A Sequential Quadratic Programming Method for Nonsmooth Optimization

Frank E. Curtis

Modeling and Optimization: Theory and Applications (MOPTA) 2009

August 21, 2009

Outline

Outline

Sequential Quadratic Programming (SQP)

Gradient Sampling (GS)

SQP-GS

Numerical Results

Concluding Remarks

Constrained Optimization of Smooth Functions

- Consider constrained optimization problems of the form

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

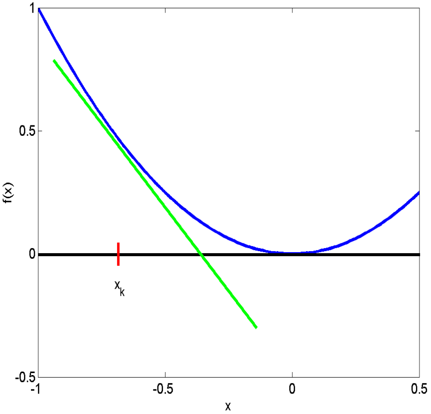
where f and c are *smooth* (equality constraints OK, too)

- At x_k, solve the SLP/SQP subproblem

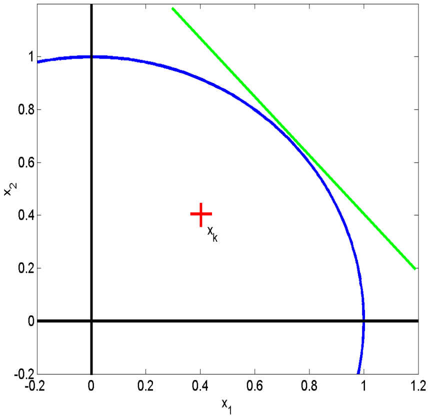
$$\min_d \; f_k + \nabla f_k^T d + \tfrac{1}{2} d^T H_k d$$
$$\text{s.t. } c_k + \nabla c_k^T d \le 0, \quad \|d\| \le \Delta_k$$

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

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

Penalty-SQP Method

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

- Solve the SQP subproblem

$$\begin{aligned} \min_d \ & \rho(f_k + \nabla f_k^T d) + \sum s^i + \tfrac{1}{2} d^T H_k d \\ \text{s.t. } & c_k + \nabla c_k^T d \le s, \quad s \ge 0 \end{aligned}$$

  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

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

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

Outline

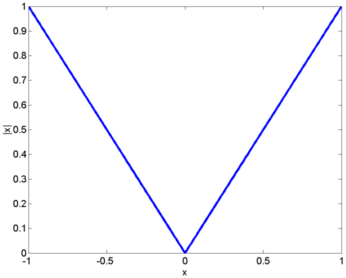
Unconstrained Optimization of Nonsmooth Functions

- 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:



The 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

$$\bar{\partial} f(x') = \bigcap_{\epsilon > 0} \text{cl conv } \nabla f(\mathbb{B}(x', \epsilon) \cap \mathcal{D})$$

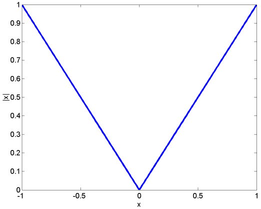
- A point x' is called Clarke stationary if $0 \in \bar{\partial} f(x')$

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: Stabilized/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

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

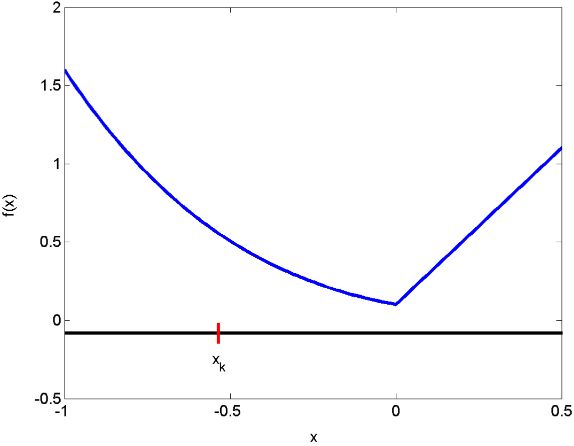
where $\mathcal{B}_k = \mathbb{B}(x_k, \epsilon) \cap \mathcal{D}$

