Sequential Quadratic Programming with Gradient Sampling for Nonconvex Nonsmooth Constrained Optimization

Frank E. Curtis, Lehigh University
Michael L. Overton, New York University

Ninth US-Mexico Workshop on Optimization and Applications

January 6, 2011

Based on "A Sequential Quadratic Programming Algorithm for Nonconvex, Nonsmooth Constrained Optimization," under review for SIAM Journal on Optimization.

Nonlinear optimization

Consider constrained optimization problems of the form:

$$
\begin{aligned}
\min_x \; & f(x) && \text{(smooth)} \\
\text{s.t.} \; & c_{\mathcal{E}}(x) = 0 && \text{(smooth)} \\
& c_{\mathcal{I}}(x) \leq 0 && \text{(smooth)}
\end{aligned}
$$

- Decades worth of algorithmic development.
- SQP, IPM, etc., with countless variations.
- Strong global and local convergence guarantees.
- Multiple popular, successful software packages.

Nonlinear optimization with nonsmoothness

Consider constrained optimization problems of the form:

$$
\begin{aligned}
\min_{x} \ & f(x) && \text{((non)smooth)} \\
\text{s.t. } & c_{\mathcal{E}}(x) = 0 && \text{(smooth)} \\
& c_{\tilde{\mathcal{E}}}(x) = 0 && \text{(nonsmooth)} \\
& c_{\mathcal{I}}(x) \leq 0 && \text{(smooth)} \\
& c_{\tilde{\mathcal{I}}}(x) \leq 0 && \text{(nonsmooth)}
\end{aligned}
$$

- Algorithms for smooth problems no longer effective theoretically/practically.
- However, so much of the structure is the same as before.
- Can we adapt nonlinear optimization technology to handle nonsmoothness?

Outline

Outline

Constrained optimization with smooth functions

- Consider constrained optimization problems of the form:

$$\begin{aligned} \min_{x} \ & f(x) && \text{(smooth)} \\ \text{s.t. } & c(x) \leq 0 && \text{(smooth)} \end{aligned}$$

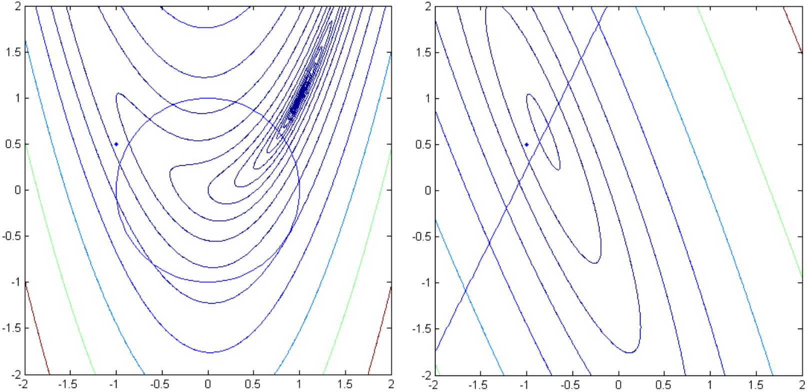
- At x_k, solve the SQP subproblem

$$\begin{aligned} \min_{d} \ & f(x_k) + \nabla f(x_k)^T d + \tfrac{1}{2} d^T H_k d \\ \text{s.t. } & c(x_k) + \nabla c(x_k)^T d \leq 0 \end{aligned}$$

to compute the search direction d_k.

SQP illustration

$$\min_x \; f(x) = 10(x_2 - x_1^2)^2 + (1 - x_1)^2 \;\; \text{s.t. } c(x) = \|x\|^2 - 1 \le 0 \;\; \text{at } x_k = (-1, \tfrac{1}{2}).$$



Inconsistent linearizations of the constraints

- The linearized constraints may be inconsistent, but we can relax the problem to

$$\min_{d} \rho(f(x_k) + \nabla f(x_k)^T d) + \sum s^i + \tfrac{1}{2} d^T H_k d$$
$$\text{s.t. } c(x_k) + \nabla c(x_k)^T d \leq s, \quad s \geq 0,$$

  i.e., a Penalty-SQP (PSQP) subproblem, where $\rho > 0$ is a penalty parameter.

- We perform a line search on the exact penalty function

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

  to promote global convergence.

