Sparse and Smoothing Methods for Nonlinear Optimization Without Derivatives

Luis Nunes Vicente
University of Coimbra

joint work with A. Bandeira (Princeton) and K. Scheinberg (Lehigh) (sparse)
R. Garmanjani (smoothing)

September 28, 2011 — The 5th Sino-Japanese Optimization Meeting

http//www.mat.uc.pt/~lnv

Why Derivative-Free Optimization?

Some of the reasons to apply derivative-free optimization are the following:

Why Derivative-Free Optimization?

Some of the reasons to apply derivative-free optimization are the following:

- Nowadays computer hardware and mathematical algorithms allows increasingly large simulations.

Why Derivative-Free Optimization?

Some of the reasons to apply derivative-free optimization are the following:

- Nowadays computer hardware and mathematical algorithms allows increasingly large simulations.
- Functions are noisy (one cannot trust derivatives or approximate them by finite differences).

Why Derivative-Free Optimization?

Some of the reasons to apply derivative-free optimization are the following:

- Nowadays computer hardware and mathematical algorithms allows increasingly large simulations.
- Functions are noisy (one cannot trust derivatives or approximate them by finite differences).
- Binary codes (source code not available) and random simulations — making automatic differentiation impossible to apply.

Why Derivative-Free Optimization?

Some of the reasons to apply derivative-free optimization are the following:

- Nowadays computer hardware and mathematical algorithms allows increasingly large simulations.
- Functions are noisy (one cannot trust derivatives or approximate them by finite differences).
- Binary codes (source code not available) and random simulations — making automatic differentiation impossible to apply.
- Legacy codes (written in the past and not maintained by the original authors).

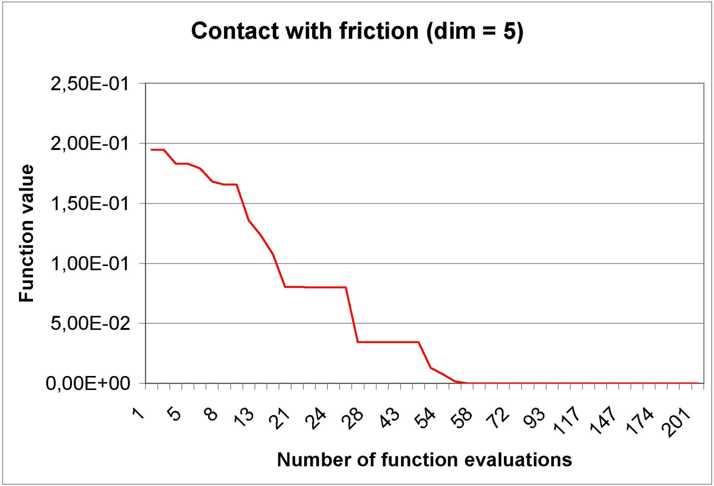
Why Derivative-Free Optimization?

Some of the reasons to apply derivative-free optimization are the following:

- Nowadays computer hardware and mathematical algorithms allows increasingly large simulations.
- Functions are noisy (one cannot trust derivatives or approximate them by finite differences).
- Binary codes (source code not available) and random simulations — making automatic differentiation impossible to apply.
- Legacy codes (written in the past and not maintained by the original authors).
- Lack of sophistication of the user (users need improvement but want to use something simple).

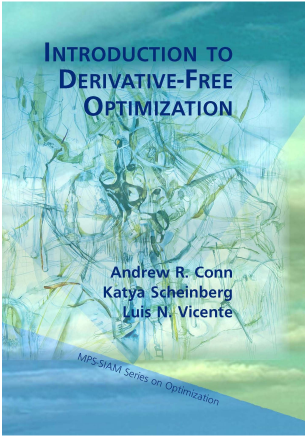
Limitations of Derivative-Free Optimization

In DFO convergence/stopping is typically slow (per function evaluation):



The book!

- A. R. Conn, K. Scheinberg, and L. N. Vicente, Introduction to Derivative-Free Optimization, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2009.



There are two main classes of rigorous methods in DFO

- Direct search methods, of directional type.

  Achieve descent by using positive spanning sets and moving in the directions of the best points.

There are two main classes of rigorous methods in DFO

- Direct search methods, of directional type.

  Achieve descent by using positive spanning sets and moving in the directions of the best points.

- Model-based methods, of local nature.

  Examples of models are polynomials or radial basis functions (RBFs).

- One typically minimizes a model m in a trust region $B_p(x;\Delta)$:

Trust-region subproblem

$$\min_{y \in B_p(x;\Delta)} m(y)$$

Model-based trust-region methods

- One typically minimizes a model m in a trust region $B_p(x;\Delta)$:

Trust-region subproblem

$$\min_{y \in B_p(x;\Delta)} \quad m(y)$$

In derivative-based optimization, one could use:

Model-based trust-region methods

- One typically minimizes a model m in a trust region $B_p(x;\Delta)$:

Trust-region subproblem

$$\min_{y \in B_p(x;\Delta)} m(y)$$

In derivative-based optimization, one could use:

1st order Taylor:

$$m(y) \ = \ f(x) + \nabla f(x)^\top (y - x)$$

Model-based trust-region methods

- One typically minimizes a model m in a trust region $B_p(x;\Delta)$:

Trust-region subproblem

$$\min_{y \in B_p(x;\Delta)} \; m(y)$$

In derivative-based optimization, one could use:

1st order Taylor:

$$m(y) \; = \; f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)^\top H (y-x)$$

Model-based trust-region methods

- One typically minimizes a model m in a trust region $B_p(x; \Delta)$:

Trust-region subproblem

$$\min_{y \in B_p(x;\Delta)} m(y)$$

In derivative-based optimization, one could use:

2nd order Taylor:

$$m(y) \ = \ f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x)$$

Fully linear models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully linear if

Fully linear models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully linear if

- It is $\mathcal{C}^1$ with Lipschitz continuous gradient.

Fully linear models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully linear if

- It is $\mathcal{C}^1$ with Lipschitz continuous gradient.
- The following error bounds hold:

$$\begin{aligned} \|\nabla f(y) - \nabla m(y)\| &\leq \kappa_{eg}\,\Delta \qquad \forall y \in B(x;\Delta) \\ |f(y) - m(y)| &\leq \kappa_{ef}\,\Delta^2 \qquad \forall y \in B(x;\Delta). \end{aligned}$$

Fully linear models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully linear if

- It is $\mathcal{C}^1$ with Lipschitz continuous gradient.
- The following error bounds hold:

$$\begin{aligned} \|\nabla f(y) - \nabla m(y)\| &\leq \kappa_{eg}\,\Delta \qquad \forall y \in B(x;\Delta) \\ |f(y) - m(y)| &\leq \kappa_{ef}\,\Delta^2 \qquad \forall y \in B(x;\Delta). \end{aligned}$$

For a class of fully linear models, the (unknown) constants $\kappa_{ef}, \kappa_{eg} > 0$ must be independent of x and Δ.

Fully linear models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully linear if

- It is $\mathcal{C}^1$ with Lipschitz continuous gradient.
- The following error bounds hold:

$$\|\nabla f(y) - \nabla m(y)\| \;\leq\; \kappa_{eg}\, \Delta \qquad \forall y \in B(x;\Delta)$$

$$|f(y) - m(y)| \;\leq\; \kappa_{ef}\, \Delta^2 \qquad \forall y \in B(x;\Delta).$$

For a class of fully linear models, the (unknown) constants $\kappa_{ef}, \kappa_{eg} > 0$ must be independent of x and Δ.

Fully linear models can be quadratic (or even nonlinear).

Fully quadratic models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully quadratic if

Fully quadratic models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully quadratic if

- It is $\mathcal{C}^2$ with Lipschitz continuous Hessian.

Fully quadratic models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully quadratic if

- It is $\mathcal{C}^2$ with Lipschitz continuous Hessian.
- The following error bounds hold:

$$
\begin{aligned}
\|\nabla^2 f(y) - \nabla^2 m(y)\| &\leq \kappa_{eh}\, \Delta & \forall y \in B(x;\Delta) \\
\|\nabla f(y) - \nabla m(y)\| &\leq \kappa_{eg}\, \Delta^2 & \forall y \in B(x;\Delta) \\
|f(y) - m(y)| &\leq \kappa_{ef}\, \Delta^3 & \forall y \in B(x;\Delta).
\end{aligned}
$$

Fully quadratic models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully quadratic if

- It is $\mathcal{C}^2$ with Lipschitz continuous Hessian.
- The following error bounds hold:

$$
\begin{aligned}
\|\nabla^2 f(y) - \nabla^2 m(y)\| &\leq \kappa_{eh}\,\Delta \qquad \forall y \in B(x;\Delta) \\
\|\nabla f(y) - \nabla m(y)\| &\leq \kappa_{eg}\,\Delta^2 \qquad \forall y \in B(x;\Delta) \\
|f(y) - m(y)| &\leq \kappa_{ef}\,\Delta^3 \qquad \forall y \in B(x;\Delta).
\end{aligned}
$$

For a class of fully quadratic models, the (unknown) constants $\kappa_{ef}, \kappa_{eg}, \kappa_{eh} > 0$ must be independent of x and Δ.

Fully quadratic models

Given a point x and a trust-region radius Δ, a model $m(y)$ around x is called fully quadratic if

- It is $\mathcal{C}^2$ with Lipschitz continuous Hessian.
- The following error bounds hold:

$$
\begin{aligned}
\|\nabla^2 f(y) - \nabla^2 m(y)\| &\leq \kappa_{eh}\,\Delta \qquad \forall y \in B(x;\Delta)\\
\|\nabla f(y) - \nabla m(y)\| &\leq \kappa_{eg}\,\Delta^2 \qquad \forall y \in B(x;\Delta)\\
|f(y) - m(y)| &\leq \kappa_{ef}\,\Delta^3 \qquad \forall y \in B(x;\Delta).
\end{aligned}
$$

For a class of fully quadratic models, the (unknown) constants $\kappa_{ef}, \kappa_{eg}, \kappa_{eh} > 0$ must be independent of x and Δ.

Fully quadratic models are only necessary for global convergence to 2nd order stationary points.

Polynomial interpolation models

Given a sample set $Y = \{y^0, y^1, \ldots, y^p\}$, a polynomial basis ϕ, and a polynomial model $m(y) = \alpha^\top \phi(y)$, the interpolating conditions form the linear system:

$$M(\phi, Y)\alpha \; = \; f(Y),$$

Polynomial interpolation models

Given a sample set $Y = \{y^0, y^1, \ldots, y^p\}$, a polynomial basis ϕ, and a polynomial model $m(y) = \alpha^\top \phi(y)$, the interpolating conditions form the linear system:

$$M(\phi, Y)\alpha \;=\; f(Y),$$

where

$$M(\phi, Y) = \left[\begin{array}{cccc} \phi_0(y^0) & \phi_1(y^0) & \cdots & \phi_p(y^0) \\ \phi_0(y^1) & \phi_1(y^1) & \cdots & \phi_p(y^1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(y^p) & \phi_1(y^p) & \cdots & \phi_p(y^p) \end{array}\right] \quad f(Y) = \left[\begin{array}{c} f(y^0) \\ f(y^1) \\ \vdots \\ f(y^p) \end{array}\right].$$

Natural/canonical basis

The natural/canonical basis appears in a Taylor expansion and is given by:

$$\bar{\phi} = \left\{ \frac{1}{2} y_1^2, ..., \frac{1}{2} y_n^2, y_1 y_2, ..., y_{n-1} y_n, y_1, ..., y_n, 1 \right\}.$$

Natural/canonical basis

The natural/canonical basis appears in a Taylor expansion and is given by:

$$\bar{\phi} = \left\{ \frac{1}{2} y_1^2, ..., \frac{1}{2} y_n^2, y_1 y_2, ..., y_{n-1} y_n, y_1, ..., y_n, 1 \right\}.$$

Under appropriate smoothness, the second order Taylor model, centered at 0, is:

$$\begin{aligned} & f(0)\,[1] + \tfrac{\partial f}{\partial x_1}(0)[y_1] + \tfrac{\partial f}{\partial x_2}(0)[y_2] \\ & + \tfrac{\partial^2 f}{\partial x_1^2}(0)[y_1^2/2] + \tfrac{\partial^2 f}{\partial x_1 x_2}(0)[y_1 y_2] + \tfrac{\partial^2 f}{\partial x_2^2}(0)[y_2^2/2]. \end{aligned}$$

Well poisedness (Λ–poisedness)

- Λ is a Λ–poisedness constant related to the geometry of Y.

Well poisedness (Λ–poisedness)

- Λ is a Λ–poisedness constant related to the geometry of Y.

The original definition of Λ–poisedness says that the maximum absolute value of the Lagrange polynomials in $B(x; \Delta)$ is bounded by Λ.

Well poisedness (Λ–poisedness)

- Λ is a Λ–poisedness constant related to the geometry of Y.

The original definition of Λ–poisedness says that the maximum absolute value of the Lagrange polynomials in $B(x;\Delta)$ is bounded by Λ.

An equivalent definition of Λ–poisedness is ($|Y| = |\alpha|$)

$$\|M(\bar{\phi}, Y_{scaled})^{-1}\| \leq \Lambda,$$

with Y_{scaled} obtained from Y such that $Y_{scaled} \subset B(0;1)$.

Well poisedness (Λ–poisedness)

- Λ is a Λ–poisedness constant related to the geometry of Y.

