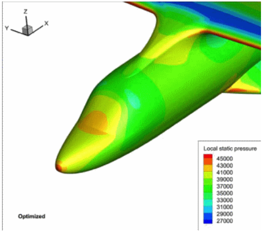
Dynamique(s) de descente pour l'optimisation multi-objectif

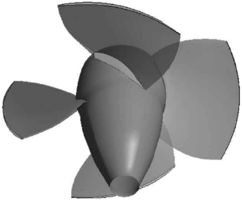
Guillaume Garrigos

Istituto Italiano di Tecnologia & Massachusetts Institute of Technology
Genova, Italie

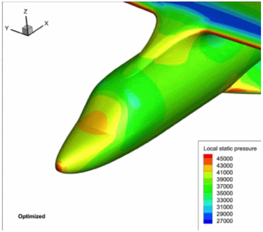
Journées SMAI-MODE
24 Mars, 2016

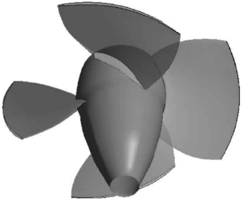
In engineering, decision sciences, it happens that various objective functions shall be minimized simultaneously: $f_1, ..., f_m : H \longrightarrow \mathbb{R}$





In engineering, decision sciences, it happens that various objective functions shall be minimized simultaneously: $f_1, ..., f_m : H \longrightarrow \mathbb{R}$





$\longrightarrow$ Needs appropriate tools: multi-objective optimization.

The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz, H Hilbert.

$$\text{Solve} \;\; \text{MIN}\; (f_1(x), ..., f_m(x)) \;:\; x \in C \subset H \text{ convex.}$$

The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz, H Hilbert.

$$\text{Solve} \;\; \text{MIN} \; (f_1(x), ..., f_m(x)) \; : \; x \in C \subset H \text{ convex.}$$

We consider the usual order(s) on $\mathbb{R}^m$:

$$a \preceq b \Leftrightarrow a_i \leq b_i \text{ for all } i = 1, ..., m,$$
$$a \prec b \Leftrightarrow a_i < b_i \text{ for all } i = 1, ..., m.$$

The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz, H Hilbert.

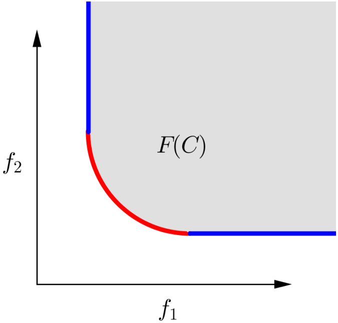
$$\text{Solve} \quad \text{MIN } (f_1(x), ..., f_m(x)) \; : \; x \in C \subset H \text{ convex.}$$

We consider the usual order(s) on $\mathbb{R}^m$:

$$a \preceq b \Leftrightarrow a_i \leq b_i \text{ for all } i = 1, ..., m,$$
$$a \prec b \Leftrightarrow a_i < b_i \text{ for all } i = 1, ..., m.$$

x is a Pareto point if
$\nexists y \in C$ such that $F(y) \precneqq F(x)$

x is a weak Pareto point if
$\nexists y \in C$ such that $F(y) \prec F(x)$



The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz.

$$\text{Solve} \;\; \text{MIN} \;\; f_1(x), ..., f_m(x) \;\; : \;\; x \in C \subset H \text{ convex.}$$

How to solve it?

The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz.

$$\text{Solve} \quad \text{MIN} \quad f_1(x), ..., f_m(x) \;:\; x \in C \subset H \text{ convex.}$$

How to solve it?

- genetic algorithm $\longrightarrow$ no theoretical guarantees.

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz.

$$\text{Solve} \quad \text{MIN} \quad f_1(x), ..., f_m(x) \quad : \quad x \in C \subset H \text{ convex.}$$

How to solve it?

- genetic algorithm $\longrightarrow$ no theoretical guarantees.
- scalarization method:

$$\bigcup_{\theta \in \Delta^m} \underset{x \in H}{\text{argmin}}\ f_\theta(x) \subset \{\text{weak Paretos}\} \subset \{\text{Paretos}\},$$

where Δ^m is the simplex unit and $f_\theta(x) := \sum_{i=1}^{m} \theta_i f_i(x)$.

The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz.

$$\text{Solve} \;\; \text{MIN} \;\; f_1(x), ..., f_m(x) \;\; : \;\; x \in C \subset H \text{ convex.}$$

We are going to present a method which:

The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz.

$$\text{Solve } \ \text{MIN } \ f_1(x), ..., f_m(x) \ : \ x \in C \subset H \text{ convex.}$$

We are going to present a method which:

- generalizes the gradient descent dynamic $\dot{x}(t) + \nabla f(x(t)) = 0$,

The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz.

$$\text{Solve } \text{MIN } f_1(x), ..., f_m(x) \ : \ x \in C \subset H \text{ convex.}$$

We are going to present a method which:

- generalizes the gradient descent dynamic $\dot{x}(t) + \nabla f(x(t)) = 0$,
- is *cooperative*, i.e. all objective functions decrease simultaneously,

The multi-objective optimization problem

Let $F = (f_1, ..., f_m) : H \to \mathbb{R}^m$ locally Lipschitz.

$$\text{Solve} \;\; \text{MIN} \;\; f_1(x), ..., f_m(x) \;\; : \;\; x \in C \subset H \text{ convex.}$$

We are going to present a method which:

- generalizes the gradient descent dynamic $\dot{x}(t) + \nabla f(x(t)) = 0$,
- is *cooperative*, i.e. all objective functions decrease simultaneously,
- is independent of any choice of parameters.

