Sparse Hessian Recovery and Trust-Region Methods based on Probabilistic Models

Luis Nunes Vicente
University of Coimbra

joint work with A. S. Bandeira (Princeton) and K. Scheinberg (Lehigh)

March 16, 2012 — 3rd Conference on OMS

`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).

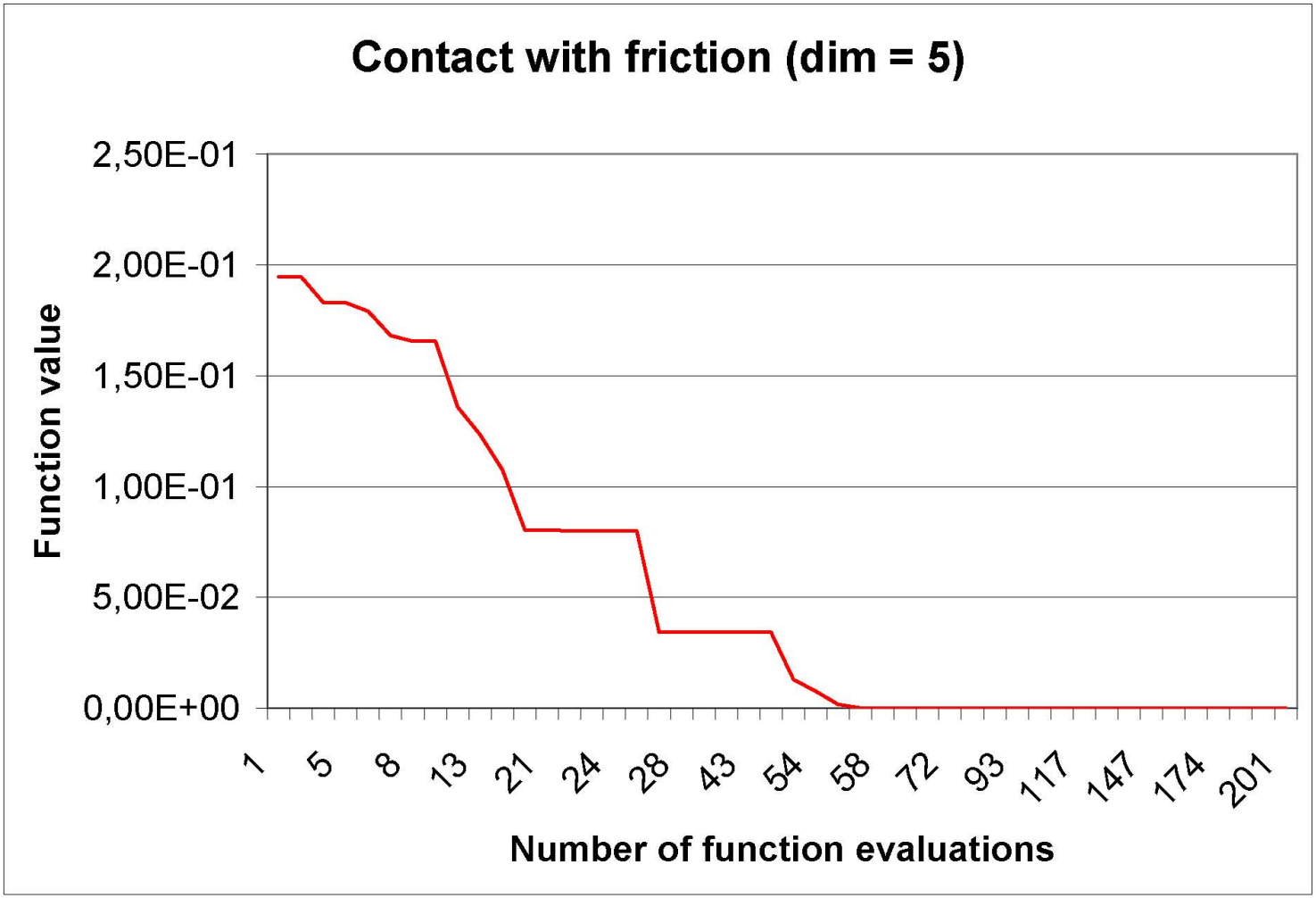
Why Derivative-Free Optimization?

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).

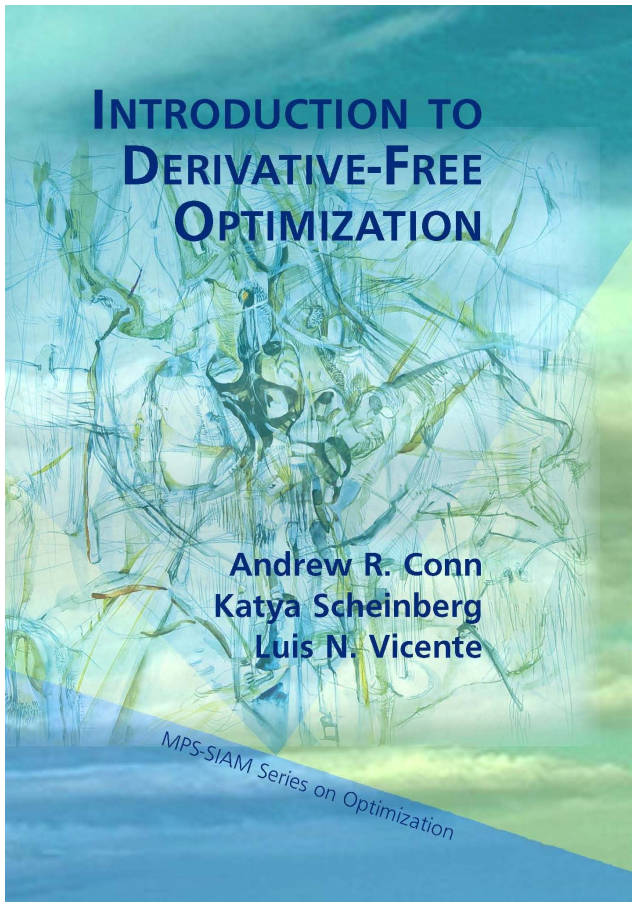
Limitations of Derivative-Free Optimization

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).

