Lecture 1. Essentials of numerical nonsmooth optimization (NSO)

M. Gaudioso,*

*DIMES-Universitá della Calabria and ICAR-CNR, Rende (CS), Italia

EUROPT Summer School

August 30th– September 1st, 2021

Outline

Nonsmooth optimization (NSO)

- Nonsmooth optimization (NSO) (or Nondifferentiable optimization (NDO)), is about problems where the objective function exhibits kinks, that is discontinuities of the derivatives.
- The simplest example of a nonsmooth function $f : \mathbb{R} \mapsto \mathbb{R}$:

$$f(x) = |x|$$

Classic calculus does not provide any characterization of the minimum point $x^* = 0$

Where NSO problems come from

- *Worst case-based* decision making: to minimize the cost under the worst possible scenario. Thus $minmax$ or $maxmin$ problems.
- $Minmaxmin$ models: worst–case formulations with both "wait and see" and "here and now" decision variables, with the realization of a scenario in the middle.
- *Right-hand-side* decomposition of large scale problems controlled by a *master* assigning resources to each subproblem.
- *Lagrangian relaxation* of Integer or Mixed-Integer programs, often coupled with efficient Lagrangian heuristics, resulting in the optimization of a piecewise affine (hence nonsmooth) function.
- *Bilevel problems*, based on a hierarchy of two autonomous decision makers.

The history of NSO

- Function approximation, Chebyshëv in XIX Century.
- Subgradient method, second half of XX Century.
- Cutting plane method, the sixties of XX Century.
- Bundle method, the seventies of XX century.

A fundamental question

Most of the functions of practical interest are differentiable almost everywhere. Thus why don't we apply the algorithms for solving differentiable problems to the nonsmooth ones?

- Convergence to a wrong point!

Wolfe's example, 1975

$$f(x_1, x_2) = \begin{cases} 5(9x_1^2 + 16x_2^2)^{\frac{1}{2}} & for\ x_1 > |x_2| \\ 9x_1 + 16|x_2| & for\ x_1 \leq |x_2| \end{cases}$$

f is differentiable everywhere but on the half-line $\{x_1, 0\}$, $x_1 \leq 0$, where $f \to -\infty$. Steepest descent converges to $(0, 0)$!

P. Wolfe / A method of conjugate subgradients

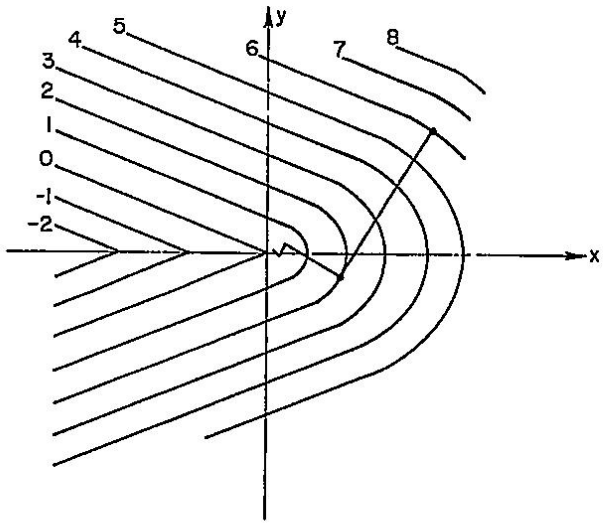


Fig. 1. Contours of f and steepest descent path.

Outline

The problem

We consider the unconstrained minimization problem

$$\min\left\{f(x) : x \in \mathbb{R}^n\right\},$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and nonsmooth.
- f is finite over $\mathbb{R}^n$, hence it is proper.
- f has finite minimum f^* which is attained at a nonempty convex compact set $M^* \subset \mathbb{R}^n$ and any point in M^*, is denoted by x^*

For a given $\epsilon > 0$, an ϵ-approximate solution is any point $x \in \mathbb{R}^n$ such that

$$f(x) < f^* + \epsilon.$$

Subgradients and Subdifferential

A *subgradient* of f at $x \in \mathbb{R}^n$ is any vector $g \in \mathbb{R}^n$ satisfying the following (subgradient)-inequality

$$f(y) \geq f(x) + g^\top (y - x) \quad \forall y \in \mathbb{R}^n.$$

The *subdifferential* of f at $x \in \mathbb{R}^n$, denoted by $\partial f(x)$, is the set of all the subgradients of f at x, i.e.,

$$\partial f(x) \triangleq \left\{ g \in \mathbb{R}^n : f(y) \geq f(x) + g^\top (y - x) \ \ \forall y \in \mathbb{R}^n \right\}.$$

At any point x where f is differentiable, the subdifferential is a singleton and it is $\partial f(x) = \{\nabla f(x)\}$.

The ϵ-subgradient

For $\epsilon \geq 0$, an *ϵ-subgradient* of f at x is any vector $g \in \mathbb{R}^n$ fulfilling

$$f(y) \geq f(x) + g^\top (y - x) - \epsilon \quad \forall y \in \mathbb{R}^n,$$

and the *ϵ-subdifferential* of f at x, denoted by $\partial f_\epsilon(x)$, is the set of all the ϵ-subgradients of f at x, i.e.,

$$\partial_\epsilon f(x) \triangleq \left\{ g \in \mathbb{R}^n : f(y) \geq f(x) + g^\top (y - x) - \epsilon \ \forall y \in \mathbb{R}^n \right\}.$$

It is

$$\partial_0 f(x) = \partial f(x).$$

Directional differentiability of convex functions

The set $\partial f(\cdot)$ is a convex, bounded and closed. The directional derivative at any x, along the direction $d \in \mathbb{R}^n$, is indicated as $f'(x,d)$ and it is

$$f'(x,d) \triangleq \lim_{t \downarrow 0} \frac{f(x+td) - f(x)}{t} = \max_{g \in \partial f(x)} g^\top d.$$

At a point x where f is differentiable and the subdifferential is a singleton, we retrace the classic formula

$$f'(x,d) = \nabla f(x)^\top d$$

Definition

Any direction $d \in \mathbb{R}^n$ is a descent direction at x if there exists $\bar{t} > 0$ such that

$$f(x+td) < f(x) \quad \forall t \in (0, \bar{t}].$$

Moreover the following equivalence holds true

$$f'(x,d) < 0 \quad \Leftrightarrow \quad d \text{ is a descent direction at } x.$$

Extension to nonconvex functions

f is *locally* Lipschitz, that is Lipschitz on every bounded set C ($\exists L$ such that $|f(x) - f(y)| \leq L\|x - y\|$, $\forall x, y \in C$)

$$\Downarrow$$

f is differentiable almost everywhere

There exists at each point x the generalized gradient (or Clarke's gradient, or subdifferential):

$$\partial_C f(x) \triangleq \text{conv}\{g : g \in \mathbb{R}^n, \nabla f(x_k) \to g,\ x_k \to x,\ x_k \notin \Omega_f\},$$

where Ω_f is the set (of zero measure) where f is not differentiable.

