Sequential convex programming,: value function and convergence

Edouard Pauwels
joint work with Jérôme Bolte

Journées MODE
Toulouse

March 23 2016

Introduction

- Local search methods for finite dimensional nonconvex optimization.

Introduction

- Local search methods for finite dimensional nonconvex optimization.
- Sequential convex programming (SCP)

$$\text{``}x_{k+1} = \arg\min_{y} h(y, x_k)\text{''}$$

Introduction

- Local search methods for finite dimensional nonconvex optimization.
- Sequential convex programming (SCP)

$$\text{“}x_{k+1} = \arg\min_{y} h(y, x_k)\text{”}$$

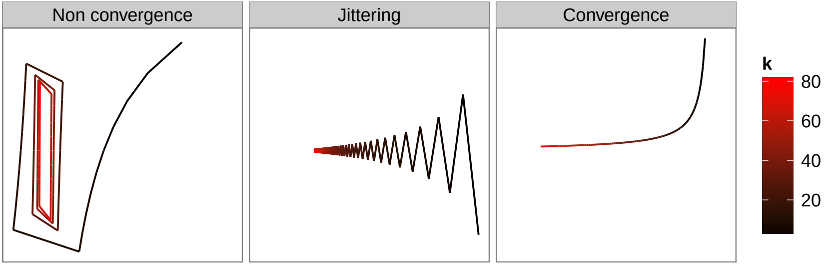
- Convergence of the iterates to a critical point?

Introduction

- Local search methods for finite dimensional nonconvex optimization.
- Sequential convex programming (SCP)

$$\text{``}x_{k+1} = \arg\min_y h(y, x_k)\text{''}$$

- Convergence of the iterates to a critical point?



The "prox-friendly" setting:

Non-smooth part (constraints or more general).

The "prox-friendly" setting:

Non-smooth part (constraints or more general).

Favorable geometries $\rightarrow$ "Proximal splitting" : Bruck, Pasty, Combettes, Pesquet, Nesterov, Beck, Teboulle, Eckstein, Tseng ...

The “prox-friendly” setting: convergence results

Non-smooth part (constraints or more general).

Favorable geometries $\rightarrow$ “Proximal splitting” : Bruck, Pasty, Combettes, Pesquet, Nesterov, Beck, Teboulle, Eckstein, Tseng . . .

Convergence with semi-algebraic data: Bolte, Attouch, Noll, Rondepierre, Chouzenoux, Pesquet, Repetti, Lewis, Sabach, Teboulle . . .

The "prox-friendly" setting: Limitations

Non-smooth part (constraints or more general).

Favorable geometries $\rightarrow$ "Proximal splitting" : Bruck, Pasty, Combettes, Pesquet, Nesterov, Beck, Teboulle, Eckstein, Tseng ...

Convergence with semi-algebraic data: Bolte, Attouch, Noll, Rondepierre, Chouzenoux, Pesquet, Repetti, Lewis, Sabach, Teboulle ...

Not all problems have tractable proximal operators

The "prox-friendly" setting: Limitations

Non-smooth part (constraints or more general).

Favorable geometries → "Proximal splitting" : Bruck, Pasty, Combettes, Pesquet, Nesterov, Beck, Teboulle, Eckstein, Tseng ...

Convergence with semi-algebraic data: Bolte, Attouch, Noll, Rondepierre, Chouzenoux, Pesquet, Repetti, Lewis, Sabach, Teboulle ...

Not all problems have tractable proximal operators

Complex geometries → nonsmoothness / constraints must be approximated:

- LP, QP, SDP: convex programming oracles.
- Large field (SQP, SQCQP, Gauss-Newton, ...).
- "prox-friendly" analysis does not apply directly:
  - More approximations → sources of oscilation.
  - convergence barely understood.

The "prox-friendly" setting: Limitations

Non-smooth part (constraints or more general).

Favorable geometries $\rightarrow$ "Proximal splitting" : Bruck, Pasty, Combettes, Pesquet, Nesterov, Beck, Teboulle, Eckstein, Tseng . . .

Convergence with semi-algebraic data: Bolte, Attouch, Noll, Rondepierre, Chouzenoux, Pesquet, Repetti, Lewis, Sabach, Teboulle . . .

