Recovering structure and using random models in derivative free optimization

Katya Scheinberg

(joint work with A. Bandeira and L.N. Vicente)

Derivative free optimization

- Unconstrained optimization problem

$$\min_{x \in \Omega} f(x)$$

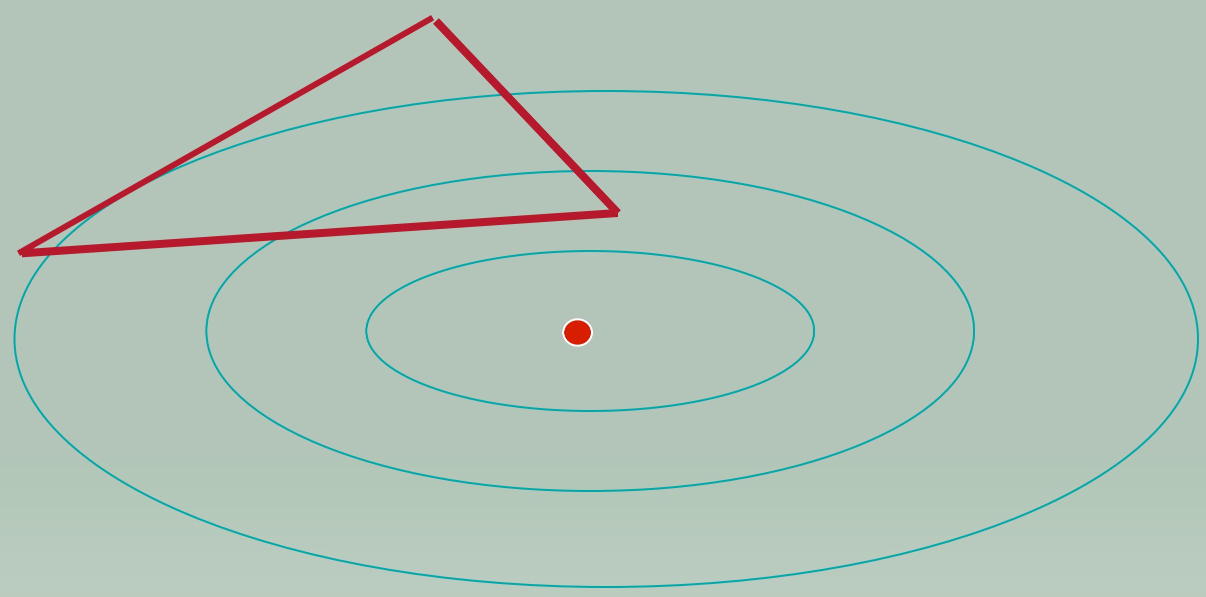
- Function $f \in C^2$ is a result of a black box computation. It is expensive to compute and no derivative information is available.
- Numerical noise is often present.

Main idea

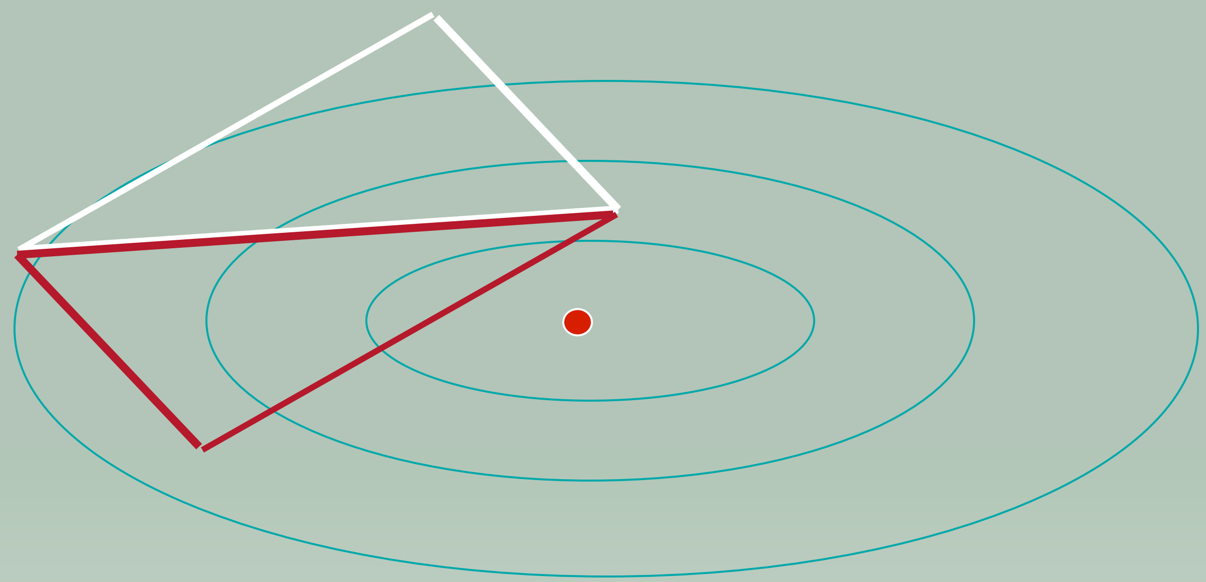
- It is common in optimization to exploit structure of the objective function to improve efficiency of the methods.
- Often structure manifests itself in the sparsity of the Hessian.
- In DFO we do not know sparsity structure, but it does not mean the structure is not there.
- With recent advances in sparse structure recovery (in particular compressed sensing) we can hope to exploit the latent structures in black box optimization.
- This requires a use of randomly sampled models
- Need new convergence theory

Algorithms

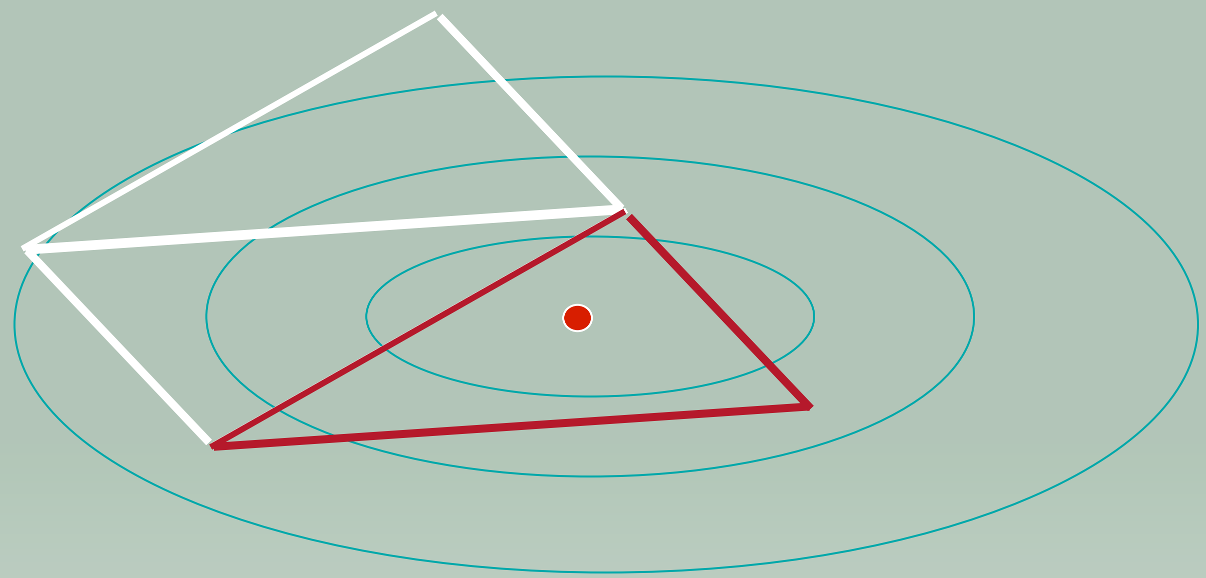
Nelder-Mead method (1965)



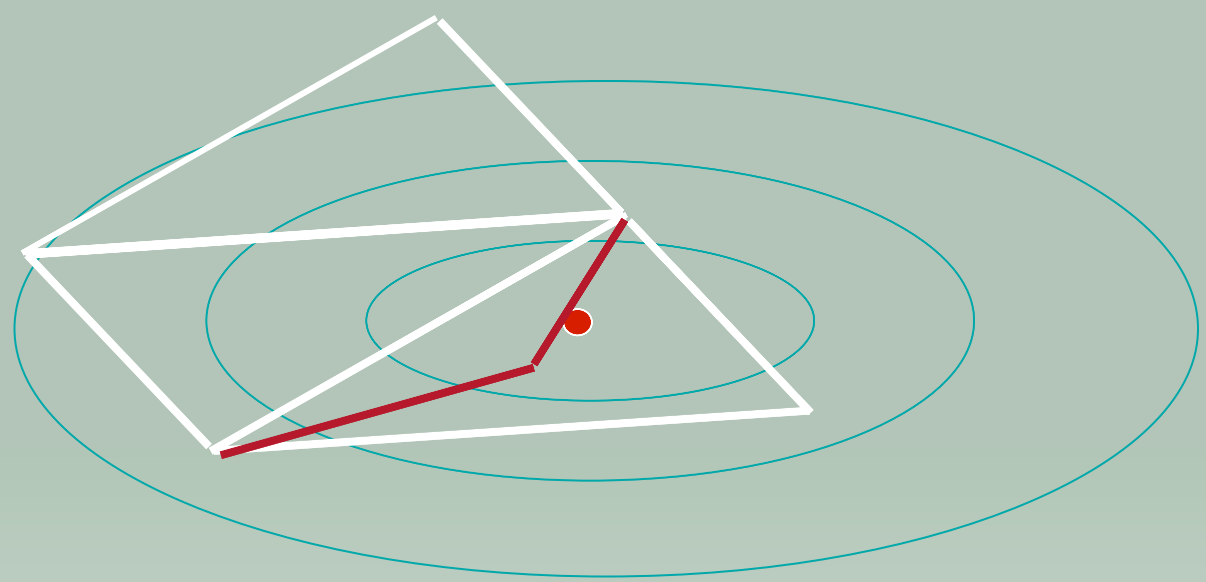
Nelder-Mead method (1965)



