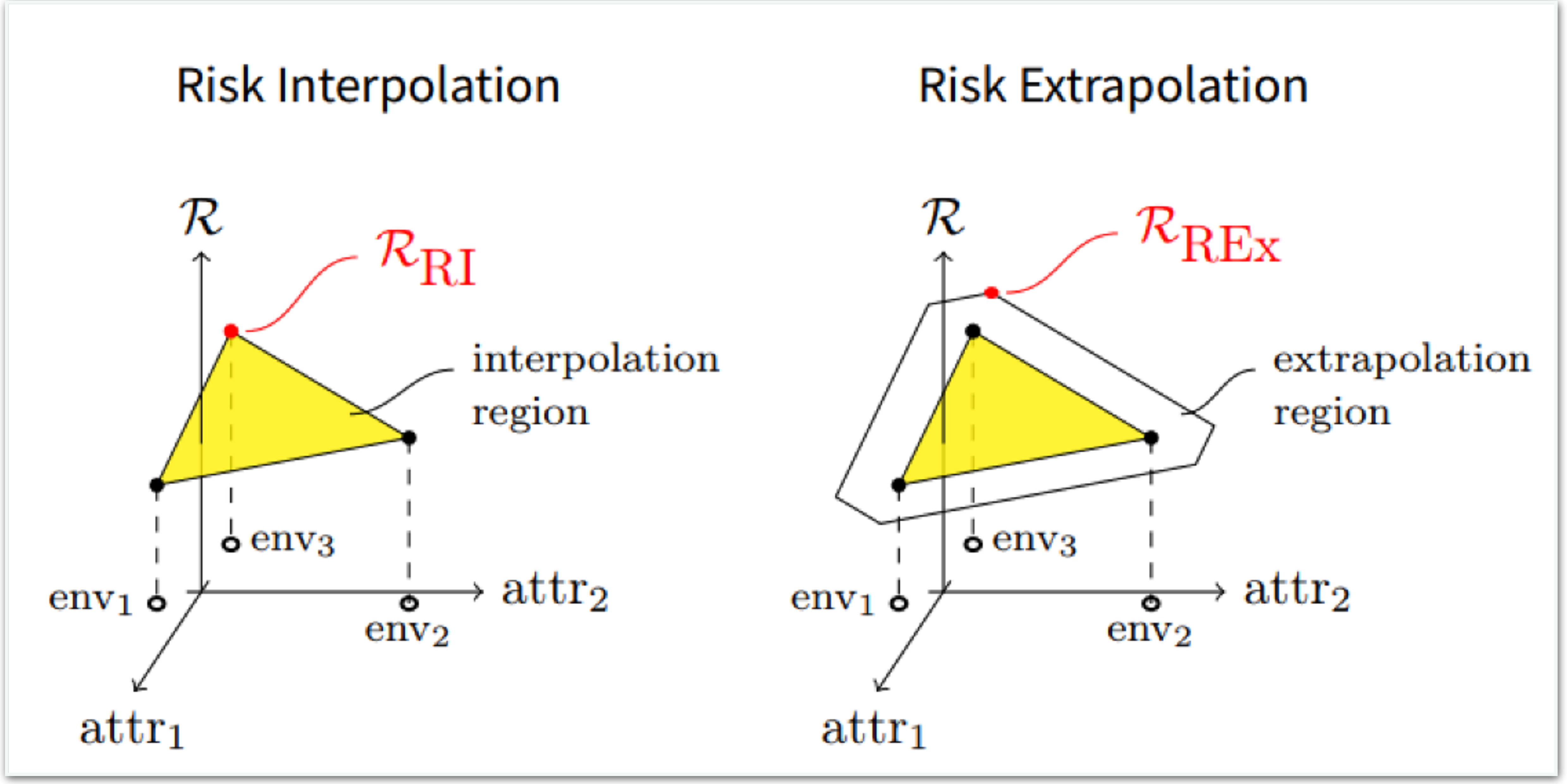
(Arjovsky et al., 2019; Bottou et al., 2019; Krueger et al., 2021)

Robustify MOO objectives

We can introduce **additional** guidance that **directly** enforces extrapolation at certain region.

$$\{\sum_{e\in\mathcal{E}_{tr}} \lambda_e \mathcal{D}_e \mid \sum_{e\in\mathcal{E}_{tr}} \lambda_e = 1, \lambda_e \geq 0, \forall e\} \Rightarrow \{\sum_{e\in\mathcal{E}_{tr}} \lambda_e \mathcal{D}_e \mid \sum_{e\in\mathcal{E}_{tr}} \lambda_e = 1, \lambda_e \leq 0, \forall e\} \Rightarrow \{\sum_{e\in\mathcal{E}_{tr}} \lambda_e \mathcal{D}_e \mid \sum_{e\in\mathcal{E}_{tr}} \lambda_e = 1, \lambda_e \leq -\beta, \forall e\}$$



