A Review of Multi-Objective Optimization: Theory and Algorithms

Suyun Liu & Luis Nunes Vicente

ISE Department, Lehigh University

April 21, 2020

Presentation outline

1. Introduction to multi-objective optimization

2. Scalarization methods (entire Pareto front)
   - Weighted-sum method
   - ϵ-constrained method

3. Gradient-based methods (single Pareto point)
   - Multi-objective steepest descent method
   - Multi-objective Newton's method

4. Outline of the various algorithmic classes

Multi-Objective Optimization

A multi-objective optimization problem (MOP) consists of 'simultaneously' optimizing several objective functions (often conflicting):

$$\begin{aligned} \min \quad & F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix} \\ \text{s.t.} \quad & x \in \Omega \end{aligned}$$

where

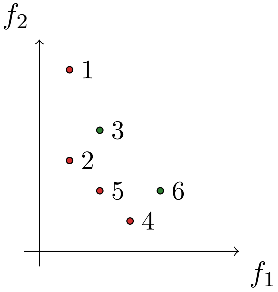
1. $\Omega \subseteq \mathbb{R}^n$ is the feasible set in decision space
2. $\mathbb{R}^m$ is the goal/objective space
3. $F(\Omega) = \{F(x) : x \in \Omega\} \subseteq \mathbb{R}^m$ is the image of the feasible set.

Pareto dominance

Let us consider a bi-objective discrete example where $\Omega = \{1, 2, 3, 4, 5, 6\}$.

The functions f_1 and f_2 are defined by:

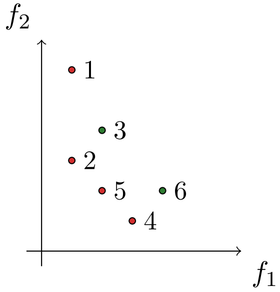
Ω	1	2	3	4	5	6
f_1	1	1	2	3	2	4
f_2	6	3	4	1	2	2



Let us consider a bi-objective discrete example where $\Omega = \{1, 2, 3, 4, 5, 6\}$.

The functions f_1 and f_2 are defined by:

Ω	1	2	3	4	5	6
f_1	1	1	2	3	2	4
f_2	6	3	4	1	2	2



1. There is no point that minimizes both functions.
2. 3 has no interest (2 is better in both objectives), the same with 6.
3. $P = \{1, 2, 4, 5\} \subset \Omega$ is the set of Pareto minimizers (or efficient or nondominated points).

Pareto dominance

Definition 1: x is a (weak) Pareto minimizer of F in Ω if

$$\nexists y \in \Omega \quad \text{such that} \quad F(y) \ < \ F(x).$$

Here, we are using an partial order induced by $\mathbb{R}^m_{++}$

$$F(x) < F(y) \Leftrightarrow F(y) - F(x) \in \mathbb{R}^m_{++}.$$

The set of (weak) Pareto minimizers is given by

$$P \ = \ \{x \in \Omega : \ \nexists y \in \Omega \ : F(y) \ < \ F(x)\}.$$

Pareto dominance

In the previous example, $P_s = \{2, 4, 5\}$ is the set of strict Pareto minimizers:

Ω	1	2	3	4	5	6
f_1	1	1	2	3	2	4
f_2	6	3	4	1	2	2

In fact, point 1 is not a strict Pareto minimizer since

$$F(2) \ \leq \ F(1) \qquad \text{and} \qquad F(2) \neq F(1).$$

Pareto dominance

Definition 2: x is a strict Pareto minimizer of F in Ω if

$$\nexists y \in \Omega : F(y) \leq F(x) \quad \text{and} \quad F(y) \neq F(x).$$

The set of strict Pareto minimizers is thus given by

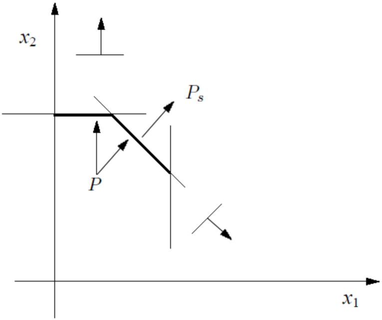
$$P_s = \{x \in \Omega : \nexists y \in \Omega : F(y) \leq F(x) \text{ and } F(y) \neq F(x)\}.$$

Theorem (Relationship between P and P_s)

$P_s \subseteq P$.

