Toward a theory of multi-objective learning

September 8, 2025

Geelon So | EnCORE Collaboration Workshop

Tobias Wegel

Junhyung Park

Fanny Yang

Motivating example: self-driving car

Goal: train a model for a self-driving car.

Single-objective risk minimization

$$\min_{f \in \mathcal{F}} R(f)$$

Single-objective risk minimization

$$\min_{f \in \mathscr{F}} R(f)$$

$f : X \to Y$ is a **model** from a model class $\mathscr{F}$

"Under situation x, the car should do y."

Single-objective risk minimization

$R(f)$ measures the **population risk** of the model f

"This is the expected fuel efficiency of f on highways."

$$\min_{f \in \mathscr{F}} R(f)$$

$f : X \to Y$ is a **model** from a model class $\mathscr{F}$

"Under situation x, the car should do y."

Single-objective risk minimization

$R(f)$ measures the **population risk** of the model f

"This is the expected fuel efficiency of f on highways."

$$\min_{f \in \mathcal{F}} R(f)$$

$f : X \to Y$ is a **model** from a model class $\mathcal{F}$

"Under situation x, the car should do y."

The learning problem

Directly optimizing R is not possible since we only have sample access to it.

Motivating example: self-driving car

Goal: train a model for a self-driving car.

What we might care about:

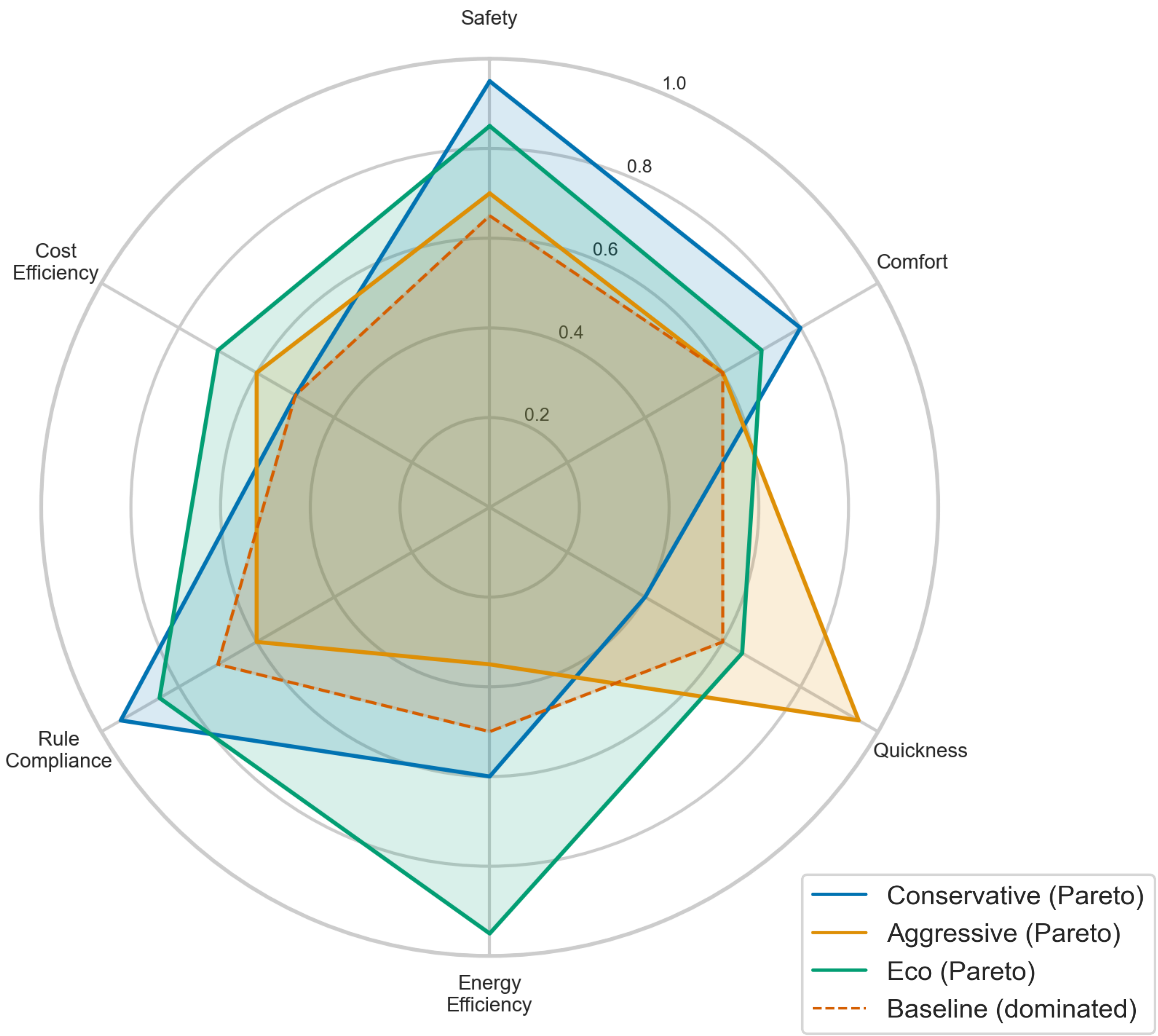
- Safety
- Efficiency
- Comfort
- [many more]