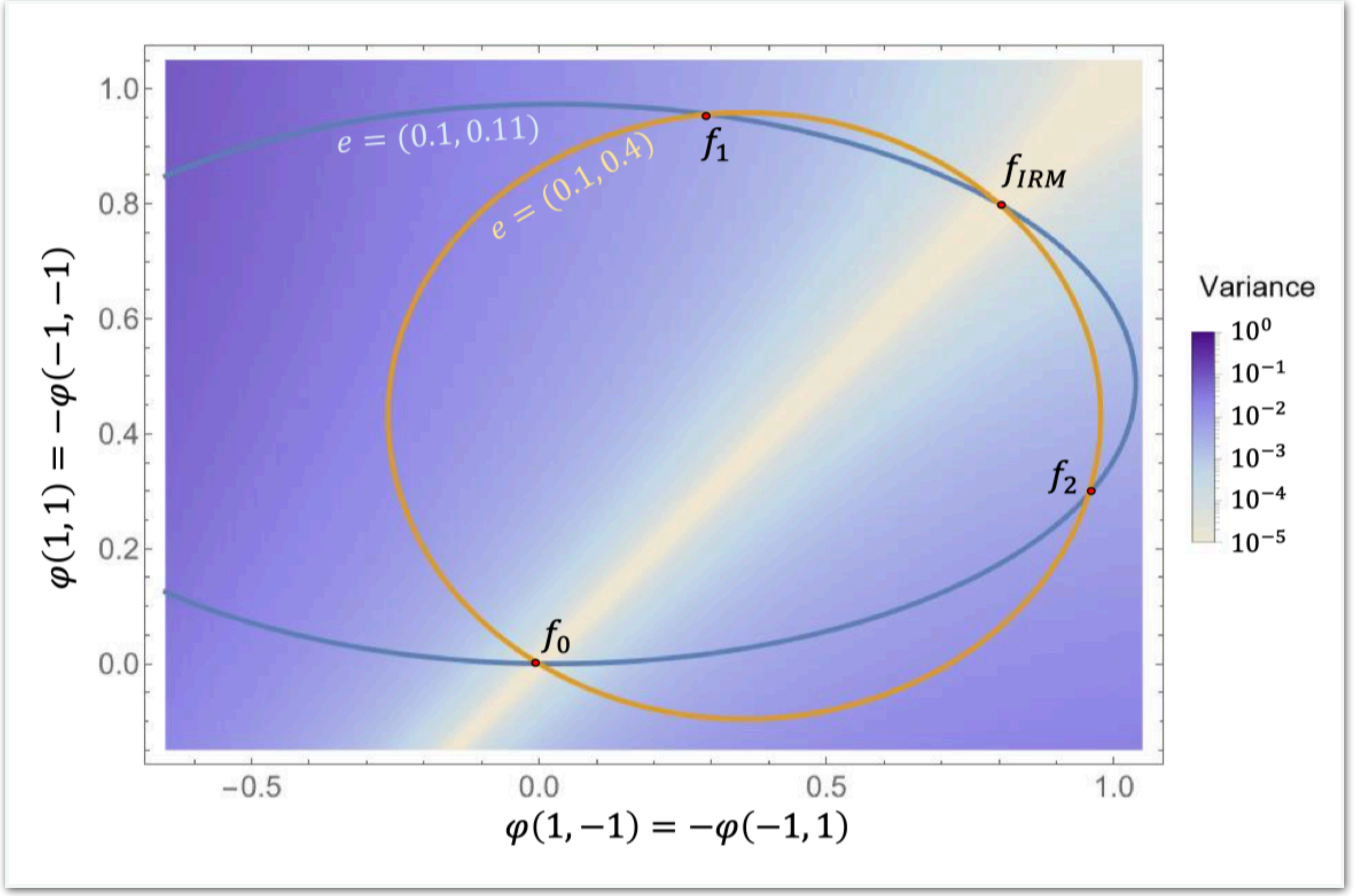


This brings us a new MOO objectives, IRMX: $\min_{f=w\cdot\varphi} \{L_1, L_2, L_{\text{IRM}}, L_{\text{REx}}\}^T$

(Arjovsky et al., 2019; Bottou et al., 2019; Krueger et al., 2021)

PAIR: PAreto Invariant Risk minimization

😋 A PAIRed journey into the adventure of extrapolation: $\min_{f=w\cdot\varphi} \{L_{\mathrm{ERM}}, L_{\mathrm{IRM}}, L_{\mathrm{REx}}\}^T$

