Optimization on the Pareto set

Geometry of multi-objective optimization

Geelon So (UCSD), geelon@ucsd.edu
Yale Theory Student Seminar — March 5, 2024

Multi-objective optimization

Decision

Buying a home

Objectives

Cost

Location

Value

Noise

Solution concept: Pareto efficiency/optimality

A Pareto efficient decision makes an optimal trade off: improving one objective necessarily comes at the cost of worsening another.

Multi-objective optimization problem

The (unconstrained) multi-objective optimization problem:

$$\min_{x \in \mathbb{R}^d} F(x).$$

- $F \equiv (f_1, \ldots, f_n) : \mathbb{R}^d \to \mathbb{R}^n$ is a collection of objectives.
- x is a decision variable.
- $F(x)$ is the outcome of the decision x.

Pareto optimal solutions

Definition

*A decision $x \in \mathbb{R}^d$ is **Pareto optimal** if for all $x' \in \mathbb{R}^d$ and $i \in [d]$,*

$$f_i(x') < f_i(x) \qquad \Longrightarrow \qquad f_j(x') > f_j(x'),$$

for some $j \in [d]$.

Notation: let $\mathrm{Pareto}(F)$ be the set of Pareto optimal solutions.

Making a single decision

At the end of the day, we often need to settle on a single decision.

Problem: in general, Pareto optimal solutions are:

- non-unique: there can be many optimal trade offs,
- partially ordered: there is usually no 'best' optimal trade off.

Thus, the problem is not very well-posed yet.

Multi-objective optimization: current approaches

Covering approach
Construct a representative subsample of the set of Pareto efficient solutions.

Issues

- Unruly geometry makes sampling difficult.
- Pareto set can be very large; not a scalable approach.

Example: a realtor selects a small collection of homes for you to inspect.

Multi-objective optimization: current approaches

Scalarization approach
Reduce to single-objective optimization:
e.g. weight objectives by importance.

Issues

- Incomparable objectives.
- Hard to design the ‘right’ scalar objective.

Example: quantify how much each additional mile to work is worth to you.

Pareto-constrained optimization

This work:

- Let $F \equiv (f_1, \ldots, f_n) : \mathbb{R}^d \to \mathbb{R}^n$ be n objective functions.
- Suppose we are given an additional preference function $f_0 : \mathbb{R}^d \to \mathbb{R}$.

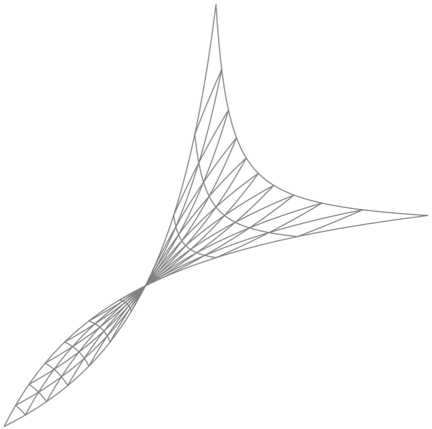
Goal: optimize f_0 constrained to the Pareto set of F,

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

Challenges of Pareto-constrained optimization

1. The Pareto set is defined implicitly.
2. The Pareto set is generally non-smooth and non-convex.
   - This is true even when the objectives are very nice.
   - Even defining an appropriate solution concept can be non-trivial.

Non-smoothness and non-convexity of Pareto set



Example
The Pareto set of three quadratics,

$$f_i(x) = \frac{1}{2}(x - c_i)^\top A_i (x - c_i).$$

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad c_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

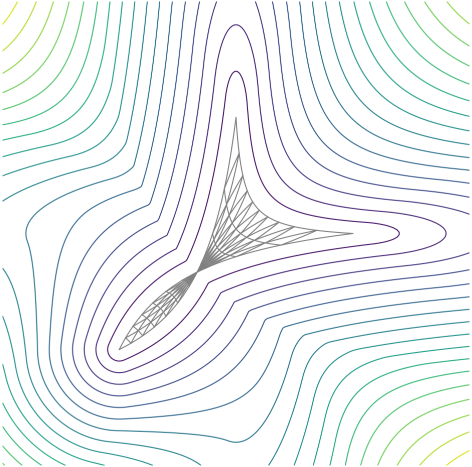
$$A_2 = \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix} \qquad c_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix} \qquad c_2 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$$

Previously observed unruliness

- singularities or self-crossings (Sheftel et al., 2013)
- needle-like extensions and knees (Kulkarni et al., 2023)

A failed attempt



Approach. Find a potential Φ where:

- $\Phi(x) \geq 0$
- $x \in \text{Pareto}(F) \iff \Phi(x) = 0.$

Difficulty. Non-smoothness of Pareto set carries over to the potential.