Multi-objective risk minimization

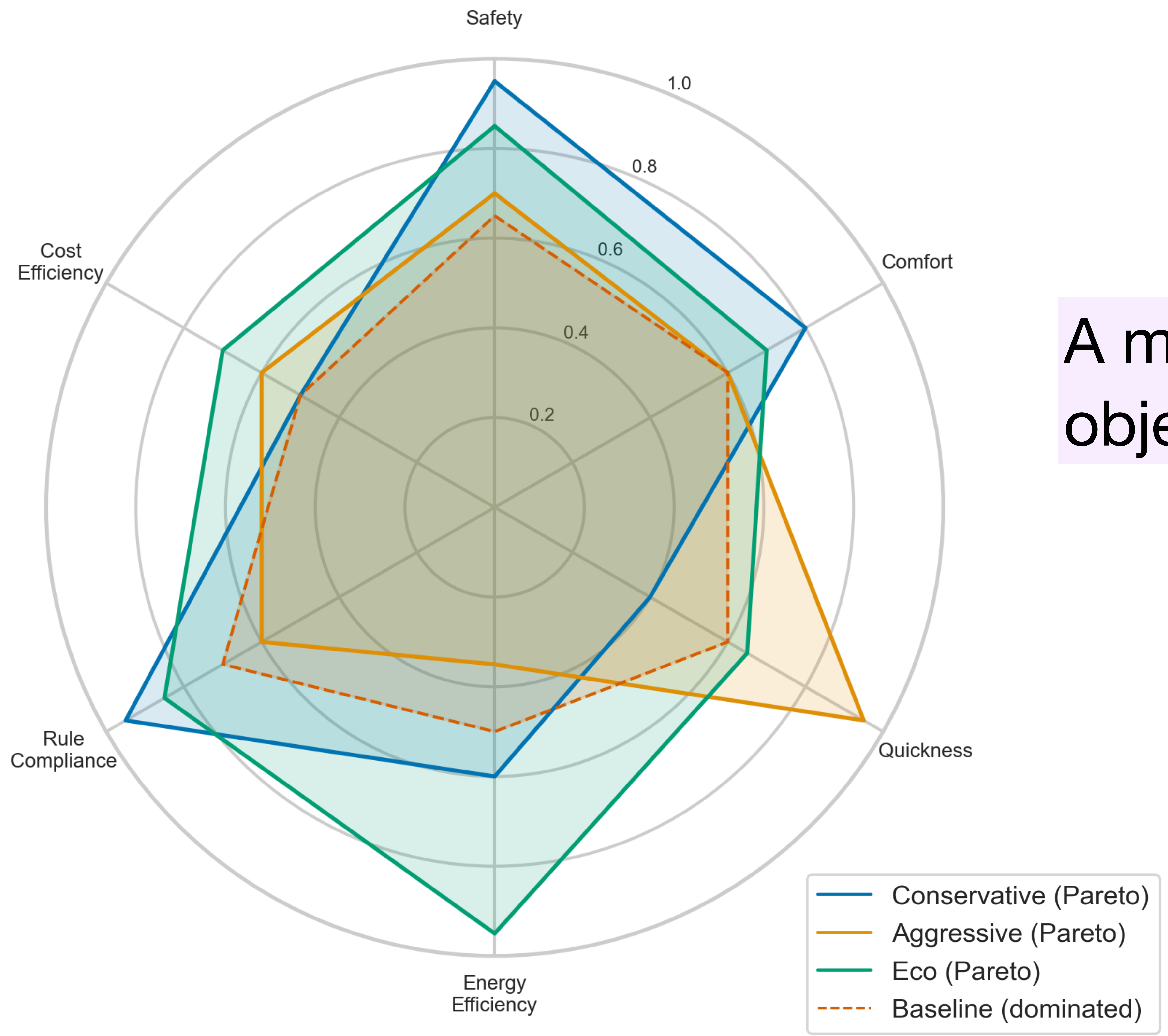
$$\min_{f \in \mathcal{F}} \underbrace{\mathbf{R}(f) \equiv \big(R_1(f), \ldots, R_K(f)\big)}$$

Now, we care about many types of risks $\mathbf{R}(f)$.