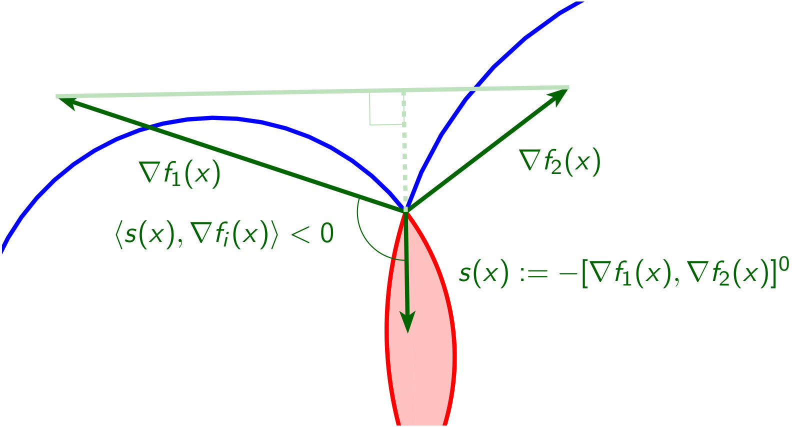
Single objective optimization:

$$x_{n+1} = x_n + \lambda_n d_n,$$

where d_n satisfies $df(x_n; d_n) < 0$ (e.g. $d_n = -\nabla f(x_n)$).

Multi-objective optimization:
Can we find d_n such that $df_i(x_n; d_n) < 0$ for all $i \in \{1, ..., m\}$?

Cornet (1981)


$\nabla f_1(x)$
$\nabla f_2(x)$
$\langle s(x), \nabla f_i(x) \rangle < 0$
$s(x) := -[\nabla f_1(x), \nabla f_2(x)]^0$

Multi-objective steepest descent

Let $F = (f_1, ..., f_m) : H \longrightarrow \mathbb{R}^m$ locally Lipschitz, $C = H$ Hilbert.

Definition

For all $x \in H$, $s(x) := -\left(\mathrm{co}\,\{\partial^c f_i(x)\}_{i=1,...,m}\right)^0$ is the (common) steepest descent direction at x.

Multi-objective steepest descent

Let $F = (f_1, ..., f_m) : H \longrightarrow \mathbb{R}^m$ locally Lipschitz, $C = H$ Hilbert.

Definition

For all $x \in H$, $s(x) := -\left(\mathrm{co}\,\{\partial^c f_i(x)\}_{i=1,...,m}\right)^0$ is the (common) steepest descent direction at x.

Remarks in the smooth case

- If $m = 1$ then $s(x) = -\nabla f_1(x)$.

Multi-objective steepest descent

Let $F = (f_1, ..., f_m) : H \longrightarrow \mathbb{R}^m$ locally Lipschitz, $C = H$ Hilbert.

Definition

For all $x \in H$, $s(x) := -\left(\mathrm{co}\,\{\partial^c f_i(x)\}_{i=1,...,m}\right)^0$ is the (common) steepest descent direction at x.

Remarks in the smooth case

- If $m = 1$ then $s(x) = -\nabla f_1(x)$.
- At each x, $s(x)$ selects a convex combination:

$$s(x) = -\sum_{i=1}^{m} \theta_i(x) \nabla f_i(x) = -\nabla f_{\theta(x)}(x) \text{ where } f_{\theta(x)} = \sum_{i=1}^{m} \theta_i(x) f_i.$$

Multi-objective steepest descent

Let $F = (f_1, ..., f_m) : H \longrightarrow \mathbb{R}^m$ locally Lipschitz, $C = H$ Hilbert.

Definition

For all $x \in H$, $s(x) := -\left(\mathrm{co}\,\{\partial^c f_i(x)\}_{i=1,...,m}\right)^0$ is the (common) steepest descent direction at x.