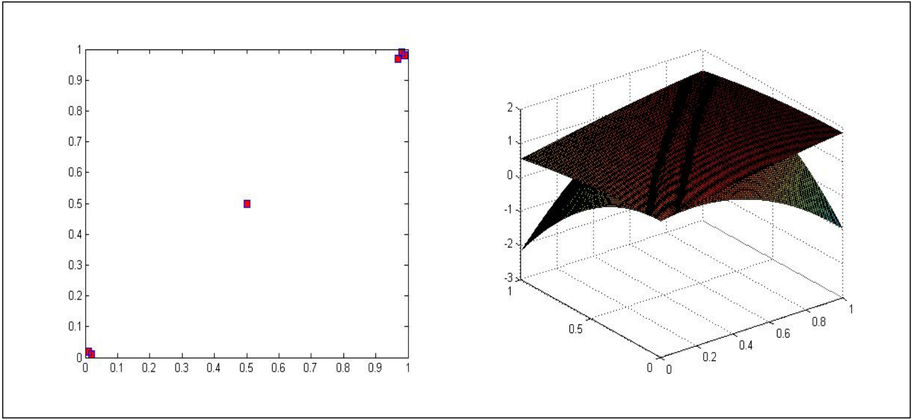
The original definition of Λ–poisedness says that the maximum absolute value of the Lagrange polynomials in $B(x;\Delta)$ is bounded by Λ.

An equivalent definition of Λ–poisedness is ($|Y| = |\alpha|$)

$$\|M(\bar{\phi}, Y_{scaled})^{-1}\| \leq \Lambda,$$

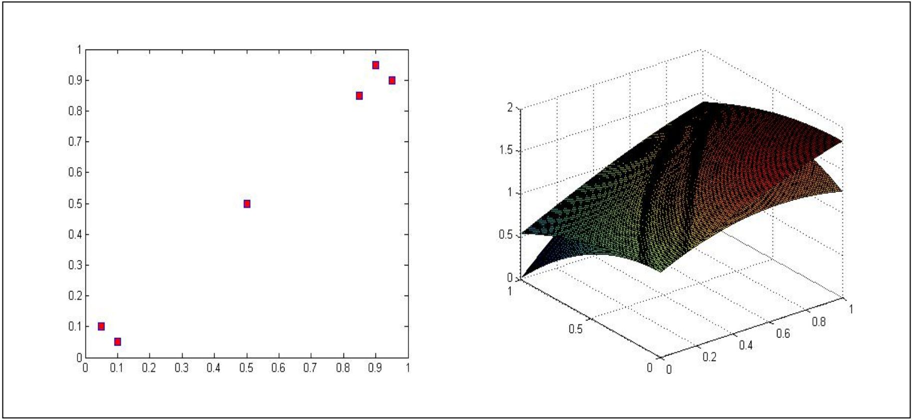
with Y_{scaled} obtained from Y such that $Y_{scaled} \subset B(0;1)$.

Non-squared cases are defined analogously (IDFO).


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
2
1
0
-1
-2
-3
1
0.5
0
0 0.2 0.4 0.6 0.8 1

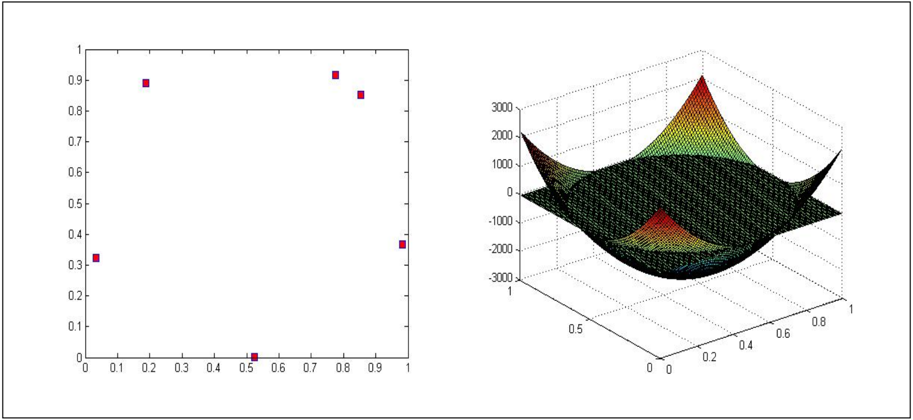
$$\Lambda = 5324.$$

A not so badly poised set



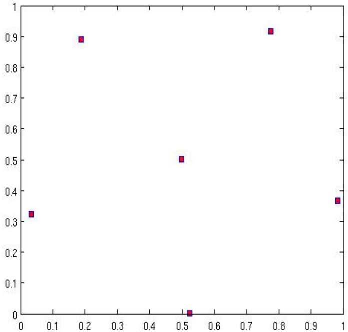
$$\Lambda = 295.$$

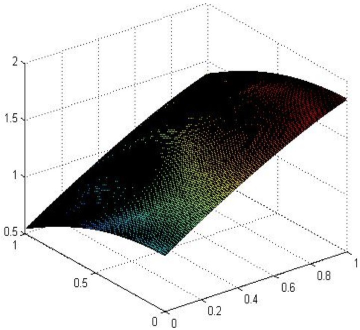
Another badly poised set



$$\Lambda = 492625.$$

An ideal set





$$\Lambda = 1.$$

The system $M(\phi, Y)\alpha = f(Y)$ can be

- Overdetermined when $|Y| > |\alpha|$.

Quadratic interpolation models

The system $M(\phi, Y)\alpha = f(Y)$ can be

- Determined when $|Y| = |\alpha|$.

  $\longrightarrow$ For $M(\phi, Y)$ to be squared one needs $N = (n+2)(n+1)/2$ evaluations of f (often too expensive).

Quadratic interpolation models

The system $M(\phi, Y)\alpha = f(Y)$ can be

- Determined when $|Y| = |\alpha|$.

  $\longrightarrow$ For $M(\phi, Y)$ to be squared one needs $N = (n+2)(n+1)/2$ evaluations of f (often too expensive).

  $\longrightarrow$ Leads to fully quadratic models when Y is well poised (the constants κ in the error bounds will depend on Λ).

Quadratic interpolation models

The system $M(\phi, Y)\alpha = f(Y)$ can be

- Determined when $|Y| = |\alpha|$.

  $\longrightarrow$ For $M(\phi, Y)$ to be squared one needs $N = (n+2)(n+1)/2$ evaluations of f (often too expensive).

  $\longrightarrow$ Leads to fully quadratic models when Y is well poised (the constants κ in the error bounds will depend on Λ).

- Underdetermined when $|Y| < |\alpha|$.

  $\longrightarrow$ Minimum Frobenius norm models (Powell, IDFO book).

Quadratic interpolation models

The system $M(\phi, Y)\alpha = f(Y)$ can be

- Determined when $|Y| = |\alpha|$.

  $\longrightarrow$ For $M(\phi, Y)$ to be squared one needs $N = (n+2)(n+1)/2$ evaluations of f (often too expensive).

  $\longrightarrow$ Leads to fully quadratic models when Y is well poised (the constants κ in the error bounds will depend on Λ).

- Underdetermined when $|Y| < |\alpha|$.

  $\longrightarrow$ Minimum Frobenius norm models (Powell, IDFO book).

  $\longrightarrow$ Other approaches?...

Underdetermined quadratic models

Let m be an underdetermined quadratic model (with Hessian H) built with less than $N = \mathcal{O}(n^2)$ points.

Underdetermined quadratic models

Let m be an underdetermined quadratic model (with Hessian H) built with less than $N = \mathcal{O}(n^2)$ points.

Theorem (IDFO book)

If Y is Λ_L–poised for linear interpolation or regression then

$$\|\nabla f(y) - \nabla m(y)\| \;\leq\; \Lambda_L \left[C_f + \|H\|\right] \Delta \qquad \forall y \in B(x;\Delta).$$

Underdetermined quadratic models

Let m be an underdetermined quadratic model (with Hessian H) built with less than $N = \mathcal{O}(n^2)$ points.

Theorem (IDFO book)

If Y is Λ_L–poised for linear interpolation or regression then

$$\|\nabla f(y) - \nabla m(y)\| \;\leq\; \Lambda_L \left[C_f + \|H\|\right] \Delta \qquad \forall y \in B(x;\Delta).$$

$\longrightarrow$ One should build models by minimizing the norm of H.

Minimum Frobenius norm models

Using $\bar{\phi}$ and separating the quadratic terms, write

$$m(y) \;=\; \alpha_Q^\top \bar{\phi}_Q(y) + \alpha_L^\top \bar{\phi}_L(y).$$

Minimum Frobenius norm models

Using $\bar{\phi}$ and separating the quadratic terms, write

$$m(y) \;=\; \alpha_Q^\top \bar{\phi}_Q(y) + \alpha_L^\top \bar{\phi}_L(y).$$

Then, build models by minimizing the entries of the Hessian ('Frobenius norm'):

$$\begin{array}{ll} \min & \frac{1}{2}\|\alpha_Q\|_2^2 \\ \text{s.t.} & M(\bar{\phi}, Y)\alpha \;=\; f(Y). \end{array}$$

Minimum Frobenius norm models

Using $\bar{\phi}$ and separating the quadratic terms, write

$$m(y) \;=\; \alpha_Q^\top \bar{\phi}_Q(y) + \alpha_L^\top \bar{\phi}_L(y).$$

Then, build models by minimizing the entries of the Hessian ('Frobenius norm'):

$$\begin{array}{rl} \min & \frac{1}{2}\|\alpha_Q\|_2^2 \\ \text{s.t.} & M(\bar{\phi}, Y)\alpha \;=\; f(Y). \end{array}$$

The solution of this convex QP problem requires a linear solve with:

$$\left[\begin{array}{cc} M_Q M_Q^\top & M_L \\ M_L^\top & 0 \end{array}\right] \quad \text{where} \quad M(\bar{\phi}, Y) \;=\; \left[\begin{array}{cc} M_Q & M_L \end{array}\right].$$

Minimum Frobenius norm models

Theorem (IDFO book)

If Y is Λ_F–poised in the minimum Frobenius norm sense then

$$\|H\| \;\leq\; C_f \Lambda_F,$$

where H is, again, the Hessian of the model.

Minimum Frobenius norm models

Theorem (IDFO book)

If Y is Λ_F–poised in the minimum Frobenius norm sense then

$$\|H\| \;\leq\; C_f \Lambda_F,$$

where H is, again, the Hessian of the model.

Putting the two theorems together yield:

$$\|\nabla f(y) - \nabla m(y)\| \;\leq\; \Lambda_L \left[C_f + C_f \Lambda_F\right] \Delta \qquad \forall y \in B(x;\Delta).$$

Minimum Frobenius norm models

Theorem (IDFO book)

If Y is Λ_F–poised in the minimum Frobenius norm sense then

$$\|H\| \;\leq\; C_f \Lambda_F,$$

where H is, again, the Hessian of the model.

Putting the two theorems together yield:

$$\|\nabla f(y) - \nabla m(y)\| \;\leq\; \Lambda_L \left[C_f + C_f \Lambda_F\right] \Delta \qquad \forall y \in B(x;\Delta).$$

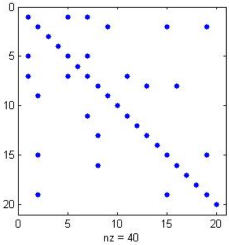
$\longrightarrow$ MFN models are fully linear.

Sparsity on the Hessian

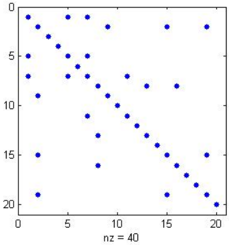
- In many problems, pairs of variables have no ‘correlation’, leading to zero second order partial derivatives in f:

Sparsity on the Hessian

- In many problems, pairs of variables have no 'correlation', leading to zero second order partial derivatives in f:



- In many problems, pairs of variables have no 'correlation', leading to zero second order partial derivatives in f:



- Thus, the Hessian $\nabla^2 m(x = 0)$ of the model (i.e., the vector α_Q in the basis $\bar{\phi}$) should be sparse.

Our main question

Question

Is it possible to build fully quadratic models by quadratic underdetermined interpolation (i.e., using less than $N = \mathcal{O}(n^2)$ points) in the SPARSE case?

Our main question

Question

Is it possible to build fully quadratic models by quadratic underdetermined interpolation (i.e., using less than $N = \mathcal{O}(n^2)$ points) in the SPARSE case?

An answer will be given by building the models using instead the ℓ_1-norm and relaxing the interpolating conditions for noisy recovery

$$\begin{aligned} \min \quad & \|\alpha_Q\|_1 \\ \text{s.t.} \quad & \|M(\bar{\phi}, Y)\alpha - f(Y)\|_2 \ \leq \ \eta. \end{aligned}$$

- Objective: Find sparse α subject to a highly underdetermined linear system $M\alpha = f$.

- Objective: Find sparse α subject to a highly underdetermined linear system $M\alpha = f$.

- $\left\{ \begin{array}{lll} \min & \|\alpha\|_0 & = |\operatorname{supp}(\alpha)| \\ \text{s.t.} & M\alpha = f & \end{array} \right.$ is NP-Hard.

Compressed sensing — sparse recovery

- Objective: Find sparse α subject to a highly underdetermined linear system $M\alpha = f$.

- $\left\{ \begin{array}{lll} \min & \|\alpha\|_0 & = |\operatorname{supp}(\alpha)| \\ \text{s.t.} & M\alpha = f & \end{array} \right.$ is NP-Hard.

