Interpolation-Based Trust-Region Methods for DFO

Luis Nunes Vicente
University of Coimbra

(joint work with A. Bandeira, A. R. Conn, S. Gratton, and K. Scheinberg)

July 27, 2010 — ICCOPT, Santiago

`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).

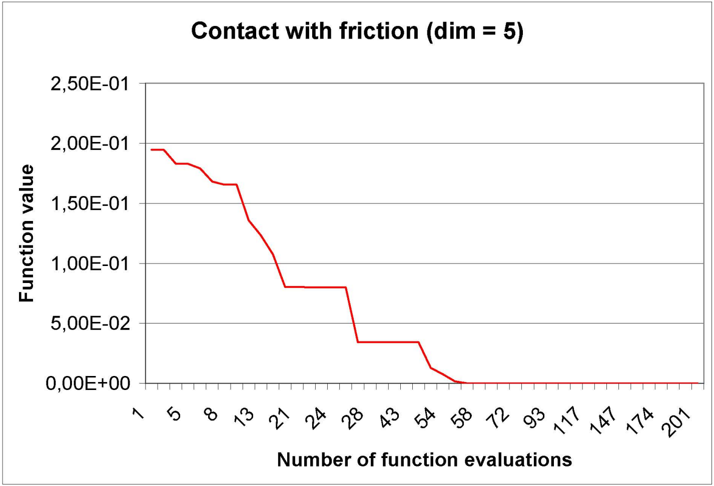
Why Derivative-Free Optimization?

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):



For a recent talk on Direct Search (8th EUROPT, 2010) see:

`http://www.mat.uc.pt/~lnv/talks`

For a recent talk on Direct Search (8th EUROPT, 2010) see:

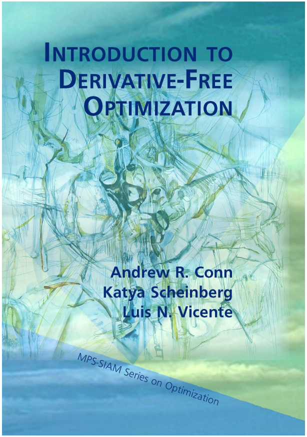
http://www.mat.uc.pt/~lnv/talks

Ana Luisa Custodio — Talk WA04 10:30.

Ismael Vaz — Talk WB04 13:30.

The book!

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



Model-based trust-region methods

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

Trust-region subproblem

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

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:

- 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:

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

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

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

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:

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

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

$$|f(y) - m(y)| \leq \kappa_{ef}\,\Delta^3 \qquad \forall y \in B(x;\Delta).$$

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

$$M(\phi, Y) = \begin{bmatrix} \phi_0(y^0) & \phi_1(y^0) & \cdots & \phi_p(y^0) \\ \phi_0(y^1) & \phi_1(y^1) & \cdots & \phi_p(y^1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(y^p) & \phi_1(y^p) & \cdots & \phi_p(y^p) \end{bmatrix} \quad f(Y) = \begin{bmatrix} f(y^0) \\ f(y^1) \\ \vdots \\ f(y^p) \end{bmatrix}.$$

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

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

Natural/canonical basis

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

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

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

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

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.

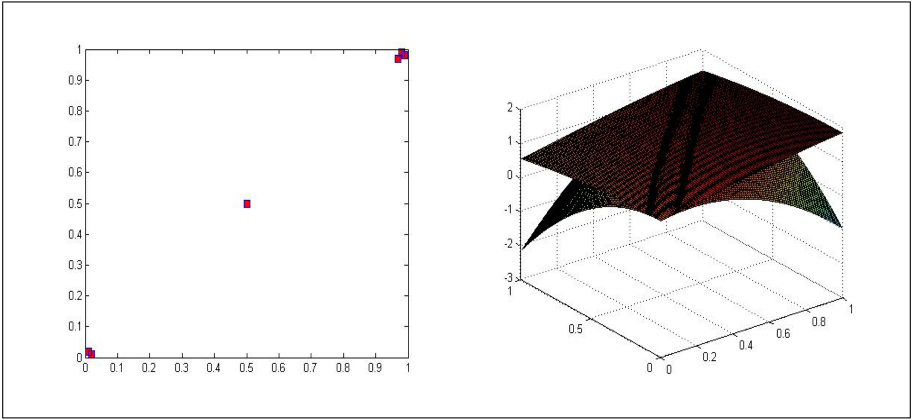
An equivalent definition of Λ–poisedness is ($|Y| = |\alpha|$)