Remarks in the smooth case

- If $m = 1$ then $s(x) = -\nabla f_1(x)$.
- At each x, $s(x)$ selects a convex combination:

$$s(x) = -\sum_{i=1}^{m} \theta_i(x)\nabla f_i(x) = -\nabla f_{\theta(x)}(x) \text{ where } f_{\theta(x)} = \sum_{i=1}^{m} \theta_i(x) f_i.$$

- $s(x)$ is the steepest descent:

$$\frac{s(x)}{\|s(x)\|} = \underset{d \in \mathbb{B}_H}{\mathrm{argmin}} \left\{ \max_{i=1,...,m} \langle \nabla f_i(x), d \rangle \right\}.$$

Algorithm:

$$x_{n+1} = x_n + \lambda_n s(x_n).$$

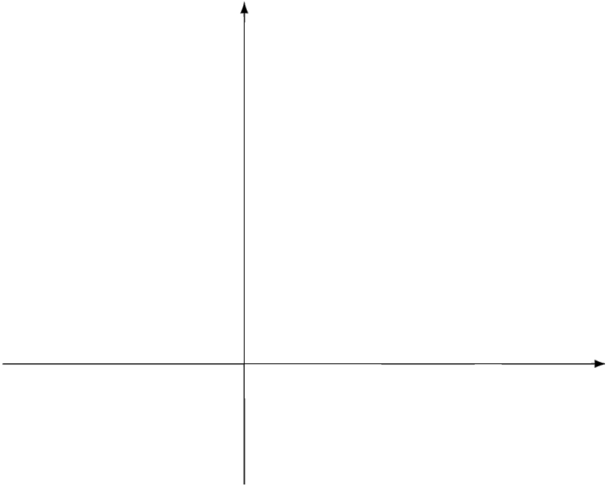
Studied in the 2000's by Svaiter, Fliege, Iusem, ...

Continuous dynamic:

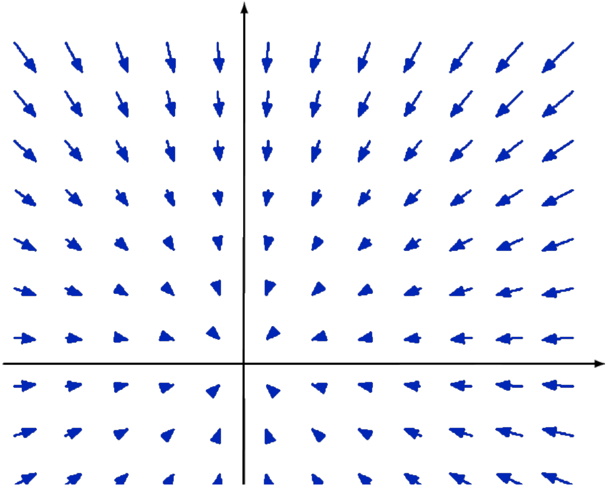
$$\text{(SD)} \quad \dot{x}(t) = s(x(t)),$$

$$\text{i.e.} \quad \text{(SD)} \quad \dot{x}(t) + \left(\mathrm{co}\,\{\partial^c f_i(x(t))\}_i\right)^0 = 0$$

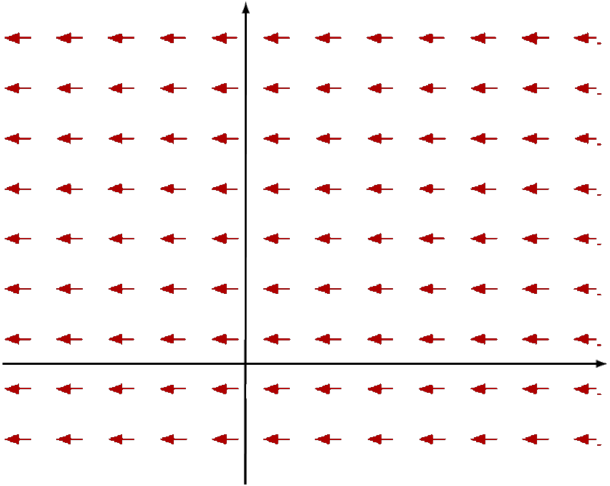
$$\text{(SD)} \quad \dot{x}(t) = s(x(t)) \quad \text{with } f_1(x) = \|x\|^2 \text{ and } f_2(x) = x_1.$$



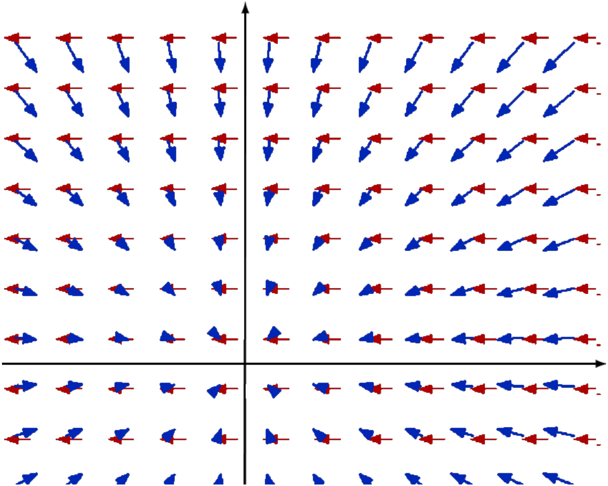
$$\text{(SD)} \quad \dot{x}(t) = s(x(t)) \text{ with } f_1(x) = \|x\|^2 \text{ and } f_2(x) = x_1.$$