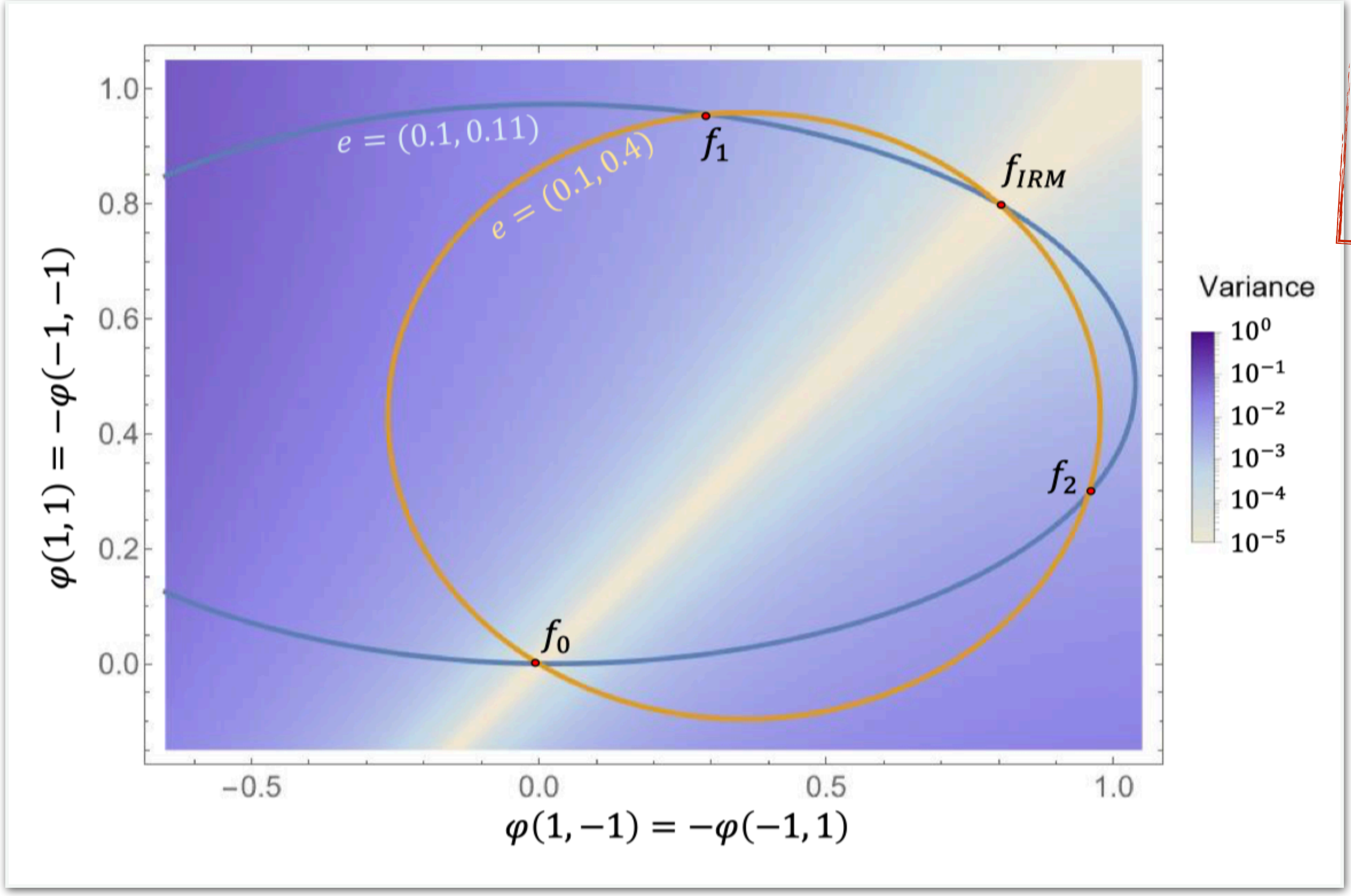


Theoretical results (Informal):

IRMX solves the IRMv1 failures under any environment settings in *(Kamath et al., 2021)*.

PAIR: PAreto Invariant Risk minimization

😋 A PAIRed journey into the adventure of extrapolation: $\min_{f=w\cdot\varphi} \{L_{\text{ERM}}, L_{\text{IRM}}, L_{\text{REx}}\}^T$



IRMX raises more challenges in hp. tuning!

Theoretical results (Informal):
IRMX solves the IRMv1 failures under any environment settings in *(Kamath et al., 2021)*.

PAIR: PAreto Invariant Risk minimization

IRMX raises more challenges in the optimization:

$$\min_{f=w\cdot\varphi} \{L_{\text{ERM}}, L_{\text{IRM}}, L_{\text{REx}}\}^T$$

- The Pareto frontier becomes **more complicated**:

PAIR: PAreto Invariant Risk minimization

IRMX raises more challenges in the optimization:

$$\min_{f=w\cdot\varphi} \{L_{\text{ERM}}, L_{\text{IRM}}, L_{\text{REx}}\}^T$$

- The Pareto frontier becomes **more complicated**:
  - ✓ The optimizer needs to be able to reach **any** Pareto optimal solutions!

e.g., MGDA algorithms *(Désidéri, 2012)*

PAIR: PAreto Invariant Risk minimization

IRMX raises more challenges in the optimization:

$$\min_{f=w\cdot\varphi} \{L_{\text{ERM}}, L_{\text{IRM}}, L_{\text{REx}}\}^T$$

- The Pareto frontier becomes **more complicated**:
  - ✓ The optimizer needs to be able to reach **any** Pareto optimal solutions!
- There can be **multiple** Pareto optimal solutions:

PAIR: PAreto Invariant Risk minimization

IRMX raises more challenges in the optimization:

$$\min_{f=w\cdot\varphi} \{L_{\text{ERM}}, L_{\text{IRM}}, L_{\text{REx}}\}^T$$

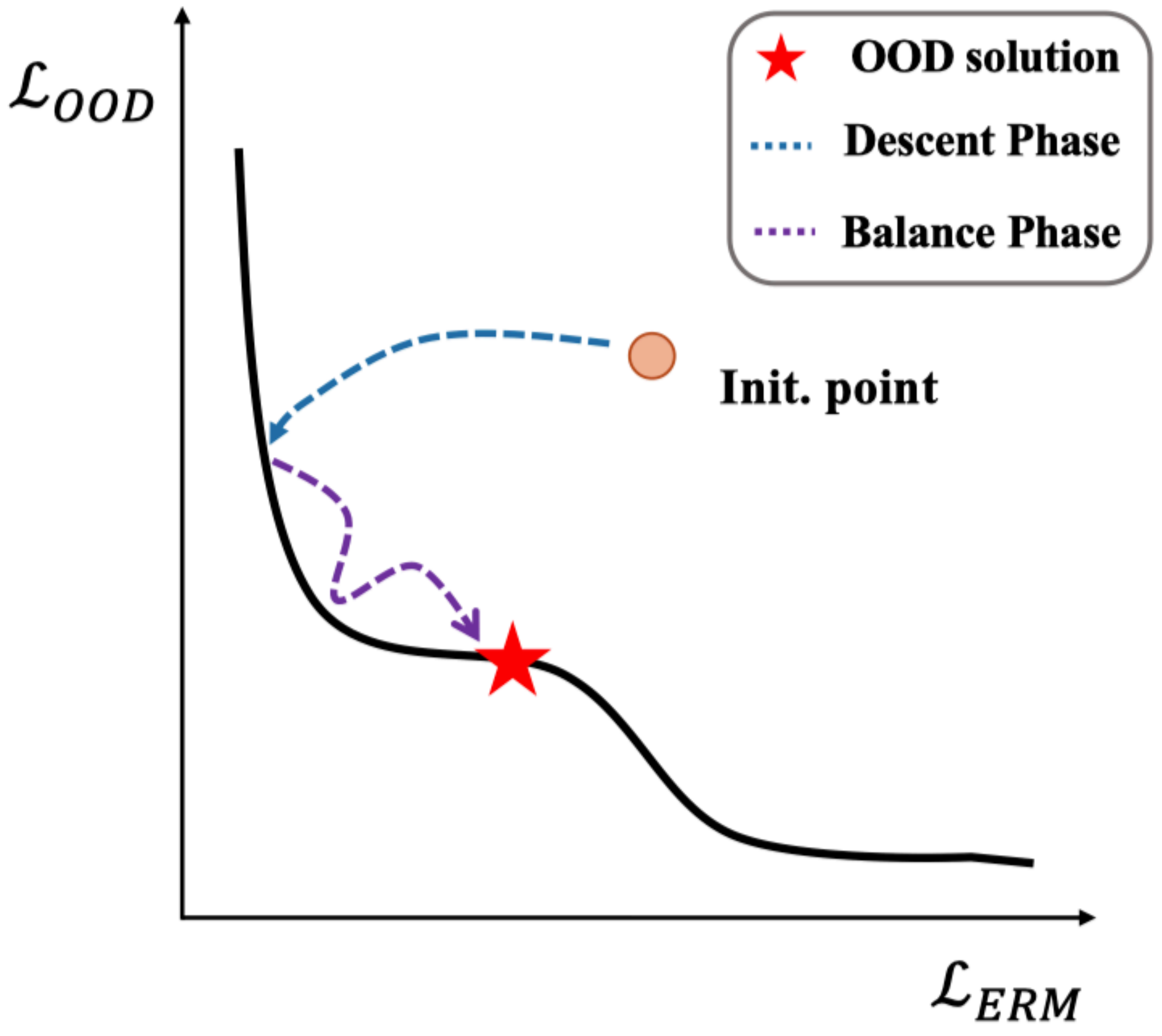
- The Pareto frontier becomes **more complicated**:
  - ✓ The optimizer needs to be able to reach **any** Pareto optimal solutions!
- There can be **multiple** Pareto optimal solutions:
  - ✓ A **preference** of each objective is required!

Exact Pareto Optimality:
Given a preference $\mathbf{p} = \{p_{\text{ERM}}, p_{\text{IRM}}, p_{\text{REx}}\}^T$ for each objective, a solution $\widehat{\mathbf{L}} = \{\widehat{L}_{\text{ERM}}, \widehat{L}_{\text{IRM}}, \widehat{L}_{\text{REx}}\}^T$ satisfies Exact Pareto Optimality iff. $p_{\text{ERM}}\widehat{L}_{\text{ERM}} = p_{\text{IRM}}\widehat{L}_{\text{IRM}} = p_{\text{REx}}\widehat{L}_{\text{REx}}$.

PAIR: PAreto Invariant Risk minimization

IRMX raises more challenges in the optimization:

$$\min_{f=w\cdot\varphi} \{L_{\text{ERM}}, L_{\text{IRM}}, L_{\text{REx}}\}^T$$



Exact Pareto optimal search

- The Pareto frontier becomes **more complicated**:
  - ✓ The optimizer needs to be able to reach **any** Pareto optimal solutions!
- There can be **multiple** Pareto optimal solutions:
  - ✓ A **preference** of each objective is required! ***PAIR-o*** as the OOD optimizer;