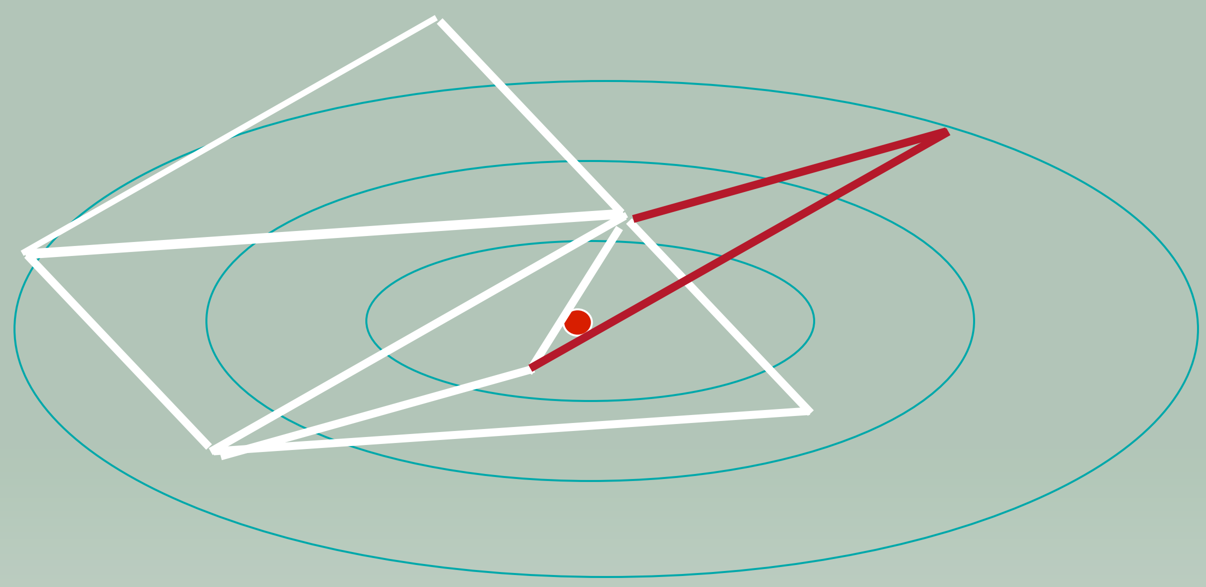
Nelder-Mead method (1965)



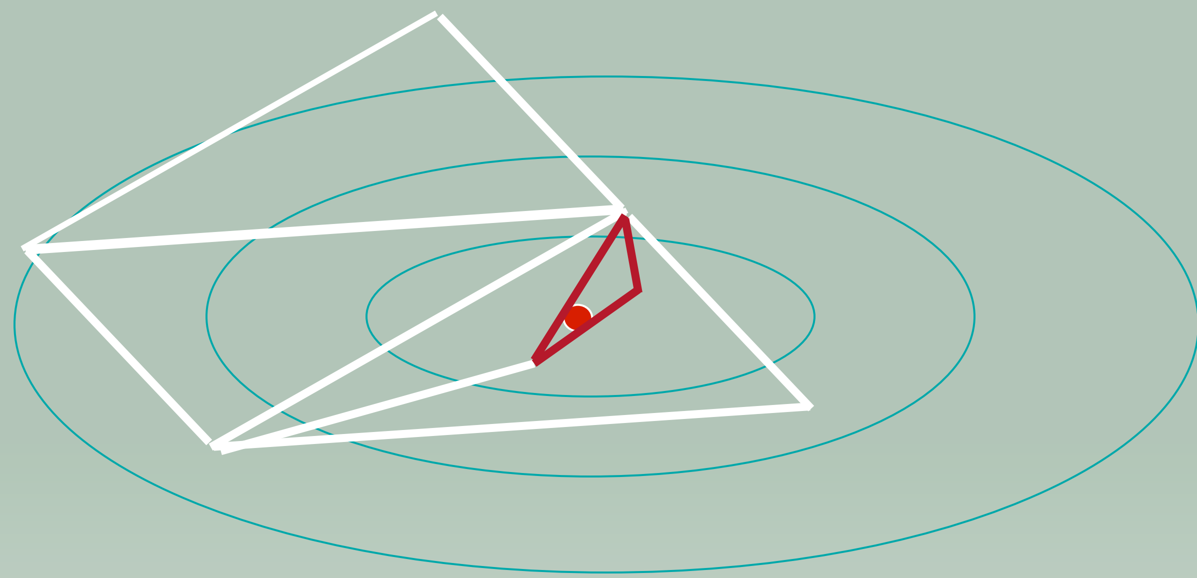
Nelder-Mead method (1965)



Nelder-Mead method (1965)

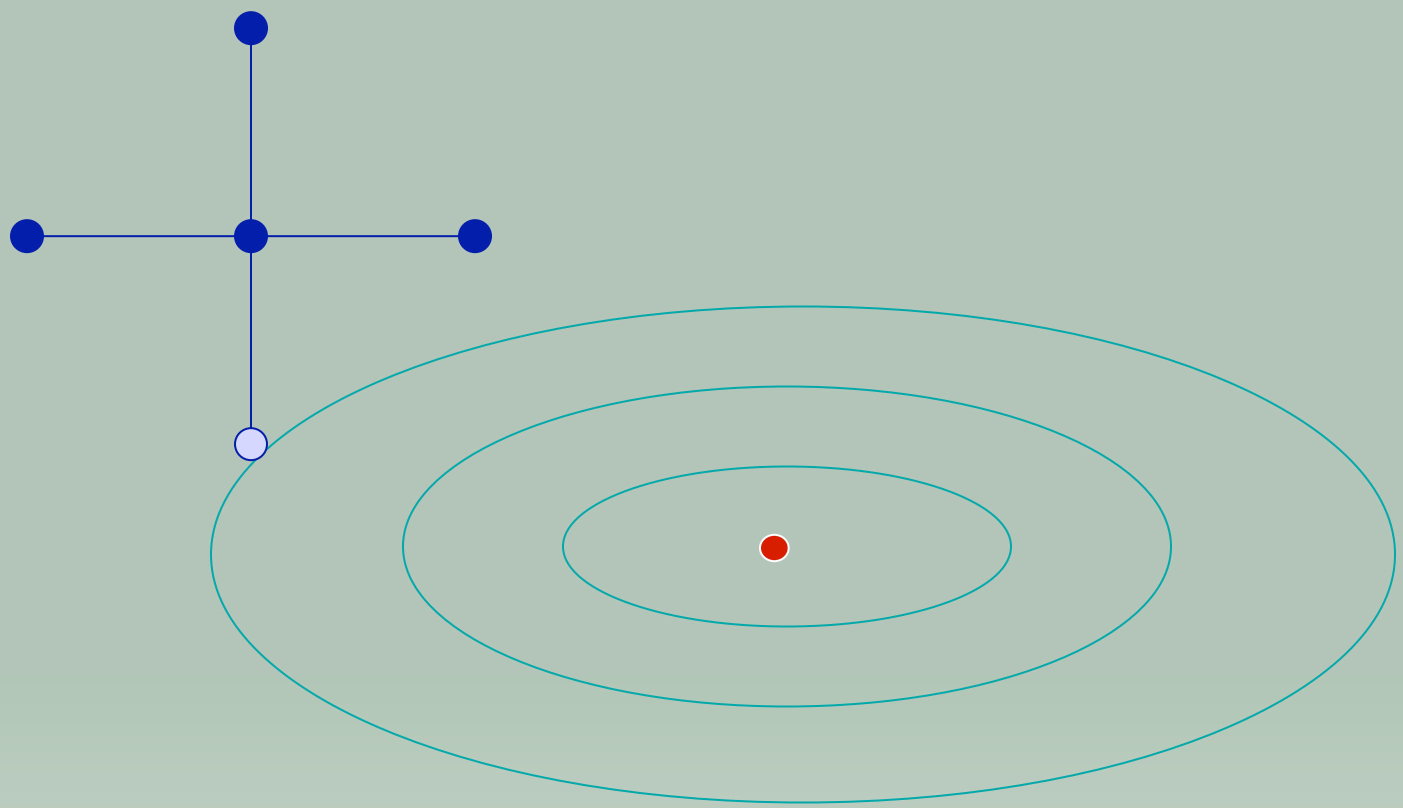


Nelder-Mead method (1965)

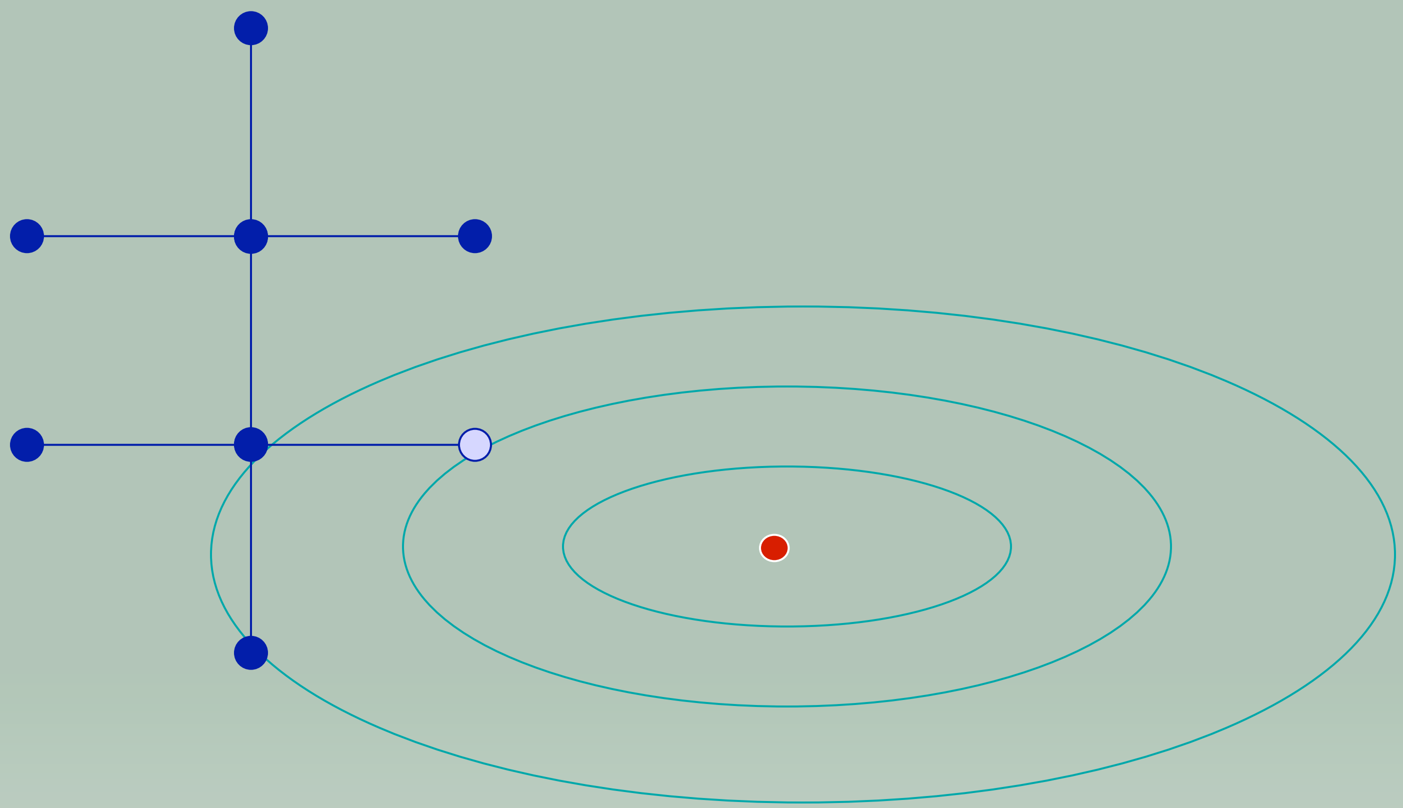


The simplex changes shape during the algorithm to adapt to curvature. But the shape can deteriorate and NM gets stuck