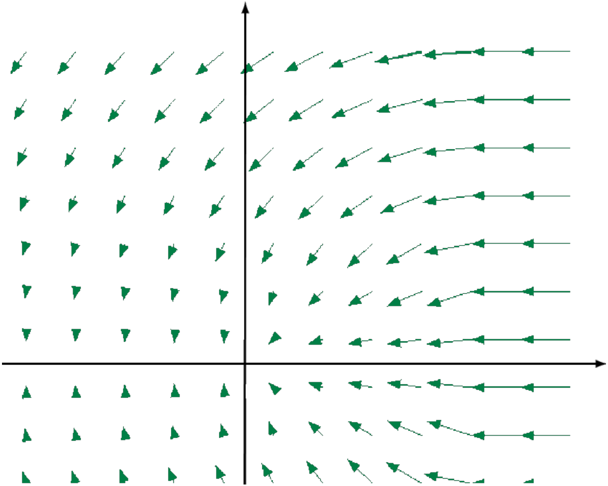
$$\text{(SD)} \quad \dot{x}(t) = s(x(t)) \quad \text{with} \quad f_1(x) = \|x\|^2 \text{ and } f_2(x) = x_1.$$



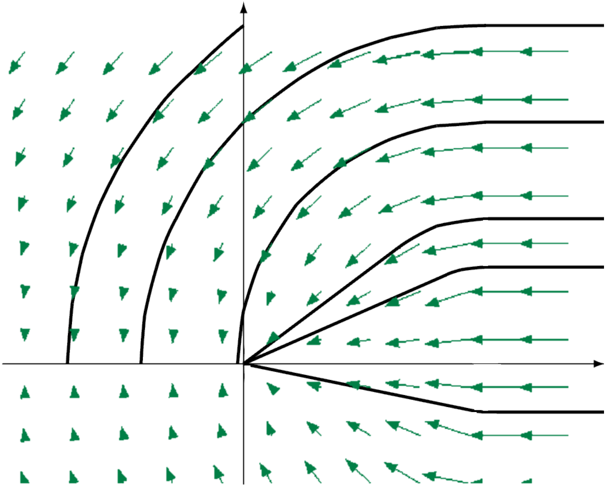
(SD) $\quad \dot{x}(t) = s(x(t))$ with $\quad f_1(x) = \|x\|^2$ and $f_2(x) = x_1$.



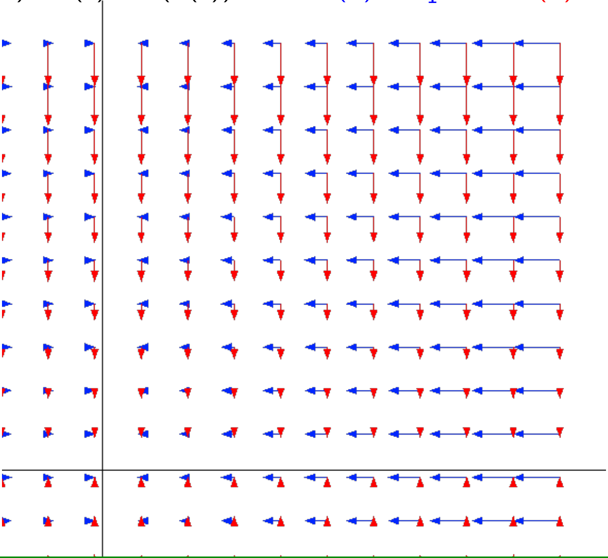
$$\text{(SD)} \quad \dot{x}(t) = s(x(t)) \quad \text{with} \qquad f_1(x) = \|x\|^2 \text{ and } f_2(x) = x_1.$$



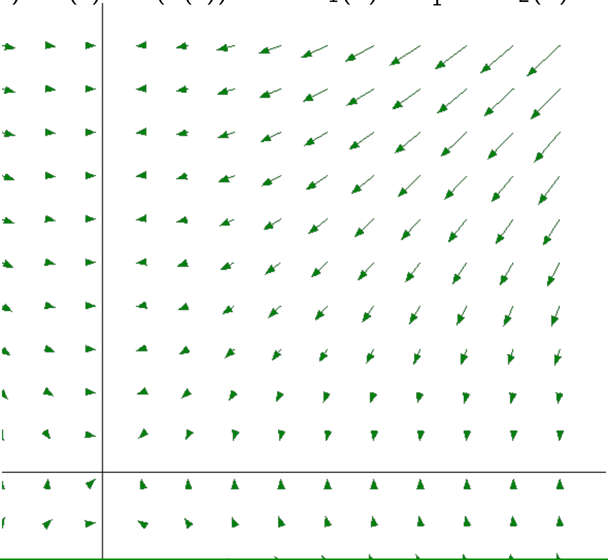
$$\text{(SD)} \quad \dot{x}(t) = s(x(t)) \quad \text{with} \qquad f_1(x) = \|x\|^2 \text{ and } f_2(x) = x_1.$$