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:

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:

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

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

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:

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

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

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

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

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

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

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:

$$\bar{\phi} = \left\{ \frac{1}{2}y_1^2, ..., \frac{1}{2}y_n^2, y_1y_2, ..., y_{n-1}y_n, y_1, ..., y_n, 1 \right\}.$$

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

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

$$+ \frac{\partial^2 f}{\partial x_1^2}(0)[y_1^2/2] + \frac{\partial^2 f}{\partial x_1 x_2}(0)[y_1y_2] + \frac{\partial^2 f}{\partial x_2^2}(0)[y_2^2/2].$$

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.

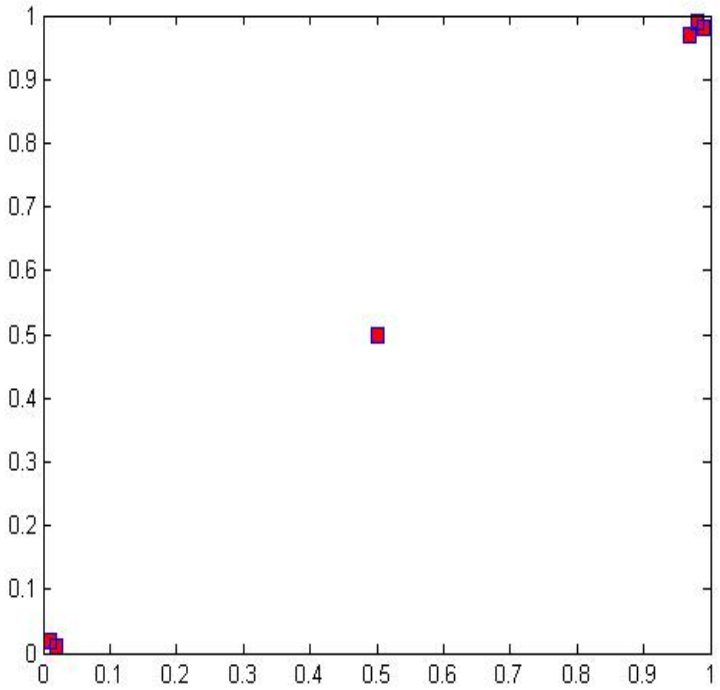
An equivalent definition of Λ–poisedness is ($|Y| = |\alpha|$)

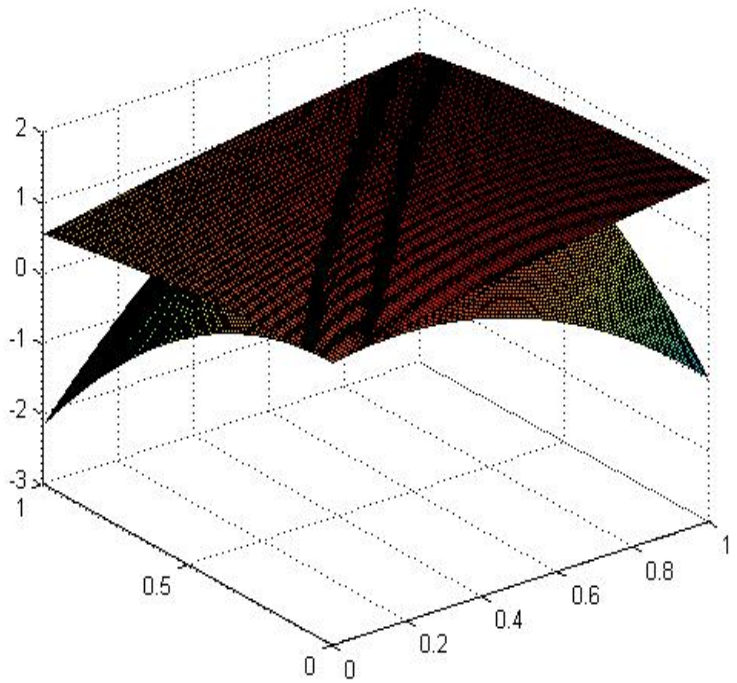
$$\|M(\bar{\phi}, Y_{scaled})^{-1}\| \leq \Lambda,$$

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

Non-squared cases are defined analogously (IDFO).

