Preference optimization on Pareto sets: On multi-objective optimization

Abhishek Roy[1,3], Geelon So[2,3] and Yi-An Ma[2]

[1]Texas A&M University, [2]University of California, San Diego, [3]Equal contribution

MOTIVATION

ML systems need to **make trade-offs** across many objectives besides accuracy.

- *Fairness-aware learning*—decisions simultaneously impact many groups of people
- *Large language models*—correctness, alignment, succinctness, reasoning quality, steerability
- *Self-driving cars*—safety, latency, fuel-efficiency
- *Portfolio optimization*—conditional values at risk

Question: How to find the most preferred trade-off?

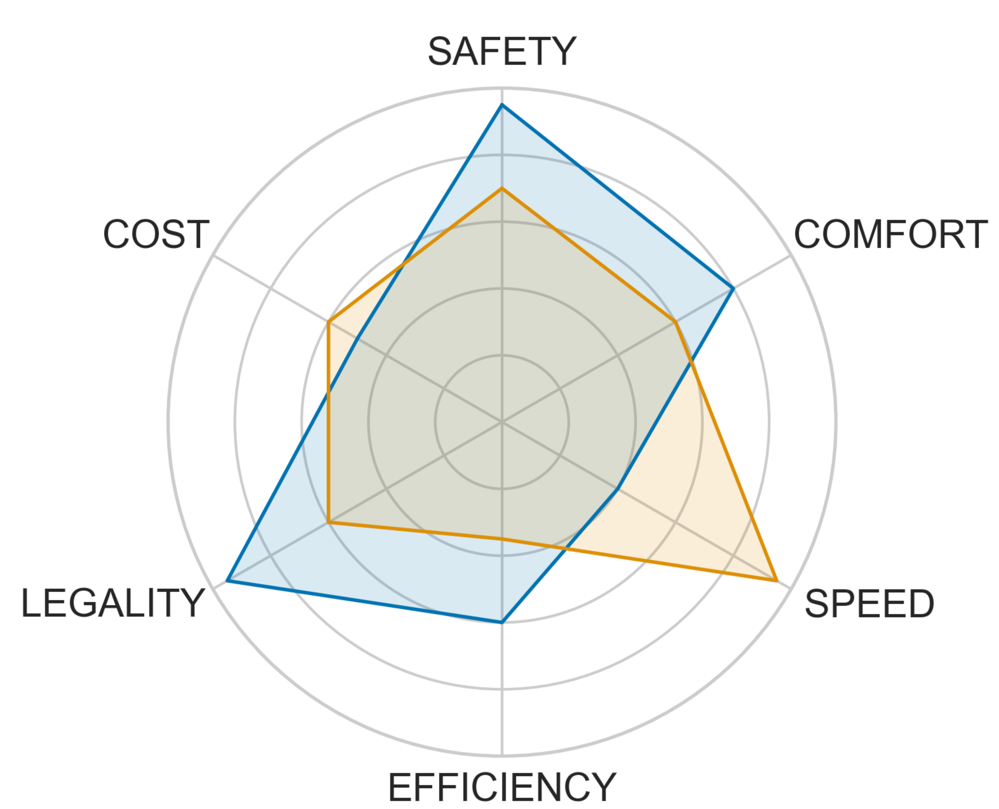

Figure. A decision is a *Pareto optimal trade-off* if improving any one objective must worsen another.

OPTIMIZATION PROBLEM

- $F \equiv (f_1, \ldots, f_n)$ is a set of n objectives on $\mathbb{R}^d$
- $\text{Pareto}(F)$ is the set of Pareto-optimal decisions
- $f_0 : \mathbb{R}^d \to \mathbb{R}$ is a preference function

Pareto-constrained preference optimization

$$\min_{x \in \text{Pareto}(F)} f_0(x)$$

CHALLENGES

- The Pareto set is **implicit**, **non-convex** and **non-smooth**, even for nice functions like quadratics.
- Problem is **NP-hard** even in the linear case (Fulöp 1993). Difficult to define **preference stationarity**.
- Preferences may come from **human feedback**.