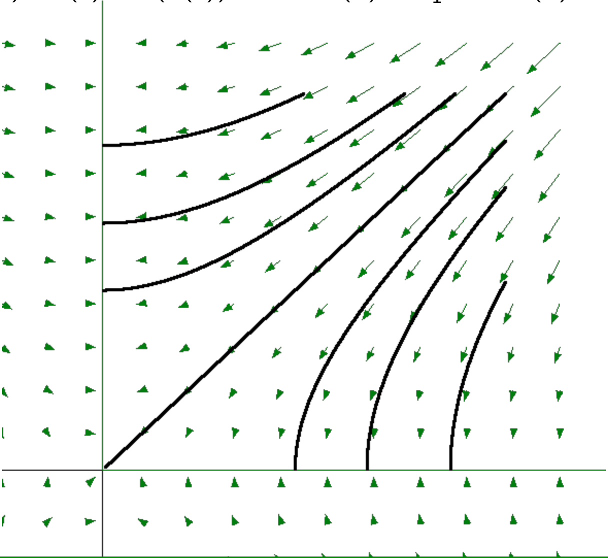
$$\text{(SD)} \quad \dot{x}(t) = s(x(t)) \quad \text{with } f_1(x) = x_1^2 \text{ and } f_2(x) = x_2^2.$$



$$\text{(SD)} \quad \dot{x}(t) = s(x(t)) \quad \text{with} \quad f_1(x) = x_1^2 \text{ and } f_2(x) = x_2^2.$$



$$\text{(SD)} \quad \dot{x}(t) = s(x(t)) \ \text{ with } \ f_1(x) = x_1^2 \text{ and } f_2(x) = x_2^2.$$



A cooperative dynamic

Let $x : \mathbb{R}_+ \longrightarrow H$ be a solution of (SD) $\dot{x}(t) = s(x(t))$.
For all $i = 1, ..., m$, the function $t \mapsto f_i(x(\cdot))$ is decreasing.

A cooperative dynamic

Let $x : \mathbb{R}_+ \longrightarrow H$ be a solution of (SD) $\dot{x}(t) = s(x(t))$.
For all $i = 1, ..., m$, the function $t \mapsto f_i(x(\cdot))$ is decreasing.

Convergence in the convex case

Assume that the objective functions are convex. Then any bounded trajectory weakly converges to a weak Pareto point.

A cooperative dynamic

Let $x : \mathbb{R}_+ \longrightarrow H$ be a solution of (SD) $\dot{x}(t) = s(x(t))$.
For all $i = 1, ..., m$, the function $t \mapsto f_i(x(\cdot))$ is decreasing.

Convergence in the convex case

Assume that the objective functions are convex. Then any bounded trajectory weakly converges to a weak Pareto point.

Existence in the convex case

Suppose that H is finite dimensional. Then, for any initial data, there exists a global solution to (SD).

- In case of convex constraint $C \subset H$:

$$\text{(SD)} \quad \dot{x}(t) + \left(N_C(x(t)) + \operatorname{co}\left\{\partial^c f_i(x(t))\right\}_i\right)^0 = 0.$$

How to discretize it properly?

- In case of convex constraint $C \subset H$:

$$(\text{SD}) \quad \dot{x}(t) + \left(N_C(x(t)) + \operatorname{co}\{\partial^c f_i(x(t))\}_i\right)^0 = 0.$$

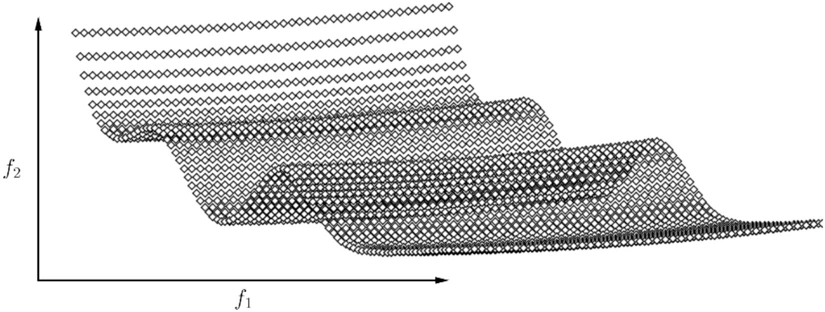
  How to discretize it properly?
- Uniqueness? Yes, if $\{\nabla f_i(x(\cdot))\}_{i=1,\dots,m}$ are affinely independants.

