PDE-Constrained and Nonsmooth Optimization

Frank E. Curtis

COR@L Seminar

October 1, 2009

Outline

Outline

PDE-constrained optimization

$$\begin{aligned} \min\ & f(x) \\ \text{s.t. } & c_{\mathcal{E}}(x) = 0 \\ & c_{\mathcal{I}}(x) \geq 0 \end{aligned}$$

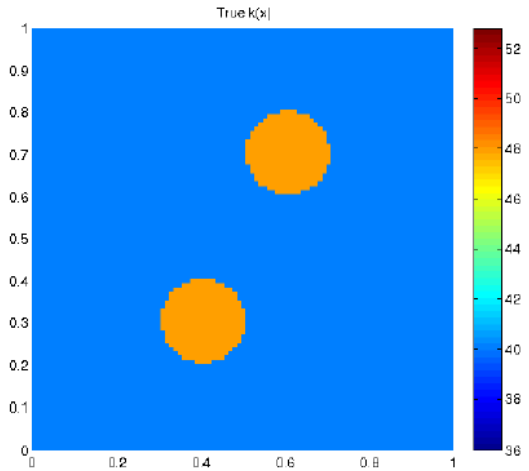
PDE-constrained optimization

$$\begin{aligned} \min\ & f(x) \\ \text{s.t. } & c_{\mathcal{E}}(x) = 0 \qquad (PDE) \\ & c_{\mathcal{I}}(x) \geq 0 \end{aligned}$$

Problem is infinite-dimensional

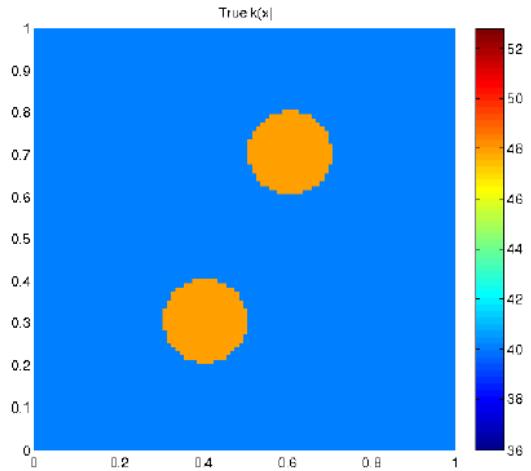
Inverse problems

Recover a parameter $k(x)$ based on data collected from propagating waves



Inverse problems

Recover a parameter $k(x)$ based on data collected from propagating waves

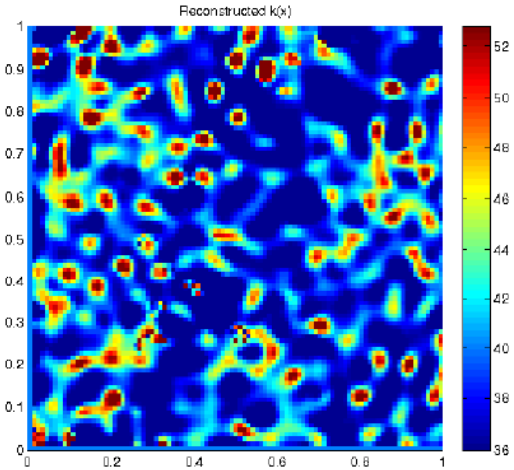


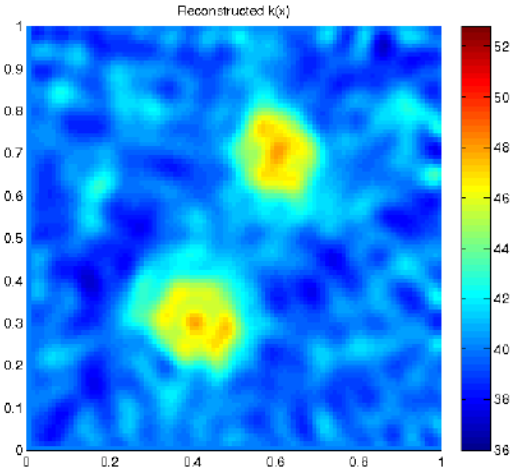
$$\min_{y,k} \frac{1}{2} \sum_j \sum_m (y_j(x_m) - y_{j,m})^2 + \alpha(\beta \|k\|^2_{L^2(\Omega)} + \|\nabla k\|^2_{(L^2(\Omega))^n})$$

$$\text{s.t.} \begin{cases} \Delta y_j(x) + S(x)k(x)^2 y_j(x) - S(x)(k_0^2 - k(x)^2) y_j^i = 0, \; x \text{ in } \Omega \\ y_j = 0, \; x \text{ on } \partial\Omega \\ l(x) \leq k(x) \leq u(x), \; x \text{ in } \Omega \end{cases}$$

Inverse problems

Recover a parameter $k(x)$ based on data collected from propagating waves



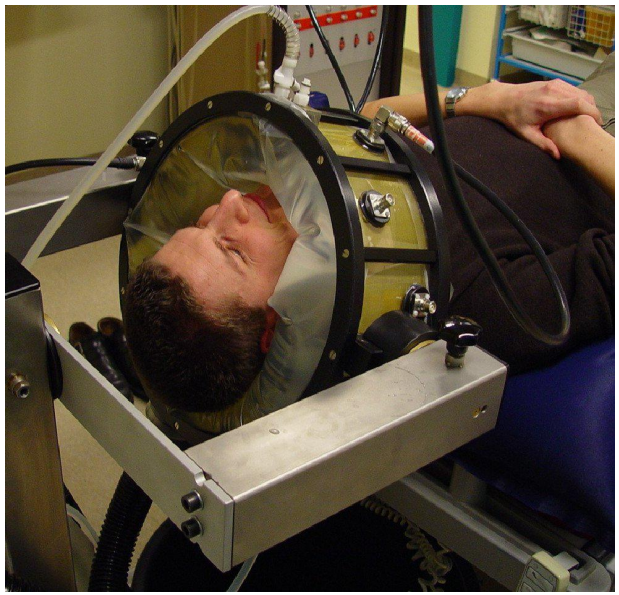


$$\min_{y,k} \frac{1}{2} \sum_j \sum_m (y_j(x_m) - y_{j,m})^2 + \alpha(\beta \|k\|^2_{L^2(\Omega)} + \|\nabla k\|^2_{(L^2(\Omega))^n})$$

$$\text{s.t.} \begin{cases} \Delta y_j(x) + S(x)k(x)^2 y_j(x) - S(x)(k_0^2 - k(x)^2) y_j^i = 0, \; x \text{ in } \Omega \\ y_j = 0, \; x \text{ on } \partial\Omega \\ l(x) \leq k(x) \leq u(x), \; x \text{ in } \Omega \end{cases}$$

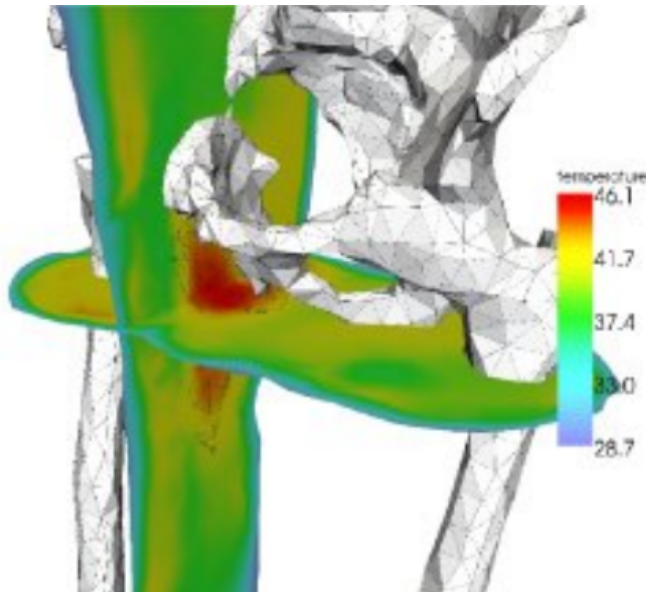
Optimal design

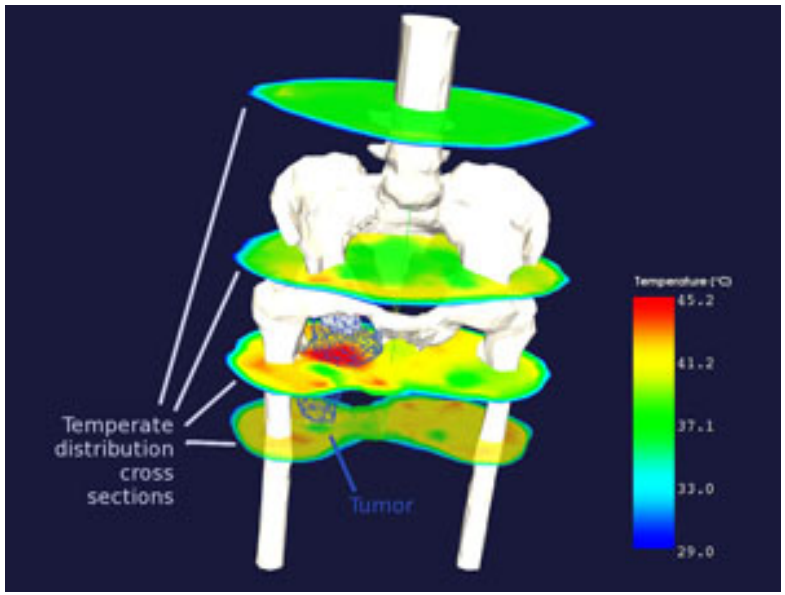
- Regional hyperthermia is a cancer therapy that aims at heating large and deeply seated tumors by means of radio wave adsorption
- Results in the killing of tumor cells and makes them more susceptible to other accompanying therapies; e.g., chemotherapy



Optimal design

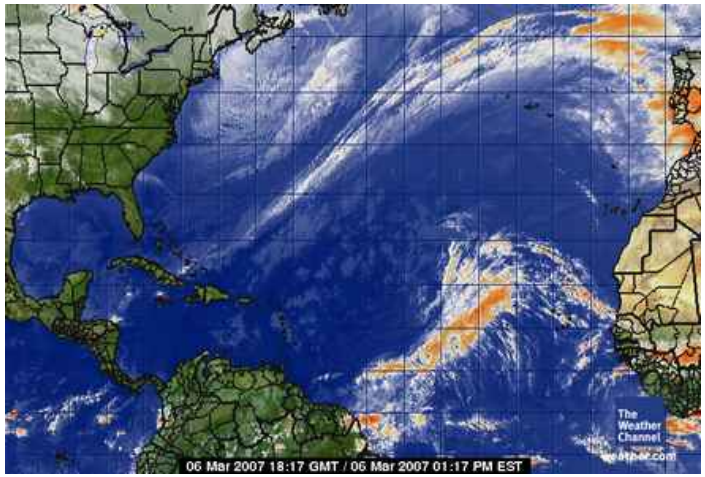
- Computer modeling can be used to help plan the therapy for each patient, and it opens the door for numerical optimization
- The goal is to heat the tumor to a target temperature of 43°C while minimizing damage to nearby cells





Parameter estimation

- Weather forecasting

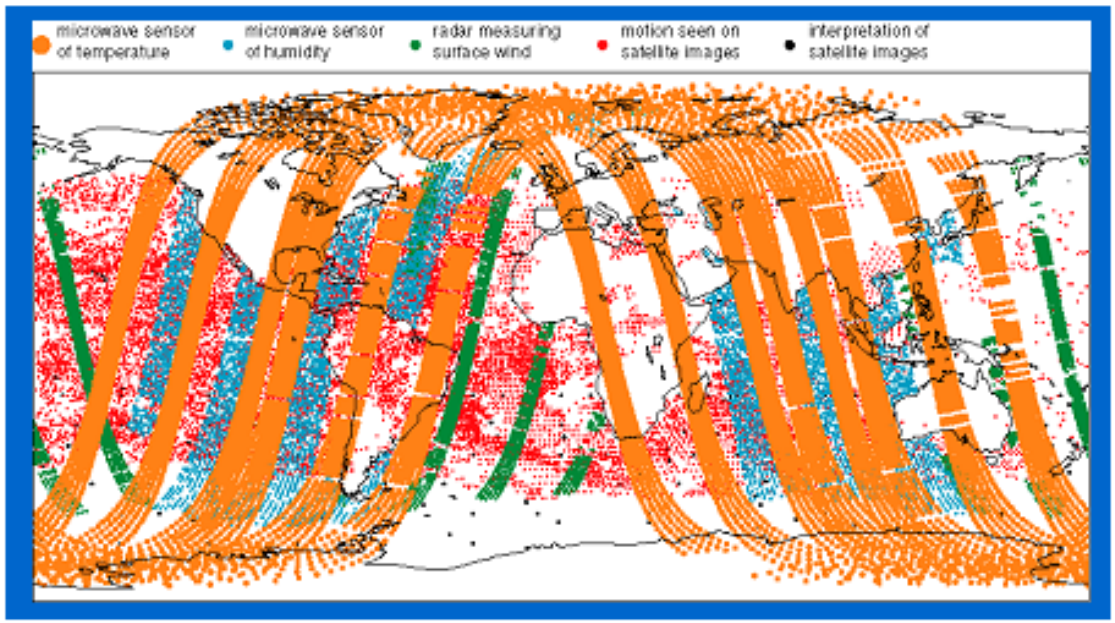


- If the initial state of the atmosphere (temperatures, pressures, wind patterns, humidities) were known at a certain point in time, then an accurate forecast could be obtained by integrating atmospheric model equations forward in time
- Flow described by Navier-Stokes and further sophistications of atmospheric physics and dynamics

Parameter estimation

Limited amount of data (satellites, buoys, planes, ground-based sensors)

- Each observation is subject to error
- Nonuniformly distributed around the globe (satellite paths, densely-populated areas)



Outline

Nonlinear equations

- Newton's method

$$\boxed{\mathcal{F}(x) = 0} \Rightarrow \boxed{\nabla \mathcal{F}(x_k) d_k = -\mathcal{F}(x_k)}$$

- Judge progress by the merit function

$$\phi(x) \triangleq \tfrac{1}{2} \|\mathcal{F}(x)\|^2$$

- Direction is one of descent since

$$\nabla \phi(x_k)^T d_k = \mathcal{F}(x_k)^T \nabla \mathcal{F}(x_k) d_k = -\|\mathcal{F}(x_k)\|^2 < 0$$

(Note the consistency between the step computation and merit function!)