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

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

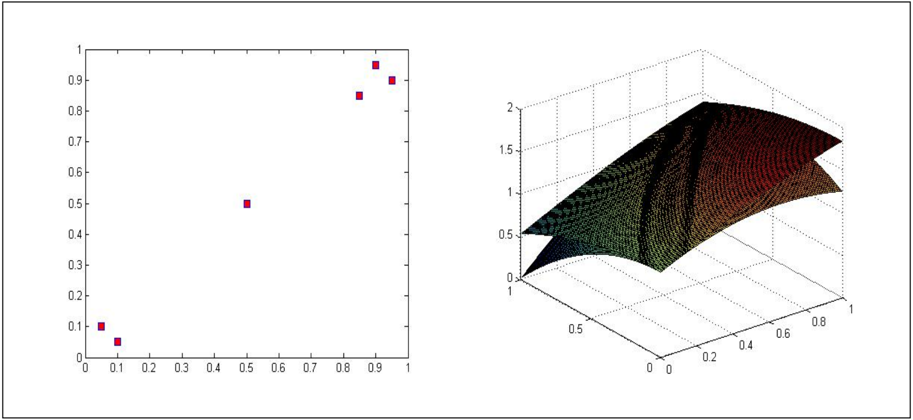
Non-squared cases are defined analogously (IDFO).

A badly poised set



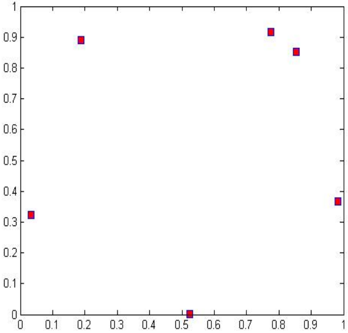
$$\Lambda = 21296.$$

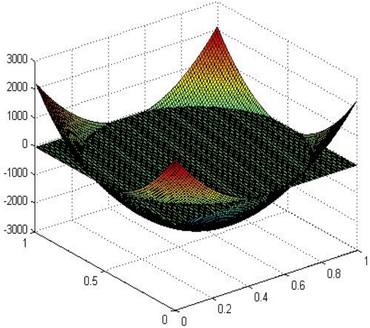
A not so badly poised set



$$\Lambda = 440.$$

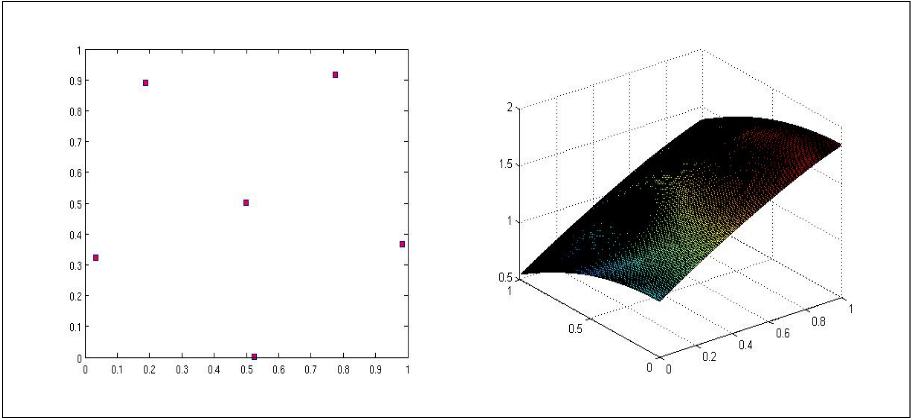
Another badly poised set





$$\Lambda = 524982.$$

An ideal set



$$\Lambda = 1.$$

Quadratic interpolation models

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

- Overdetermined when $|Y| > |\alpha|$. See talk on Direct Search!

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

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

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

Minimum Frobenius norm models

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

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

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

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

Minimum Frobenius norm models

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

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

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:

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

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:

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

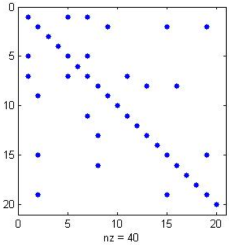
$\longrightarrow$ MFN models are fully linear.

Sparsity on the Hessian

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

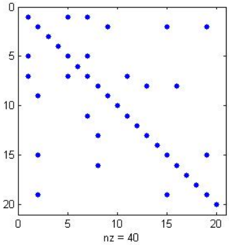
Sparsity on the Hessian

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



Sparsity on the Hessian

- 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 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?

Compressed sensing — sparse recovery

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

Compressed sensing — sparse recovery

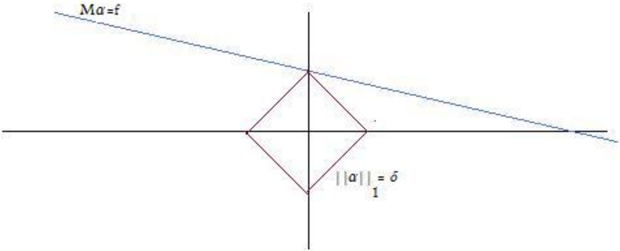
- Objective: Find sparse α subject to a highly underdetermined linear system $M\alpha = f$.
- $\begin{cases} \min & \|\alpha\|_0 = |\operatorname{supp}(\alpha)| \\ \text{s.t.} & M\alpha = f \end{cases}$ is NP-Hard.