Solution concept: Pareto optimality

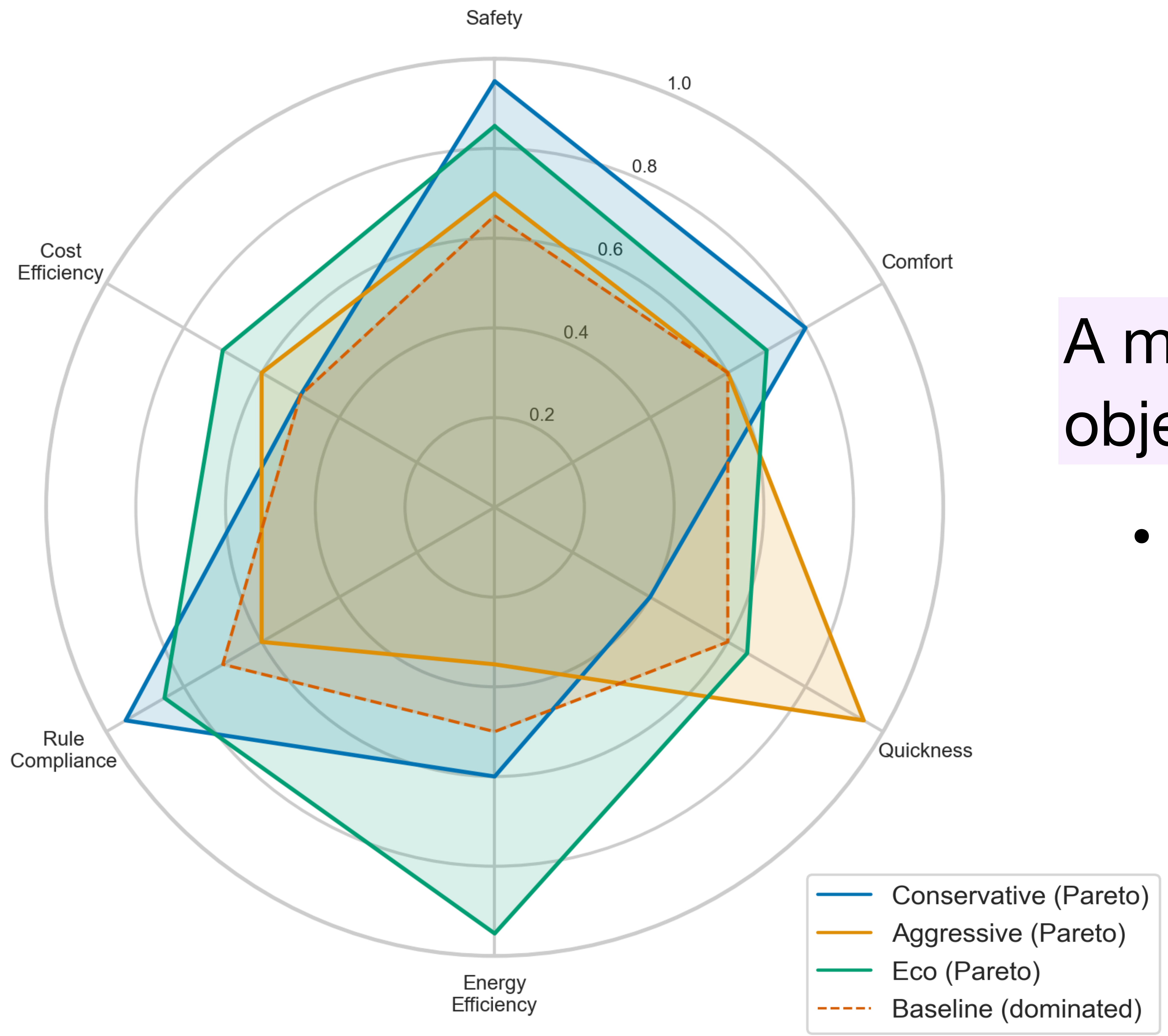


Solution concept: Pareto optimality



A model is **Pareto optimal** if improving one objective must come at a cost of another.

Solution concept: Pareto optimality



A model is **Pareto optimal** if improving one objective must come at a cost of another.

- The specific type of *trade-off* is not usually known beforehand.

Question: How much data is needed to learn all Pareto optimal models?

Question: How much data is needed to learn all Pareto optimal models?

- What if we already know something about how to solve the individual tasks?

The statistical learning setting

Let $X \times Y$ be a data space.