Pareto dominance

Case (a): $P_s \subsetneq P$

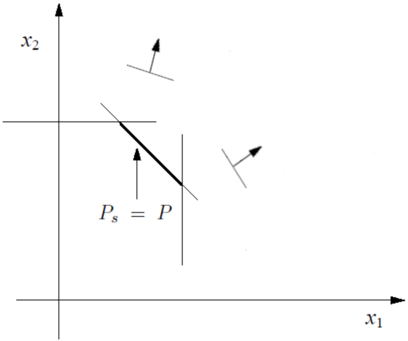


$$\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \leq 0.5, \quad x_2 \leq 0.75, \quad x_1 + x_2 \leq 1, \quad x_1, x_2 \geq 0\}$$

$$f_1(x_1, x_2) = -x_2$$

$$f_2(x_1, x_2) = x_2 - x_1$$

Case (b): $P = P_s$



$$\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \leq 0.5, \quad x_2 \leq 0.75, \quad x_1 + x_2 \leq 1, \quad x_1, x_2 \geq 0\}$$

$$f_1(x_1, x_2) = -0.5x_1 - x_2$$

$$f_2(x_1, x_2) = -2x_1 - x_2$$

Pareto dominance

The existence of points in P and P_s can be guaranteed in a classical way.

Theorem (existence and compactness)

If Ω is compact and F is $\mathbb{R}^m$-continuous, then

1. P *is nonempty and compact.*
2. P_s *is nonempty.*

Pareto dominance

The existence of points in P and P_s can be guaranteed in a classical way.

Theorem (existence and compactness)

If Ω is compact and F is $\mathbb{R}^m$-continuous, then

1. *P is nonempty and compact.*
2. *P_s is nonempty.*

Definition 3: x is local (strict) Pareto minimizer if there is a neighborhood $V \subseteq \Omega$ of x such that the point x is (strictly) nondominated.

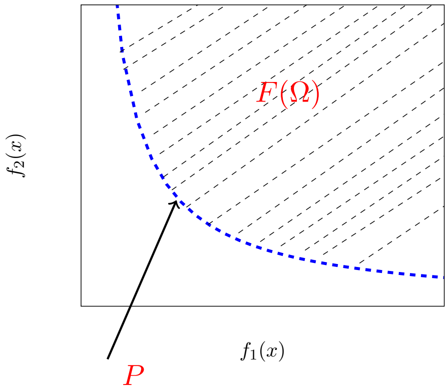
Property 1: If Ω is convex and F is $\mathbb{R}^m$-convex, every local Pareto minimizer is a global Pareto minimizer.

Pareto fronts

Recall the image of the feasible set Ω:

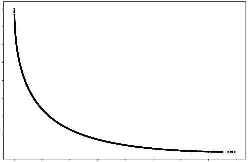
$$F(\Omega) = \{F(x) : x \in \Omega\}$$

Proposition 1: $F(x), x \in P$ is always on the boundary of $F(\Omega)$.

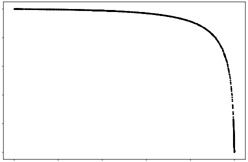


Pareto front

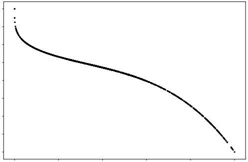
Denote Pareto front by $F(P) = \{F(x) : x \in P\}$.



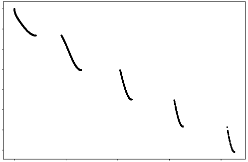
(a) SP1



(b) FF1



(c) JOS2



(d) ZDT3

Figure: Different geometry shapes of Pareto fronts: (a) Convex; (b) Concave; (c) Mixed (neither convex nor concave); (d) Disconnected.

Presentation outline

1. Introduction to multi-objective optimization
2. Scalarization methods (entire Pareto front)
   - Weighted-sum method
   - ϵ-constrained method
3. Gradient-based methods (single Pareto point)
   - Multi-objective steepest descent method
   - Multi-objective Newton's method
4. Outline of the various algorithmic classes

Weighted-sum method

According to a pre-defined preference given by a set of non-negative weights $\mu_1, \ldots, \mu_m$, the general weighted-sum method is to solve

$$\min \sum_{i=1}^{m} \mu_i f_i(x) \qquad \text{s.t. } x \in \Omega$$

Assume that

$$x_* \in \operatorname*{argmin}_{x \in \Omega} \sum_{i=1}^{m} \mu_i f_i(x).$$

Property 2:

1. If μ_i's are not all zero (non-negative scalarization), then $x_* \in P$.
2. If μ_i's are all positive (positive scalarization), then $x_* \in P_s$.