Model function and line search

- A model of the penalty function is given by

$$q(d;\rho,x_k,H_k) \triangleq \rho(f(x_k)+\nabla f(x_k)^T d)+\sum \max\{c^i(x_k)+\nabla c^i(x_k)^T d,0\}+\tfrac{1}{2}d^T H_k d.$$

- Solving the PSQP subproblem is equivalent to minimizing $q(d;\rho,x_k,H_k)$.
- The reduction in $q(\cdot;\rho,x_k,H_k)$ yielded by d_k is

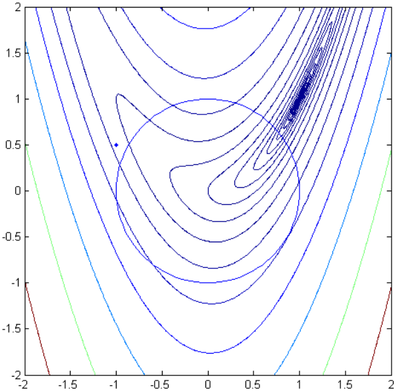
$$\Delta q(d_k;\rho,x_k,H_k) \triangleq q(0;\rho,x_k,H_k) - q(d_k;\rho,x_k,H_k).$$

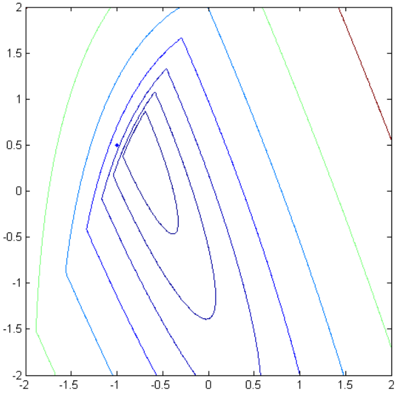
- We impose the sufficient decrease condition

$$\phi(x_k+\alpha_k d_k;\rho) \leq \phi(x_k;\rho) - \eta\alpha_k \Delta q(d_k;\rho,x_k,H_k).$$

PSQP illustration

$$\min_{x} \ \phi(x;\rho) = \rho(10(x_2 - x_1^2)^2 + (1 - x_1)^2) + \max\{x_1^2 + x_2^2 - 1, 0\} \ \text{ at } x_k = (-1, \tfrac{1}{2}).$$





PSQP method

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

- Solve the PSQP subproblem

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

  to compute d_k.

- Backtrack from $\alpha_k \leftarrow 1$ to satisfy the sufficient decrease condition

$$\phi(x_k + \alpha_k d_k; \rho) \leq \phi(x_k; \rho) - \eta \alpha_k \Delta q(d_k; \rho, x_k, H_k).$$

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

Sketch of convergence theory for PSQP

Assume that the following conditions hold:

- $\{x_k\}$ is contained in a convex set over which f and c and their first derivatives are bounded and Lipschitz continuous.
- $\{H_k\}$ are symmetric positive definite, bounded above in norm, and bounded away from singularity.

Then, $\{x_k\}$ converges to a stationary point of $\phi(x;\rho)$.

Outline

Unconstrained optimization of nonsmooth functions

- Consider unconstrained optimization problems of the form:

$$\min_{x} \; f(x) \quad \text{(nonsmooth, locally Lipschitz)}$$

(ϵ-)subdifferentials

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

Gradient sampling (GS)

- At x_k, suppose we approximate the ϵ-subdifferential

$$\partial f(x_k, \epsilon) = \mathrm{cl\ conv}\ \partial f(\mathbb{B}(x_k, \epsilon) \cap \mathcal{D})$$

by sampling gradients in a finite set (with $x_{k0} := x_k$)

$$\mathcal{B}_k := \{x_{k0}, x_{k1}, \ldots, x_{kp}\} \subset \mathbb{B}(x_k, \epsilon) \cap \mathcal{D}.$$

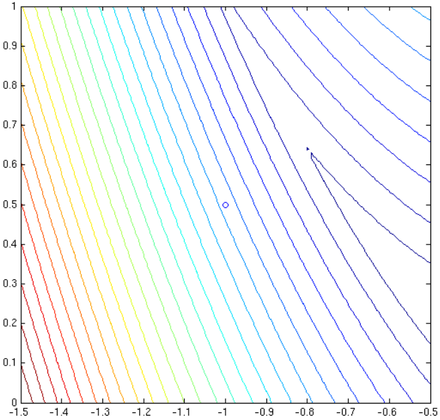
- An approximate steepest descent step is then obtained by solving

$$\begin{aligned} \min_d\ & \tfrac{1}{2}\|d\|^2 \\ \text{s.t. } & d = -\sum \lambda^i \nabla f(x_{ki}) \\ & 1 = \sum \lambda^i \\ & \lambda^i \geq 0. \end{aligned}$$

That is, $d = -g$, where g is the projection of 0 onto $\mathrm{conv}\{\nabla f(x_{ki})\}$.