Equality constrained optimization

- Consider

$$\min_{x \in \mathbb{R}^n} f(x)$$
$$\text{s.t. } c(x) = 0$$

- Lagrangian is

$$\mathcal{L}(x, \lambda) \triangleq f(x) + \lambda^T c(x)$$

so the first-order optimality conditions are

$$\nabla \mathcal{L}(x, \lambda) = \begin{bmatrix} \nabla f(x) + \nabla c(x) \lambda \\ c(x) \end{bmatrix} \triangleq \mathcal{F}(x, \lambda) = 0$$

Merit function

- Simply minimizing

$$\varphi(x,\lambda) = \tfrac{1}{2}\|\mathcal{F}(x,\lambda)\|^2 = \tfrac{1}{2}\left\|\begin{bmatrix}\nabla f(x) + \nabla c(x)\lambda \\ c(x)\end{bmatrix}\right\|^2$$

  is generally inappropriate for constrained optimization

- We use the merit function

$$\phi(x;\pi) \triangleq f(x) + \pi\|c(x)\|$$

  where π is a penalty parameter

Minimizing a penalty function

Consider the penalty function for

$$\min\ (x-1)^2,\ \text{s.t. } x=0 \quad \text{i.e.} \quad \phi(x;\pi) = (x-1)^2 + \pi|x|$$

for different values of the penalty parameter π

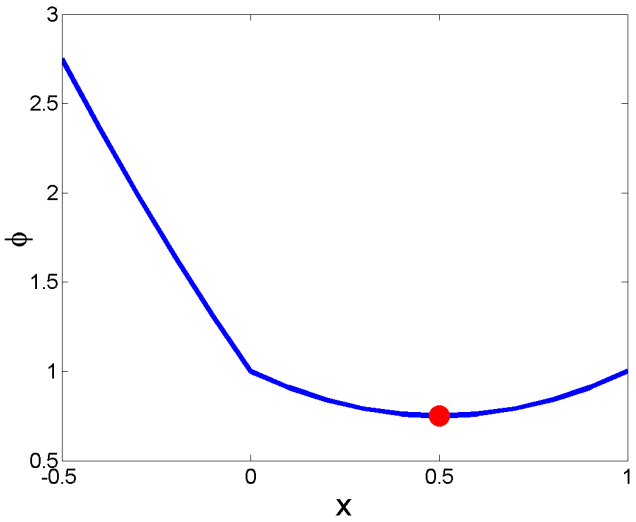


Figure: $\pi = 1$

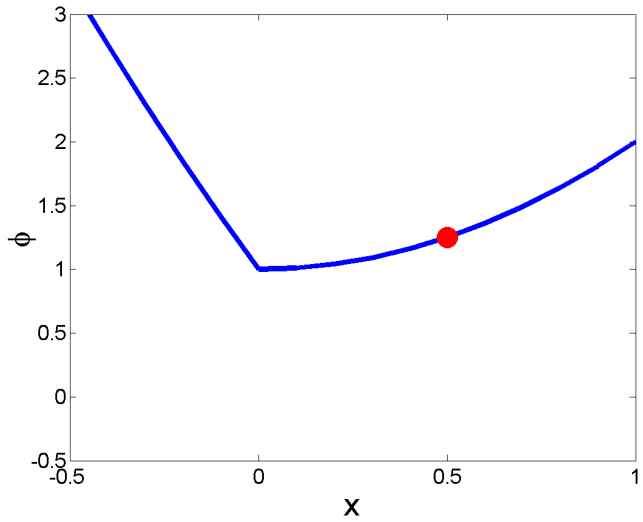


Figure: $\pi = 2$

Algorithm 0: Newton method for optimization

(Assume the problem is sufficiently convex and regular)
for $k = 0, 1, 2, \ldots$

- Solve the primal-dual (Newton) equations

$$\begin{bmatrix} H(x_k, \lambda_k) & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k)\lambda_k \\ c(x_k) \end{bmatrix}$$

- Increase π, if necessary, so that $D\phi_k(d_k; \pi_k) \ll 0$ (e.g., $\pi_k \geq \|\lambda_k + \delta_k\|$)
- Backtrack from $\alpha_k \leftarrow 1$ to satisfy the Armijo condition

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k D\phi_k(d_k; \pi_k)$$

- Update iterate $(x_{k+1}, \lambda_{k+1}) \leftarrow (x_k, \lambda_k) + \alpha_k(d_k, \delta_k)$

Convergence of Algorithm 0

Assumption

The sequence $\{(x_k, \lambda_k)\}$ *is contained in a convex set* Ω *over which* f, c, *and their first derivatives are bounded and Lipschitz continuous. Also,*

- *(Regularity)* $\nabla c(x_k)^T$ *has full row rank with singular values bounded below by a positive constant*
- *(Convexity)* $u^T H(x_k, \lambda_k) u \geq \mu \|u\|^2$ *for* $\mu > 0$ *for all* $u \in \mathbb{R}^n$ *satisfying* $u \neq 0$ *and* $\nabla c(x_k)^T u = 0$

Theorem

(Han (1977)) The sequence $\{(x_k, \lambda_k)\}$ *yields the limit*

$$\lim_{k \to \infty} \left\| \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k) \lambda_k \\ c(x_k) \end{bmatrix} \right\| = 0$$

Outline

Large-scale primal-dual algorithms

- Computational issues:
  - Large matrices to be stored
  - Large matrices to be factored
- Algorithmic issues:
  - The problem may be nonconvex
  - The problem may be ill-conditioned
- Computational/Algorithmic issues:
  - No matrix factorizations makes difficulties more difficult

Nonlinear equations

- Compute
$$\nabla \mathcal{F}(x_k) d_k = -\mathcal{F}(x_k) + r_k$$
requiring (Dembo, Eisenstat, Steihaug (1982))
$$\|r_k\| \leq \kappa \|\mathcal{F}(x_k)\|, \quad \kappa \in (0,1)$$
- Progress judged by the merit function
$$\phi(x) \triangleq \tfrac{1}{2}\|\mathcal{F}(x)\|^2$$
- Again, note the consistency...

$$\nabla \phi(x_k)^T d_k = \mathcal{F}(x_k)^T \nabla \mathcal{F}(x_k) d_k = -\|\mathcal{F}(x_k)\|^2 + \mathcal{F}(x_k)^T r_k \leq (\kappa - 1)\|\mathcal{F}(x_k)\|^2 < 0$$

Optimization

- Compute

$$\begin{bmatrix} H(x_k, \lambda_k) & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k)\lambda_k \\ c(x_k) \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

  satisfying

$$\left\| \begin{bmatrix} \rho_k \\ r_k \end{bmatrix} \right\| \leq \kappa \left\| \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k)\lambda_k \\ c(x_k) \end{bmatrix} \right\|, \quad \kappa \in (0, 1)$$

- If κ is not sufficiently small (e.g., 10^{-3} vs. 10^{-12}), then d_k may be an ascent direction for our merit function; i.e.,

$$D\phi_k(d_k; \pi_k) > 0 \quad \text{for all } \pi_k \geq \pi_{k-1}$$

- Our work begins here... inexact Newton methods for optimization
- We cover the convex case, nonconvexity, irregularity, inequality constraints

Model reductions

- Define the model of $\phi(x;\pi)$:

$$m(d;\pi) \triangleq f(x) + \nabla f(x)^T d + \pi(\|c(x) + \nabla c(x)^T d\|)$$

- d_k is acceptable if

$$\begin{aligned}\Delta m(d_k;\pi_k) \triangleq\ & m(0;\pi_k) - m(d_k;\pi_k) \\ =\ & -\nabla f(x_k)^T d_k + \pi_k(\|c(x_k)\| - \|c(x_k) + \nabla c(x_k)^T d_k\|) \gg 0\end{aligned}$$

- This ensures $D\phi_k(d_k;\pi_k) \ll 0$ (and more)

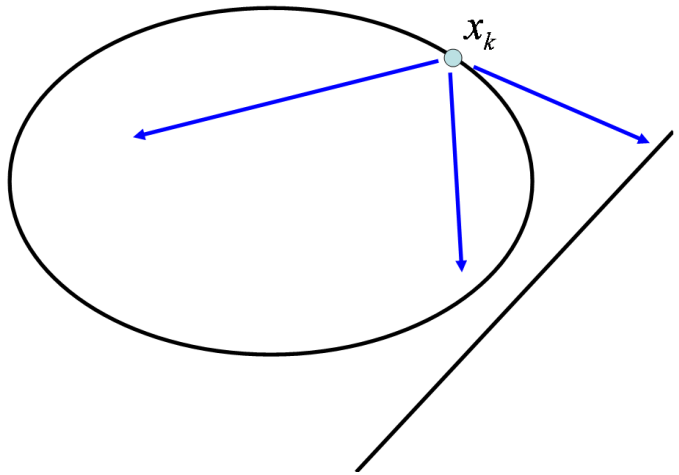
Termination test 1

The search direction (d_k, δ_k) is acceptable if

$$\left\| \begin{bmatrix} \rho_k \\ r_k \end{bmatrix} \right\| \leq \kappa \left\| \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k)\lambda_k \\ c(x_k) \end{bmatrix} \right\|, \quad \kappa \in (0,1)$$

and if for $\pi_k = \pi_{k-1}$ and some $\sigma \in (0,1)$ we have

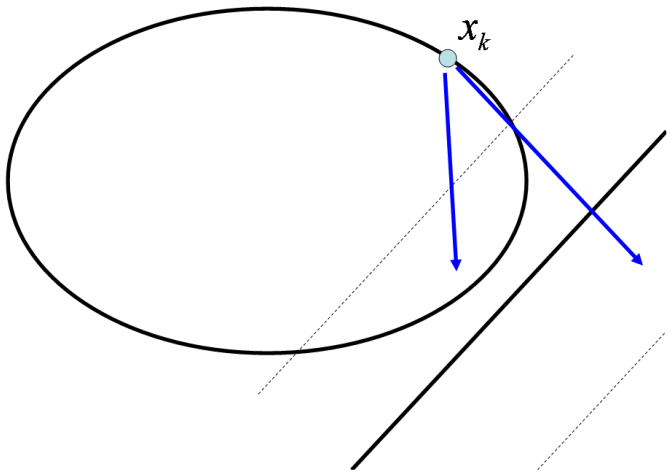
$$\Delta m(d_k; \pi_k) \geq \underbrace{\max\{\tfrac{1}{2} d_k^T H(x_k, \lambda_k) d_k, 0\} + \sigma \pi_k \max\{\|c(x_k)\|, \|r_k\| - \|c(x_k)\|\}}_{\geq 0 \text{ for any } d}$$



Termination test 2

The search direction (d_k, δ_k) is acceptable if

$$\begin{aligned} \|\rho_k\| &\leq \beta \|c(x_k)\|, \quad \beta > 0 \\ \text{and} \quad \|r_k\| &\leq \epsilon \|c(x_k)\|, \quad \epsilon \in (0,1) \end{aligned}$$



Increasing the penalty parameter π then yields

$$\Delta m(d_k; \pi_k) \geq \underbrace{\max\{\tfrac{1}{2} d_k^T H(x_k, \lambda_k) d_k, 0\} + \sigma \pi_k \|c(x_k)\|}_{\geq 0 \text{ for any } d}$$

Algorithm 1: Inexact Newton for optimization

(Byrd, Curtis, Nocedal (2008))

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

- Iteratively solve

$$\begin{bmatrix} H(x_k, \lambda_k) & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k)\lambda_k \\ c(x_k) \end{bmatrix}$$

  until termination test 1 or 2 is satisfied

- If only termination test 2 is satisfied, increase π so

$$\pi_k \geq \max \left\{ \pi_{k-1}, \frac{\nabla f(x_k)^T d_k + \max\{\frac{1}{2} d_k^T H(x_k, \lambda_k) d_k, 0\}}{(1-\tau)(\|c(x_k)\| - \|r_k\|)} \right\}$$

- Backtrack from $\alpha_k \leftarrow 1$ to satisfy

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) - \eta \alpha_k \Delta m(d_k; \pi_k)$$

- Update iterate $(x_{k+1}, \lambda_{k+1}) \leftarrow (x_k, \lambda_k) + \alpha_k (d_k, \delta_k)$

Convergence of Algorithm 1

Assumption

The sequence $\{(x_k, \lambda_k)\}$ *is contained in a convex set* Ω *over which* f, c, *and their first derivatives are bounded and Lipschitz continuous. Also,*

- *(Regularity)* $\nabla c(x_k)^T$ *has full row rank with singular values bounded below by a positive constant*
- *(Convexity)* $u^T H(x_k, \lambda_k) u \geq \mu \|u\|^2$ *for* $\mu > 0$ *for all* $u \in \mathbb{R}^n$ *satisfying* $u \neq 0$ *and* $\nabla c(x_k)^T u = 0$

Theorem

(Byrd, Curtis, Nocedal (2008)) The sequence $\{(x_k, \lambda_k)\}$ *yields the limit*

$$\lim_{k \to \infty} \left\| \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k) \lambda_k \\ c(x_k) \end{bmatrix} \right\| = 0$$

Handling nonconvexity and rank deficiency

- There are two assumptions we aim to drop:
  - *(Regularity)* $\nabla c(x_k)^T$ *has full row rank with singular values bounded below by a positive constant*
  - *(Convexity)* $u^T H(x_k, \lambda_k) u \geq \mu \|u\|^2$ *for* $\mu > 0$ *for all* $u \in \mathbb{R}^n$ *satisfying* $u \neq 0$ *and* $\nabla c(x_k)^T u = 0$

  e.g., the problem is not regular if it is infeasible, and it is not convex if there are maximizers and/or saddle points
- Without them, Algorithm 1 may stall or may not be well-defined

No factorizations means no clue

- We might not store or factor

$$\begin{bmatrix} H(x_k, \lambda_k) & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix}$$

  so we might not know if the problem is nonconvex or ill-conditioned

- Common practice is to perturb the matrix to be

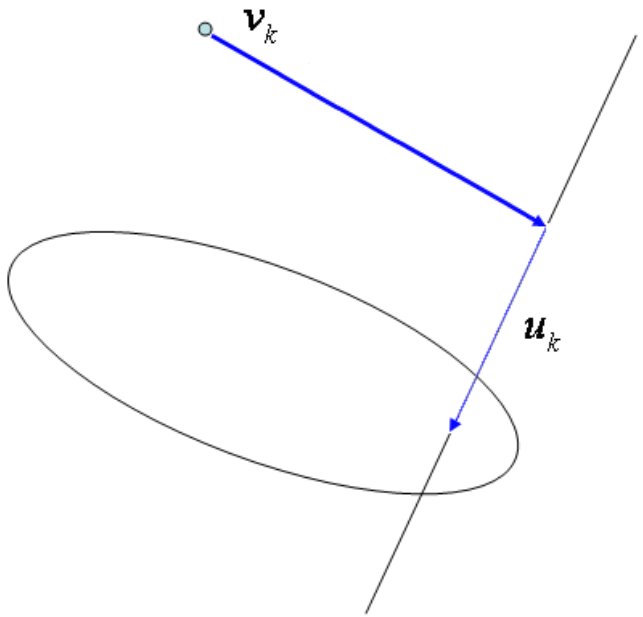
$$\begin{bmatrix} H(x_k, \lambda_k) + \xi_1 I & \nabla c(x_k) \\ \nabla c(x_k)^T & -\xi_2 I \end{bmatrix}$$

  where ξ_1 convexifies the model and ξ_2 regularizes the constraints

- Poor choices of ξ_1 and ξ_2 can have terrible consequences in the algorithm

Our approach for global convergence

- Decompose the direction d_k into a normal component (toward the constraints) and a tangential component (toward optimality)



- We impose a specific type of trust region constraint on the v_k step in case the constraint Jacobian is (near) rank deficient

Handling nonconvexity

- In computation of $d_k = v_k + u_k$, convexify the Hessian as in

$$\begin{bmatrix} H(x_k, \lambda_k) + \xi_1 I & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix}$$

  by monitoring iterates

- Hessian modification strategy: Increase ξ_1 whenever

$$\begin{aligned} \|u_k\|^2 &> \psi \|v_k\|^2, \quad \psi > 0 \\ \tfrac{1}{2} u_k^T (H(x_k, \lambda_k) + \xi_1 I) u_k &< \theta \|u_k\|^2, \quad \theta > 0 \end{aligned}$$

Algorithm 2: Inexact Newton (regularized)

(Curtis, Nocedal, Wächter (2009))
for $k = 0, 1, 2, \ldots$

- Approximately solve

$$\min \ \tfrac{1}{2}\|c(x_k) + \nabla c(x_k)^T v\|^2, \quad \text{s.t. } \|v\| \leq \omega \|(\nabla c(x_k))c(x_k)\|$$

  to compute v_k satisfying Cauchy decrease

- Iteratively solve

$$\begin{bmatrix} H(x_k, \lambda_k) + \xi_1 I & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k)\lambda_k \\ -\nabla c(x_k)^T v_k \end{bmatrix}$$

  until termination test 1 or 2 is satisfied, increasing ξ_1 as described

- If only termination test 2 is satisfied, increase π so

$$\pi_k \geq \max \left\{ \pi_{k-1}, \frac{\nabla f(x_k)^T d_k + \max\{\frac{1}{2} u_k^T (H(x_k, \lambda_k) + \xi_1 I) u_k, \theta \|u_k\|^2\}}{(1-\tau)(\|c(x_k)\| - \|c(x_k) + \nabla c(x_k)^T d_k\|)} \right\}$$

- Backtrack from $\alpha_k \leftarrow 1$ to satisfy

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) - \eta \alpha_k \Delta m(d_k; \pi_k)$$

- Update iterate $(x_{k+1}, \lambda_{k+1}) \leftarrow (x_k, \lambda_k) + \alpha_k (d_k, \delta_k)$

Convergence of Algorithm 2

Assumption

The sequence $\{(x_k, \lambda_k)\}$ is contained in a convex set Ω over which f, c, and their first derivatives are bounded and Lipschitz continuous

Theorem

(Curtis, Nocedal, Wächter (2009)) If all limit points of $\{\nabla c(x_k)^T\}$ have full row rank, then the sequence $\{(x_k, \lambda_k)\}$ yields the limit

$$\lim_{k \to \infty} \left\| \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k)\lambda_k \\ c(x_k) \end{bmatrix} \right\| = 0.$$

Otherwise,

$$\lim_{k \to \infty} \|(\nabla c(x_k))c(x_k)\| = 0$$

and if $\{\pi_k\}$ is bounded, then

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

Handling inequalities

- Interior point methods are attractive for large applications
- Line-search interior point methods that enforce

$$c(x_k) + \nabla c(x_k)^T d_k = 0$$

  may fail to converge globally (Wächter, Biegler (2000))
- Fortunately, the trust region subproblem we use to regularize the constraints also saves us from this type of failure!