Weighted-sum method

In the convex case the non-negative scalarization of P is necessary and sufficient:

Theorem (sufficient and necessary condition)

Assume that F is $\mathbb{R}^m$-convex ($f_1, \ldots, f_m$ are convex) on Ω convex. Then,

$$x_* \in P$$

if and only if

$$\exists\ \mu_1, \ldots, \mu_m \geq 0 \text{ not all zero} \quad x_* \in \operatorname*{argmin}_{x \in \Omega} \sum_{i=1}^{m} \mu_i f_i(x).$$

Weighted-sum method

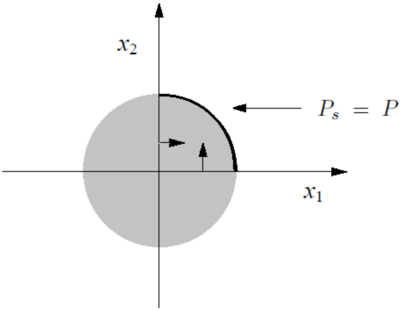
However, the positive scalarization of P_s is not necessary.

Consider the example where:

$$m = 2, \quad f_i(x) \; = \; -x_i, \; i = 1, 2 \quad \text{and} \quad \Omega \; = \; \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}.$$

In this example we have

$$P_s \; = \; P \; = \; \{x \in \mathbb{R}^2 : \; x_1^2 + x_2^2 = 1, \; x_1, x_2 \geq 0\}.$$



Weighted-sum method

Thus $(1,0) \in P_s$. However

$$\mu_1 f_1(1,0) + \mu_2 f_2(1,0) \;=\; -\mu_1$$

and

$$\min_{y\in\Omega}\; \mu_1 f_1(y) + \mu_2 f_2(y) \;\;=\;\; \min_{y\in\Omega}\; -\mu_1 y_1 - \mu_2 y_2$$

are only equal when $\mu_1 > 0$ and $\mu_2 = 0$.

Therefore, just by varying the positive weight combinations, one might not necessarily capture the whole P_s.

However, in the strictly convex case, the non-negative scalarization is also necessary for P_s.

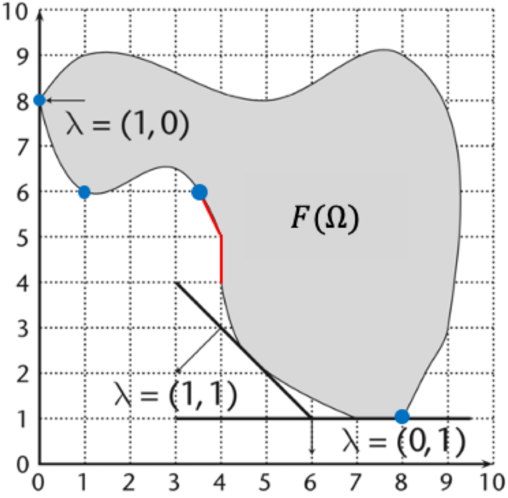
Theorem ($P_s = P$ in strictly convex case)

Let F be $\mathbb{R}^m$-strictly convex ($f_1, \ldots, f_m$ are strictly convex) on Ω convex. Then

$$P_s = P.$$

By varying all non-negative weight combinations, we are able to get the whole P and P_s.