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 $M\alpha = f$.
- $\begin{cases} \min & \|\alpha\|_0 = |\operatorname{supp}(\alpha)| \\ \text{s.t.} & M\alpha = f \end{cases}$ 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 $2\delta_{2s} + \delta_s < 1$ then it can be recovered by ℓ_1-minimization:

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

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 $2\delta_{2s} + \delta_s < 1$ then it can be recovered by ℓ_1-minimization:

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

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

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?

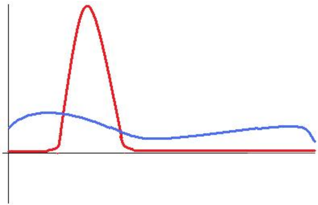
- Choose orthonormal bases.

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.
- Avoid localized functions ($\|\phi_i\|_{L^\infty}$ should be uniformly bounded).

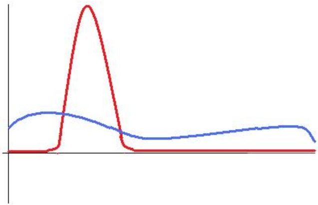


Bounded orthonormal expansions (Rauhut)

Question

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

- Choose orthonormal bases.
- Avoid localized functions ($\|\phi_i\|_{L^\infty}$ should be uniformly bounded).



- 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*.*
- $\frac{p}{\log p} \geq c_1 K^2 s (\log s)^2 \log N$*.*

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.
- $\frac{p}{\log p} \geq c_1 K^2 s (\log s)^2 \log N$.

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* $\|\phi_i\|_{L^\infty} \leq K$.
- *each point of* Y *is drawn independently according to* μ.
- $\frac{p}{\log p} \geq c_1 K^2 s (\log s)^2 \log N$.

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*

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.
- $\frac{p}{\log p} \geq c_1 K^2 s (\log s)^2 \log N$.

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*

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

Sparse orthonormal bounded expansion recovery

Theorem (Rauhut, 2010)

If

- *ϕ is orthonormal in a probability measure μ and $\|\phi_i\|_{L^\infty} \leq K$.*
- *each point of Y is drawn independently according to μ.*
- $\frac{p}{\log p} \geq c_1 K^2 s (\log s)^2 \log N$.

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

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

Then,

$$\|\bar{\alpha} - \alpha^*\|_2 \leq \frac{C}{\sqrt{p}}\eta.$$

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:

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

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

Proposition (Bandeira, Scheinberg, and Vicente, 2010)

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, 2010)

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

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

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

Proposition (Bandeira, Scheinberg, and Vicente, 2010)

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).

Hessian sparse recovery

Let us look again at

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

where

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

Let us look again at

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

where

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

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

Hessian sparse recovery

Let us look again at

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

where

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

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

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

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

where

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

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

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

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

Hessian sparse recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2010)

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

Hessian sparse recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2010)

If
- *the Hessian of f at 0 is $s-$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, 2010)

If
- *the Hessian of f at 0 is $s-$sparse.*
- *Y is a random sample set chosen w.r.t. the uniform measure on $B_\infty(0;\Delta)$.*
- $\frac{p}{\log p} \geq 9c_1(s+n+1)\log^2(s+n+1)\log \mathcal{O}(n^2)$.

Theorem (Bandeira, Scheinberg, and Vicente, 2010)

If
- *the Hessian of f at 0 is $s-$sparse.*
- *Y is a random sample set chosen w.r.t. the uniform measure on $B_\infty(0;\Delta)$.*
- $\frac{p}{\log p} \geq 9c_1(s+n+1)\log^2(s+n+1)\log\mathcal{O}(n^2)$.

Then, with high probability, the quadratic

Hessian sparse recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2010)

If

- *the Hessian of f at 0 is $s-$sparse.*
- *Y is a random sample set chosen w.r.t. the uniform measure on $B_\infty(0;\Delta)$.*
- $\frac{p}{\log p} \geq 9c_1(s+n+1)\log^2(s+n+1)\log \mathcal{O}(n^2)$.

Then, with high probability, the quadratic

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

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

An answer to our question

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

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

An answer to our question

- For instance, when the number of non-zeros of the Hessian is $s = \mathcal{O}(n)$, 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.

Open question

- Generalize the result above when minimizing only the ℓ_1-norm of the Hessian (α_Q) rather than of the whole α.

  $\longrightarrow$ Numerical simulations have shown that such approach is (slightly) advantageous.

Remarks

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

Remarks

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

- Solve

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

Remarks

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

- Solve

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

- 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 approach we:

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

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

- Use deterministic sampling.