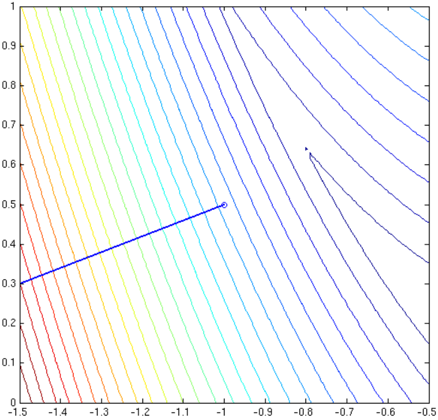
GS illustration

$$\min_x \; f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \;\; \text{at } x_k = (-1, \tfrac{1}{2}).$$



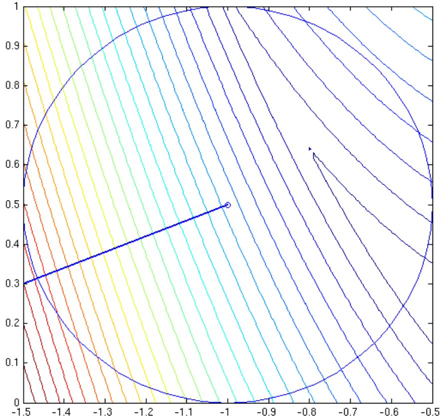
GS illustration

$$\min_{x}\ f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \ \text{ at } x_k = (-1, \tfrac{1}{2}).$$



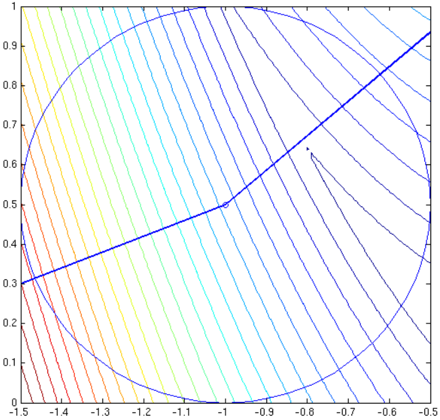
GS illustration

$$\min_{x} \; f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \;\text{ at } x_k = (-1, \tfrac{1}{2}).$$



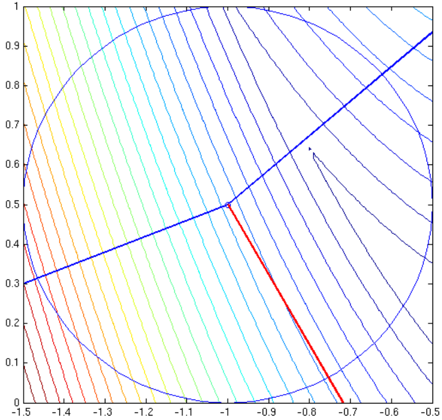
GS illustration

$$\min_x \ f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \ \text{ at } x_k = (-1, \tfrac{1}{2}).$$



GS illustration

$$\min_x \ f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \ \text{ at } x_k = (-1, \tfrac{1}{2}).$$



GS method

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

- Sample $p = n + 1$ points $\{x_{k1}, \dots, x_{kp}\}$ in $\mathbb{B}(x_k, \epsilon) \cap \mathcal{D}$.
- Solve the GS subproblem

$$\begin{aligned} \min_d \ & \tfrac{1}{2}\|d\|^2 \\ \text{s.t. } & d = -\sum \lambda^i \nabla f(x_{ki}) \\ & 1 = \sum \lambda^i \\ & \lambda^i \geq 0. \end{aligned}$$

  to compute d_k.
- Backtrack from $\alpha_k \leftarrow 1$ to satisfy the sufficient decrease condition

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

- Update $x_{k+1} \approx x_k + \alpha_k d_k$ (to ensure $x_{k+1} \in \mathcal{D}$).
- If $\|d_k\|^2 \leq \epsilon^2$, then reduce ϵ.

Sketch of convergence theory for GS

Assume that the following condition holds:

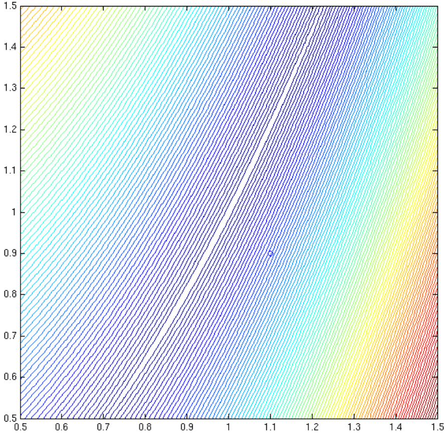
- f is locally Lipschitz, bounded below, and continuously differentiable in an open set $\mathcal{D}$ of $\mathbb{R}^n$.

Then, with probability 1, every cluster point of $\{x_k\}$ is stationary for f.

(See Burke, Lewis, and Overton (2005) and Kiwiel (2007).)