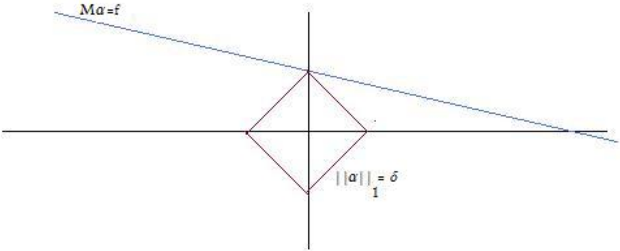
- $\left\{ \begin{array}{ll} \min & \|\alpha\|_1 \\ \text{s.t.} & M\alpha = f \end{array} \right.$ often recovers sparse solutions.

Compressed sensing — sparse recovery

- Objective: Find sparse α subject to a highly underdetermined linear system $M\alpha = f$.

- $\begin{cases} \min & \|\alpha\|_0 \quad = |\operatorname{supp}(\alpha)| \\ \text{s.t.} & M\alpha = f \end{cases}$ is NP-Hard.

- $\begin{cases} \min & \|\alpha\|_1 \\ \text{s.t.} & M\alpha = f \end{cases}$ often recovers sparse solutions.



Restricted isometry property

Definition (RIP)

The RIP Constant of order s of M $(p \times N)$ is the smallest δ_s such that

$$(1-\delta_s)\|\alpha\|_2^2 \leq \|M\alpha\|_2^2 \leq (1+\delta_s)\|\alpha\|_2^2$$

for all $s-$sparse α ($\|\alpha\|_0 \leq s$).

Restricted isometry property

Definition (RIP)

The RIP Constant of order s of M $(p \times N)$ is the smallest δ_s such that

$$(1-\delta_s)\|\alpha\|_2^2 \leq \|M\alpha\|_2^2 \leq (1+\delta_s)\|\alpha\|_2^2$$

for all $s-$sparse α ($\|\alpha\|_0 \leq s$).

Theorem (Candès, Tao, 2005, 2006)

If $\bar{\alpha}$ is $s-$sparse and M satisfies RIP of order $2s$ with $\delta_{2s} < \frac{1}{3}$, then $\bar{\alpha}$ can be recovered by ℓ_1-minimization:

$$\begin{array}{ll} \min & \|\alpha\|_1 \\ \text{s.t.} & M\alpha = M\bar{\alpha}. \end{array}$$

Restricted isometry property

Definition (RIP)

The RIP Constant of order s of M $(p \times N)$ is the smallest δ_s such that

$$(1-\delta_s)\|\alpha\|_2^2 \leq \|M\alpha\|_2^2 \leq (1+\delta_s)\|\alpha\|_2^2$$

for all $s-$sparse α ($\|\alpha\|_0 \leq s$).

Theorem (Candès, Tao, 2005, 2006)

If $\bar{\alpha}$ is $s-$sparse and M satisfies RIP of order $2s$ with $\delta_{2s} < \frac{1}{3}$, then $\bar{\alpha}$ can be recovered by ℓ_1-minimization:

$$\begin{array}{ll} \min & \|\alpha\|_1 \\ \text{s.t.} & M\alpha = M\bar{\alpha}. \end{array}$$

i.e., the optimal solution α^* of this problem is unique and given by $\alpha^* = \bar{\alpha}$.

Compressed sensing — noisy sparse recovery

Theorem (Candès 2009)

Let $M \in \mathbb{R}^{p \times N}$ *satisfy RIP of order* $2s$ *with*

$$\delta_{2s} \; < \; \sqrt{2} - 1.$$

For every $s-$*sparse vector* $\bar{\alpha} \in \mathbb{R}^N$, *let noisy measurements* $f = M\bar{\alpha} + \epsilon$ *be given satisfying* $\|\epsilon\|_2 \leq \eta$.

Let α^* *be a solution of*

$$\min_{\alpha \in \mathbb{R}^N} \|\alpha\|_1 \quad \text{s.t.} \quad \|M\alpha - f\|_2 \; \leq \eta.$$

Then

$$\|\alpha^* - \bar{\alpha}\|_2 \; \leq \; c_{total} \, \eta,$$

for a constant c_{total} *only depending on the RIP constant.*

Theorem (Jacques 2010, Bandeira, Scheinberg, and Vicente 2011)

Let $M = (M_1, M_2) \in \mathbb{R}^{p\times(N-r)} \times \mathbb{R}^{p\times r}$ *satisfy RIP of order* $2(s-r)$ *with*

$$\delta_{2(s-r)} < \sqrt{2} - 1.$$

For every $(s-r)-$*sparse vector* $\bar{\alpha}_1$, *with* $\bar{\alpha} = (\bar{\alpha}_1, \bar{\alpha}_2)$, *let noisy measurements* $f = M\bar{\alpha} + \epsilon$ *be given satisfying* $\|\epsilon\|_2 \leq \eta$.

Let $\alpha^* = (\alpha_1^*, \alpha_2^*)$ *be a solution of*

$$\min_{\alpha\in\mathbb{R}^N} \|\alpha_1\|_1 \quad \text{s.t.} \quad \|M\alpha - f\|_2 \leq \eta.$$

Then

$$\|\alpha^* - \bar{\alpha}\|_2 \leq c_{partial}\, \eta,$$

for a constant $c_{partial}$ *only depending on the RIP constant.*

Random matrices

- It is hard to find deterministic matrices that satisfy the RIP for large s.

Random matrices

- It is hard to find deterministic matrices that satisfy the RIP for large s.

- Using Random Matrix Theory it is possible to prove RIP for

$$p = \mathcal{O}(s \log N).$$

  - Matrices with Gaussian entries.
  - Matrices with Bernoulli entries.
  - Uniformly chosen subsets of discrete Fourier transform.
  - $\cdots$

Bounded orthonormal expansions (Rauhut)

Question

How to find a basis ϕ and a sample set Y such that $M(\phi, Y)$ satisfies the RIP?

Bounded orthonormal expansions (Rauhut)

Question

How to find a basis ϕ and a sample set Y such that $M(\phi, Y)$ satisfies the RIP?

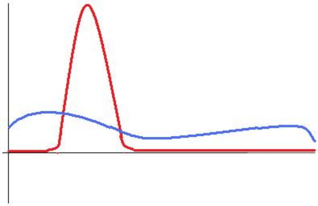
- Choose orthonormal bases (leads to uncorrelated matrix entries).

Bounded orthonormal expansions (Rauhut)

Question

How to find a basis ϕ and a sample set Y such that $M(\phi, Y)$ satisfies the RIP?

- Choose orthonormal bases (leads to uncorrelated matrix entries).
- Avoid localized functions ($\|\phi_i\|_{L^\infty}$ should be uniformly bounded) — to avoid zeros in matrix entries.

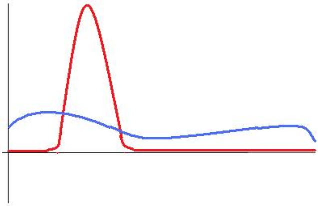


Bounded orthonormal expansions (Rauhut)

Question

How to find a basis ϕ and a sample set Y such that $M(\phi, Y)$ satisfies the RIP?

- Choose orthonormal bases (leads to uncorrelated matrix entries).
- Avoid localized functions ($\|\phi_i\|_{L^\infty}$ should be uniformly bounded) — to avoid zeros in matrix entries.



- Select Y randomly.

Sparse orthonormal bounded expansion recovery

Theorem (Rauhut, 2010)

If • ϕ *is orthonormal in a probability measure* μ *and* $\|\phi_i\|_{L^\infty} \leq K$.

Sparse orthonormal bounded expansion recovery

Theorem (Rauhut, 2010)

If
- ϕ *is orthonormal in a probability measure* μ *and* $\|\phi_i\|_{L^\infty} \leq K$.
- *each point of* Y *is drawn independently according to* μ.

Sparse orthonormal bounded expansion recovery

Theorem (Rauhut, 2010)

If
- ϕ *is orthonormal in a probability measure* μ *and* $\|\phi_i\|_{L^\infty} \leq K$.
- *each point of* Y *is drawn independently according to* μ.
- $\frac{p}{\log p} \geq c\, K^2 s (\log s)^2 \log N$.

Sparse orthonormal bounded expansion recovery

Theorem (Rauhut, 2010)

If
- *ϕ is orthonormal in a probability measure μ and $\|\phi_i\|_{L^\infty} \leq K$.*
- *each point of Y is drawn independently according to μ.*
- $\frac{p}{\log p} \geq c\, K^2 s (\log s)^2 \log N$.

Then, with high probability, for every $s-$sparse vector $\bar{\alpha}$:

Sparse orthonormal bounded expansion recovery

Theorem (Rauhut, 2010)

If

- *ϕ is orthonormal in a probability measure μ and $\|\phi_i\|_{L^\infty} \le K$.*
- *each point of Y is drawn independently according to μ.*
- *$\frac{p}{\log p} \ge c\, K^2 s (\log s)^2 \log N$.*

Then, with high probability, for every $s-$sparse vector $\bar{\alpha}$:

Given noisy samples $f = M(\phi, Y)\bar{\alpha} + \epsilon$ with $\|\epsilon\|_2 \le \eta$, let α^* be the solution of

Sparse orthonormal bounded expansion recovery

Theorem (Rauhut, 2010)

If
- ϕ *is orthonormal in a probability measure* μ *and* $\|\phi_i\|_{L^\infty} \leq K$.
- *each point of* Y *is drawn independently according to* μ.
- $\frac{p}{\log p} \geq c\, K^2 s (\log s)^2 \log N$.

Then, with high probability, for every $s-$*sparse vector* $\bar{\alpha}$*:*

Given noisy samples $f = M(\phi, Y)\bar{\alpha} + \epsilon$ *with* $\|\epsilon\|_2 \leq \eta$, *let* α^* *be the solution of*

$$\min \|\alpha\|_1 \quad \mathrm{s.t.} \quad \|M(\phi, Y)\alpha - f\|_2 \leq \eta.$$

Sparse orthonormal bounded expansion recovery

Theorem (Rauhut, 2010)

If
- ϕ *is orthonormal in a probability measure* μ *and* $\|\phi_i\|_{L^\infty} \le K$.
- *each point of* Y *is drawn independently according to* μ.
- $\frac{p}{\log p} \ge c\, K^2 s (\log s)^2 \log N$.

Then, with high probability, for every $s-$*sparse vector* $\bar{\alpha}$:

Given noisy samples $f = M(\phi, Y)\bar{\alpha} + \epsilon$ *with* $\|\epsilon\|_2 \le \eta$, *let* α^* *be the solution of*

$$\min \|\alpha\|_1 \quad \text{s.t.} \quad \|M(\phi, Y)\alpha - f\|_2 \le \eta.$$

Then,

$$\|\alpha^* - \bar{\alpha}\|_2 \le c_{total}\, \eta.$$

What basis do we need for sparse Hessian recovery?

Remember the second order Taylor model

$$f(0)\,[1] + \frac{\partial f}{\partial x_1}(0)[y_1] + \frac{\partial f}{\partial x_2}(0)[y_2]$$

$$+ \frac{\partial^2 f}{\partial x_1^2}(0)[y_1^2/2] + \frac{\partial^2 f}{\partial x_1 x_2}(0)[y_1 y_2] + \frac{\partial^2 f}{\partial x_2^2}(0)[y_2^2/2].$$

What basis do we need for sparse Hessian recovery?

Remember the second order Taylor model

$$f(0)\,[1] + \frac{\partial f}{\partial x_1}(0)[y_1] + \frac{\partial f}{\partial x_2}(0)[y_2]$$

$$+ \frac{\partial^2 f}{\partial x_1^2}(0)[y_1^2/2] + \frac{\partial^2 f}{\partial x_1 x_2}(0)[y_1 y_2] + \frac{\partial^2 f}{\partial x_2^2}(0)[y_2^2/2].$$

So, we want something like the natural/canonical basis:

$$\bar{\phi} = \left\{ \frac{1}{2}y_1^2, ..., \frac{1}{2}y_n^2, y_1 y_2, ..., y_{n-1} y_n, y_1, ..., y_n, 1 \right\}.$$

An orthonormal basis for quadratics (appropriate for sparse Hessian recovery)

Proposition (Bandeira, Scheinberg, and Vicente, 2011)

The following basis ψ for quadratics is orthonormal (w.r.t. the uniform measure on $B_\infty(0;\Delta)$) and satisfies $\|\psi_\iota\|_{L^\infty} \leq 3$.

An orthonormal basis for quadratics (appropriate for sparse Hessian recovery)

Proposition (Bandeira, Scheinberg, and Vicente, 2011)

The following basis ψ for quadratics is orthonormal (w.r.t. the uniform measure on $B_\infty(0;\Delta)$) and satisfies $\|\psi_\iota\|_{L^\infty} \leq 3$.

$$\left\{ \begin{array}{rcl} \psi_0(u) & = & 1 \\ \psi_{1,i}(u) & = & \frac{\sqrt{3}}{\Delta} u_i \\ \psi_{2,ij}(u) & = & \frac{3}{\Delta^2} u_i u_j \\ \psi_{2,i}(u) & = & \frac{3\sqrt{5}}{2} \frac{1}{\Delta^2} u_i^2 - \frac{\sqrt{5}}{2}. \end{array} \right.$$

An orthonormal basis for quadratics (appropriate for sparse Hessian recovery)

Proposition (Bandeira, Scheinberg, and Vicente, 2011)

The following basis ψ *for quadratics is orthonormal (w.r.t. the uniform measure on* $B_\infty(0;\Delta)$*) and satisfies* $\|\psi_\iota\|_{L^\infty} \leq 3$*.*

$$\begin{cases} \psi_0(u) & = & 1 \\ \psi_{1,i}(u) & = & \frac{\sqrt{3}}{\Delta} u_i \\ \psi_{2,ij}(u) & = & \frac{3}{\Delta^2} u_i u_j \\ \psi_{2,i}(u) & = & \frac{3\sqrt{5}}{2} \frac{1}{\Delta^2} u_i^2 - \frac{\sqrt{5}}{2}. \end{cases}$$

$\longrightarrow$ ψ is very similar to the canonical basis, and preserves the sparsity of the Hessian (at 0).