Optimality conditions

Function f convex:

$$x^* \text{ is a } minimum \Leftrightarrow 0 \in \partial f(x)$$

Function f locally Lipschitz:

$$x^* \text{ is a } minimum \Rightarrow 0 \in \partial_C f(x)$$

Descent directions

For smooth functions, $\nabla f(x)$ provides complete information about the directional derivative:

$$f'(x,d) = \nabla f(x)^\top d.$$

For nonsmooth functions, calculation of the directional derivative, requires a maximization process over the entire subdifferential,

$$f'(x,d) = \max_{g \in \partial f(x)} g^\top d$$

To be a descent one, any direction has to form an obtuse angle with all the subgradients

The steepest descent direction:

$$d^* = \arg\min \left\{ f'(x,d) : \|d\| \leq 1,\ d \in \mathbb{R}^n \right\}.$$

The stepest descent

To calculate the steepest descent apply the $minmax$ theorem and solve

$$\min_{\|d\|\leq 1} \max_{g\in\partial f(x)} g^\top d = \max_{g\in\partial f(x)} \min_{\|d\|\leq 1} g^\top d = \max_{g\in\partial f(x)} -\|g\| = -\min_{g\in\partial f(x)} \|g\|,$$

that is find the minimum norm point in the compact convex set $\partial f(x)$.

Most of the available NSO methods do not follow the steepest descent philosophy.

Outline

Classification of methods

We adopt the following classification of the NSO iterative algorithms:

- *Single-point* models: the new iterate x_{k+1} is calculated by exploiting only the available information at x_k (typically the objective function value and a subgradient).
- *Multi-point* models: they exploit similar local information about x_k, but also data coming from several other points (typically $x_1, \ldots, x_{k-1}$), no matter how far from x_k they are.

A local approximation of f versus the approximation of S_k, the level set at x_k, $S_k \triangleq \{x : f(x) \leq f(x_k)\}$.

Monotonicity of the sequence $f(x_k)$ is not a *must* in NSO!

Outline

The subgradient method with constant stepsize (SM)

The basic iteration scheme of the subgradient method (N.Z. Shor) is:

$$x_{k+1} = x_k - t\frac{g_k}{\|g_k\|},$$

where $g_k \in \partial f(x_k)$ and $t > 0$ is a constant *stepsize*.

At any $x \notin M^*$ the anti–subgradient direction $d = -g$, $g \in \partial f(x)$, although it is $g^\top d = -\|g\|^2 < 0$, is not necessarily a descent direction.

Definition

Any direction d is a *minimum approaching* direction at x, if there exists $\bar{t} > 0$ such that

$$\|x + td - x^*\| < \|x - x^*\|, \quad \forall t \in (0, \bar{t}).$$

Minimum approaching direction

Proposition

$d = -g$ is a minimum approaching direction.

Proof.

Convexity of $f \Rightarrow f(x^*) \geq f(x) + g^\top(x^* - x) \Rightarrow g^\top(x^* - x) < 0$.
It is

$$\begin{aligned} \|x + td - x^*\|^2 &= \|x - tg - x^*\|^2 \\ &= \|x - x^*\|^2 + t(t\|g\|^2 + 2g^\top(x^* - x)) \\ &< \|x - x^*\|^2 \end{aligned}$$

for every $0 < t < \bar{t} = \dfrac{2g^\top(x - x^*)}{\|g\|^2}$). □

An example

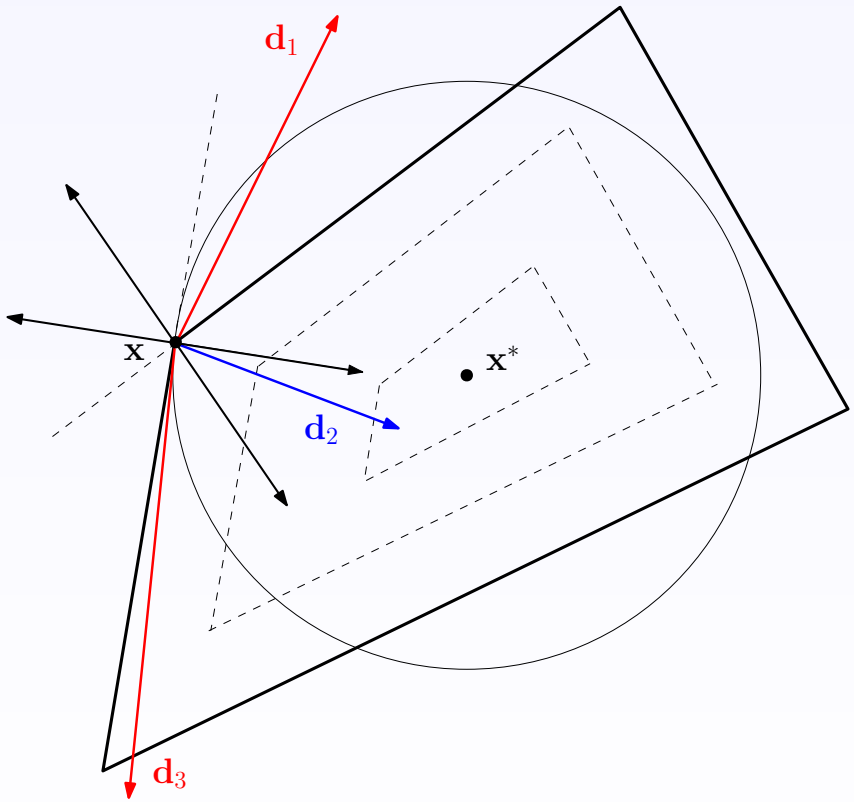


Figure: Descent and/or minimum approaching directions

Convergence of SM with constant stepsize

Theorem

For every $\epsilon > 0$ and $x^* \in M^*$ there exist a point $\bar{x}$ and an index $\bar{k}$ such that