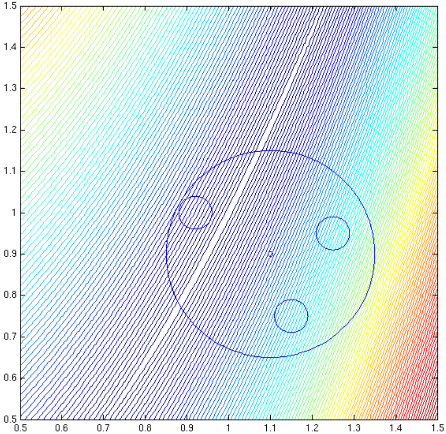
GS theory illustration

$$\min_{x} \; f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \;\; \text{at } x_k = (1.1, 0.9).$$



GS theory illustration

$$\min_{x} \; f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \quad \text{at } x_k = (1.1, 0.9).$$



Dual problem for search direction

The GS subproblem is equivalent to

$$\begin{aligned} \max_{d}\ & f(x_k) - \tfrac{1}{2}\|d\|^2 \\ \text{s.t. } & d = -\sum \lambda^i \nabla f(x_{ki}) \\ & 1 = \sum \lambda^i \\ & \lambda^i \geq 0. \end{aligned}$$

The dual of this QP reveals an alternative definition of d_k:

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

Equivalently:

$$\min_{d}\ f(x_k) + \max_{x \in \mathcal{B}_k} \{\nabla f(x)^T d\} + \tfrac{1}{2}\|d\|^2$$

Dual problem for search direction (more general)

The GS subproblem is equivalent to $(H_k \succ 0)$

$$
\begin{aligned}
\max_{d}\ & f(x_k) - \tfrac{1}{2} d^T H_k d \\
\text{s.t. } & H_k d = -\sum \lambda^i \nabla f(x_{ki}) \\
& 1 = \sum \lambda^i \\
& \lambda^i \geq 0.
\end{aligned}
$$

The dual of this QP reveals an alternative definition of d_k:

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

Equivalently:

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

Outline

Constrained optimization of nonsmooth functions

- Consider constrained optimization problems of the form

$$\min_x \; f(x) \qquad \text{(nonsmooth, locally Lipschitz)}$$
$$\text{s.t. } c(x) \leq 0 \quad \text{(nonsmooth, locally Lipschitz)}$$

- We may consider applying an unconstrained technique directly to

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

but can we do better by maintaining the framework of SQP?

SQP and GS

- The SQP subproblem (for a smooth constrained problem) is

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

- The GS subproblem (for a nonsmooth objective) is

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

SQP and GS

- The SQP subproblem (for a smooth constrained problem) is

$$\begin{aligned} \min_{d} \ & \rho z + \sum s^i + \tfrac{1}{2} d^T H_k d \\ \text{s.t. } & f(x_k) + \nabla f(x_k)^T d \le z \\ & c(x_k) + \nabla c(x_k)^T d \le s, \ s \ge 0. \end{aligned}$$

- The GS subproblem (for a nonsmooth objective) is

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

- The SQP-GS subproblem (for a nonsmooth constrained problem) is

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

SQP-GS in more detail

- The SQP-GS subproblem is

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

where

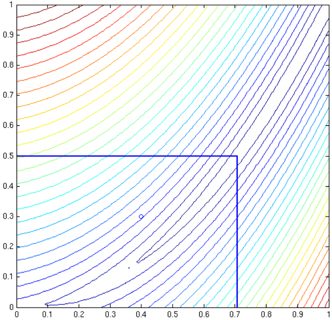
$$\begin{aligned} \mathcal{B}_k^f &= \{x_{k0}, x_{k1}, \dots, x_{kp}\} &\subset\ & \mathbb{B}(x_k, \epsilon) \cap \mathcal{D}^f \\ \mathcal{B}_k^{c^i} &= \{x_{ki0}, x_{ki1}, \dots, x_{kip}\} &\subset\ & \mathbb{B}(x_k, \epsilon) \cap \mathcal{D}^{c^i} \text{ for } i = 1, \dots, m. \end{aligned}$$

- This is equivalent to

$$\begin{aligned} \min_d \ & q(d; \rho, x_k, H_k, \mathcal{B}_k^f, \mathcal{B}_k^{c^1}, \dots, \mathcal{B}_k^{c^m}) := \\ & \rho \max_{x \in \mathcal{B}_k^f} (f(x_k) + \nabla f(x)^T d) + \sum \max_{x \in \mathcal{B}_k^{c^i}} \max\{c^i(x_k) + \nabla c^i(x)^T d, 0\} + \tfrac{1}{2} d^T H_k d. \end{aligned}$$

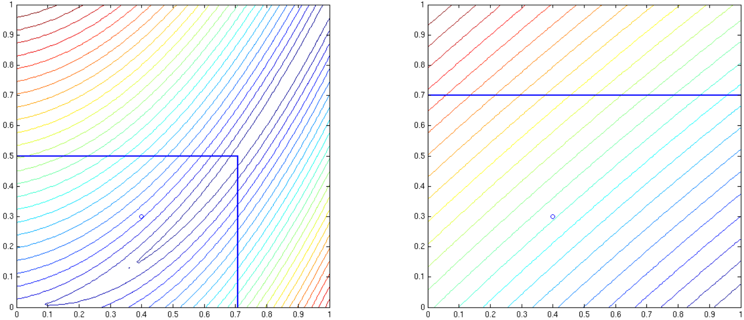
SQP-GS illustration

$$\min_{x} \; f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \;\text{ s.t. }\; c(x) = \max\{\sqrt{2}x_1, 2x_2\} - 1 \leq 0 \;\text{ at }\; x_k = (\tfrac{2}{5}, \tfrac{3}{10}).$$