Let us look again at

$$\min \|\alpha_Q\|_1 \quad \text{s. t.} \quad \|M(\phi, Y)\alpha - f\|_2 \le \eta,$$

where

$$f = M(\psi, Y)\bar{\alpha} + \epsilon.$$

Hessian sparse recovery

Let us look again at

$$\min \|\alpha_Q\|_1 \quad \text{s. t.} \quad \|M(\phi, Y)\alpha - f\|_2 \ \leq \eta,$$

where

$$f \ = \ M(\psi, Y)\bar{\alpha} + \epsilon.$$

So, the 'noisy' data is $f = f(Y)$.

Hessian sparse recovery

Let us look again at

$$\min \|\alpha_Q\|_1 \quad \text{s.t.} \quad \|M(\phi, Y)\alpha - f\|_2 \le \eta,$$

where

$$f = M(\psi, Y)\bar{\alpha} + \epsilon.$$

So, the 'noisy' data is $f = f(Y)$.

What we are trying to recover is the 2nd order Taylor model $\bar{\alpha}^\top \psi(y)$.

Hessian sparse recovery

Let us look again at

$$\min \|\alpha_Q\|_1 \quad \text{s.t.} \quad \|M(\phi, Y)\alpha - f\|_2 \ \le \eta,$$

where

$$f \ = \ M(\psi, Y)\bar{\alpha} + \epsilon.$$

So, the ‘noisy’ data is $f = f(Y)$.

What we are trying to recover is the 2nd order Taylor model $\bar{\alpha}^\top \psi(y)$.

Thus, in $\|\epsilon\| \le \eta$, one has $\eta = \mathcal{O}(\Delta^3)$.

Hessian sparse recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2011)

If
- *the Hessian of f at 0 is $h-$sparse.*

Hessian sparse recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2011)

If
- *the Hessian of f at 0 is $h-$sparse.*
- *Y is a random sample set chosen w.r.t. the uniform measure on $B_\infty(0;\Delta)$.*

Hessian sparse recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2011)

If
- *the Hessian of f at 0 is $h-$sparse.*
- *Y is a random sample set chosen w.r.t. the uniform measure on $B_\infty(0;\Delta)$.*
- $\frac{p}{\log p} \geq 9c\,(h+n+1)\log^2(h+n+1)\log \mathcal{O}(n^2)$.

Hessian sparse recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2011)

If

- *the Hessian of f at 0 is $h-$sparse.*
- *Y is a random sample set chosen w.r.t. the uniform measure on $B_\infty(0;\Delta)$.*
- $\frac{p}{\log p} \geq 9c\,(h+n+1)\log^2(h+n+1)\log \mathcal{O}(n^2)$.

Then, with high probability, the quadratic

Hessian sparse recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2011)

If
- *the Hessian of f at 0 is $h-$sparse.*
- *Y is a random sample set chosen w.r.t. the uniform measure on $B_\infty(0;\Delta)$.*
- $\frac{p}{\log p} \geq 9c\,(h+n+1)\log^2(h+n+1)\log \mathcal{O}(n^2)$.

Then, with high probability, the quadratic

$$q^* = \sum \alpha_\iota^* \psi_\iota$$

obtained by solving the noisy and partial ℓ_1-minimization problem is a fully quadratic model for f (with error constants not depending on Δ).

An answer to our main question

- For instance, when the number of non-zeros of the Hessian is $h = \mathcal{O}(n)$, we are able to construct fully quadratic models with

$$\mathcal{O}(n \log^4 n) \quad \text{points.}$$

An answer to our main question

- For instance, when the number of non-zeros of the Hessian is $h = \mathcal{O}(n)$, we are able to construct fully quadratic models with

$$\mathcal{O}(n \log^4 n) \quad \text{points.}$$

- Also, we recover both the function and its sparsity structure.

However, the Theorem only provides motivation because, in a practical Optimization approach we:

Remarks

However, the Theorem only provides motivation because, in a practical Optimization approach we:

- Solve

$$\begin{array}{ll} \min & \|\alpha_Q\|_1 \\ \text{s.t.} & M(\bar{\phi}_Q, Y)\alpha_Q + M(\bar{\phi}_L, Y)\alpha_L \;=\; f(Y). \end{array}$$

Remarks

However, the Theorem only provides motivation because, in a practical Optimization approach we:

- Solve
$$\begin{array}{rl} \min & \|\alpha_Q\|_1 \\ \text{s. t.} & M(\bar{\phi}_Q, Y)\alpha_Q + M(\bar{\phi}_L, Y)\alpha_L \; = \; f(Y). \end{array}$$

- Deal with small n (from the DFO setting) and the bound we obtain is asymptotical.

Remarks

However, the Theorem only provides motivation because, in a practical Optimization approach we:

- Solve

$$\begin{array}{ll} \min & \|\alpha_Q\|_1 \\ \text{s. t.} & M(\bar{\phi}_Q, Y)\alpha_Q + M(\bar{\phi}_L, Y)\alpha_L \;=\; f(Y). \end{array}$$

- Deal with small n (from the DFO setting) and the bound we obtain is asymptotical.
- Use deterministic sampling.

A practical interpolation-based trust-region solver

We have tested the effect of minimum ℓ_1-norm Hessian models in a practical trust-region DFO algorithm:

- New sample points are only defined by the trust-region step $x + \Delta x$ (no model management iterations).

A practical interpolation-based trust-region solver

We have tested the effect of minimum ℓ_1-norm Hessian models in a practical trust-region DFO algorithm:

- New sample points are only defined by the trust-region step $x + \Delta x$ (no model management iterations).
- Quadratic underdetermined models are built by minimum ℓ_1 or Frobenius norm minimization.

A practical interpolation-based trust-region solver

We have tested the effect of minimum ℓ_1-norm Hessian models in a practical trust-region DFO algorithm:

- New sample points are only defined by the trust-region step $x + \Delta x$ (no model management iterations).
- Quadratic underdetermined models are built by minimum ℓ_1 or Frobenius norm minimization.
- Points too far from the current iterate are thrown away (sort of a criticality step).

A practical interpolation-based trust-region solver

We have tested the effect of minimum ℓ_1-norm Hessian models in a practical trust-region DFO algorithm:

- New sample points are only defined by the trust-region step $x + \Delta x$ (no model management iterations).
- Quadratic underdetermined models are built by minimum ℓ_1 or Frobenius norm minimization.
- Points too far from the current iterate are thrown away (sort of a criticality step).
- Trust-region radius is not reduced when the sample set has less than $n + 1$ points.

Performance profiles (accuracy of 10^{-4} in function values)

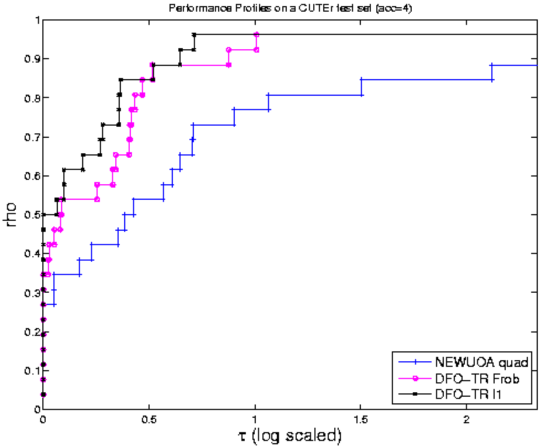


Figure: Performance profiles comparing DFO−TR (ℓ_1 and Frobenius) and NEWUOA (Powell) in a test set from CUTEr (Fasano et al.).

Performance profiles (accuracy of 10^{-6} in function values)

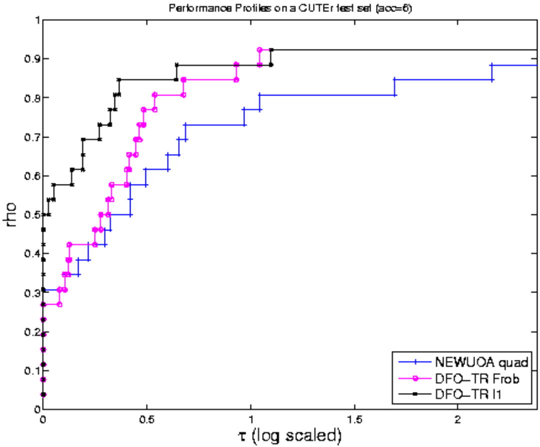


Figure: Performance profiles comparing DFO-TR (ℓ_1 and Frobenius) and NEWUOA (Powell) in a test set from CUTEr (Fasano et al.).

Concluding remarks

- Optimization is a fundamental tool in Compressed Sensing. However, this work shows that CS can also be ‘applied to’ Optimization.

Concluding remarks

- Optimization is a fundamental tool in Compressed Sensing. However, this work shows that CS can also be 'applied to' Optimization.

- In a sparse scenario, we were able to construct fully quadratic models with samples of size $\mathcal{O}(n \log^4 n)$ instead of the classical $\mathcal{O}(n^2)$.

Concluding remarks

- Optimization is a fundamental tool in Compressed Sensing. However, this work shows that CS can also be 'applied to' Optimization.

- In a sparse scenario, we were able to construct fully quadratic models with samples of size $\mathcal{O}(n \log^4 n)$ instead of the classical $\mathcal{O}(n^2)$.

- We proposed a practical DFO method (using ℓ_1-minimization) that was able to outperform state-of-the-art methods in several numerical tests (in the already 'tough' DFO scenario where n is small).

Open questions

- Improve the efficiency of the model ℓ_1-minimization, by properly warmstarting it (currently we solve it as an LP using `lipsol` by Y. Zhang).

Open questions

- Improve the efficiency of the model ℓ_1-minimization, by properly warmstarting it (currently we solve it as an LP using `lipsol` by Y. Zhang).

- Study the convergence properties of possibly stochastic interpolation-based trust-region methods.

References

- A. Bandeira, K. Scheinberg, and L. N. Vicente, Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization, 2011.
- A. Bandeira, K. Scheinberg, and L. N. Vicente, On partially sparse recovery, 2011.
- A. R. Conn, K. Scheinberg, and L. N. Vicente, Global convergence of general derivative-free trust-region algorithms to first and second order critical points, SIAM J. Optim., 20 (2009) 387–415.
- S. Gratton and L. N. Vicente, A surrogate management framework using rigorous trust-regions steps, 2011.

Direct-search methods

Definition

- *Sample the objective function at a finite number of points at each iteration.*
- *Base actions on those function values.*

Direct-search methods

Definition

- *Sample the objective function at a finite number of points at each iteration.*
- *Base actions on those function values.*

- Direct search of directional type: Achieve descent by using positive spanning sets and moving in the directions of the best points.

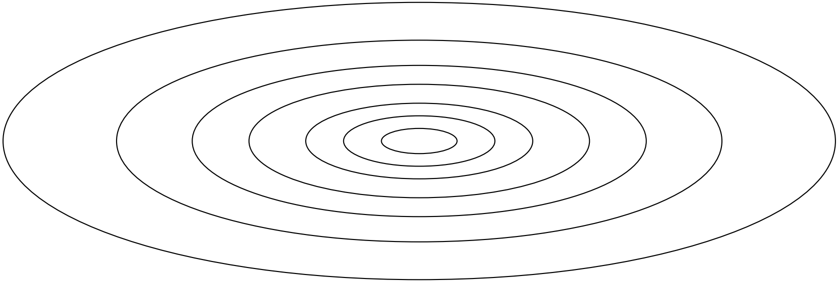
Direct-search methods

Definition

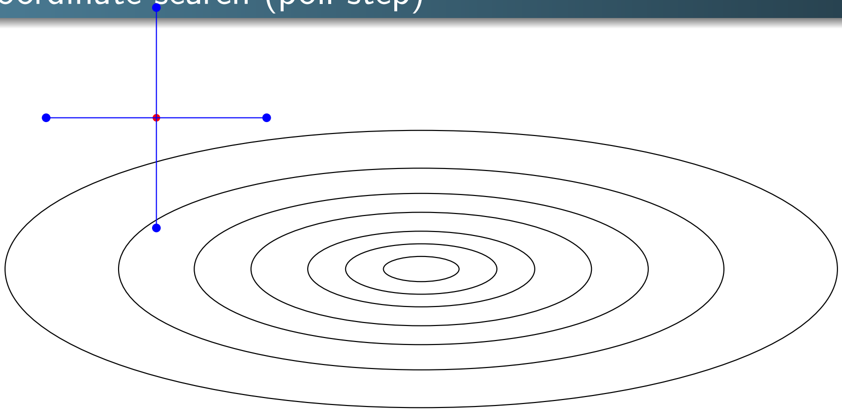
- *Sample the objective function at a finite number of points at each iteration.*
- *Base actions on those function values.*

- Direct search of directional type: Achieve descent by using positive spanning sets and moving in the directions of the best points.
- These methods do not necessarily depend on derivative approximation or model building (although they can be made much more efficient when doing so).