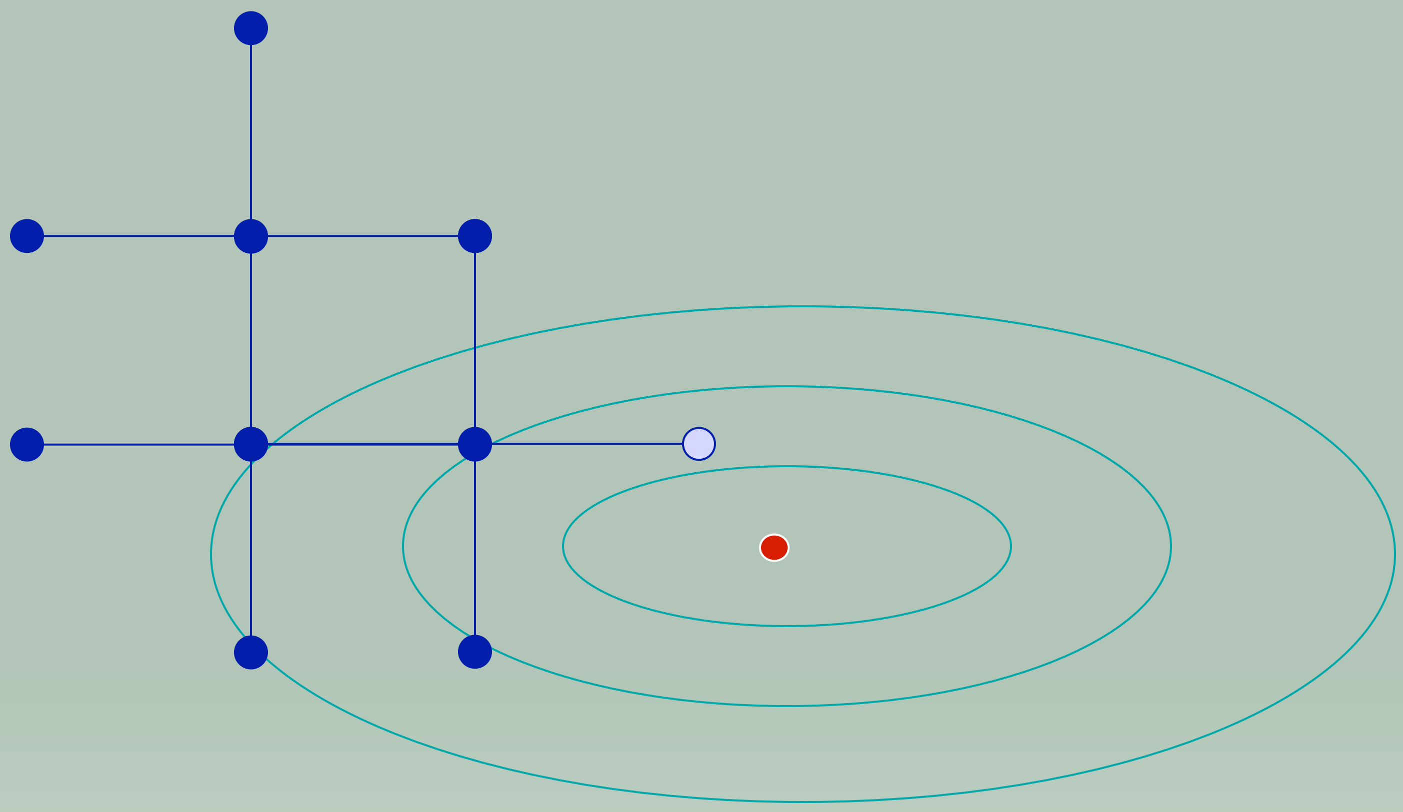
Direct Search methods (early 1990s)



