A Quasi-Newton Gradient Sampling Algorithm for Nonsmooth Optimization

Frank E. Curtis, Lehigh University

involving joint work with

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

PIMS Workshop on Numerical Linear Algebra and Optimization

8 August, 2013

Outline

Context

Emphasis today on structured problems

- structure often means convex
- seeking sparsity, low matrix rank, low total variation, ...

This talk focuses on unstructured problems

- problems may be nonconvex
- general-purpose algorithms needed

Randomized algorithms for deterministic optimization:

- No gradient info? e.g., simulation-based
- Only some gradient info? e.g., machine learning
- Only some subgradient info? e.g., unstructured nonsmooth

Good theory, computational flexibility, ...

Contributions

Gradient sampling[1] (GS):

- general-purpose nonconvex, nonsmooth optimization
- global convergence guarantees (w.p.1)
- good performance in practice (but expensive!)
- extensions to derivative-free & constrained optimization[2]

Contributions in this talk[3]:

- dramatically reduced per-iteration & overall cost
- dynamic transition from BFGS[4]
- extended methodology to constrained optimization
- global convergence guarantees (w.p.1)
- promising numerical results, and more coming

Open question:

- algorithmic termination for constrained optimization

---

[1] Burke, Lewis, Overton (2005); Kiwiel (2007)
[2] Kiwiel (2010); Tang, Liu, Jian, Li (2012); Hare, Nutini (2013)
[3] Curtis, Que (2012); Curtis, Overton (2012)
[4] Broyden (1970); Fletcher (1970); Goldfarb (1970); Shanno (1970)

Unconstrained nonconvex, nonsmooth optimization

Let f be locally Lipschitz in $\mathbb{R}^n$ and continuously differentiable in an open, dense subset $\mathcal{D}$ of $\mathbb{R}^n$, then consider the minimization problem

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

Define

$$\mathbb{B}_\epsilon(\overline{x}) := \{x : \|x - \overline{x}\|_2 \leq \epsilon\}$$

A point $\overline{x}$ is stationary if

$$0 \in \partial f(\overline{x}) := \bigcap_{\epsilon > 0} \operatorname{cl}\operatorname{conv} \nabla f(\mathbb{B}_\epsilon(\overline{x}) \cap \mathcal{D})$$

A point $\overline{x}$ is ϵ-stationary if

$$0 \in \partial_\epsilon f(\overline{x}) := \operatorname{cl}\operatorname{conv} \partial f(\mathbb{B}_\epsilon(\overline{x}))$$

GS idea

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

$$\begin{array}{lcccccll} X_k & := & \{ & x_{k0}, & x_{k1}, & \cdots, & x_{kp} & \} & \text{(sample points)} \\ G_k & := & [& g_{k0} & g_{k1} & \cdots & g_{kp} &] & \text{(sample gradients)} \end{array}$$

ϵ_k-subdifferential approximated by convex hull of sampled gradients:

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

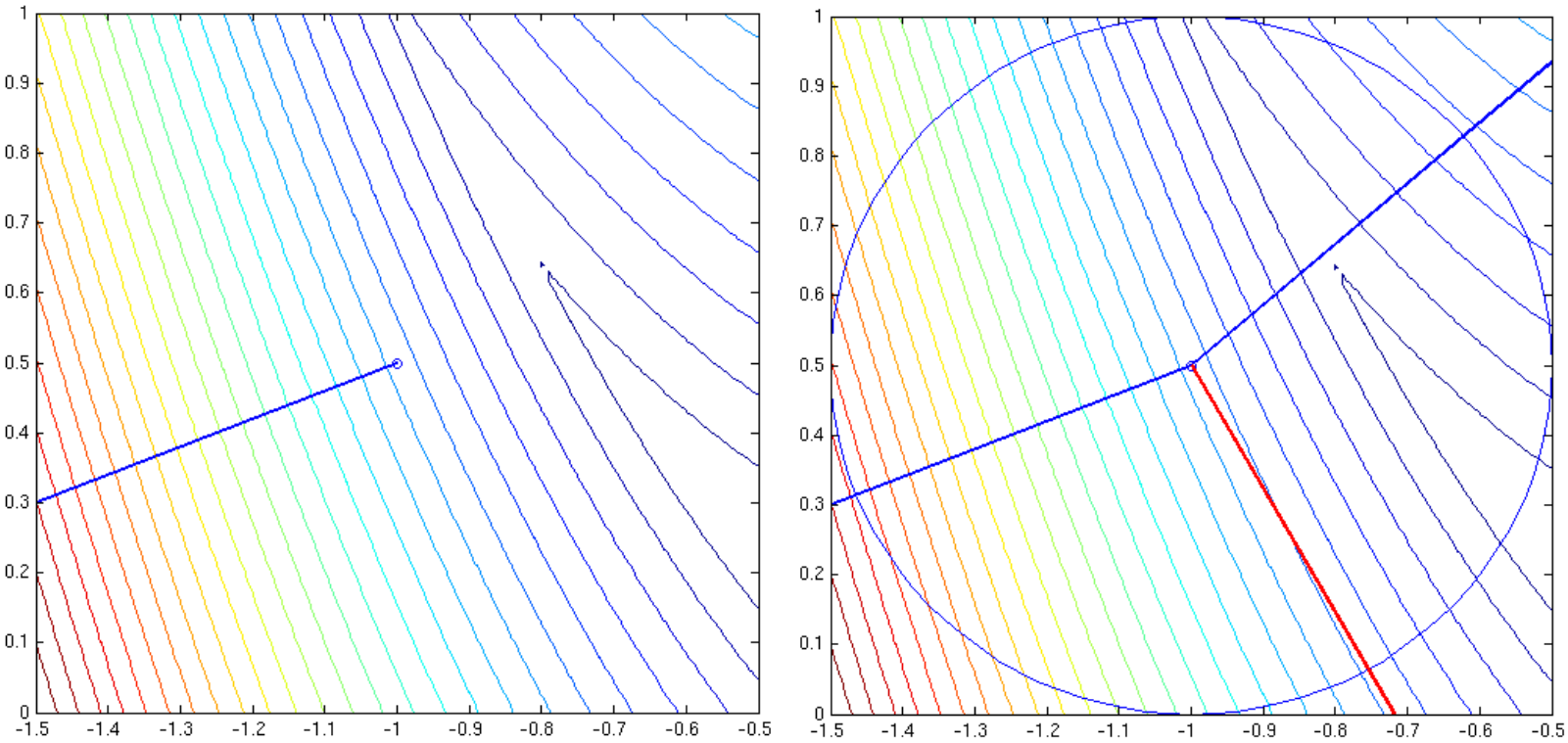
Projection of origin onto convex hull of sampled gradients:

