Structured nonsmooth optimization:

connections between Riemannian geometry, $\mathcal{U}$-Lagrangian theory, and SQP methods

Jérôme MALICK

INRIA Cornell

Continuous Optimization Seminar – Feb. 2006

Outline

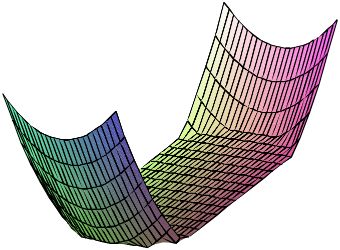
1. Smooth substructure
2. Newton of Riemannian geometry
3. Newton of the $\mathcal{U}$-Lagrangian theory
4. Newton of constrained optimization

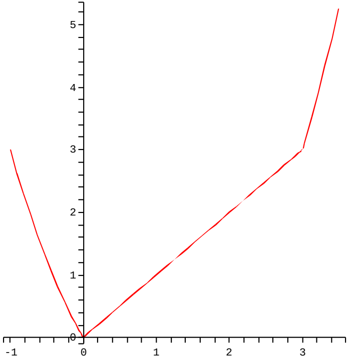
Outline

1. **Smooth substructure**
2. Newton of Riemannian geometry
3. Newton of the $\mathcal{U}$-Lagrangian theory
4. Newton of constrained optimization

Example 1 : max function

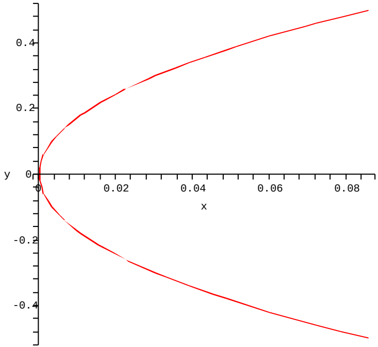
$$f(x, y) = \max\{\, x,\ (x-1)^2 + y^2 - 1 \,\}$$

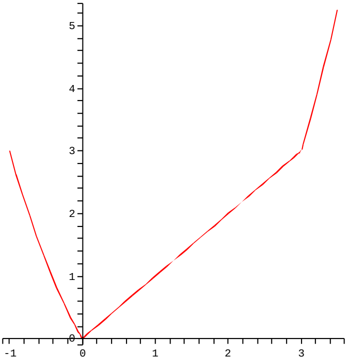




Example 1 : max function

$$f(x, y) = \max\{\, x,\ (x-1)^2 + y^2 - 1 \,\}.$$





Example 1 : max function

$$f(x) = \max\{ f_i(x), \quad i = 1, \ldots, n\} \qquad \text{with } f_i \text{ smooth}$$

$$I(x) = \operatorname{argmax}\{ f_i(x), \quad i = 1, \ldots, n\}$$

Under classical assumptions,

- $\mathcal{M} = \{x \in \mathbb{R}^n, \quad I(x) = I(\bar{x})\}$ is a smooth manifold
- f is smooth on $\mathcal{M}$
- $\widetilde{f}(x) = f_i(x)$ (for $i \in I(\bar{x})$) is a smooth representation of f

Example 2 : λ_1

In the space $\mathcal{S}_n$ of symmetric matrices

$$\mathcal{M}_r = \{A \in \mathcal{S}_n, \quad \lambda_1(A) = \cdots = \lambda_r(A) > \lambda_{r+1}(A)\}$$

The largest eigenvalue $\lambda_1 : \mathcal{S}_n \to \mathbb{R}$ is smooth on $\mathcal{M}_r$

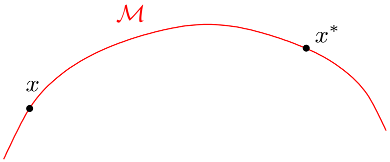
$\widetilde{\lambda}_1$ is a smooth representation of λ_1

$$\widetilde{\lambda}_1(X) = \frac{1}{r}(\lambda_1(X) + \cdots + \lambda_r(X))$$

General situation

$f : \mathbb{R}^n \to \mathbb{R}$ nonsmooth
but smooth on $\mathcal{M}$

$x^* \in \mathcal{M} \subset \mathbb{R}^n$



Using this sub-smoothness in minimization algorithms ?

