Nonsmooth Optimization via Gradient Sampling

Frank E. Curtis, Lehigh University

involving joint work with

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

Foundations of Computational Mathematics (FoCM) Conference

July 6, 2011

Outline

Gradient Sampling (GS)

Adaptive Variable-Metric GS

Numerical Results

Final Remarks

Outline

Gradient Sampling (GS)

Adaptive Variable-Metric GS

Numerical Results

Final Remarks

Unconstrained optimization of nonsmooth functions

Consider the unconstrained problem

$$\boxed{\min_{x} \; f(x)}$$

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

Unconstrained optimization of nonsmooth functions

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

$$\mathbb{B}(x', \epsilon) := \{x \mid \|x - x'\| \leq \epsilon\}$$

- x' is stationary if

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

- x' is ϵ-stationary if

$$0 \in \partial f(x', \epsilon) = \operatorname{cl}\operatorname{conv} \partial f(\mathbb{B}(x', \epsilon))$$

Gradient sampling (GS) idea

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

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

Then, the ϵ-subdifferential is approximated by the convex hull of nearby gradients:

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

Gradient sampling (GS) idea

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

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

Then, the ϵ-subdifferential is approximated by the convex hull of nearby gradients:

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

- Approximate ϵ-steepest descent step obtained from

$$\begin{aligned} \min_{\lambda} \ & \tfrac{1}{2}\|G_k \lambda\|^2 \\ \text{s.t. } & e^T \lambda = 1, \ \lambda \geq 0 \end{aligned}$$

That is, $d_k = -G_k \lambda_k$ is the projection of 0 onto $\operatorname{conv}\{g_{k0}, g_{k1}, \ldots, g_{kp}\}$

GS method

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

- Sample $p \geq n + 1$ points $\{x_{k1}, \ldots, x_{kp}\} \subset \mathbb{B}(x_k, \epsilon) \cap \mathcal{D}$
- Compute $d_k \leftarrow -G_k\lambda_k$ by solving the quadratic optimization (QO) subproblem

$$\begin{aligned} \min_{\lambda} \ & \tfrac{1}{2}\|G_k\lambda\|^2 \\ \text{s.t. } & e^T\lambda = 1, \ \lambda \geq 0 \end{aligned}$$

- 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

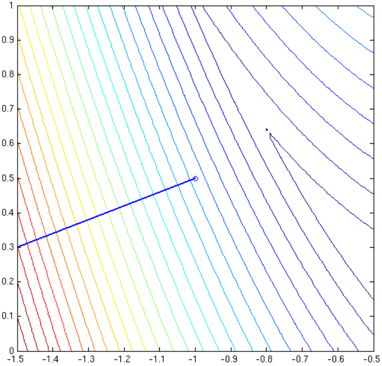
Global convergence of GS

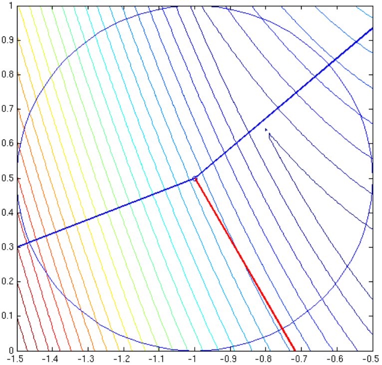
Theorem: Let f be locally Lipschitz and continuously differentiable on an open dense $\mathcal{D} \subset \mathbb{R}^n$. Then, w.p.1, $f(x_k) \downarrow \infty$ or every cluster point of $\{x_k\}$ is stationary for f

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

GS illustration

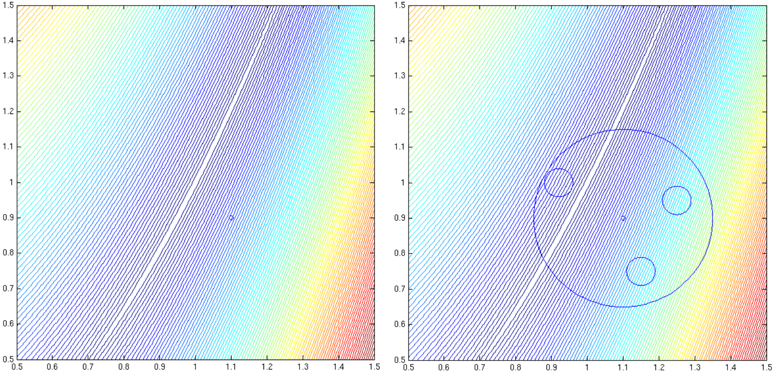
$$\min_x \; f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2 \quad \text{at } x_k = (-1, \tfrac{1}{2})$$