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

$d_k = -g_k$ is an approximate ϵ_k-steepest descent step

GS illustration

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



GS method

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

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

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

- Backtrack from $\alpha_k \leftarrow 1$ to obtain sufficient decrease:

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

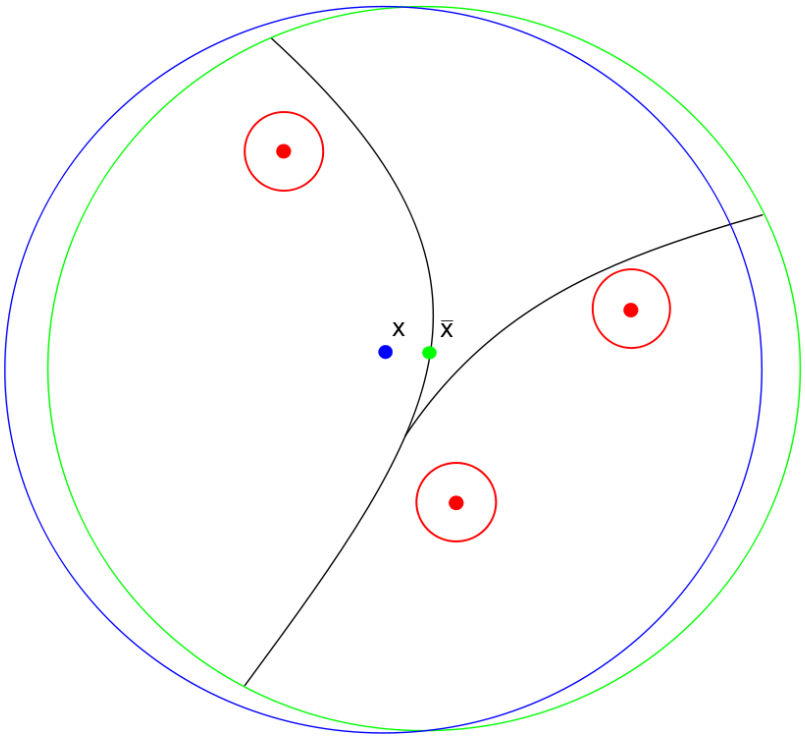
- Update $x_{k+1} \approx x_k + \alpha_k d_k$ (to ensure $x_{k+1} \in \mathcal{D}$)
- If $\|d_k\| \leq \epsilon_k$, then choose $\epsilon_{k+1} < \epsilon_k$; else, set $\epsilon_{k+1} \leftarrow \epsilon_k$

GS global convergence

Theorem: Let f be locally Lipschitz in $\mathbb{R}^n$ and continuously differentiable in an open, dense subset $\mathcal{D}$ of $\mathbb{R}^n$. Then, w.p.1, either

- $f(x_k) \to -\infty$, or
- every cluster point of $\{x_k\}$ is stationary for f

Proof idea: At x_k, either a sufficient descent direction is produced or



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

GS issues

Practical limitations:

- $p \geq n + 1$ gradient evaluations per iteration
- subproblems distinct; solved from scratch
- steepest descent method(?)
- constraints: exact penalty?

Proposed enhancements:

- adaptive sampling[5]; $O(1)$ gradients per iteration
- warm/hot-started subproblem solves
- quasi-Newton "Hessian" approximations
- SQP framework for nonconvex, nonsmooth constraints

Want to use BFGS and gradient sampling ...

[5] Kiwiel (2010)

GS issues

Practical limitations:

- $p \geq n + 1$ gradient evaluations per iteration
- subproblems distinct; solved from scratch
- steepest descent method(?)
- constraints: exact penalty?