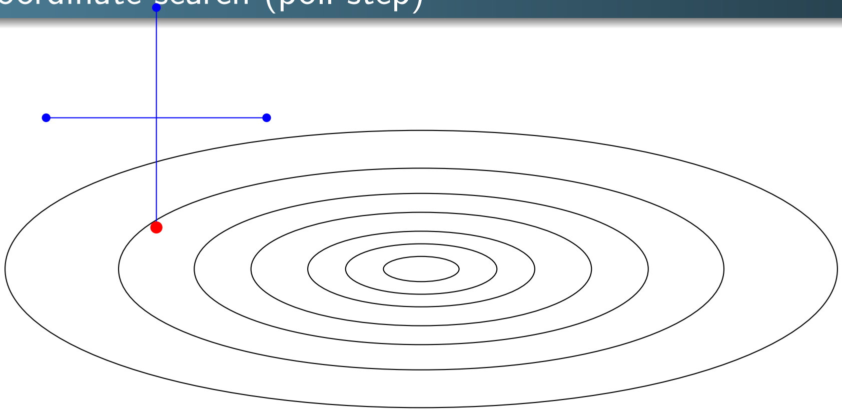
Coordinate search (poll step)



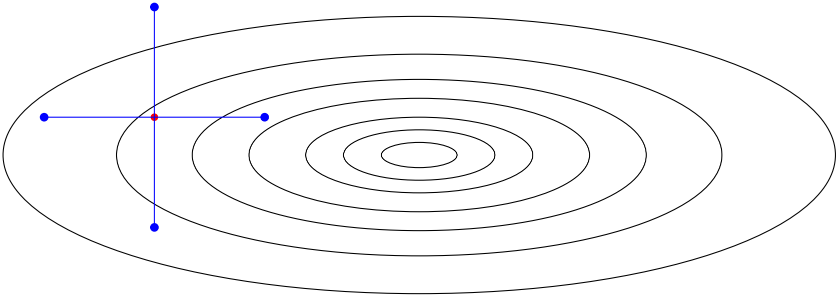
Coordinate search (poll step)



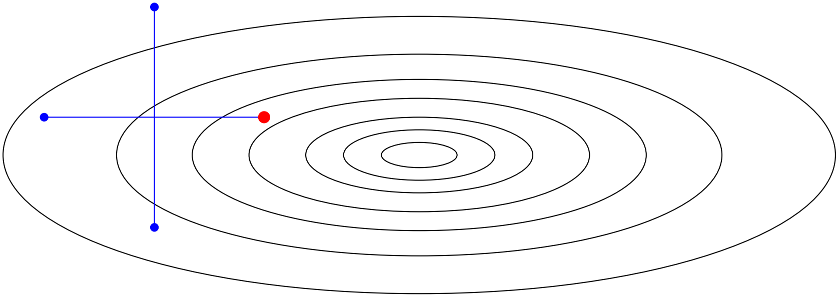
Coordinate search (poll step)



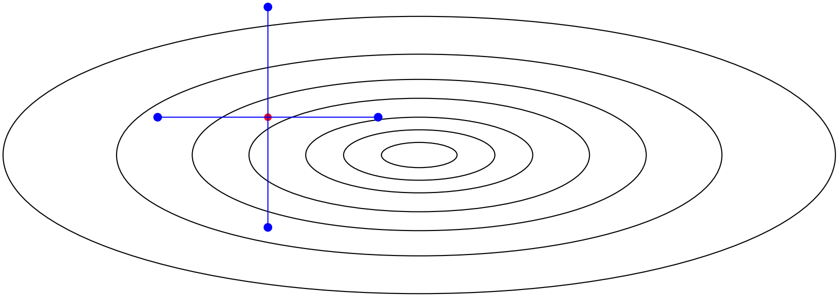
Coordinate search (poll step)



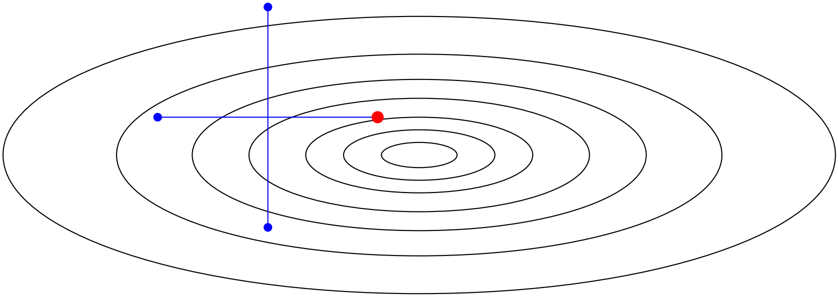
Coordinate search (poll step)



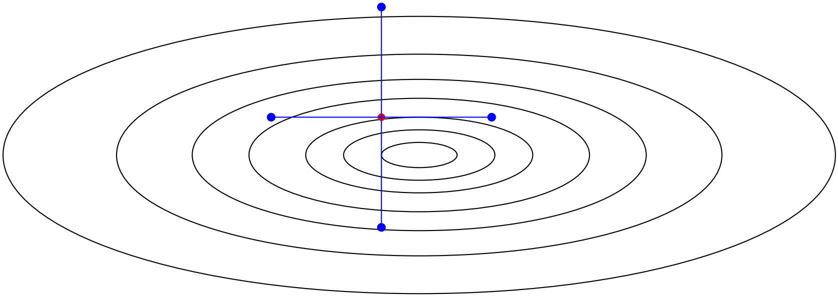
Coordinate search (poll step)



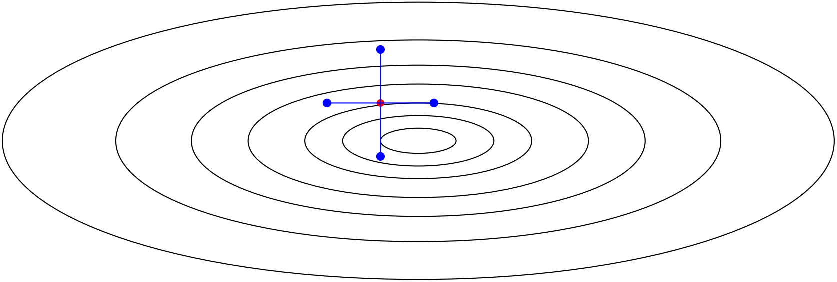
Coordinate search (poll step)



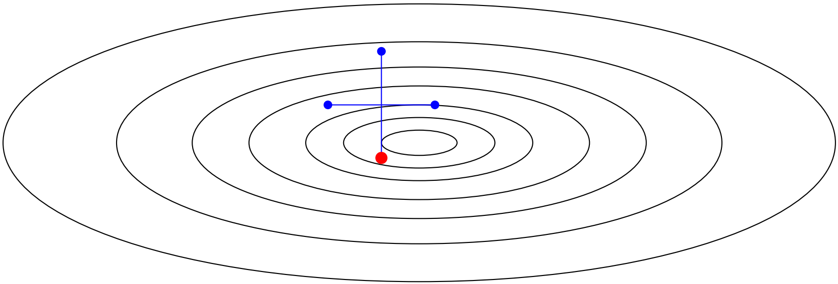
Coordinate search (poll step)



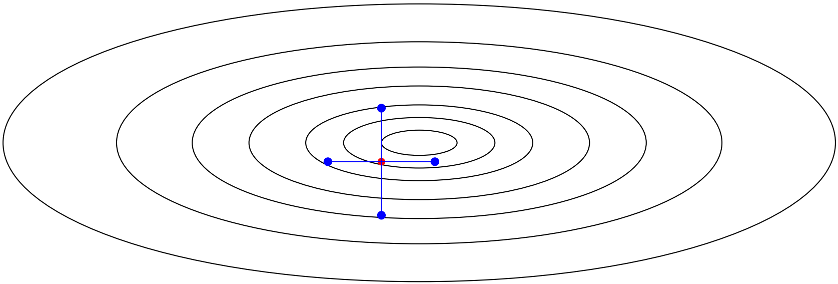
Coordinate search (poll step)



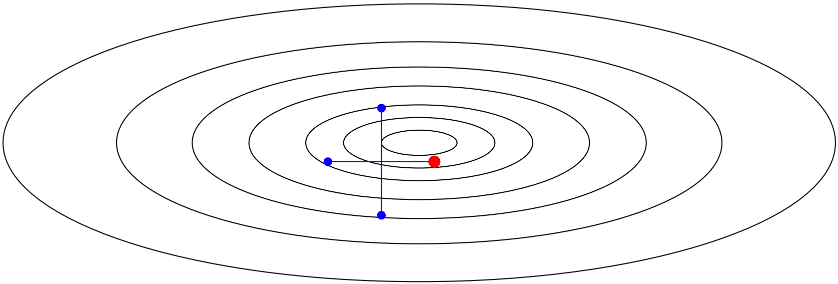
Coordinate search (poll step)



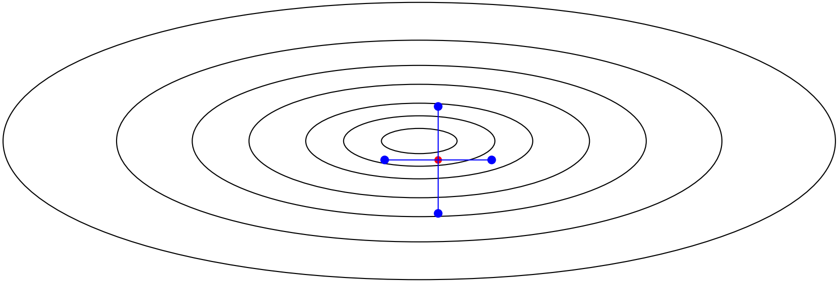
Coordinate search (poll step)



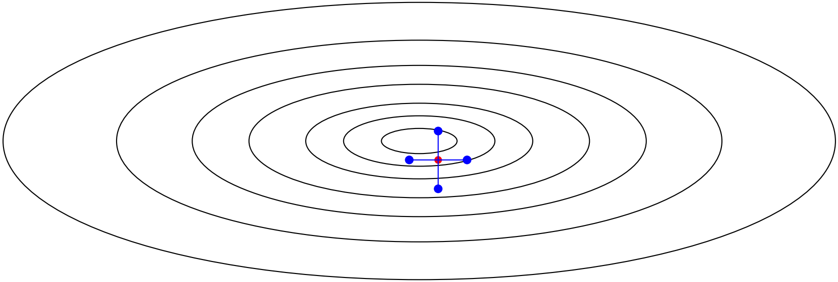
Coordinate search (poll step)



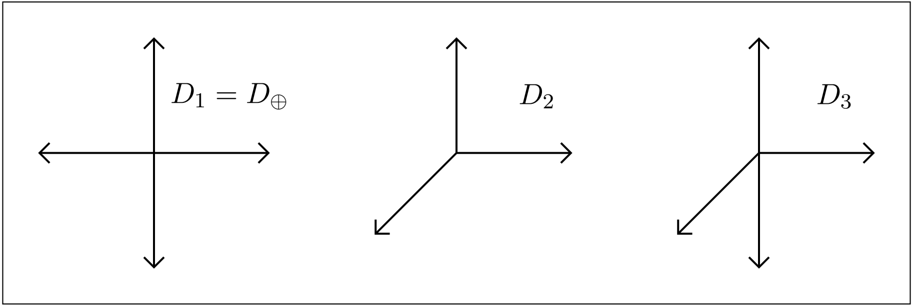
Coordinate search (poll step)



Coordinate search (poll step)



Positive spanning sets / positive bases



All of them are positive spanning sets (since they span $\mathbb{R}^n$ $(n = 2)$ with nonnegative coefficients).

D_1 and D_2 are positive bases.

Our problem setting

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f : \mathbb{R}^n \to \mathbb{R}$$

f is at least locally Lipschitz continous

Forcing function

A forcing function $\rho(\cdot)$ is a positive and monotonically nondecreasing function such that

$$\lim_{\alpha \downarrow 0} \frac{\rho(\alpha)}{\alpha} \;=\; 0.$$

Forcing function

A forcing function $\rho(\cdot)$ is a positive and monotonically nondecreasing function such that

$$\lim_{\alpha \downarrow 0} \frac{\rho(\alpha)}{\alpha} \;=\; 0.$$

In this talk, we will consider $\rho(\alpha) = \alpha^p$, with $p > 1$.

Forcing function

A forcing function $\rho(\cdot)$ is a positive and monotonically nondecreasing function such that

$$\lim_{\alpha \downarrow 0} \frac{\rho(\alpha)}{\alpha} \; = \; 0.$$

In this talk, we will consider $\rho(\alpha) = \alpha^p$, with $p > 1$.

A simple example of a forcing function is when $p = 2$: $\rho(\alpha) = \alpha^2$.

A class of direct-search methods

Initialization: Choose x_0 and $\alpha_0 > 0$.

A class of direct-search methods

Initialization: Choose x_0 and $\alpha_0 > 0$.

For $k = 0, 1, 2, \ldots$

(1) Search step (optional): Try to compute a point x with

$$f(x) < f(x_k) - \rho(\alpha_k)$$

by evaluating the function f at a finite number of points.

A class of direct-search methods

Initialization: Choose x_0 and $\alpha_0 > 0$.