Non-convexity in weighted-sum method



ϵ-constrained method

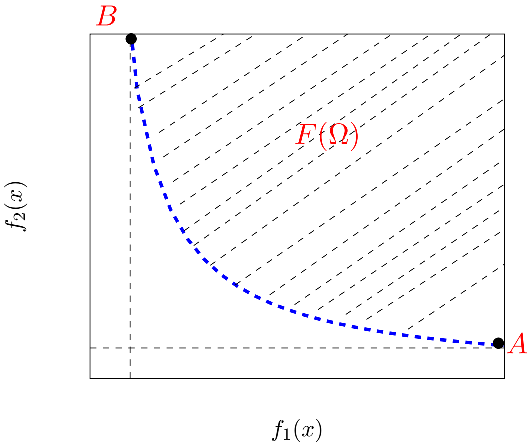
The original MOP is converted into a constrained problem by optimizing an objective from the satisfaction of the other

$$\begin{aligned} \min \quad & f_1(x) \\ \text{s.t.} \quad & x \in \Omega, \\ & f_2(x) \le \epsilon. \end{aligned}$$

In this case, P can be computed solving these problems for

$$\epsilon \in \left[\min_{y \in \Omega} f_2(y), \; f_2(\operatorname*{argmin}_{y \in \Omega} f_1(y)) \right].$$

ϵ-constrained method



1. $\epsilon = \min_{x \in \Omega} f_2(x)$, the optimal solution corresponds to A.
2. $\epsilon = f_2(\operatorname{argmin}_{y \in \Omega} f_1(y))$, the optimal solution corresponds to B.

ϵ-constrained method

ϵ-constrained method does not require any convexity assumption.

Consider the general ϵ-constrained problem ($\epsilon \in \mathbb{R}^m$)

$$
\begin{aligned}
\min \quad & f_l(x) \\
\text{s.t.} \quad & f_i(x) \leq \epsilon_i, \forall i = 1, \ldots, m, \text{ and } i \neq l \\
& x \in \Omega.
\end{aligned}
\tag{1}
$$

Theorem (sufficient and necessary condition)

1. *Let ϵ be such that the feasible region of (1) is nonempty for a certain l. If x_* is an optimal solution of problem (1), then $x_* \in P$.*
2. *A feasible point $x_* \in \Omega$ is in P_s if and only if there is a vector $\epsilon_* \in \mathbb{R}^m$ such that x_* is an optimal solution for all problems (1), $l = 1, \ldots, m$.*

Presentation outline

1. Introduction to multi-objective optimization

2. Scalarization methods (entire Pareto front)
   - Weighted-sum method
   - ϵ-constrained method

3. Gradient-based methods (single Pareto point)
   - Multi-objective steepest descent method
   - Multi-objective Newton's method

4. Outline of the various algorithmic classes

First order necessary condition

Consider a MOP

$$\min F(x) \quad x \in \mathbb{R}^n.$$

where we assume $F : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable.

Pareto first-order stationary condition: x is Pareto stationary for F if

$$\forall d \in \mathbb{R}^n, \text{we have } JF(x)d \not< 0.$$

where

$$JF(x) = \begin{bmatrix} \nabla f_1(x)^\top \\ \vdots \\ \nabla f_m(x)^\top \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Equivalently,

$$\max_{i=1,\ldots,m} \nabla f_i(x)^\top d \geq 0, \quad \forall d \in \mathbb{R}^n.$$

First order necessary condition

Equivalently, if the convex hull of $\nabla f_i(x)$'s contains the origin, i.e.,

$$\exists \lambda \in \Delta^m \text{ such that } \sum_{i=1}^{m} \lambda_i \nabla f_i(x_k) = 0$$

where $\Delta^m = \{\lambda : \sum_{i=1}^m \lambda_i = 1, \lambda_i \geq 0, \forall i = 1, ..., m\}$ is the m-simplex set.

Note: when F is $\mathbb{R}^m$-convex, $x \in P \Leftrightarrow x$ is Pareto first-order stationary.

Line search

For any non-stationary point x, there exists $d \in \mathbb{R}^n$ such that $\nabla f_i(x)^\top d < 0, \forall i = 1, \ldots, m$.

One further has

$$\lim_{t \to 0} \frac{f_i(x + td) - f_i(x)}{t} \ = \ \nabla f_i(x)^\top d \ < \ 0, \quad \forall i$$

i.e., $\exists t_0$ such that $F(x + td) < F(x)$ holds for all $t \in (0, t_0]$.

Lemma (sufficient decrease condition)

Given any $\sigma \in (0, 1)$, *there exists* $\bar{t}_0 > 0$ *such that*

$$F(x + td) < F(x) + \sigma\, t\, JF(x)d \quad \forall t \in (0, \bar{t}_0]$$

The multi-objective steepest descent method

Steepest descent direction is computed by (Fliege and Svaiter, 2000)

$$d(x) = \mathrm{argmax}_{d\in\mathbb{R}^n} \min_{i=1,\ldots,m} -\nabla f_i(x)^\top d + \frac{1}{2}\|d\|^2.$$

This subproblem is uniformly convex.