Proposed enhancements:

- adaptive sampling[5]; $O(1)$ gradients per iteration
- warm/hot-started subproblem solves
- quasi-Newton "Hessian" approximations
- SQP framework for nonconvex, nonsmooth constraints

Want to use BFGS and gradient sampling . . . without gradient sampling[6]

[5] Kiwiel (2010)
[6] Mitchell (ICCOPT, 2013)

Search direction computation

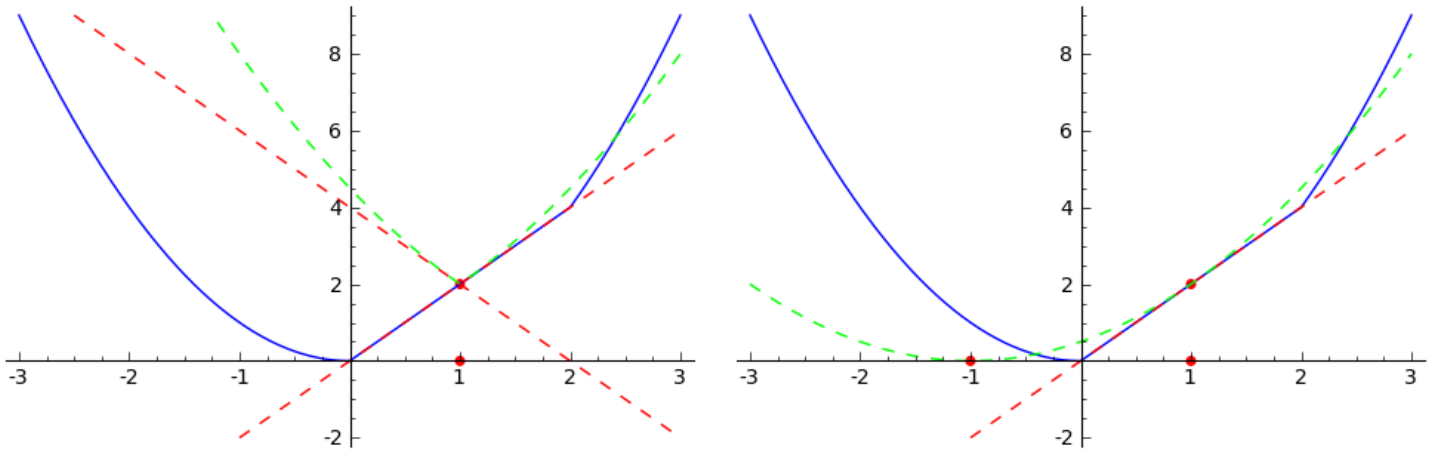
At x_k, suppose we have

$$\begin{array}{lcccccccc} X_k & := & \{ & x_{k0}, & x_{k1}, & \cdots, & x_{kp_k} & \} & \text{(sample points)} \\ G_k & := & [& g_{k0} & g_{k1} & \cdots & g_{kp_k} &] & \text{(sample gradients)} \end{array}$$

Given $H_k \succ 0$ and $W_k := H_k^{-1}$, consider the search direction subproblems

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

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



With $p_k = 0$, we have the quasi-Newton step $d_k \leftarrow -W_k \nabla f(x_k)$

Line search, iterate update, and sampling radius update

Forward/backtracking line search to satisfy the Wolfe conditions

$$f(x_k) - f(x_k + \alpha_k d_k) > \underline{\eta}\alpha_k \|d_k\|^2_{H_k} \qquad // \ \underline{\eta} \in (0,1)$$
$$v^T d_k \geq \overline{\eta} \nabla f(x_k)^T d_k, \text{ where } v \in \partial f(x_k + \alpha d_k) \qquad // \ \overline{\eta} \in (\underline{\eta}, 1)$$

If necessary, perturb $x_k + \alpha_k d_k$ to find $x_{k+1} \in \mathcal{D}$ satisfying

$$f(x_k) - f(x_{k+1}) > \underline{\eta}\alpha_k \|d_k\|^2_{H_k}$$
$$\nabla f(x_{k+1})^T d_k \geq \overline{\eta} \nabla f(x_k)^T d_k$$
$$\|x_k + \alpha_k d_k - x_{k+1}\|_2 \leq \min\{\alpha_k, \epsilon_k\} \|d_k\|_2$$

Reduce the sampling radius (i.e., choose $\epsilon_{k+1} < \epsilon_k$) if

$$\|d_k\|^2_{H_k} \leq \nu \epsilon_k \qquad // \ \nu > 0$$
$$\|d_k\|^2_{H_k} \geq \underline{\xi} \epsilon_k \|d_k\|_2 \qquad // \ \underline{\xi} \in (0,1)$$
$$\alpha_k > 0$$

Finite termination if $f_k(\alpha) := f(x_k + \alpha d_k) - f(x_k)$ weakly lower semismooth[7]

[7] Lewis, Overton (2012); Mifflin (1977); Lemaréchal (1981)

Sample point generation

At x_k, suppose we had

$$\begin{array}{lcccccccl} X_k & := & \{ & x_{k0}, & x_{k1}, & \cdots, & x_{kp} & \} & \text{(sample points)} \\ G_k & := & [& g_{k0} & g_{k1} & \cdots & g_{kp} &] & \text{(sample gradients)} \end{array}$$

If curvature is bounded and step-size sufficiently large in that

$$\begin{aligned} \underline{\xi}\epsilon_k\|d_k\|_2^2 \leq \|d_k\|_{H_k}^2 \leq \overline{\xi}\epsilon_k^{-1}\|d_k\|_2^2 \quad & // \ 0 < \underline{\xi} < \overline{\xi} \\ \underline{\alpha} \leq \alpha_k \quad & // \ 0 < \underline{\alpha} \end{aligned}$$

then erase sample set (i.e., $X_{k+1} \leftarrow \{x_{k+1}\}$ and $p_{k+1} \leftarrow 0$); else,