For $k = 0, 1, 2, \ldots$

(1) Search step (optional): Try to compute a point x with

$$f(x) < f(x_k) - \rho(\alpha_k)$$

by evaluating the function f at a finite number of points.

If such a point is found then set $x_{k+1} = x$, declare the iteration and the search step successful, and skip the poll step.

A class of direct-search methods

(2) Poll step: Choose a positive spanning set D_k.

Order the set of poll points $P_k = \{x_k + \alpha_k d : \, d \in D_k\}$ and start evaluating f following the chosen order.

A class of direct-search methods

(2) Poll step: Choose a positive spanning set D_k.

Order the set of poll points $P_k = \{x_k + \alpha_k d : \, d \in D_k\}$ and start evaluating f following the chosen order.

If a point $x_k + \alpha_k d_k$ is found such that

$$f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k)$$

then stop polling, set $x_{k+1} = x_k + \alpha_k d_k$, and declare the iteration and the poll step successful.

(2) Poll step: Choose a positive spanning set D_k.

Order the set of poll points $P_k = \{x_k + \alpha_k d : d \in D_k\}$ and start evaluating f following the chosen order.

If a point $x_k + \alpha_k d_k$ is found such that

$$f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k)$$

then stop polling, set $x_{k+1} = x_k + \alpha_k d_k$, and declare the iteration and the poll step successful.

Otherwise declare the iteration (and the poll step) unsuccessful and set $x_{k+1} = x_k$.

(3) Step size update: If the iteration was successful then maintain or increase the step size parameter: $\alpha_{k+1} \in [\alpha_k, \gamma\alpha_k]$.

(3) Step size update: If the iteration was successful then maintain or increase the step size parameter: $\alpha_{k+1} \in [\alpha_k, \gamma\alpha_k]$.

Otherwise decrease the step size parameter: $\alpha_{k+1} \in [\beta_1\alpha_k, \beta_2\alpha_k]$.

A class of direct-search methods

(3) Step size update: If the iteration was successful then maintain or increase the step size parameter: $\alpha_{k+1} \in [\alpha_k, \gamma\alpha_k]$.

Otherwise decrease the step size parameter: $\alpha_{k+1} \in [\beta_1\alpha_k, \beta_2\alpha_k]$.

The parameters are chosen at initialization: $0 < \beta_1 \leq \beta_2 < 1$, and $\gamma \geq 1$.

Behavior of the step size parameter

Assumption

The level set $L(x_0) = \{x \in \mathbb{R}^n : \ f(x) \leq f(x_0)\}$ *is bounded. The function* f *is bounded below in* $L(x_0)$*.*

Behavior of the step size parameter

Assumption

The level set $L(x_0) = \{x \in \mathbb{R}^n : \; f(x) \leq f(x_0)\}$ *is bounded. The function* f *is bounded below in* $L(x_0)$*.*

Lemme (IDFO book or SIAM Review 2003 survey on DS)

There exists a point x_* *and a subsequence* K *of unsuccessful iterations such that*

$$\lim_{k \in K} x_k = x_* \quad \textit{and} \quad \lim_{k \in K} \alpha_k = 0.$$

Behavior of the step size parameter

Assumption

The level set $L(x_0) = \{x \in \mathbb{R}^n : \; f(x) \leq f(x_0)\}$ *is bounded. The function* f *is bounded below in* $L(x_0)$*.*

Lemme (IDFO book or SIAM Review 2003 survey on DS)

There exists a point x_* *and a subsequence* K *of unsuccessful iterations such that*

$$\lim_{k \in K} x_k = x_* \quad \textit{and} \quad \lim_{k \in K} \alpha_k = 0.$$

From such a result, one can then prove global convergence results (some form of stationarity independently of the starting point).

The question that interest us (smooth case)

Question

Given $\epsilon \in (0,1)$*, how many iterations* k *are needed to reach*

$$\|\nabla f(x_{k+1})\| \;\leq\; \epsilon.$$

Worst case complexity of direct search (smooth case)

Assumption

The norm of the vectors of any positive spanning set D_k are bounded above and away from zero.

The cosine measure of D_k is bounded away from zero.

Worst case complexity of direct search (smooth case)

Assumption

The norm of the vectors of any positive spanning set D_k are bounded above and away from zero.

The cosine measure of D_k is bounded away from zero.

Theorem (Lewis, Tolda, and Torczon 2003)

Let D_k be a positive spanning set.

Assume that ∇f is Lipschitz continuous (with constant $L_f > 0$).

If $f(x_k + \alpha_k d) \geq f(x_k) - \rho(\alpha_k)$, for all $d \in D_k$, then

$$\|\nabla f(x_k)\| \leq C(L_f, \textit{bounds}) \times \alpha_k \qquad \ldots \text{ in the case } \rho(\alpha) = \alpha^2.$$

Note that global convergence is deduced from here: $\|\nabla f(x_k)\| \underset{K}{\longrightarrow} 0$.

Worst case complexity of DS (smooth case)

Suppose that all iterations ℓ satisfy $\|\nabla f(x_\ell)\| \geq \epsilon$, $\ell = 0, \ldots, k$.

Worst case complexity of DS (smooth case)

Suppose that all iterations ℓ satisfy $\|\nabla f(x_\ell)\| \geq \epsilon$, $\ell = 0, \ldots, k$.

Then we obtain a lower bound on the step size:

$$\alpha_\ell \;\geq\; \frac{\epsilon}{C}, \quad \ell = 0, \ldots, k.$$

Worst case complexity of DS (smooth case)

Suppose that all iterations ℓ satisfy $\|\nabla f(x_\ell)\| \geq \epsilon$, $\ell = 0, \ldots, k$.

Then we obtain a lower bound on the step size:

$$\alpha_\ell \;\geq\; \frac{\epsilon}{C}, \quad \ell = 0, \ldots, k.$$

From sufficient decrease on succ. iteration, $f(x_{\ell+1}) \leq f(x_\ell) - \alpha_\ell^2$.

Worst case complexity of DS (smooth case)

Suppose that all iterations ℓ satisfy $\|\nabla f(x_\ell)\| \geq \epsilon$, $\ell = 0, \ldots, k$.

Then we obtain a lower bound on the step size:

$$\alpha_\ell \;\geq\; \frac{\epsilon}{C}, \quad \ell = 0, \ldots, k.$$

From sufficient decrease on succ. iteration, $f(x_{\ell+1}) \leq f(x_\ell) - \alpha_\ell^2$.

From this, we thus get a bound on the # of succ. iterations to reach $\|\nabla f(x_{k+1})\| \leq \epsilon$.

Worst case complexity of DS (smooth case)

Suppose that all iterations ℓ satisfy $\|\nabla f(x_\ell)\| \geq \epsilon$, $\ell = 0, \ldots, k$.

Then we obtain a lower bound on the step size:

$$\alpha_\ell \;\geq\; \frac{\epsilon}{C}, \quad \ell = 0, \ldots, k.$$

From sufficient decrease on succ. iteration, $f(x_{\ell+1}) \leq f(x_\ell) - \alpha_\ell^2$.

From this, we thus get a bound on the # of succ. iterations to reach $\|\nabla f(x_{k+1})\| \leq \epsilon$.

The # of unsucc. iterations is a function of the # succ. iterations.

Worst case complexity of DS (smooth case)

Suppose that all iterations ℓ satisfy $\|\nabla f(x_\ell)\| \geq \epsilon$, $\ell = 0, \ldots, k$.

Then we obtain a lower bound on the step size:

$$\alpha_\ell \;\geq\; \frac{\epsilon}{C}, \quad \ell = 0, \ldots, k.$$

From sufficient decrease on succ. iteration, $f(x_{\ell+1}) \leq f(x_\ell) - \alpha_\ell^2$.

From this, we thus get a bound on the # of succ. iterations to reach $\|\nabla f(x_{k+1})\| \leq \epsilon$.

The # of unsucc. iterations is a function of the # succ. iterations.

Thus...

Worst case complexity of DS (smooth case)

Theorem (LNV 2010)

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(\epsilon^{-\frac{p}{\min(p-1,1)}}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

Worst case complexity of DS (smooth case)

Theorem (LNV 2010)

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(\epsilon^{-\frac{p}{\min(p-1,1)}}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The number of function evaluations is $\mathcal{O}\left((n+1)\epsilon^{-\frac{p}{\min(p-1,1)}}\right)$.

Worst case complexity of DS (smooth case)

Theorem (LNV 2010)

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(\epsilon^{-\frac{p}{\min(p-1,1)}}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The number of function evaluations is $\mathcal{O}\left((n+1)\epsilon^{-\frac{p}{\min(p-1,1)}}\right)$.
- One obtains $\mathcal{O}\left(\epsilon^{-2}\right)$ for $p = 2$ as in steepest descent (Nesterov).

Worst case complexity of DS (smooth case)

Theorem (LNV 2010)

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(\epsilon^{-\frac{p}{\min(p-1,1)}}\right)$$

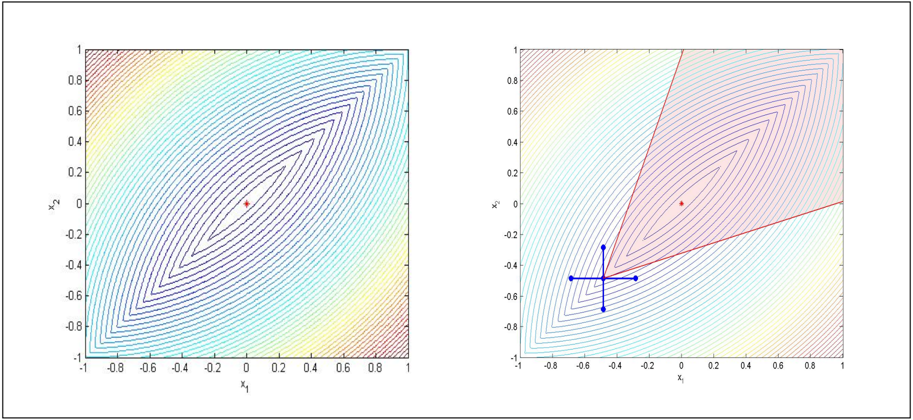
iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The number of function evaluations is $\mathcal{O}\left((n+1)\epsilon^{-\frac{p}{\min(p-1,1)}}\right)$.
- One obtains $\mathcal{O}\left(\epsilon^{-2}\right)$ for $p = 2$ as in steepest descent (Nesterov).

Reference:

- L. N. Vicente, Worst case complexity of direct search, preprint 10-17, Dept. of Mathematics, Univ. Coimbra, 2010.

Difficulties in the nonsmooth case



The cone of descent directions at the poll center is shaded.

Difficulties in the nonsmooth case

Thus, one needs to use an infinite number of polling directions.

Difficulties in the nonsmooth case

Thus, one needs to use an infinite number of polling directions.

This does not pose a problem to global convergence, which can be guaranteed a.e. in the unit sphere (see Audet and Dennis 2006, Vicente and Custódio 2011).

Difficulties in the nonsmooth case

Thus, one needs to use an infinite number of polling directions.

This does not pose a problem to global convergence, which can be guaranteed a.e. in the unit sphere (see Audet and Dennis 2006, Vicente and Custódio 2011).

But it does create a problem for worst case complexity:

Difficulties in the nonsmooth case

Thus, one needs to use an infinite number of polling directions.

This does not pose a problem to global convergence, which can be guaranteed a.e. in the unit sphere (see Audet and Dennis 2006, Vicente and Custódio 2011).

But it does create a problem for worst case complexity:

Since it is difficult to measure how many iterations are needed to find a descent one ...

Difficulties in the nonsmooth case

Thus, one needs to use an infinite number of polling directions.

This does not pose a problem to global convergence, which can be guaranteed a.e. in the unit sphere (see Audet and Dennis 2006, Vicente and Custódio 2011).

But it does create a problem for worst case complexity:

Since it is difficult to measure how many iterations are needed to find a descent one ...

... and thus to relate some form of stationarity (Clarke) to the step size.

One possible fix: Smoothing functions

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be a locally Lipschitz continuous function.*

One possible fix: Smoothing functions

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be a locally Lipschitz continuous function.*

We call $\tilde{f} : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}$ *a smoothing function of* f *if*

1. $\tilde{f}(\cdot, \mu)$ *is continuously differentiable in* $\mathbb{R}^n$ *for any* $\mu \in \mathbb{R}^{++}$*,*

One possible fix: Smoothing functions

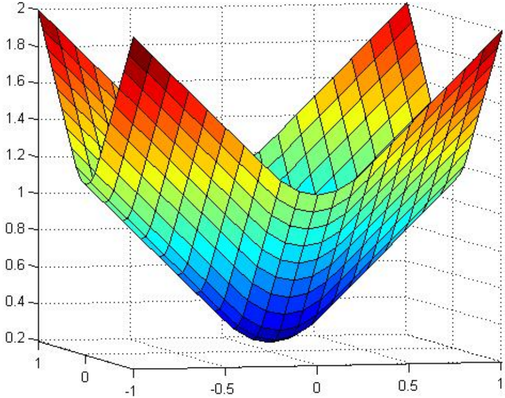
Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be a locally Lipschitz continuous function.*

We call $\tilde{f} : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}$ *a smoothing function of* f *if*

1. $\tilde{f}(\cdot, \mu)$ *is continuously differentiable in* $\mathbb{R}^n$ *for any* $\mu \in \mathbb{R}^{++}$*,*
2. *and, for any* $x \in \mathbb{R}^n$*,*

$$\lim_{z \to x, \mu \downarrow 0} \tilde{f}(z, \mu) = f(x).$$



A smoothing function of $|x_1| + |x_2|$ for $\mu = 0.5$.