Toward Newton Methods

$$\begin{cases} \min & f(x) \\ & x \in \mathbb{R}^n \end{cases} \qquad \begin{cases} \min & f(x) \\ & x \in \mathcal{M} \end{cases}$$

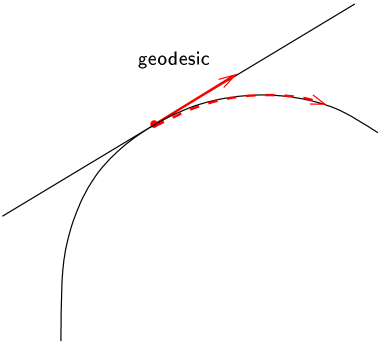
Toward Newton Methods

$$\begin{cases} \min & f(x) \\ & x \in \mathbb{R}^n \end{cases} \qquad \begin{cases} \min & f(x) \\ & x \in \mathcal{M} \end{cases}$$

Newton step $\quad x_+ = x - [\nabla^2 f_{|\mathcal{M}}(x)]^{-1} \nabla f_{|\mathcal{M}}(x)$

- how to differentiate $f_{|\mathcal{M}}$?
- how to move on $\mathcal{M}$?

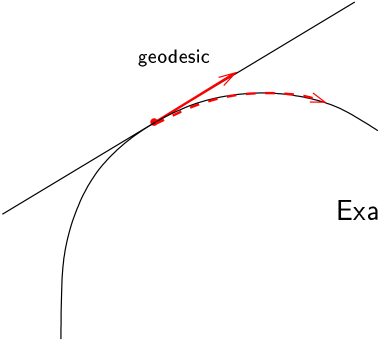
Geodesics



$$x \in \mathcal{M} \qquad u \in \mathrm{T}_{\mathcal{M}}(x)$$

$$\gamma(x, u, \cdot) \colon \mathbb{R} \to \mathcal{M}$$

Geodesics



$$x \in \mathcal{M} \qquad u \in \mathrm{T}_{\mathcal{M}}(x)$$

$$\gamma(x, u, \cdot) \colon \mathbb{R} \to \mathcal{M}$$

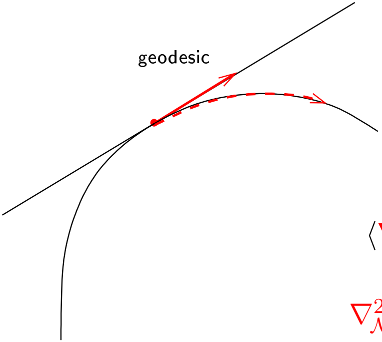
Example : $\mathcal{M} = \mathbb{R}^n$

$$\mathrm{T}_{\mathcal{M}}(x) = \mathbb{R}^n$$

$$\gamma(x, u, t) = x + tu$$

M.P. Do Carmo.
Riemannian Geometry.
Mathematics : Theory and Applications. Birkhaüser, 1992.

Geodesics



$$x \in \mathcal{M} \qquad u \in \mathrm{T}_{\mathcal{M}}(x)$$

$$\gamma(x, u, \cdot) \colon \mathbb{R} \to \mathcal{M}$$

$$\langle \nabla_{\mathcal{M}} f(x), u \rangle = \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(x, u, t)) \Big|_{t=0}$$

$$\nabla^2_{\mathcal{M}} f(x)(u, u) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\gamma(x, u, t)) \Big|_{t=0}$$

M.P. Do Carmo.
Riemannian Geometry.
Mathematics : Theory and Applications. Birkhaüser, 1992.

Gradient and subdifferential

Fréchet subdifferential

$$\hat{\partial} f(x) = \{v \in \mathbb{R}^n, \quad f(y) \geqslant f(x) + \langle v, y - x \rangle + o(\|y - x\|)\}$$

Gradient and subdifferential

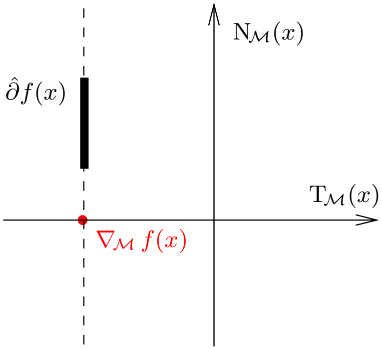
Fréchet subdifferential

$$\hat{\partial} f(x) = \{v \in \mathbb{R}^n, \quad f(y) \geqslant f(x) + \langle v, y - x\rangle + o(\|y - x\|)\}$$

If $\hat{\partial} f(x) \neq \varnothing$

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}\left(\hat{\partial} f(x)\right)$$

(but not $\partial f(x)$)



Steepest descent

Theorem (Riemannian Gradient and subdifferential)