GS illustration

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



Global convergence of GS

Recall the GS (dual) subproblem:

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

Here is the corresponding primal subproblem:

$$\min_{d}\ q(d; X_k) := f(x_k) + \max_{x\in X_k}\{\nabla f(x)^T d\} + \tfrac{1}{2}\|d\|^2$$

Solving this subproblem yields

$$\Delta q(d_k; X_k) := q(0; X_k) - q(d_k; X_k) = \tfrac{1}{2}\|d_k\|^2$$

Also consider the subproblem

$$\min_{d}\ \widetilde{q}(d; x', \epsilon) := f(x') + \max_{x\in\mathbb{B}(x',\epsilon)\cap\mathcal{D}}\{\nabla f(x)^T d\} + \tfrac{1}{2}\|d\|^2$$

Global convergence of GS

Let

$$\mathcal{S}(x_k, \epsilon) = \prod_1^p (\mathbb{B}(x_k, \epsilon) \cap \mathcal{D})$$

and

$$\mathcal{T}(x_k, \epsilon, x', \omega) = \{X_k \in \mathcal{S}(x_k, \epsilon) \mid \Delta q(d_k; X_k) \leq \Delta \widetilde{q}(d'; x', \epsilon) + \omega\}$$

Lemma: 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 $\mathcal{T} \subset \mathcal{T}(x_k, \epsilon, x', \omega)$

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

Sketch of proof: Follows mainly from Carathéodory's theorem

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

Sketch of proof: If $\epsilon \nrightarrow 0$, then for all large k

$$\Delta q(d_k; X_k) = \tfrac{1}{2}\|d_k\|^2 > \epsilon^2/2$$

However, with probability 1, this will not occur

- $\epsilon \nrightarrow 0$ implies $x_k \to x'$. If x' is ϵ-stationary, then w.p.1 we will obtain a sample set yielding $\Delta q(d_k; X_k) \leq \epsilon^2/2$, contradicting the above
- $\epsilon \nrightarrow 0$ also implies $\alpha_k \to 0$. If x' is not ϵ-stationary, then w.p.1 we obtain a subsequence with α_k bounded away from zero, contradicting $\alpha_k \to 0$

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

Practical issues

Practical limitations of GS:

- $p \geq n + 1$ gradient evaluations per iteration
- All subproblems solved from scratch
- Behaves like steepest descent(?)

Practical issues

Practical limitations of GS:

- $p \geq n + 1$ gradient evaluations per iteration
- All subproblems solved from scratch
- Behaves like steepest descent(?)

Proposed solutions:

- Adaptive sampling; $O(1)$ gradients per iteration: Kiwiel (2010)
- Warm-started subproblem solves
- "Hessian" approximations for quadratic term

Outline

Gradient Sampling (GS)

Adaptive Variable-Metric GS

Numerical Results

Final Remarks

Adaptive sampling (AGS)

At x_k, we had

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

At x_{k+1}, we

- maintain sample points still within radius ϵ
- throw out gradients outside of radius
- sample 1 (or some) new gradients

How can we maintain global convergence?

Adaptive sampling (AGS)

At x_k, we had

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

At x_{k+1}, we

- maintain sample points still within radius ϵ
- 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

Recall the GS (dual) subproblem:

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

Here is the corresponding primal subproblem:

$$\min_{d} \; f(x_k) + \max_{x \in X_k}\{\nabla f(x)^T d\} + \tfrac{1}{2} d^T d$$

Primal-dual pair of subproblems (variable-metric)

Recall the GS (dual) subproblem:

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

Here is the corresponding primal subproblem:

$$\min_{d} \;\; f(x_k) + \max_{x \in X_k}\{\nabla f(x)^T d\} + \tfrac{1}{2} d^T H_k^{-1} d$$

How should H_k be chosen?

Quasi-Newton updating (AGS-BFGSa)

Consider the model

$$m_{k+1}(d) = f(x_{k+1}) + \nabla f(x_{k+1})^T d + \tfrac{1}{2} d^T H_{k+1}^{-1} d$$

Matching the gradients of f and m_{k+1} at x_k yields the secant equation

$$H_{k+1}(\nabla f(x_{k+1}) - \nabla f(x_k)) = x_{k+1} - x_k$$

Minimizing changes in $\{H_k\}$ yields the well-known BFGS update

Questions:

- Effective within GS?
- Making best use of info?
- Ill-conditioning: Bad or good?