$$\|\bar{x} - x^*\| < \frac{t}{2}(1+\epsilon)$$

and

$$f(x_{\bar{k}}) = f(\bar{x}).$$

Pictorial representation

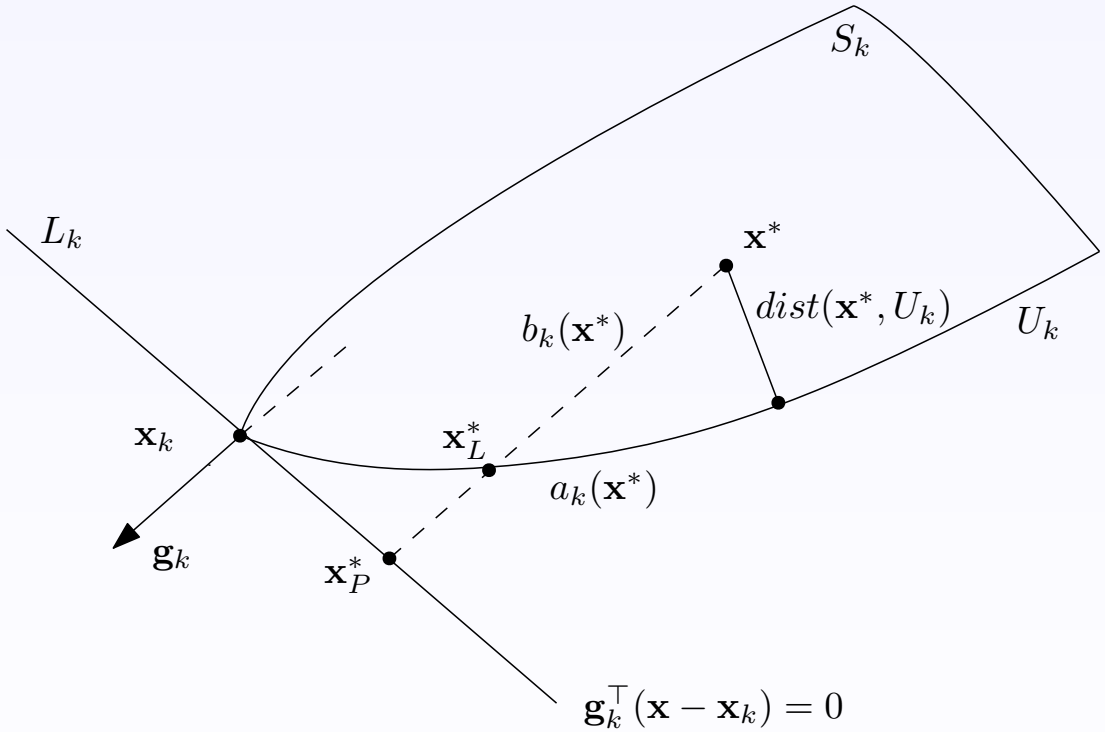


Figure: Convergence of the subgradient method

Proof

Proof.

Let $U_k = \{x \mid f(x) = f(x_k)\}$, $L_k = \{x \mid g_k^\top (x - x_k) = 0\}$ and $a_k(x^*) = \|x^* - x_P^*\|$. Observe

$$a_k(x^*) \geq b_k(x^*) = \|x^* - x_L^*\| \text{ and } b_k(x^*) \geq dist(x^*, U_k)$$

.

Since $a_k(x^*) = \dfrac{g_k^\top (x_k - x^*)}{\|g_k\|}$, it follows $dist(x^*, L_k) \leq b_k(x^*) \leq a_k(x^*) = \dfrac{g_k^\top (x_k - x^*)}{\|g_k\|}$.

Thus we obtain

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^* - t\frac{g_k}{\|g_k\|}\|^2 = \|x_k - x^*\|^2 + t^2 - 2ta_k(x^*)$$

$$\leq \|x_k - x^*\|^2 + t^2 - 2tb_k(x^*).$$

Next, suppose that $b_k(x^*) \geq t(1+\epsilon)/2, \ \forall k$. We would have:

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \epsilon t^2 \leq \|x_1 - x^*\|^2 - \epsilon k t^2,$$

which contradicts $\|x_{k+1} - x^*\|^2 \geq 0$ for all k. □ □

Classic subgradient method (SM)

Poor performance with constant stepsize. Thus subgradient method (SM) with adjustable stepsize:

$$x_{k+1} = x_k - t_k \frac{x_k}{\|x_k\|},$$

Theorem

Let f be convex and assume that M^* is bounded and nonempty, with $f^* = f(x^*)$. If the stepsize sequence $\{t_k\}$ satisfies the conditions

$$\lim_{k\to\infty} t_k = 0 \qquad \text{and} \qquad \sum_{k=1}^{\infty} t_k = \infty,$$

then either there exists an index k^* such that $x_{k^*} \in M^*$ or $\lim_{k\to\infty} f(x_k) = f^*$.

Possible choices of t_k are $t_k = c/k$, where c is any positive constant or $t_k = \frac{f(x_k)-f^*}{\|g_k\|}$ (the latter is known as Polyak's stepsize. Whenever f^* is unknown, it is usually replaced by any lower bound).

Speed of convergence

Proposition

Assume that

i) f *admits a sharp minimum, i.e., there exists* $\mu > 0$ *such that* $f(x) \geq f^* + \mu\|x - x^*\|$, *for every* $x \in \mathbb{R}^n$, *and*

ii) *the minimum value* f^* *is known, so that the Polyak stepsize can be calculated.*

Then, the subgradient method has linear convergence rate $q = \sqrt{1 - \frac{\mu^2}{c^2}}$, *where* c *is any upper bound on the norm of* g_k. *Moreover an* ϵ*-approximate solution is achieved in* $O(\frac{1}{\epsilon^2})$ *iterations.*

Remark

The monotonicity of the sequence $\{f(x_k)\}$ *is not ensured in (SM) while the minimum approaching nature is apparent. Note that the convergence rate* q *can be arbitrarily close to* 1.

SM variants

Improved (SM) approaches have been designed for classes of *weakly* structured problems.

- Nesterov's smoothing method (the bound on the number of iterations improves from $O(\frac{1}{\epsilon^2})$ to $O(\frac{1}{\epsilon})$).
- Nemirovski's Mirror Descent Algorithm (MDA)

Outline

Piecewise affine approximation of a convex function

Next iterate x_{k+1} is calculated on the basis of x_k and of several other points.

It is exploited the property that a convex function is the pointwise supremum of an infinite number of affine ones:

$$f(x) = \sup\left\{f(y) + g(y)^\top (x - y) : y \in \mathbb{R}^n\right\}, \quad g(y) \in \partial f(y).$$

Taking any finite set of points $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$, and letting

$$\ell_i(x) \triangleq f(x_i) + g_i^\top (x - x_i),\ g_i \in \partial f(x_i),\ i \in \{1, \ldots, k\}$$

obtain an approximation of f:

$$f_k(x) \triangleq \max\left\{\ell_i(x) : i \in \{1, \ldots, k\}\right\}.$$

Properties of f_k:

$$f_k(x) \leq f(x) \qquad \forall x \in \mathbb{R}^n,$$

$$f_k(x) = f(x_i), \quad g_i \in \partial f_k(x_i) \qquad \forall i \in \{1, \ldots, k\},$$

The cutting plane function

Summing up, f_k, the cutting plane function:

- is a *global* approximation of f. as it minorizes f everywhere, while interpolating it at points $x_1, \ldots, x_k$.
- is convex and piecewise affine, being the pointwise maximum of the affine functions $\ell_i(x)$, the linearizations of f rooted at $x_1, \ldots, x_k$

Remark

Even for the same set of points $x_1, \ldots, x_k$, the model function f_k can be not unique, since the subdifferential $\partial f(\cdot)$ is a multifunction.