If f is partly-smooth at $\bar{x} \in \mathcal{M}$ such that $0 \in \operatorname{ri} \partial f(\bar{x})$, then, for $x \in \mathcal{M}$ around $\bar{x}$,

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\partial f(x)}(0)$$

Steepest descent

Theorem (Riemannian Gradient and subdifferential)

If f is partly-smooth at $\bar{x} \in \mathcal{M}$ such that $0 \in \operatorname{ri} \partial f(\bar{x})$, then, for $x \in \mathcal{M}$ around $\bar{x}$,

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\partial f(x)}(0)$$

Steepest descent of smooth $f_{|\mathcal{M}}$
$=$ Steepest descent of nonsmooth f

Outline

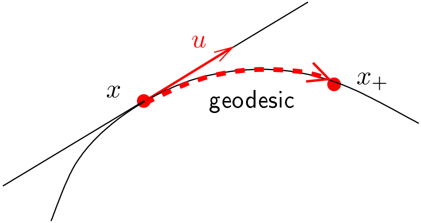
Riemannian Newton

1. tangential direction

$$u = -[\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x)$$

2. next iterate on $\mathcal{M}$

$$x_+ = \gamma(x, u, 1)$$



D. Gabay.
Minimizing a differentiable function over a differentiable manifold.
J. Optimization Theory Appl., 37(2) :177–219, June 1982.

S. T. Smith.
Optimization techniques on Riemannian manifolds.
Fields Inst. Comm., 3 :113–136, 1994.

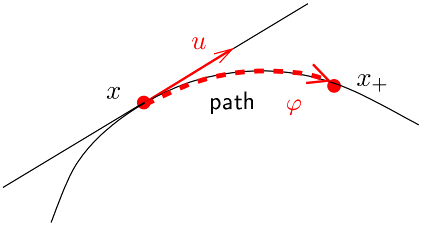
Extensions

1. tangential direction

$$u = -[\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x)$$

2. next iterate on $\mathcal{M}$

$$x_+ = \varphi(x, u)$$



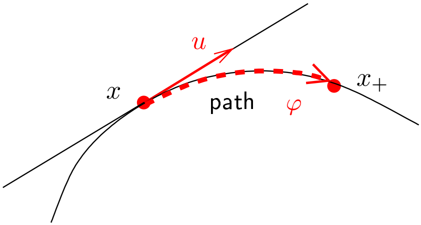
Extensions

1. tangential direction

$$u = -[\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x)$$

2. next iterate on $\mathcal{M}$

$$x_+ = \varphi(x, u)$$



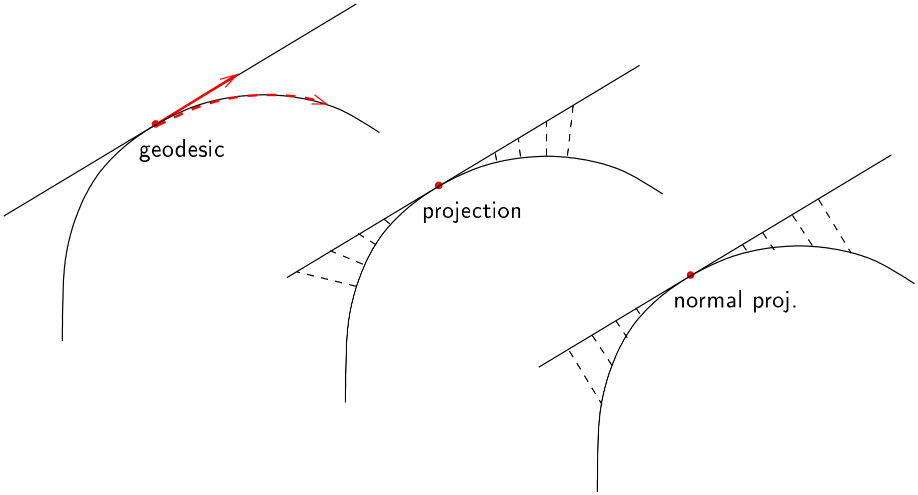
Theorem (Local convergence)

Under classical assumptions and

$$\varphi(x, tu) = \gamma(x, u, t) + o(t^2)$$

then the local quadratic convergence is maintained

Admissible paths



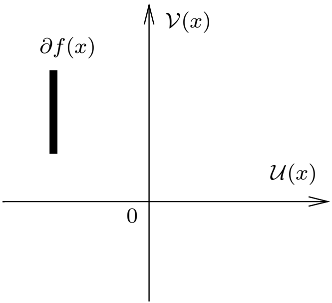
Outline

$\mathcal{VU}$-decomposition

f convex function
$x \in \mathbb{R}^n$

$\mathcal{V}(x) =$ subspace parallel to $\partial f(x)$

$$\mathcal{U}(x) = \mathcal{V}(x)^{\perp}$$
$$= \{d \in \mathbb{R}^n,\ f'(x,d) = -f'(x,-d)\}$$



C. Lemaréchal, F. Oustry and C. Sagastizábal.
The $\mathcal{U}$-Lagrangian of a convex function.
Trans. AMS, 352(2) :711–729, 1999.