Quasi-Newton updating (AGS-BFGSb)

Consider BFGS, but instead of updating between iterations, update during

- For each k, initialize $H_k \leftarrow I$
- Imagine moving along each $d_{ki} = x_{ki} - x_k$ and apply BFGS update

Quasi-Newton updating (AGS-BFGSc)

Our model is actually more like

$$m_k(d) = f(x_k) + \max_{x \in X_k}\{\nabla f(x)^T d\} + \tfrac{1}{2} d^T H_{k+1}^{-1} d$$

If we knew the optimal dual solution in advance, then m_k shares a minimizer with

$$\widetilde{m}_k(d) = f(x_k) + \lambda_k^T G_k^T d + \tfrac{1}{2} d^T H_{k+1}^{-1} d$$

Matching the gradients of f and m_k at x_{k-1} yields the secant equation

$$H_{k+1}(G_k \lambda_k - G_{k-1}\lambda_{k-1}) = x_k - x_{k-1}$$

Minimizing changes in $\{H_k\}$ yields a BFGS-like update

Overestimation (AGS-over)

Suppose we also have function values at sample points

- Try to choose H_k so that the following model overestimates f:

$$m_k(d) = f(x_k) + \max_{x \in X_k} \{\nabla f(x)^T d\} + \tfrac{1}{2} d^T H_k^{-1} d$$

- If $m_k(d_{ki}) < f(x_{ki})$, then "lift" H_k so that $m_k(d_{ki}) = f(x_{ki})$
- Updates we use have the form $H_k \leftarrow M^T H_k M$ where

$$M = \frac{1}{(1+\gamma)^{1/n}} \left(I + \frac{\gamma}{d_{ki}^T d_{ki}} d_{ki} d_{ki}^T \right)$$

- This update ensures contours maintain the same volume

Outline

Gradient Sampling (GS)

Adaptive Variable-Metric GS

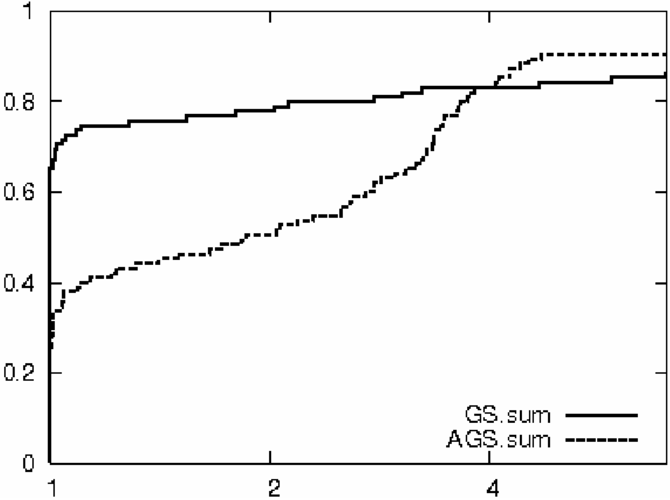
Numerical Results

Final Remarks

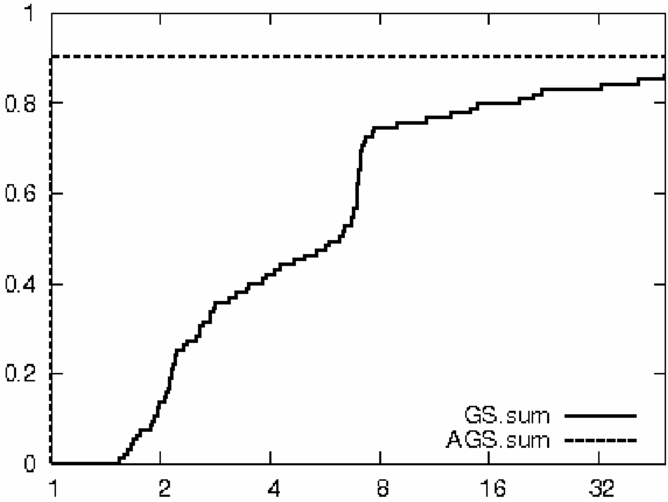
Implementation and test details

- Matlab implementation
- QO solver adapted from Kiwiel (1986)
- Test problems from Haarala (2004) with $n = 10$
- GS: $p = 2n$ gradients per iteration
- AGS: 2 gradient evaluations per iteration
- AGS: $p = 2n$ required for line search
- Optimality tolerance set to 1e−4