The cutting plane (CP) function

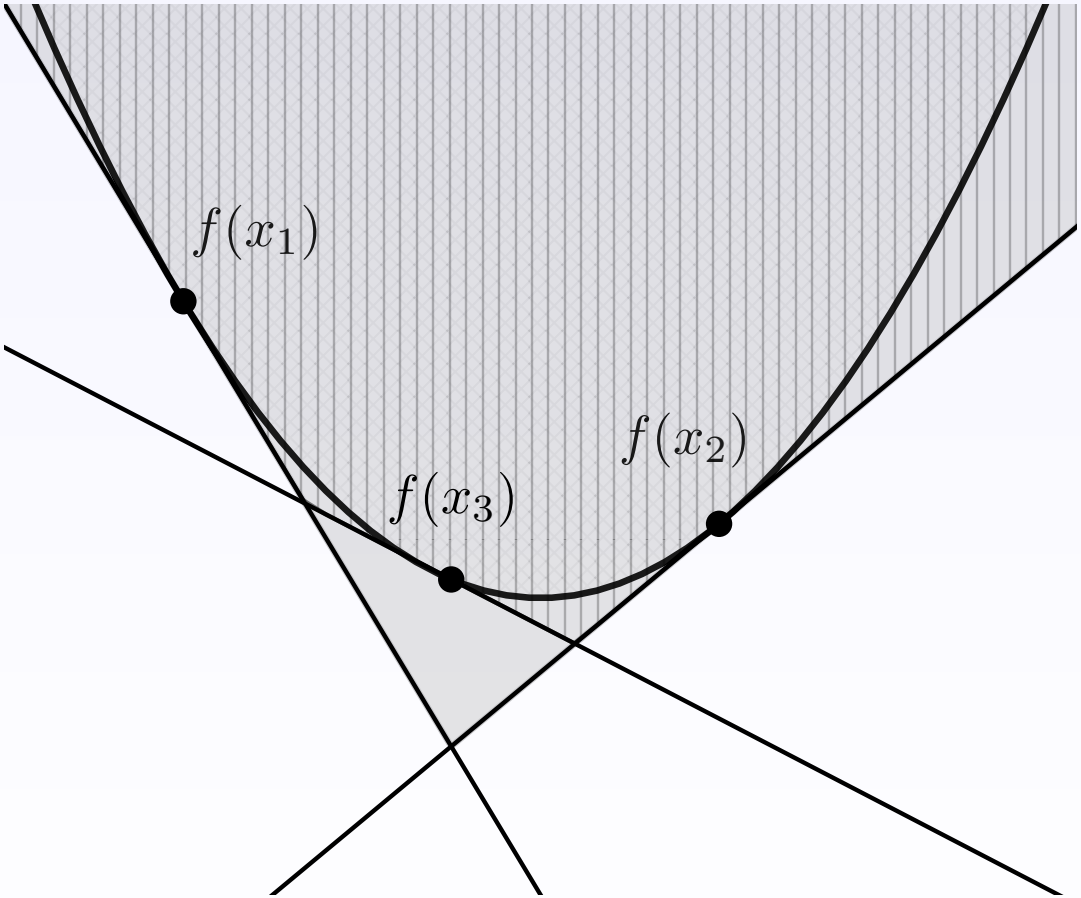


Figure: Cutting plane

Use f_k to select the next trial point:

$$x_{k+1} = \arg\min \{f_k(x) : x \in \mathbb{R}^n\},$$

Still a nonsmooth problem, equivalent to an LP (defined in $\mathbb{R}^{n+1}$).

$$LP \quad \begin{cases} \min\limits_{w,x} & w \\ & w \geq f(x_i) + g_i^T(x - x_i) \quad \forall i \in \{1, \ldots, k\} \end{cases}$$

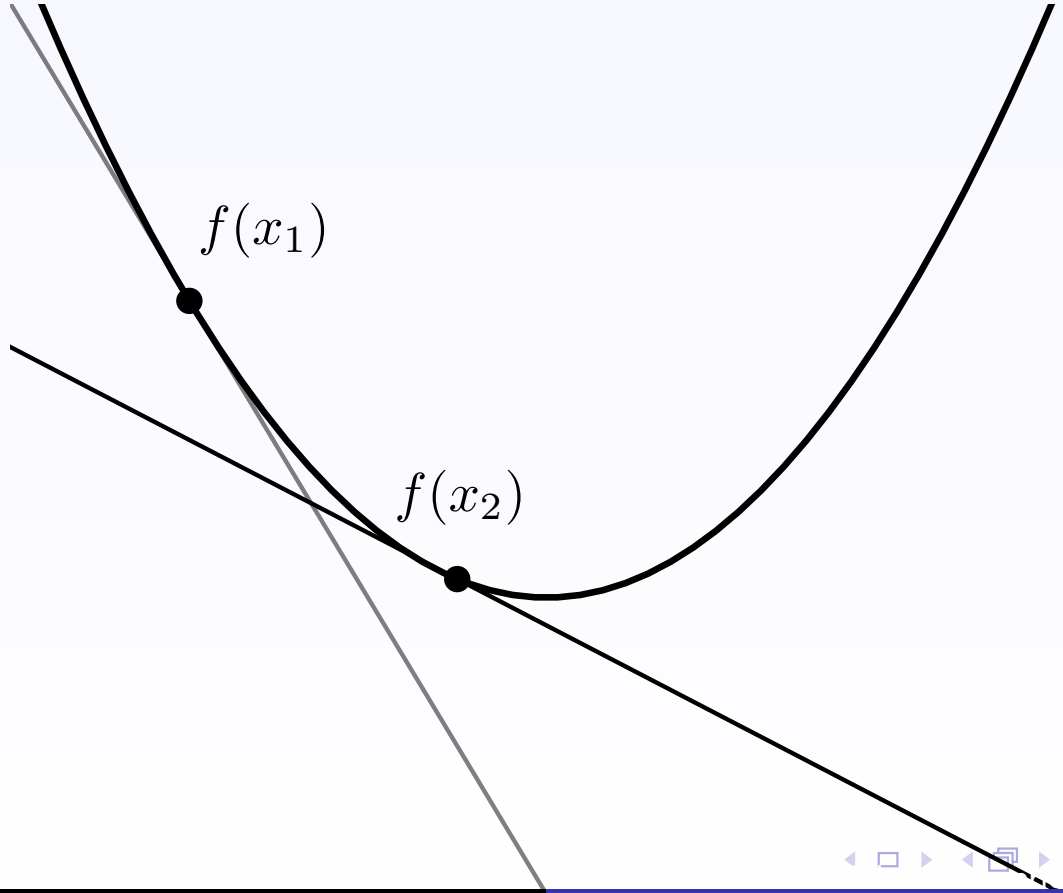
with optimal solution denoted by (x_{k+1}, w_k), $w_k = f_k(x_{k+1})$.
Boundedness of the LP requires dual feasibility:

$$\max \left\{ \sum_{i=1}^{k} \lambda_i (f(x_i) - g_i^\top x_i) : \sum_{i=1}^{k} \lambda_i = 1, \quad \underbrace{\sum_{i=1}^{k} \lambda_i g_i = 0, \ \lambda \geq 0}_{0 \in \operatorname{conv}\{g_i : i \in \{1, \ldots, k\}\}} \right\}.$$

A hard-to-test *qualification* of points $x_1, \ldots, x_k$.