$\mathcal{U}$-Lagrangian

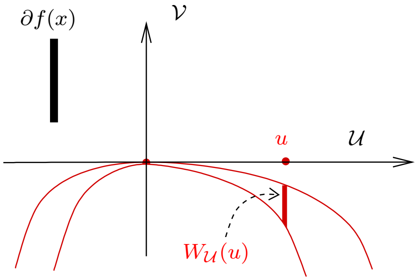
For $g \in \partial f(x)$

$$L_{\mathcal{U}}(u) = \min_{v \in \mathcal{V}} \{ f(x+u+v) - \langle g, v \rangle \}$$

$$W_{\mathcal{U}}(u) = \operatorname{argmin}_{v \in \mathcal{V}} \{ f(x+u+v) - \langle g, v \rangle \}$$

$\{x + u + W_{\mathcal{U}}(u),\ u \in \mathcal{U}\}$
so-called cornicopia

R. Mifflin and C. Sagastizábal
$\mathcal{VU}$-decomposition derivatives for convex max-functions.
In *Ill-Posed Variational Problems and Regularization Techniques*, Springer, 1999.



Algorithms

1. $\mathcal{U}$-Newton

   C. Lemaréchal, F. Oustry and C. Sagastizábal.
   The $\mathcal{U}$-Lagrangian of a convex function.
   Trans. AMS, 352(2) :711–729, 1999.

2. Projected $\mathcal{U}$-Newton for $f = \lambda_1$

   F. Oustry.
   The $\mathcal{U}$-Lagrangian of the maximum eigenvalue function.
   SIAM J. Optim., 9(2) :526–549, 1999.

3. Proximal $\mathcal{U}$-Newton

   R. Mifflin and C. Sagastizábal.
   A proximal $\mathcal{VU}$-algorithm for convex minimization.
   Math. Prog. B, 104 :609-633, 2005.

An additional assumption

- f admits fast-track

R. Mifflin and C. Sagastizábal.
Proximal points are on the fast track.
J. Convex Anal., 9(2) :563–579, 2002.

- f is partly smooth

A. S. Lewis.
Active sets, nonsmoothness and sensitivity.
SIAM J. Optim., 13 :702–725, 2003.

Equivalence

to be partly-smooth $\Longleftrightarrow$ to admit a fast-track

J. Malick and S.A. Miller.
Newton Methods for Nonsmooth Convex Minimization : Connections among $\mathcal{U}$-Lagrangian, Riemannian Newton and SQP Methods.
Math. Prog. B 104 :609-633, 2005.

W.L. Hare.
Recent functions and sets of smooth substructure : Relationships and examples.
To appear in Journal of Computational Optimization and Applications, 2006.

Partial smoothness

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ *is partly-smooth relative to the submanifold* $\mathcal{M} \subset \mathbb{R}^n$ *if*

1. *smoothness :* $f_{|\mathcal{M}}$ *is smooth*
2. *Clarke regularity :* f *is regular*
3. *sharpness : the affine subspace* $\operatorname{aff} \partial f(x)$ *is parallel to* $N_{\mathcal{M}}(x)$
4. *subdifferential continuity :* $(\partial f)_{|\mathcal{M}}$ *is continuous*

Partial smoothness

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ *is partly-smooth relative to the submanifold* $\mathcal{M} \subset \mathbb{R}^n$ *if*

1. *smoothness :* $f_{|\mathcal{M}}$ *is smooth*
2. *Clarke regularity :* f *is regular*
3. *sharpness : the affine subspace* $\operatorname{aff} \partial f(x)$ *is parallel to* $N_{\mathcal{M}}(x)$
4. *subdifferential continuity :* $(\partial f)_{|\mathcal{M}}$ *is continuous*