Interpolation-based trust-region methods

Trust-region methods for DFO typically:

- attempt to form quadratic models (by interpolation/regression and using polynomials or radial basis functions)

$$m_k(x_k + s) \;=\; f(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$$

  based on (well poised) sample sets.

Interpolation-based trust-region methods

Trust-region methods for DFO typically:

- attempt to form quadratic models (by interpolation/regression and using polynomials or radial basis functions)

$$m_k(x_k + s) \;=\; f(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$$

  based on (well poised) sample sets.

$\longrightarrow$ Well poisedness ensures fully linear or fully quadratic models.

Interpolation-based trust-region methods

Trust-region methods for DFO typically:

- attempt to form quadratic models (by interpolation/regression and using polynomials or radial basis functions)

$$m_k(x_k + s) \;=\; f(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$$

  based on (well poised) sample sets.

$\longrightarrow$ Well poisedness ensures fully linear or fully quadratic models.

- Calculate a step s_k by approximately solving the trust-region subproblem

$$\min_{s \in B_2(x_k;\Delta_k)} \quad m_k(x_k + s).$$

- Set x_{k+1} to $x_k + s_k$ (success) or to x_k (unsuccess) and update Δ_k depending on the value of

$$\rho_k \;=\; \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

- Set x_{k+1} to $x_k + s_k$ (success) or to x_k (unsuccess) and update Δ_k depending on the value of

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

- Attempt to accept steps based on simple decrease, i.e., if

$$\rho_k > 0 \iff f(x_k + s_k) < f(x_k).$$

Interpolation-based trust-region methods (IDFO)

- Reduce Δ_k only if ρ_k is small and the model is FL/FQ — unsuccessful iterations.

Interpolation-based trust-region methods (IDFO)

- Reduce Δ_k only if ρ_k is small and the model is FL/FQ — unsuccessful iterations.

- Accept new iterates based on simple decrease $(\rho_k > 0)$ as long as the model is FL/FQ — acceptable iterations.

Interpolation-based trust-region methods (IDFO)

- Reduce Δ_k only if ρ_k is small and the model is FL/FQ — unsuccessful iterations.

- Accept new iterates based on simple decrease $(\rho_k > 0)$ as long as the model is FL/FQ — acceptable iterations.

- Allow for model-improving iterations (when ρ_k is not large enough and the model is not certifiably FL/FQ).

Interpolation-based trust-region methods (IDFO)

- Reduce Δ_k only if ρ_k is small and the model is FL/FQ — unsuccessful iterations.

- Accept new iterates based on simple decrease $(\rho_k > 0)$ as long as the model is FL/FQ — acceptable iterations.

- Allow for model-improving iterations (when ρ_k is not large enough and the model is not certifiably FL/FQ).

  $\longrightarrow$ Do not reduce Δ_k.

Interpolation-based trust-region methods (IDFO)

- Incorporate a criticality step (1st or 2nd order) when the ‘stationarity’ of the model is small.

- Incorporate a criticality step (1st or 2nd order) when the ‘stationarity’ of the model is small.

  $\longrightarrow$ Internal cycle of reductions of Δ_k — until model is well poised in $B(x_k; \|g_k\|)$.

Behavior of the trust-region radius

Due to the criticality step, one has for successful iterations:

$$f(x_k) - f(x_{k+1}) \;\geq\; \mathcal{O}(\|g_k\| \min\{\|g_k\|, \Delta_k\}) \geq \; \mathcal{O}(\Delta_k^2).$$

Behavior of the trust-region radius

Due to the criticality step, one has for successful iterations:

$$f(x_k) - f(x_{k+1}) \;\geq\; \mathcal{O}(\|g_k\| \min\{\|g_k\|, \Delta_k\}) \geq \; \mathcal{O}(\Delta_k^2).$$

Thus:

Theorem (Conn, Scheinberg, and Vicente, 2009)

The trust-region radius converges to zero:

$$\lim_{k \to +\infty} \Delta_k \;=\; 0.$$

Behavior of the trust-region radius

Due to the criticality step, one has for successful iterations:

$$f(x_k) - f(x_{k+1}) \;\geq\; \mathcal{O}(\|g_k\| \min\{\|g_k\|, \Delta_k\}) \geq \; \mathcal{O}(\Delta_k^2).$$

Thus:

Theorem (Conn, Scheinberg, and Vicente, 2009)

The trust-region radius converges to zero:

$$\lim_{k \to +\infty} \Delta_k \;=\; 0.$$

$\longrightarrow$ Similar to direct-search methods where $\liminf_{k \to +\infty} \alpha_k = 0$.