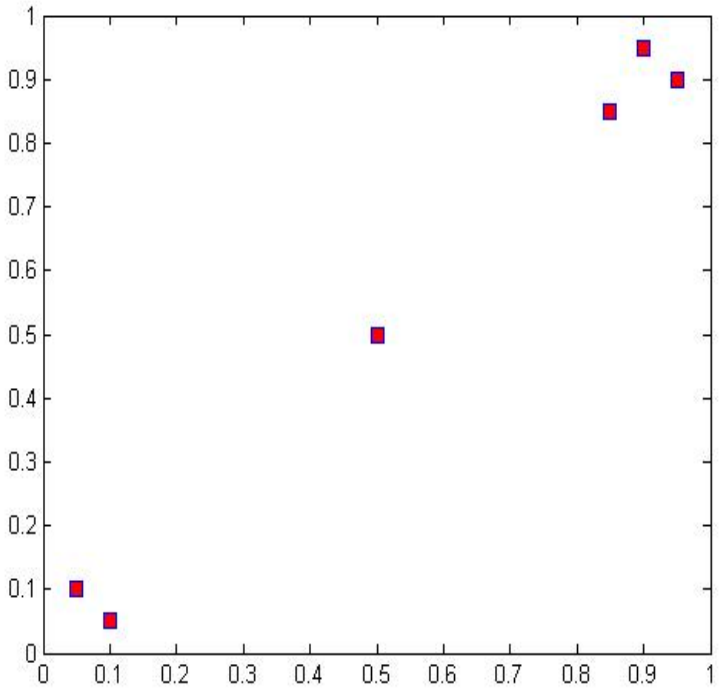
A badly poised set

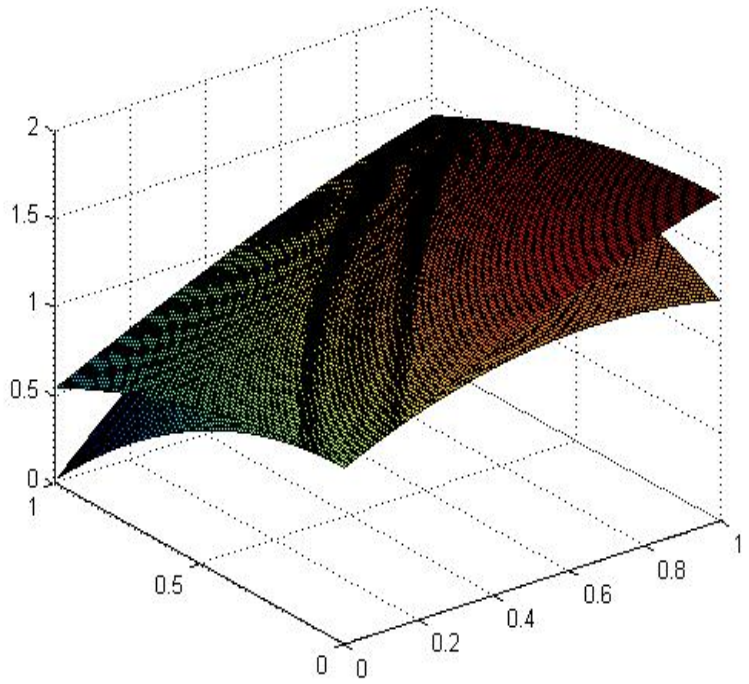




$$\Lambda = 5324.$$

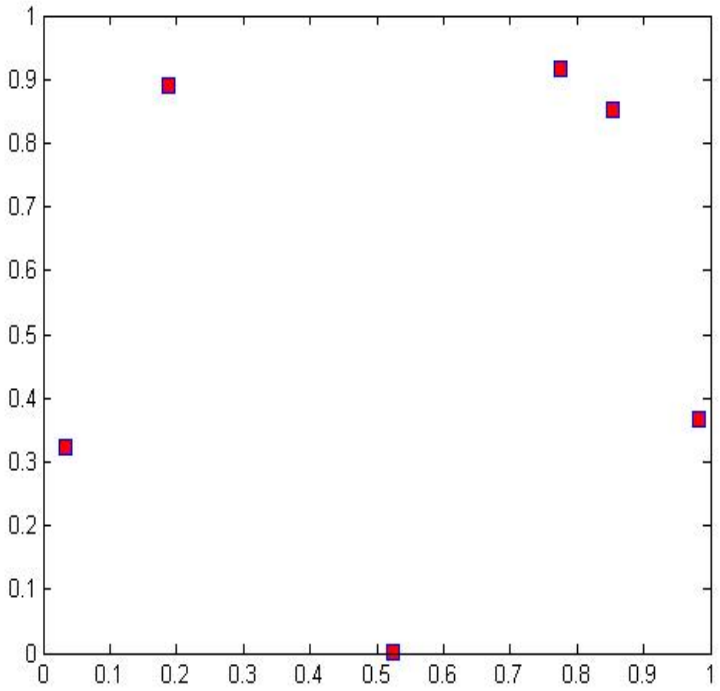
A not so badly poised set

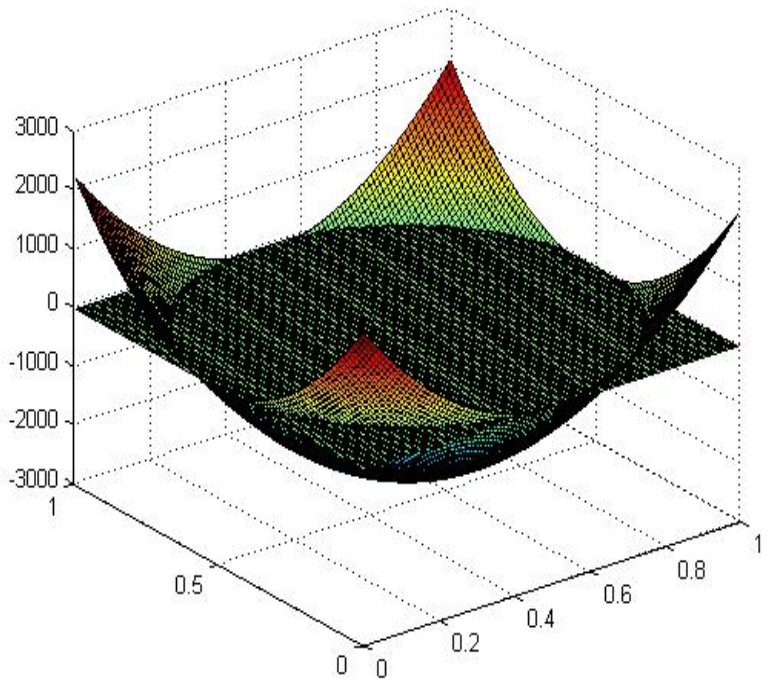




$$\Lambda = 295.$$

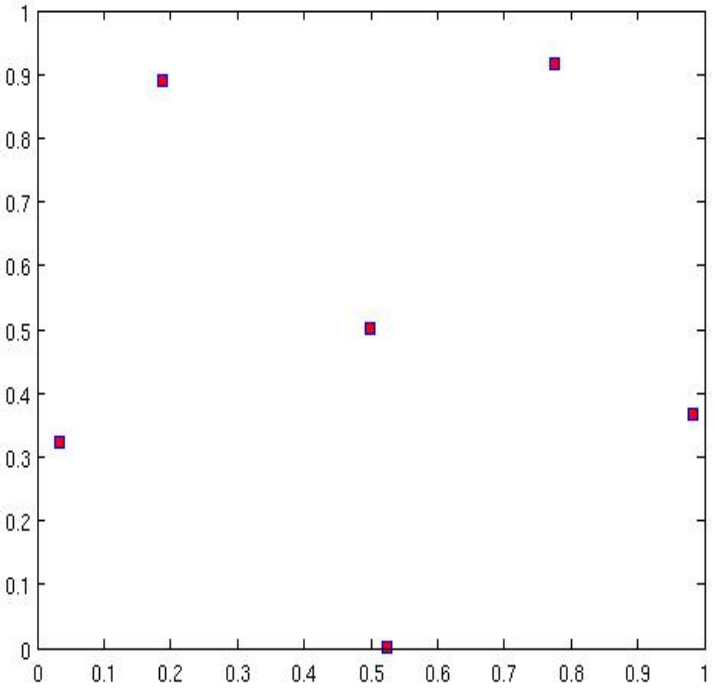
Another badly poised set

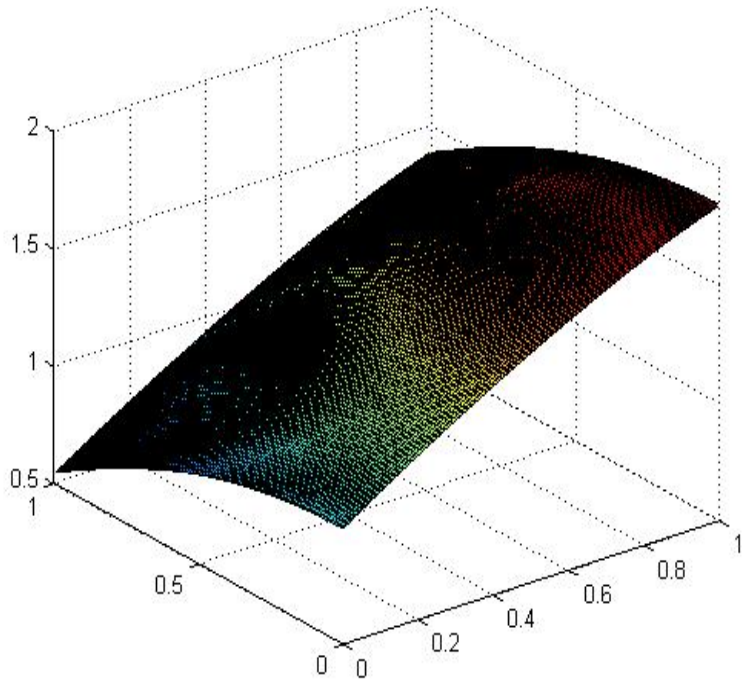




$$\Lambda = 492625.$$

An ideal set





$$\Lambda = 1.$$

Quadratic interpolation models

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

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 i.e. ($\bar{M}_L$ well conditioned in $M(\bar{\phi}, Y_{scaled}) = [\bar{M}_Q \ \bar{M}_L]$) then

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 i.e. ($\bar{M}_L$ well conditioned in $M(\bar{\phi}, Y_{scaled}) = [\bar{M}_Q \ \bar{M}_L]$) 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 