Algorithm 2 (Interior-point version)

- Apply Algorithm 2 to the logarithmic-barrier subproblem

$$\min \; f(x) - \mu \sum_{i=1}^{q} \ln s^i, \quad \text{s.t. } c_{\mathcal{E}}(x) = 0, \; c_{\mathcal{I}}(x) - s = 0$$

for $\mu \to 0$

- Define

$$\begin{bmatrix} H(x_k, \lambda_{\mathcal{E},k}, \lambda_{\mathcal{I},k}) & 0 & \nabla c_{\mathcal{E}}(x_k) & \nabla c_{\mathcal{I}}(x_k) \\ 0 & \mu I & 0 & -S_k \\ \nabla c_{\mathcal{E}}(x_k)^T & 0 & 0 & 0 \\ \nabla c_{\mathcal{I}}(x_k)^T & -S_k & 0 & 0 \end{bmatrix} \begin{bmatrix} d_k^x \\ d_k^s \\ \delta_{\mathcal{E},k} \\ \delta_{\mathcal{I},k} \end{bmatrix}$$

so that the iterate update has

$$\begin{bmatrix} x_{k+1} \\ s_{k+1} \end{bmatrix} \leftarrow \begin{bmatrix} x_k \\ s_k \end{bmatrix} + \alpha_k \begin{bmatrix} d_k^x \\ S_k d_k^s \end{bmatrix}$$

- Incorporate a fraction-to-the-boundary rule in the line search and a slack reset in the algorithm to maintain $s \geq \max\{0, c_{\mathcal{I}}(x)\}$

Convergence of Algorithm 2 (interior-point)

Assumption

The sequence $\{(x_k, \lambda_{\mathcal{E},k}, \lambda_{\mathcal{I},k})\}$ is contained in a convex set Ω over which f, $c_{\mathcal{E}}$, $c_{\mathcal{I}}$, and their first derivatives are bounded and Lipschitz continuous

Theorem

(Curtis, Schenk, Wächter (2009))

- *For a given μ, Algorithm 2 yields the same limits as in the equality constrained case*
- *If Algorithm 2 yields a sufficiently accurate solution to the barrier subproblem for each $\{\mu_j\} \to 0$ and if the linear independence constraint qualification (LICQ) holds at a limit point $\bar{x}$ of $\{x_j\}$, then there exist Lagrange multipliers $\bar{\lambda}$ such that the first-order optimality conditions of the nonlinear program are satisfied*

Outline

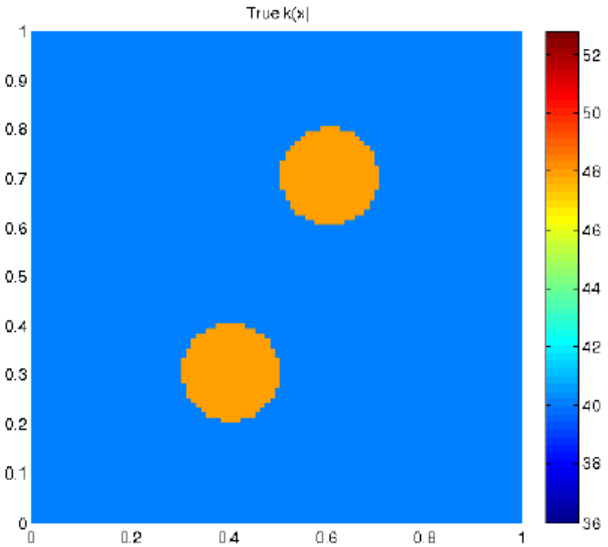
Implementation details

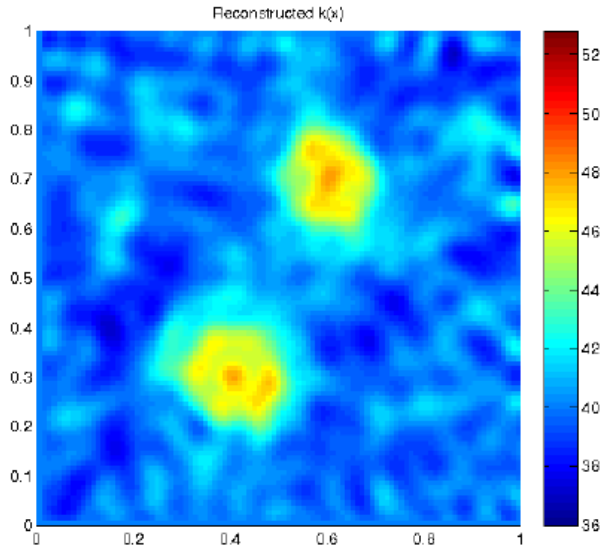
- Incorporated in `IPOPT` software package (Wächter)
  - `inexact_algorithm yes`
- Linear systems solved with `PARDISO` (Schenk)
  - SQMR (Freund (1994))
- Preconditioning in `PARDISO`
  - incomplete multilevel factorization with inverse-based pivoting
  - stabilized by symmetric-weighted matchings
- Optimality tolerance: 1e-8

CUTEr and COPS collections

- 745 problems written in AMPL
- 645 solved successfully
- 42 "real" failures
- Robustness between 87%-94%
- Original `IPOPT`: 93%

Helmholtz





N	n	p	q	# iter	CPU sec (per iter)
32	14724	13824	1800	37	807.823 (21.833)
64	56860	53016	7688	25	3741.42 (149.66)
128	227940	212064	31752	20	54581.8 (2729.1)

Boundary control

$$\begin{aligned} \min\ & \tfrac{1}{2}\int_\Omega (y(x)-y_t(x))^2 dx \\ \text{s.t.}\ & -\nabla\cdot(e^{y(x)}\cdot\nabla y(x)) = 20 \ \text{ in } \Omega \\ & y(x) = u(x) \ \text{ on } \partial\Omega \\ & 2.5 \le u(x) \le 3.5 \ \text{ on } \partial\Omega \end{aligned}$$

where

$$y_t(x) = 3 + 10x_1(x_1-1)x_2(x_2-1)\sin(2\pi x_3)$$

N	n	p	q	# iter	CPU sec (per iter)
16	4096	2744	2704	13	2.8144 (0.2165)
32	32768	27000	11536	13	103.65 (7.9731)
64	262144	238328	47632	14	5332.3 (380.88)

Original `IPOPT` with $N = 32$ requires 238 seconds per iteration

Hyperthermia treatment planning

$$
\begin{aligned}
\min\ & \tfrac{1}{2}\int_\Omega (y(x)-y_t(x))^2 dx \\
\text{s.t.}\ & -\Delta y(x) - 10(y(x)-37) = u^* M(x) u \quad \text{in } \Omega \\
& 37.0 \le y(x) \le 37.5 \quad \text{on } \partial\Omega \\
& 42.0 \le y(x) \le 44.0 \quad \text{in } \Omega_0
\end{aligned}
$$

where

$$u_j = a_j e^{i\phi_j}, \quad M_{jk}(x) = < E_j(x), E_k(x) >, \quad E_j = \sin(j x_1 x_2 x_3 \pi)$$

N	n	p	q	# iter	CPU sec (per iter)
16	4116	2744	2994	68	22.893 (0.3367)
32	32788	27000	13034	51	3055.9 (59.920)

Original `IPOPT` with $N = 32$ requires 408 seconds per iteration

Groundwater modeling

$$\begin{aligned}
\min\ & \tfrac{1}{2}\int_\Omega (y(x)-y_t(x))^2 dx + \tfrac{1}{2}\alpha \int_\Omega [\beta(u(x)-u_t(x))^2 + |\nabla(u(x)-u_t(x))|^2]dx \\
\text{s.t.}\ & -\nabla\cdot(e^{u(x)}\cdot\nabla y_i(x)) = q_i(x) \ \text{ in } \Omega, \quad i=1,\ldots,6 \\
& \nabla y_i(x)\cdot n = 0 \ \text{ on } \partial\Omega \\
& \int_\Omega y_i(x)dx = 0, \quad i=1,\ldots,6 \\
& -1 \le u(x) \le 2 \ \text{ in } \Omega
\end{aligned}$$

where

$$q_i = 100\sin(2\pi x_1)\sin(2\pi x_2)\sin(2\pi x_3)$$

N	n	p	q	# iter	CPU sec (per iter)
16	28672	24576	8192	18	206.416 (11.4676)
32	229376	196608	65536	20	1963.64 (98.1820)
64	1835008	1572864	524288	21	134418. (6400.85)

Original `IPOPT` with $N = 32$ requires approx. 20 hours for the first iteration

Outline

Summary

- We have a new framework for inexact Newton methods for optimization
- Convergence results are as good (and sometimes better) than exact methods
- Preliminary numerical results are encouraging

Future work

- Tune the method for specific applications
- Incorporate useful techniques such as filters, second-order corrections, specialized preconditioners
- Use (approximate) elimination techniques so that larger (e.g., time-dependent) problems can be solved

Outline

Constrained optimization of smooth functions

- Consider constrained optimization problems of the form

$$\min_x \ f(x)$$
$$\text{s.t. } c(x) \leq 0$$

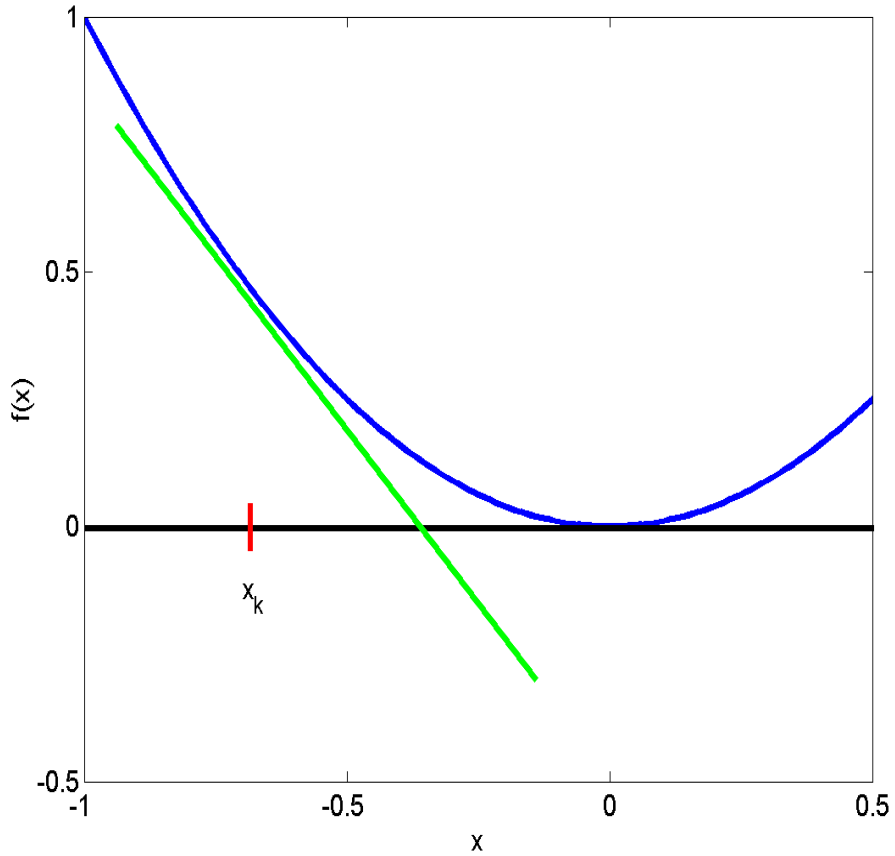
  where f and c are *smooth* (equality constraints OK, too)

- At x_k, solve the SLP/SQP subproblem

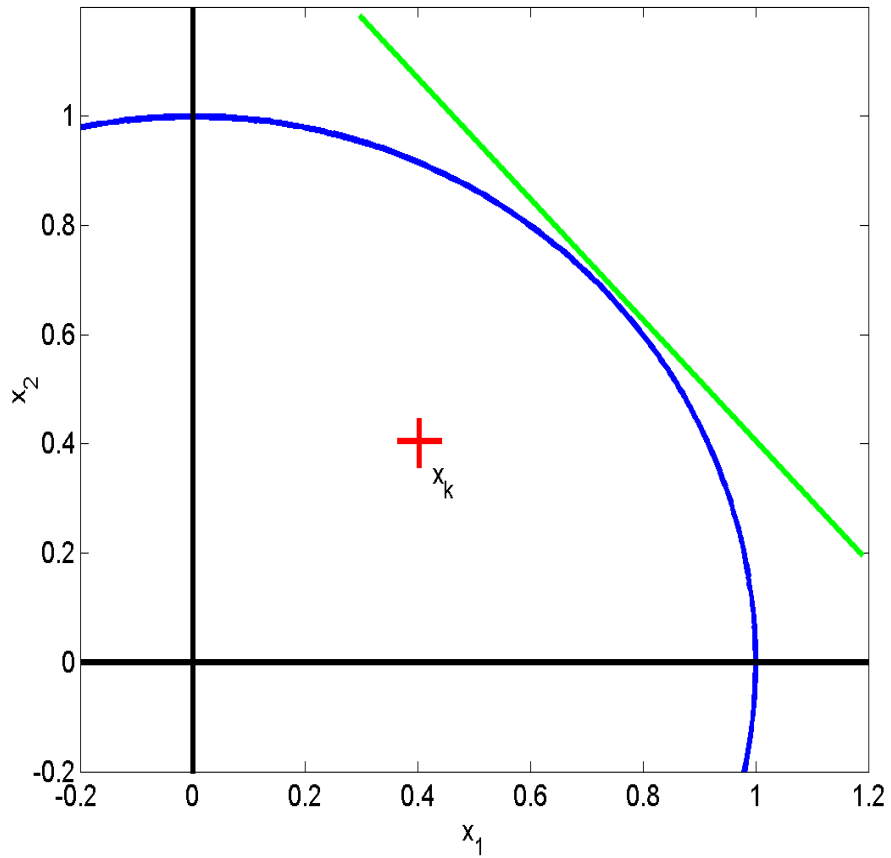
$$\min_d \ f_k + \nabla f_k^T d + \tfrac{1}{2} d^T H_k d$$
$$\text{s.t. } c_k + \nabla c_k^T d \leq 0, \quad \|d\| \leq \Delta_k$$

  to compute the search direction d_k

SQP illustration: Objective model



SQP illustration: Constraint model



Practicalities

- Since the linearized constraints may be inconsistent, we solve

$$\min_d \ \rho(f_k + \nabla f_k^T d) + \sum s^i + \tfrac{1}{2} d^T H_k d$$
$$\text{s.t. } c_k + \nabla c_k^T d \leq s, \quad s \geq 0,$$

  where $\rho > 0$ is a *penalty parameter*

- We perform a line search on the penalty function

$$\phi(x; \rho) \triangleq \rho f(x) + \sum \max\{0, c^i(x)\}$$

  to promote global convergence

Line search

- A model of the penalty function is given by

$$q_k(d;\rho) \triangleq \rho(f_k + \nabla f_k^T d) + \sum \max\{0, c_k^i + \nabla c_k^{i\,T} d\} + \tfrac{1}{2} d^T H_k d$$