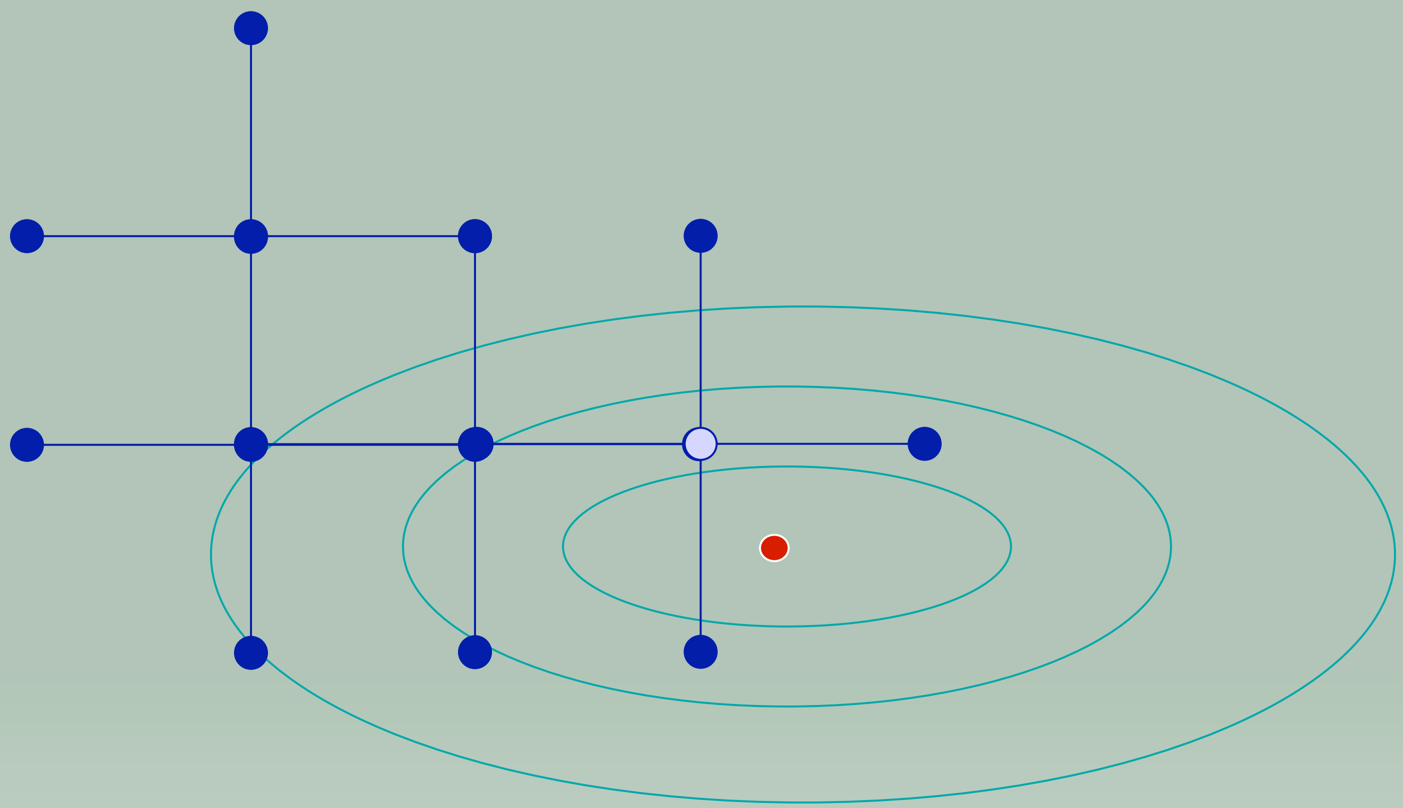
Direct Search methods



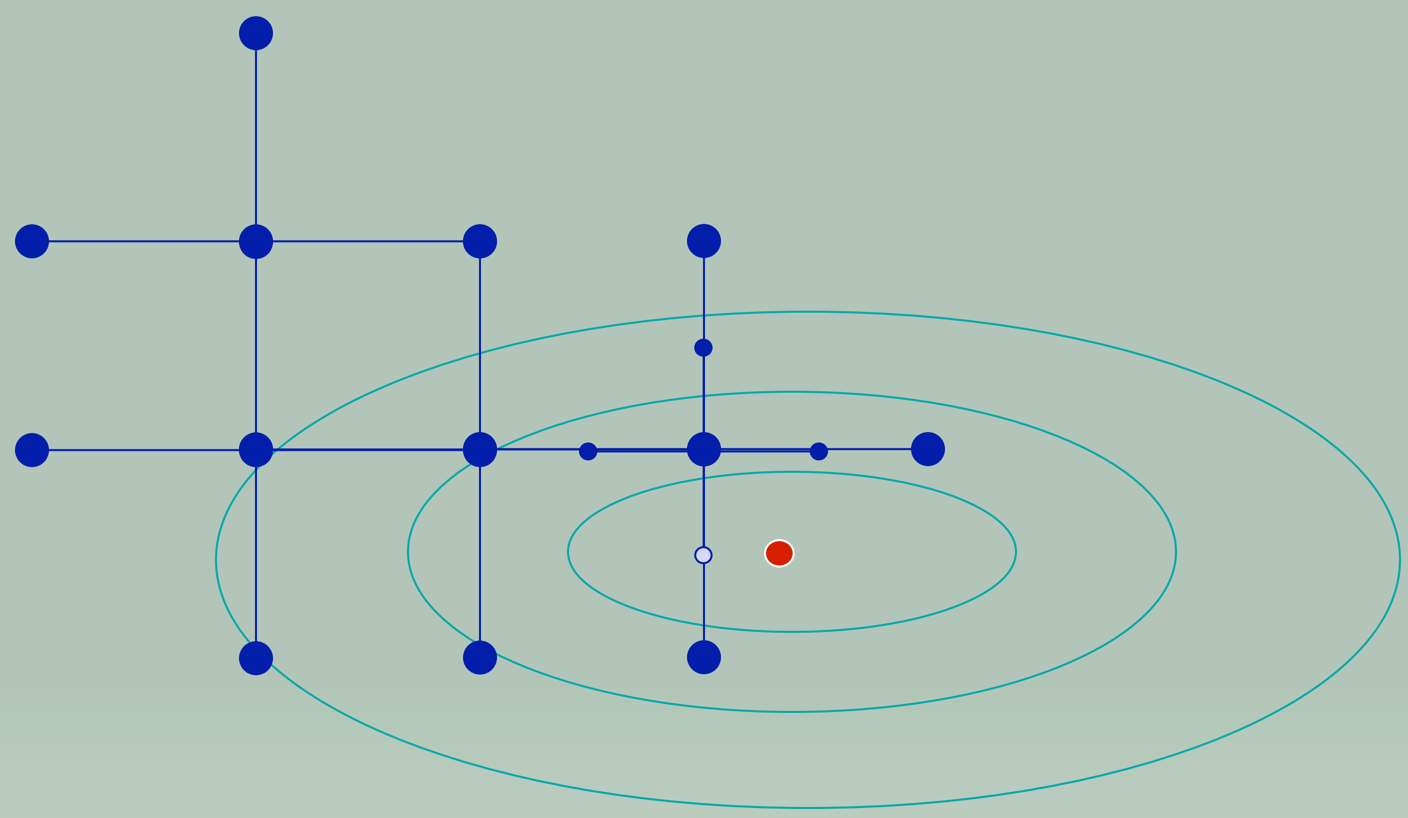
Direct Search methods



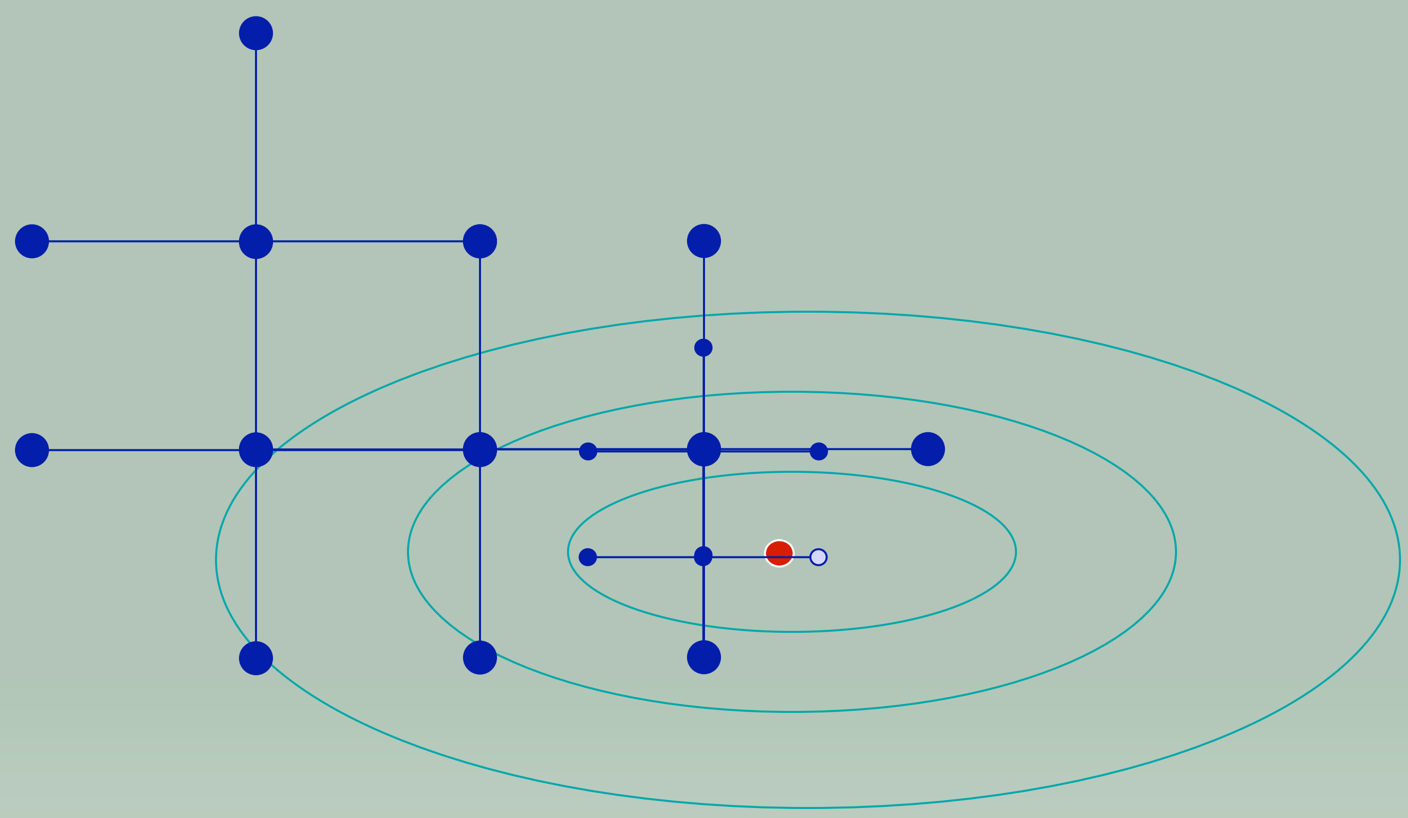
Direct Search method



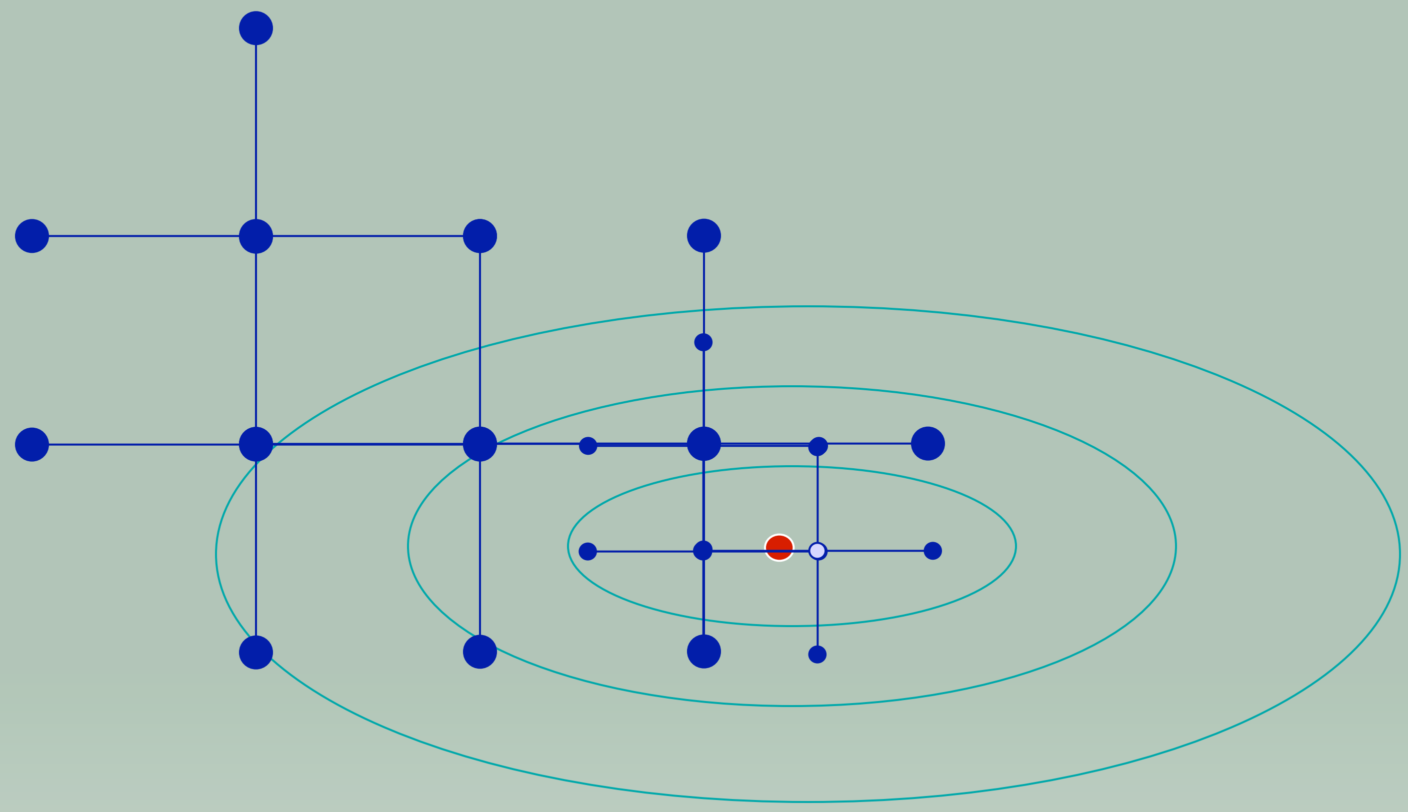
Direct Search method



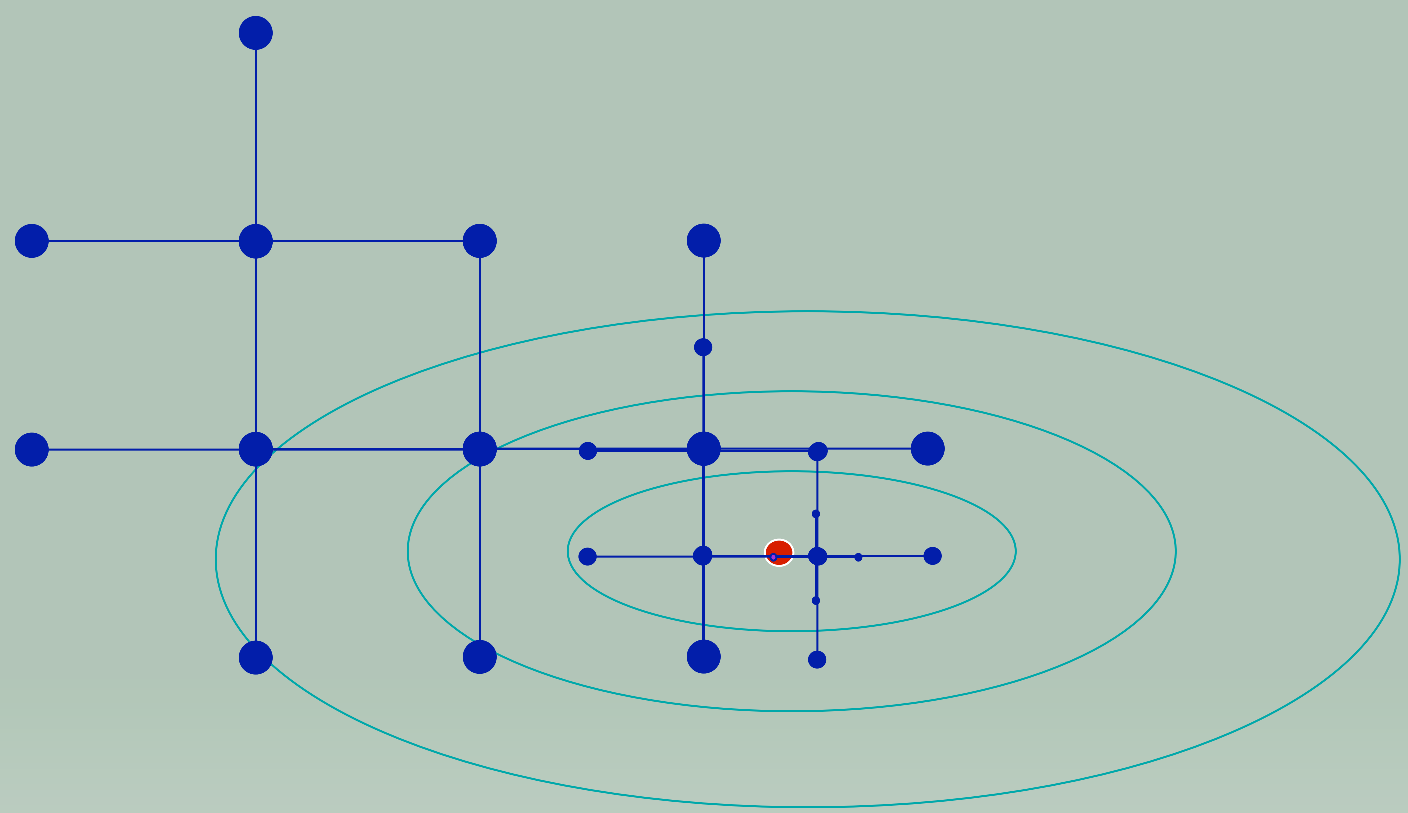
Direct Search method



Direct Search method

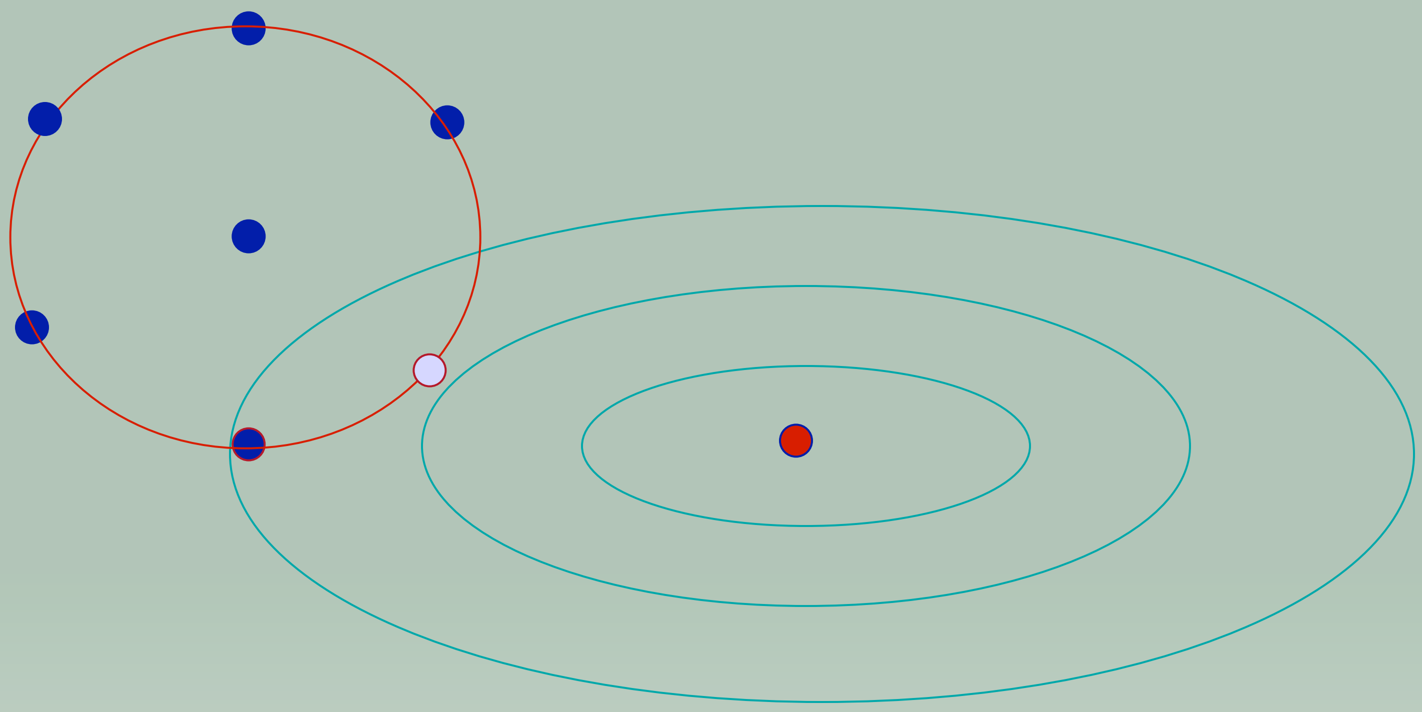


Direct Search method

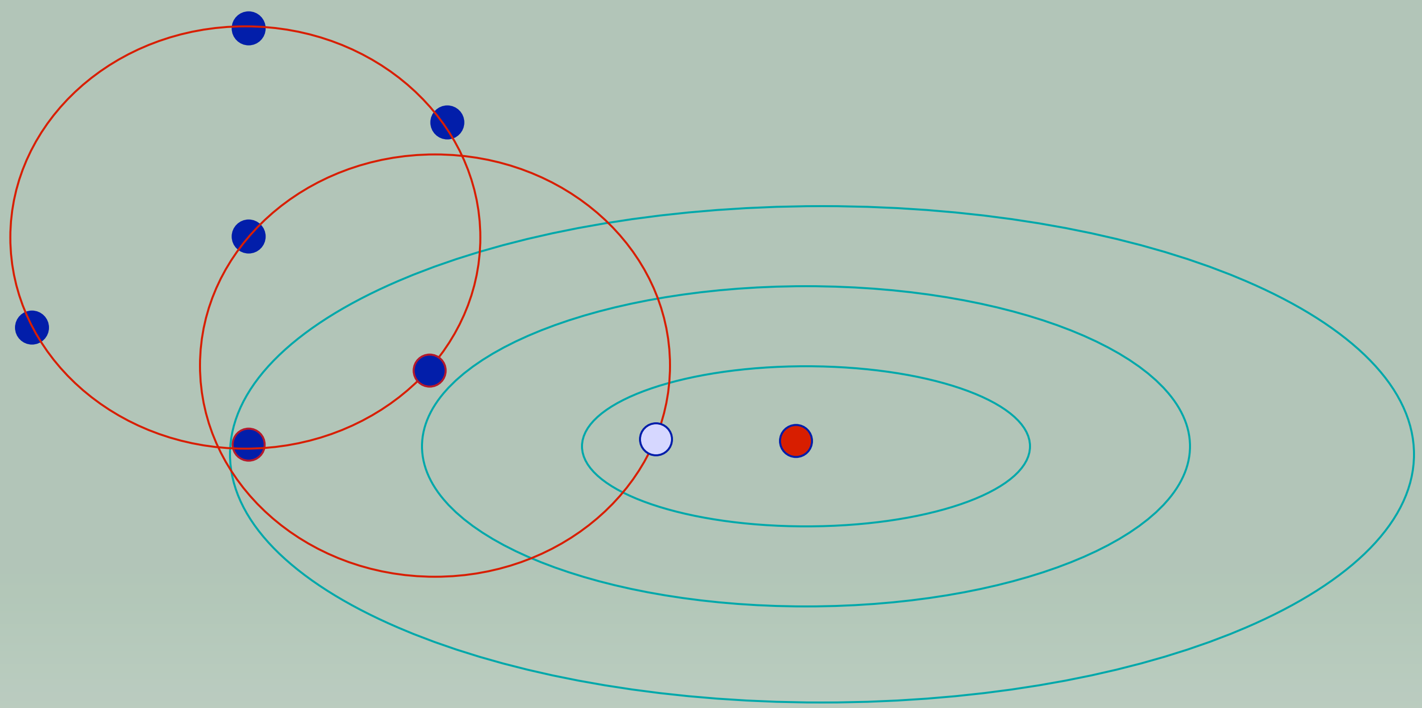


Fixed pattern, never deteriorates: theoretically convergent, but slow

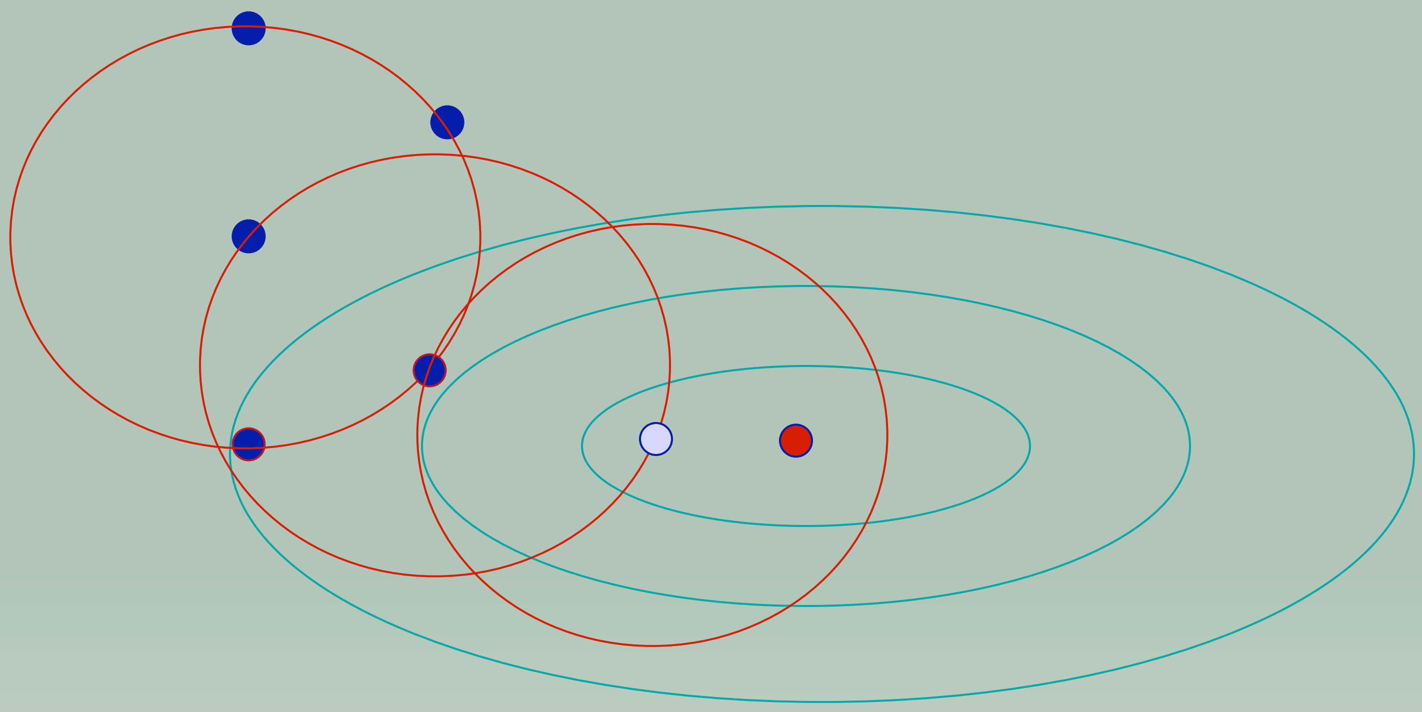
Model based trust region methods (late 90s)



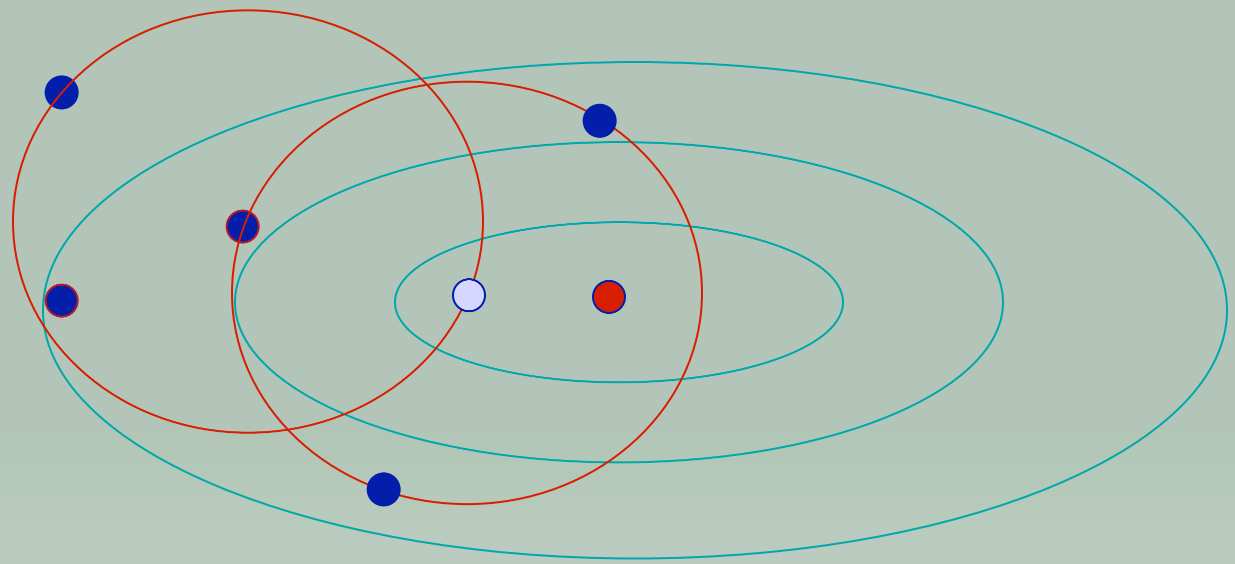
Model based trust region methods



Model based trust region methods



Model Based trust region methods



Exploits curvature, flexible efficient steps, uses second order models.

What do we want?

- Get as much curvature information as possible.
- Economize on function evaluations.
- Have models which we can optimize (i.e quadratic for now).

Basic Trust Region Algorithm

Initialize: Choose a class of models, initialize x_0, $m_0(x)$, Δ_0. Choose $\eta > 0$ and other parameters.

Criticality step: If $\|g_k(x_k)\| \leq \epsilon_c$ then make sure we have a good model in $B(x_k, \rho_k)$ for some $\rho_k \leq \mu \|g_k(x_k)\|$.

Compute Step: Compute s_k from $\min_{\|s\| \leq \Delta_k} m_k(x_k + s)$ evaluate $f(x_k + s_k)$ and $r_k = (f(x_k) - f(x_k + s_k))/(m(x_k) - m(x_k + s_k))$.

Accept step: If $r_k \geq \eta$ then $x_{k+1} = x_k + s_k$.

TR Update: If $r_k < \eta_1$ and the model is good, decrease Δ. If $r_k < \eta_1$ and the model is not good, improve the model. Otherwise may increase Δ_k.