- discard gradients outside of radius ϵ_{k+1} about x_{k+1}
- maintain sample points within radius; warm/hot-starting
- sample ≥ 1 new gradient(s)
- discard "old gradients" so $p_{k+1} \leq n+1$

Overall,

$$X_{k+1} \leftarrow (X_k \cap \mathbb{B}_{\epsilon_{k+1}}(x_{k+1})) \cup \{x_{k+1}\} \cup \overline{X}_{k+1}$$

where $\overline{X}_{k+1} \subset \mathbb{B}_{\epsilon_{k+1}}(x_{k+1}) \cap \mathcal{D}$

Quasi-Newton updating

If curvature is bounded and step-size sufficiently large in that

$$\underline{\xi}\epsilon_k\|d_k\|_2^2 \leq \|d_k\|_{H_k}^2 \leq \overline{\xi}\epsilon_k^{-1}\|d_k\|_2^2 \quad // \; 0 < \underline{\xi} < \overline{\xi}$$
$$\underline{\alpha} \leq \alpha_k \quad // \; 0 < \underline{\alpha}$$

then standard BFGS update; else, L-BFGS update with pairs satisfying

$$\max\{\|s_j\|_2^2, \|y_j\|_2^2\} \leq \sigma \quad // \; \sigma > 0$$
$$s_j^T y_j \geq \gamma \quad // \; \gamma > 0$$

Theorem[8]: Initializing $H_{k+1} \leftarrow \mu_k I \succ 0$, after m updates we have for any $d \in \mathbb{R}^n$

$$\left(\frac{2^m}{\mu_k}\left(1+\frac{\sigma^2}{\gamma^2}\right)^m + \frac{\sigma}{\gamma}\left(\frac{2^m\left(1+\frac{\sigma^2}{\gamma^2}\right)^m - 1}{2\left(1+\frac{\sigma^2}{\gamma^2}\right) - 1}\right)\right)^{-1}\|d\|_2^2 \leq \|d\|_{H_{k+1}}^2 \leq \left(\mu_k + \frac{m\sigma}{\gamma}\right)\|d\|_2^2$$

[8] Curtis, Que (2012)

BFGS-GS method

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

- Compute $d_k \leftarrow -W_k G_k y_k$ via the dual subproblem

$$\min_y \ \tfrac{1}{2}\|G_k y\|^2_{W_k}$$
$$\text{s.t. } e^T y = 1, \ y \geq 0$$

- Forward/backtracking line search to obtain α_k satisfying Wolfe conditions
- Perturb (if necessary) to obtain new iterate x_{k+1}
- Set sampling radius ϵ_{k+1}
- Set sample set X_{k+1}
- Compute any unknown elements of G_{k+1}
- Set (L-)BFGS "Hessian" approximation W_{k+1}

Adaptive sampling

Testing adaptive sampling on 26 problems; 5000 gradient evaluation limit

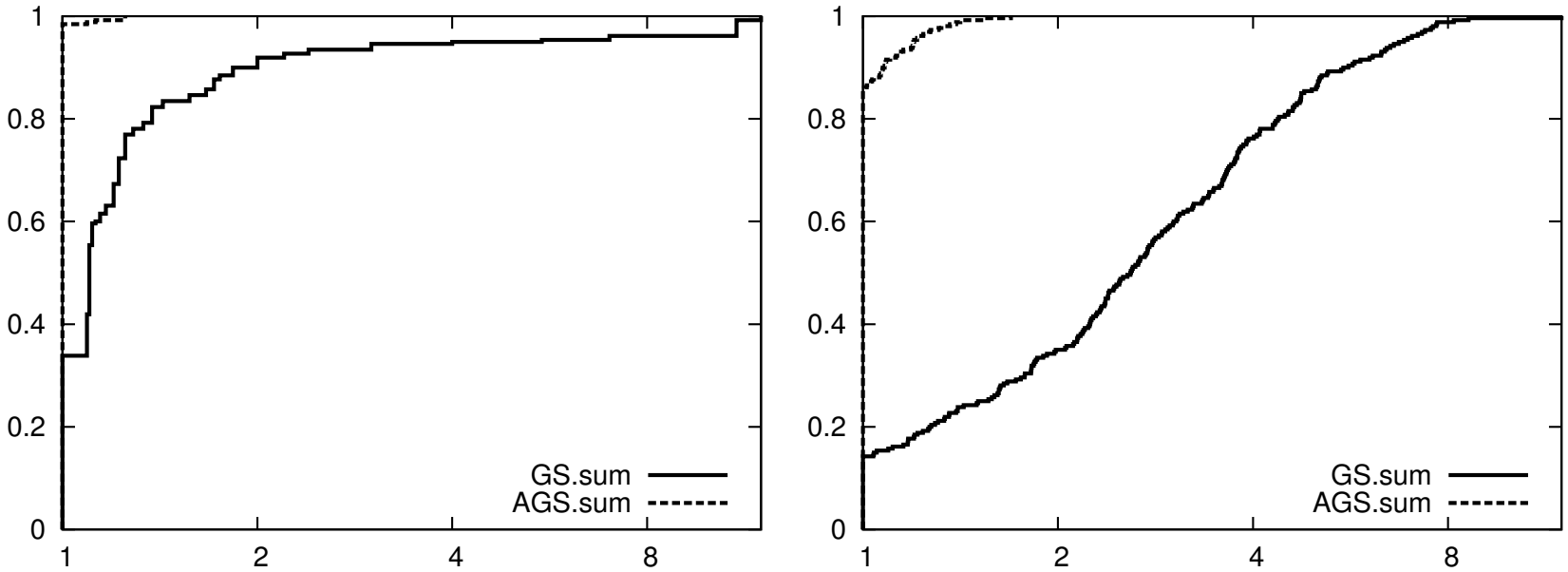


Figure: Final sampling radius (left) and average QP per nonlinear iteration (right)