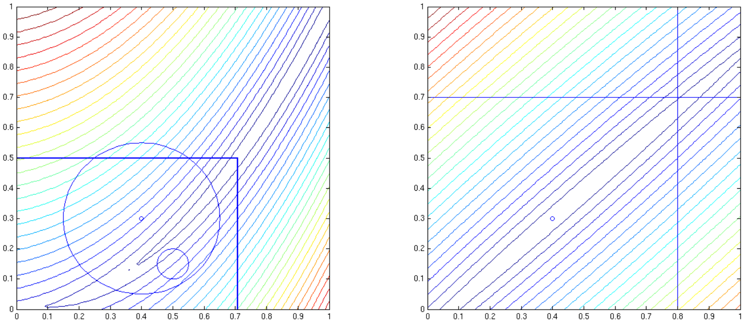
SQP-GS illustration

$$\min_x \; f(x) = 10|x_2 - x_1^2| + (1-x_1)^2 \;\text{ s.t. }\; c(x) = \max\{\sqrt{2}x_1, 2x_2\} - 1 \le 0 \;\text{ at }\; x_k = (\tfrac{2}{5}, \tfrac{3}{10}).$$



SQP-GS illustration

$$\min_x \; f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \;\text{ s.t. }\; c(x) = \max\{\sqrt{2}x_1, 2x_2\} - 1 \le 0 \;\text{ at }\; x_k = (\tfrac{2}{5}, \tfrac{3}{10}).$$



SQP-GS method

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

- Sample $p = n + 1$ points for each function to generate $\mathcal{B}_k := \{\mathcal{B}_k^f, \mathcal{B}_k^{c^1}, \ldots, \mathcal{B}_k^{c^m}\}$.
- Solve the SQP-GS subproblem

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

  to compute d_k.
- Backtrack from $\alpha_k \leftarrow 1$ to satisfy

$$\phi(x_k + \alpha_k d_k; \rho) \le \phi(x_k; \rho) - \eta \alpha_k \Delta q(d_k; \rho, x_k, H_k, \mathcal{B}_k).$$

- Update $x_{k+1} \approx x_k + \alpha_k d_k$ (to ensure $x_{k+1} \in \mathcal{D}^f \cap \mathcal{D}^{c^1} \cap \cdots \cap \mathcal{D}^{c^m}$)
- If $\Delta q(d_k; \rho, x_k, H_k, \mathcal{B}_k) \le \epsilon^2$, then reduce ϵ.

Convergence theory for SQP-GS

Assume that the following conditions hold:

- f and c^i, $i = 1, \ldots, m$, are locally Lipschitz and continuously differentiable on open dense subsets of $\mathbb{R}^n$.
- $\{x_k\}$ and all generated sample points are contained in a convex set over which f and c^i, $i = 1, \ldots, m$, and their first derivatives are bounded.
- $\{H_k\}$ are symmetric positive definite, bounded above in norm, and bounded away from singularity.

Convergence theory for SQP-GS

Define the subproblem

$$\min_{d} \widetilde{q}(d;\rho,x',H',\epsilon) := \rho \max_{x\in\mathbb{B}(x',\epsilon)\cap\mathcal{D}^f}(f(x')+\nabla f(x)^T d) + \sum \max_{x\in\mathbb{B}(x',\epsilon)\cap\mathcal{D}^{c^i}} \max\{c^i(x')+\nabla c^i(x)^T d, 0\} + \tfrac{1}{2}d^T H' d.$$

x' is ϵ-stationary if the solution d' to the above yields

$$\Delta\widetilde{q}(d';\rho,x',H',\epsilon) = 0.$$

x' is stationary if it is ϵ-stationary for all $\epsilon > 0$.

Convergence theory for SQP-GS

Lemma 1:

- If $\Delta q(d_k; \rho, x_k, H_k, \mathcal{B}_k) = \frac{1}{2} d_k^T H_k d_k = 0$, then x_k is ϵ-stationary.

Lemma 2:

- The directional derivative of the penalty function satisfies

$$\phi'(d_k; \rho, x_k) \leq -d_k^T H_k d_k < 0,$$

and so d_k is a descent direction for $\phi(x; \rho)$ at x_k.