What is a "good" model?

We need Taylor-like behavior of first or second order models

A model is called fully linear in $B(x,\Delta)$ if

$$\|\nabla f(x+s) - \nabla m(x+s)\| \;\leq\; \kappa_{eg}\,\Delta, \quad \forall s \in B(0;\Delta),$$

$$|f(x+s) - m(x+s)| \;\leq\; \kappa_{ef}\,\Delta^2, \quad \forall s \in B(0;\Delta),$$

for some fixed κ_{eg} and κ_{ef} independent of x and Δ.

What is a "better" model?

We need Taylor-like behavior of first or second order models

A model is called fully quadratic in $B(x, \Delta)$ if

$$\begin{aligned}
\|\nabla^2 f(x+s) - \nabla^2 m(x+s)\| &\leq \kappa_{eh}\, \Delta, \quad \forall s \in B(0;\Delta),\\
\|\nabla f(x+s) - \nabla m(x+s)\| &\leq \kappa_{eg}\, \Delta^2, \quad \forall s \in B(0;\Delta),\\
|f(x+s) - m(x+s)| &\leq \kappa_{ef}\, \Delta^3, \quad \forall s \in B(0;\Delta),
\end{aligned}$$

for some fixed κ_{eh}, κ_{eg}, κ_{ef} independent of x and Δ.

Convergence results

Fully linear models – "first order methods" and convergence to a stationary point

Fully quadratic models - "second order methods" and convergence to the local minimum

Conn, S. and Vicente, 2008.

Polynomial models

Polynomial Interpolation

Given a polynomial basis $\phi = (\phi_1(x), \ldots, \phi_q(x))$ any polynomial $m(x)$ is expressed as

$$m(x) \; = \; \sum_{k=1}^{q} \alpha_k \phi_k(x)$$

Given an interpolation set $Y = \{y^1, \ldots, y^p\}$ the interpolation conditions are

$$m(y^i) \; = \; \sum_{k=1}^{q} \alpha_k \phi_k(y^i) = f(y^i) \quad \forall i = 1, \ldots, p.$$

The coefficient matrix of the system is:

$$M(\phi, Y) = \begin{bmatrix} \phi_1(y^1) & \phi_2(y^1) & \cdots & \phi_q(y^1) \\ \phi_1(y^2) & \phi_2(y^2) & \cdots & \phi_q(y^2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1(y^p) & \phi_2(y^p) & \cdots & \phi_q(y^p) \end{bmatrix} \qquad (p = q).$$

Special case - monomial quadratic basis

Specifically for $\bar{\phi} = \{1, x_1, \cdots, x_n, \frac{1}{2}x_1^2, x_1x_2, \cdots, \frac{1}{2}x_n^2\}$

$$M(\bar{\phi}, Y) = M = \begin{bmatrix} 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{2}(y_n^1)^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{2}(y_n^p)^2 \end{bmatrix}$$

Interpolation model:

$$\text{find } \alpha : \ M\alpha = f(Y)$$

$$m(x) = \sum_{i=1}^{q} \alpha_i \bar{\phi}_i(x) = \frac{1}{2} x^\top H x + g^\top x + \kappa$$

- $\kappa = \alpha_1$
- $g = (\alpha_2, \ldots, \alpha_{n+1})$
- $H_{ij} = \alpha_{n+(i-1)*n+j+1}$

Fully quadratic model

$$M(\bar{\phi}, Y) = M = \begin{bmatrix} 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{2}(y_n^1)^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{2}(y_n^p)^2 \end{bmatrix}$$

Need p=(n+1)(n+2)/2 interpolation points!!!

Interpolation model:

$$\text{find } \alpha : \quad M\alpha = f(Y)$$

$$m(x) = \sum_{i=1}^{q} \alpha_i \bar{\phi}_i(x) = \frac{1}{2} x^\top H x + g^\top x + \kappa$$

- $\kappa = \alpha_1$
- $g = (\alpha_2, \ldots, \alpha_{n+1})$
- $H_{ij} = \alpha_{n+(i-1)*n+j+1}$

Underdetermined quadratic model

$$M(\bar{\phi}, Y) = M = \begin{bmatrix} 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{2}(y_n^1)^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{2}(y_n^p)^2 \end{bmatrix}$$

Consider p<(n+1)(n+2)/2 interpolation points!!!

Interpolation model:

$$\text{find } \alpha : \; M\alpha = f(Y)$$

Interpolation model is not unique – many choices, which to pick?