- Φ is not analytic near singularity; Taylor series a poor approximate.

Geometry of the Pareto set

Pareto stationarity

Definition

Let $f_1, \ldots, f_n$ be smooth. A point $x \in \mathbb{R}^d$ is Pareto stationary if zero is a convex combination:

$$\sum_{i \in [n]} w_i \nabla f_i(x) = 0,$$

for some $w_1, \ldots, w_n \geq 0$ such that $w_1 + \cdots w_n = 1$.

Notation: let Δ^{n-1} denote the $(n-1)$-dimensional simplex and for all $w \in \Delta^{n-1}$,

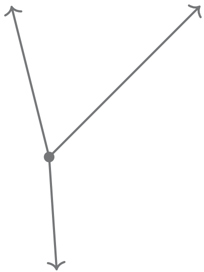
$$f_w(x) := \sum_{i \in [n]} w_i f_i(x).$$

Therefore, x is Pareto stationary if and only if $\nabla f_w(x) = 0$ for some $w \in \Delta^{n-1}$.

Pareto optimality $\Longrightarrow$ Pareto stationarity

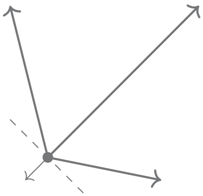
Claim. If x is not Pareto stationary, then there is a descent direction for all objectives.

Proof. Given vectors $v_1, \ldots, v_n$, Gordan's theorem states that there are two alternatives:



Zero is a convex combination:

$$w_1 v_1 + \cdots + w_n v_n = 0.$$



All vectors lie in some half-space:

$$u^\top v_i < 0.$$

Strict convexity + Pareto stationarity $\implies$ Pareto optimality

Claim. If $f_1, \ldots, f_n$ are strictly convex and x is Pareto stationary, then x is Pareto optimal.

Pareto optimality $\iff$ Pareto stationarity (under strict convexity)

Proposition

Let $f_1, \ldots, f_n$ be smooth and strictly convex. Then:

$$\mathrm{Pareto}(F) = \{x : \nabla f_w(x) = 0 \text{ for some } w \in \Delta^{n-1}\}.$$

Pareto manifold

Definition

*Let $f_1, \ldots, f_n$ be smooth and strictly convex. The **Pareto manifold** $\mathcal{P}(F)$ is defined:*

$$\mathcal{P}(F) = \{(x, w) : \nabla f_w(x) = 0\},$$

where (x, w) ranges over $\mathbb{R}^d \times \Delta^{n-1}$.

Claims:

- Pareto(F) recovered by projecting $\mathcal{P}(F)$ onto $\mathbb{R}^d$.
- $\mathcal{P}(F)$ is a smooth submanifold of $\mathbb{R}^d \times \Delta^{n-1}$.
- In fact, it is diffeomorphic to Δ^{n-1}.

Proof of smoothness structure

1. The Pareto manifold $\mathcal{P}(F)$ is the zero set of a smooth function:

$$(x, w) \mapsto \nabla f_w(x).$$

2. The Jacobian with respect to x at $(x, w) \in \mathcal{P}(F)$ is invertible:

$$\nabla^2 f_w(x) \succ 0.$$

3. By the implicit function theorem, there is a smooth map $x^* : \Delta^{n-1} \to \mathbb{R}^d$, so that:

$$(x, w) = \big(x^*(w), w\big), \qquad \forall (x, w) \in \mathcal{P}(F).$$

4. In fact, we can also deduce:

$$x^*(w) \equiv x_w := \operatorname*{arg\,min}_{x \in \mathbb{R}^d} f_w(x) \qquad \text{and} \qquad \nabla x^*(w) = -\nabla^2 f_w(x_w)^{-1} \nabla F(x_w).$$