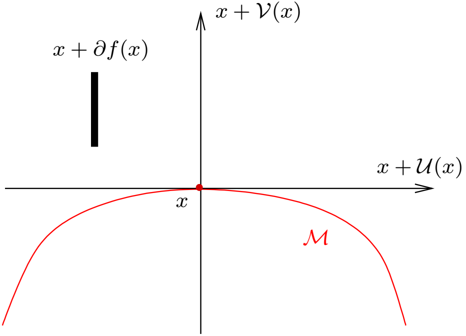
$\longrightarrow$ interpretation of objects of the $\mathcal{U}$-Lagrangian theory

Interpretation of $\mathcal{U}$ and $\mathcal{V}$

cornicopia
$=$ sub-manifold $\mathcal{M}$

$$\mathcal{V}(x) = \mathrm{N}_{\mathcal{M}}(x)$$

$$\mathcal{U}(x) = \mathrm{T}_{\mathcal{M}}(x)$$



Interpretation of derivatives of $L_{\mathcal{U}}$

$L_{\mathcal{U}}$ is smooth on $\mathcal{U}$ and

$$\nabla L_{\mathcal{U}}(0) = \nabla_{\mathcal{M}} f(x)$$

$$\nabla^2 L_{\mathcal{U}}(0) = \nabla^2_{\mathcal{M}} f(x)$$

(if $g = \nabla_{\mathcal{M}} f(x) \in \partial f(x)$, guaranteed in practice)

Point proximal

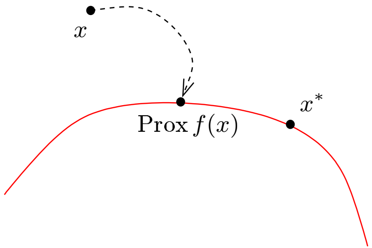
$$\operatorname{Prox} f(x) = \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ f(y) + \frac{1}{2} \|y - x\|^2 \right\}$$

Point proximal

$$\operatorname{Prox} f(x) = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2}\|y - x\|^2 \right\}$$

If x close to x^* such that $0 \in \operatorname{ri} \partial f(x^*)$, then

- $\operatorname{Prox} f(x) \in \mathcal{M}$
- $\operatorname{Prox} f(\cdot)$ smooth



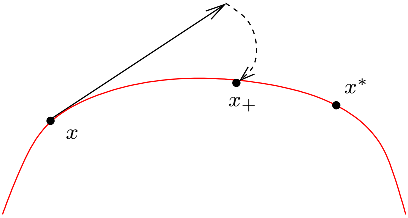
Interpretation of algorithms

$\mathcal{U}$-Newton methods follow the same two-step process :

1. Tangential prediction

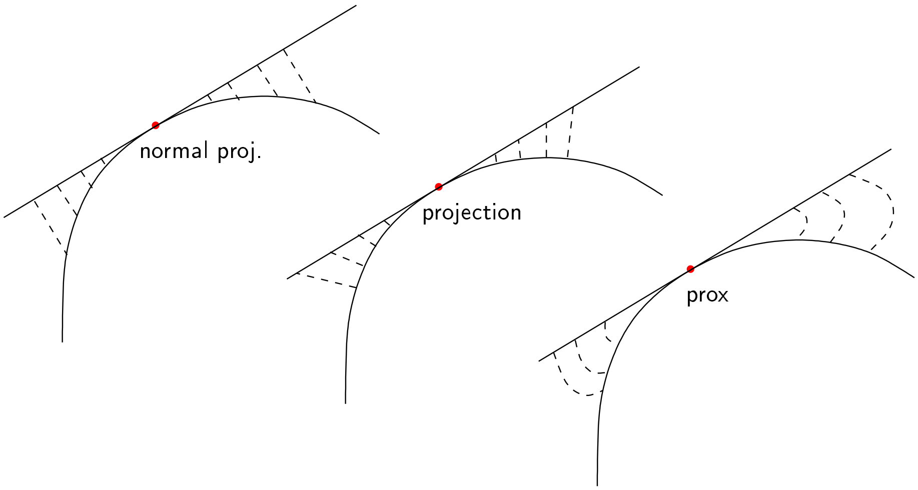
$$u = -[\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x)$$

2. Correction to $x + u$ to make it back to $\mathcal{M}$



concrete versions of Riemannian Newton

Different $\mathcal{U}$-Newton



Outline

Reformulation

$$\begin{cases} \min & f(x) \\ & x \in \mathcal{M} \end{cases} \qquad \begin{cases} \min & \widetilde{f}(x) \\ & \Phi(x) = 0 \end{cases}$$