- However, we can only approximate this step by solving

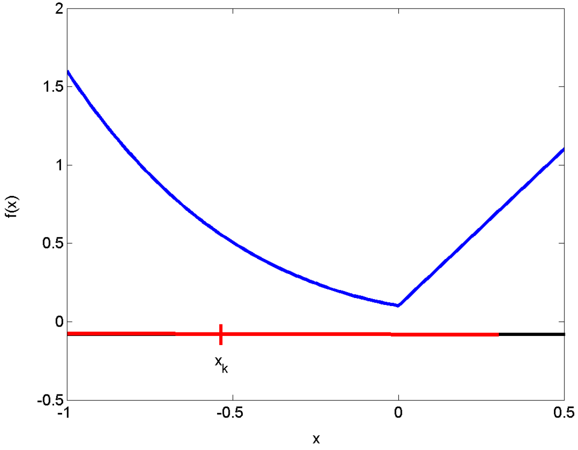
$$\min_d \; f_k + \max_{x \in \mathcal{B}_k} \{\nabla f(x)^T d\} + \tfrac{1}{2} d^T H_k d$$

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

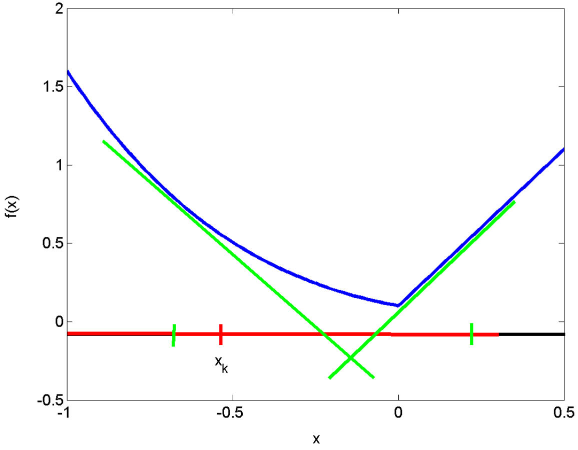
GS Illustration: Objective model



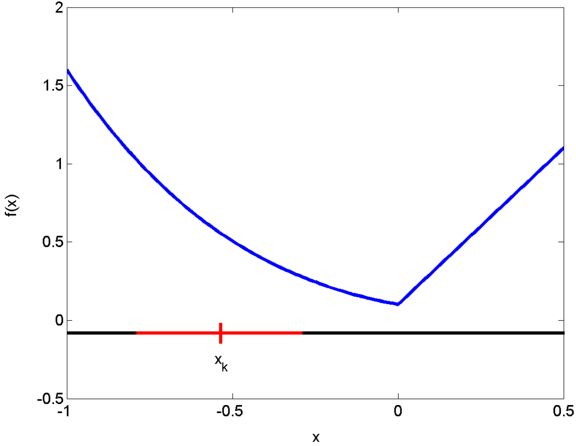
GS Illustration: Objective model



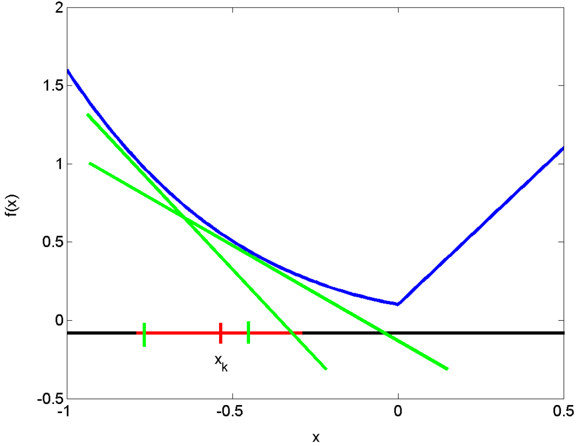
GS Illustration: Objective model



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

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

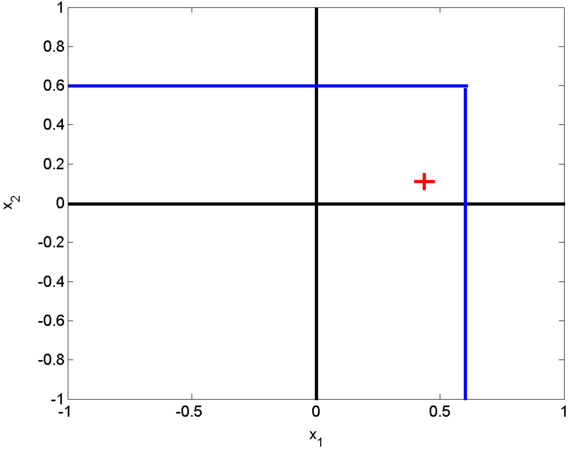
  where $\mathcal{B}_k^i = \{x_k, x_{k1}^i, \ldots, x_{kp}^i\} \subset \mathbb{B}(x_k, \epsilon)$ for $i = 0, \ldots, m$