An example of unboundedness

An example where the cutting plane function is unbounded (derivatives $f'(x_1)$ and $f'(x_2)$ both negative, thus $0 \notin [f'(x_1), f'(x_2)]$).



To avoid unboundedness put the problem in an (artificial) constrained optimization setting.

$$\min\left\{f(x):\ x\in Q\subset\mathbb{R}^n\right\},$$

Thus, the cutting-plane iteration becomes

$$x_{k+1}=\arg\min\left\{f_k(x):\ x\in Q\right\},$$

$f_k(x)$ provides a lower bound on f^*, the optimal value of f the sequence $\{f_k^*\}$ is monotonically nondecreasing and thus the lower bound becomes increasingly better.

(General) cutting plane method (CPM)

We state now the cutting plane method in a slightly more general setting (relax $x_{k+1} = \arg\min_{x \in Q} f_k(x)$ with $x_{k+1} \in S_k^* \triangleq \{x : x \in Q,\ f_k(x) \leq f^*\}$.

Algorithm: (General) cutting plane method (CPM)

Step 1 A starting point $x_1 \in \mathbb{R}^n$ and a stopping tolerance $\epsilon > 0$ are given. Calculate $g_1 \in \partial f(x_1)$, build $f_1(x)$, and set $k = 1$.

Step 2 Calculate $x_{k+1} \in S_k^* \triangleq \{x : x \in Q,\ f_k(x) \leq f^*\} \neq \emptyset$

Step 3 **If** $f(x_{k+1}) - f_k(x_{k+1}) < \epsilon$ **Then** set $x^* = x_{k+1}$ and **Exit**. **Else** Calculate $g_{k+1} \in \partial f(x_{k+1})$ and build $f_{k+1}(x)$

Step 4 Set $k = k + 1$ and **Go To** Step 2

Convergence

Theorem

CPM terminates at an ϵ-optimal point.

Proof.

Observe first that, since $f_k(x_{k+1}) \leq f^*$, satisfaction of the stopping condition at Step 3 implies $f(x_{k+1}) - f^* < \epsilon$ (ϵ-optimality).
Assume that the stopping condition $f(x_{k+1}) - f_k(x_{k+1}) \geq \epsilon$, $\forall k$.is never satisfied.

From convexity of f it is, for every $i \in \{1, \ldots, k\}$

$$\begin{aligned} f(x_i) &\geq f(x_{k+1}) + g_{k+1}^\top (x_i - x_{k+1}) \\ &\geq f_k(x_{k+1}) + \epsilon + g_{k+1}^\top (x_i - x_{k+1}) \\ &\geq f(x_i) + g_i^\top (x_{k+1} - x_i) + \epsilon + g_{k+1}^\top (x_i - x_{k+1}), \end{aligned}$$

which imply

$$0 \geq \epsilon - 2L_Q \|x_{k+1} - x_i\| \qquad \forall i \in \{1, \ldots, k\},$$

that is the sequence $\{x_k\}$ does not have an accumulation point, contradicting compactness of Q. □

Some comments on CPM

- The rationale of the choice of x_{k+1} is to stay *well inside* into the level set of f_k, possibly accommodating for *inexact* solution of LP.
- CPM is intrinsically nonmonotone (no objective function decrease is guaranteed at each iteration). Only monotonicity of the lower bound ($f_{k+1}(x_{k+1}) \geq f_k(x_k)$).

Drawbacks:

1. Convergence proof is based on the hypothesis of infinite storage capacity (size of LP increases at each iteration).
2. numerical instability.

Example

Take $\min\{\frac{1}{2}x^2 : x \in \mathbb{R}\}$, with $x^* = 0$. Let $k = 2$, $x_1 = -1$ and $x_2 = 0.01$, with x_2 *rather close* to the minimizer. Then it is $x_3 = -0.495$, with *much bigger distance* from the minimizer.

Outline

Bundle methods as the evolution of CP

Bundle B_k is the cumulated information available at iteration k about points scattered throughout the function domain: a set of point/function/subgradient triplets:

$$B_k \triangleq \left\{ \left(x_i, f(x_i), g_i \right) : \, g_i \in \partial f(x_i), \, i \in \{1, \ldots, k\} \right\}.$$

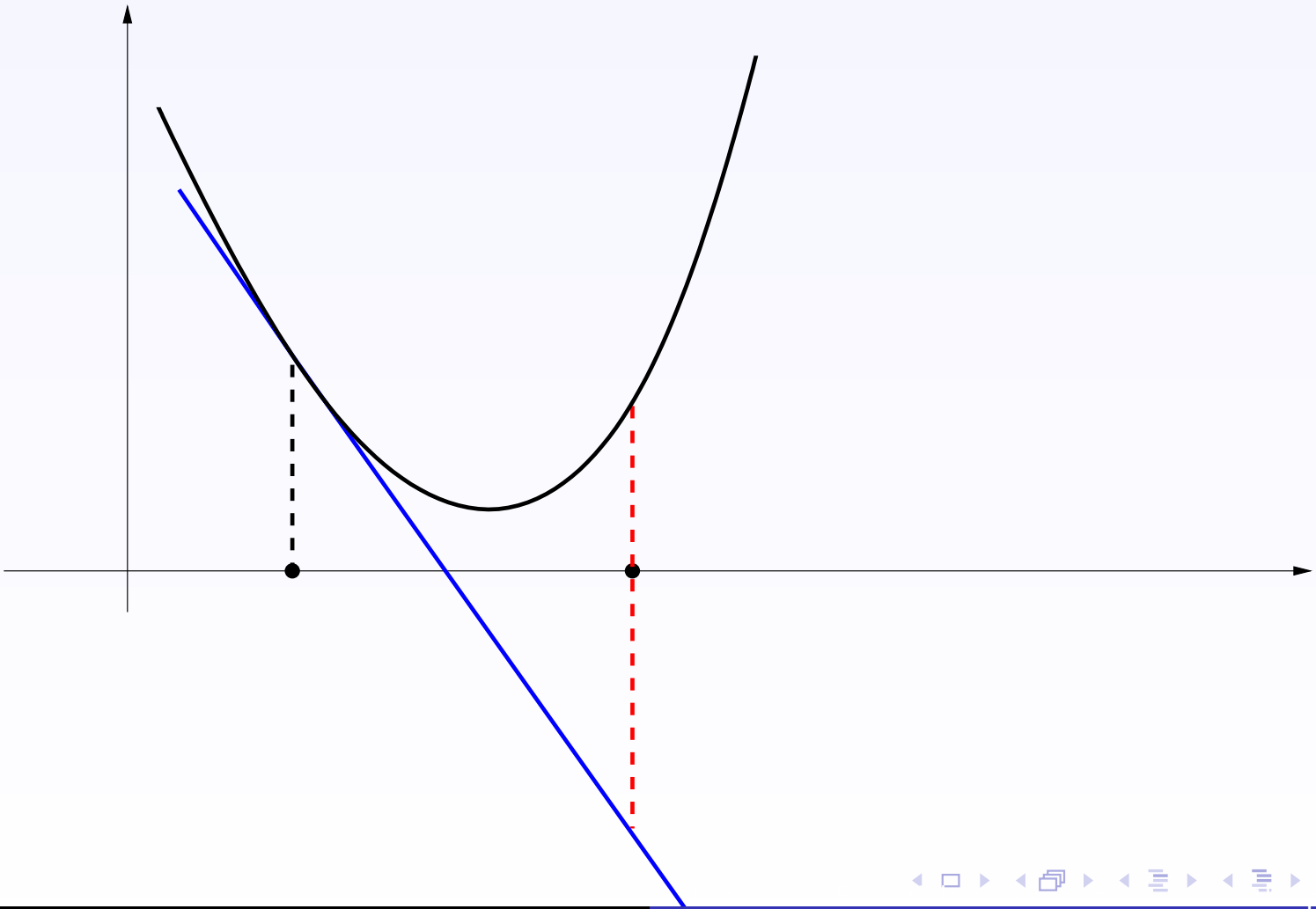
One among points x_i is the *stability center*, say $\overline{x}_k$, often referred to as the *incumbent*, with the *incumbent value* $f(\overline{x}_k)$.
Introduce the change of variables $x = \overline{x}_k + d$ and rewrite the cutting plane function $f_k(x)$ as the *difference function*:

$$f_k(\overline{x}_k + d) - f(\overline{x}_k) = \max \left\{ g_i^\top d - \alpha_i : i \in \{1, \ldots, k\} \right\}$$

where α_i, $i \in \{1, \ldots, k\}$, is the *linearization error*, associated to the affine expansion $\ell_i(x)$ at $\overline{x}_k$:

$$\alpha_i \triangleq f(\overline{x}_k) - \left(f(x_i) + g_i^\top (\overline{x}_k - x_i) \right) \geq 0.$$

The linearization error



For $x \in \mathbb{R}^n$, $i \in \{1, \ldots, k\}$ and $g_i \in \partial f(x_i)$ the subgradient transport property holds:

$$\begin{aligned} f(x) &\geq f(x_i) + g_i^\top (x - x_i) \\ &= f(\overline{x}_k) - f(\overline{x}_k) + f(x_i) + g_i^\top (x - \overline{x}_k + \overline{x}_k - x_i) \\ &= f(\overline{x}_k) + g_i^\top (x - \overline{x}_k) - \alpha_i, \end{aligned}$$

i.e.,

$$g_i \in \partial f(x_i) \quad \Rightarrow \quad g_i \in \partial_{\alpha_i} f(\overline{x}_k).$$

Even points far from the stability center provide approximate information on its differential properties.
The bundle B_k, instead of *triplets*, can be considered as made up of *couples* as follows

$$B_k \triangleq \left\{ (g_i, \alpha_i) : g_i \in \partial_{\alpha_i} f(\overline{x}_k),\ i \in \{1, \ldots, k\} \right\}.$$

Under the transformation of variables and the choice of the stability center, the LP becomes

$$\min\left\{v : v \geq g_i^\top d - \alpha_i \ \forall i \in \{1,\dots,k\},\, d \in \mathbb{R}^n,\, v \in \mathbb{R}\right\}$$

with optimal solution (d_k, v_k):

$$d_k = x_{k+1} - \overline{x}_k, \quad \text{and} \quad v_k = w_k - f(\overline{x}_k) = f_k(x_{k+1}) - f(\overline{x}_k).$$

$v_k \leq 0$ is the *predicted reduction* at $x_{k+1} = \overline{x}_k + d_k$.

Bundle methods elaborate on CP as they ensure:

i) Well-posedness of the subproblem.

ii) Stabilization of the next iterate.

Algorithm: (Conceptual) Bundle Method (BM)

Step 1 A starting point $x_1 \in \mathbb{R}^n$ and $g_1 \in \partial f(x_1)$ are given. Set $\overline{x}_1 = x_1$, $\alpha_1 = 0$, $B_1 = \{(g_1, \alpha_1)\}$ and $k = 1$.

Step 2 Solve an appropriate variant of the CP subproblem

Step 3 **If** the solution certifies approximate optimality of point $\overline{x}_k$ **Then** set $x^* = \overline{x}_k$ and **EXIT**.

Step 4 **Else** compare the *expected* and the *actual* reduction at $x_{k+1} = \overline{x}_k + td_k$ for $t = 1$, or possibly for $t \in (0, 1]$

Step 5 **If** a sufficient decrease in function f is achieved at x_{k+1} **Then** update the stability center $\overline{x}_{k+1} = x_{k+1}$, calculate $g_{k+1} \in \partial f(x_{k+1})$ and set $\alpha_{k+1} = 0$. Update the linearization errors and the bundle $B_{k+1} = B_k \cup \{(g_{k+1}, \alpha_{k+1})\}$, set $k = k + 1$ and **Go To** Step 2

Step 6 **Else** Set $\overline{x}_{k+1} = \overline{x}_k$ (stability center unchanged), calculate $g_{k+1} \in \partial f(x_{k+1})$ and set set $\alpha_{k+1} = f(\overline{x}_{k+1}) - \big(f(x_{k+1}) + g_{k+1}^\top(\overline{x}_{k+1} - x_{k+1})\big)$. Update the bundle $B_{k+1} = B_k \cup \{(g_{k+1}, \alpha_{k+1})\}$, set $k = k + 1$ and **Go To** Step 2.

Bundle methods classification

Previous schema leaves open a number of algorithmic decisions that lead to quite different methods, mainly based on the choice of the subproblem at Step 2.
The approaches are substantially three:

- Proximal BM;
- Trust region BM;
- Level BM.

The common rationale is to induce some limitation on the distance between two successive iterates to gather both well-posedness and stability.
Some approach-specific parameters control the actual magnitude of the limitation (their tuning is the real crucial issue affecting the numerical performance of all BM variants).