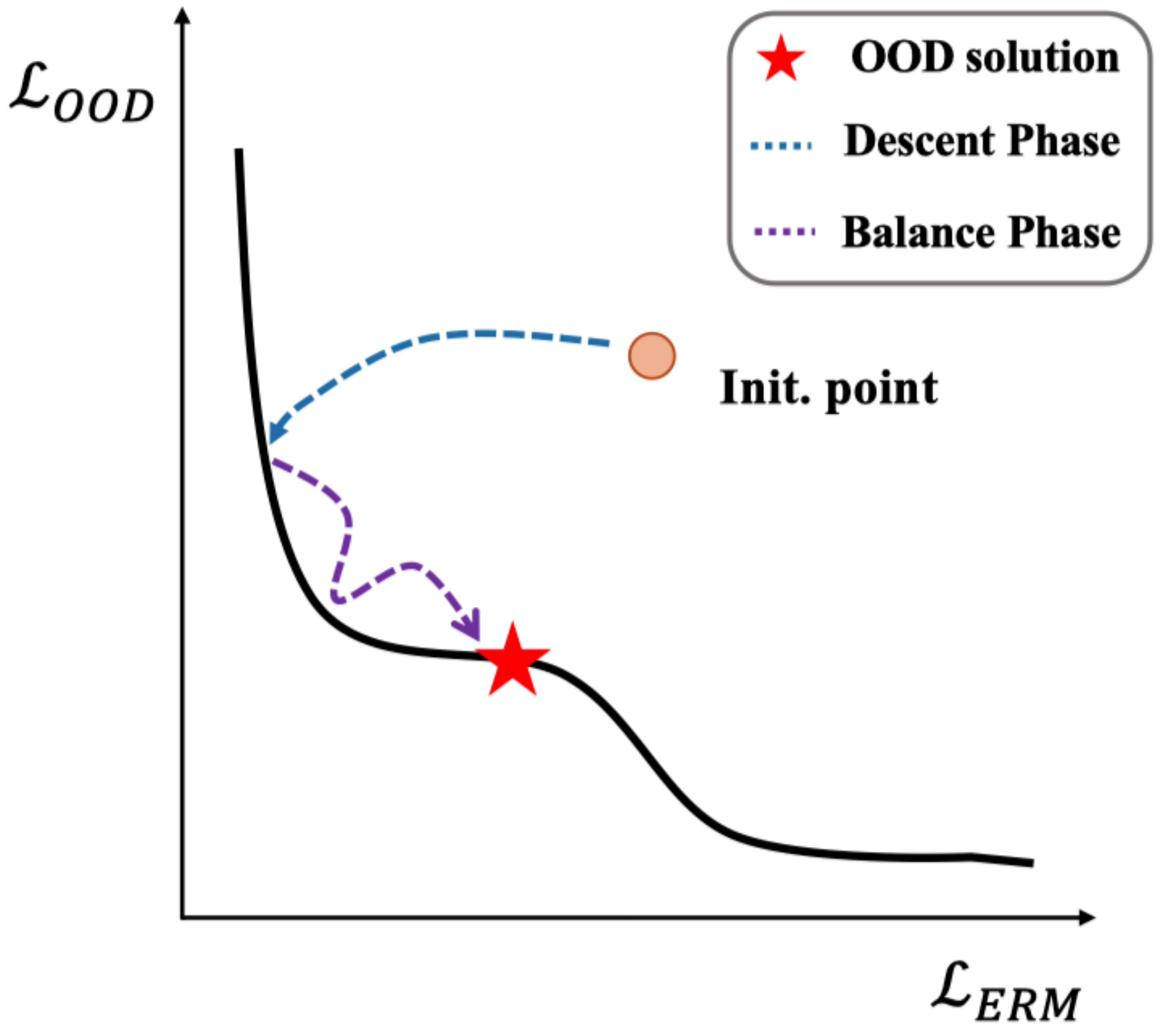
Theoretical results (Informal):
Under mild assumptions, let f_{OOD} be the desired OOD solution w.r.t. an underlying preference $\mathbf{p}_{\text{OOD}}$, PAIR-o converges and approximates to f_{OOD} for any approximated $\hat{\mathbf{p}}_{\text{OOD}}$.

(Mahapatra & Rajan 2020)

PAIR: PAreto Invariant Risk minimization

IRMX raises more challenges in the optimization:

$$\min_{f=w\cdot\varphi} \{L_{\text{ERM}}, L_{\text{IRM}}, L_{\text{REx}}\}^T$$



Exact Pareto optimal search

- The Pareto frontier becomes **more complicated**:
  - ✓ The optimizer needs to be able to reach **any** Pareto optimal solutions!
- There can be **multiple** Pareto optimal solutions:
  - ✓ A **preference** of each objective is required! ***PAIR-o*** as the OOD optimizer;
  - ✓ It also motivates a new model selection criteria, by selecting models that maximally satisfy the Exact Pareto Optimality! ***PAIR-s*** as the OOD model selector;