i.e. ($\bar{M}_L$ well conditioned in $M(\bar{\phi}, Y_{scaled}) = [\bar{M}_Q \ \bar{M}_L]$) 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'):

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

Minimum Frobenius norm models

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

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

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

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

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

$$\begin{bmatrix} M_Q M_Q^\top & M_L \\ M_L^\top & 0 \end{bmatrix} \quad \text{where} \quad M(\bar{\phi}, Y) \;=\; \begin{bmatrix} M_Q & M_L \end{bmatrix}.$$

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:

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

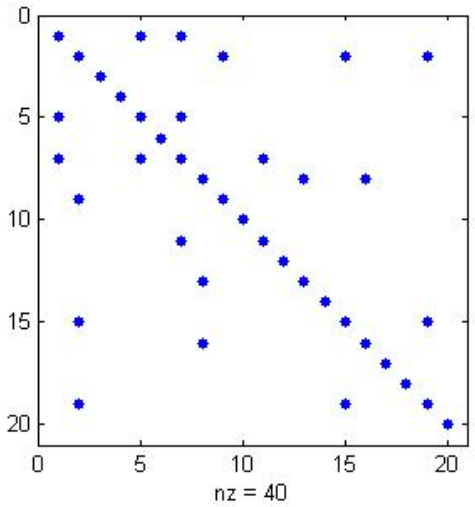
$\longrightarrow$ MFN models are fully linear.