Pareto-constrained optimization

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

Pareto-constrained optimization: high-level idea

$\text{Pareto}(F)$	$\mathcal{P}(F)$	Δ^{n-1}
Problem definition	Smoothness structure	Theory and algorithms
$\min_{x \in \text{Pareto}(F)} f_0(x)$	$\min_{(x,w) \in \mathcal{P}(F)} f_0(x)$	$\min_{w \in \Delta^{n-1}} f_0(x^*(w))$

- Pulling back to the simplex overcomes non-smoothness and non-convexity.
- However, the problem remains implicit, since $x^*(w)$ is implicitly defined.
  - This is an instance of a bilevel optimization problem:

$$\min_{w \in \Delta^{n-1}} f_0 \left(\arg\min_{x \in \mathbb{R}^d} f_w(x) \right).$$

Solution concepts

Given objectives $f_1, \ldots, f_n$ and a preference function f_0, we say:

- A point $x \in \mathbb{R}^d$ is preference optimal if it minimizes:

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

- A point $x \in \mathbb{R}^d$ is preference stationary if:
  1. x minimizes f_w for some $w \in \Delta^{n-1}$, and
  2. for all $w' \in \Delta^{n-1}$,

$$-\nabla(f_0 \circ x^*)(w)^\top (w' - w) \leq 0.$$

Preference optimality $\Longrightarrow$ preference stationarity

Proposition (Necessary condition)

If x is preference optimal, then it is preference stationary.

Proof.

Standard from convex optimization, see Nesterov (2013) for example. $\square$

Preference stationarity is a second-order condition

Expanding out the preference stationarity condition, we obtain:

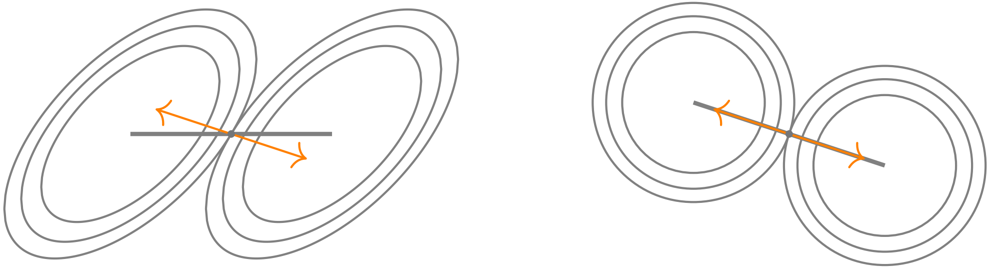
$$\nabla f_0(x_w)\, \nabla^2 f_w(x_w)^{-1}\, \nabla F(x_w)(w' - w) \leq 0,$$

which relies on second-order information about the objectives.

Question: is second-order information necessary?

- Yes. First order information ∇F doesn't tell us how the Pareto set curves.

Necessity of second-order information



Two Pareto sets (thick gray) with the same first-order information (orange vectors).

Necessary first-order conditions are trivial

Proposition

If a first-order condition is necessary for preference optimality, then it is trivial.

- A first-order condition only looks at $\nabla f_0(x), \nabla f_1(x), \ldots, \nabla f_n(x)$.
- It is necessary if it holds whenever x is preference optimal.
- Informally, it is trivial if it holds for almost all sets of first-order information.

Implication: first-order conditions are either (i) wrong at times, or (ii) uninformative.

Theory and algorithms

Estimating the gradient

Ideally, we could perform gradient descent on $f_0 \circ x^*$ using the chain rule:

$$\nabla x^*(w) = -\nabla^2 f_w(x_w)^{-1} \nabla F(x_w).$$

Because x_w is implicit, let us define the (computable) approximation:

$$\widehat{\nabla} x^*(x, w) := -\nabla^2 f_w(x)^{-1} \nabla F(x).$$

Two goals:

- Analysis of algorithms that make use of this approximation.
- Design of an algorithm that robustly makes use of this approximation.

Assumptions

We assume that the objectives $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ satisfy:

- μ-strong convexity and L-Lipschitz smoothness,
- L_H-Lipschitz continuity of the Hessians,
- minimizers are contained in the r-ball, so that:

$$\arg\min_{x \in \mathbb{R}^d} f_i \in B(0, r).$$

We also assume that the preference $f_0 : \mathbb{R}^d \to \mathbb{R}$ satisfies:

- L_0-Lipschitz smoothness.