Proximal BM (PBM)

The subproblem at Step 2 is

$$\min\left\{v+\frac{1}{2}\gamma_k\|d\|^2 : v \geq g_i^\top d - \alpha_i \ \forall i \in \{1,\ldots,k\},\ d \in \mathbb{R}^n,\ v \in \mathbb{R}\right\}$$

where $\gamma_k > 0$ is the adjustable *proximity* parameter. It is a quadratic programming problem with a unique minimizer.
Wolfe's dual is

$$\min\left\{\frac{1}{2\gamma_k}\|\sum_{i=1}^k \lambda_i g_i\|^2 + \sum_{i=1}^k \lambda_i\alpha_i : e^\top\lambda = 1,\ \lambda \geq 0,\ \lambda \in \mathbb{R}^k\right\}$$

Letting (d_k, v_k) and $\lambda^{(k)}$ be, respectively, the primal and dual optimal solutions, it is:

$$d_k = -\frac{1}{\gamma_k}\sum_{i=1}^k \lambda_i^{(k)} g_i \quad v_k = -\frac{1}{\gamma_k}\|\sum_{i=1}^k \lambda_i^{(k)} g_i\|^2 - \sum_{i=1}^k \lambda_i^{(k)}\alpha_i,$$

Approximate optimality test

Primal-dual relations provide possible certification of the (approximate) optimality of the stability center at Step 3 of BM.
Let $g(\lambda) = \sum_{i=1}^{k} \lambda_i^{(k)} g_i$ and suppose $v_k \geq -\epsilon$ for some $\epsilon > 0$ Thus it is

$$\|g(\lambda)\| \leq \sqrt{\gamma_k \epsilon} \triangleq \delta \quad \text{and} \quad \sum_{i=1}^{k} \lambda_i^{(k)} \alpha_i \leq \epsilon.$$

Thus, taking into account $g(\lambda) \in \partial_\epsilon f(\overline{x}_k)$, we obtain

$$f(x) \geq f(\overline{x}_k) + g(\lambda)^\top (x - \overline{x}_k) - \epsilon \geq f(\overline{x}_k) - \delta \|x - \overline{x}_k\| - \epsilon, \quad \forall x \in \mathbb{R}^n,$$

which can be interpreted as an approximate optimality condition at point $\overline{x}_k$, provided that δ is not *too big*. Thus we conclude that $\{\gamma_k\}$:

- must be bounded from above,
- must be not too small (otherwise PBM would behave similarly to CP).

Checking for objective function decrease

At Step 4 it is $v_k < -\epsilon$ and the agreement between the predicted and the actual reduction is checked via the inequality

$$f(\overline{x}_k + d_k) - f(\overline{x}_k) \leq m v_k,$$

where $m \in (0, 1)$ is the *sufficient decrease* parameter.

If the condition is fulfilled (*Serious Step*), the new stability center is $\overline{x}_{k+1} = \overline{x}_k + d_k$ and a new iteration starts after updating the bundle.

If, instead, a sufficient decrease has not been attained, two implementations are possible: with or without *line search*. We discuss only the no-line search case

Null Step and bundle enrichment

Once the attempt to get a satisfactory decent has failed, the stability center remains unchanged (*Null Step*).

The couple (g_{k+1}, α_{k+1}) is joined to the bundle, with $g_{k+1} \in \partial f(x_{k+1})$, and $\alpha_{k+1} = f(\overline{x}_k) - f(x_{k+1}) + g_{k+1}^\top d_k$.

The updated model in terms of difference function becomes

$$\begin{aligned} f_{k+1}(\overline{x}_k + d) - f(\overline{x}_k) &= \max\left\{ g_i^\top d - \alpha_i : i \in \{1, \ldots, (k+1)\} \right\} \\ &= \max\left\{ f_k(\overline{x}_k + d) - f(\overline{x}_k),\ g_{k+1}^\top d - \alpha_{k+1} \right\}, \end{aligned}$$

providing a more accurate estimate of the objective function f, at least around point x_{k+1}.

Sketch of a termination proof

Typical convergence proofs are based on the following three facts:

a) The objective function reduction, every time the stability center is updated, is bounded away from zero (consequence of $v_k < -\epsilon$ and of the sufficient decrease).

b) Since it is assumed that the function has finite minimum, from a) it follows that only a finite number of stability center updates may take place.

c) Corresponding to a sequence of null steps, $\{v_k\}$ is monotonically increasing and it is $\{v_k\} \to 0$. Consequently, the stopping test $v_k \geq -\epsilon$ is satisfied after a finite number of null steps.

Trust region BM

The approach consists in solving at Step 2 of BM the classic Cutting Plane with the addition of a trust region constraint, based on $\Delta_k > 0$, the trust region parameter:

$$\min \left\{ v : v \geq g_i^\top d - \alpha_i,\ \forall i \in \{1, \ldots, k\},\ \|d\| \leq \Delta_k,\ d \in \mathbb{R}^n,\ v \in \mathbb{R} \right\}.$$

The issues:

- The choice of the norm in the trust region constraint: usually the ℓ_1 or the ℓ_∞ norm are taken, so that the subproblem is still LP.
- The setting of the trust region parameter: Δ_k not *too small* (next iterate too close to the stability center), nor *too large* (loss of the stabilizing effect). In general set $\Delta_k \in [\Delta_{min}, \Delta_{max}]$, and adopt some tuning heuristics.

Trust region BM may accommodate for a line search too.

Proximal Level BM (PLBM)

The subproblem looks for the closest point to $\overline{x}_k$ where the difference function takes a sufficiently negative value, dictated by the reduction parameter $\theta_k > 0$.

$$\min \left\{ \frac{1}{2}\|d\|^2 : \; g_i^\top d - \alpha_i \leq -\theta_k \; \forall i \in \{1, \ldots, k\}, \; d \in \mathbb{R}^n \right\}$$

Setting of θ_k is the key issue:

- Calculate the optimal value of f_k

$$f_k^* = \min \left\{ f_k(x) : \; x \in Q \subset \mathbb{R}^n \right\}$$

- Set $\theta_k = \mu\Gamma(k)$, where $\Gamma(k) = f_k(\overline{x}_k) - f_k^*$ and $\mu \in (0, 1)$.

Remark

Setting of θ_k appears more amenable than choosing γ_k and Δ_k in PBM and TBM. Here we deal with function values while in PBM and TBM we try tu capture secon order information!

Making BM implementable

Convergence proved under unlimited bundle size is a major drawback of BM: solved, thanks to subgradient aggregation.

Construct the aggregate couple (g_a, α_a):

$$g_a = \sum_{i=1}^{k} \lambda_i^{(k)} g_i \qquad \text{and} \qquad \alpha_a = \sum_{i=1}^{k} \lambda_i \alpha_i^{(k)}.$$

Define first the single-constraint aggregate program replacing all bundle couples (g_i, α_i) by (g_a, α_a),

$$\min \left\{ v + \frac{1}{2}\gamma_k \|d\|^2 : v \geq g_a^\top d - \alpha_a,\ d \in \mathbb{R}^n,\ v \in \mathbb{R} \right\},$$

and note that the optimal solution (d_a, v_a) is exactly (d_k, v_k):

$$d_a = -\frac{1}{\gamma_k} g_a = d_k, \quad v_a = g_a^\top d_a - \alpha_a = v_k.$$

Then add (g_{k+1}, α_{k+1}), $g_{k+1} \in \partial f(x_{k+1})$ (the *poorman* bundle).

