Nonconvex, Nonsmooth Optimization via Gradient Sampling

Frank E. Curtis, Lehigh University

involving joint work with

Michael L. Overton, New York University
Xiaocun Que, Lehigh University

SIAM Conference on Imaging Science

May 20, 2012

Outline

Outline

Optimization research: Structured vs. unstructured problems

Emphasis today on solving structured optimization problems.

- In most cases, structure means convex.
- Often goes further, e.g., seeking sparsity, low matrix rank, low total variation, etc.
- First-order methods, optimal algorithms, regularization, etc.

My work has focused on unstructured optimization problems.

- For one thing, unstructured means nonconvex.
- General-purpose algorithms are the "go-to" methods for new problems.
- General-purpose algorithms are all we have for very hard problems.

(Disclaimer: In this talk, I do not address global optimization.)

Deterministic optimization methods based on randomized models

Unconstrained minimization of an objective function $f : \mathbb{R}^n \to \mathbb{R}$:

- No gradient info available? e.g., objective values from simulations
- Only some gradient info available? e.g., large-scale machine learning
- Subdifferential not available? e.g., any unstructured nonsmooth problem

Randomized algorithms offer computational flexibility, as well as other benefits.

Contributions

Gradient sampling is a general-purpose method for nonconvex, nonsmooth problems.

- We dramatically reduce per-iteration and overall computational cost.
- Nothing is lost in terms of global convergence guarantees.
- We extend the methodology and theory to constrained optimization.
- Numerical results are promising and will improve with further enhancements.

Unconstrained nonconvex, nonsmooth optimization

Consider the unconstrained problem

$$\boxed{\min_x \ f(x)}$$

where f is locally Lipschitz and continuously differentiable in (dense) $\mathcal{D} \subset \mathbb{R}^n$.

- Let
$$\mathbb{B}_\epsilon(\overline{x}) := \{x \mid \|x - \overline{x}\| \leq \epsilon\}.$$
- $\overline{x}$ is stationary if
$$0 \in \partial f(\overline{x}) := \bigcap_{\epsilon > 0} \operatorname{cl} \operatorname{conv} \nabla f(\mathbb{B}_\epsilon(\overline{x}) \cap \mathcal{D}).$$
- $\overline{x}$ is ϵ-stationary if
$$0 \in \partial_\epsilon f(\overline{x}) := \operatorname{cl} \operatorname{conv} \partial f(\mathbb{B}_\epsilon(\overline{x})).$$

Gradient sampling (GS) idea

At x_k, let $x_{k0} := x_k$ and sample $\{x_{k1}, \ldots, x_{kp}\} \subset \mathbb{B}_\epsilon(x_k) \cap \mathcal{D}$, yielding:

$$\begin{aligned} X_k &:= \begin{Bmatrix} x_{k0}, & x_{k1}, & \cdots, & x_{kp} \end{Bmatrix} & \text{(sample points)} \\ G_k &:= \begin{bmatrix} g_{k0} & g_{k1} & \cdots & g_{kp} \end{bmatrix} & \text{(sample gradients)} \end{aligned}$$

The ϵ-subdifferential is approximated by the convex hull of the sampled gradients:

$$\begin{aligned} \partial_\epsilon f(x_k) &= \operatorname{cl}\operatorname{conv}\partial f(\mathbb{B}_\epsilon(x_k)) \\ &\approx \operatorname{conv}\{g_{k0}, g_{k1}, \ldots, g_{kp}\} \end{aligned}$$

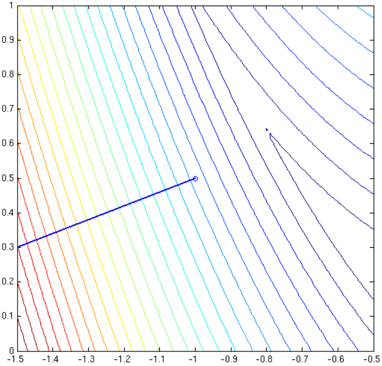
- Compute the projection of 0 onto the convex hull of the sampled gradients:

$$\boxed{g_k := \operatorname{Proj}(0 | \operatorname{conv}\{g_{k0}, g_{k1}, \ldots, g_{kp}\})}$$

Then, $d_k = -g_k$ is an approximate ϵ-steepest descent step.