Sparsity on the Hessian

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

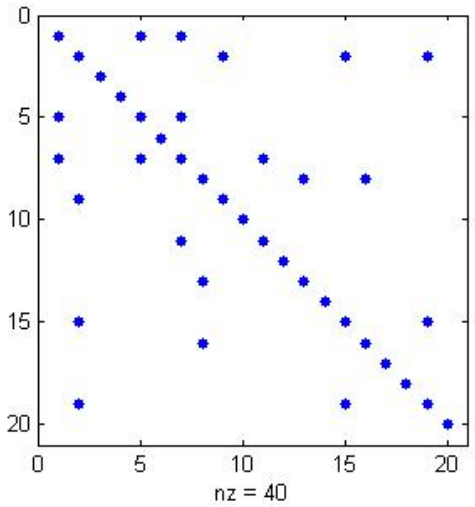
Sparsity on the Hessian

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



Sparsity on the Hessian

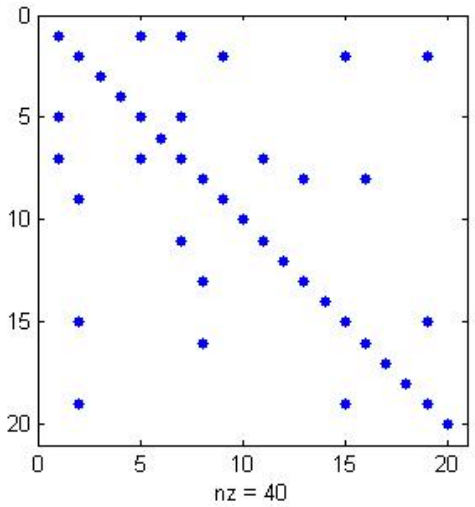
- In many problems, pairs of variables have no 'correlation', leading to zero second order partial derivatives in f:



- Thus, the Hessian $H = \nabla^2 m$ of the model should be sparse,

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 $H = \nabla^2 m$ of the model should be sparse,

  i.e., the vector α_Q in the basis $\bar{\phi}$ should be sparse (assuming $x = 0$ w.l.o.g.).

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

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

Compressed sensing — sparse recovery

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

Compressed sensing — sparse recovery

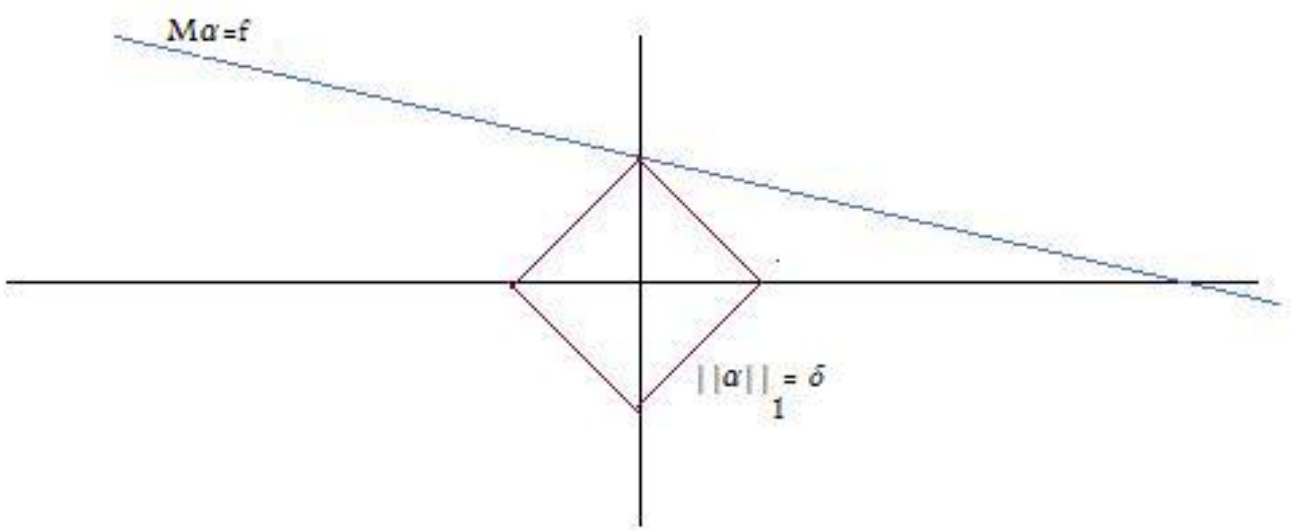
- Objective: Find sparse α subject to a highly underdetermined linear system $M\alpha = f$.
- $\begin{cases} \min & \|\alpha\|_0 = |\operatorname{supp}(\alpha)| \\ \text{s.t.} & M\alpha = f \end{cases}$ is NP-Hard.

Compressed sensing — sparse recovery

- Objective: Find sparse α subject to a highly underdetermined linear system $M\alpha = f$.
- $\begin{cases} \min & \|\alpha\|_0 = |\operatorname{supp}(\alpha)| \\ \text{s.t.} & M\alpha = f \end{cases}$ is NP-Hard.
- $\begin{cases} \min & \|\alpha\|_1 \\ \text{s.t.} & M\alpha = f \end{cases}$ often recovers sparse solutions.

Compressed sensing — sparse recovery

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

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.*

Compressed sensing — noisy PARTIALLY sparse recovery

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.
  - ...

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?