- Solving the SQP subproblem is equivalent to minimizing $q_k(d;\rho)$
- The reduction in $q_k(d;\rho)$ yielded by d_k is

$$\Delta q_k(d_k;\rho) \triangleq q_k(0;\rho) - q_k(d_k;\rho)$$

- We impose the sufficient decrease condition

$$\phi(x_k + \alpha_k d_k;\rho) \leq \phi(x_k;\rho) - \eta\alpha_k \Delta q_k(d_k;\rho)$$

Penalty-SQP method

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

- Solve the SQP subproblem

$$\min_d \ \rho(f_k + \nabla f_k^T d) + \sum s^i + \tfrac{1}{2} d^T H_k d$$
$$\text{s.t. } c_k + \nabla c_k^T d \leq s, \quad s \geq 0$$

  or, equivalently, solve

$$\min_d \ q_k(d; \rho) \triangleq \rho(f_k + \nabla f_k^T d) + \sum \max\{0, c_k^i + \nabla {c_k^i}^T d\} + \tfrac{1}{2} d^T H_k d$$

  to compute d_k

- Backtrack from $\alpha_k = 1$ to satisfy

$$\phi(x_k + \alpha_k d_k; \rho) \leq \phi(x_k; \rho) - \eta \alpha_k \Delta q_k(d_k; \rho)$$

- Update $x_{k+1} \leftarrow x_k + \alpha_k d_k$

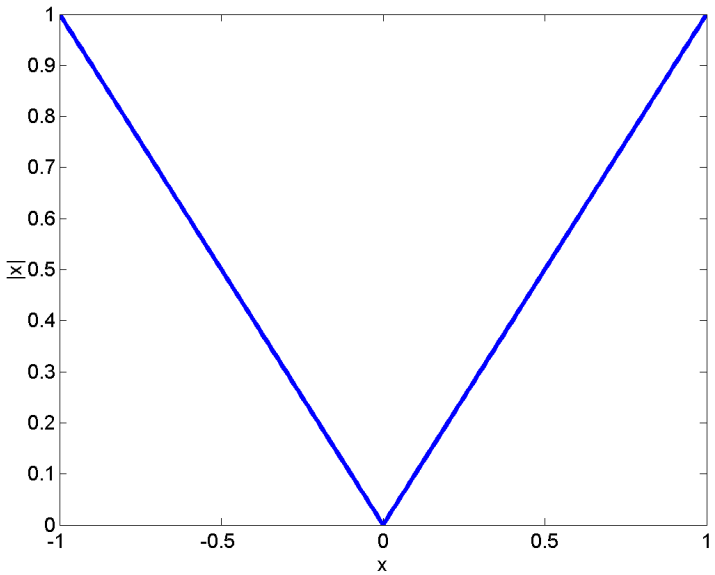
Outline

Unconstrained optimization of nonsmooth functions

- Consider the unconstrained optimization problem

$$\min_x \ f(x)$$

  where f may be nonsmooth (but is at least locally Lipschitz)
- The prototypical example is the absolute value function:



Clarke subdifferential

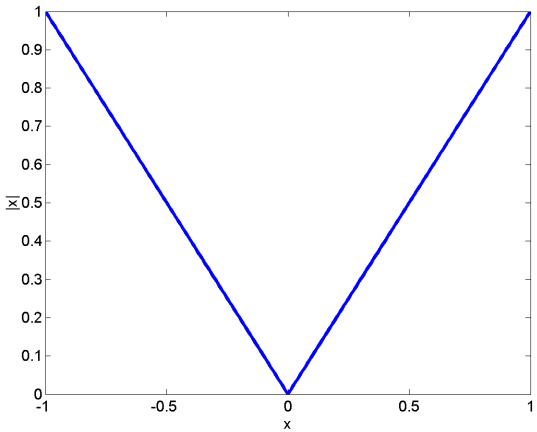
- Suppose f is differentiable over an open dense set $\mathcal{D}$
- Let
$$\mathbb{B}(x', \epsilon) \triangleq \{x \mid \|x - x'\| \leq \epsilon\}$$
- The Clarke subdifferential is
$$\bar{\partial} f(x') = \bigcap_{\epsilon > 0} \text{cl conv } \nabla f(\mathbb{B}(x', \epsilon) \cap \mathcal{D})$$
- A point x' is called Clarke stationary if $0 \in \bar{\partial} f(x')$

ϵ-stationarity

- The Clarke ϵ-subdifferential is given by

$$\bar{\partial} f(x', \epsilon) = \text{cl conv } \bar{\partial} f(\mathbb{B}(x', \epsilon) \cap \mathcal{D})$$

- A point x' is called ϵ-stationary if $0 \in \bar{\partial} f(x', \epsilon)$



- ... find ϵ-stationary point, reduce ϵ, find ϵ-stationary point,...

Gradient sampling: Robust steepest descent

- (Burke, Lewis, Overton, 2005)
- We restrict iterates to the open dense set $\mathcal{D}$
- Ideally, at x_k, for a given ϵ we would solve

$$\min_d \; f_k + \max_{x \in \mathcal{B}_k} \{\nabla f(x)^T d\} + \tfrac{1}{2} d^T H_k d$$

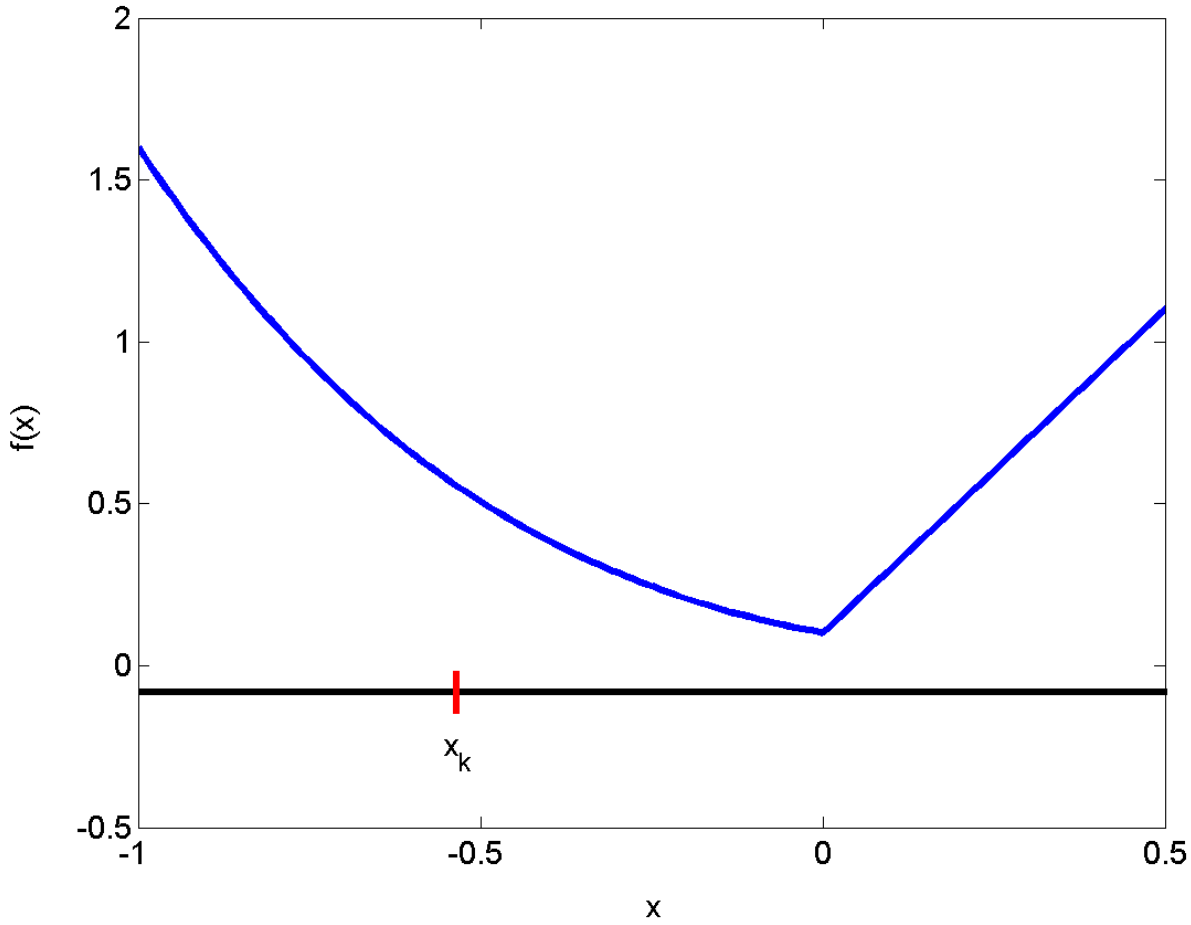
where $\mathcal{B}_k = \mathbb{B}(x_k, \epsilon) \cap \mathcal{D}$

- However, we can only approximate this step by solving

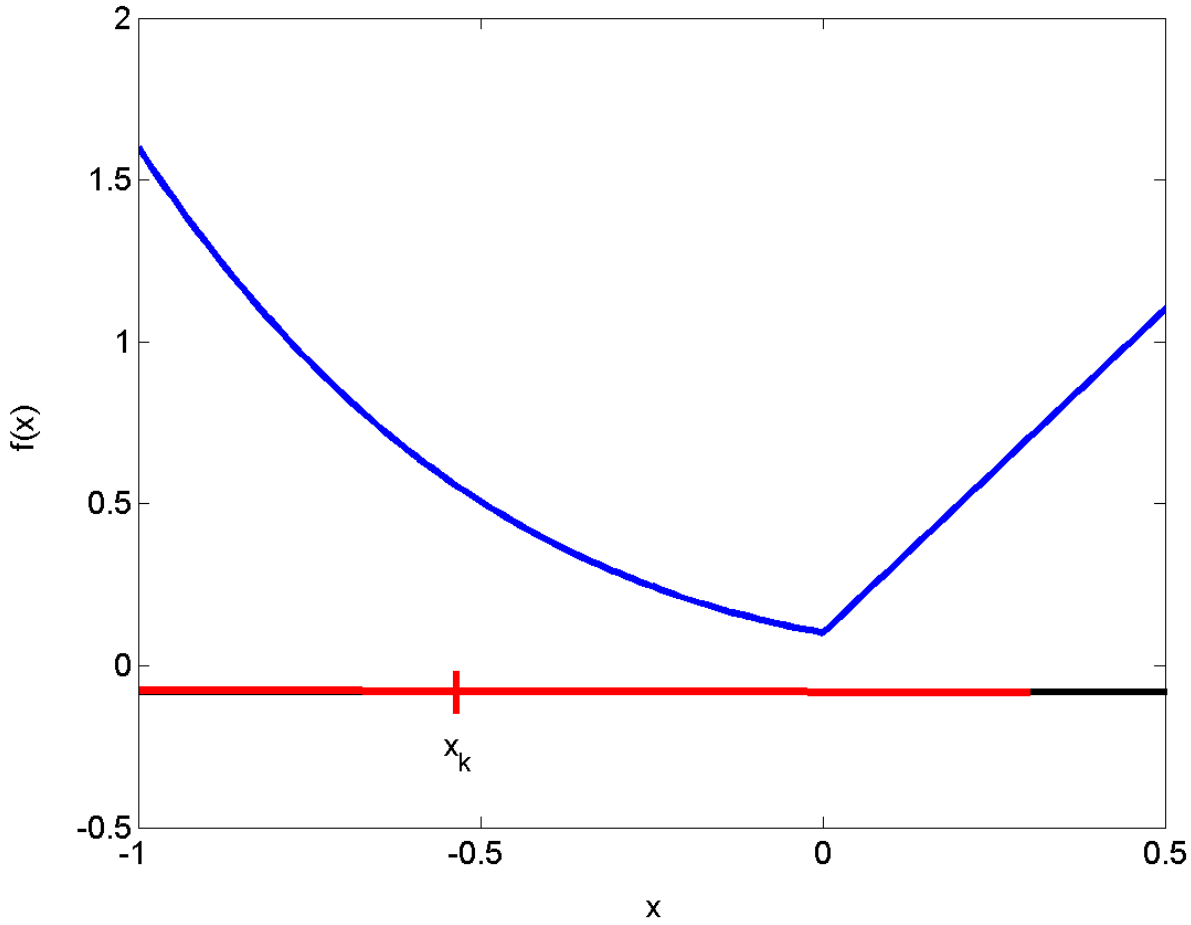
$$\min_d \; f_k + \max_{x \in \mathcal{B}_k} \{\nabla f(x)^T d\} + \tfrac{1}{2} d^T H_k d$$

where $\mathcal{B}_k = \{x_k, x_{k1}, \ldots, x_{kp}\} \subset \mathbb{B}(x_k, \epsilon) \cap \mathcal{D}$