- In case of convex constraint $C \subset H$:

$$\text{(SD)} \quad \dot{x}(t) + \left(N_C(x(t)) + \mathrm{co}\ \{\partial^c f_i(x(t))\}_i\right)^0 = 0.$$

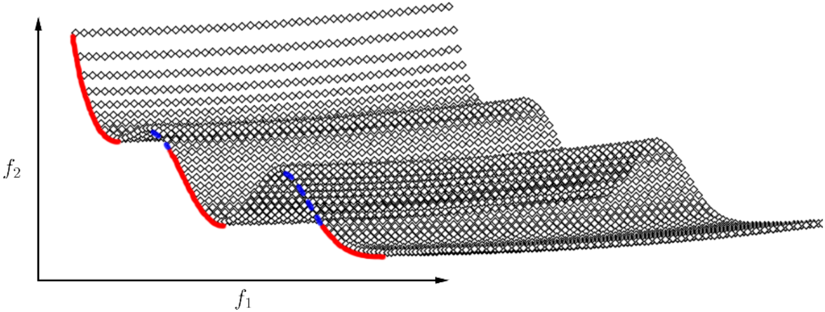
  How to discretize it properly?
- Uniqueness? Yes, if $\{\nabla f_i(x(\cdot))\}_{i=1,\dots,m}$ are affinely independants.
- Convergence to Pareto points? Guaranteed by endowing $\mathbb{R}^m$ with a different order (but some of the Paretos might be lost in the operation).

$$f_1(x,y) = x + y$$
$$f_2(x,y) = x^2 + y^2 + \frac{1}{x} + 3e^{-100(x-0.3)^2} + 3e^{-100(x-0.6)^2}$$
$$(x,y) \in C = [0.1,1]^2$$



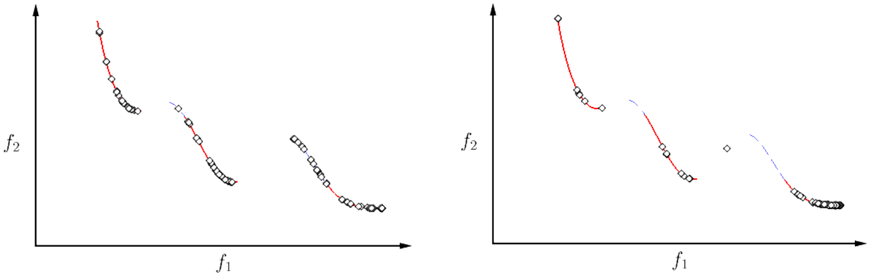
Plot of $F(C)$, $F = (f_1, f_2) : C \longrightarrow \mathbb{R}^2$.

$$f_1(x,y) = x + y$$
$$f_2(x,y) = x^2 + y^2 + \frac{1}{x} + 3e^{-100(x-0.3)^2} + 3e^{-100(x-0.6)^2}$$
$$(x,y) \in C = [0.1, 1]^2$$



Plot of $F(C)$, $F = (f_1, f_2) : C \longrightarrow \mathbb{R}^2$ and its pareto front.

$$f_1(x,y) = x + y$$
$$f_2(x,y) = x^2 + y^2 + \frac{1}{x} + 3e^{-100(x-0.3)^2} + 3e^{-100(x-0.6)^2}$$
$$(x,y) \in C = [0.1, 1]^2$$



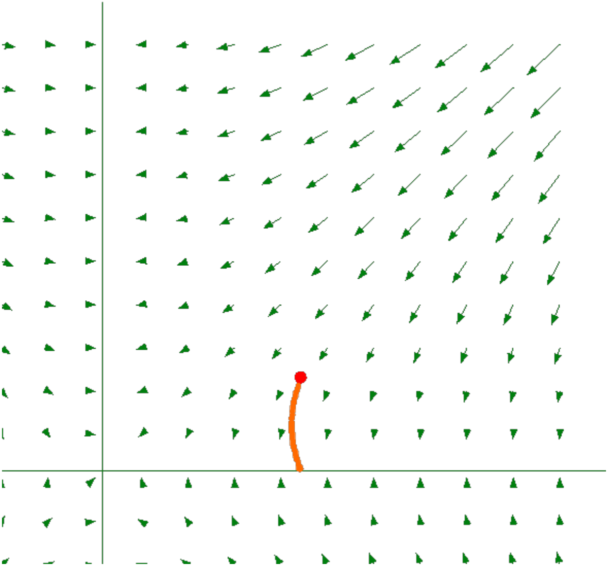
Gradient method (Right) vs Scalar method (Left). 100 samples.

Can we select, among the weak Paretos (= the zeros of $x \mapsto s(x)$) the closest to a desired state?

Can we select, among the weak Paretos (= the zeros of $x \mapsto s(x)$) the closest to a desired state?

$\rightarrow$ Tikhonov regularization

$$\dot{x}(t) - s(x(t)) + \varepsilon(x(t) - x_d) = 0, \varepsilon > 0.$$



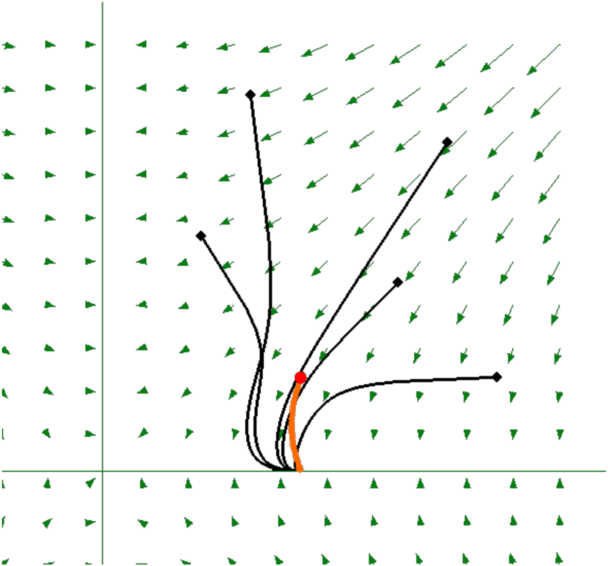
Can we select, among the weak Paretos (= the zeros of $x \mapsto s(x)$) the closest to a desired state?