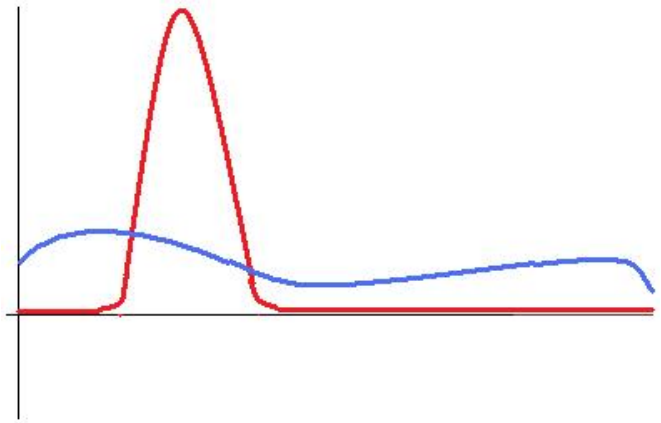
- Choose orthonormal bases (leads to uncorrelated matrix entries).

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.

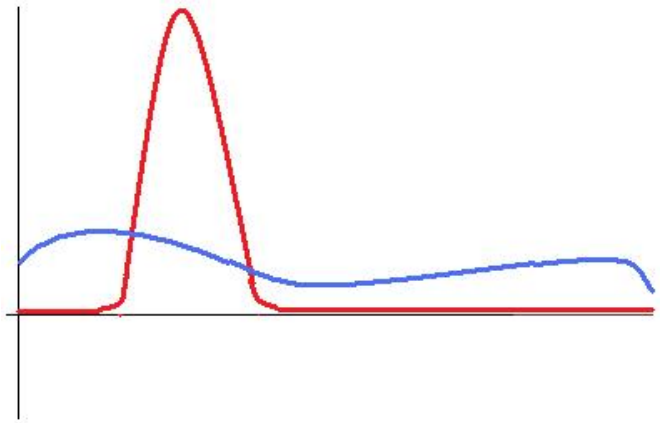


Bounded orthonormal expansions (Rauhut)

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* $\|\phi_i\|_{L^\infty} \leq K$.
- *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*

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*

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

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

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,

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

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 $\|\psi_\iota\|_{L^\infty} \leq 3$.

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

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

Proposition (Bandeira, Scheinberg, and Vicente, 2011)

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

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

$\longrightarrow \psi$ is very similar to the canonical basis, and thus "preserves" the sparsity of the Hessian.

Sparse Hessian recovery

Let us look again at

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

where

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

Sparse Hessian 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 $f = f(Y)$.

Sparse Hessian 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 $f = f(Y)$.

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

Sparse Hessian 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 $f = f(Y)$.

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

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

Sparse Hessian recovery

Theorem (Bandeira, Scheinberg, and Vicente, 2011)

If

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

Sparse Hessian 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* $B_\infty(0;\Delta)$.

Sparse Hessian 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* $B_\infty(0;\Delta)$.
- $\frac{p}{\log p} \geq 9c\,(h+n+1)\log^2(h+n+1)\log \mathcal{O}(n^2)$.

Sparse Hessian 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

Sparse Hessian 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

$$q^* = \sum \alpha_i^* \psi_i$$

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.

Remarks

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

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

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

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)

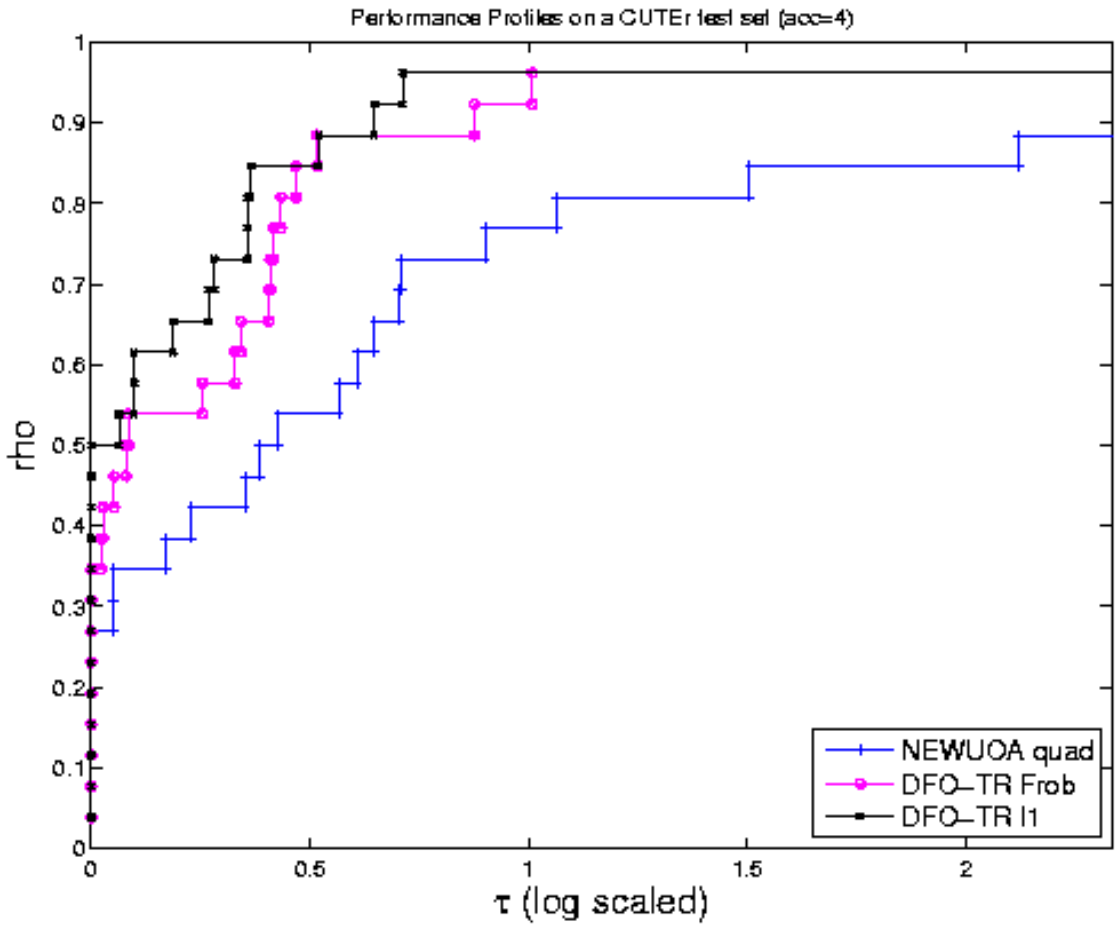


Figure: Performance profiles comparing `DFO-TR` (ℓ_1 and Frobenius) and `NEWUOA` (Powell) in a test set from `CUTEr` (Fasano et al.).