Implications of assumptions

1. the diameter and curvature of the Pareto set can be controlled
2. the approximation error $\left\|\widehat{\nabla} x^*(x, w) - \nabla x^*(w)\right\|$ can also be controlled

Majorizing surrogates

Definition

A majorizing surrogate $g : \Delta^{n-1} \to \mathbb{R}$ *of the composition* $f_0 \circ x^*$ *is a map:*

$$g(w) \leq (f_0 \circ x^*)(w), \qquad \forall w \in \Delta^{n-1}.$$

A family of majorizing surrogates

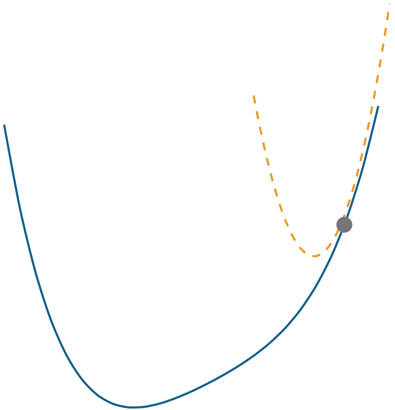
Proposition

Suppose the above assumptions hold. The following majorizes $f_0 \circ x^*$,

$$g(w'; x, w) := f(x_w) + \nabla f_0(x)^\top \widehat{\nabla} x^*(x, w)(w' - w) + \frac{cn}{2}\|w' - w\|_2^2 + \mathrm{err}(x, w).$$

- This yields a family of majorizing surrogates parametrized by (x, w).
- The constant c and error function $\mathrm{err}(x, w)$ can be computed explicitly.
- As x approaches x_w, the error term shrinks and the upper bound becomes tighter.

Majorization-minimization



Majorization-minimization
For $k = 0, 1, \ldots$

- Compute a majorizing surrogate at x_k.
- Set x_{k+1} to minimize the surrogate.

Ideally, the surrogate is tangent to the objective at x_k, and locally remains a good upper bound.

- This ensures guaranteed progress.

A majorization-minimization approach

Pareto majorization-minimization (PMM).
Initialize $(x_0, w_0) \in \mathbb{R}^d \times \Delta^{n-1}$. For $k = 0, 1, \ldots, K-1$:

- Compute the majorizing surrogate $g(\cdot\,; x_k, w_k)$.
- Compute approximate minimizers:

$$w_{k+1} \leftarrow \widehat{\operatorname*{arg\,min}_{w \in \Delta^{n-1}}} g(w; x_k, w_k) \qquad \text{and} \qquad x_{k+1} \leftarrow \widehat{\operatorname*{arg\,min}_{x \in \mathbb{R}^d}} f_{w_{k+1}}(x).$$

The subroutines are easy: $\operatorname*{arg\,min}_{w \in \Delta^{n-1}} g(w)$ is a quadratic program; $f_w(x)$ is strongly convex.

Convergence analysis

Theorem (Convergence of PMM)

The Pareto majorization-minimization method achieves ε-stationarity in $O(\varepsilon^{-2})$ iterations, provided that the subroutines achieve:

- ε*-stationarity for* $\arg\min_{w \in \Delta^{n-1}} g(w)$
- ε^2*-optimality for* $\arg\min_{x \in \mathbb{R}^d} f_w(x)$.

Extensions

Ongoing work

- Optimization with dueling feedback
- Analysis of two-timescale projected gradient descent/mirror descent
- Sampling from the Pareto set

Collaborators

Abhishek Roy
UC San Diego

Yian Ma
UC San Diego

Thank you!

Paper at *https://arxiv.org/abs/2308.02145*.

References

Aditya Kulkarni, Maximilian Kohns, Michael Bortz, Karl-Heinz Küfer, and Hans Hasse. Regularities of pareto sets in low-dimensional practical multi-criteria optimisation problems: Analysis, explanation, and exploitation. *Optimization and Engineering*, 24(3):1611–1632, 2023.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Abhishek Roy*, Geelon So*, and Yi-An Ma. Optimization on pareto sets: On a theory of multi-objective optimization. *arXiv preprint arXiv:2308.02145*, 2023.

Hila Sheftel, Oren Shoval, Avi Mayo, and Uri Alon. The geometry of the p areto front in biological phenotype space. *Ecology and evolution*, 3(6):1471–1483, 2013.