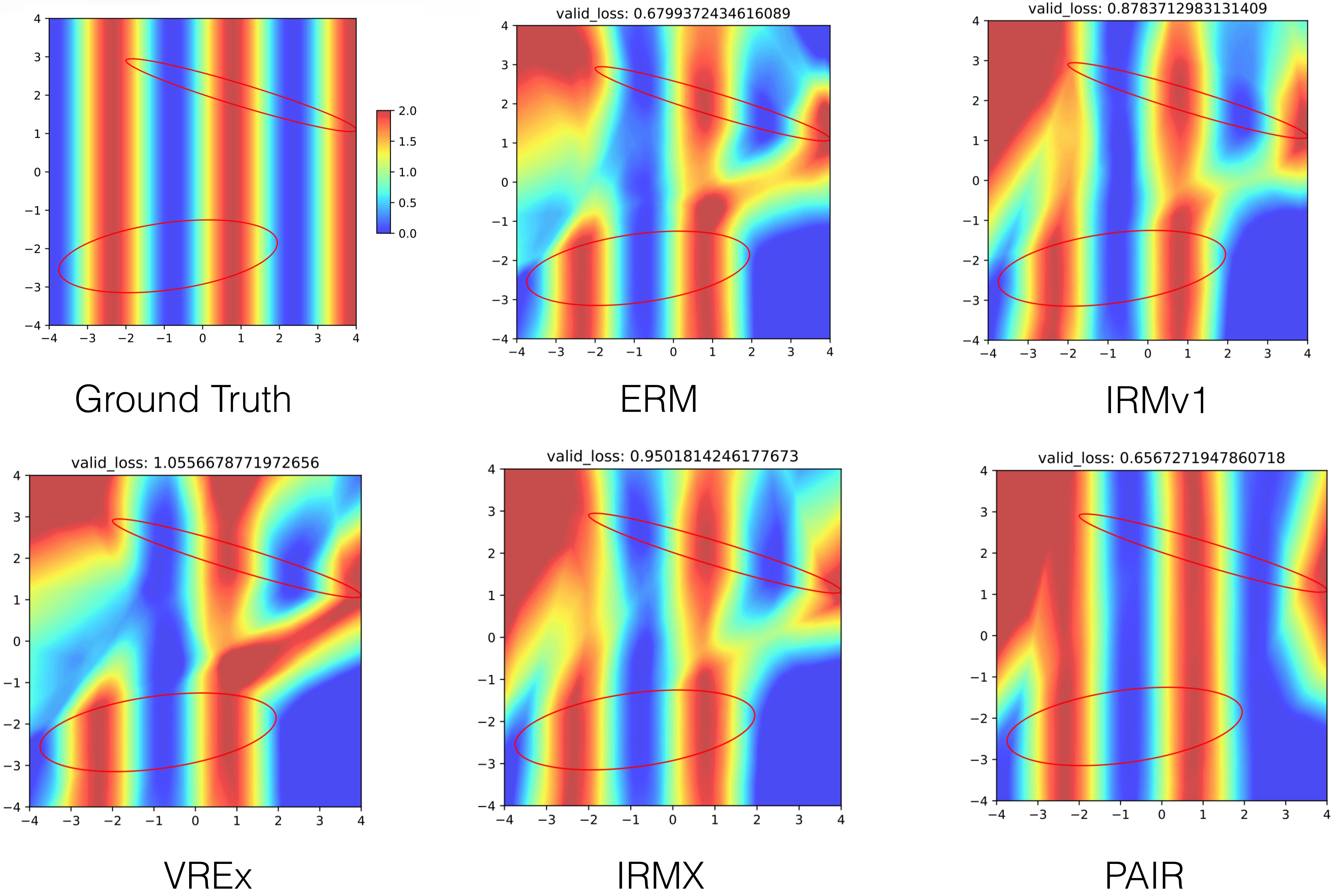
(Gulrajani & Lopez-Paz, 2021)

Causal Invariance Recovery Tests

Regression target:
$Y = \sin(X_1) + 1$, only depends on the x-axis;

Training envs:
Two elliptical regions (Gaussian distributions) marked in red;