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

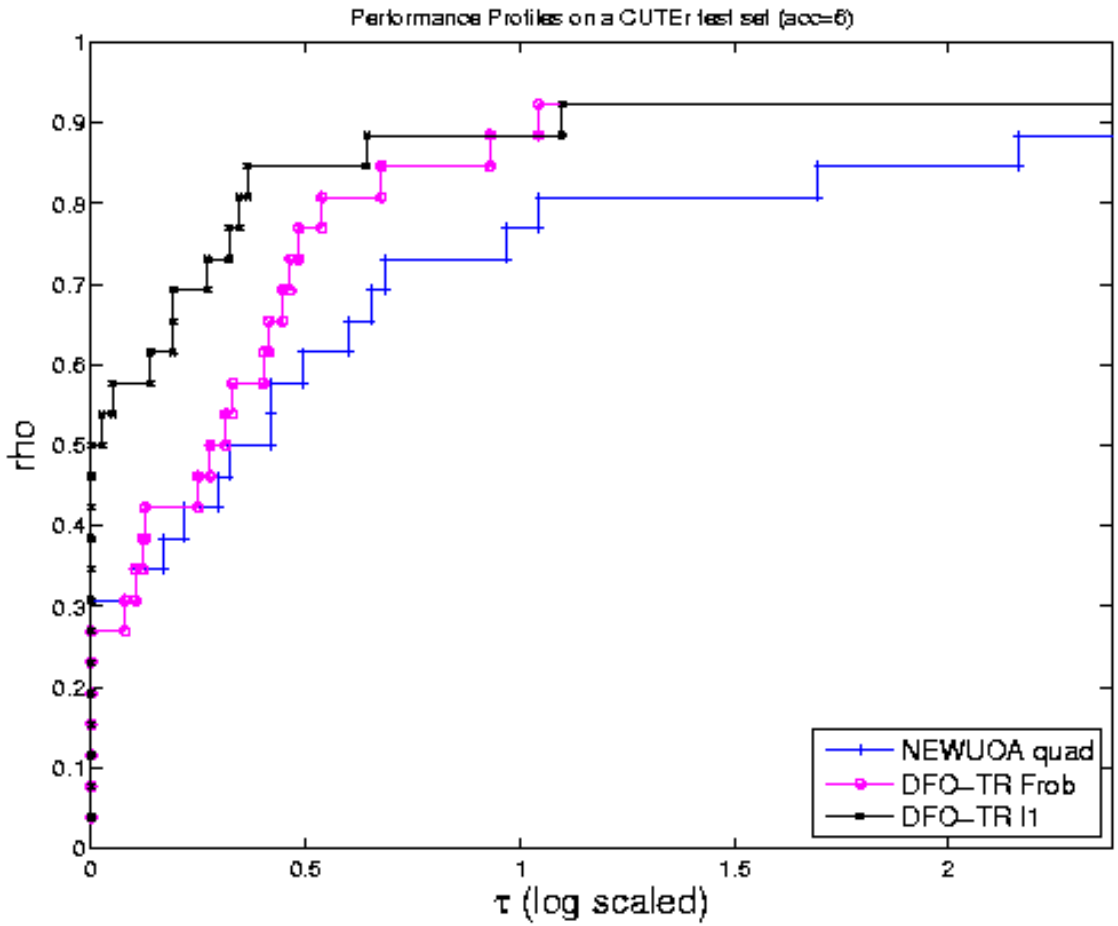


Figure: Performance profiles comparing `DFO-TR` (ℓ_1 and Frobenius) and `NEWUOA` (Powell) in a test set from `CUTEr` (Fasano et al.).

Concluding remarks (sparse Hessian recovery)

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

Concluding remarks (sparse Hessian recovery)

- 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 (sparse Hessian recovery)

- 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 trust-region methods based on probabilistic models.

TRM based on probabilistic models

Let models be built iteratively in some random fashion.

TRM based on probabilistic models

Let models be built iteratively in some random fashion.

We will consider random models M_k, and then use the notation $m_k = M_k(\omega_k)$ for their realizations.

TRM based on probabilistic models

Let models be built iteratively in some random fashion.

We will consider random models M_k, and then use the notation $m_k = M_k(\omega_k)$ for their realizations.

The key assumption for convergence will be then that these models exhibit good accuracy with sufficiently high probability.