- This is equivalent to

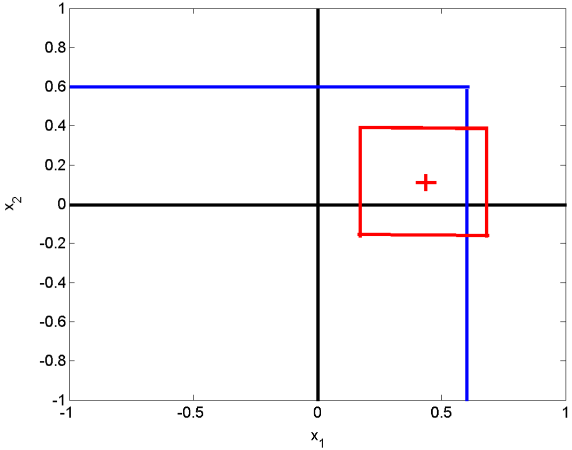
$$\min_d \ \rho \max_{x \in \mathcal{B}_k^0} (f_k + \nabla f(x)^T d) + \sum \max_{x \in \mathcal{B}_k^i} \max\{0, c_k^i + \nabla c^i(x)^T d, 0\} + \tfrac{1}{2} d^T H_k d$$

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

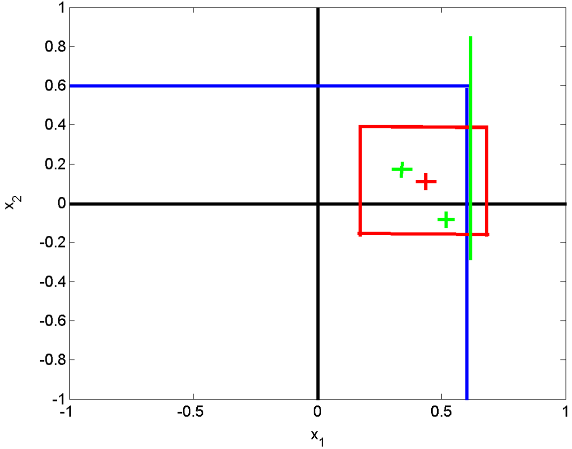
SQP-GS Illustration: Constraint model



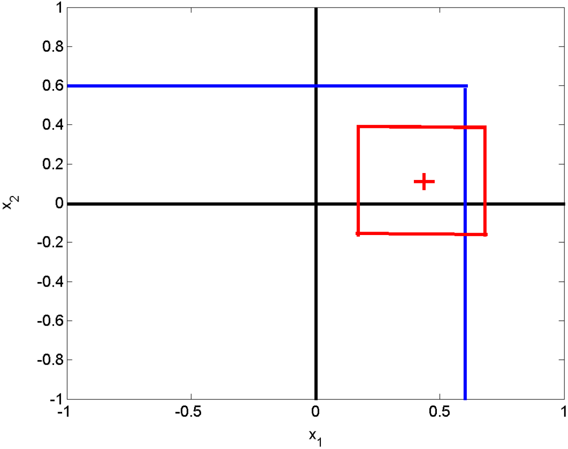
SQP-GS Illustration: Constraint model



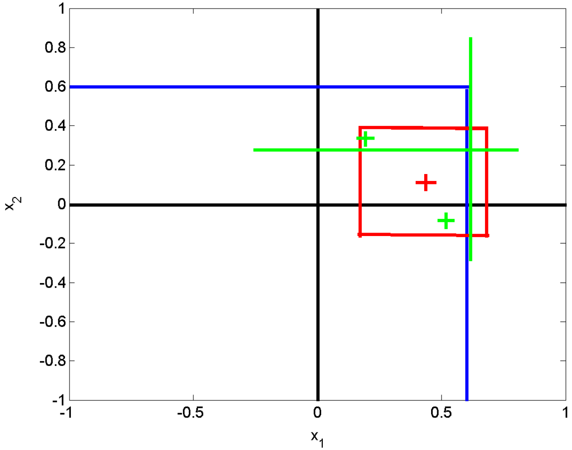
SQP-GS Illustration: Constraint model



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

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

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

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

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

Sequential Quadratic Programming (SQP)