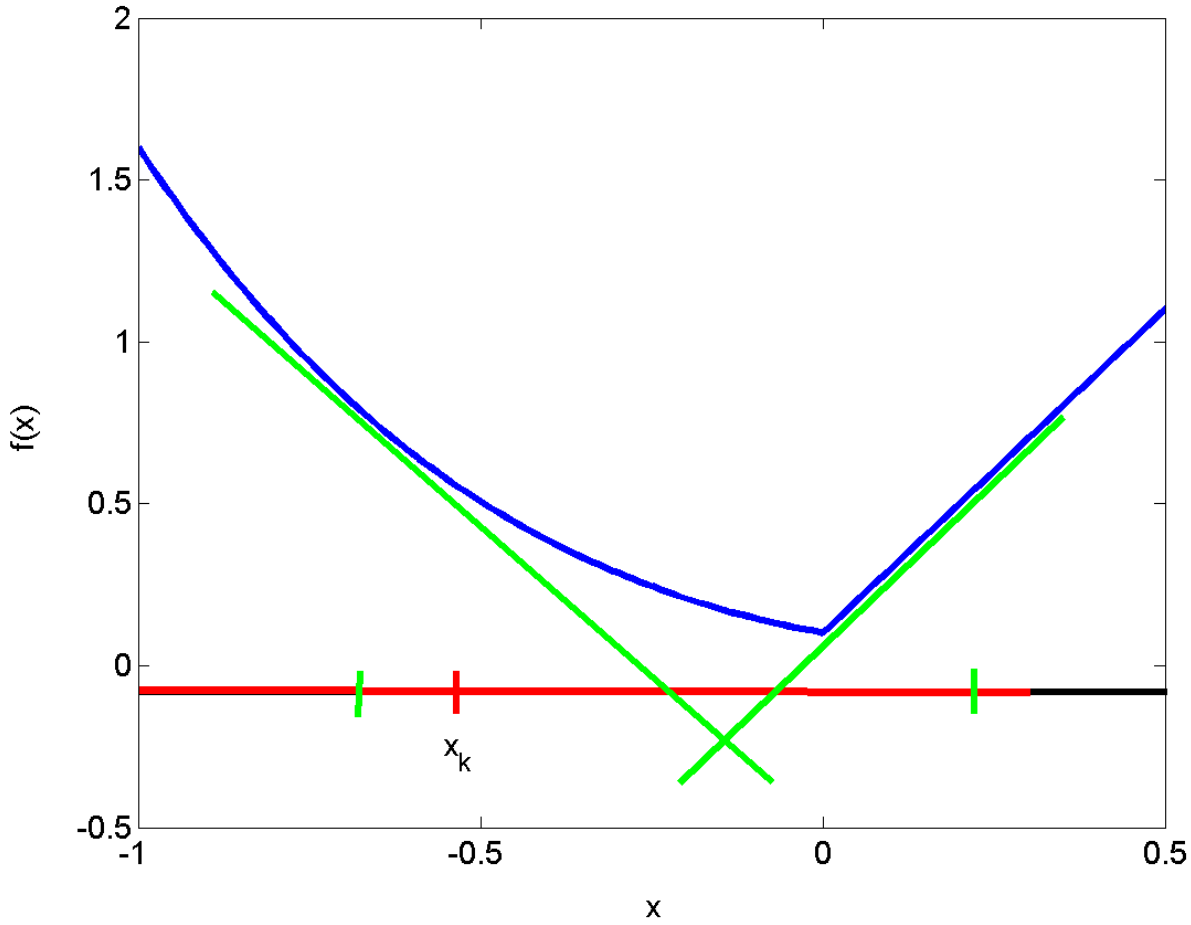
GS illustration: Objective model



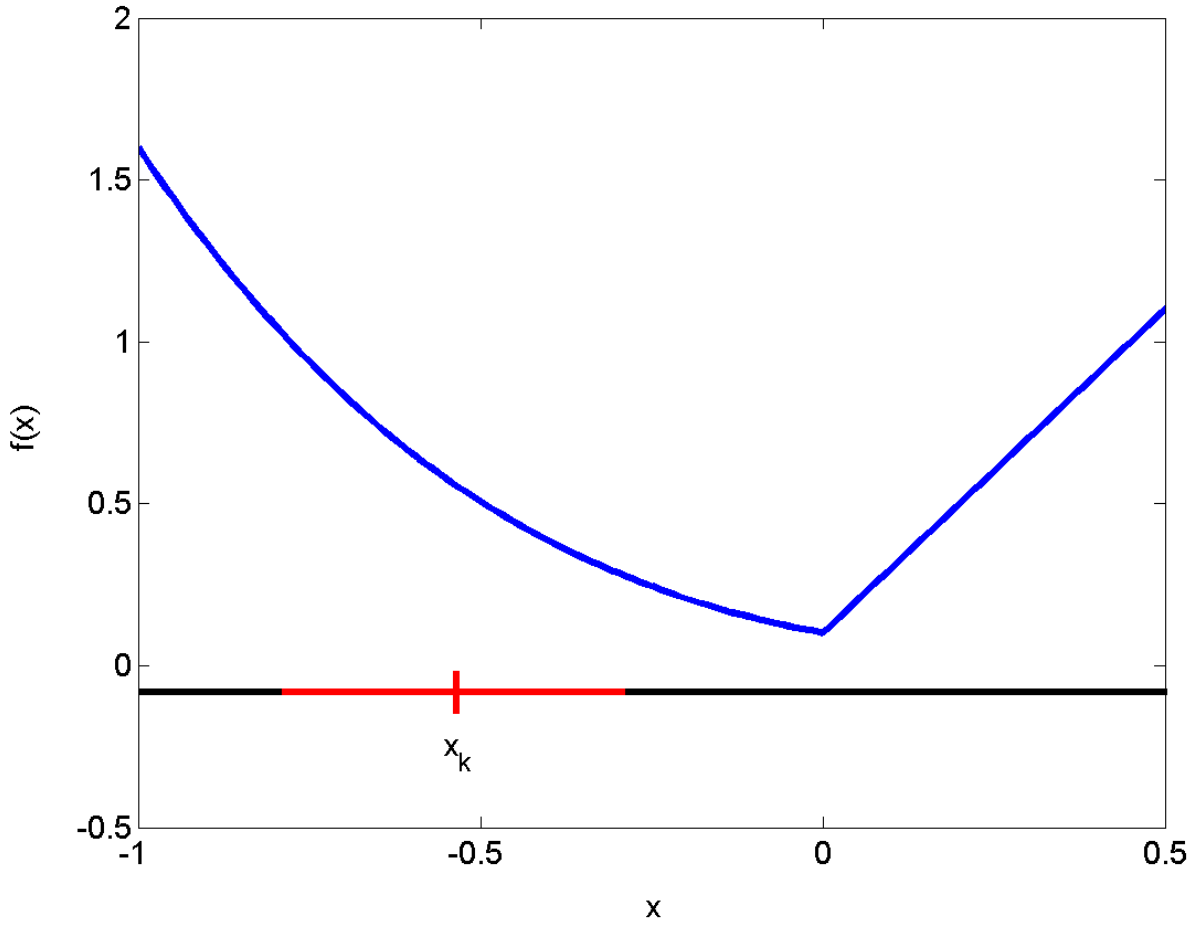
GS illustration: Objective model



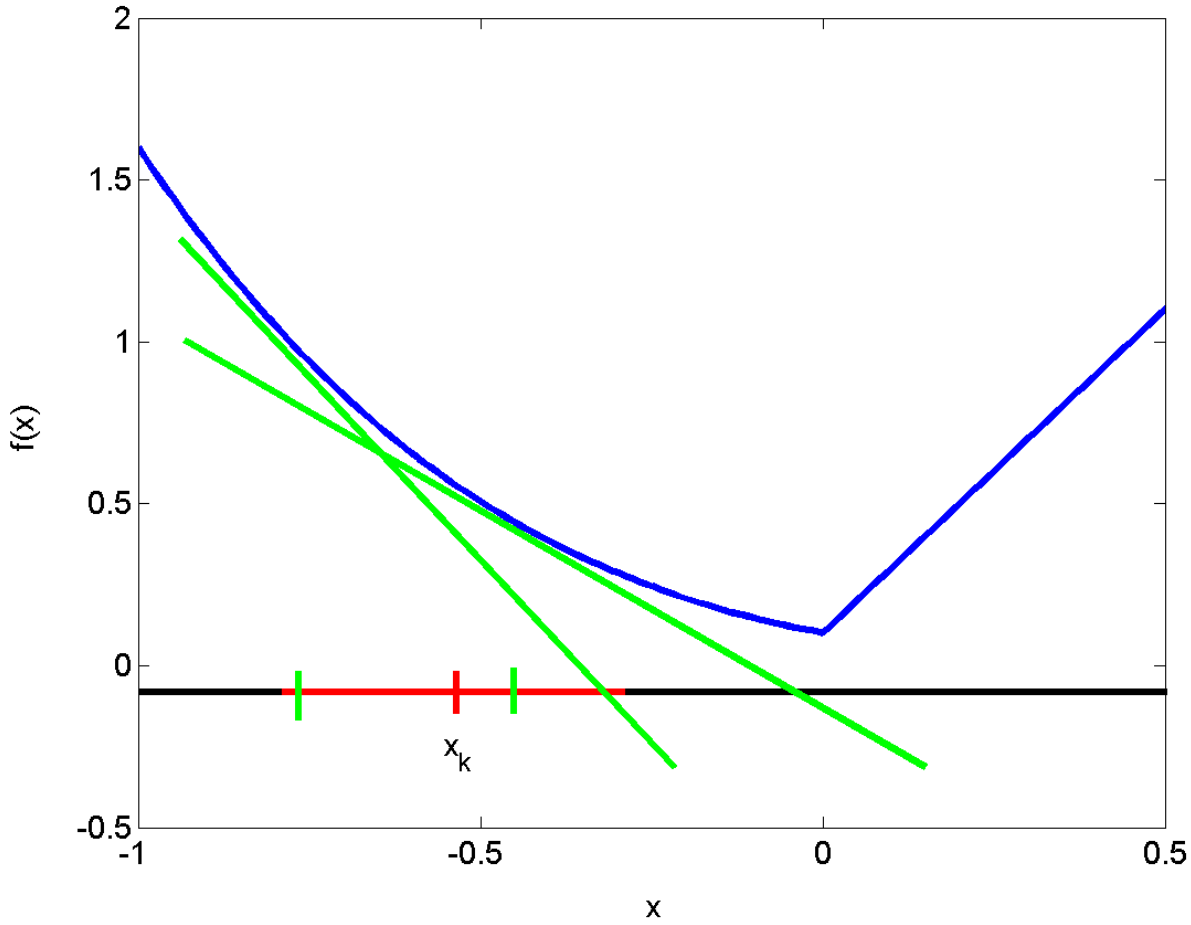
GS illustration: Objective model



GS illustration: Objective model



GS illustration: Objective model



GS method

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

- Sample points $\{x_{k1}, \ldots, x_{kp}\}$ in $\mathbb{B}(x_k, \epsilon) \cap \mathcal{D}$
- Solve the GS subproblem

$$\min_d \; f_k + \max_{x \in \mathcal{B}_k} \{\nabla f(x)^T d\} + \tfrac{1}{2} d^T H_k d$$

  to compute d_k
- Backtrack from $\alpha_k = 1$ to satisfy

$$f(x_k + \alpha_k d_k) \leq f(x_k) - \eta \alpha_k \|d_k\|^2$$

- Update $x_{k+1} \approx x_k + \alpha_k d_k$ (to ensure $x_{k+1} \in \mathcal{D}$)
- If $\|d_k\| \leq \epsilon$, then reduce ϵ

Outline

Constrained optimization of nonsmooth functions

- Consider constrained optimization problems of the form

$$\min_x \ f(x)$$
$$\text{s.t. } c(x) \leq 0$$

where f and c may be *nonsmooth* (equality constraints OK, too)

- We may consider solving

$$\min_x \ \phi(x;\rho) \triangleq \rho f(x) + \sum \max\{0, c^i(x)\}$$

or even

$$\min_x \ \varphi(x;\rho) \triangleq \rho f(x) + \max_i \max\{0, c^i(x)\}$$

but this makes me... :-(

SQP and GS

- The SQP subproblem is

$$\begin{aligned} \min_d \ & \rho z + \sum s^i + \tfrac{1}{2} d^T H_k d \\ \text{s.t. } & f_k + \nabla f_k^T d \leq z \\ & c_k + \nabla c_k^T d \leq s, \ s \geq 0 \end{aligned}$$

- The GS subproblem is

$$\begin{aligned} \min_d \ & z + \tfrac{1}{2} d^T H_k d \\ \text{s.t. } & f_k + \nabla f(x)^T d \leq z, \ \forall \ x \in \mathcal{B}_k \end{aligned}$$

SQP-GS

- The SQP-GS subproblem is

$$\begin{aligned} \min_{d,z,s} \ & \rho z + \sum s^i + \tfrac{1}{2} d^T H_k d \\ \text{s.t. } & f_k + \nabla f(x)^T d \leq z, \ \forall\, x \in \mathcal{B}_k^0 \\ & c_k^i + \nabla c^i(x)^T d \leq s^i, \ s^i \geq 0, \ \forall\, x \in \mathcal{B}_k^i, \ i = 1, \ldots, m \end{aligned}$$

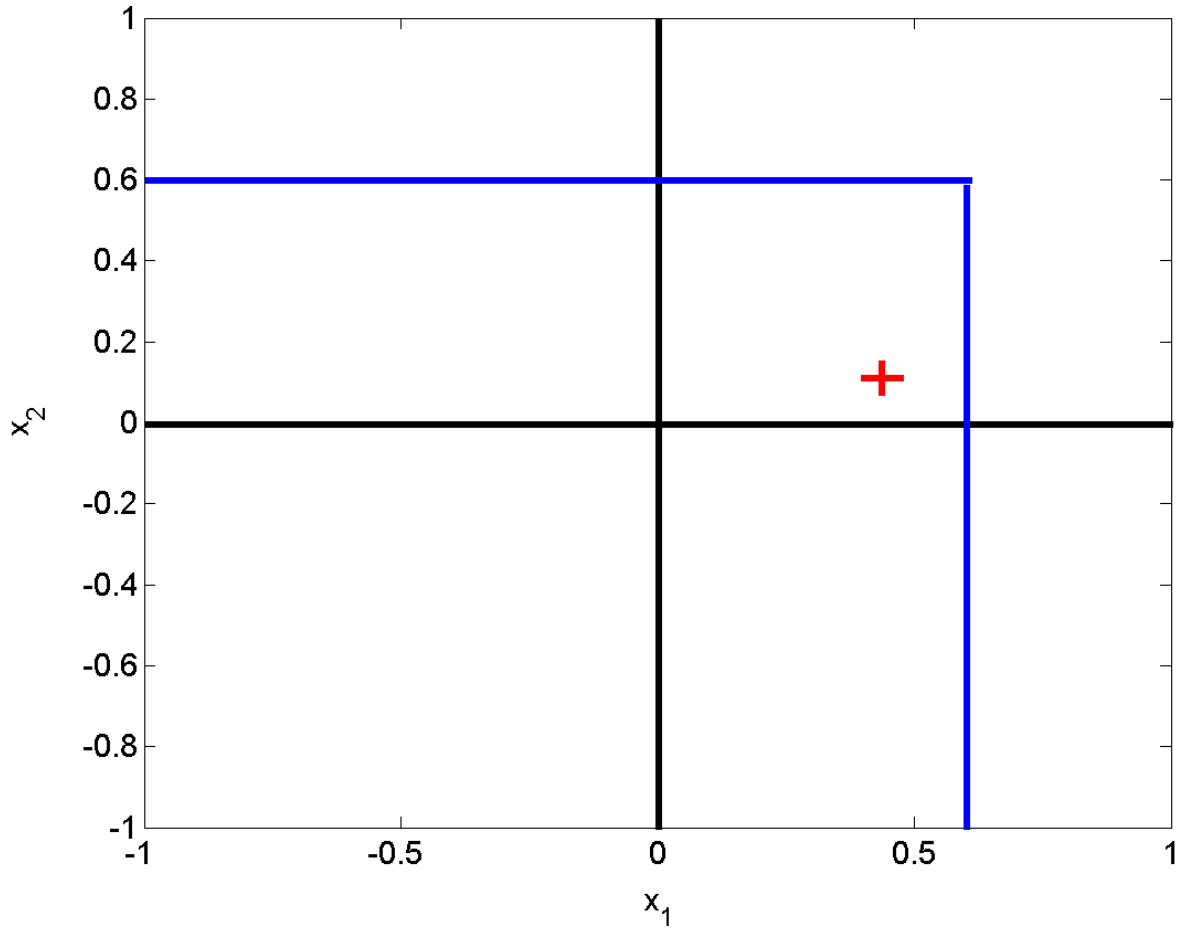
  where $\mathcal{B}_k^i = \{x_k, x_{k1}^i, \ldots, x_{kp}^i\} \subset \mathbb{B}(x_k, \epsilon)$ for $i = 0, \ldots, m$