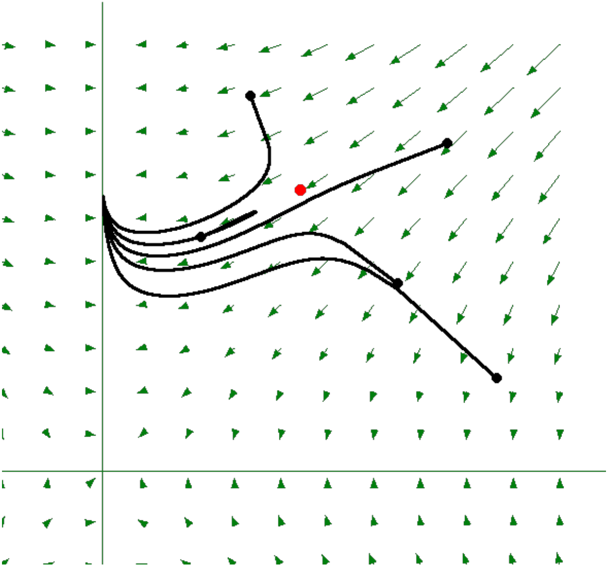
$\rightarrow$ Diagonal Tikhonov regularization

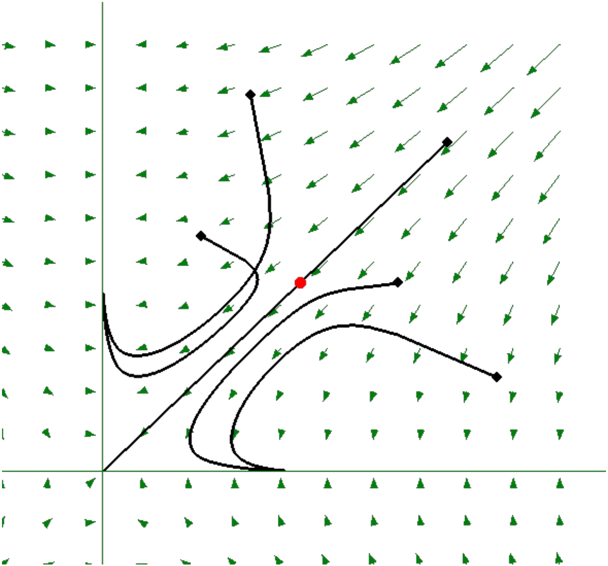
$$\dot{x}(t) - s(x(t)) + \varepsilon(t)(x(t) - x_d) = 0,$$

$$\varepsilon(t) \downarrow 0, \int_0^{\infty} \varepsilon(t)\ dt = +\infty.$$

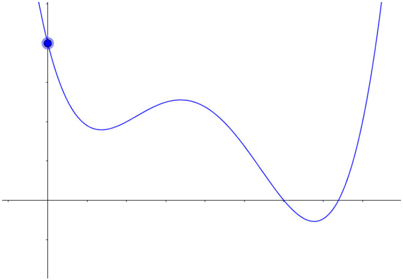
See the works of Attouch, Cabot, Czarnecki, Peypouquet (...) in the monotone case.

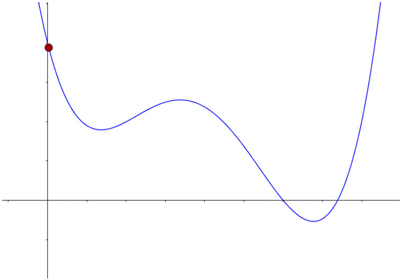






What about inertial dynamics?





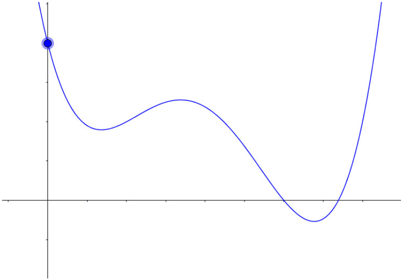
$$\dot{x}(t) + \nabla f(x(t)) = 0$$

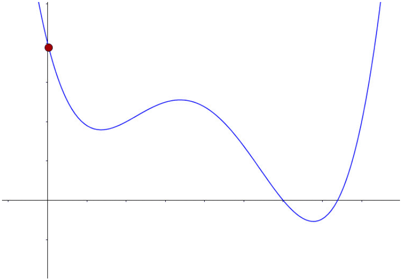
$$x_{n+1} = x_n - \lambda \nabla f(x_n)$$

$$\ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0$$

$$x_{n+1} = y_n - \lambda \nabla f(y_n)$$
$$y_{n+1} = x_{n+1} + (1-\gamma)(x_{n+1} - x_n)$$

What about inertial dynamics?