Convergence theory for SQP-GS

Let

$$\mathcal{S}(x_k,\epsilon) = \left(\prod_1^p \mathbb{B}(x_k,\epsilon)\cap\mathcal{D}^f, \prod_1^p \mathbb{B}(x_k,\epsilon)\cap\mathcal{D}^{c^1},\dots,\prod_1^p \mathbb{B}(x_k,\epsilon)\cap\mathcal{D}^{c^m}\right)$$

and

$$\mathcal{T}(\rho,x_k,\epsilon,x',\omega) = \{\mathcal{B}_k\in\mathcal{S}(x_k,\epsilon) \mid \Delta q(d_k;\rho,x_k,H_k,\mathcal{B}_k)\le\Delta\widetilde{q}(d';\rho,x',H_k,\epsilon)+\omega\}.$$

Lemma 3:

- For any $\omega > 0$, there exists $\zeta > 0$ and a nonempty set $\mathcal{T}$ such that for all $x_k \in \mathbb{B}(x',\zeta)$ we have $\mathcal{T}\subset\mathcal{T}(\rho,x_k,\epsilon,x',\omega)$.

That is, in a sufficiently small neighborhood of x', there exists a set of sample sets revealing $\Delta\widetilde{q}(d';\rho,x',H_k,\epsilon)$ to an arbitrary accuracy.

Convergence theory for SQP-GS

Theorem:

- With probability 1, every cluster point of $\{x_k\}$ is stationary for $\phi(x;\rho)$.

Sketch of proof:

- If $\epsilon \nrightarrow 0$, then for all large k

$$\Delta q(d_k;\rho,x_k,H_k,\mathcal{B}_k) > \epsilon^2.$$

However, with probability 1, this will not occur.

  - $\epsilon \nrightarrow 0$ implies $x_k \rightarrow x'$. If x' is ϵ-stationary, then w.p.1 we will obtain a sample set yielding $\Delta q(d_k;\rho,x_k,H_k,\mathcal{B}_k) \leq \epsilon^2/2$, contradicting the above.
  - $\epsilon \nrightarrow 0$ also implies $\alpha_k \rightarrow 0$. If x' is not ϵ-stationary, then w.p.1 we will obtain a subsequence of iterations yielding α_k bounded away from zero, contradicting $\alpha_k \rightarrow 0$.

Thus, with probability 1, $\epsilon \rightarrow 0$ and any cluster point x' is stationary for $\phi(x;\rho)$.

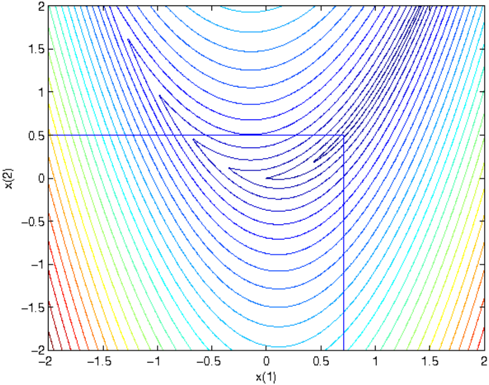
Outline

Implementation

- Prototype implementation in MATLAB.
- QP subproblems solved with MOSEK.
- BFGS approximations of Hessian of $\phi(x; \rho)$. (See Lewis and Overton (2009).)
- ρ decreased only when ϵ decreased.

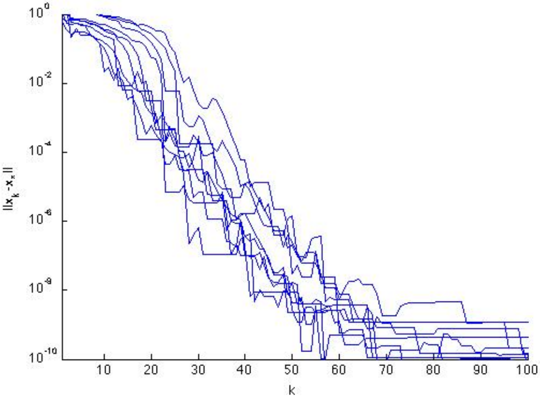
Example 1: Nonsmooth Rosenbrock

$$\min_{x} \ 10|x_1^2 - x_2| + (1 - x_1)^2 \ \text{ s.t. } \ \max\{\sqrt{2}x_1, 2x_2\} \leq 1.$$



Example 1: Nonsmooth Rosenbrock

$$\min_{x} \; 10|x_1^2 - x_2| + (1 - x_1)^2 \;\; \text{s.t.} \;\; \max\{\sqrt{2}x_1, 2x_2\} \le 1.$$



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

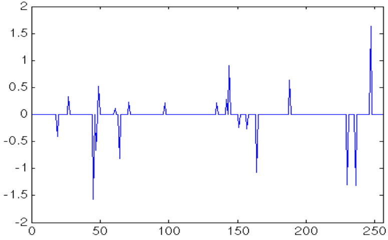
N	n	K	f (SQP-GS)	f (GS)
2	4	1	1.00000e+00	1.00000e+00
4	16	2	7.46287e-01	7.46286e-01
6	36	3	6.33612e-01	6.33477e-01
8	64	4	5.59883e-01	5.58820e-01
10	100	5	2.17882e-01	2.17193e-01
12	144	6	1.22556e-01	1.22226e-01
14	196	7	8.19025e-02	8.01010e-02
16	256	8	5.65826e-02	5.57912e-02