GS illustration

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



GS method

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

- Sample $p \geq n+1$ points $\{x_{k1}, \ldots, x_{kp}\} \subset \mathbb{B}_\epsilon(x_k) \cap \mathcal{D}$.
- Compute $d_k \leftarrow -g_k$ by computing the projection

$$g_k = \text{Proj}(0 | \,\text{conv}\{g_{k0}, g_{k1}, \ldots, g_{kp}\}).$$

- 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 $\|d_k\| \leq \epsilon$, then reduce ϵ.

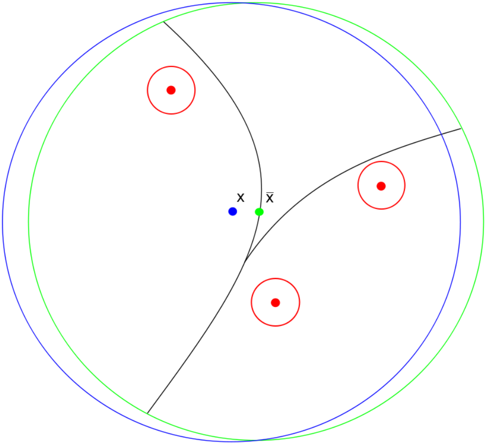
(See Burke, Lewis, and Overton (2005) and Kiwiel (2007).)

Global convergence of GS

Theorem: Let f be locally Lipschitz and continuously differentiable on an open dense $\mathcal{D} \subset \mathbb{R}^n$. Then, w.p.1, $f(x_k) \to -\infty$ or every cluster point of $\{x_k\}$ is stationary for f.

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

Illustration of critical part of proof



$\exists \{y_{ki}\}_{i=1,\ldots,p}$ and $\delta > 0$ such that $\text{Proj}(0|\{\nabla f(y_{ki} + O(\delta))\}) \approx \text{Proj}(0|\partial_\epsilon f(\overline{x}))$

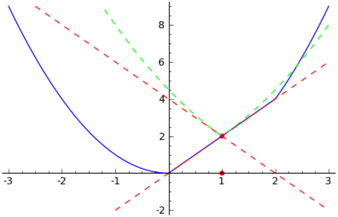
Local models in GS

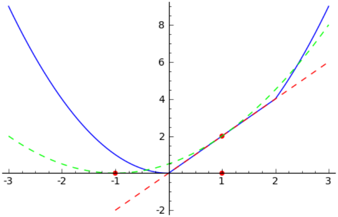
Computing the projection is equivalent to solving the dual subproblem:

$$\max_{\lambda} f(x_k) - \tfrac{1}{2}\|G_k\lambda\|^2$$
$$\text{s.t. } e^T\lambda = 1,\ \lambda \geq 0.$$

The corresponding primal subproblem is to compute d_k in the solution to

$$\min_{z,d} z + \tfrac{1}{2}\|d\|^2$$
$$\text{s.t. } f(x_k)e + G_k^T d \leq ze.$$





Outline

Gradient Sampling

Enhancements

Numerical Experiments

Summary

Practical issues

Practical limitations of original GS method:

- $p \geq n+1$ gradient evaluations per iteration
- All subproblems solved from scratch
- Behaves like steepest descent(?)
- Does not allow constraints

Proposed enhancements:

- Adaptive sampling; only $O(1)$ gradients per iteration (Kiwiel (2010))
- Warm-started subproblem solves
- "Hessian" approximations for quadratic term
- SQP framework to allow nonconvex, nonsmooth constraints

Adaptive Gradient Sampling (AGS)

At x_k, we had:

$$\begin{aligned} X_k &:= \{x_{k0}, \quad x_{k1}, \quad \cdots, \quad x_{kp}\} \quad &\text{(sample points)} \\ G_k &:= \begin{bmatrix} g_{k0} & g_{k1} & \cdots & g_{kp} \end{bmatrix} \quad &\text{(sample gradients)} \end{aligned}$$

At x_{k+1}, we

- maintain sample points still within radius ϵ; (this allows warm-starting!)
- throw out gradients outside of radius;
- sample 1 (or some) new gradients.

How can we maintain global convergence?

- If sample size is at least $n+1$, then proceed as usual; else, truncate line search.