Regularized quadratic models

$$M(\bar{\phi}, Y) = M = \begin{bmatrix} 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{2}(y_n^1)^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{2}(y_n^p)^2 \end{bmatrix}$$

p<(n+1)(n+2)/2 – underdetermined system

"Robust" interpolation model

$$\min_{\alpha} \quad \|\alpha\|_2$$

$$\text{s.t.} \quad M\alpha = f(Y)$$

$$m(x) = \frac{1}{2}x^\top H x + g^\top x + \kappa$$

- $\kappa = \alpha_1$
- $g = (\alpha_2, \ldots, \alpha_{n+1})$
- $H_{ij} = \alpha_{n+(i-1)*n+j+1}$

Minimum Frobenius Norm models

$$M(\bar{\phi}, Y) = M = \left[\overbrace{\begin{matrix} 1 & y_1^1 & \cdots & y_n^1 \\ \vdots & \vdots & & \vdots \\ 1 & y_1^p & \cdots & y_n^p \end{matrix}}^{M_L} \quad \overbrace{\begin{matrix} \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{2}(y_n^1)^2 \\ \vdots & \vdots & & \vdots \\ \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{2}(y_n^p)^2 \end{matrix}}^{M_Q} \right]$$

Minimum Frob norm of the Hessian model

$$\begin{array}{ll} \min\limits_{\alpha} & \|\alpha_Q\|_2 \\ \text{s.t.} & M_L \alpha_L + M_Q \alpha_Q = f(Y) \end{array}$$

$$m(x) = \tfrac{1}{2} x^\top H x + g^\top x + \kappa$$

- $\alpha_L \to (k, g)$
- $\alpha_Q \to H$

Convergence result for MFN models

Minimum Frobenius norm quadratic models are fully linear under appropriate conditions, hence can guarantee convergence to the stationary point.

Conn, S. and Vicente, 2008.

Usefulness and limitation

In practice using MFN quadratic models is by far superior to using fully quadratic models, since good second order information can be recovered from just a few extra interpolation points.

In theory MFN quadratic models have not been shown to be better than linear models, unless $p=(n+1)(n+1)/2$.

Question: can we consistently build fully quadratic interpolation models with $p<(n+1)(n+1)/2$ points?

Example

$$\min f(x) = \sum_{i}^{n}((x_i^2 - x_n^2)^2 - 4x_i)$$

$$\nabla^2_{ij} f(x) = 0, \;\; \forall i \neq j, j \neq n$$

$$(f(y^0), \nabla_{ij} f(y^0)) \to \alpha_L \quad \nabla^2 f(y^0) \to \alpha_Q$$

$$m(x) = (\bar{\phi}_L^\top, \bar{\phi}_Q^\top) \begin{pmatrix} \alpha_L \\ \alpha_Q \end{pmatrix} \longleftarrow$$

Taylor model

(α_L, α_Q) has only 2n+n nonzeros

3n points are enough to recover the fully quadratic model

Colson, Toint, 2004

But usually we do not the sparsity structure of the Hessian. Moreover, it may depend of the region of local approximation…

We want to recover the sparse model by using few sample points

Sounds familiar? – use compressed sensing ideas!

Minimum Frobenius Norm models

$$M(\bar{\phi}, Y) = M = \begin{array}{cc} \overbrace{}^{M_L} & \overbrace{\phantom{\tfrac{1}{2}(y_1^1)^2 \quad y_1^1 y_2^1 \quad \cdots \quad \tfrac{1}{2}(y_n^1)^2}}^{M_Q} \end{array}$$

$$\left[\begin{array}{cccccccc} 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{2}(y_n^1)^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{2}(y_n^p)^2 \end{array} \right]$$

Minimum Frob norm of the Hessian model

$$\begin{array}{ll} \min_{\alpha} & \|\alpha_Q\|_2 \\ \text{s.t.} & M_L \alpha_L + M_Q \alpha_Q = f(Y) \end{array}$$

$$m(x) = \tfrac{1}{2} x^\top H x + g^\top x + \kappa$$

- $\alpha_L \rightarrow (k, g)$
- $\alpha_Q \rightarrow H$

Sparse quadratic interpolation models

$$M(\bar{\phi}, Y) = M = \begin{array}{c} \begin{array}{cc} \overbrace{\hspace{8em}}^{M_L} & \overbrace{\hspace{14em}}^{M_Q} \end{array} \\ \left[\begin{array}{cccccccc} 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{2}(y_n^1)^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{2}(y_n^p)^2 \end{array} \right] \end{array}$$

Sparse interpolation model

$$\begin{array}{ll} \min_{\alpha} & \|\alpha_Q\|_1 \\ \text{s.t.} & M_L \alpha_L + M_Q \alpha_Q = f(Y) \end{array}$$

$$m(x) = \tfrac{1}{2} x^\top H x + g^\top x + \kappa$$

- $\alpha_L \rightarrow (k, g)$
- $\alpha_Q \rightarrow H$

Sparse interpolation model recovery

Recovery by using the l_1-norm

Recovering sparse solution, x such that Ax=b

given matrix $A \in R^{m \times n}$, m<<n

The system is underdetermined, but if card(x)=s<m, can recover signal,

$$\min \quad ||x||_0$$
$$s.t. \quad Ax = b.$$

Under certain conditions of matrix *A* (RIP) recover *x* from

$$\min \quad ||x||_1$$
$$s.t. \quad Ax = b.$$

Candes, Tao, Donoho…..

Partial recovery by the l_1-norm

Assume x_1 is dense and but if card(x_2)=s-r<m-r, then recover signal,

$$\min \quad ||x_2||_0$$
$$s.t. \quad A_1x_1 + A_2x_2 = b.$$

Under modified conditions of matrix A (partial RIP) recover x from

$$\min \quad ||x_2||_1$$
$$s.t. \quad A_1x_1 + A_2x_2 = b.$$

RIP => Partial RIP

Vaswavi &Lu' 10,

Bandeira, S and Vicente' 10

Sparse recovery for interpolation

We want to recover (partially) sparse vector α such that

$$M(\phi, Y)\alpha = f(Y) \qquad (*)$$

We need *M(*ϕ*,Y)* to satisfy (partial) RIP. How can this be done? Choose appropriate *Y* and ϕ

Definition 1. *A "suitable basis" $\phi = \{\phi_1, ..., \phi_q\}$ is an orthonormal basis, in the domain $\mathcal{D}$ for the measure μ, satisfying the K-boundedness condition; i.e.,*

$$\int_{\mathcal{D}} \phi_i(x)\phi_j(x)d\mu(x) = \delta_{ij}$$

and $\max_{x\in\mathcal{D}} |\phi_j(x)| \le K$, for all $i, j = 1, \ldots, q$, for which the solution of (*) is expected to be sparse.

Random matrix property

Theorem 1. *Let ϕ be a "suitable basis". Let a sample set $Y = \{y^1, ..., y^p\} \subset \mathcal{D}$ be chosen randomly (i.i.d) according to the probability measure μ. If the number of samples p satisfies*

$$\begin{aligned} \frac{p}{\log p} &\geq c_1 K^2 s (\log s)^2 \log q \\ p &\geq c_2 K^2 s \log\left(\frac{1}{\varepsilon}\right), \end{aligned}$$

Then, the sparse solution is recovered with probability at least $1-\varepsilon$ ($M(\phi, Y)$ satisfies the RIP property)

Rauhut' 10

Suitable basis

Definition 1. *We define the basis ψ as the following $(n+1)(n+2)/2$ polynomials:*

- $$\psi_1(x) = 1$$
- $$\psi_{1+i}(x) = \frac{\sqrt{3}}{\Delta} x_i$$
- $$\psi_{n+1+(i-1)*n+j}(x) = \frac{3}{\Delta^2} x_i x_j$$
- $$\psi_{n+2+(i-1)*n}(x) = \frac{3\sqrt{5}}{2} \frac{1}{\Delta^2} x_i^2 - \frac{\sqrt{5}}{2}.$$

ψ(*x*) is *K*-bounded and orthonormal on a hypercube or radius Δ centered at zero.

Sparse quadratic interpolation models

$$M(\bar{\phi}, Y) = M = \left[\begin{array}{cccc|cccc} \multicolumn{4}{c}{\overbrace{\hphantom{1 \quad y_1^1 \quad \cdots \quad y_n^1}}^{M_L}} & \multicolumn{4}{c}{\overbrace{\hphantom{\frac{1}{2}(y_1^1)^2 \quad y_1^1 y_2^1 \quad \cdots \quad \frac{1}{2}(y_n^1)^2}}^{M_Q}} \\ 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{2}(y_n^1)^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{2}(y_n^p)^2 \end{array}\right]$$

Sparse interpolation model

$$\begin{array}{ll} \min_{\alpha} & \|\alpha_Q\|_1 \\ \text{s.t.} & M_L\alpha_L + M_Q\alpha_Q = f(Y) \end{array}$$

$$m(x) = \tfrac{1}{2}x^\top Hx + g^\top x + \kappa$$

- $\alpha_L \to (k, g)$
- $\alpha_Q \to H$

We should not assume that interpolation model is exactly sparse

Sparse quadratic approximation models

We assume that there exists a fully quadratic model *$m^*(x)$* of *$f(x)$* with sparse Hessian.

$$m^*(x) = \sum_{i=0}^{p} \alpha_i^* \phi(x)$$

where $\|M(\phi, Y)\alpha^* - f(Y)\| \leq O(\Delta^3)$, and α_Q^* is sparse

We seek α: (α may not equal α*)

$$\min_{\alpha} \quad \|\alpha_Q\|_1$$
$$\text{s.t.} \quad \|M_L \alpha_L + M_Q \alpha_Q - f(Y)\| \leq O(\Delta^3)$$

Noisy recovery using random points

Theorem 1. *Under the same assumptions, with probability at least $1-\varepsilon$, $\varepsilon \in (0,1)$, the following holds for every $s-$sparse vector x:*
Let noisy samples $f(Y) = M(\phi, Y)x + \epsilon$ with

$$\|\epsilon\|_2 \leq \eta$$

be given, for any positive η, and let x^* be the solution of the noisy ℓ_1-minimization problem with $A = M(\phi, Y)$. Then,

$$\|x - x^*\|_2 \leq \frac{d}{\sqrt{p}}\eta$$

for some universal constant $d > 0$.

Rauhut' 10

Main theorem

Theorem 1. *Let* $m(x) = \sum_i \alpha\psi_i(x)$ *be an* s*-sparse fully quadratic model of* f *on* $\Delta \in (0, \Delta_{\max}]$. *Given* p *random points,* $Y = \{y^1, ..., y^p\}$, *chosen uniformly in* $B_\infty(0;\Delta)$, *with*

$$\frac{p}{\log p} \geq c\,(s+n+1)\log^2(s+n+1)\log n, \tag{1}$$

with probability larger than $1 - n^{-\gamma \log p}$, *the solution* m^* *to the* ℓ_1*-minimization problem is a fully quadratic model of* f *on* $B_\infty(0;\Delta)$.