Not all problems have tractable proximal operators

Complex geometries $\rightarrow$ nonsmoothness / constraints must be approximated:

- LP, QP, SDP: convex programming oracles.
- Large field (SQP, SQCQP, Gauss-Newton, . . .).
- "prox-friendly" analysis does not apply directly:
  - More approximations $\rightarrow$ sources of oscilation.
  - convergence barely understood.

Convergence to critical points in "non-prox-friendly" settings?

Outline

1. Existing results: gradient methods with semi-algebraic data
2. Complex geometries: sequential convex programming
3. Implicit gradient steps: the value function

Favorable geometries: gradient methods

Gradient descent: $\min_x f(x)$, (f smooth)

$$s(x_k - x_{k+1}) = \nabla f(x_k)$$

Favorable geometries: gradient methods

Gradient descent: $\min_x f(x)$, (f smooth)

$$s(x_k - x_{k+1}) = \nabla f(x_k)$$

Proximal point: $\min_x g(x)$ (g non-smooth)

$$x_{k+1} \in \mathrm{prox}_{g/s}(x_k) \qquad (\arg\min_y g(y)/s + \frac{1}{2}\|y - x_k\|_2^2)$$

$$s(x_k - x_{k+1}) \in \partial g(x_{k+1})$$

Favorable geometries: gradient methods

Gradient descent: $\min_x f(x)$, (f smooth)

$$s(x_k - x_{k+1}) = \nabla f(x_k)$$

Proximal point: $\min_x g(x)$ (g non-smooth)

$$x_{k+1} \in \mathrm{prox}_{g/s}(x_k) \qquad (\arg\min_y g(y)/s + \frac{1}{2}\|y - x_k\|_2^2)$$

$$s(x_k - x_{k+1}) \in \partial g(x_{k+1})$$

Forward-Backward: $\min_x f(x) + g(x)$ (f smooth, g non-smooth)

$$x_{k+1} \in \mathrm{prox}_{g/s}(x_k - 1/s\nabla f(x_k))$$

$$s(x_k - x_{k+1}) \in \partial g(x_{k+1}) + \nabla f(x_k)$$

Favorable geometries: gradient methods

Gradient descent: $\min_x f(x)$, (f smooth)

$$s(x_k - x_{k+1}) = \nabla f(x_k)$$

Proximal point: $\min_x g(x)$ (g non-smooth)

$$x_{k+1} \in \mathrm{prox}_{g/s}(x_k) \qquad (\arg\min_y g(y)/s + \frac{1}{2}\|y - x_k\|_2^2)$$

$$s(x_k - x_{k+1}) \in \partial g(x_{k+1})$$

Forward-Backward: $\min_x f(x) + g(x)$ (f smooth, g non-smooth)

$$x_{k+1} \in \mathrm{prox}_{g/s}(x_k - 1/s\nabla f(x_k))$$

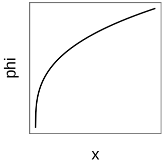
$$s(x_k - x_{k+1}) \in \partial g(x_{k+1}) + \nabla f(x_k)$$

Relation between $x_k - x_{k+1}$ and (sub)-gradient of the objective.

KL property (Łojasiewicz 63, Kurdyka 98)

Desingularizing functions on $(0, r_0)$

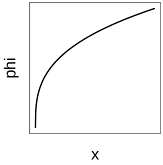
- $\varphi \in C([0, r_0), \mathbb{R}_+)$,
- $\varphi \in C^1(0, r_0)$, $\varphi' > 0$,
- φ concave and $\varphi(0) = 0$.



KL property (Łojasiewicz 63, Kurdyka 98)

Desingularizing functions on $(0, r_0)$

- $\varphi \in C([0, r_0), \mathbb{R}_+)$,
- $\varphi \in C^1(0, r_0)$, $\varphi' > 0$,
- φ concave and $\varphi(0) = 0$.