Gradient Sampling (GS)

SQP-GS

Numerical Results

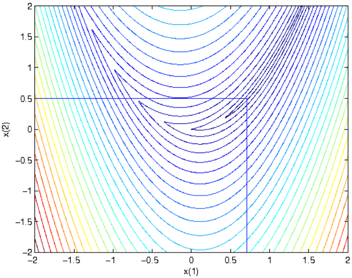
Concluding Remarks

Implementation

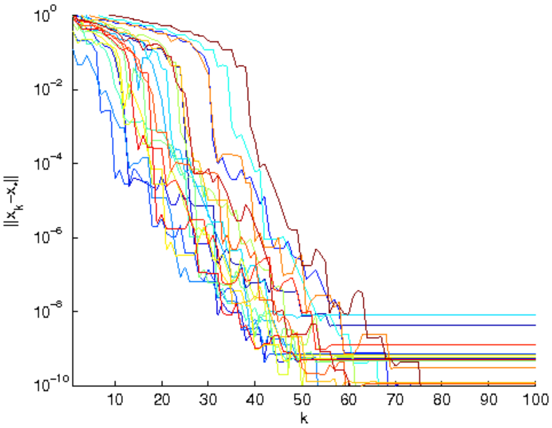
- Prototype implementation in MATLAB (available soon?)
- QP subproblems solved with MOSEK
- BFGS approximations of Hessian of penalty function
  - (Lewis and Overton, 2009)
- ρ decreased conservatively

Example 1: Nonsmooth Rosenbrock

$$\min_x \; 8|x_1^2 - x_2| + (1 - x_1)^2$$
$$\text{s.t.} \;\; \max\{\sqrt{2}x_1, 2x_2\} \le 1$$



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

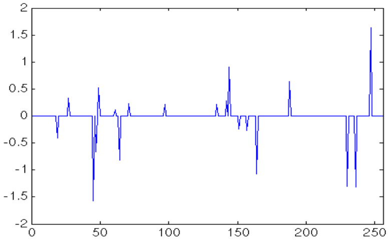
N	n	K	f (SQP-GS)	f (GS)
2	4	1	1.00000e+00	1.00000e+00
4	16	2	7.46296e-01	7.46286e-01
6	36	3	6.33589e-01	6.33477e-01
8	64	4	5.60165e-01	5.58820e-01
10	100	5	2.20724e-01	2.17193e-01
12	144	6	1.24820e-01	1.22226e-01
14	196	7	8.21835e-02	8.01010e-02
16	256	8	5.73762e-02	5.57912e-02

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

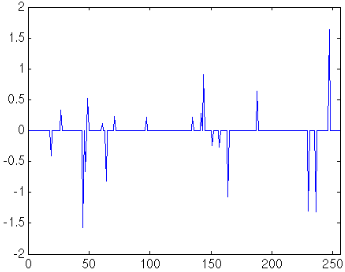
Recover a sparse signal by solving

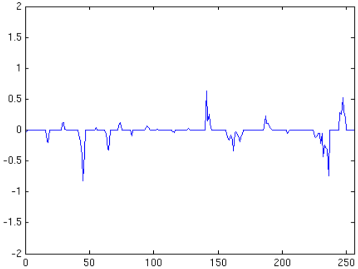
$$\min_x \|x\|_1$$
$$\text{s.t. } Ax = b$$

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



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



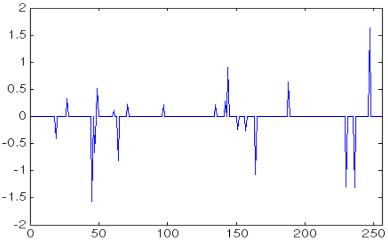


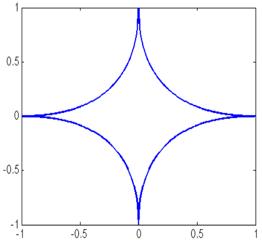
Example 3(b): Compressed sensing ($\ell_{0.5}$ norm)

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)