Conclusion: we can construct fully quadratic models of functions with sparse Hessians with O(n) sample points (with high probability).

Bandeira, S and Vicente' 10

New paradigm for "good" models

Probabilistic Taylor-like behavior of first or second order models

A random model is called fully linear in $B(x,\Delta)$
if with probability $1-\delta$

$$\|\nabla f(x+s) - \nabla m(x+s)\| \;\leq\; \kappa_{eg}\,\Delta, \quad \forall s \in B(0;\Delta),$$

$$|f(x+s) - m(x+s)| \;\leq\; \kappa_{ef}\,\Delta^2, \quad \forall s \in B(0;\Delta),$$

for some fixed κ_{eg} and κ_{ef} independent of x, Δ and δ.

What is a "better" model?

We need Taylor-like behavior of first or second order models

A model is called fully quadratic in $B(x, \Delta)$ if
with probability at least $1 - \delta$

$$\|\nabla^2 f(x+s) - \nabla^2 m(x+s)\| \;\leq\; \kappa_{eh}\,\Delta, \quad \forall s \in B(0;\Delta),$$

$$\|\nabla f(x+s) - \nabla m(x+s)\| \;\leq\; \kappa_{eg}\,\Delta^2, \quad \forall s \in B(0;\Delta),$$

$$|f(x+s) - m(x+s)| \;\leq\; \kappa_{ef}\,\Delta^3, \quad \forall s \in B(0;\Delta),$$

for some fixed κ_{eh}, κ_{eg}, κ_{ef} independent of x, Δ and δ.

So what about convergence?

The previous theory does not apply as it relies on knowing if the model is “good”.

Consider a simple TR Algorithm

Initialize: Choose a class of models, initialize x_0, $m_0(x)$, Δ_0. Choose $\eta_1 > 0$, $\eta_2 > 0$ and $\gamma > 1$.

Model selection step Build a random model $m_k(x)$ which is fully-linear in $B(x_k, \Delta_k)$ with probability $1 - \delta$.

Compute Step: Compute s_k from $\min_{\|s\| \le \Delta_k} m_k(x_k + s)$ evaluate $f(x_k + s_k)$ and $r_k = (f(x_k) - f(x_k + s_k))/(m(x_k) - m(x_k + s_k))$.

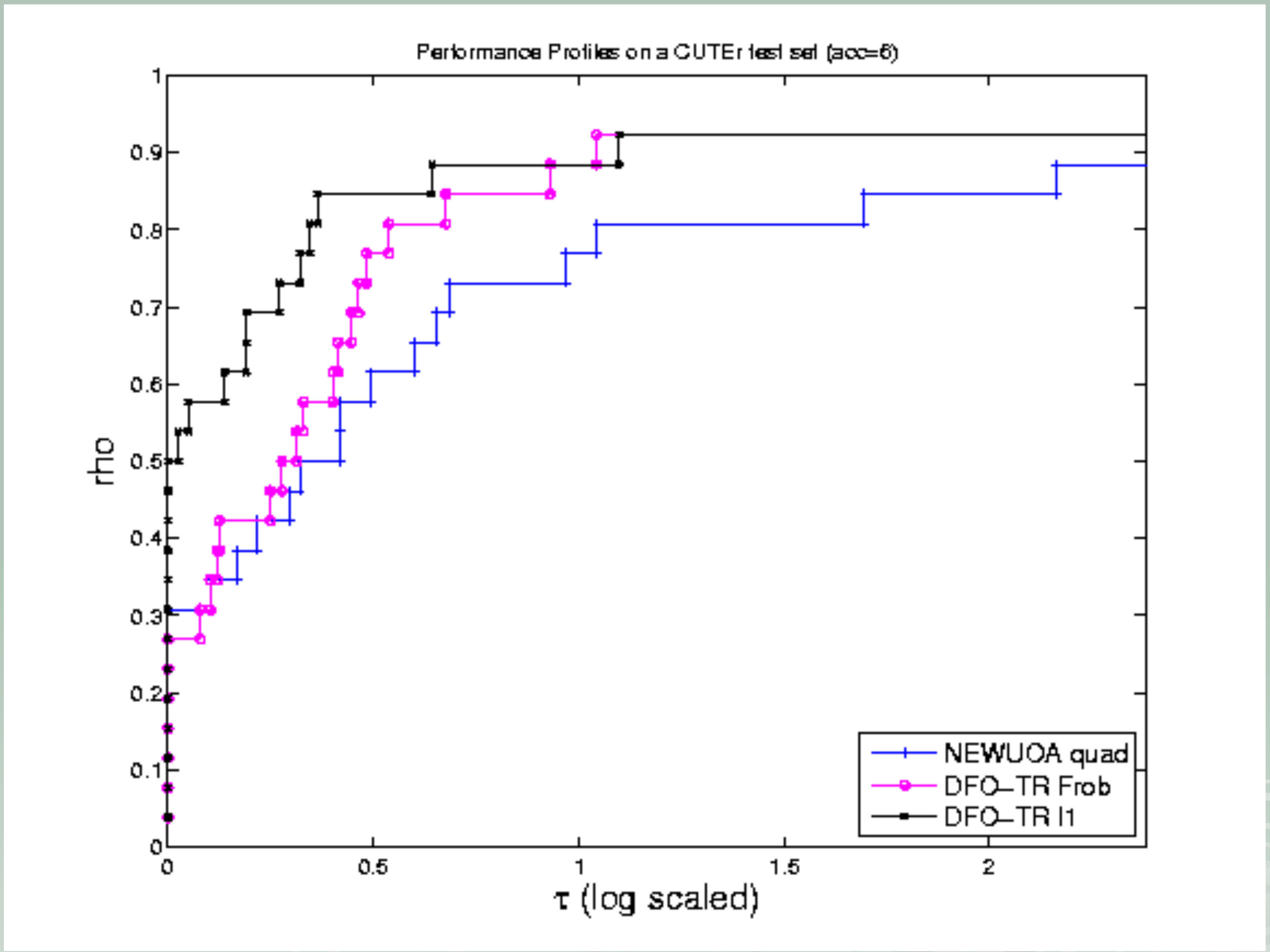
Successful step: If $r_k \ge \eta_1$ and $\nabla m_k(x_k) \ge \eta_2 \Delta_k$ then $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \gamma \Delta_k$.

Unsuccessful step: If $r_k < \eta_1$ or $\nabla m_k(x_k) < \eta_2 \Delta_k$ then $x_{k+1} = x_k$ and $\Delta_{k+1} = \gamma^{-1} \Delta_k$.

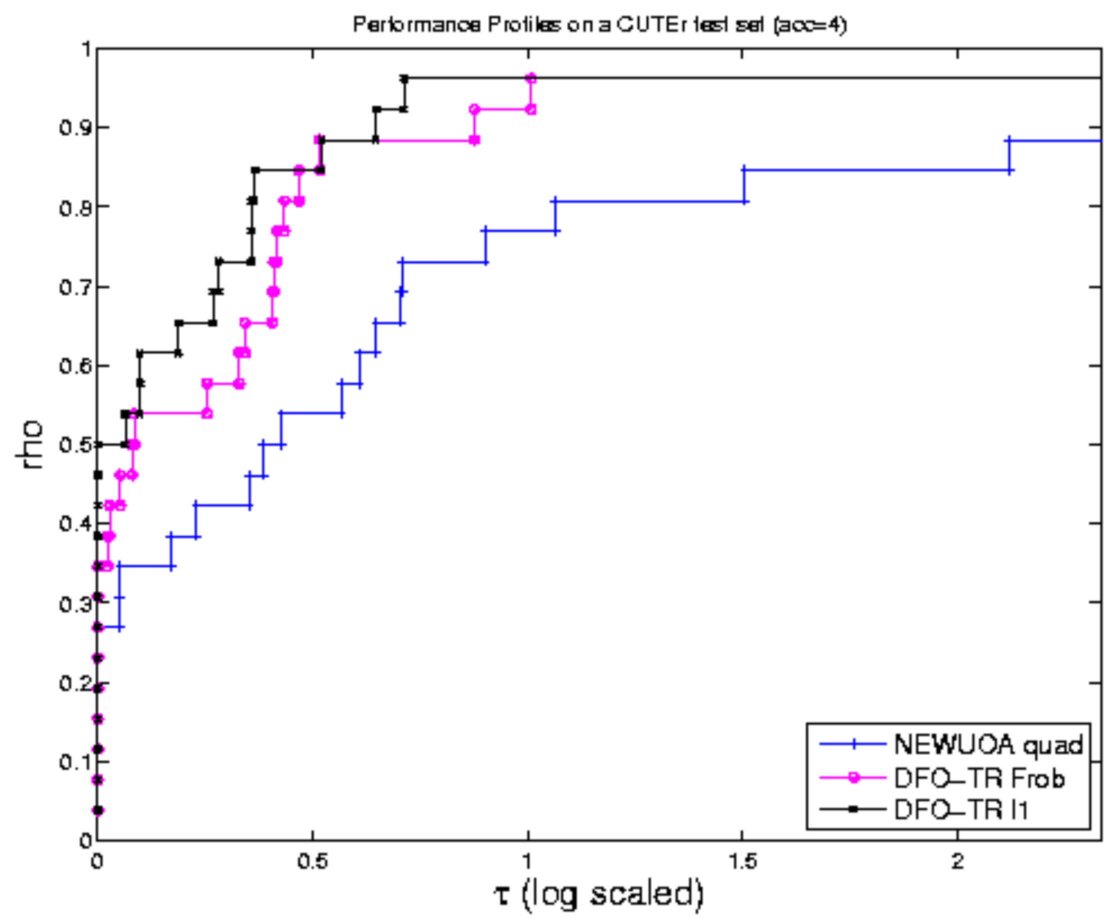
Result: if δ <0.5 then with probability 1

$$\liminf \nabla f(x_k) = 0$$

Comparison for Δ_{min}=10^{-6}



Comparison for Δ_{min}=10^{-4}



Work to do

- Complete convergence theory based on random models.
- Improving the results using new partial recovery results.
- Extending to different models.
- Recovering other types of structure.
- Efficient implementation.

Thank you!