Primal-dual pair of subproblems (variable-metric)

Recall the primal-dual pair of GS subproblems:

$$\boxed{\begin{aligned} &\max_{z,d}\ z + \tfrac{1}{2}d^T d \\ &\ \text{s.t. } f(x_k)e + G_k^T d \leq ze \end{aligned}} \qquad \boxed{\begin{aligned} &\max_{\lambda}\ f(x_k) - \tfrac{1}{2}\lambda^T G_k^T G_k \lambda \\ &\ \text{s.t. } e^T\lambda = 1,\ \lambda \geq 0 \end{aligned}}$$

Introduce second order terms with "Hessian" approximations:

$$\boxed{\begin{aligned} &\max_{z,d}\ z + \tfrac{1}{2}d^T H_k d \\ &\ \text{s.t. } f(x_k)e + G_k^T d \leq ze \end{aligned}} \qquad \boxed{\begin{aligned} &\max_{\lambda}\ f(x_k) - \tfrac{1}{2}\lambda^T G_k^T H_k{}^{-1} G_k \lambda \\ &\ \text{s.t. } e^T\lambda = 1,\ \lambda \geq 0 \end{aligned}}$$

How should H_k be chosen?

- We propose quasi-Newton and "overestimation" updating schemes that maintain positive definite and bounded "Hessians".

Global convergence of AGS

Theorem: Let $\sigma, \gamma > 0$ be user-defined constants. Then, for any k, after all updates have been performed for AGS-LBFGS for sample points 1 through $p_k \leq p$, the following holds for any $d \in \mathbb{R}^n$:

$$\left(2^p\left(1+\frac{\sigma}{\gamma^2}\right)^p \mu_k + \frac{1}{\gamma}\left(\frac{2^p\left(1+\frac{\sigma}{\gamma^2}\right)^p - 1}{2\left(1+\frac{\sigma}{\gamma^2}\right) - 1}\right)\right)^{-1} \|d\|^2 \leq d^T H_k d \leq \left(\mu_k + \frac{p\sigma}{\gamma}\right)\|d\|^2.$$

Theorem: Let $\rho \geq 1/2$ be a user-defined constant. Then, for any k, after all updates have been performed for AGS-over for sample points 1 through $p_k \leq p$, the following holds for any $d \in \mathbb{R}^n$:

$$\mu_k \|d\|^2 \leq d^T H_k d \leq \mu_k (2\rho)^p \|d\|^2.$$

Theorem: Let f be locally Lipschitz and continuously differentiable on an open dense $\mathcal{D} \subset \mathbb{R}^n$. Then, w.p.1, $f(x_k) \to -\infty$ or every cluster point of $\{x_k\}$ is stationary for f.

(See Curtis and Que (2011).)

Nonlinear constrained 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 constrained 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?

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-GS in a flash

- The SQP-GS subproblem is

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

where X_k is composed of

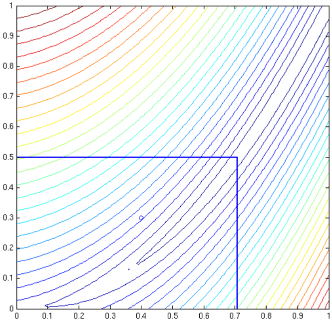
$$\begin{array}{rcccl} X_k^f & = & \{x_k, x_{k1}^f, \ldots, x_{kp}^f\} & \subset & \mathbb{B}_\epsilon(x_k) \cap \mathcal{D}^f \\ \text{and } X_k^{c^i} & = & \{x_k, x_{k1}^{c^i}, \ldots, x_{kp}^{c^i}\} & \subset & \mathbb{B}_\epsilon(x_k) \cap \mathcal{D}^{c^i} \text{ for } i = 1, \ldots, m. \end{array}$$

- This is equivalent to minimizing a model of an exact penalty function $\phi_\rho(x)$:

$$q_\rho(d; X_k, H_k) :=$$
$$\rho \max_{x \in X_k^f} (f(x_k) + \nabla f(x)^T d) + \sum \max_{x \in X_k^{c^i}} \max\{c^i(x_k) + \nabla c^i(x)^T d, 0\} + \tfrac{1}{2} d^T H_k d.$$