Quasi-Newton updating

Testing adaptive sampling on 26 problems; 5000 gradient evaluation limit

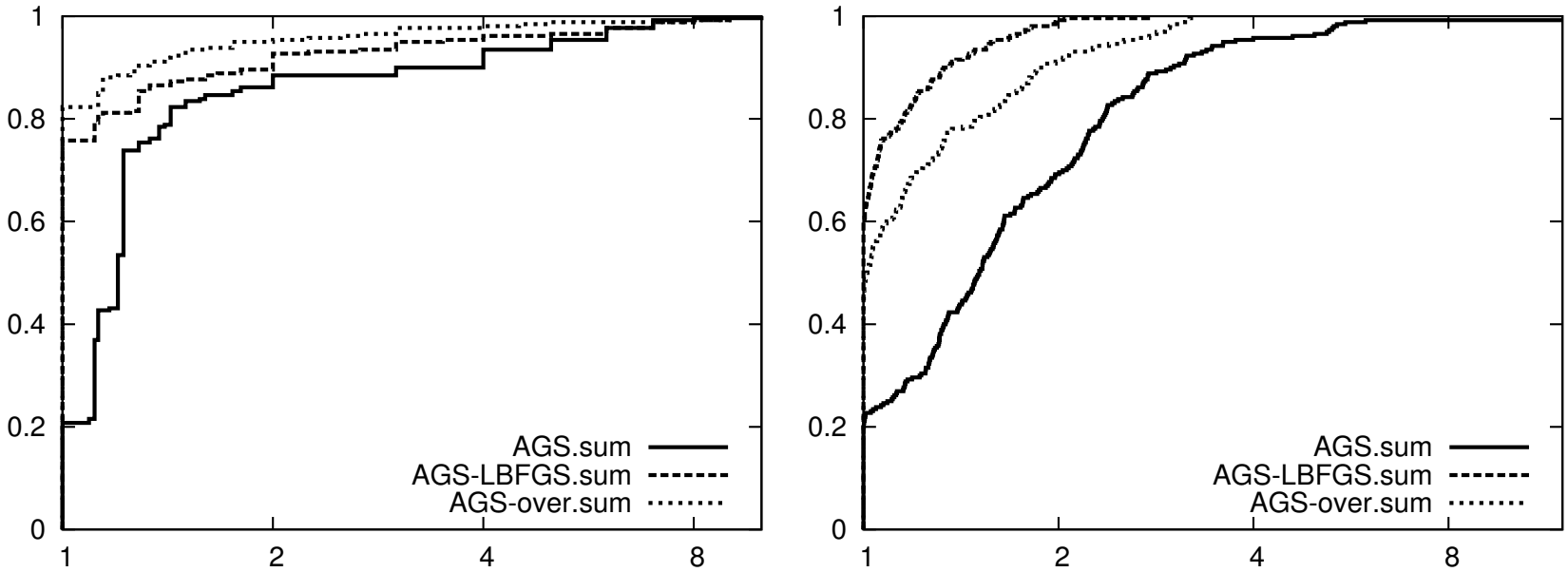


Figure: Final sampling radius (left) and average QP per nonlinear iteration (right)

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

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

where X_k is composed of

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

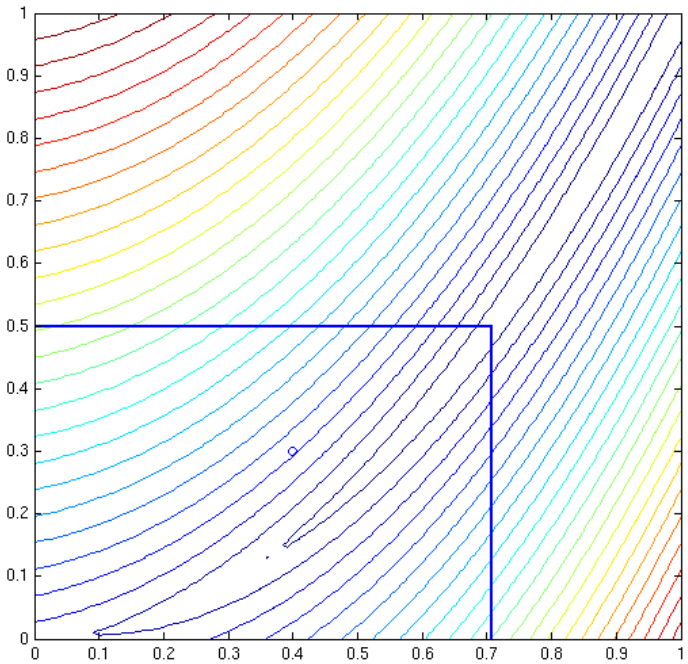
Equivalent to minimizing a model of $\phi_\rho(x) := \rho f(x) + \| \max\{c(x), 0\} \|_1$:

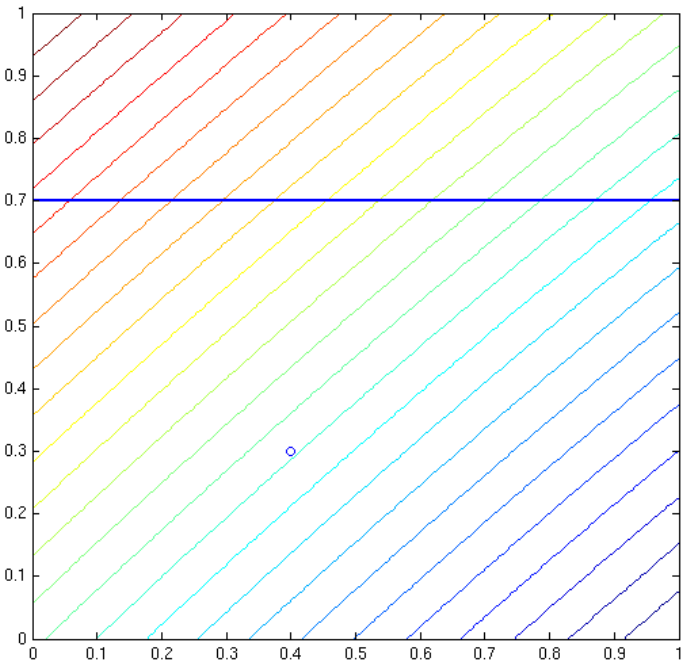
$$\begin{aligned} & q_\rho(d; X_k, H_k) := \\ & \rho \max_{x \in X_k^f} (f(x_k) + \nabla f(x)^T d) + \sum \max_{x \in X_k^{c^i}} \max\{c^i(x_k) + \nabla c^i(x)^T d, 0\} + \tfrac{1}{2} d^T H_k d \end{aligned}$$

SQP-GS illustration

$$\min_x \; 10|x_2 - x_1^2| + (1 - x_1)^2 \quad \text{s.t.} \;\; \max\{\sqrt{2}x_1, 2x_2\} - 1 \leq 0$$

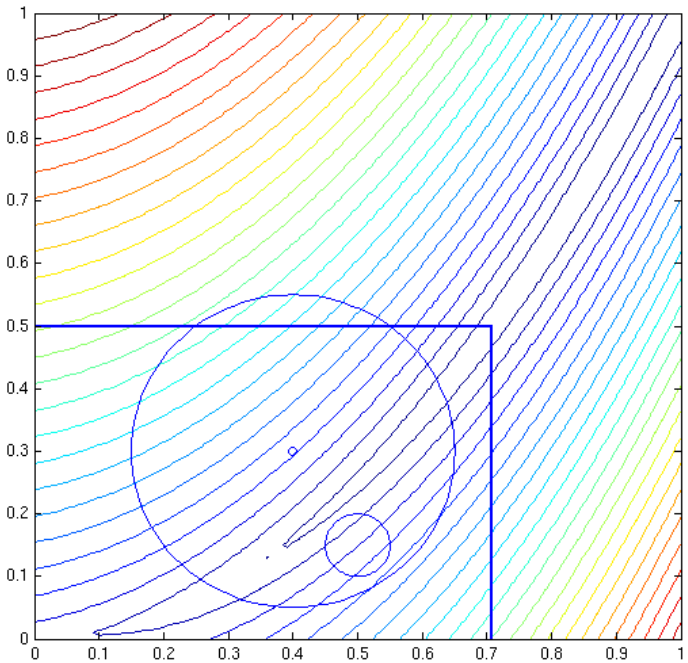
at $x_k = (\frac{2}{5}, \frac{3}{10})$

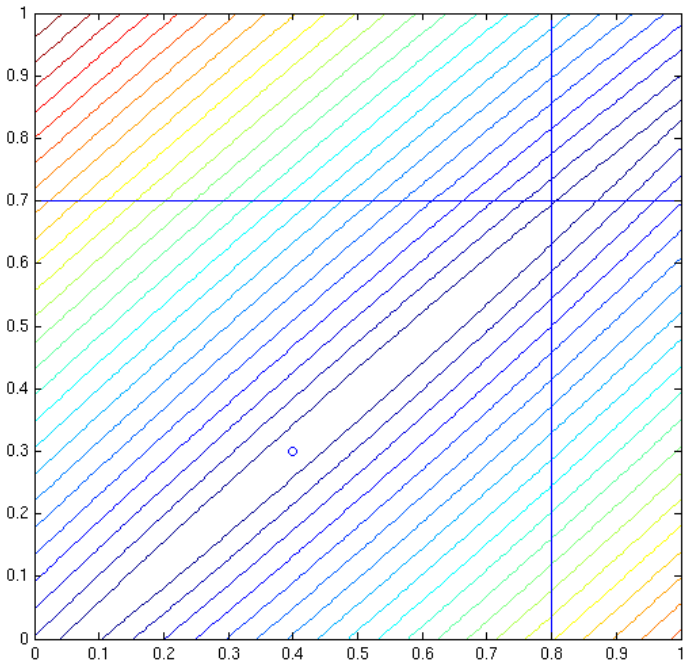




SQP-GS illustration

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





SQP-GS method

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

- Sample $p \geq n + 1$ points for each function to get $X_k = \{X_k^f, X_k^{c^1}, \ldots, X_k^{c^m}\}$
- 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 $\Delta q_\rho(d_k; X_k, H_k) \leq \frac{1}{2}\epsilon_k^2$, then set $\epsilon_{k+1} < \epsilon_k$; else, set $\epsilon_{k+1} \leftarrow \epsilon_k$
- If $\epsilon_{k+1} < \epsilon_k$ and x_k is not sufficiently feasible, then reduce ρ