Analysis of TR methods (1st order)

Using fully linear models:

Theorem (Conn, Scheinberg, and Vicente, 2009)

If ∇f *is Lips. continuous and* f *is bounded below on* $L(x_0)$ *then*

$$\lim_{k \to +\infty} \|\nabla f(x_k)\| \ = \ 0.$$

Analysis of TR methods (1st order)

Using fully linear models:

Theorem (Conn, Scheinberg, and Vicente, 2009)

If ∇f is Lips. continuous and f is bounded below on $L(x_0)$ then

$$\lim_{k \to +\infty} \|\nabla f(x_k)\| \;=\; 0.$$

$\longrightarrow$ Valid also for simple decrease (acceptable iterations).

Analysis of TR methods (2nd order)

Using fully quadratic models:

Theorem (Conn, Scheinberg, and Vicente, 2009)

If $\nabla^2 f$ is Lips. continuous and f is bounded below on $L(x_0)$ then

$$\lim_{k \to +\infty} \max \left\{ \|\nabla f(x_k)\|, -\lambda_{min}[\nabla^2 f(x_k)] \right\} \; = \; 0.$$

Analysis of TR methods (2nd order)

Using fully quadratic models:

Theorem (Conn, Scheinberg, and Vicente, 2009)

If $\nabla^2 f$ is Lips. continuous and f is bounded below on $L(x_0)$ then

$$\lim_{k \to +\infty} \max \left\{ \|\nabla f(x_k)\|, -\lambda_{min}[\nabla^2 f(x_k)] \right\} \; = \; 0.$$

$\longrightarrow$ Valid also for simple decrease (acceptable iterations).

Analysis of TR methods (2nd order)

Using fully quadratic models:

Theorem (Conn, Scheinberg, and Vicente, 2009)

If $\nabla^2 f$ is Lips. continuous and f is bounded below on $L(x_0)$ then

$$\lim_{k \to +\infty} \max \left\{ \|\nabla f(x_k)\|, -\lambda_{min}[\nabla^2 f(x_k)] \right\} \ = \ 0.$$

$\longrightarrow$ Valid also for simple decrease (acceptable iterations).

$\longrightarrow$ Going from $\liminf$ to $\lim$ requires changing the update of Δ_k.

Sample set management

Recently, Fasano, Morales, and Nocedal (2009) suggested an one-point exchange:

Sample set management

Recently, Fasano, Morales, and Nocedal (2009) suggested an one-point exchange:

- In successful iterations:

$$Y_{k+1} \;=\; Y_k \cup \{x_k + s_k\} \setminus \{y_{out}\}.$$

where $y_{out} = \operatorname{argmax} \|y - x_k\|_2$.

Sample set management

Recently, Fasano, Morales, and Nocedal (2009) suggested an one-point exchange:

- In successful iterations:

$$Y_{k+1} \;=\; Y_k \cup \{x_k + s_k\} \setminus \{y_{out}\}.$$

where $y_{out} = \text{argmax}\, \|y - x_k\|_2$.

- In the unsuccessful case:

$$Y_{k+1} \;=\; Y_k \cup \{x_k + s_k\} \setminus \{y_{out}\} \quad \text{if} \quad \|y_{out} - x_k\| \geq \|s_k\|.$$

Sample set management

Recently, Fasano, Morales, and Nocedal (2009) suggested an one-point exchange:

- In successful iterations:

$$Y_{k+1} = Y_k \cup \{x_k + s_k\} \setminus \{y_{out}\}.$$

  where $y_{out} = \operatorname{argmax} \|y - x_k\|_2$.

- In the unsuccessful case:

$$Y_{k+1} = Y_k \cup \{x_k + s_k\} \setminus \{y_{out}\} \quad \text{if} \quad \|y_{out} - x_k\| \geq \|s_k\|.$$

- Do not perform model-improving iterations.

Sample set management

Recently, Fasano, Morales, and Nocedal (2009) suggested an one-point exchange:

- In successful iterations:

$$Y_{k+1} \;=\; Y_k \cup \{x_k + s_k\} \setminus \{y_{out}\}.$$

where $y_{out} = \text{argmax}\,\|y - x_k\|_2$.

- In the unsuccessful case:

$$Y_{k+1} \;=\; Y_k \cup \{x_k + s_k\} \setminus \{y_{out}\} \quad \text{if} \quad \|y_{out} - x_k\| \geq \|s_k\|.$$

- Do not perform model-improving iterations.

They observed sample sets not badly poised!

Self-correcting geometry

Later, Scheinberg and Toint (2009) proposed:

Self-correcting geometry

Later, Scheinberg and Toint (2009) proposed:

- A self-correcting geometry approach based on one-point exchanges, globally convergent to first-order stationary points ($\liminf$).

Self-correcting geometry

Later, Scheinberg and Toint (2009) proposed:

- A self-correcting geometry approach based on one-point exchanges, globally convergent to first-order stationary points ($\liminf$).
- In the unsuccessful case, y_{out} is not only based on $\|y - x_k\|_2$, but also on the values of the Lagrange polynomials at $x_k + s_k$.