Its dual problem is

$$\lambda(x) \;=\; \underset{\lambda\in\mathbb{R}^m}{\mathrm{argmin}} \;\|\sum_{i=1}^{m} \lambda_i \nabla f_i(x)\|^2 \;\text{ s.t. }\; \lambda \in \Delta^m.$$

And we have

$$d(x) = -\sum_{i=1}^{m} (\lambda(x))_i \nabla f_i(x).$$

Note: when $m = 1$, one recovers $d(x) = -\nabla f_1(x)$.

The multi-objective steepest descent method

Steepest descent direction is computed by (Fliege and Svaiter, 2000)

$$d(x) = \operatorname*{argmin}_{d \in \mathbb{R}^n} \max_{i=1,\ldots,m} \nabla f_i(x)^\top d + \frac{1}{2}\|d\|^2.$$

This subproblem is uniformly convex.

Its dual problem is

$$\lambda(x) \; = \; \operatorname*{argmin}_{\lambda \in \mathbb{R}^m} \; \|\sum_{i=1}^{m} \lambda_i \nabla f_i(x)\|^2 \;\; \text{s.t.} \;\; \lambda \in \Delta^m.$$

And we have

$$d(x) = -\sum_{i=1}^{m} (\lambda(x))_i \nabla f_i(x).$$

Note: when $m = 1$, one recovers $d(x) = -\nabla f_1(x)$.

Multi-objective steepest descent method

Let $\theta(x)$ be the optimal value of the subproblem

$$\theta(x) = \max_{i=1,\ldots,m} \nabla f_i(x)^\top d(x) + \frac{1}{2}\|d(x)\|^2.$$

Proposition (Fliege and Svaiter (2000))

1. $\theta(x) \leq 0, \ \forall\, x \in \mathbb{R}^n$
2. *The following conditions are equivalent:*
   - x *is non-stationary*
   - $\theta(x) < 0$
   - $d(x) \neq 0$

Hence, x is stationary if and only if $\theta(x) = 0$ (or if and only if $d(x) = 0$).

The multi-objective steepest descent algorithm

Algorithm 1 MSDM with backtracking

1: Choose $\sigma \in (0,1)$ and $x_0 \in \mathbb{R}^n$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Compute d_k by solving a convex constrained subproblem

$$\begin{aligned} \min_{\beta, d} \quad & \beta + \frac{1}{2}\|d\|^2 \\ \text{s.t.} \quad & \nabla f_i(x_k)^\top d \leq \beta, i = 1, \ldots, m. \end{aligned}$$

4: If $\theta(d_k) = 0$, then stop.
5: Choose stepsize α_k as the largest $\alpha \in \{1/2^j : j \in \mathbb{N}\}$ such that

$$F(x_k + \alpha d_k) \leq F(x_k) + \sigma\, \alpha\, JF(x_k) d_k.$$

6: Update iterate $x_{k+1} = x_k + \alpha_k d_k$

Convergence and complexity of MSDM

Theorem (Lip. continuous gradients, Fliege and Svaiter (2000))

Let $\{x_k\}$ *be a sequence generated by Algorithm 1. Every accumulation point of the sequence, if any, is a stationary point.*

Theorem (F is $\mathbb{R}^m$-nonconvex, Fliege et al. (2019))

Assume at least one of functions $f_i, i = 1, \ldots, m$, *is bounded below, the sequence* $\{x_k\}$ *generated by Algorithm 1 satisfies*

$$\min_{0 \leq i \leq k-1} \|d_i\| \; \leq \; \mathcal{O}(1/\sqrt{k}).$$

Correspondingly, for the non-stationarity measure, we have

$$|\theta(x_k)| \leq \mathcal{O}(1/\sqrt{k}).$$

Convergence and complexity of MSDM

Assume the sequence $\{x_k\}$ converges to x_* associated with the weights λ^*.