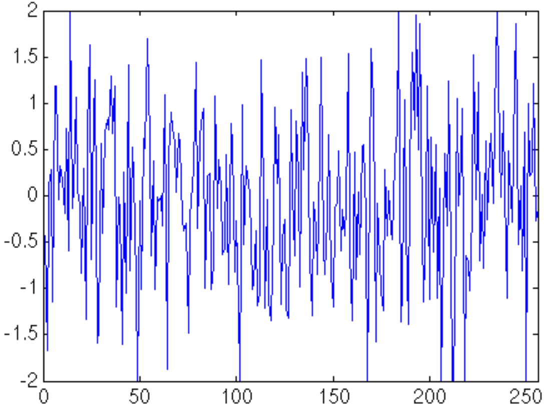


Figure: k = 1

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

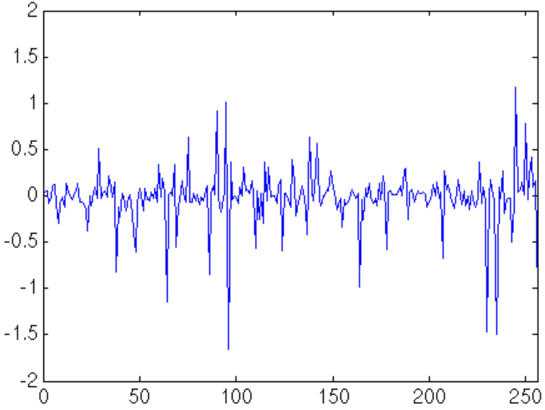


Figure: k = 10

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

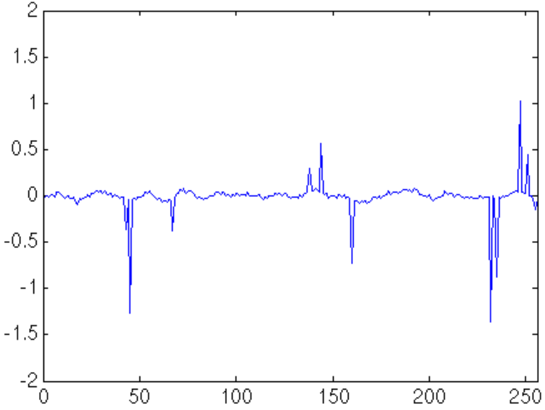


Figure: k = 25

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

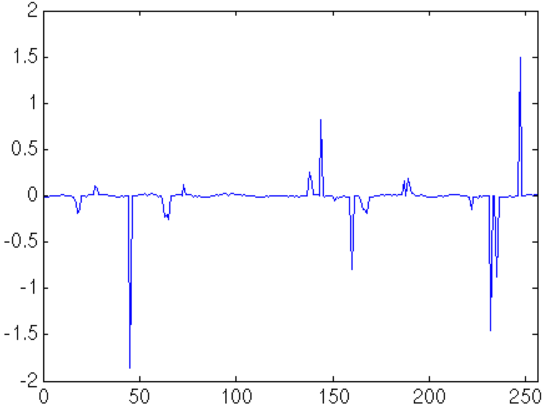


Figure: k = 50

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

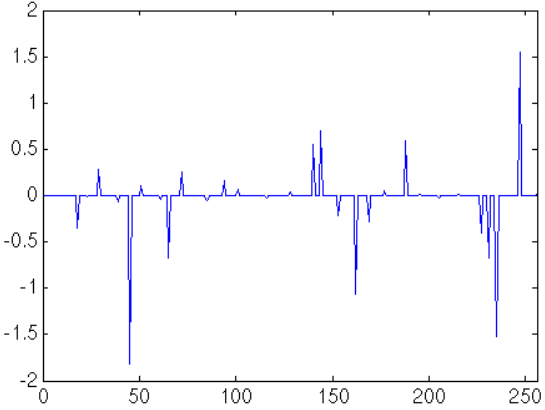
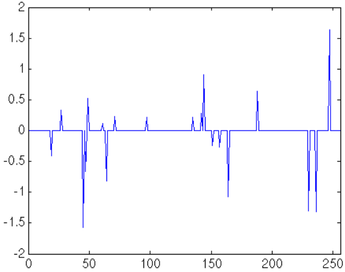
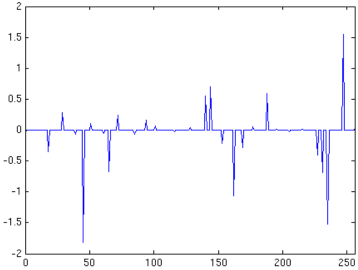


Figure: k = 200

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





Outline

Sequential Quadratic Programming (SQP)

Gradient Sampling (GS)

SQP-GS

Numerical Results

Concluding Remarks

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