- This is equivalent to

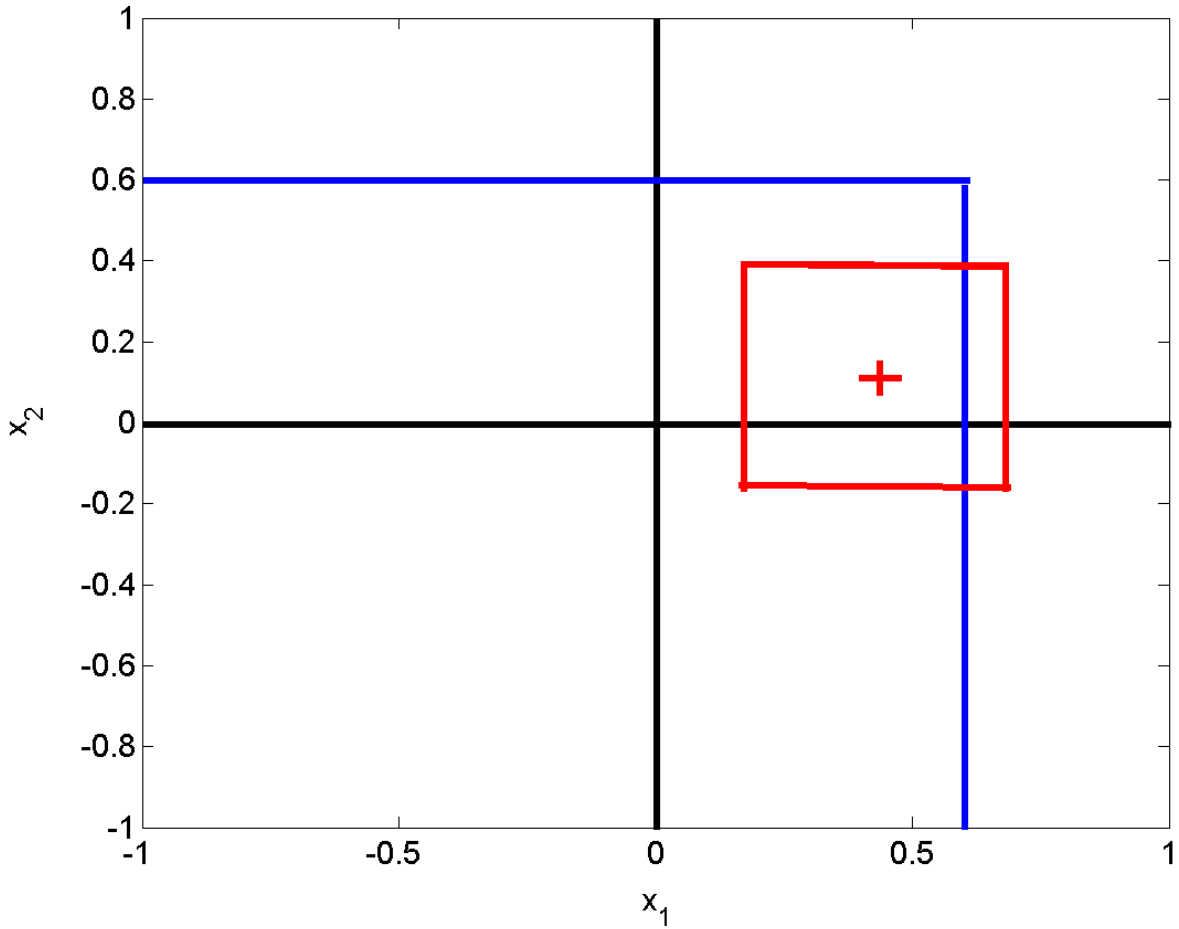
$$\min_d \ \rho \max_{x \in \mathcal{B}_k^0} (f_k + \nabla f(x)^T d) + \sum \max_{x \in \mathcal{B}_k^i} \max\{0, c_k^i + \nabla c^i(x)^T d, 0\} + \tfrac{1}{2} d^T H_k d$$

  i.e., $\min_d q_k(d; \rho)$, where now $q_k(d; \rho)$ is a *robust* model of $\phi(x; \rho)$

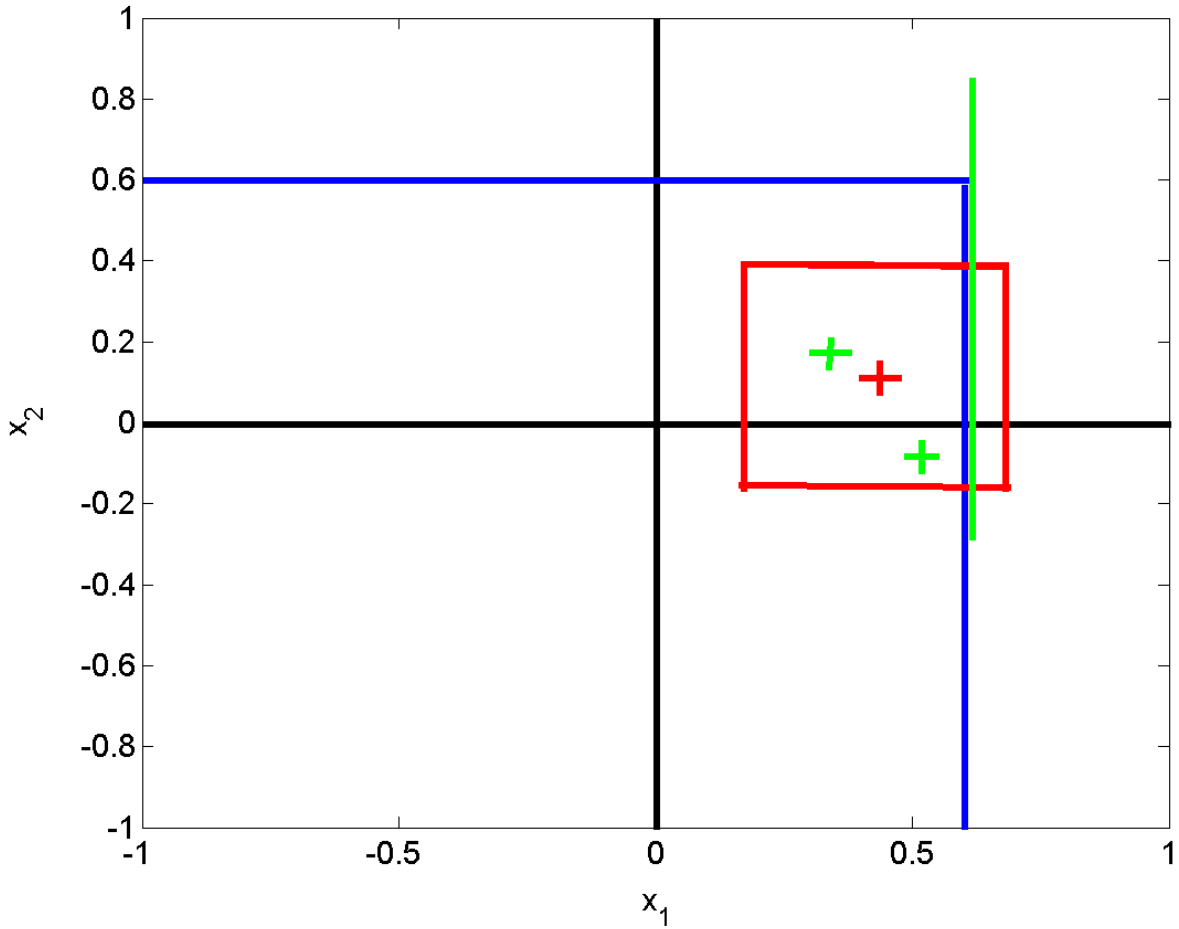
SQP-GS illustration: Constraint model



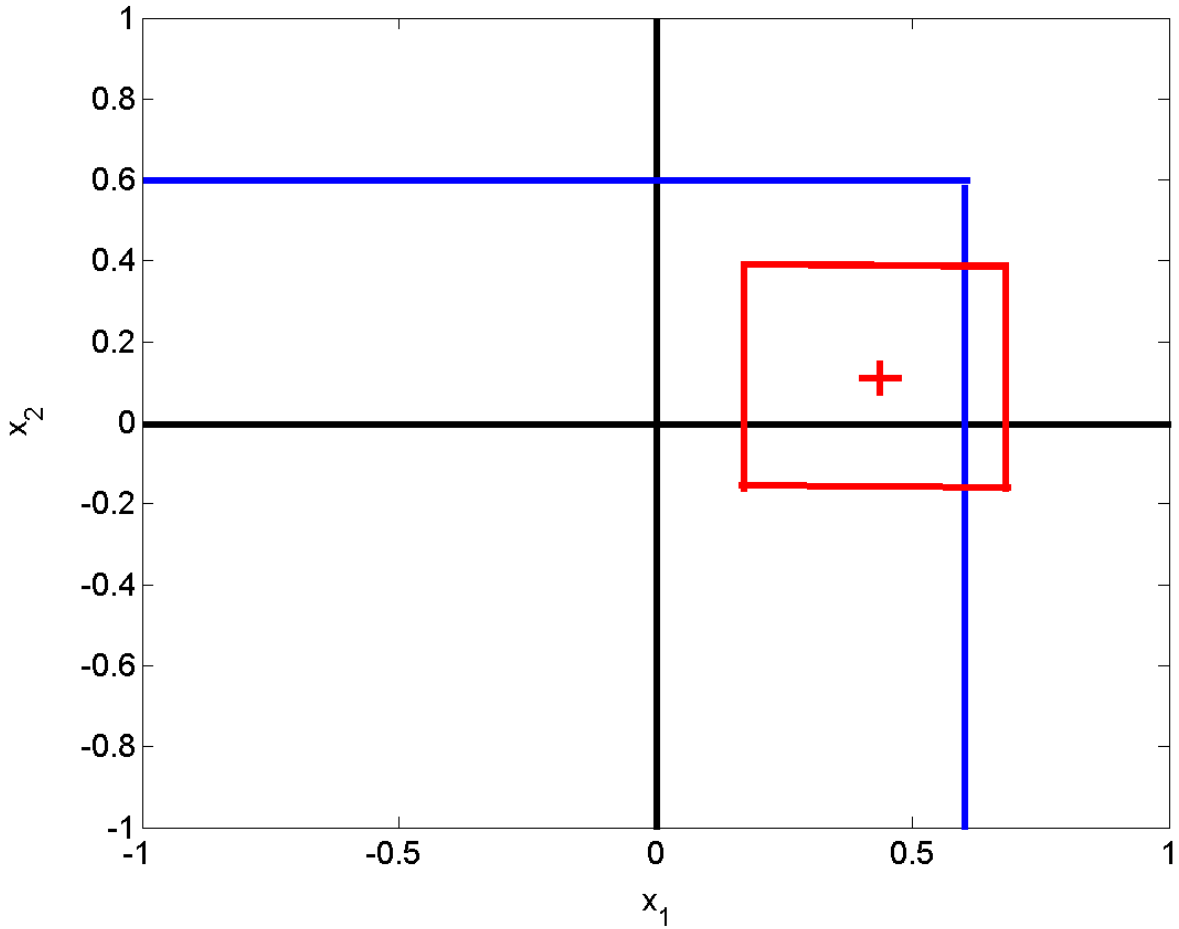
SQP-GS illustration: Constraint model



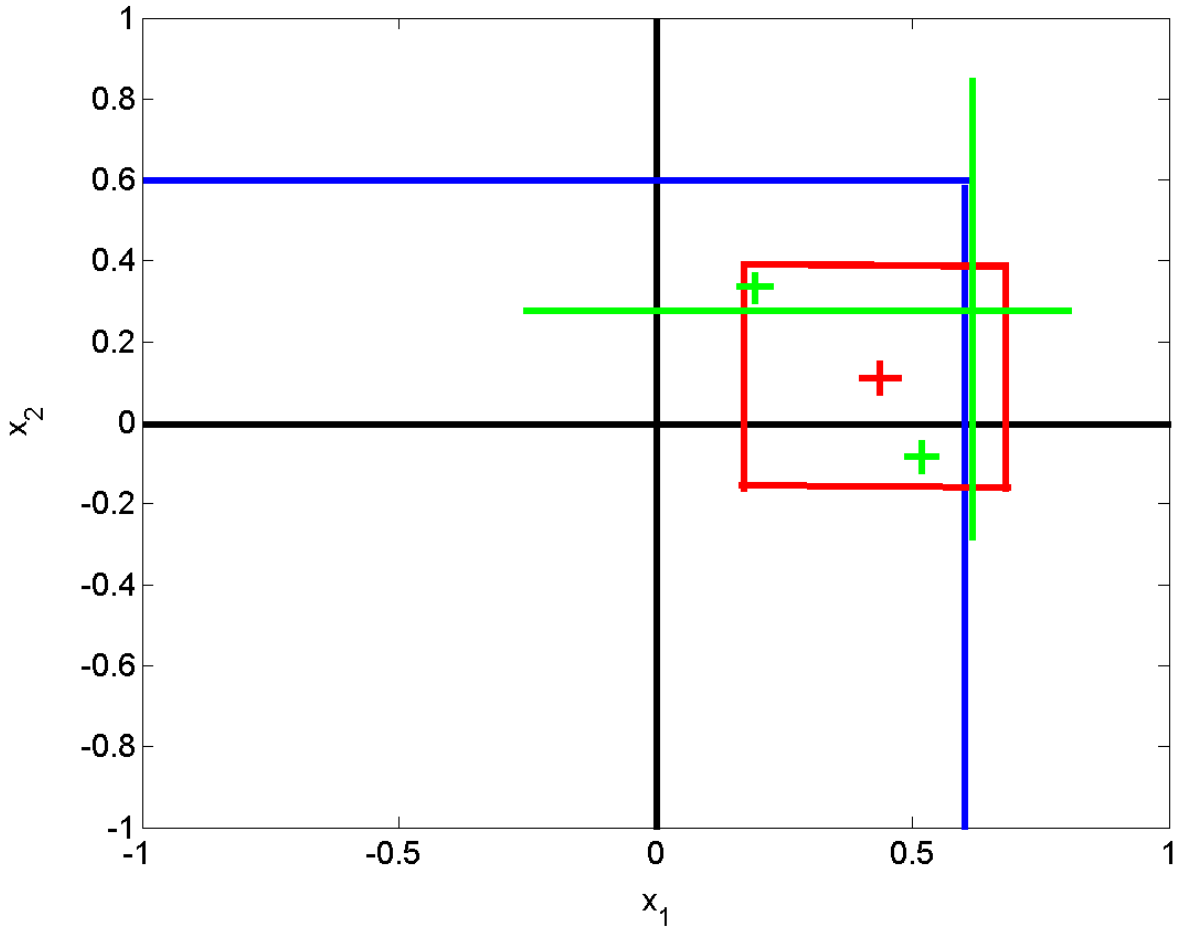
SQP-GS illustration: Constraint model



SQP-GS illustration: Constraint model



SQP-GS illustration: Constraint model



SQP-GS method

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

- Sample points $\{x_{k1}^i, \ldots, x_{kp}^i\}$ in $\mathbb{B}(x_k, \epsilon) \in \mathcal{D}^i$ for $i = 0, \ldots, m$
- Solve the SQP-GS subproblem

$$\begin{aligned} \min_{d,z,s} \ & \rho z + \sum s^i + \tfrac{1}{2} d^T H_k d \\ \text{s.t. } & f_k + \nabla f(x)^T d \leq z, \ \forall\, x \in \mathcal{B}_k^0 \\ & c_k^i + \nabla c^i(x)^T d \leq s^i, \ s^i \geq 0, \ \forall\, x \in \mathcal{B}_k^i, \ i = 1, \ldots, m \end{aligned}$$

  to compute d_k

- Backtrack from $\alpha_k = 1$ to satisfy

$$\phi(x_k + \alpha_k d_k; \rho) \leq \phi(x_k; \rho) - \eta \alpha_k \Delta q_k(d_k; \rho)$$

- Update $x_{k+1} \approx x_k + \alpha_k d_k$ (to ensure $x_{k+1} \in \cap_i \mathcal{D}^i$)
- If $\Delta q_k(d_k; \rho) \leq \epsilon$, then reduce ϵ

Global convergence

- Assumption 1: The functions f and c^i, $i = 1, \ldots, m$, are locally Lipschitz and continuously differentiable on open dense subsets of $\mathbb{R}^n$
- Assumption 2: The sequence of iterates and sample points are contained in a convex set over which the functions f and c^i, $i = 1, \ldots, m$, and their first derivatives are bounded
- Assumption 3: For universal constants $\overline{\xi} \geq \underline{\xi} > 0$, the Hessian matrices satisfy $\underline{\xi}\|d\|^2 \leq d^T H_k d \leq \overline{\xi}\|d\|^2$ for all $d \in \mathbb{R}^n$

Global convergence

- Lemma 1: $\Delta q_k(d_k;\rho) = 0$ if and only if x_k is ϵ-stationary
- Lemma 2: The one-sided directional derivative of the penalty function satisfies
$$\phi'(d_k;\rho) \leq d_k^T H_k d_k < 0$$
and so d_k is a descent direction for $\phi(x;\rho)$ at x_k
- Lemma 3: Suppose the sample size is $p \geq n+1$. If the current iterate x_k is sufficiently close to a stationary point x' of the penalty function $\phi(x;\rho)$, then there exists a nonempty open set of sample sets such that the solution to the SQP-GS subproblem d_k yields an arbitrarily small $\Delta q_k(d_k;\rho)$
  - Carathéodory's Theorem
- Theorem: With probability one, every cluster point of $\{x_k\}$ is feasible and stationary for $\phi(x;\rho)$