1. F is $\mathbb{R}^m$-strongly convex
   - A linear rate in terms of iterates: $\|x_k - x_*\| \leq \mathcal{O}(c^k), c \in (0,1)$
   - A linear rate for optimality gap using weighted-sum function: $\sum_{i=1}^m \lambda_i^* f_i(x_k) - \sum_{i=1}^m \lambda_i^* f_i(x_*) \leq \mathcal{O}(c^k)$.

2. F is $\mathbb{R}^m$-convex: $\mathcal{O}(1/k)$ sublinear rate for optimality gap defined by a weaker form of weighted-sum function

$$\sum_{i=1}^m \bar{\lambda}_i^{k-1} f_i(x_k) - \sum_{i=1}^m \bar{\lambda}_i^{k-1} f_i(x_*) \leq \mathcal{O}(1/k)$$

where $\bar{\lambda}^{k-1} = \frac{1}{k} \sum_{l=1}^{k-1} \lambda_i^l$.

Multi-objective Newton's method

Assume F is $\mathbb{R}^m$-strongly convex and twice continuous differentiable.

Newton direction $s(x)$ is computed by (Fliege et al., 2009)

$$s(x) = \underset{s \in \mathbb{R}^n}{\operatorname{argmin}} \max_{i=1,\dots,m} \nabla f_i(x)^\top s + \frac{1}{2} s^\top \nabla^2 f_i(x) s$$

Here, we are approximating $\max_{i=1,\dots,m} f_i(x+s) - f_i(x)$ using maximum over local quadratic model.

The subproblem can be framed into a convex quadratically constrained problem:

$$\begin{aligned} \min \quad & t \\ \text{s.t.} \quad & \nabla f_i(x)^\top s + \tfrac{1}{2} s^\top \nabla^2 f_i(x) s - t \le 0, \quad \forall i = 1, \dots, m \\ & (t, s) \in \mathbb{R} \times \mathbb{R}^n \end{aligned} \tag{2}$$

Multi-objective Newton's method

Lemma (Newton direction, Fliege et al. (2009))

$$s(x) = -\left[\sum_{i=1}^{m} \lambda(x)_i \nabla^2 f_i(x)\right]^{-1} \sum_{i=1}^{m} \lambda(x)_i \nabla f_i(x)$$

where $\lambda(x)$ is the Lagrange coefficient associated with problem (2).

Let $t(x)$ be the optimal value of the subproblem

$$t(x) = \max_{i=1,\ldots,m} \nabla f_i(x)^\top s(x) + \frac{1}{2} s(x)^\top \nabla^2 f_i(x) s(x)$$

Multi-objective Newton's method

Proposition (Fliege et al. (2009))

1. $\forall x \in \mathbb{R}^n, t(x) \leq 0$
2. *The following conditions are equivalent:*
   - x *is not Pareto stationary*
   - $t(x) < 0$
   - $s(x) \neq 0$

Hence, x is stationary if and only if $t(x) = 0$ (or if and only if $s(x) = 0$).

Algorithm 2 MNM with backtracking

1: Choose $\sigma \in (0,1)$ and $x_0 \in \mathbb{R}^n$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Compute s_k by solving a convex constrained subproblem

$$\begin{array}{ll} \min & t \\ \text{s.t.} & \nabla f_i(x_k)^\top s + \frac{1}{2} s^\top \nabla^2 f_i(x_k) s - t \leq 0, \quad \forall i = 1, \ldots, m \\ & (t, s) \in \mathbb{R} \times \mathbb{R}^n \end{array}$$

4: If $t_k = 0$, then stop.
5: Choose stepsize α_k as the largest $\alpha \in \{1/2^j : j \in \mathbb{N}\}$ such that

$$F(x_k + \alpha s_k) \leq F(x_k) + \sigma\, \alpha\, JF(x_k) s_k.$$

6: Update iterate $x_{k+1} = x_k + \alpha_k d_k$

Multi-objective Newton's method

Theorem (Local quadratic convergence rate, Fliege et al. (2009))

Assume the Hessians $\nabla^2 f_i, \forall i$ are uniformly positive definite and Lipschitz continuous.