Outline

Variable metric

Tuning of the proximity parameter γ_k in PBM has a strong impact on the performance.
The quadratic term $\frac{1}{2}\gamma_k\|d\|^2$ would be seen as the single–parameter positive definite approximation $\gamma_k I$ of the Hessian at point x_k
To introduce some variable metric, the simplest idea would be to solve:

$$\min\left\{v+\frac{1}{2}d^\top B_k d:\ v\geq x_i^\top d-\alpha_i\ \forall i\in\{1,\ldots,k\},\ d\in\mathbb{R}^n,\ v\in\mathbb{R}\right\},$$

where B_k is any positive definite Quasi–Newton matrix.

A more sophisticated approach works as follows:

- Calculate the Moreau-Yosida regularization of f related to point x_k.
- Apply any Quasi-Newton approach to such regularization.

Methods of Centers and variants

Still derived from cutting plane, by assuming a *set-oriented* instead of a *function-oriented* approach.
Given the cutting plane function:

- Define the *localization set*:

$$F_k(z_k) = \{(x, v) : x \in Q,\ f_k(x) \leq v \leq z_k\} \subset \mathbb{R}^{n+1},$$

  where z_k is any upper bound on f^* (e.g., $z_k = f(x_k)$). Note $(x^*, f^*) \in F_k(z_k)$

- Construct a nested sequence of sets $F_k(z_k)$ by introducing a *cut* at each iteration.

- To obtain substantial volume reduction in passing from $F_k(z_k)$ to $F_{k+1}(z_{k+1})$ (shrinking as fast as possible around (x^*, f^*)), look for a *central* cut, i.e., a cut generated on the basis of some notion of center of $F_k(z_k)$.

Calculating the center

- *Center of Gravity Method*: for a convex set C any hyperplane passing through the center of gravity generates a cut which reduces the volume by at least $(1 - e^{-1})$. Hard to calculate!
- *Analytic Center Cutting Plane Method*: the analytic center is a point that maximizes the *product of distances to all faces* of $F_k(z_k)$. It is the unique maximizer of the potential function

$$\psi_k(x, v) = \log(z_k - v) + \sum_{i=1}^{k} \log[v - f(x_i) - g_i^\top (x - x_i)].$$

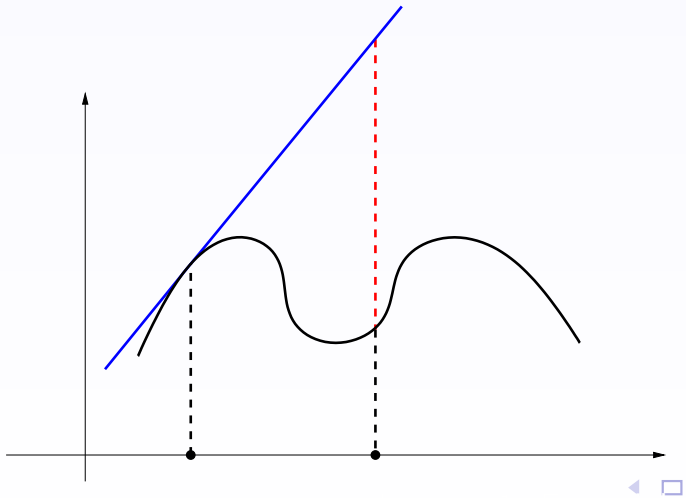
Outline

Exporting cutting plane and bundle

The approximation function f_k is still defined, provided $g_i \in \partial_C f(x)$.

$$f_k(\overline{x}_k + d) = f(\overline{x}_k) + \max\left\{x_i^\top d - \alpha_i : i \in \{1, \ldots, k\}\right\}$$

- it is no longer ensured that f_k is a lower approximation of f;
- f_k does not necessarily interpolates f at points x_i, $i \in \{1, \ldots, k\}$, even at $\overline{x}_k$, in case some α_i is negative,



Possible strategies

Replace the linearization error α_i with $\alpha_i = \max\left\{\alpha_i,\, \sigma\|\overline{x}_k - x_i\|^2\right\} \geq 0$, for some $\sigma > 0$ (interpolation at $\overline{x}_k$ ensured).
Alternatively proceed to the bundle splitting:

$$I_+ = \{i \in I :\, \alpha_i \geq 0\} \qquad \text{and} \qquad I_- = \{i \in I :\, \alpha_i < 0\}.$$

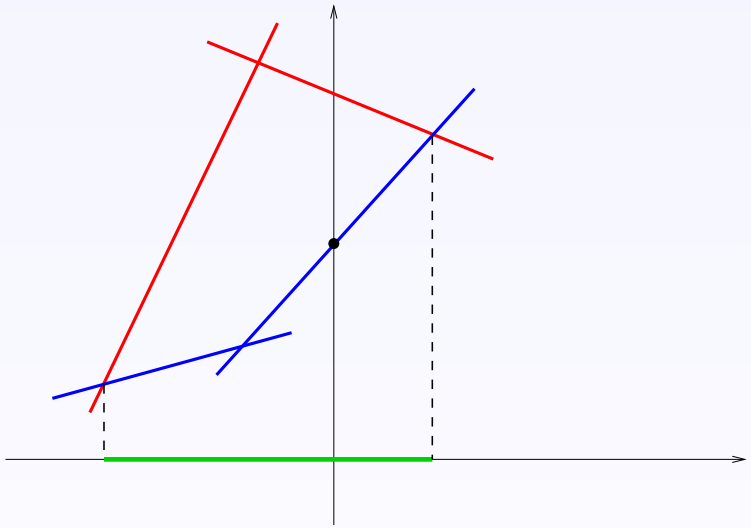
The bundles defined by I_+ and I_- refer to points that exhibit, respectively, a "convex behavior" and a "concave behavior" w.r.t. $\overline{x}_k$.
Define two piecewise affine functions (convex and concave, respectively):

$$\Delta^+(d) \triangleq \max\left\{x_i^\top d - \alpha_i :\, i \in I_+\right\}$$

and

$$\Delta^-(d) \triangleq \min\left\{x_i^\top d - \alpha_i :\, i \in I_-\right\}.$$

Around $d = 0$ (i.e., around the stability center $\overline{x}_k$) it is $\Delta^+(d) < \Delta^-(d)$.



The green area: $\{d : \Delta^+(d) \leq \Delta^-(d)\}$ Thus solve:

$$\min_d \left\{ \Delta^+(d) + \gamma_k \frac{1}{2} \|d\|^2 : \Delta^+(d) \leq \Delta^-(d) \right\},$$

that is the quadratic program:

$$\min_{d,v} \left\{ v + \gamma_k \frac{1}{2} \|d\|^2 : v \geq x_i^\top d - \alpha_i \ \forall i \in I_+, \ v \leq x_i^\top d - \alpha_i \ \forall i \in I_- \right\}.$$

Outline

References

M. Gaudioso, G. Giallombardo and G. Miglionico, *Essentials of numerical nonsmooth optimization*, 4OR, 18, pp. 1-47, 2020.