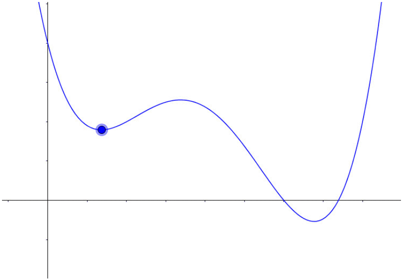
$$\dot{x}(t) + \nabla f(x(t)) = 0$$

$$\ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0$$

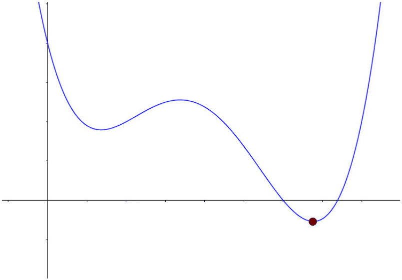
$$x_{n+1} = x_n - \lambda \nabla f(x_n)$$

$$x_{n+1} = y_n - \lambda \nabla f(y_n)$$
$$y_{n+1} = x_{n+1} + (1-\gamma)(x_{n+1} - x_n)$$

What about inertial dynamics?



$$\dot{x}(t) + \nabla f(x(t)) = 0$$



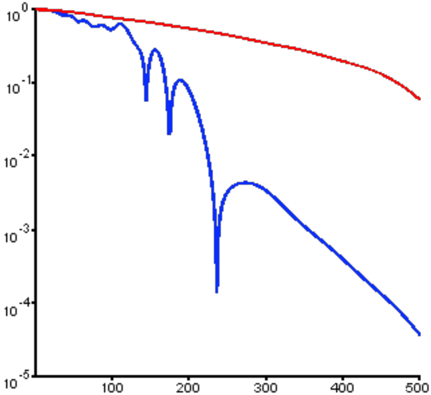
$$\ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0$$

Inertia promotes

- Faster trajectories (varying γ),
- Exploratory properties.

Convergence rates : empirical observation

$$f_1(x)=\left(\sum_{i=1}^{10} x_i^2-10cos(2\pi x_i)+10\right)^{\frac{1}{4}},\ f_2(x)=\left(\sum_{i=1}^{10}(x_i-1.5)^2-10cos(2\pi(x_i-1.5))+10\right)^{\frac{1}{4}}$$



Convergence rate of $\|F(x^n) - F(x^\infty)\|_\infty$:
Steepest Descent vs Inertial Steepest Descent

Inertial (multi-objective) Steepest Descent

Let $f_1, ..., f_m$ be smooth, with L-Lipschitz gradient.

$$\text{(ISD)}\ \ddot{x}(t) = -\gamma \dot{x}(t) + s(x(t)).$$

Let $f_1, ..., f_m$ be smooth, with L-Lipschitz gradient.

$$\text{(ISD)}\ \ddot{x}(t) = -\gamma \dot{x}(t) + s(x(t)).$$

Example: $f_1(x) = \|x\|^2$ and $f_2(x) = x_1$.

Let $f_1, ..., f_m$ be smooth, with L-Lipschitz gradient.

$$\text{(ISD)} \; m\ddot{x}(t) = -\gamma \dot{x}(t) + s(x(t)).$$

Assume that $\gamma \geq L$.

Let $f_1, \dots, f_m$ be smooth, with L-Lipschitz gradient.

$$\text{(ISD)} \ m\ddot{x}(t) = -\gamma\dot{x}(t) + s(x(t)).$$

Assume that $\gamma \geq L$.

Existence

Suppose that H is finite dimensional. Then, for any initial data, there exists a global solution to (ISD).

Let $f_1, ..., f_m$ be smooth, with L-Lipschitz gradient.

$$\text{(ISD)}\ m\ddot{x}(t) = -\gamma \dot{x}(t) + s(x(t)).$$

Assume that $\gamma \geq L$.

Existence

Suppose that H is finite dimensional. Then, for any initial data, there exists a global solution to (ISD).

Convergence in the convex case

Let $f_1, ..., f_m$ be convex. Then, any bounded trajectory weakly converges to a weak Pareto point.

Conclusion

The steepest descent provides a flexible tool once adapted to multi-objective optimization problems.

Conclusion

The steepest descent provides a flexible tool once adapted to multi-objective optimization problems.

Open questions:

- Understand the asymptotic behaviour of

$$\dot{x}(t) - s(x(t)) + \varepsilon(t)x(t) = 0$$

(the set of weak Paretos is non-convex).

Conclusion

The steepest descent provides a flexible tool once adapted to multi-objective optimization problems.

Open questions:

- Understand the asymptotic behaviour of

$$\dot{x}(t) - s(x(t)) + \varepsilon(t)x(t) = 0$$

  (the set of weak Paretos is non-convex).
- Having convergence rates for first and second-order dynamics (the critical values are not unique).

Thank you for your attention !