A class of smoothing DS methods

Initialization: Choose a function $r(\cdot)$ such that $\lim_{\mu \downarrow 0} r(\mu) = 0$.

Choose $\mu_0 > 0$ and $\sigma \in (0, 1)$.

Choose $x_0 \in \mathbb{R}^n$.

For $k = 0, 1, \ldots$

- Apply DS to $\tilde{f}(\cdot, \mu_k)$ (starting from $y_{0,k} = x_k$) generating points $y_{0,k}, \ldots, y_{j,k}$ until $\alpha_{j+1,k} < r(\mu_k)$.
- Set $x_{k+1} = y_{j,k}$ and decrease the smoothing parameter: $\mu_{k+1} \in (0, \sigma \mu_k]$.

Global convergence of smoothing DS

Assumption (for all k)

The level sets $L(y_{0,k}) = \{y \in \mathbb{R}^n : \ \tilde{f}(y, \mu_k) \leq \tilde{f}(y_{0,k}, \mu_k)\}$ *are bounded. The functions* $\tilde{f}(\cdot, \mu_k)$ *are bounded below in* $L(y_{0,k})$.

Global convergence of smoothing DS

Assumption (for all k)

The level sets $L(y_{0,k}) = \{y \in \mathbb{R}^n : \tilde{f}(y, \mu_k) \leq \tilde{f}(y_{0,k}, \mu_k)\}$ *are bounded.*
The functions $\tilde{f}(\cdot, \mu_k)$ *are bounded below in* $L(y_{0,k})$.

If we let DS run forever for a given k, then $\liminf_{j \to +\infty} \alpha_{j,k} = 0$.

Global convergence of smoothing DS

Assumption (for all k)

The level sets $L(y_{0,k}) = \{y \in \mathbb{R}^n : \ \tilde{f}(y, \mu_k) \leq \tilde{f}(y_{0,k}, \mu_k)\}$ *are bounded. The functions* $\tilde{f}(\cdot, \mu_k)$ *are bounded below in* $L(y_{0,k})$.

If we let DS run forever for a given k, then $\liminf_{j \to +\infty} \alpha_{j,k} = 0$.

Thus, one always reaches the stopping criterion and μ_k is decreased.

Global convergence of smoothing DS

Assumption (for all k)

The level sets $L(y_{0,k}) = \{y \in \mathbb{R}^n : \ \tilde{f}(y, \mu_k) \leq \tilde{f}(y_{0,k}, \mu_k)\}$ *are bounded. The functions* $\tilde{f}(\cdot, \mu_k)$ *are bounded below in* $L(y_{0,k})$.

If we let DS run forever for a given k, then $\liminf_{j \to +\infty} \alpha_{j,k} = 0$.

Thus, one always reaches the stopping criterion and μ_k is decreased.

Theorem

The smoothing parameter goes to zero:

$$\lim_{k \to \infty} \mu_k \ = \ 0.$$

Global convergence of smoothing DS

Let j_k be the unsucc. internal DS iteration that achieves the stopping criterion $\alpha_{j_k+1,k} < r(\mu_k)$.

Global convergence of smoothing DS

Let j_k be the unsucc. internal DS iteration that achieves the stopping criterion $\alpha_{j_k+1,k} < r(\mu_k)$.

After having proved that μ_k goes to zero, one then obtains:

Global convergence of smoothing DS

Let j_k be the unsucc. internal DS iteration that achieves the stopping criterion $\alpha_{j_k+1,k} < r(\mu_k)$.

After having proved that μ_k goes to zero, one then obtains:

Theorem

1. $\lim\limits_{k\to+\infty} \alpha_{j_k,k} = 0.$
2. *There exists a point* x^* *and a subsequence* $K \subseteq \{j_1, j_2, \ldots\}$ *of unsucc. DS iterates such that* $x_k = y_{j_k,k} \underset{K}{\longrightarrow} x_*$.

Global convergence of smoothing DS

Now, $\|\nabla_x \tilde{f}(x_k, \mu_k)\| \;\leq\; C(L_{\tilde{f}})\alpha_{j_k}$

Global convergence of smoothing DS

Now, $\|\nabla_x \tilde{f}(x_k, \mu_k)\| \;\leq\; C(L_{\tilde{f}})\alpha_{j_k} \;\leq\; C(L_{\tilde{f}})r(\mu_k)$.

Now, $\|\nabla_x \tilde{f}(x_k, \mu_k)\| \;\leq\; C(L_{\tilde{f}})\alpha_{j_k} \;\leq\; C(L_{\tilde{f}})r(\mu_k)$.

Thus, choosing $r(\cdot)$ appropriately (i.e., $r(\mu) = \mu^2$ when $L_{\tilde{f}} = \frac{1}{\mu}$),

Global convergence of smoothing DS

Now, $\|\nabla_x \tilde{f}(x_k, \mu_k)\| \;\leq\; C(L_{\tilde{f}})\alpha_{j_k} \;\leq\; C(L_{\tilde{f}}) r(\mu_k)$.

Thus, choosing $r(\cdot)$ appropriately (i.e., $r(\mu) = \mu^2$ when $L_{\tilde{f}} = \frac{1}{\mu}$),

Theorem

$$\lim_{k \in K} \|\nabla_x \tilde{f}(x_k, \mu_k)\| \;=\; 0$$

and x_* is stationary point associated with the smoothing function $\tilde{f}$.

Now, $\|\nabla_x \tilde{f}(x_k, \mu_k)\| \;\leq\; C(L_{\tilde{f}})\alpha_{j_k} \;\leq\; C(L_{\tilde{f}})r(\mu_k)$.

Thus, choosing $r(\cdot)$ appropriately (i.e., $r(\mu) = \mu^2$ when $L_{\tilde{f}} = \frac{1}{\mu}$),

Theorem

$$\lim_{k \in K} \|\nabla_x \tilde{f}(x_k, \mu_k)\| \;=\; 0$$

and x_* is stationary point associated with the smoothing function $\tilde{f}$.

Definition

We say that x^* is a stationary point associated with the smoothing function $\tilde{f}$ if $0 \in G_{\tilde{f}}(x_*)$, where

$$G_{\tilde{f}}(x_*) = \{v : \; \exists N : \; x \xrightarrow[N]{} x_*, \;\; \mu \downarrow 0 \;\; \textit{with} \;\; \nabla_x \tilde{f}(x, \mu) \xrightarrow[N]{} v\}.$$

Clarke subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

What is true stationarity?

Clarke subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

What is true stationarity?

Definition

Let f be Lipschitz cont. near x_*. The Clarke subdifferential is given by:

$$\partial f(x_*) \;=\; \{d \in \mathbb{R}^n : f^\circ(x_*; v) \geq v^\top d \;\; \forall v \in \mathbb{R}^n\},$$

where the Clarke generalized directional derivative is defined by

$$f^\circ(x_*; v) \;=\; \limsup_{x \to x_* \;\; t \downarrow 0} \frac{f(x + tv) - f(x)}{t}.$$

Clarke subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

What is true stationarity?

Definition

Let f be Lipschitz cont. near x_*. The Clarke subdifferential is given by:

$$\partial f(x_*) \; = \; \{d \in \mathbb{R}^n : f^\circ(x_*; v) \geq v^\top d \;\; \forall v \in \mathbb{R}^n\},$$

where the Clarke generalized directional derivative is defined by

$$f^\circ(x_*; v) \; = \; \limsup_{x \to x_* \; t \downarrow 0} \frac{f(x + tv) - f(x)}{t}.$$

Definition

We say that x_* is a Clarke stationary point if $0 \in \partial f(x_*)$.

Clarke subdifferential (alternative characterization)

Theorem

Let f be Lipschitz continuous near x_*.

Let D_f be the subset of $\mathbb{R}^n$ where f is differentiable.

Clarke subdifferential (alternative characterization)

Theorem

Let f be Lipschitz continuous near x_*.

Let D_f be the subset of $\mathbb{R}^n$ where f is differentiable.

Then the Clarke subdifferential can be given by

$$\partial f(x_*) \;=\; \mathrm{co}\{\lim \nabla f(x) :\; x \to x_*,\; x \in D_f\},$$

where co represents the convex hull.

How to construct smoothing functions

Definition

We say that the sequence $\{\psi^\mu : \mathbb{R}^n \to \mathbb{R}^+, \mu \in \mathbb{R}^{++}\}$ *is a mollifier if:*

- $B^\mu = \{z : \psi^\mu(z) > 0\}$ *converges to* $\{0\}$, *as* $\mu \downarrow 0$,
- $\int_{\mathbb{R}^n} \psi^\mu(z)dz = 1.$

How to construct smoothing functions

Definition

We say that the sequence $\{\psi^\mu : \mathbb{R}^n \to \mathbb{R}^+, \mu \in \mathbb{R}^{++}\}$ *is a mollifier if:*

- $B^\mu = \{z : \psi^\mu(z) > 0\}$ *converges to* $\{0\}$, *as* $\mu \downarrow 0$,
- $\int_{\mathbb{R}^n} \psi^\mu(z) dz = 1$.

Now consider the averaged functions

$$\tilde{f}(x, \mu) = \int_{\mathbb{R}^n} f(x - z)\psi^\mu(z) dz = \int_{\mathbb{R}^n} f(z)\psi^\mu(x - z) dz.$$

How to construct smoothing functions

Definition

We say that the sequence $\{\psi^\mu : \mathbb{R}^n \to \mathbb{R}^+, \mu \in \mathbb{R}^{++}\}$ *is a mollifier if:*

- $B^\mu = \{z : \psi^\mu(z) > 0\}$ *converges to* $\{0\}$*, as* $\mu \downarrow 0$,
- $\int_{\mathbb{R}^n} \psi^\mu(z)dz = 1$.

Now consider the averaged functions

$$\tilde{f}(x, \mu) = \int_{\mathbb{R}^n} f(x - z)\psi^\mu(z)dz = \int_{\mathbb{R}^n} f(z)\psi^\mu(x - z)dz.$$

If the mollifiers $\{\psi^\mu\}$ are bounded and continuous on $\mathbb{R}^n$, then $\tilde{f}$ is a smoothing function of f and one has the gradient consistency property

$$\partial f(x_*) \;=\; \text{co}\, G_{\tilde{f}}(x_*).$$

How to construct smoothing functions

In particular, mollifiers can be built from density functions.

How to construct smoothing functions

In particular, mollifiers can be built from density functions.

Let B be a bounded set and $\psi : B \longrightarrow \mathbb{R}^+$ be a density function with $\int_B \psi(z)dz = 1$. The following is a mollifier:

$$\psi^\mu(z) = \begin{cases} \frac{\psi(x/\mu)}{\mu^n} & \text{if } z \in \mu B, \\ 0 & \text{otherwise.} \end{cases}$$

How to construct smoothing functions

In particular, mollifiers can be built from density functions.

Let B be a bounded set and $\psi : B \longrightarrow \mathbb{R}^+$ be a density function with $\int_B \psi(z)dz = 1$. The following is a mollifier:

$$\psi^\mu(z) = \begin{cases} \frac{\psi(x/\mu)}{\mu^n} & \text{if } z \in \mu B, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $\nabla_x \tilde{f}(\cdot, \mu)$ is Lipschitz continuous with constant

$$L_{\tilde{f}} \;=\; \mathcal{O}\left(\frac{1}{\mu^2}\right).$$

How to construct smoothing functions

In particular, mollifiers can be built from density functions.

Let B be a bounded set and $\psi : B \longrightarrow \mathbb{R}^+$ be a density function with $\int_B \psi(z)dz = 1$. The following is a mollifier:

$$\psi^\mu(z) = \begin{cases} \frac{\psi(x/\mu)}{\mu^n} & \text{if } z \in \mu B, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $\nabla_x \tilde{f}(\cdot, \mu)$ is Lipschitz continuous with constant

$$L_{\tilde{f}} \ = \ \mathcal{O}\left(\frac{1}{\mu^2}\right).$$

There are other forms of building smoothing functions such that

$$L_{\tilde{f}} \ = \ \mathcal{O}\left(\frac{1}{\mu}\right).$$

Worst case complexity of smoothing DS

Theorem

Any smoothing DS (based on sufficient decrease) takes at most (when $r(\mu) = \mu^q$)

$$\mathcal{O}\left((-\log \xi)\xi^{-pq}\right)$$

iterations to reduce μ below $\xi \in (0,1)$.

After such effort, the gradient of $\tilde{f}$ is $\mathcal{O}\left(\xi^{q-1} + \xi^{(p-1)q}\right)$.

Worst case complexity of smoothing DS

Theorem

Any smoothing DS (based on sufficient decrease) takes at most (when $r(\mu) = \mu^q$*)*

$$\mathcal{O}\left((-\log \xi)\xi^{-pq}\right)$$

iterations to reduce μ *below* $\xi \in (0,1)$*.*

After such effort, the gradient of $\tilde{f}$ *is* $\mathcal{O}\left(\xi^{q-1} + \xi^{(p-1)q}\right)$*.*

Thus, as long as $q \geq 2$ and $(p-1)q \geq 1$, one arrives at a gradient of $\tilde{f}$ of $\mathcal{O}(\xi)$.

Worst case complexity of smoothing DS

Theorem

Any smoothing DS (based on sufficient decrease) takes at most (when $r(\mu) = \mu^q$)

$$\mathcal{O}\left((-\log \xi)\xi^{-pq}\right)$$

iterations to reduce μ below $\xi \in (0,1)$.