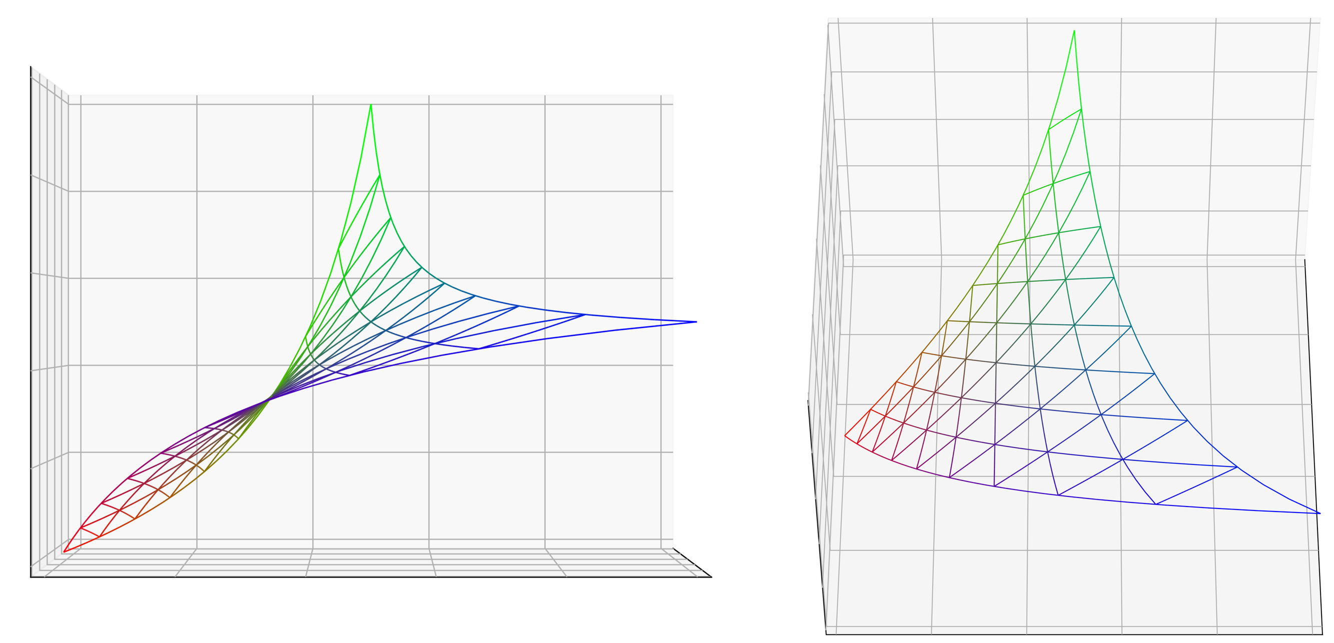

Figure. Two different projections of a Pareto manifold. The left projection yields the original Pareto set (notice its singularity).

THE PARETO MANIFOLD

- The Pareto set lives in the *decision* space, and can have poor geometry.
- We define the Pareto manifold $\mathscr{P}(F)$, which lives in a joint *decision–trade-off* space.
- When objectives are strongly convex, the Pareto manifold is *diffeomorphic* to the $(n-1)$-simplex.

Definition. The *linear scalarization* of F with weights $\beta \in \Delta^{n-1}$ is the objective:

$$f_\beta(x) := \sum_{i \in [n]} \beta_i f_i(x)$$

and its solution is $x_\beta := \arg\min_{x \in \mathbb{R}^d} f_\beta(x)$.

Definition. The *Pareto manifold* is the set:

$$\mathscr{P}(F) = \{(x_\beta, \beta) : \beta \in \Delta^{n-1}\}.$$

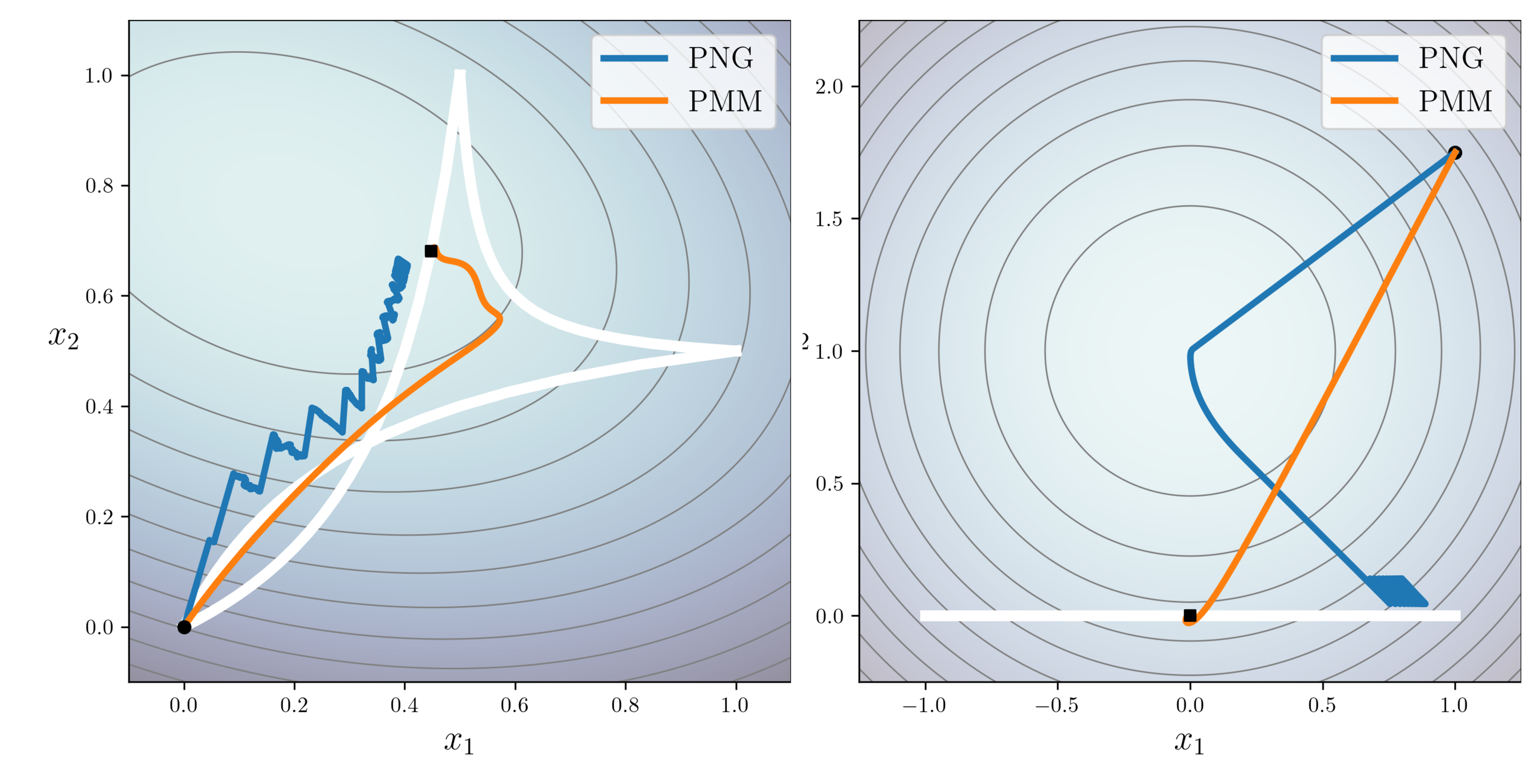

Figure. Comparison of learning dynamics of **PMM** (ours) and **PNG** (Ye and Liu, 2022). Dynamics begin at the black dot. The ground truth solution is marked by a black square.

ALGORITHMIC IDEA

- Lift the optimization problem to the joint decision–trade-off space containing the Pareto manifold:

$$\min_{(x_\beta, \beta) \in \mathscr{P}(F)} f_0(x_\beta)$$

- Implicit function theorem yields $\nabla_\beta x_\beta$, enabling gradient-descent based approach.

MAIN RESULTS

- We introduce meaningful and computationally-tractable approximate notions of **stationarity**.
- We develop an algorithm using either **gradient** or **dueling preference feedback.**
- It converges to ε-stationarity with **O(1/ε²) iterations** under standard assumptions from optimization.

TAKEAWAY

- Multi-objective optimization may be more natural in the joint decision–trade-off space.
- The Pareto set can provide structure in preference learning (dimensionality of d reduced to n).