Convergence theory for SQP-GS

Theorem[9]: 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
- $\rho \to 0$ and every cluster point of $\{x_k\}$ is stationary for ϕ_0

[9] Curtis, Overton (2012)

Nonsmooth Rosenbrock

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

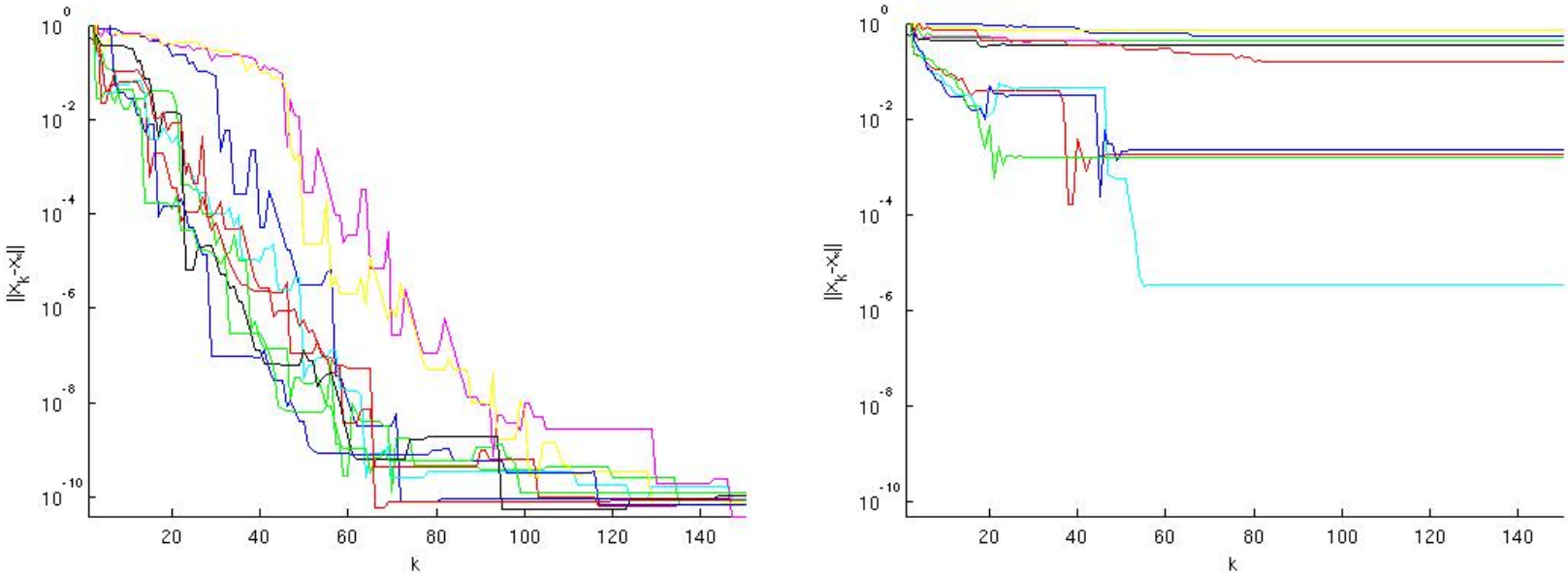


Figure: With gradient sampling (left) and without sampling (right)