Let x_0 be sufficiently close to a Pareto stationary point x_*. The sequence $\{x_k\}$ generated by Algorithm 2 satisfies

1. *$\{x_k\}$ converges to x_* with a q-quadratic rate.*
2. *$\|s(x_k)\|$ converges to 0 with a r-superlinear rate.*

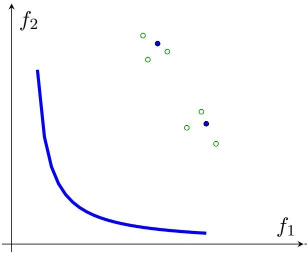
Presentation outline

Deterministic multi-objective optimization

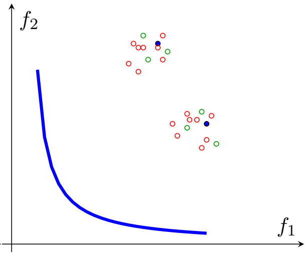
1. A priori methods: preference selection before optimization
    - weighted-sum methods (non-convexity is an issue)
    - ϵ-constrained methods (infeasibility is an issue)
    - other methods based on utility functions or expressions of preference: reference point methods, goal programming...
2. A posteriori methods: preference selection after optimization

   Most of them work by iteratively updating lists of non-dominated points:
    - evolutionary algorithms (e.g., NSGA-II and AMOSA) which have no theoretical convergence guarantee.
    - mathematical programming based algorithms (e.g., Section 3 of this talk), convergence guaranteed for one point on the Pareto front.

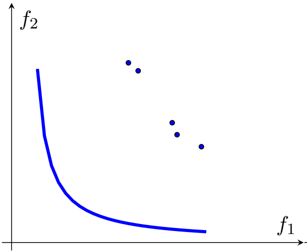
Illustration of a list updating strategy



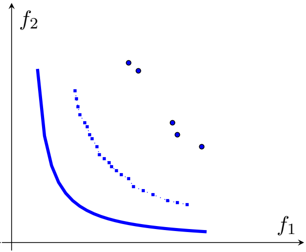
(a) Adding perturbed points.



(b) Applying descent steps.



(c) Removing dominated pts.



(d) Moving front.

Stochastic multi-objective optimization

- "Multi-objective methods": they convert the original problem into an approximated deterministic multi-objective one (e.g., using SAA).
- "Stochastic methods": they convert the original problem into a single-objective stochastic one (e.g., by the weighting method).

Purity: an accuracy measure

P_1: a set of computed Pareto minimizers by solver 1

P_2: a set of computed Pareto minimizers by solver 2

$\bar{P}$: the set of nondominated points in $P_1 \cup P_2$

$$\text{Purity}(P_1) = |P_1 \cap \bar{P}|/|\bar{P}| \in [0,1]$$

which calculates the percentage of nondominated solutions.

Maximum size of holes

P: the set of N computed Pareto minimizers

Assume each list of objective function values $\{f_{i,j}\}_{j=1}^{N}$ is sorted in order

$$\Gamma(P) \; = \; \max_{i \in \{1,\ldots,m\}} \left(\max_{j \in \{1,\ldots,N\}} \{\delta_{i,j}\} \right),$$

where $\delta_{i,j} = f_{i,j+1} - f_{i,j}$.

Spread

$$\Delta(P) \;=\; \max_{i\in\{1,\ldots,m\}} \left(\frac{\delta_{i,0} + \delta_{i,N} + \sum_{j=1}^{N-1} |\delta_{i,j} - \bar{\delta}_i|}{\delta_{i,0} + \delta_{i,N} + (N-1)\bar{\delta}_i} \right),$$

where two extreme points indexed by 0 and $N+1$ are added, and $\bar{\delta}_i$ is the average of $\delta_{i,j}$ over $j = 1, \ldots, N-1$.

The lower Γ and Δ are, the more well distributed the Pareto front is.

References

M. Ehrgott. Multicriteria Optimization, volume 491. Springer Science & Business Media, Berlin, 2005.

J. Fliege and B. F. Svaiter. Steepest descent methods for multicriteria optimization. Math. Methods Oper. Res., 51:479–494, 2000.

J. Fliege, L. G. Drummond, and B. F. Svaiter. Newton's method for multiobjective optimization. SIAM J. Optim., 20:602–626, 2009.

J. Fliege, A. I. F. Vaz, and L. N. Vicente. Complexity of gradient descent for multiobjective optimization. Optim. Methods Softw., 34:949–959, 2019.

E. H. Fukuda and L. M. G. Drummond. A survey on multiobjective descent methods. Pesquisa Operacional, 34:585–620, 2014.