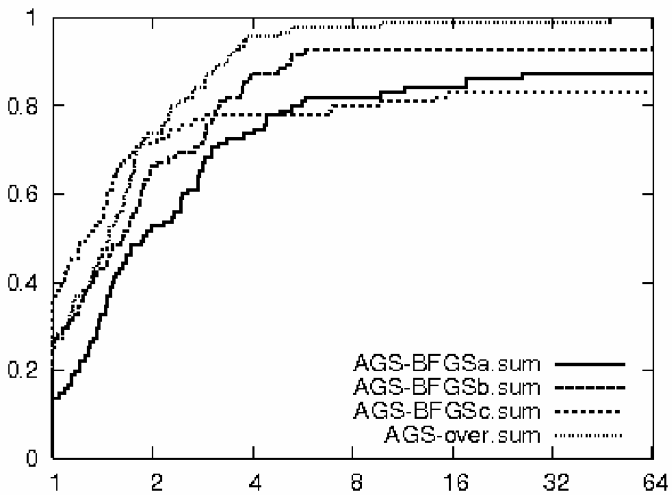
GS vs. AGS: Iterations



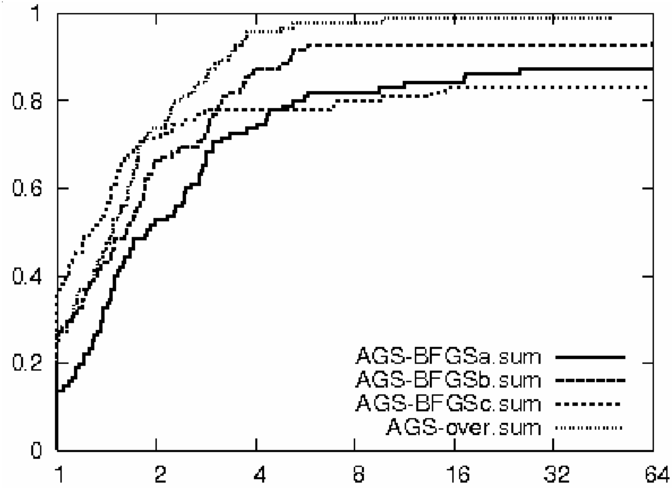
GS vs. AGS: Gradient evaluations



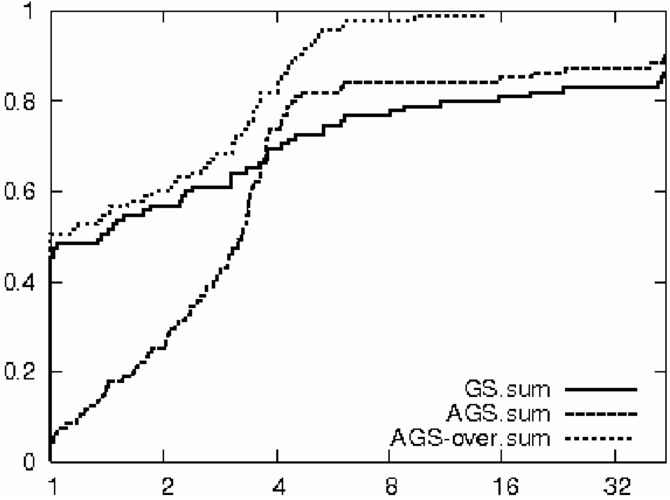
Hessian options: Iterations



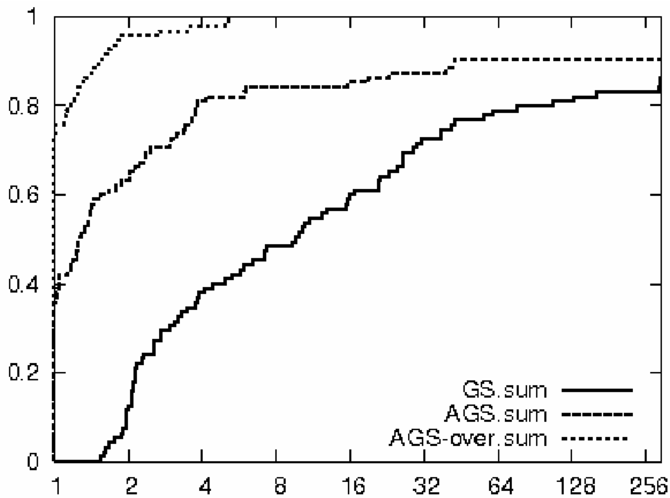
Hessian options: Gradient evaluations



GS vs. AGS vs. AGS-over: Iterations



GS vs. AGS vs. AGS-over: Gradient evaluations



Outline

Gradient Sampling (GS)

Adaptive Variable-Metric GS

Numerical Results

Final Remarks

Summary

We set out to improve the practicality of 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

These goals can be achieved with adaptive sampling and variable-metric variants

- $O(1)$ gradient evaluations required per iteration
- Subproblem solver warm-started effectively
- Hessian updating schemes improve overall iteration count

Future work

- C++ implementation
- Convergence theory for $H_k \succ 0$ (essentially finished)
- Hessian update that maintains $H_k \succ 0$(?)
- Extend to SQP methods for constrained problems (Curtis and Overton, 2011(?))