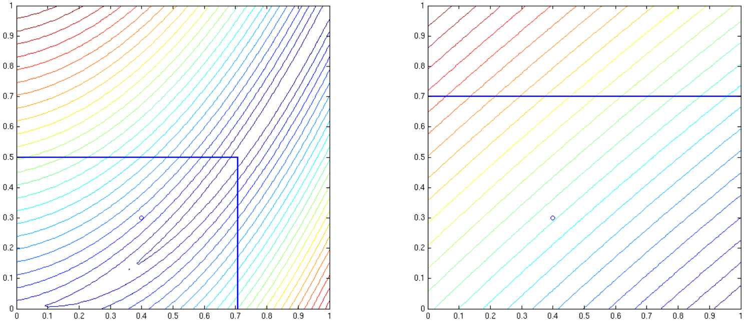
SQP-GS illustration

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



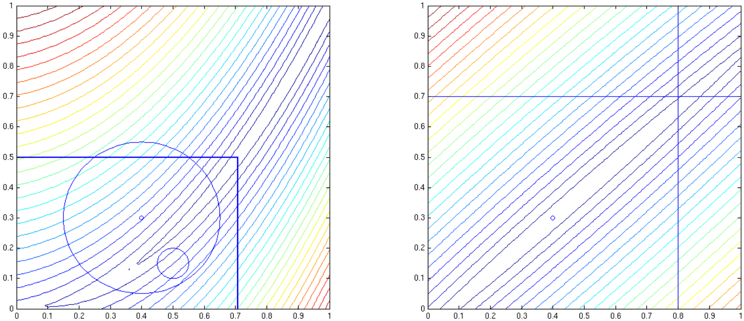
SQP-GS illustration

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



SQP-GS illustration

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



SQP-GS method

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

- Sample $p \geq n+1$ points for each function to generate $X_k = \{X_k^f, X_k^{c^1}, \ldots, X_k^{c^m}\}$.
- Compute d_k by solving the SQP-GS subproblem

$$\begin{aligned} \min_{z,d,s} \ & \rho z + e^T s + \tfrac{1}{2} d^T H_k d \\ \text{s.t. } & f(x_k) + \nabla f(x)^T d \leq z, \ \text{for } x \in X_k^f \\ & c^i(x_k) + \nabla c^i(x)^T d \leq s^i, \ s^i \geq 0, \ \text{for } x \in X_k^{c^i}, \ i = 1, \ldots, m \end{aligned}$$

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

$$\phi_\rho(x_k + \alpha_k d_k) \leq \phi_\rho(x_k) - \eta \alpha_k \Delta q_\rho(d_k; X_k, H_k).$$

- 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 $\Delta q_\rho(d_k; X_k, H_k) \leq \frac{1}{2}\epsilon^2$, then reduce ϵ.
- If ϵ has been reduced and x_k is not sufficiently feasible, then reduce ρ.

Convergence theory for SQP-GS

Theorem: Suppose 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.

Then, w.p.1, one of the following holds true:

- $\rho = \rho_* > 0$ for all large k and every cluster point of $\{x_k\}$ is stationary for ϕ_{ρ_*}. Moreover, with K defined as the infinite subsequence of iterates during which ϵ is decreased, all cluster points of $\{x_k\}_{k \in K}$ are feasible for the optimization problem.
- $\rho \to 0$ and every cluster point of $\{x_k\}$ is stationary for ϕ_0.

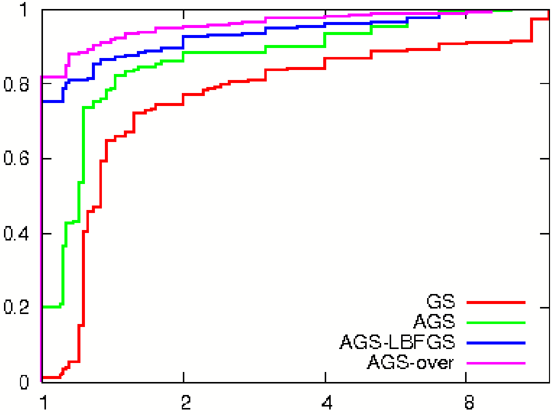
Outline

AGS: Implementation and test details

- Matlab implementation
- QO solver adapted from Kiwiel (1986)
- 26 test problems from Haarala (2004) with $n = 50$
- Each problem run with 10 random starting points
- GS: $p = 2n$ gradients per iteration
- AGS: $p = 2n$ required for full line search, but only 5 gradients per iteration

Performance profile for final ϵ

Limit of 5000 gradient evaluations: GS, 49 iters.; AGS, 833 iters.



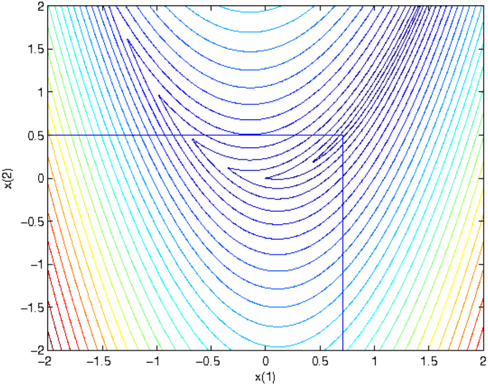
Final $\epsilon \in \{10^{-1}, \ldots, 10^{-12}\}$; performance profile for $\log_{10} \epsilon + 13$.