Self-correcting geometry

Later, Scheinberg and Toint (2009) proposed:

- A self-correcting geometry approach based on one-point exchanges, globally convergent to first-order stationary points ($\liminf$).
- In the unsuccessful case, y_{out} is not only based on $\|y - x_k\|_2$, but also on the values of the Lagrange polynomials at $x_k + s_k$.
- They showed that, if Δ_k is small compared to $\|g_k\|$, then the step either improves the function or the geometry/poisedness of the model.

Self-correcting geometry

Later, Scheinberg and Toint (2009) proposed:

- A self-correcting geometry approach based on one-point exchanges, globally convergent to first-order stationary points ($\liminf$).
- In the unsuccessful case, y_{out} is not only based on $\|y - x_k\|_2$, but also on the values of the Lagrange polynomials at $x_k + s_k$.
- They showed that, if Δ_k is small compared to $\|g_k\|$, then the step either improves the function or the geometry/poisedness of the model.
- In their approach, model-improving iterations are not needed.

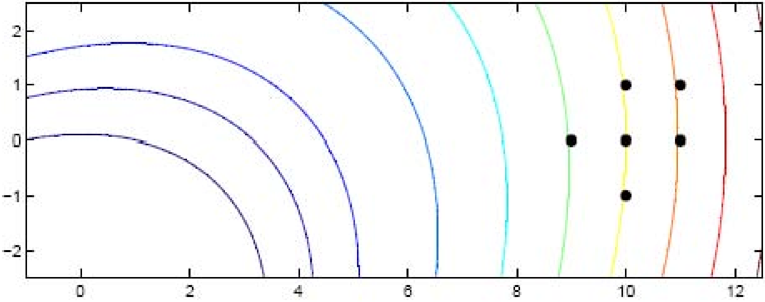
Self-correcting geometry

Later, Scheinberg and Toint (2009) proposed:

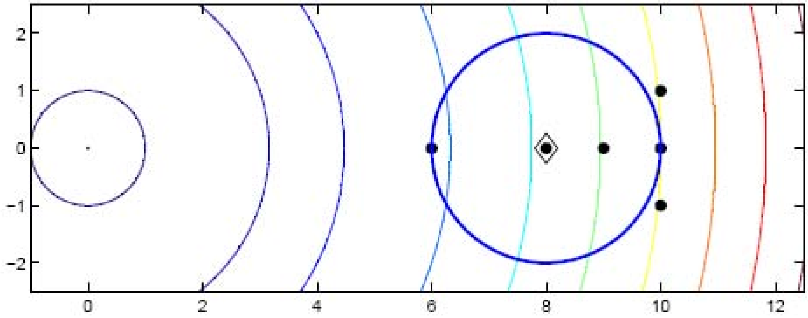
- A self-correcting geometry approach based on one-point exchanges, globally convergent to first-order stationary points ($\liminf$).
- In the unsuccessful case, y_{out} is not only based on $\|y - x_k\|_2$, but also on the values of the Lagrange polynomials at $x_k + s_k$.
- They showed that, if Δ_k is small compared to $\|g_k\|$, then the step either improves the function or the geometry/poisedness of the model.
- In their approach, model-improving iterations are not needed.

They showed, however, that the criticality step is indeed necessary.

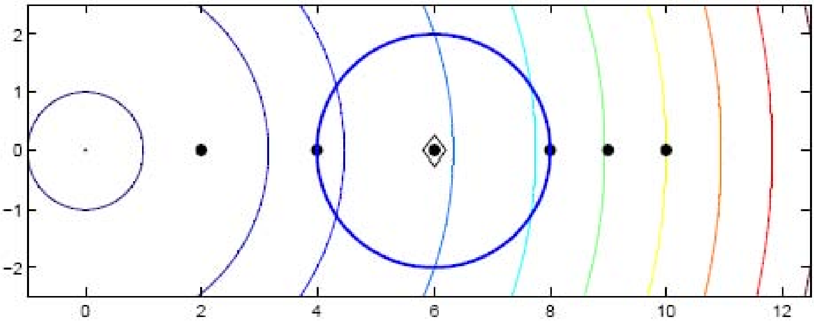
Without criticality step... (Scheinberg and Toint)



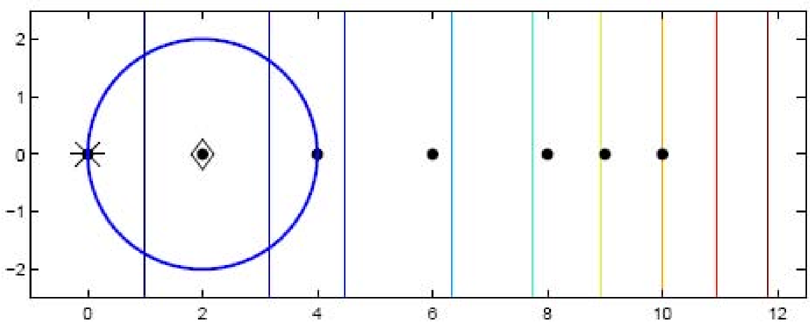
Without criticality step... (Scheinberg and Toint)



Without criticality step... (Scheinberg and Toint)



Without criticality step... (Scheinberg and Toint)



A practical interpolation-based trust-region method

A practical interpolation-based trust-region method

Model building:

- If $|Y_k| = N = \mathcal{O}(n^2)$, use determined quadratic interpolation.