Outline

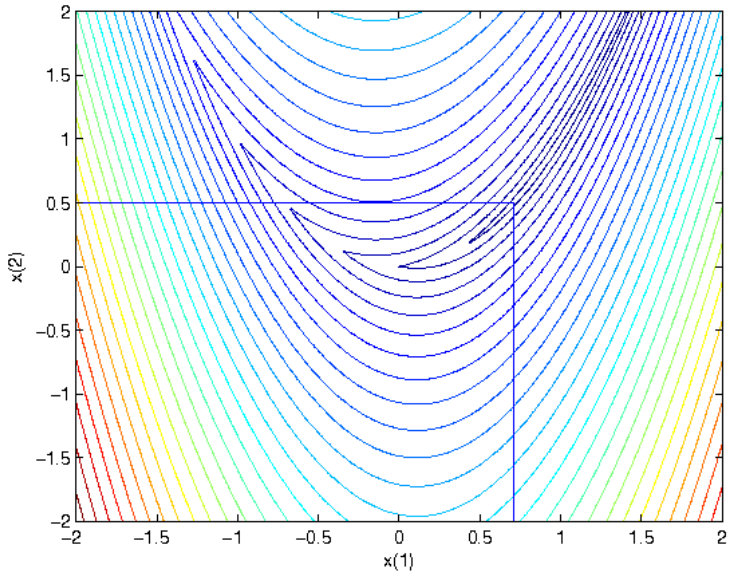
Implementation

- Prototype implementation in MATLAB (available soon?)
- QP subproblems solved with MOSEK
- BFGS approximations of Hessian of penalty function
  - (Lewis and Overton, 2009)
- ρ decreased conservatively

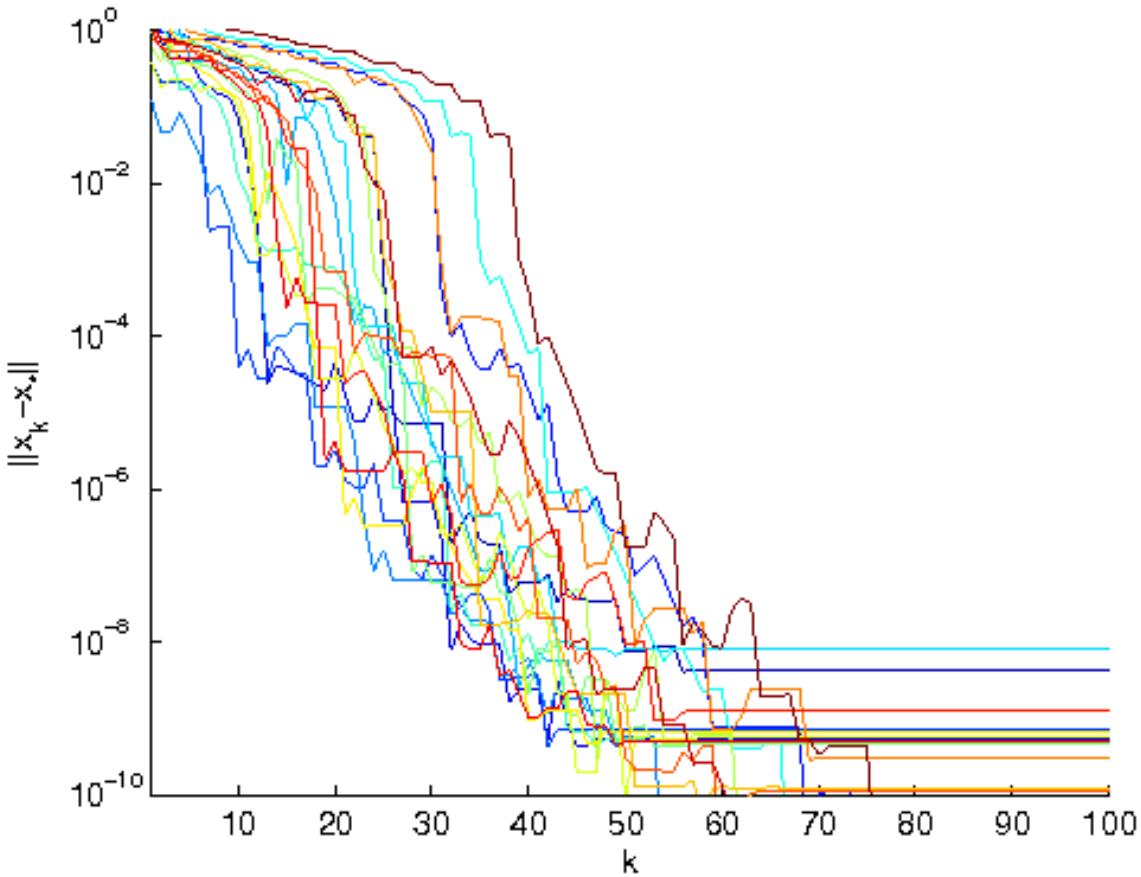
Example 1: Nonsmooth Rosenbrock

$$\min_x \; 8|x_1^2 - x_2| + (1 - x_1)^2$$

$$\text{s.t.} \;\; \max\{\sqrt{2}x_1, 2x_2\} \leq 1$$



Example 1: Nonsmooth Rosenbrock



Example 2: Entropy minimization

Find a $N \times N$ matrix X that solves

$$\min_X \ \ln\left(\prod_{j=1}^{K} \lambda_j(A \circ X^T X)\right)$$
$$\text{s.t. } \|X_j\| = 1, \ j = 1, \ldots, N$$

where $\lambda_j(M)$ denotes the jth largest eigenvalue of M, A is a real symmetric $N \times N$ matrix, $\circ$ denotes the Hadamard matrix product, and X_j denotes the jth column of X

Example 2: Entropy minimization

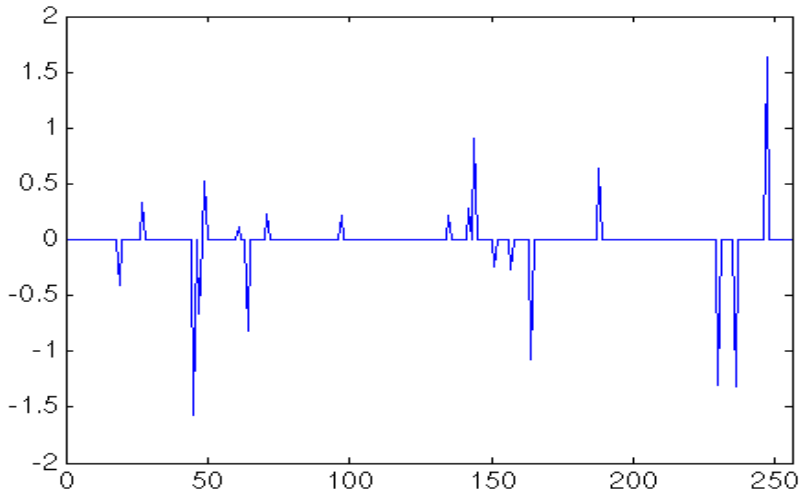
N	n	K	f (SQP-GS)	f (GS)
2	4	1	1.00000e+00	1.00000e+00
4	16	2	7.46296e-01	7.46286e-01
6	36	3	6.33589e-01	6.33477e-01
8	64	4	5.60165e-01	5.58820e-01
10	100	5	2.20724e-01	2.17193e-01
12	144	6	1.24820e-01	1.22226e-01
14	196	7	8.21835e-02	8.01010e-02
16	256	8	5.73762e-02	5.57912e-02

Example 3(a): Compressed sensing (ℓ_1 norm)

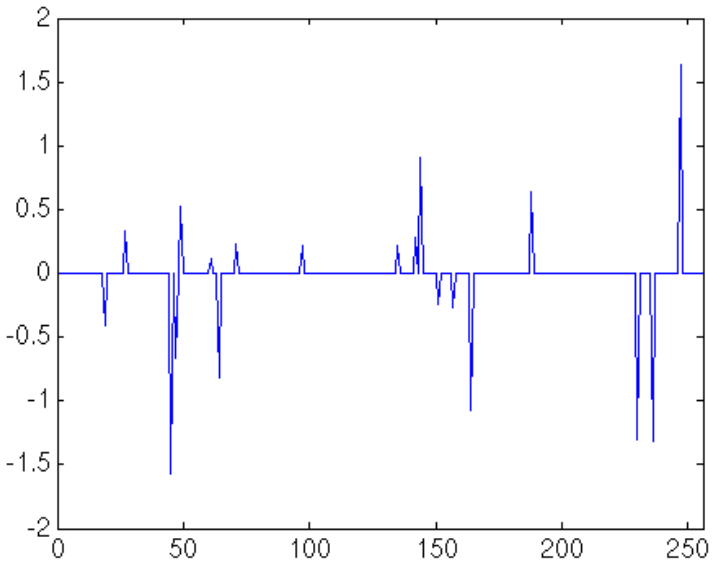
Recover a sparse signal by solving

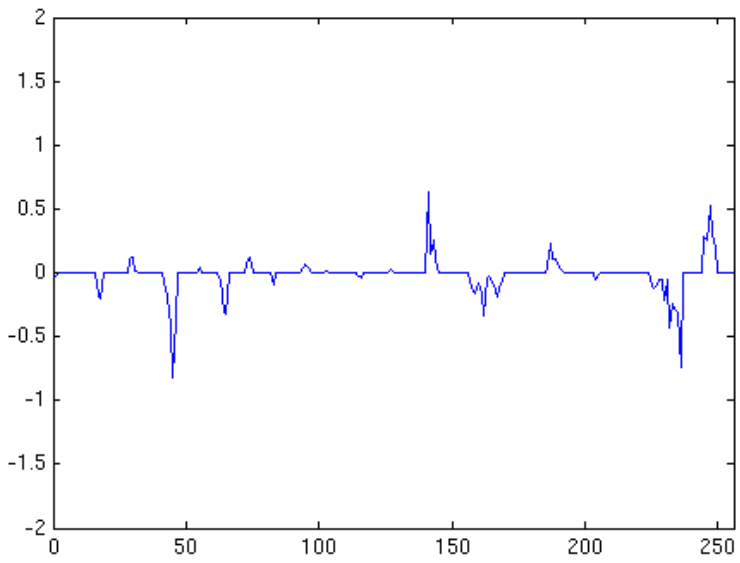
$$\min_x \ \|x\|_1$$
$$\text{s.t. } Ax = b$$

where A is a 64×256 submatrix of a discrete cosine transform (DCT) matrix



Example 3(a): Compressed sensing (ℓ_1 norm)



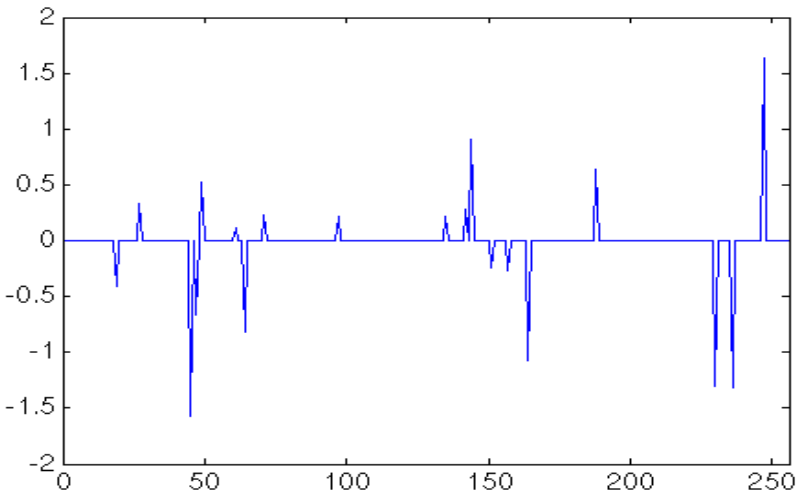


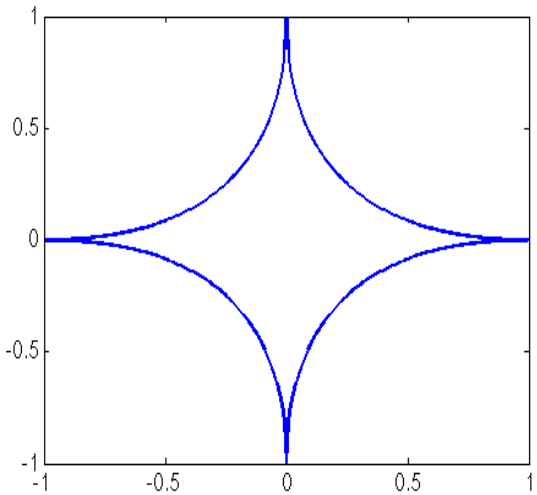
Example 3(b): Compressed sensing ($\ell_{0.5}$ norm)

Recover a sparse signal by solving

$$\min_x \ \|x\|_{0.5}$$
$$\text{s.t. } Ax = b$$

where A is a 64×256 submatrix of a discrete cosine transform (DCT) matrix





Example 3(b): Compressed sensing ($\ell_{0.5}$ norm)

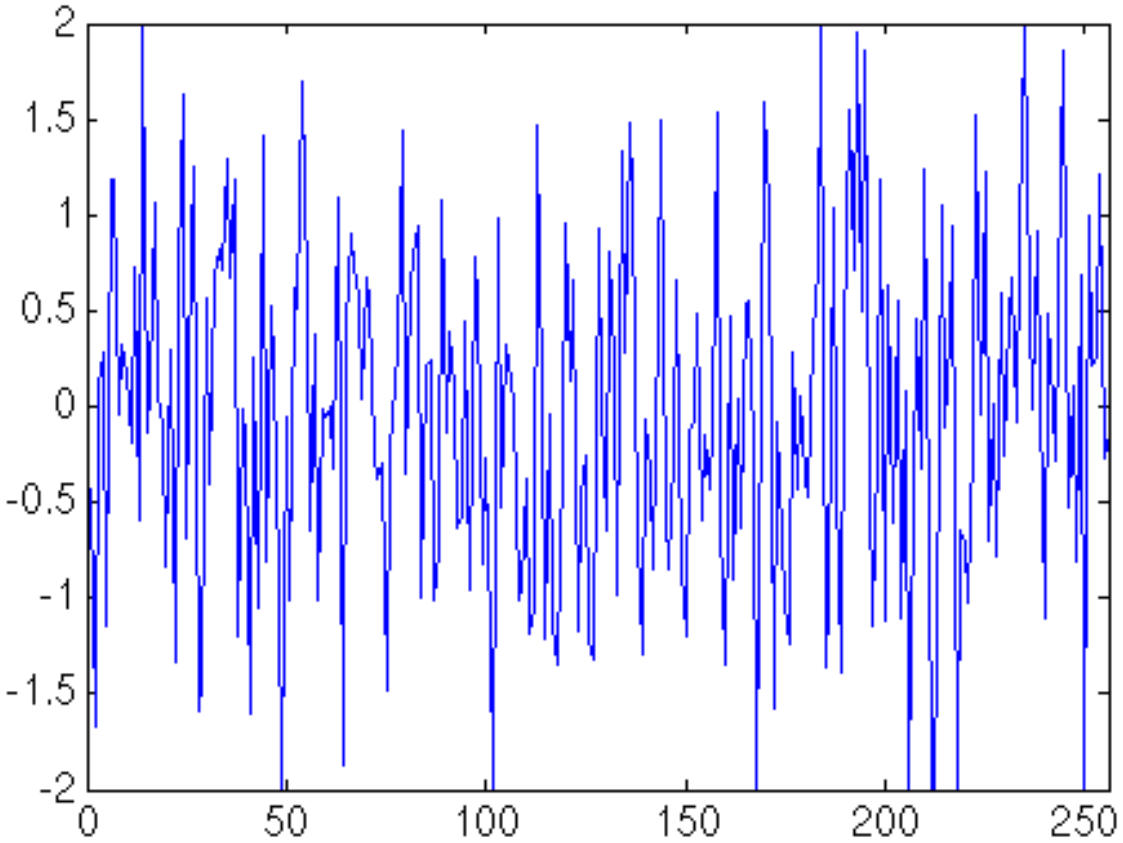


Figure: k = 1

Example 3(b): Compressed sensing ($\ell_{0.5}$ norm)