SQP-GS Implementation

- Matlab implementation
- QO subproblems solved with MOSEK
- BFGS approximations of Hessian of $\phi_\rho(x)$ (as in AGS-LBFGS)
- $p = 2n$ gradients per iteration

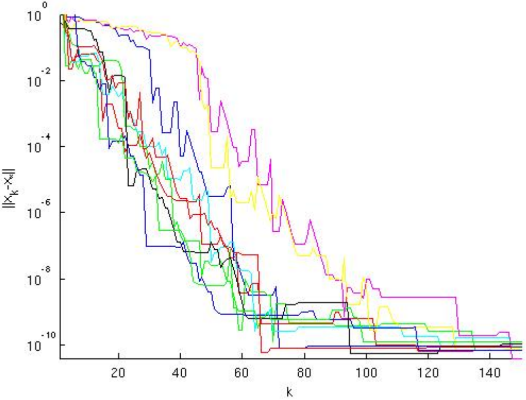
Example 1: Nonsmooth Rosenbrock

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



Example 1: Nonsmooth Rosenbrock

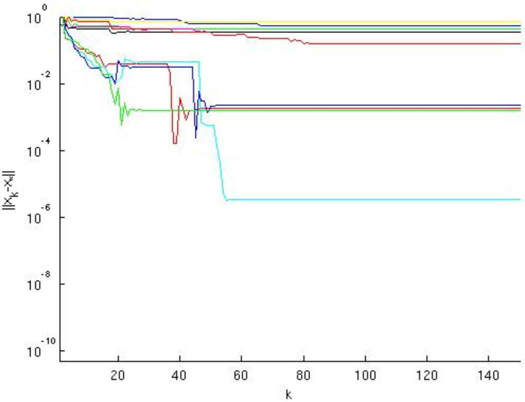
$$\min_{x} \; 10|x_1^2 - x_2| + (1 - x_1)^2 \quad \text{s.t.} \; \max\{\sqrt{2}x_1, 2x_2\} \leq 1.$$



Plot of distance to solution

Example 1: Nonsmooth Rosenbrock

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



Plot of distance to solution (no sampling)

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

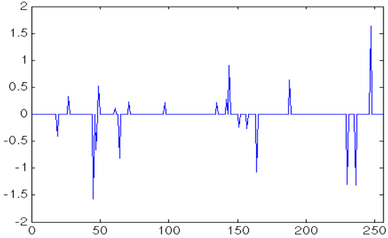
N	K	n	Objective	Infeasibility	Final ϵ	Opt. error
2	1	4	1.0000e+00	3.1752e-14	5.9605e-09	7.6722e-12
4	2	16	7.4630e-01	2.8441e-07	4.8828e-05	1.1938e-04
6	3	36	6.3359e-01	2.1149e-06	9.7656e-05	8.7263e-02
8	4	64	5.5832e-01	2.0492e-05	9.7656e-05	2.7521e-03
10	5	100	2.1841e-01	9.8364e-06	7.8125e-04	9.6041e-03
12	6	144	1.2265e-01	1.8341e-04	7.8125e-04	6.0492e-03
14	7	196	8.4650e-02	1.6692e-04	7.8125e-04	7.1461e-03
16	8	256	6.5051e-02	6.4628e-04	1.5625e-03	3.1596e-03

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

Recover a sparse signal by solving

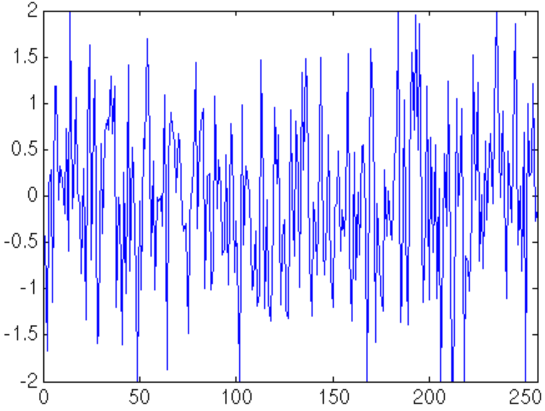
$$\min_x \|x\|_{0.5}$$
$$\text{s.t. } Ax = b$$