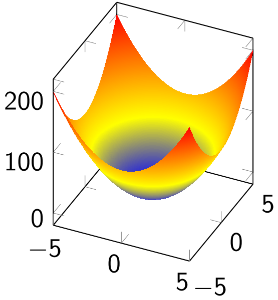
Definition

F_0 has the KL property at $\bar{x}$ ($F_0(\bar{x}) = 0$) if there exists $\epsilon > 0$ and a desingularizing function φ such that,

$$\mathrm{dist}(\partial(\varphi \circ F_0)(x), 0) \geq 1,$$
$$\forall x,\ ||x - \bar{x}|| \leq \epsilon,\ F_0(\bar{x}) < F_0(x) < F_0(\bar{x}) + \epsilon.$$

Illustration F_0 and $\varphi \circ F_0$

F_0 and $\varphi \circ F_0$



Parameterize with φ sharpens the function $\longrightarrow$

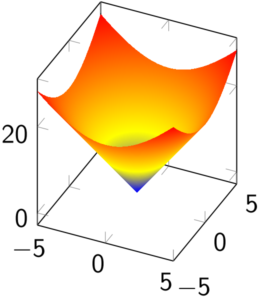
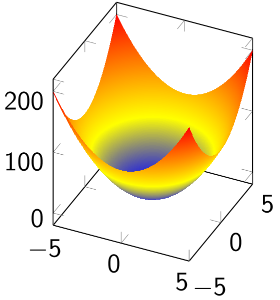
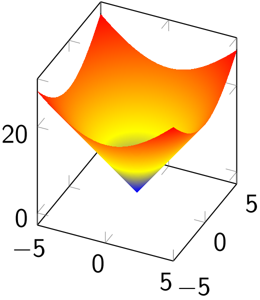


Illustration F_0 and $\varphi \circ F_0$

F_0 and $\varphi \circ F_0$



Parameterize with φ sharpens the function $\longrightarrow$



Theorem (2006, Bolte-Daniilidis-Lewis)

KL inequality holds for all lower-semicontinuous semi-algebraic functions (and many more).

Finite length property (general recipe)

(Attouch, Bolte, Svaiter, Sabach, Teboulle) ...
$A, B > 0$:

Sufficient decrease:	$f(x_{k+1}) + A\|\|x_{k+1} - x_k\|\|^2 \leq f(x_k)$
Step length:	$B\|\|\nabla f(x_k)\|\| \leq \|\|x_{k+1} - x_k\|\|$

Finite length property (general recipe)

(Attouch, Bolte, Svaiter, Sabach, Teboulle) ...
$A, B > 0$:

Sufficient decrease:	$f(x_{k+1}) + A\|\|x_{k+1} - x_k\|\|^2 \leq f(x_k)$
Step length:	$B\|\|\nabla f(x_k)\|\| \leq \|\|x_{k+1} - x_k\|\|$
Tameness:	f is semi-algebraic (KL inequality)

Finite length property (general recipe)

(Attouch, Bolte, Svaiter, Sabach, Teboulle) ...
$A, B > 0$:

Sufficient decrease:	$f(x_{k+1}) + A\|\|x_{k+1} - x_k\|\|^2 \leq f(x_k)$
Step length:	$B\|\|\nabla f(x_k)\|\| \leq \|\|x_{k+1} - x_k\|\|$
Tameness:	f is semi-algebraic (KL inequality)
Coercivity	$\{x;\ f(x) \leq f(x_0)\}$ is compact.

Finite length property (general recipe)

(Attouch, Bolte, Svaiter, Sabach, Teboulle) ...
$A, B > 0$:

Sufficient decrease:	$f(x_{k+1}) + A\|\|x_{k+1} - x_k\|\|^2 \leq f(x_k)$
Step length:	$B\|\|\nabla f(x_k)\|\| \leq \|\|x_{k+1} - x_k\|\|$
Tameness:	f is semi-algebraic (KL inequality)
Coercivity	$\{x;\ f(x) \leq f(x_0)\}$ is compact.

Finite length: $\sum \|\|x_{k+1} - x_k\|\|$ is bounded. $\{x_k\}$ converges to a critical point.

Finite length property (general recipe)

(Attouch, Bolte, Svaiter, Sabach, Teboulle) ...
$A, B > 0$:

Sufficient decrease:	$f(x_{k+1}) + A\|\|x_{k+1} - x_k\|\|^2 \leq f(x_k)$
Step length:	$B\|\|\nabla f(x_k)\|\| \leq \|\|x_{k+1} - x_k\|\|$
Tameness:	f is semi-algebraic (KL inequality)
Coercivity	$\{x;\ f(x) \leq f(x_0)\}$ is compact.

Finite length: $\sum ||x_{k+1} - x_k||$ is bounded. $\{x_k\}$ converges to a critical point.