A practical interpolation-based trust-region method

Model building:

- If $|Y_k| = N = \mathcal{O}(n^2)$, use determined quadratic interpolation.
- Otherwise use ℓ_1 $(p = 1)$ or Frobenius $(p = 2)$ minimum norm quadratic interpolation:

$$\begin{array}{ll} \min & \frac{1}{p}\|\alpha_Q\|_p^p \\ \text{s. t.} & M(\bar{\phi}_L, Y_{scaled})\alpha_L + M(\bar{\phi}_Q, Y_{scaled})\alpha_Q = f(Y_{scaled}). \end{array}$$

A practical interpolation-based trust-region method

Model building:

- If $|Y_k| = N = \mathcal{O}(n^2)$, use determined quadratic interpolation.
- Otherwise use ℓ_1 $(p = 1)$ or Frobenius $(p = 2)$ minimum norm quadratic interpolation:

$$\begin{array}{ll} \min & \frac{1}{p}\|\alpha_Q\|_p^p \\ \text{s. t.} & M(\bar{\phi}_L, Y_{scaled})\alpha_L + M(\bar{\phi}_Q, Y_{scaled})\alpha_Q = f(Y_{scaled}). \end{array}$$

Sample set update — one starts with $|Y_0| = \mathcal{O}(n)$:

A practical interpolation-based trust-region method

Model building:

- If $|Y_k| = N = \mathcal{O}(n^2)$, use determined quadratic interpolation.
- Otherwise use ℓ_1 $(p = 1)$ or Frobenius $(p = 2)$ minimum norm quadratic interpolation:

$$\begin{array}{ll} \min & \frac{1}{p}\|\alpha_Q\|_p^p \\ \text{s.t.} & M(\bar{\phi}_L, Y_{scaled})\alpha_L + M(\bar{\phi}_Q, Y_{scaled})\alpha_Q = f(Y_{scaled}). \end{array}$$

Sample set update — one starts with $|Y_0| = \mathcal{O}(n)$:

- If $|Y_k| < N = \mathcal{O}(n^2)$, set $Y_{k+1} = Y_k \cup \{x_k + s_k\}$.

A practical interpolation-based trust-region method

Model building:

- If $|Y_k| = N = \mathcal{O}(n^2)$, use determined quadratic interpolation.
- Otherwise use ℓ_1 $(p = 1)$ or Frobenius $(p = 2)$ minimum norm quadratic interpolation:

$$\begin{array}{ll} \min & \frac{1}{p}\|\alpha_Q\|_p^p \\ \text{s.t.} & M(\bar{\phi}_L, Y_{scaled})\alpha_L + M(\bar{\phi}_Q, Y_{scaled})\alpha_Q = f(Y_{scaled}). \end{array}$$

Sample set update — one starts with $|Y_0| = \mathcal{O}(n)$:

- If $|Y_k| < N = \mathcal{O}(n^2)$, set $Y_{k+1} = Y_k \cup \{x_k + s_k\}$.
- Otherwise as in Fasano et al., but with $y_{out} = \operatorname{argmax} \|y - x_{k+1}\|_2$.

A practical interpolation-based trust-region method

Model building:

- If $|Y_k| = N = \mathcal{O}(n^2)$, use determined quadratic interpolation.
- Otherwise use ℓ_1 $(p = 1)$ or Frobenius $(p = 2)$ minimum norm quadratic interpolation:

$$\begin{array}{ll} \min & \frac{1}{p}\|\alpha_Q\|_p^p \\ \text{s. t.} & M(\bar{\phi}_L, Y_{scaled})\alpha_L + M(\bar{\phi}_Q, Y_{scaled})\alpha_Q = f(Y_{scaled}). \end{array}$$

Sample set update — one starts with $|Y_0| = \mathcal{O}(n)$:

- If $|Y_k| < N = \mathcal{O}(n^2)$, set $Y_{k+1} = Y_k \cup \{x_k + s_k\}$.
- Otherwise as in Fasano et al., but with $y_{out} = \operatorname{argmax} \|y - x_{k+1}\|_2$.

'Criticality step': If Δ_k is very small, discard points far away from the trust region.