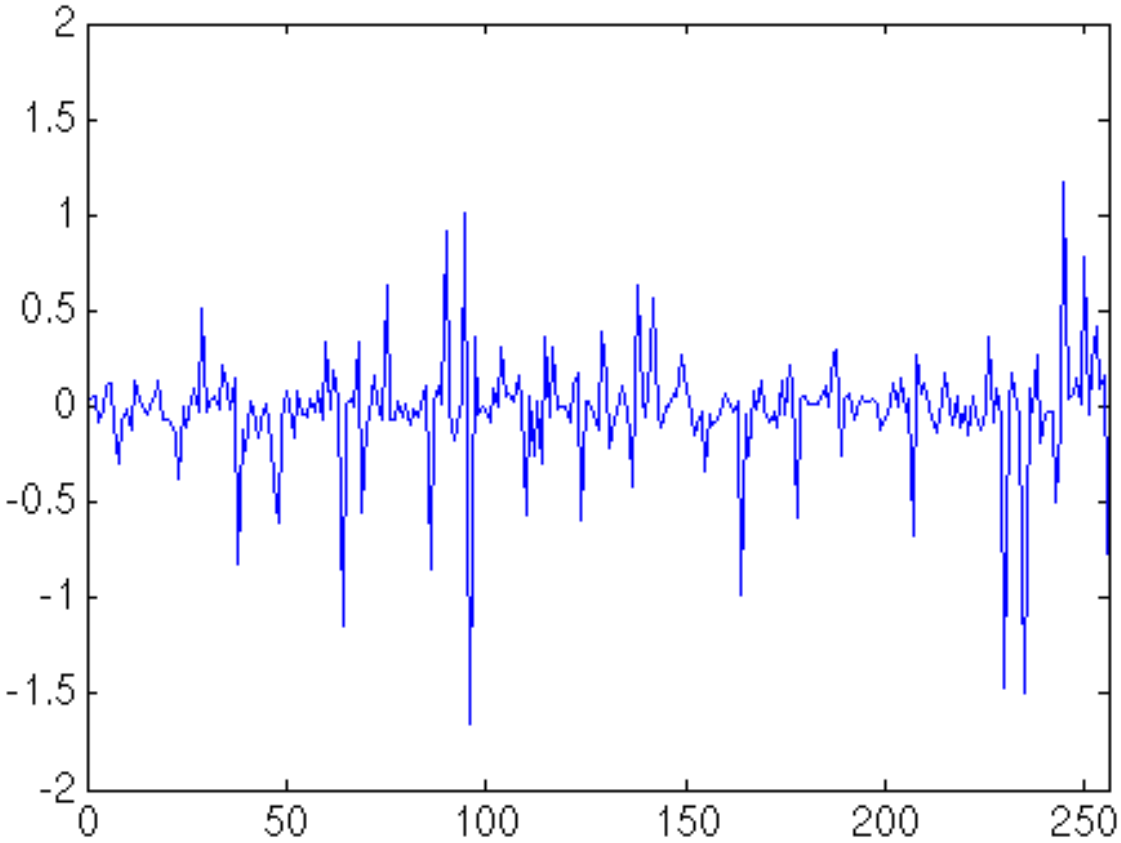


Figure: k = 10

Example 3(b): Compressed sensing ($\ell_{0.5}$ norm)

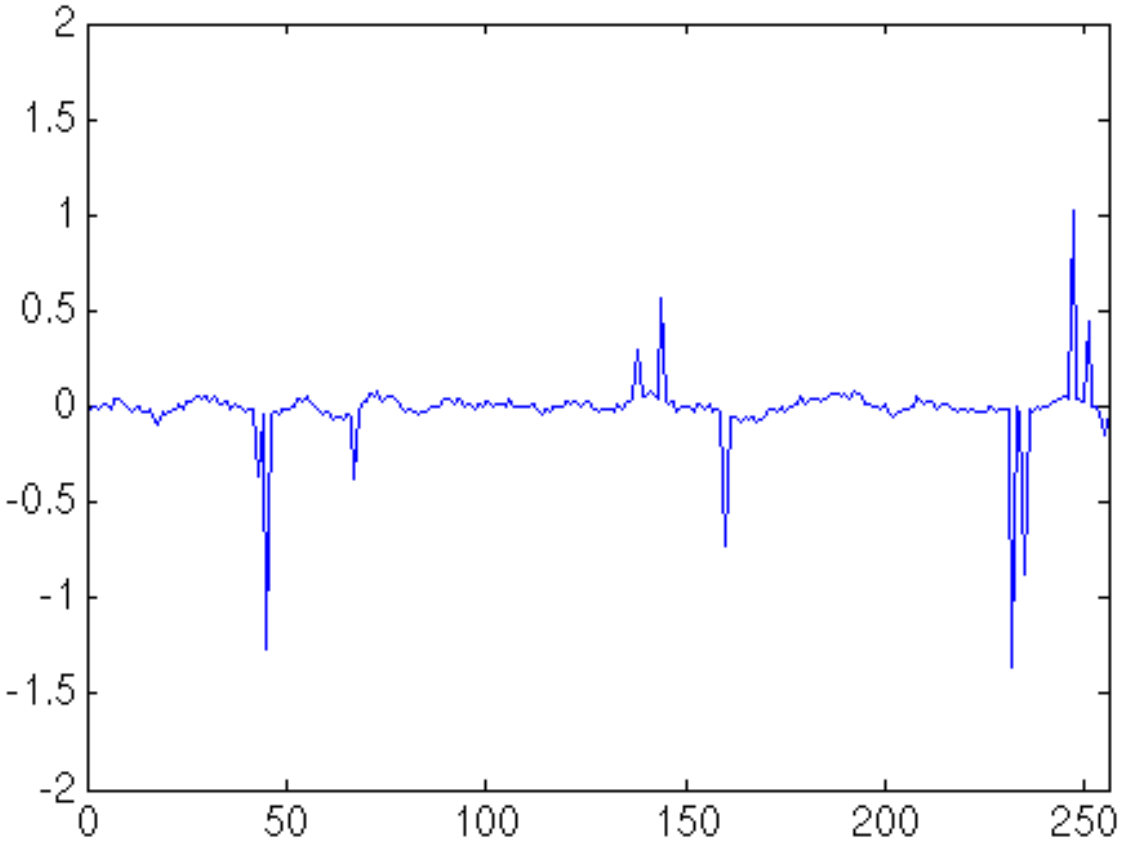


Figure: k = 25

Example 3(b): Compressed sensing ($\ell_{0.5}$ norm)

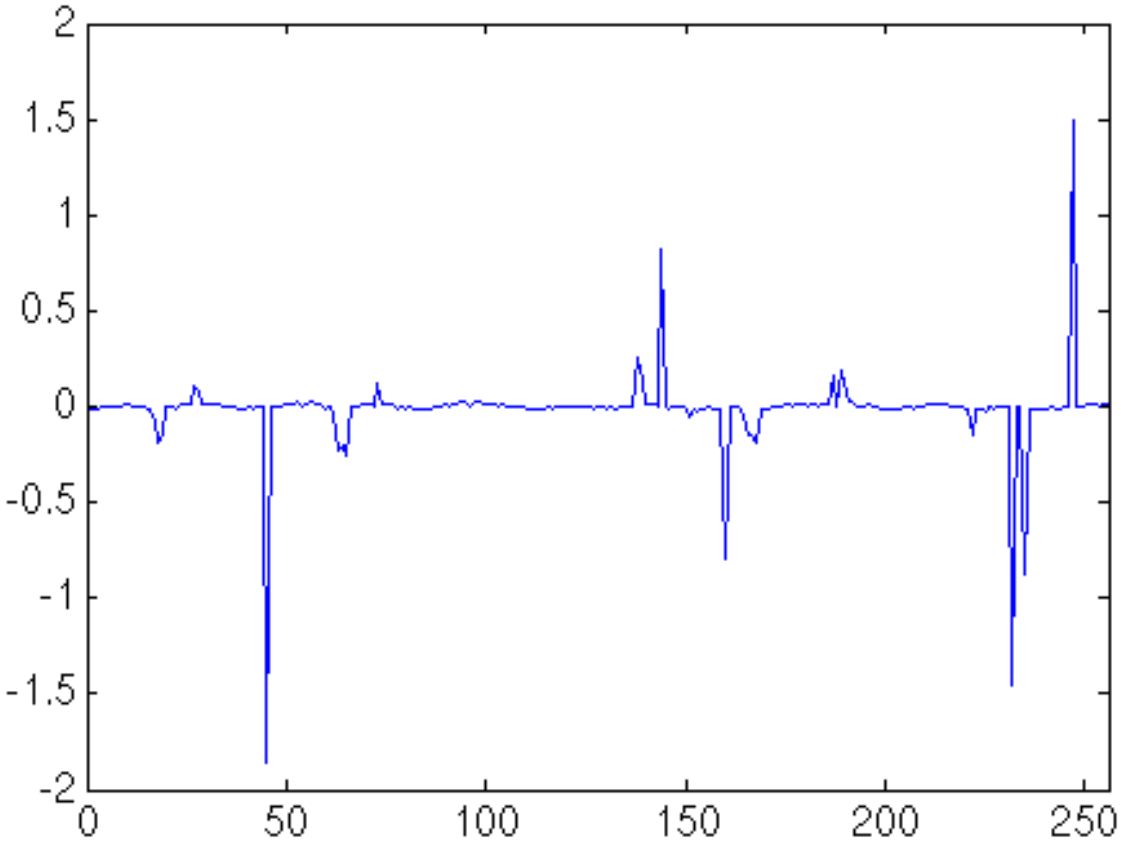


Figure: k = 50

Example 3(b): Compressed sensing ($\ell_{0.5}$ norm)

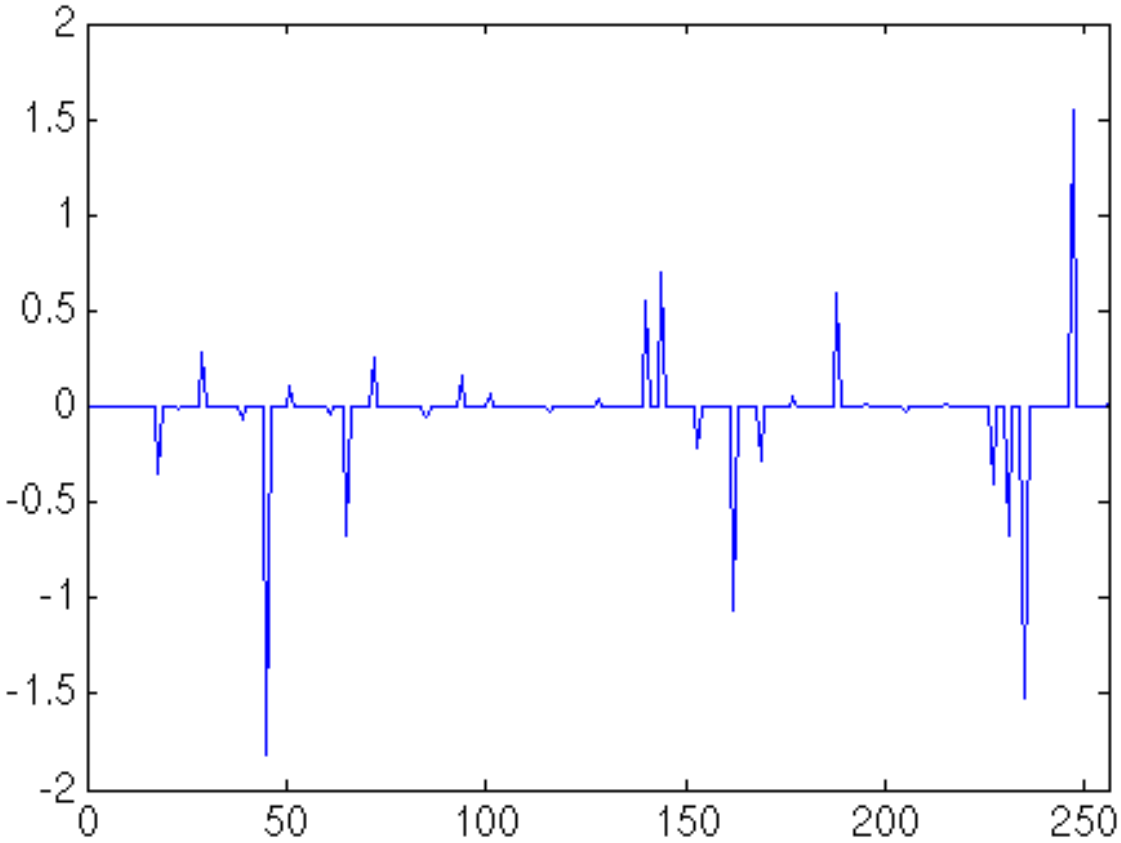
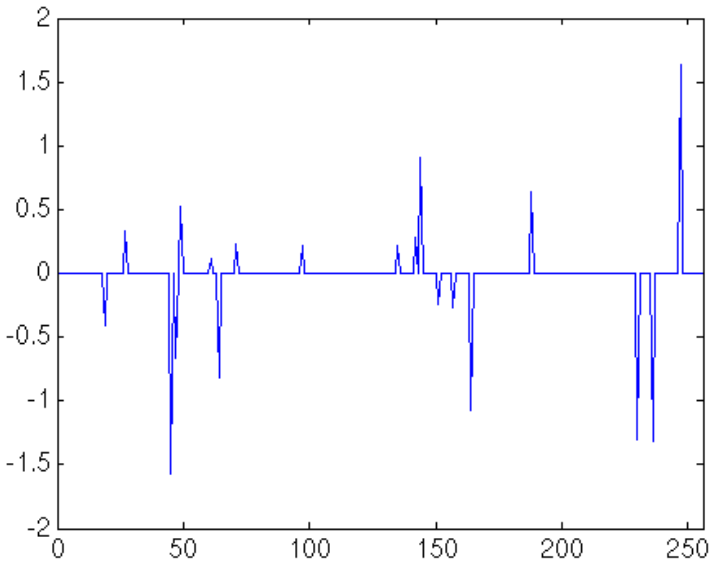
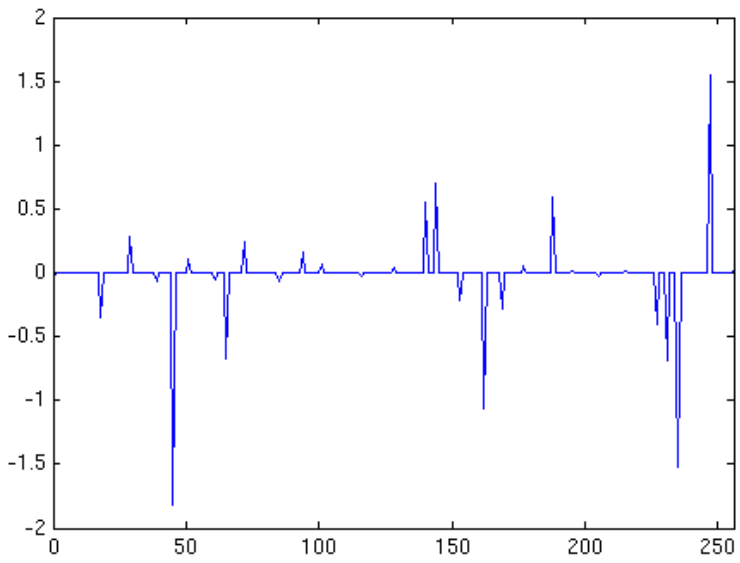


Figure: k = 200

Example 3(b): Compressed sensing ($\ell_{0.5}$ norm)





Outline

Summary

- We have presented a globally convergent algorithm for the solution of constrained, nonsmooth, and nonconvex optimization problems
- The algorithm follows a penalty-SQP framework and uses Gradient Sampling to make the search direction calculation robust
- Preliminary results are encouraging

Future work

- Tune updates for ϵ and ρ
- Allow for special handling of smooth/convex/linear functions
- Investigate SLP vs. SQP
- Extensions for particular applications; e.g., specialized sampling

Thanks!!