Remark: There exist counterexamples for functions which are not semi-algebraic.

Outline

Examples in non linear programming

Approximate local models: approximation and localization

Exact penalization: $\min_x f(x) + \beta \max\{f_i(x)\}_{i=0\ldots m}$ (f, f_i smooth).

$$
\begin{aligned}
x_{k+1} = & \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k \rangle \\
& + \beta \max\{f_i(x_k) + \langle \nabla f_i(x_k), y - x_k \rangle\}_{i=1\ldots m} + s\|y - x_k\|_2^2
\end{aligned}
$$

Examples in non linear programming

Approximate local models: approximation and localization

Exact penalization: $\min_x f(x) + \beta \max\{f_i(x)\}_{i=0\dots m}$ $(f,\ f_i$ smooth$)$.

$$x_{k+1} = \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \beta \max\{f_i(x_k) + \langle \nabla f_i(x_k), y - x_k\rangle\}_{i=1\dots m} + s\|y - x_k\|_2^2$$

Moving ball: $\min_x f(x)$ s.t. $\max_{i=1\dots m} f_i(x) \leq 0.$ $(f,\ f_i$ smooth$)$.

$$x_{k+1} = \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + s\|y - x_k\|_2^2$$

$$\text{s.t. } \max_{i=1\dots m} f_i(x_k) + \langle \nabla f_i(x_k), y - x_k\rangle + s\|y - x_k\|_2^2 \leq 0$$

Examples in non linear programming

Approximate local models: approximation and localization

Exact penalization: $\min_x f(x) + \beta \max\{f_i(x)\}_{i=0\ldots m}$ (f, f_i smooth).

$$x_{k+1} = \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \beta \max\{f_i(x_k) + \langle \nabla f_i(x_k), y - x_k\rangle\}_{i=1\ldots m} + s\|y - x_k\|_2^2$$

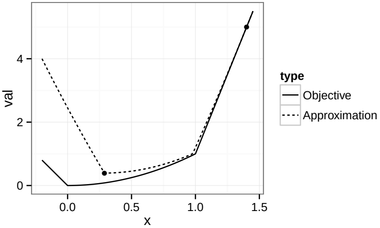
Moving ball: $\min_x f(x)$ s.t. $\max_{i=1\ldots m} f_i(x) \leq 0.$ (f, f_i smooth).

$$x_{k+1} = \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + s\|y - x_k\|_2^2$$
$$\text{s.t. } \max_{i=1\ldots m} f_i(x_k) + \langle \nabla f_i(x_k), y - x_k\rangle + s\|y - x_k\|_2^2 \leq 0$$

Gauss-Newton: $\min_x g(F(x))$ (F smooth, g convex)

$$x_{k+1} = \arg\min_y g(F(x_k) + \nabla F(x_k)(y - x_k)) + s\|y - x_k\|_2^2$$

A gradient method?



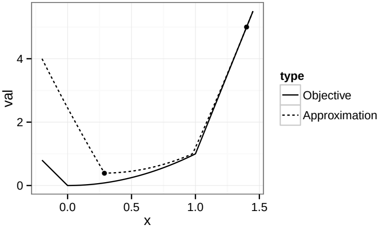
Objective:

$$H(x) = \max_{i=1\ldots m} f_i(x)$$

Approximation:

$$h_s(y,x) = \max_{i=1\ldots m} f_i(x) + \langle \nabla f_i(x), y - x \rangle + s\|y - x\|_2^2$$

A gradient method?



Objective:

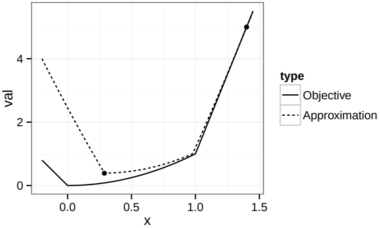
$$H(x) = \max_{i=1\ldots m} f_i(x)$$

Approximation:

$$h_s(y, x) = \max_{i=1\ldots m} f_i(x) + \langle \nabla f_i(x), y - x \rangle + s\|y - x\|_2^2$$

$x_{k+1} = \arg\min_y h_s(y, x_k)$: a gradient method?

A gradient method?