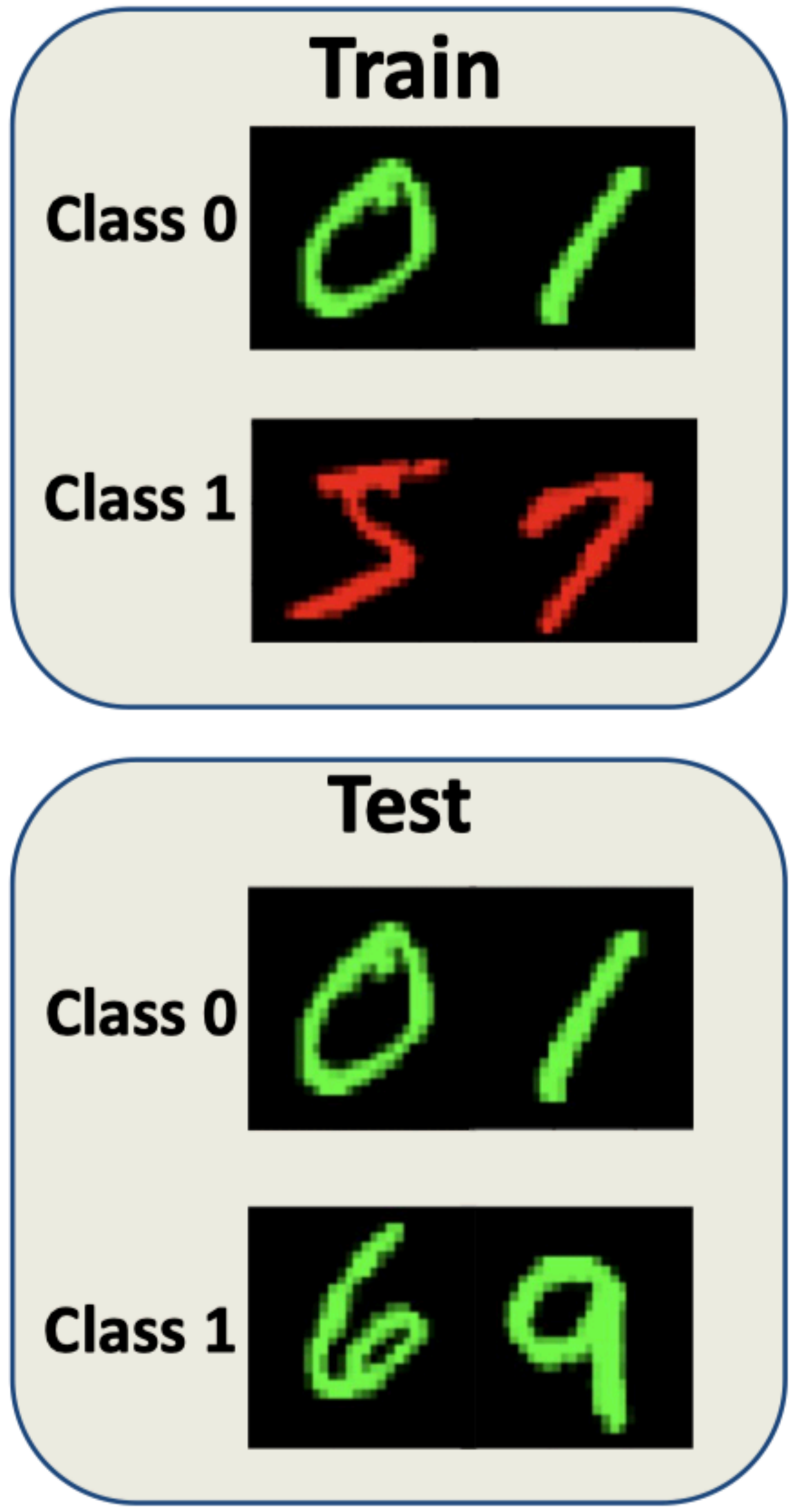
Invariance:
The **overlapped** x-axis region, i.e., $[-2,2]$.



Proof-of-Concept Experiments

Table 1: OOD Performance on ColoredMNIST

Method	CMNIST	CMNIST-m	Avg.
ERM	17.1 ± 0.9	73.3 ± 0.9	45.2
IRMv1	67.3 ± 1.9	76.8 ± 3.2	72.1
V-REx	68.6 ± 0.7	82.9 ± 1.3	75.8
IRMX	65.8 ± 2.9	81.6 ± 2.0	73.7
PAIR-o$_f$	68.6 ± 0.9	$\mathbf{83.7} \pm 1.2$	76.2
PAIR-o$_\varphi$	68.6 ± 0.8	$\mathbf{83.7} \pm 1.2$	76.2
PAIR-o$_w$	$\mathbf{69.2} \pm \mathbf{0.7}$	$\mathbf{83.7} \pm 1.2$	**76.5**
Oracle	72.2 ± 0.2	86.5 ± 0.3	79.4
Optimum	75	90	82.5
Chance	50	50	50



(Arjovsky et al., 2019; Zhang et al., 2022)

PAIR as the optimizer

Table 2: OOD generalization performances on WILDS benchmark.

	CAMELYON17	CIVILCOMMENTS	FMOW	IWILDCAM	POVERTYMAP	RXRX1	AVG. RANK(↓)†
	Avg. acc. (%)	Worst acc. (%)	Worst acc. (%)	Macro F1	Worst Pearson r	Avg. acc. (%)	
ERM	70.3 (±*6.4*)	56.0 (±*3.6*)	32.3 (±*1.25*)	30.8 (±*1.3*)	0.45 (±*0.06*)	29.9 (±*0.4*)	4.50
CORAL	59.5 (±*7.7*)	65.6 (±*1.3*)	31.7 (±*1.24*)	**32.7** (±*0.2*)	0.44 (±*0.07*)	28.4 (±*0.3*)	5.50
GroupDRO	68.4 (±*7.3*)	70.0 (±*2.0*)	30.8 (±*0.81*)	23.8 (±*2.0*)	0.39 (±*0.06*)	23.0 (±*0.3*)	6.83
IRMv1	64.2 (±*8.1*)	66.3 (±*2.1*)	30.0 (±*1.37*)	15.1 (±*4.9*)	0.43 (±*0.07*)	8.2 (±*0.8*)	7.67
V-REx	71.5 (±*8.3*)	64.9 (±*1.2*)	27.2 (±*0.78*)	27.6 (±*0.7*)	0.40 (±*0.06*)	7.5 (±*0.8*)	7.00
Fish	74.3 (±*7.7*)	73.9 (±*0.2*)	34.6 (±*0.51*)	24.8 (±*0.7*)	0.43 (±*0.05*)	10.1 (±*1.5*)	4.33
LISA	**74.7** (±*6.1*)	70.8 (±*1.0*)	33.5 (±*0.70*)	24.0 (±*0.5*)	**0.48** (±*0.07*)	**31.9** (±*0.8*)	2.67
IRMX	67.0 (±*6.6*)	74.3 (±*0.8*)	33.7 (±*0.78*)	26.6 (±*0.9*)	0.45 (±*0.04*)	28.7 (±*0.2*)	4.00
PAIR-o	74.0 (±*7.0*)	**75.2** (±*0.7*)	**35.5** (±*1.13*)	27.9 (±*0.7*)	0.47 (±*0.06*)	28.8 (±*0.1*)	**2.17**

† Averaged rank is reported because of the dataset heterogeneity. A lower rank is better.