The statistical learning setting

Let $X \times Y$ be a data space. For each objective $k \in [K]$:

- Let R_k measure the risk of a standard supervised learning task:

The statistical learning setting

Let $X \times Y$ be a data space. For each objective $k \in [K]$:

- Let R_k measure the risk of a standard supervised learning task:
  - P_k a data distribution

The statistical learning setting

Let $X \times Y$ be a data space. For each objective $k \in [K]$:

- Let R_k measure the risk of a standard supervised learning task:
  - P_k a data distribution
  - $\ell_k(y, \hat{y})$ measures the loss of predicting $\hat{y}$ when correct answer is y

$$R_k(f) = \mathbb{E}_{P_k}\left[\ell_k\big(y, f(x)\big)\right]$$

The statistical learning setting

Let $X \times Y$ be a data space. For each objective $k \in [K]$:

- Let R_k measure the risk of a standard supervised learning task:
  - P_k a data distribution
  - $\ell_k(y, \hat{y})$ measures the loss of predicting $\hat{y}$ when correct answer is y

$$R_k(f) = \mathbb{E}_{P_k}\left[\ell_k\big(y, f(x)\big)\right]$$

  - $f_k^\star$ is the Bayes-optimal model minimizing R_k

A negative result

Theorem. Let $\mathscr{F}$ be a model class. To ε-learn all Pareto optimal models, we need:

$$\tilde{\Theta}\left(\frac{\mathrm{VC}(\mathscr{F}) \cdot K}{\varepsilon^2}\right) \text{ samples,}$$

A negative result

Theorem. Let $\mathcal{F}$ be a model class. To ε-learn all Pareto optimal models, we need:

$$\tilde{\Theta}\left(\frac{\mathrm{VC}(\mathcal{F}) \cdot K}{\varepsilon^2}\right) \text{ samples,}$$

even if we know $f_k^\star$ and have unlimited unlabeled data from P_k for each $k \in [K]$.

A negative result

Theorem. Let $\mathscr{F}$ be a model class. To ε-learn all Pareto optimal models, we need:

$$\tilde{\Theta}\left(\frac{\mathrm{VC}(\mathscr{F}) \cdot K}{\varepsilon^2}\right) \text{ samples,}$$

even if we know $f_k^\star$ and have unlimited unlabeled data from P_k for each $k \in [K]$.

Intuition: ability to drive fast + ability to be safe $\not\Rightarrow$ ability to drive fast safely

A negative result

Theorem. Let $\mathcal{F}$ be a model class. To ε-learn all Pareto optimal models, we need:

$$\tilde{\Theta}\left(\frac{\mathrm{VC}(\mathcal{F}) \cdot K}{\varepsilon^2}\right) \text{ samples,}$$

even if we know $f_k^\star$ and have unlimited unlabeled data from P_k for each $k \in [K]$.

Intuition: ability to drive fast + ability to be safe $\not\Rightarrow$ ability to drive fast safely

Problem: loss functions such as the zero-one loss can be “uninformative”.

A positive result

Theorem. Let $\mathscr{F}$ be a joint model class,

A positive result

Theorem. Let $\mathscr{F}$ be a joint model class, and let $\mathscr{H}_k \ni f_k^\star$ be model class that contains the Bayes-optimal model.

A positive result

Theorem. Let $\mathscr{F}$ be a joint model class, and let $\mathscr{H}_k \ni f_k^\star$ be model class that contains the Bayes-optimal model. If the losses ℓ_k are Bregman losses:

A positive result

Theorem. Let $\mathscr{F}$ be a joint model class, and let $\mathscr{H}_k \ni f_k^\star$ be model class that contains the Bayes-optimal model. If the losses ℓ_k are Bregman losses:

$$O\left(\frac{\sum_k \mathrm{VC}(\mathscr{H}_k)}{\varepsilon^4}\right) \text{ labeled samples,} \quad O\left(\frac{\mathrm{VC}(\mathscr{F})}{\varepsilon^2}\right) \text{unlabeled samples}$$

are enough to ε-learn all Pareto-optimal models.

A positive result

Theorem. Let $\mathscr{F}$ be a joint model class, and let $\mathscr{H}_k \ni f_k^\star$ be model class that contains the Bayes-optimal model. If the losses ℓ_k are Bregman losses:

$$O\left(\frac{\sum_k \mathrm{VC}(\mathscr{H}_k)}{\varepsilon^4}\right) \text{ labeled samples, } \quad O\left(\frac{\mathrm{VC}(\mathscr{F})}{\varepsilon^2}\right) \text{ unlabeled samples}$$

are enough to ε-learn all Pareto-optimal models.

Importantly, the label sample complexity does not the complexity of the joint class $\mathscr{F}$ in which the good trade-offs are possible.

Takeaways

- Multi-objective learning (MOL) problems are ubiquitous in practice.
- Learning good trade-offs can be much harder than solving the individual tasks.
- Structure in loss/feedback important for efficient multi-objective generalization.

Thanks!

On the sample complexity of semi-supervised multi-objective learning

https://arxiv.org/abs/2508.17152