Objective:

$$H(x) = \max_{i=1\dots m} f_i(x)$$

Approximation:

$$h_s(y,x) = \max_{i=1\dots m} f_i(x) + \langle \nabla f_i(x), y - x\rangle + s\|y - x\|_2^2$$

$$x_{k+1} = \arg\min_y h_s(y, x_k)\text{: a gradient method?}$$

Main difficulty: track activity

- $I(x) := \arg\max_i f_i(x)$ (subgradients of H).
- $I(x_{k+1})$ and $I(x_k)$ are very hard to connect.
- No relation between $x_{k+1} - x_k$ and elements in $\partial H(x_k)$ or $\partial H(x_{k+1})$.
- Same issues for all the previous methods.

Outline

1. Existing results: gradient methods with semi-algebraic data

2. Complex geometries: sequential convex programming

3. Implicit gradient steps: the value function

Toward a link between SCP and gradient methods

Gradient descent is an SCP method:

$$
\begin{aligned}
& s(x_k - x_{k+1}) = \nabla f(x_k) \\
\Leftrightarrow \quad & x_{k+1} = \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{s}{2}\|y - x_k\|_2^2
\end{aligned}
$$

Toward a link between SCP and gradient methods

Gradient descent is an SCP method:

$$
\begin{aligned}
& s(x_k - x_{k+1}) = \nabla f(x_k) \\
\Leftrightarrow \quad & x_{k+1} = \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{s}{2}\|y - x_k\|_2^2
\end{aligned}
$$

An identity from Moreau:

g convex, lower semicontinuous,

$G: x \to \min_y g(y) + \frac{1}{2}\|y - x\|_2^2$ (value function of prox_g).

$$
\begin{aligned}
& x_{k+1} = \mathrm{prox}_g(x_k) \qquad (\arg\min_y g(y) + \frac{1}{2}\|y - x_k\|_2^2) \\
\Leftrightarrow \quad & x_k - x_{k+1} = \nabla G(x_k)
\end{aligned}
$$

Toward a link between SCP and gradient methods

Gradient descent is an SCP method:

$$
\begin{aligned}
& s(x_k - x_{k+1}) = \nabla f(x_k) \\
\Leftrightarrow \quad & x_{k+1} = \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{s}{2}\|y - x_k\|_2^2
\end{aligned}
$$

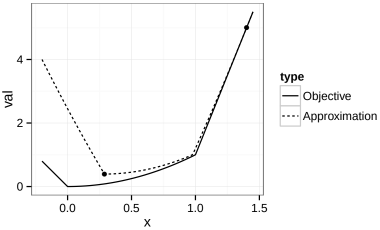
An identity from Moreau:

g convex, lower semicontinuous,
$G\colon x \to \min_y g(y) + \frac{1}{2}\|y - x\|_2^2$ (value function of prox_g).

$$
\begin{aligned}
& x_{k+1} = \mathrm{prox}_g(x_k) \qquad (\arg\min_y g(y) + \frac{1}{2}\|y - x_k\|_2^2) \\
\Leftrightarrow \quad & x_k - x_{k+1} = \nabla G(x_k)
\end{aligned}
$$

A prox step is an implicit gradient step on its value function.
Can we extend to more general SCP?

SCP: strongly convex tangent approximation, example



Objective: (each f_i is $\mathcal{C}^2$, semi-algebraic)

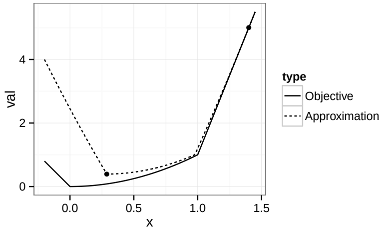
$$H(x) = \max_{i=1\ldots m} f_i(x)$$

Approximation:

$$h_s(y,x) = \max_{i=1\ldots m} f_i(x) + \langle \nabla f_i(x), y - x \rangle + s\|y - x\|_2^2$$

$$x_{k+1} = \arg\min_y h_s(y, x_k) \qquad (P_s(x))$$

SCP: strongly convex tangent approximation, example



Objective: (each f_i is $\mathcal{C}^2$, semi-algebraic)

$$H(x) = \max_{i=1\ldots m} f_i(x)$$

Approximation:

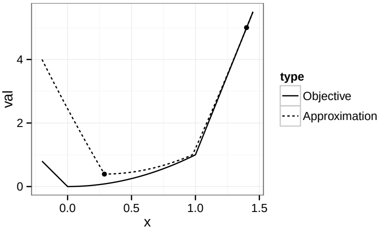
$$h_s(y,x) = \max_{i=1\ldots m} f_i(x) + \langle \nabla f_i(x), y - x\rangle + s\|y - x\|_2^2$$

$$x_{k+1} = \arg\min_y h_s(y, x_k) \qquad (P_s(x))$$

The value function: $V_s(x)$ = value of $P_s(x)$

- Critical points of V_s are exactly critical points of H.
- $\mathrm{dist}\,(0, \partial V_s(x_k)) \leq C\|x_{k+1} - x_k\|$ locally
- $V_s(x_k) + D\|x_k - x_{k+1}\|^2 \leq V_s(x_{k-1})$ locally (for suitable s).
- V_s is semi-algebraic $\rightarrow$ nonsmooth KL property.

SCP: strongly convex tangent approximation, example



Objective: (each f_i is $\mathcal{C}^2$, semi-algebraic)

$$H(x) = \max_{i=1\ldots m} f_i(x)$$

Approximation:

$$h_s(y,x) = \max_{i=1\ldots m} f_i(x) + \langle \nabla f_i(x), y - x\rangle + s\|y - x\|_2^2$$

$$x_{k+1} = \arg\min_y h_s(y, x_k) \qquad (P_s(x))$$

The value function: $V_s(x) =$ value of $P_s(x)$

- Critical points of V_s are exactly critical points of H.
- $\text{dist}\,(0, \partial V_s(x_k)) \leq C||x_{k+1} - x_k||$ locally
- $V_s(x_k) + D||x_k - x_{k+1}||^2 \leq V_s(x_{k-1})$ locally (for suitable s).
- V_s is semi-algebraic $\rightarrow$ nonsmooth KL property.

Implicit (sub)-gradient step on the value function $\rightarrow$ back to charted territory.

Non linear programming with semi-algebraic data

SQP and SQCQP from (Bolte-P. 2014): General convergence result

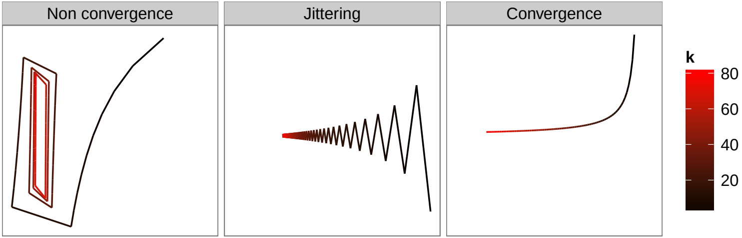
- Exact penalization SQP: Sℓ^1-QP (Fletcher 1985), ESQM (Auslender 2013).
- Inner approximating methods: Moving Ball (Auslender *et. al* (2010)).

Ongoing work (P. 2016):

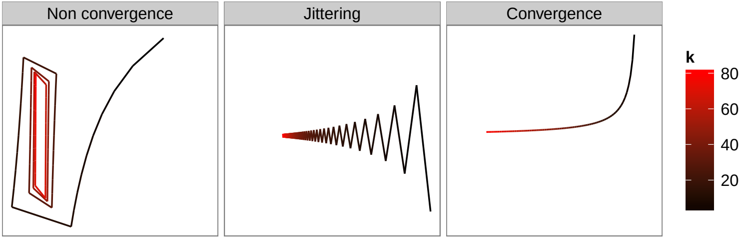
- Composite Gauss-Newton (Burke 1985):
  $\min_x \; g(F(x))$, $g \colon \mathbb{R}^m \to \mathbb{R}$ convex finite valued, $F \colon \mathbb{R}^n \to \mathbb{R}^m$ $\mathcal{C}^2$.

$$x_{k+1} = \arg\min_y g(F(x_k) + \nabla F(x_k)(y - x_k)) + s\|y - x_k\|_2^2$$

Conclusion



Conclusion



- **First general convergence result** for SCP methods (complex geometry).
- **Abstract SCP:** strongly convex tangent approximations of tame objective.
  - $\rightarrow$ implicit subgradient method on the value function.
- **More details:** J. Bolte and E. Pauwels. Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. MOR 2016. http://arxiv.org/abs/1409.8147