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

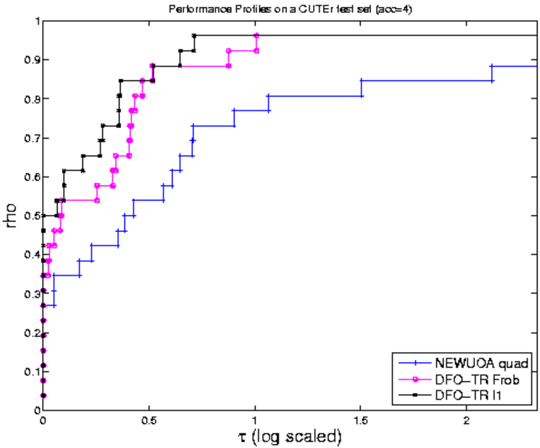


Figure: Performance profiles comparing `DFO−TR` (ℓ_1 and Frobenius) and `NEWUOA` (Powell) in a test set from `CUTEr` (Fasano et al.).

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

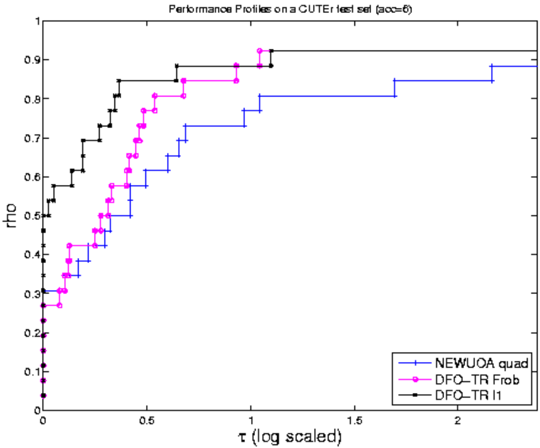


Figure: Performance profiles comparing DFO−TR (ℓ_1 and Frobenius) and NEWUOA (Powell) in a test set from CUTEr (Fasano et al.).

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.

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.

- Investigate ℓ_1-minimization techniques in statistical models (like Kriging for interpolating 'sparse' data sets), but applied to Optimization.

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.

- Investigate ℓ_1-minimization techniques in statistical models (like Kriging for interpolating 'sparse' data sets), but applied to Optimization.

- Develop a globally convergent model-based trust-region method for non-smooth functions.

References

- A. Bandeira, K. Scheinberg, and L. N. Vicente, Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization, in preparation, 2010.
- 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.
- G. Fasano, J. L. Morales, and J. Nocedal, On the geometry phase in model-based algorithms for derivative-free optimization, Optim. Methods Softw., 24 (2009) 145–154.
- K. Scheinberg and Ph. L. Toint, Self-correcting geometry in model-based algorithms for derivative-free unconstrained optimization, 2009.

Other recent work

- S. Gratton, Ph. L. Toint, and A. Tröltzsch, An active-set trust-region method for derivative-free nonlinear bound-constrained optimization, 2010. Talk WA02 11:30

Other recent work

- S. Gratton, Ph. L. Toint, and A. Tröltzsch, An active-set trust-region method for derivative-free nonlinear bound-constrained optimization, 2010. Talk WA02 11:30
- S. Gratton and L. N. Vicente, A surrogate management framework using rigorous trust-regions steps, 2010.

Other recent work

- S. Gratton, Ph. L. Toint, and A. Tröltzsch, An active-set trust-region method for derivative-free nonlinear bound-constrained optimization, 2010. Talk WA02 11:30
- S. Gratton and L. N. Vicente, A surrogate management framework using rigorous trust-regions steps, 2010.
- M. J. D. Powell, The BOBYQA algorithm for bound constrained optimization without derivatives, 2009.

Other recent work

- S. Gratton, Ph. L. Toint, and A. Tröltzsch, An active-set trust-region method for derivative-free nonlinear bound-constrained optimization, 2010. Talk WA02 11:30
- S. Gratton and L. N. Vicente, A surrogate management framework using rigorous trust-regions steps, 2010.
- M. J. D. Powell, The BOBYQA algorithm for bound constrained optimization without derivatives, 2009.
- S. M. Wild and C. Shoemaker, Global convergence of radial basis function trust region derivative-free algorithms, 2009.

Optimization 2011 (July 24–27, Portugal)

2011
optimization
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Caparica *Campus*
July 24-27

plenary speakers

Gilbert Laporte | HEC Montréal
New trends in vehicle routing

Jean Bernard Lasserre | LAAS-CNRS, Toulouse
Moments and semidefinite relaxations for parametric optimization

José Mario Martínez | State University of Campinas
Unifying inexact restoration, SQP, and augmented Lagrangian methods

Mauricio G.C. Resende | AT&T Labs - Research
Using metaheuristics to solve real optimization problems in telecommunications

Nick Sahinidis | Carnegie Mellon University
Recent advances in nonconvex optimization

Stephen J. Wright | University of Wisconsin
Algorithms and applications in sparse optimization