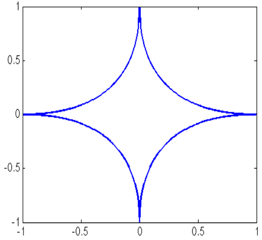
Example 3: $\ell_{0.5}$ norm minimization

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: $\ell_{0.5}$ norm minimization

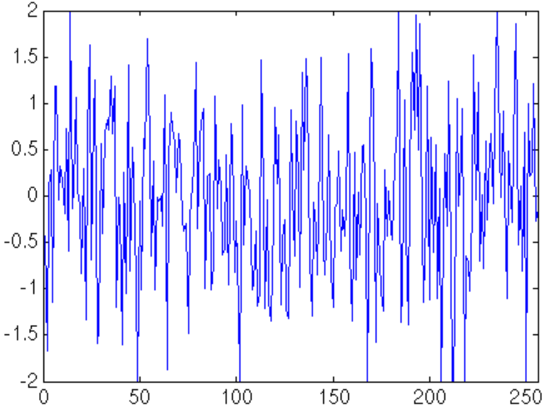


Figure: k = 1

Example 3: $\ell_{0.5}$ norm minimization

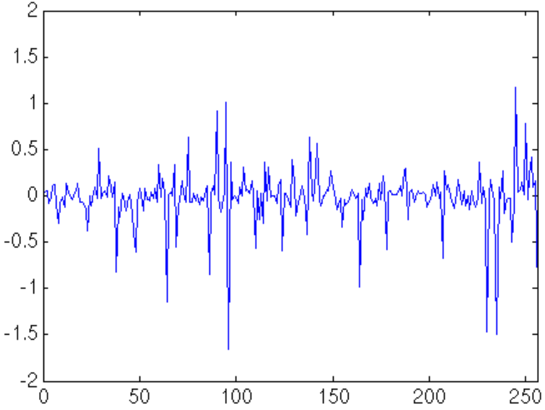


Figure: k = 10

Example 3: $\ell_{0.5}$ norm minimization

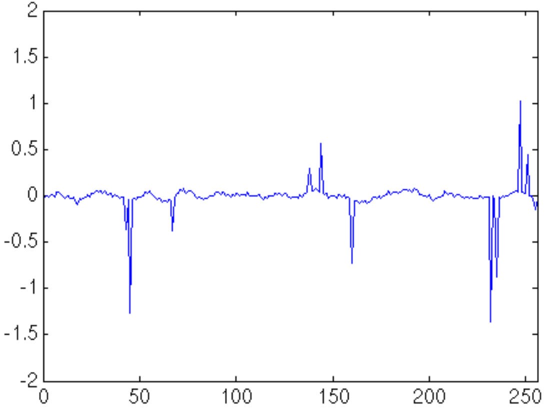


Figure: k = 25

Example 3: $\ell_{0.5}$ norm minimization

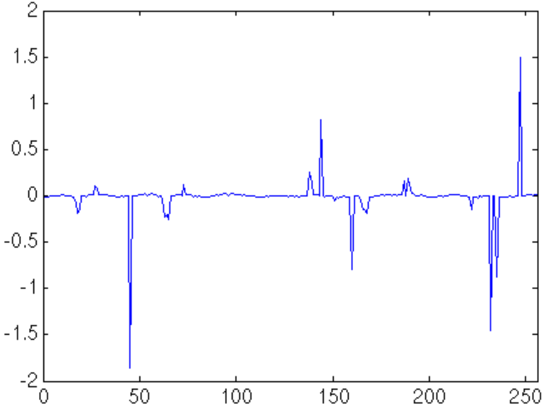


Figure: k = 50

Example 3: $\ell_{0.5}$ norm minimization

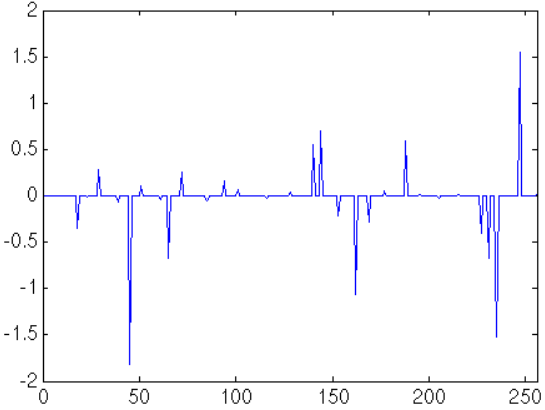


Figure: k = 200

Outline

Sequential Quadratic Programming (SQP)

Gradient Sampling (GS)

SQP-GS

Numerical Results

Future Work

Summary

Implementation details

- Already includes option for sampling only for nonsmooth functions.
- SQP-GS, SLP-GS, and trust regions:

$$d_k^T H_k d_k \quad \text{and/or} \quad \|d_k\| \leq \Delta_k$$

- Quasi-Newton Hessian approximations, penalty function or Lagrangian.
- SQP-IQP vs. SQP-EQP vs. SQP-IQP-EQP.
- Special handling of convex/linear functions.
- Tuned updates for the sampling radius ϵ.