$\ell_{0.5}$ 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

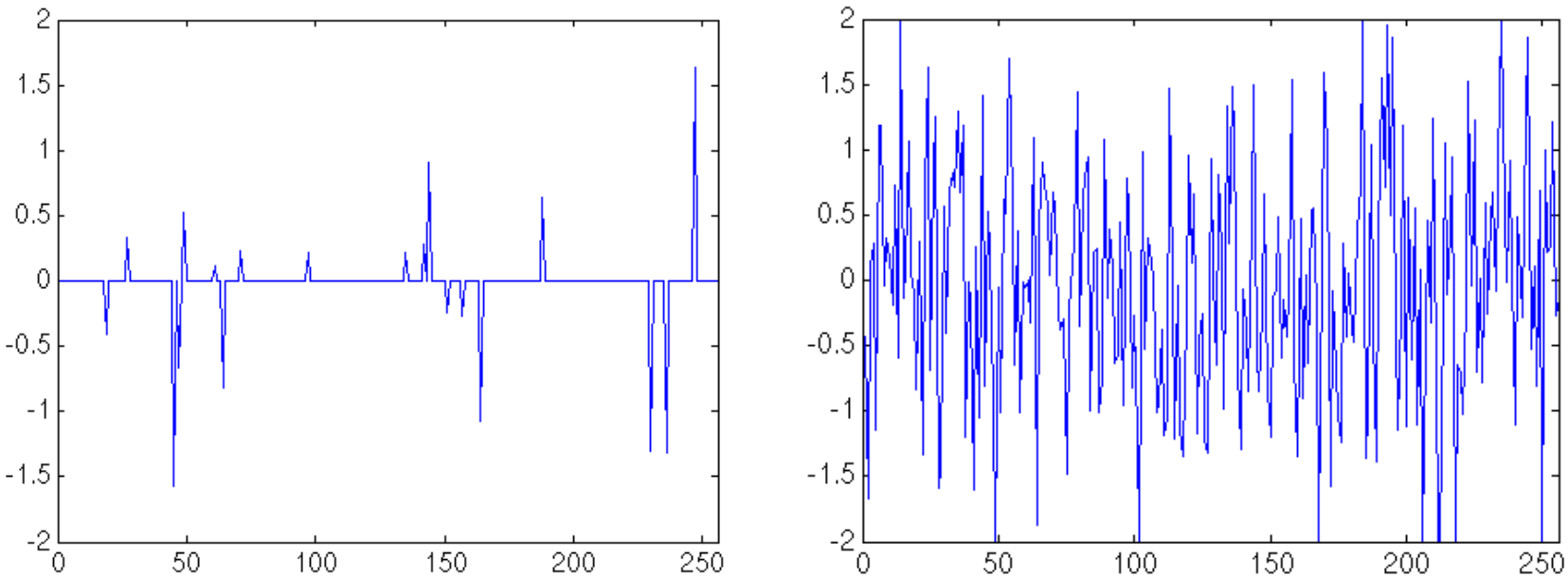


Figure: True solution (left) and iterate at $k = 1$ (right)

$\ell_{0.5}$ 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

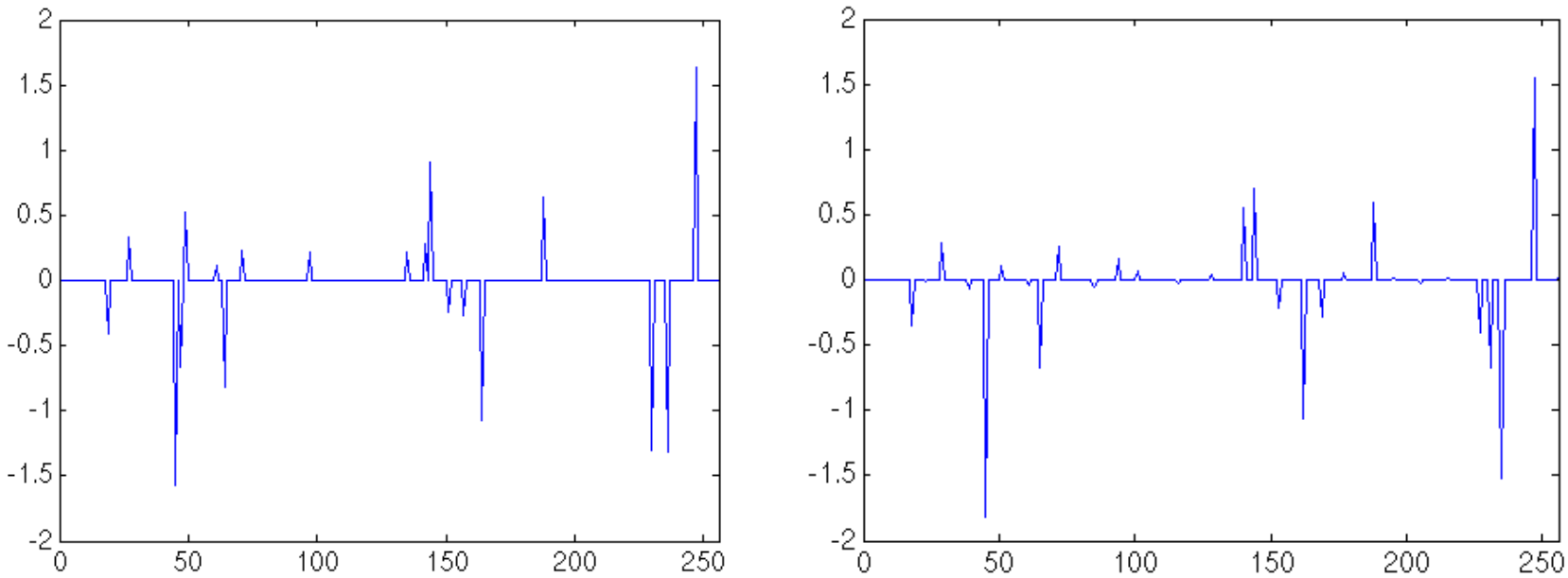


Figure: True solution (left) and iterate at $k = 200$ (right)

Contributions

Gradient sampling (GS):

- general-purpose nonconvex, nonsmooth optimization
- global convergence guarantees (w.p.1)
- good performance in practice (but expensive!)
- extensions to derivative-free & constrained optimization

Contributions:

- dramatically reduced per-iteration & overall cost
- natural transition from BFGS
- extended methodology to constrained optimization
- global convergence guarantees (w.p.1)
- promising numerical results, and more coming

Open question

To test for ϵ_k-stationarity, GS employs the condition[10]

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

Can a random sampling of gradients be used to test *constrained* optimality?

- Idea 1: Use GS ideas to test for stationarity of an exact penalty function:

$$\partial\phi_\rho(x) \ni 0$$

- Idea 2: Use SQP-GS ideas to approximate the optimal Lagrange multiplier λ and test for stationarity of the Lagrangian with respect to x:

$$\begin{aligned} \partial_x L(x,\lambda) &\ni 0 \\ \lambda^j c^j(x) &= 0, \;\; j = 1, \ldots, m \end{aligned}$$

[10] Burke, Lewis, Overton (2002)

Thanks!

References:

- F. E. Curtis and X. Que, "A Quasi-Newton Gradient Sampling Algorithm for Nonsmooth, Nonconvex Optimization," Lehigh University, working paper.
- F. E. Curtis and X. Que, "An Adaptive Gradient Sampling Algorithm for Nonsmooth Optimization," *Optimization Methods and Software*, DOI: 10.1080/10556788.2012.714781, 2012.
- 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.