PAIR re-empowers IRMv1 and achieves new state-of-the-arts across ***6 challenging realistic datasets***.

PAIR as the model selector

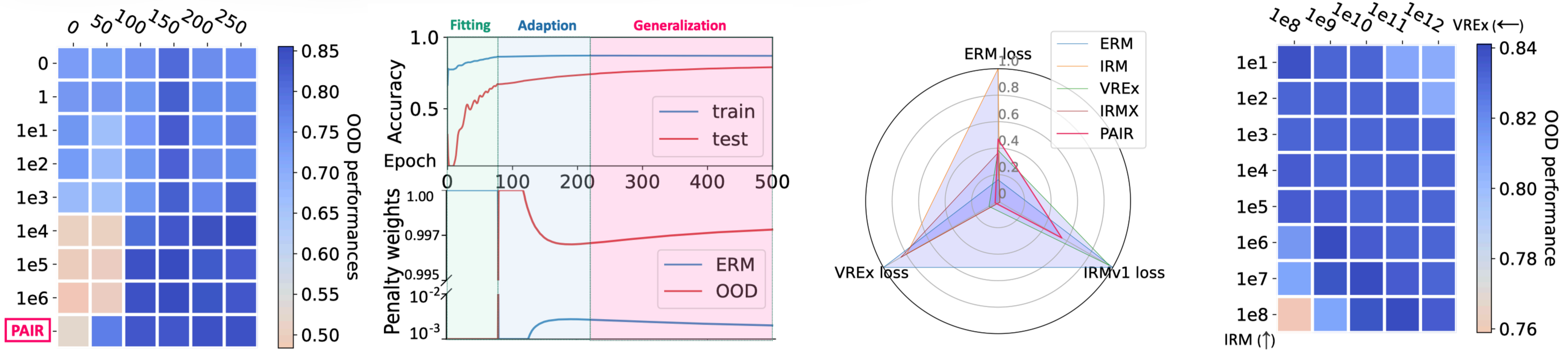
Table 3: OOD generalization performances using DOMAINBED evaluation protocol.

	PAIR-s	COLOREDMNIST † +90%	+80%	10%	Δ wr.	PACS ‡ A	C	P	S	Δ wr.	TERRAINCOGNITA † L100	L38	L43	L46	Δ wr.
ERM		71.0	**73.4**	10.0		87.2	79.5	95.5	76.9		46.7	**41.8**	57.4	39.7	
DANN		71.0	**73.4**	10.0		86.5	79.9	97.1	75.3		46.1	41.2	56.7	35.6	
DANN	✓	71.6	73.3	10.9	+0.9	87.0	81.4	96.8	77.5	+2.2	43.1	41.1	55.2	38.7	+3.1
GroupDRO		72.6	73.1	9.9		87.7	82.1	98.0	79.6		48.4	40.3	57.9	40.0	
GroupDRO	✓	**72.7**	73.2	13.0	+3.1	86.7	**83.2**	**97.8**	81.4	+1.8	48.4	40.3	57.9	40.0	+0.0
IRMv1		72.3	72.6	9.9		82.3	80.8	95.8	78.9		48.4	35.6	55.4	40.1	
IRMv1	✓	67.4	64.8	**24.2**	+14.3	85.3	81.7	97.4	79.7	+0.8	40.4	38.3	48.8	37.0	+1.4
Fishr		72.2	73.1	9.9		**88.4**	82.2	97.7	81.6		49.2	40.6	57.9	40.4	
Fishr	✓	69.1	70.9	22.6	+12.7	87.4	82.6	97.5	**82.2**	+0.6	**51.0**	40.7	**58.2**	**40.8**	+0.3

†Using the training domain validation accuracy. ‡Using the test domain validation accuracy.

PAIR-s substantially improves the worst environment performance of all representative OOD methods up to ***10%***.

How PAIR mitigates the optimization dilemma



(a) PAIR v.s. IRMX (b) Optimization trajectory (c) Normalized losses (d) Preference sensitivity

(a). PAIR **alleviates** the exhaustive parameter tuning efforts;

(b), (c). PAIR **adaptively** tunes the penalty weights towards **better** OOD solutions;

(d). PAIR is also **robust** to preference choices;

Summary

We provided a new understanding of the optimization dilemma in OOD generalization from the Multi-Objective Optimization perspective.

We attributed the failures of OOD optimization to the compromised robustness of relaxed OOD objectives and the unreliable optimization scheme.

We highlighted the importance of trading-off the ERM and OOD objectives and proposed a new optimization scheme PAIR to mitigate the dilemma.

Paper Code

Thank you!

Contact: yqchen@cse.cuhk.edu.hk