After such effort, the gradient of $\tilde{f}$ is $\mathcal{O}\left(\xi^{q-1} + \xi^{(p-1)q}\right)$.

Thus, as long as $q \geq 2$ and $(p-1)q \geq 1$, one arrives at a gradient of $\tilde{f}$ of $\mathcal{O}(\xi)$.

Optimal choices consist of $q = 2$ and $p = 3/2$, leading to a worst case cost of

$$\mathcal{O}\left((-\log(\xi))\xi^{-3}\right).$$

A smoothing function for $\|F(\cdot)\|_1$

Chen and Zhou have introduced the following smoothing function of $|t|$:

$$\tilde{s}(t,\mu) = \int_{-\infty}^{\infty} |t - \mu\tau|\rho(\tau)d\tau,$$

where $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ is a piecewise continuous density function with a finite number of pieces satisfying

$$\rho(\tau) = \rho(-\tau) \quad \text{and} \quad \int_{-\infty}^{\infty} |\tau|\rho(\tau)d\tau < \infty.$$

A smoothing function for $\|F(\cdot)\|_1$

Chen and Zhou have introduced the following smoothing function of $|t|$:

$$\tilde{s}(t,\mu) = \int_{-\infty}^{\infty} |t - \mu\tau| \rho(\tau) d\tau,$$

where $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ is a piecewise continuous density function with a finite number of pieces satisfying

$$\rho(\tau) = \rho(-\tau) \text{ and } \int_{-\infty}^{\infty} |\tau|\rho(\tau)d\tau < \infty.$$

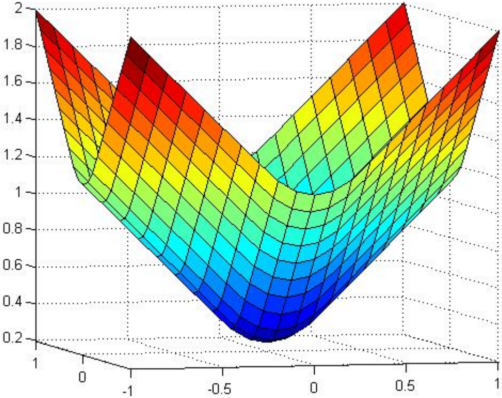
Using the density corresponding to the so-called Steklov mollifier,

$$\rho(\tau) = \begin{cases} 1 & \text{if } \tau \in [-\frac{1}{2}, \frac{1}{2}], \\ 0 & \text{otherwise,} \end{cases}$$

one obtains

$$\tilde{s}(t,\mu) = \begin{cases} \frac{t^2}{\mu} + \frac{\mu}{4} & \text{if } t \in [-\frac{\mu}{2}, \frac{\mu}{2}], \\ |t| & \text{otherwise.} \end{cases}$$

Example of this smoothing function



The smoothing function $\tilde{s}(x_1, \mu) + \tilde{s}(x_2, \mu)$ of $|x_1| + |x_2|$ for $\mu = 0.5$.

A smoothing function for $\|F(\cdot)\|_1$

Proposition (Chen and Zhou 2010)

(i) $\tilde{s}$ *is a smoothing function of* $|\cdot|$.

A smoothing function for $\|F(\cdot)\|_1$

Proposition (Chen and Zhou 2010)

(i) $\tilde{s}$ *is a smoothing function of* $|\cdot|$.

(ii) $\nabla_t \tilde{s}(\cdot, \mu)$ *is Lipschitz continuous with* $L_{\tilde{s}} = \mathcal{O}\left(\frac{1}{\mu}\right)$.

A smoothing function for $\|F(\cdot)\|_1$

Proposition (Chen and Zhou 2010)

(i) $\tilde{s}$ *is a smoothing function of* $|\cdot|$.

(ii) $\nabla_t \tilde{s}(\cdot, \mu)$ *is Lipschitz continuous with* $L_{\tilde{s}} = \mathcal{O}\left(\frac{1}{\mu}\right)$.

(iii) $\tilde{s}$ *is gradient consistent:*

$$\left\{ \lim_{t \to 0, \mu \downarrow 0} \tilde{s}'(t, \mu) \right\} \; = \; [-1, 1] \; = \; \partial |\cdot|(0).$$

A smoothing function for $\|F(\cdot)\|_1$

Now, let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1. Then $\tilde{s}(f)$ is a smoothing function of $|f|$.

A smoothing function for $\|F(\cdot)\|_1$

Now, let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1. Then $\tilde{s}(f)$ is a smoothing function of $|f|$.

Let $F : \mathbb{R}^n \to \mathbb{R}^m$ with $F(x) = (F_1(x), \ldots, F_m(x))$, where each F_i is C^1 and ∇F_i is Lips. continuous.

A smoothing function for $\|F(\cdot)\|_1$

Now, let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1. Then $\tilde{s}(f)$ is a smoothing function of $|f|$.

Let $F : \mathbb{R}^n \to \mathbb{R}^m$ with $F(x) = (F_1(x), \ldots, F_m(x))$, where each F_i is C^1 and ∇F_i is Lips. continuous.

Theorem

(i) $\tilde{F} = \sum_{i=1}^m \tilde{s}(F_i)$ *is a smoothing function of* $\|F\|_1 = \sum_{i=1}^m |F_i|$.

A smoothing function for $\|F(\cdot)\|_1$

Now, let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1. Then $\tilde{s}(f)$ is a smoothing function of $|f|$.

Let $F : \mathbb{R}^n \to \mathbb{R}^m$ with $F(x) = (F_1(x), \ldots, F_m(x))$, where each F_i is C^1 and ∇F_i is Lips. continuous.

Theorem

(i) $\tilde{F} = \sum_{i=1}^m \tilde{s}(F_i)$ *is a smoothing function of* $\|F\|_1 = \sum_{i=1}^m |F_i|$.

(i) $\tilde{F}$ *satisfies the gradient consistent property*

$$\left\{ \lim_{x \to x^*, \mu \downarrow 0} \nabla_x \tilde{F}(x, \mu) \right\} \;=\; \partial \|F(x_*)\|_1.$$

A smoothing function for $\|F(\cdot)\|_1$

Now, let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1. Then $\tilde{s}(f)$ is a smoothing function of $|f|$.

Let $F : \mathbb{R}^n \to \mathbb{R}^m$ with $F(x) = (F_1(x), \ldots, F_m(x))$, where each F_i is C^1 and ∇F_i is Lips. continuous.

Theorem

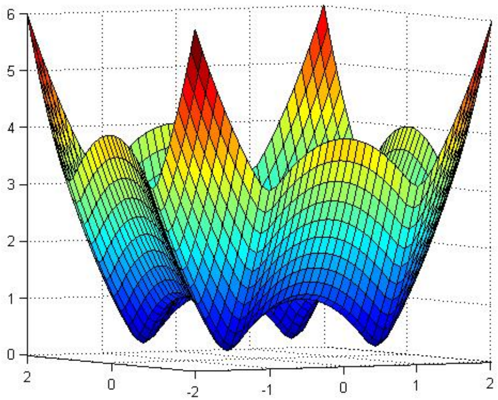
(i) $\tilde{F} = \sum_{i=1}^m \tilde{s}(F_i)$ *is a smoothing function of* $\|F\|_1 = \sum_{i=1}^m |F_i|$.

(i) $\tilde{F}$ *satisfies the gradient consistent property*

$$\left\{ \lim_{x \to x^*, \mu \downarrow 0} \nabla_x \tilde{F}(x, \mu) \right\} = \partial \|F(x_*)\|_1.$$

(iii) For each μ, $\nabla_x \tilde{F}(\cdot, \mu)$ *is Lipschitz cont. with constant* $L_{\tilde{F}} = \mathcal{O}\left(\frac{1}{\mu}\right)$.

Example of this smoothing function



The smoothing function $\tilde{F}(x_1, x_2, \mu)$
for $\|F(x_1, x_2)\|_1 = \|(x_1^2 - 1, x_2^2 - 1)\|_1$ and $\mu = 0.5$.

Some numerical experiments

We have tested the smoothing direct-search approach on the MATLAB direct-search `sid-psm` code:

- A. L. Custódio and L. N. Vicente, *Using sampling and simplex derivatives in pattern search methods*, SIAM Journal on Optimization, 18 (2007), 537-555.
- A. L. Custódio, H. Rocha, and L. N. Vicente, *Incorporating minimum Frobenius norm models in direct search*, Computational Optimization and Applications, 46 (2010) 265–278.

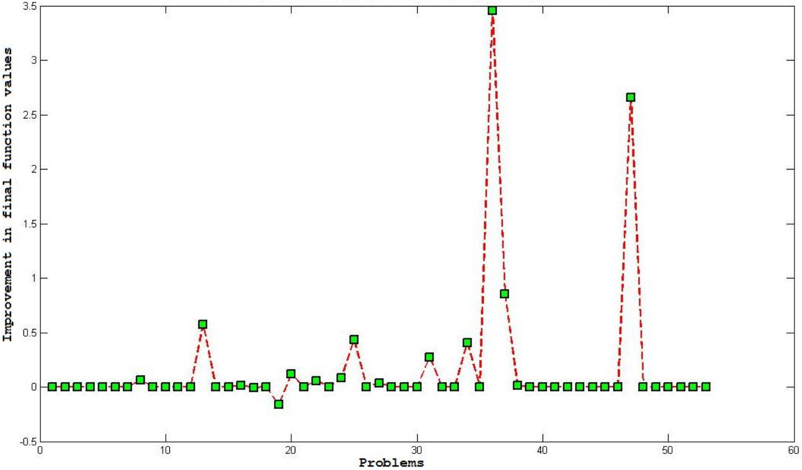
Some numerical experiments

We have tested the smoothing direct-search approach on the MATLAB direct-search `sid-psm` code:

- A. L. Custódio and L. N. Vicente, *Using sampling and simplex derivatives in pattern search methods*, SIAM Journal on Optimization, 18 (2007), 537-555.
- A. L. Custódio, H. Rocha, and L. N. Vicente, *Incorporating minimum Frobenius norm models in direct search*, Computational Optimization and Applications, 46 (2010) 265–278.

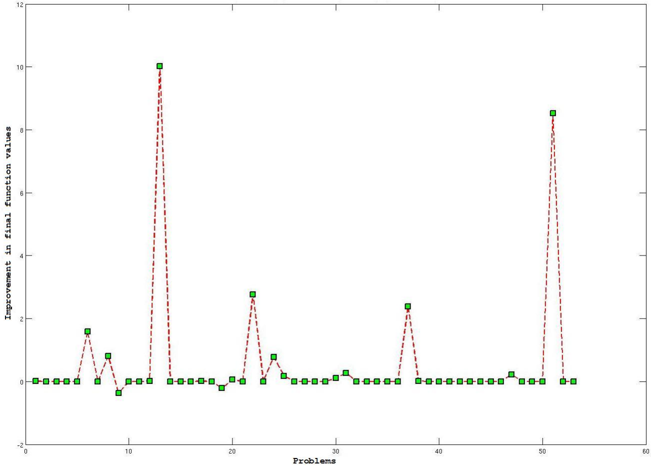
We tested the piecewise-linear problems $(\min \|F(\cdot)\|_1)$ from:

- J. J. Moré and S. M. Wild, *Benchmarking derivative-free optimization algorithms*, SIAM Journal on Optimization, 20 (2009), 172–191.



Smoothing DS with $\mu_0 = 10^{-2}$ vs DS
(no search step, cycling polling).

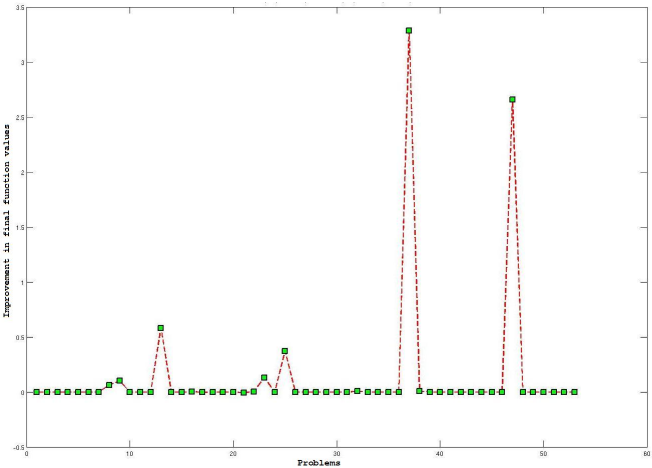
Some numerical experiments



Smoothing DS with $\mu_0 = 10^{-2}$ vs DS

(search step using smoothing function with $\mu = 10^{-4}$, cycling polling).

Some numerical experiments



Smoothing DS with $\mu_0 = 10^{-2}$ vs DS
(no search step, polling using simplex gradient of smoothing function with $\mu = 10^{-4}$).

Conclusions

We have developed a smoothing direct-search approach using smoothing functions.

Conclusions

We have developed a smoothing direct-search approach using smoothing functions.

We have proved that the smoothing DS method is globally convergent.

Conclusions

We have developed a smoothing direct-search approach using smoothing functions.

We have proved that the smoothing DS method is globally convergent.

Smoothing DS is costly but seems able to find better solutions.

Conclusions

We have developed a smoothing direct-search approach using smoothing functions.

We have proved that the smoothing DS method is globally convergent.

Smoothing DS is costly but seems able to find better solutions.

We have derived a complexity worst case bound for direct-search methods in the non-smooth case.