Penalty parameter updates

Conservative update:

- For fixed ρ, update $\epsilon \to 0$ to find a stationary point for $\phi(x;\rho)$. Decrease ρ if infeasible, and resolve.

Moderate update (used in current tests):

- Define a forcing sequence $\theta \to 0$ for monitoring feasibility violations. Whenever ϵ is decreased, decrease ρ if violation exceeds current θ.

Aggressive update:

- (Steering rules). During every iteration, decrease ρ until the computed search direction yields sufficient progress toward linearized feasibility.

(All options already implemented.)

IP-GS

- "Redundant" constraints with unique slacks produce the log-barrier subproblem

$$\min_{x,s} \; f(x) - \mu \sum \sum \ln s^{ij}$$
$$\text{s.t. } c^i(x) + s^{ij} = 0,\; i \in \mathcal{I},\; j = 0, \dots, p.$$

- Newton step corresponds to solving the linear system

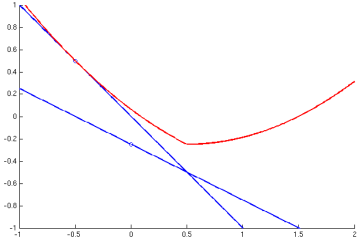
$$\begin{bmatrix} H_k & 0 & J_k^T \\ 0 & \Omega_k & I \\ J_k & I & 0 \end{bmatrix} \begin{bmatrix} d_k^x \\ d_k^s \\ \delta_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) + J_k^T \lambda_k \\ \lambda_k - \mu S_k^{-1} e \\ c(x_k) \otimes e + s \end{bmatrix},$$

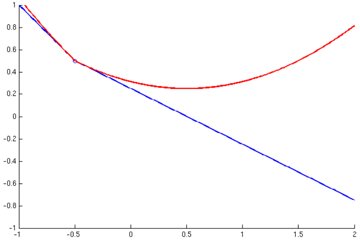
where

$$J_k := [\nabla c^i(x_k^{ij})].$$

Bundle methods vs. GS

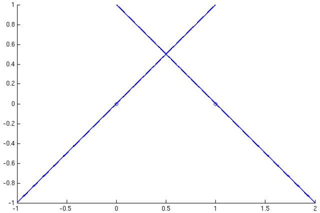
$(BM)\ \min\ z + \frac{1}{2}\|d\|^2$ $\text{s.t. } f(x_j) + \nabla f(x_j)^T d \leq z,\ \forall x_j \in \mathcal{B}_k$	$(GS)\ \min\ z + \frac{1}{2}\|d\|^2$ $\text{s.t. } f(x_k) + \nabla f(x_j)^T d \leq z,\ \forall x_j \in \mathcal{B}_k$

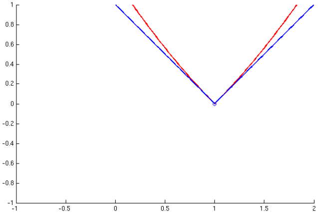




Bundle methods vs. GS

(BM) min $z + \frac{1}{2}\|d\|^2$ s.t. $f(x_j) + \nabla f(x_j)^T d \leq z, \ \forall x_j \in \mathcal{B}_k$	(GS) min $z + \frac{1}{2}\|d\|^2$ s.t. $f(x_k) + \nabla f(x_j)^T d \leq z, \ \forall x_j \in \mathcal{B}_k$





Bundle methods merged with gradient "sampling"

Define the subproblem to be

$$\begin{aligned} &\min\ z + \tfrac{1}{2}\|d\|^2 \\ &\text{s.t. } f(x_k) + \nabla f(x_j)^T d \le z,\ \forall x_j \in \mathcal{B}_k \end{aligned}$$

with $\mathcal{B}_k$ defined at previous iterates. We essentially have:

- a bundle method for nonconvex problems that replaces

$$f(x_j) + \nabla f(x_j)^T d \le z,\ \forall x_j \in \mathcal{B}_k$$

  with

$$f(x_k) + \nabla f(x_j)^T d \le z,\ \forall x_j \in \mathcal{B}_k;$$

- a cheaper GS method that replaces sampled gradients with historical info.

Outline

Summary

- Globally convergent method for nonconvex nonsmooth constrained optimization.
- Penalty-SQP with Gradient Sampling to capture information of nonsmoothness.
- Preliminary numerical results are encouraging.
- Extensions can carry nonlinear optimization technology to nonsmooth problems.