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



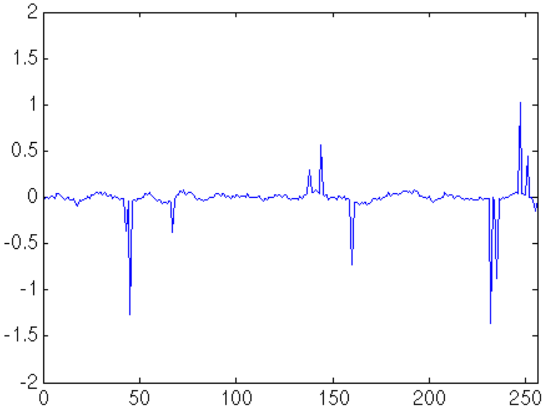
(Use $\ell_{0.5}$ norm as ℓ_1 does not recover sparse solution.)

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



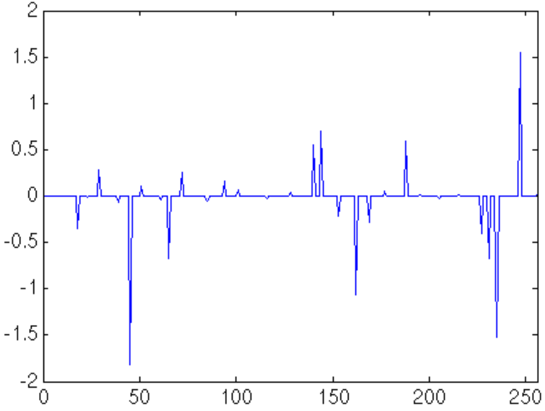
$k = 1$

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



$k = 25$

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



$k = 200$

Example 4: Robust optimization

Find the robust minimizer of a linear objective s.t. an uncertain quadratic constraint:

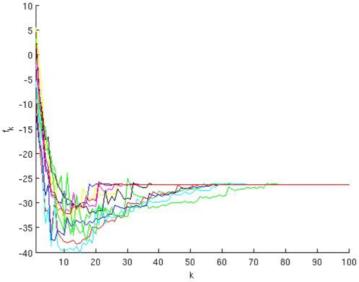
$$\min_x \ f^T x \quad \text{s.t.} \ x^T A x + b^T x + c \leq 0, \ \forall (A, b, c) \in \mathcal{U},$$

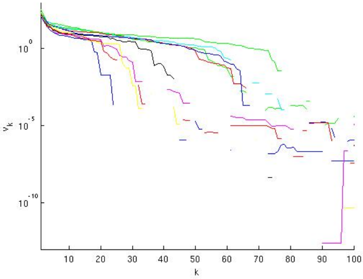
where $f \in \mathbb{R}^n$ and for each (A, b, c) in the uncertainty set

$$\mathcal{U} := \left\{ (A, b, c) : (A, b, c) = (A^{(0)}, b^{(0)}, c^{(0)}) + \sum_{i=1}^{10} u^i (A^{(i)}, b^{(i)}, c^{(i)}), \ u^T u \leq 1 \right\}$$

$A \in \mathbb{R}^{n \times n}$ is positive semidefinite, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

Example 4: Robust optimization





Plot of function values (left) and constraint violation values (right)

Outline

Summary

We set out to improve the practicality and enhance GS methods.

- We aimed to reduce overall gradient evaluations.
- We aimed to reduce the cost of the subproblem solves.
- We aimed to maintain convergence guarantees.
- We aimed to extend the methodology to constrained optimization.

The first goals can be achieved with adaptive sampling and Hessian approximations:

- $O(1)$ gradient evaluations required per iteration
- Subproblem solver warm-started effectively
- Hessian updating schemes improve performance
- Global convergence guarantees maintained

Last goal can be achieved in a SQP-GS framework with constraint gradient sampling:

- Subproblem solve is still a QO per iteration
- Global convergence guarantees maintained

Thanks!

References:

- F. E. Curtis and X. Que, "An Adaptive Gradient Sampling Algorithm for Nonsmooth Optimization," in 2^{nd} review for *Optimization Methods and Software*.
- F. E. Curtis and M. L. Overton, "A Sequential Quadratic Programming Algorithm for Nonconvex, Nonsmooth Constrained Optimization," *SIAM Journal on Optimization*, Volume 22, Issue 2, pg. 474-500, 2012.