Reformulation

$$\begin{cases} \min & f(x) \\ & x \in \mathcal{M} \end{cases} \qquad \begin{cases} \min & \widetilde{f}(x) \\ & \Phi(x) = 0 \end{cases}$$

Note : Riemannian gradient = projected gradient

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)} \left(\nabla \widetilde{f}(x) \right)$$

Reformulation

$$\begin{cases} \min & f(x) \\ & x \in \mathcal{M} \end{cases} \qquad \begin{cases} \min & \widetilde{f}(x) \\ & \Phi(x) = 0 \end{cases}$$

Note : Riemannian gradient = projected gradient

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)} \left(\nabla \widetilde{f}(x) \right)$$

Lagrangian $\qquad L(y, \lambda) = \widetilde{f}(x) + \lambda^\top \Phi(x)$

Reformulation

$$\begin{cases} \min & f(x) \\ & x \in \mathcal{M} \end{cases} \qquad \begin{cases} \min & \widetilde{f}(x) \\ & \Phi(x) = 0 \end{cases}$$

Note : Riemannian gradient = projected gradient

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)} \left(\nabla \widetilde{f}(x) \right)$$

Lagrangian $\qquad L(y, \lambda) = \widetilde{f}(x) + \lambda^\top \Phi(x)$

SQP
$$\begin{cases} \min & \nabla \widetilde{f}(x)^\top d + \frac{1}{2} d^\top \nabla^2 L(x, \lambda) d \\ & \Phi(x) + \mathrm{D}\Phi(x) d = 0 \end{cases}$$

Dual parameter

Choose $\lambda(x)$ such that

$$\min_{\lambda} \left\| \nabla \widetilde{f}(x) + \mathrm{D}\Phi(x)^{\top} \lambda \right\|^2$$

Dual parameter

Choose $\lambda(x)$ such that

$$\min_{\lambda} \left\| \nabla \widetilde{f}(x) + \mathrm{D}\Phi(x)^{\top} \lambda \right\|^2$$

Proposition (SQP-step = $\mathcal{U}$-step)

If $x \in \mathcal{M}$ then $x_+ - x \in \mathcal{U}$ = prediction step

- $\nabla_x L(x, \lambda(x)) = \nabla L_{\mathcal{U}}(0) = \nabla_{\mathcal{M}} f(x)$
- $\nabla_x^2 L(x, \lambda(x)) = \nabla^2 L_{\mathcal{U}}(0) = \nabla_{\mathcal{M}}^2 f(x)$

$\mathcal{V}$-step ?

$\mathcal{V}$-step ?

SQP needs a (nonsmooth) merit function

$$q(x) = \widetilde{f}(x) + \kappa \|\Phi(x)\|$$

$\Longrightarrow$ so-called Maratos effect

$\mathcal{V}$-step ?

SQP needs a (nonsmooth) merit function

$$q(x) = \widetilde{f}(x) + \kappa \|\Phi(x)\|$$

$\Longrightarrow$ so-called Maratos effect

The remedy appears like a correction step = $\mathcal{V}$-step

D. Q. Mayne.
On the use of exact penalty functions to determine step length in optimization algorithms.
In *Lecture Notes in Mathematics*, volume 773, pages 98–109. Springer Verlag, 1980.

Conclusion

- Nonsmooth interpretations of geometrical objects
- Geometrical interpretations of nonsmooth objects
- Connections between Riemannian Newton, $\mathcal{U}$-Newton and SQP

Conclusion

- Nonsmooth interpretations of geometrical objects
- Geometrical interpretations of nonsmooth objects
- Connections between Riemannian Newton, $\mathcal{U}$-Newton and SQP

J. Malick and S.A. Miller.
Newton Methods for Nonsmooth Convex Minimization
Math. Prog. B 104 :609-633, 2005.

A. Daniilidis, W.L. Hare and J. Malick.
Geometrical Interpretation of Methods for Structured Nonsmooth Optimization
to be accepted in *Optimization*, 2006.

Conclusion

- Nonsmooth interpretations of geometrical objects
- Geometrical interpretations of nonsmooth objects
- Connections between Riemannian Newton, $\mathcal{U}$-Newton and SQP

J. Malick and S.A. Miller.
Newton Methods for Nonsmooth Convex Minimization
Math. Prog. B 104 :609-633, 2005.

A. Daniilidis, W.L. Hare and J. Malick.
Geometrical Interpretation of Methods for Structured Nonsmooth Optimization
to be accepted in *Optimization*, 2006.

thanks !