TRM based on probabilistic models

Assumption

We say that a sequence of random models $\{M_k\}$ *is* (p)*-probabilistically* $(\kappa_{eg}, \kappa_{ef})$*-fully linear for a corresponding sequence* $\{B(X_k, \Delta_k)\}$ *if the events*

TRM based on probabilistic models

Assumption

We say that a sequence of random models $\{M_k\}$ *is* (p)*-probabilistically* $(\kappa_{eg}, \kappa_{ef})$*-fully linear for a corresponding sequence* $\{B(X_k, \Delta_k)\}$ *if the events*

$$S_k = \{M_k \text{ is a } (\kappa_{eg}, \kappa_{ef})\text{-fully linear model of } f \text{ on } B(X_k, \Delta_k)\}$$

TRM based on probabilistic models

Assumption

We say that a sequence of random models $\{M_k\}$ *is* (p)*-probabilistically* $(\kappa_{eg}, \kappa_{ef})$*-fully linear for a corresponding sequence* $\{B(X_k, \Delta_k)\}$ *if the events*

$$S_k \;=\; \{M_k \text{ is a } (\kappa_{eg}, \kappa_{ef})\text{-fully linear model of } f \text{ on } B(X_k, \Delta_k)\}$$

satisfy the following submartingale-like condition

$$P(S_k | \sigma(M_0, \ldots, M_{k-1})) \;\geq\; p. \qquad (implied\ by\ P(S_k) \geq p)$$

TRM based on probabilistic models

Assumption

We say that a sequence of random models $\{M_k\}$ *is* (p)*-probabilistically* $(\kappa_{eg}, \kappa_{ef})$*-fully linear for a corresponding sequence* $\{B(X_k, \Delta_k)\}$ *if the events*

$$S_k \;=\; \{M_k \text{ is a } (\kappa_{eg}, \kappa_{ef})\text{-fully linear model of } f \text{ on } B(X_k, \Delta_k)\}$$

satisfy the following submartingale-like condition

$$P(S_k | \sigma(M_0, \ldots, M_{k-1})) \;\geq\; p. \qquad (implied\ by\ P(S_k) \geq p)$$

Furthermore, if $p \geq \frac{1}{2}$*, then we say that the random models are probabilistically* $(\kappa_{eg}, \kappa_{ef})$*-fully linear.*

TRM based on probabilistic models

Fix three positive parameters η_1, η_2, γ, with $\gamma > 1$.

TRM based on probabilistic models

Fix three positive parameters η_1, η_2, γ, with $\gamma > 1$.

Algorithm (Iteration k)

Approximate the function f in $B(x_k, \delta_k)$ with m_k.

TRM based on probabilistic models

Fix three positive parameters η_1, η_2, γ, with $\gamma > 1$.

Algorithm (Iteration k)

Approximate the function f in $B(x_k, \delta_k)$ with m_k.

Compute a step s_k by solving $\min_{s \in B(0,\delta_k)} m(x_k + s)$.

TRM based on probabilistic models

Fix three positive parameters η_1, η_2, γ, with $\gamma > 1$.

Algorithm (Iteration k)

Approximate the function f in $B(x_k, \delta_k)$ with m_k.

Compute a step s_k by solving $\min_{s \in B(0,\delta_k)} m(x_k + s)$.

Let

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

TRM based on probabilistic models

Fix three positive parameters η_1, η_2, γ, with $\gamma > 1$.

Algorithm (Iteration k)

Approximate the function f in $B(x_k, \delta_k)$ with m_k.

Compute a step s_k by solving $\min_{s \in B(0,\delta_k)} m(x_k + s)$.

Let

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

If $\rho_k \geq \eta_1$ and $\|g_k\| \geq \eta_2 \delta_k$,

TRM based on probabilistic models

Fix three positive parameters η_1, η_2, γ, with $\gamma > 1$.

Algorithm (Iteration k)

Approximate the function f in $B(x_k, \delta_k)$ with m_k.

Compute a step s_k by solving $\min_{s \in B(0,\delta_k)} m(x_k + s)$.

Let

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

If $\rho_k \geq \eta_1$ and $\|g_k\| \geq \eta_2 \delta_k$, set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \gamma \delta_k$.

TRM based on probabilistic models

Fix three positive parameters η_1, η_2, γ, with $\gamma > 1$.

Algorithm (Iteration k)

Approximate the function f in $B(x_k, \delta_k)$ with m_k.

Compute a step s_k by solving $\min_{s \in B(0,\delta_k)} m(x_k + s)$.

Let

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

If $\rho_k \geq \eta_1$ and $\|g_k\| \geq \eta_2 \delta_k$, set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \gamma \delta_k$.

Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \gamma^{-1} \delta_k$.

Lemma

For every realization of the algorithm,

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

TRM based on probabilistic models

Theorem

Suppose that the model sequence $\{M_k\}$ *is probabilistically* $(\kappa_{eg}, \kappa_{ef})$*-fully linear for some* $\kappa_{eg}, \kappa_{ef} > 0$.

TRM based on probabilistic models

Theorem

Suppose that the model sequence $\{M_k\}$ is probabilistically $(\kappa_{eg}, \kappa_{ef})$-fully linear for some $\kappa_{eg}, \kappa_{ef} > 0$.

Let $\{X_k\}$ be a sequence of random iterates generated by the algorithm.

TRM based on probabilistic models

Theorem

Suppose that the model sequence $\{M_k\}$ *is probabilistically* $(\kappa_{eg}, \kappa_{ef})$*-fully linear for some* $\kappa_{eg}, \kappa_{ef} > 0$.

Let $\{X_k\}$ *be a sequence of random iterates generated by the algorithm.*

Then, almost surely,

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

TRM based on probabilistic models

It is also possible to prove a.s. convergence to second-order critical points.

References

- A. S. Bandeira, K. Scheinberg, and L. N. Vicente, Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization, to appear in Mathematical Programming.

- A. S. Bandeira, K. Scheinberg, and L. N. Vicente, On partially sparse recovery, 2011.

- A. S. Bandeira, K. Scheinberg, and L. N. Vicente, Trust-region methods based on probabilistic models, in preparation.