Nonconvex, Nonsmooth Optimization by Gradient Sampling

Frank E. Curtis, Lehigh University

involving joint work with

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

Northwestern IEMS Department Seminar

March 6, 2012
Outline

Motivations

Gradient Sampling (GS)

Adaptive GS

SQP-GS

Future Work
Outline

Motivations

Gradient Sampling (GS)

Adaptive GS

SQP-GS

Future Work
Nonlinear/convex optimization research

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.
- Nemirovski, Nesterov, Wright, ...
- d’Aspremont, Lan, Recht, Yin, ...
- Focus on large-scale problems needing only an approximate solution.
- First-order methods, optimal algorithms, regularization, ...
I am interested in algorithms for **unstructured** nonlinear optimization.

- For one thing, unstructured means nonconvex.
- Other work: Inexact Newton methods for large-scale optimization.
- Other work: Model/data inconsistencies leading to infeasibility and degeneracy.
- This talk: Enhancing practical NLO methods for handling nonsmoothness.

Widespread use of optimization requires accommodating algorithms.

- Accommodating algorithms can be the “go-to” methods for new problems.
- Accommodating algorithms are all we have for very hard problems.
Deterministic optimization methods based on randomized models

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

- 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 and other benefits.

- DFO: randomization leads to better poised models.
- SO: (batch) stochastic gradient methods have nice practical/theoretical behavior.
- UO: gradient sampling...
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.
Outline

Motivations

Gradient Sampling (GS)

Adaptive GS

SQP-GS

Future Work
Unconstrained nonconvex, nonsmooth optimization

Consider the unconstrained problem

\[
\min_x f(x)
\]

where \( f \) is locally Lipschitz and continuously differentiable in (dense) \( D \subset \mathbb{R}^n \).
Unconstrained nonconvex, nonsmooth optimization

Consider the unconstrained problem

$$\min_x f(x)$$

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

- Let
  $$B_\epsilon(x) := \{x \mid \|x - x\| \leq \epsilon\}$$

- $x$ is stationary if
  $$0 \in \partial f(x) := \bigcap_{\epsilon > 0} \text{cl conv} \nabla f(B_\epsilon(x) \cap D)$$

- $x$ is $\epsilon$-stationary if
  $$0 \in \partial_\epsilon f(x) := \text{cl conv} \partial f(B_\epsilon(x))$$
Gradient sampling (GS) idea

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

$X_k := \{x_{k0}, x_{k1}, \ldots, x_{kp}\}$  (sample points)

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

The $\epsilon$-subdifferential is approximated by the convex hull of the sampled gradients:

$\partial_\epsilon f(x_k) = \text{cl conv } \partial f(B_\epsilon(x_k))$

$\approx \text{conv}\{g_{k0}, g_{k1}, \ldots, g_{kp}\}$
Gradient sampling (GS) idea

At $x_k$, let $x_0 := x_k$ and sample $\{x_{k1}, \ldots, x_{kp}\} \subset B_\epsilon(x_k) \cap D$, yielding:

\[ X_k := \{x_0, x_{k1}, \ldots, x_{kp}\} \quad \text{(sample points)} \]

\[ G_k := \begin{bmatrix} g_0 & g_{k1} & \cdots & g_{kp} \end{bmatrix} \quad \text{(sample gradients)} \]

The $\epsilon$-subdifferential is approximated by the convex hull of the sampled gradients:

\[ \partial_\epsilon f(x_k) = \text{cl conv} \partial f(B_\epsilon(x_k)) \approx \text{conv}\{g_0, g_{k1}, \ldots, g_{kp}\} \]

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

\[ g_k := \text{Proj}(0 \mid \text{conv}\{g_0, g_{k1}, \ldots, g_{kp}\}) \]

Then, $d_k = -g_k$ is an approximate $\epsilon$-steepest descent step.
GS illustration

\[
\min_x 10|x_2 - x_1^2| + (1 - x_1)^2 \quad \text{at } x_k = (-1, \frac{1}{2})
\]
**GS illustration**

\[
\min_{\mathbf{x}} 10|\mathbf{x}_2 - \mathbf{x}_1^2| + (1 - \mathbf{x}_1)^2 \text{ at } \mathbf{x}_k = (1.1, 0.9)
\]
GS method

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

- Sample $p \geq n + 1$ points $\{x_{k1}, \ldots, x_{kp}\} \subset B_\epsilon(x_k) \cap \mathcal{D}$.
- 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 $\epsilon$.

(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

Near $\bar{x}$, the GS algorithm ideally computes $\text{Proj}(0 | \partial_\epsilon f(\bar{x}))$. 
Illustration of critical part of proof

By continuity, there exists \( \{y_{ki}\}_{i=1,...,p} \) such that

\[
\text{Proj}(0|\{\nabla f(y_{ki})\}) \approx \text{Proj}(0|\partial_{\epsilon} f(\bar{x})).
\]
Illustration of critical part of proof

The same holds for sufficiently small neighborhoods about the $y_{ki}$'s.
Illustration of critical part of proof

Far from $\bar{x}$, the algorithm does not necessarily approximate $\text{Proj}(0|\partial_\epsilon f(\bar{x}))$ well.
Illustration of critical part of proof

However, it can in a sufficiently small neighborhood of $\bar{x}$. 
Local models in GS

Computing the projection is equivalent to solving the dual subproblem:

$$\max_{\lambda} f(x_k) - \frac{1}{2} \|G_k \lambda\|^2$$

s.t. $$e^T \lambda = 1, \lambda \geq 0.$$  

The corresponding primal subproblem is to compute $$d_k$$ to minimize

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

If all gradients about $$x$$ were available, then we would ideally compute $$\bar{d}$$ minimizing

$$q(d; B_\epsilon(x) \cap D) = f(x) + \max_{x \in B_\epsilon(x) \cap D} \{\nabla f(x)^T d\} + \frac{1}{2} \|d\|^2.$$
Critical lemma

Let the sample space be

\[ S_\varepsilon(x_k) := \{x_k\} \times \prod_{1}^{p}(B_\varepsilon(x_k) \cap D) \]

and consider the set

\[ T_{\varepsilon,\omega}(x_k, \bar{x}) = \{X_k \in S_\varepsilon(x_k) \mid \Delta q(d_k; X_k) \leq \Delta q(\bar{d}; B_\varepsilon(\bar{x}) \cap D) + \omega\}. \]

Lemma: For any \( \omega > 0 \), there exists \( \zeta > 0 \) and a nonempty set \( T \) such that for all \( x_k \in B(\bar{x}, \zeta) \) we have \( T \subset T_{\varepsilon,\omega}(x_k, \bar{x}) \).

(That is, in a sufficiently small neighborhood of \( \bar{x} \), there exists a sample set revealing \( \Delta q(\bar{d}; B_\varepsilon(\bar{x}) \cap D) \) with arbitrarily good, though not necessarily perfect, accuracy.)

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

**Sketch of proof:** If $f(x_k) \nrightarrow -\infty$, then

$$\alpha_k \Delta q(d_k; X_k) \to 0.$$ 

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

$$\Delta q(d_k; X_k) = \frac{1}{2} \|d_k\|^2 > \frac{1}{2} \epsilon^2,$$  \hspace{1cm} (\star) 

and it can be shown that $x_k \to \overline{x}$ and $\alpha_k \to 0$. However, w.p.1, this will not occur:

- If $\overline{x}$ is $\epsilon$-stationary, then w.p.1 we will obtain a sample set in $\mathcal{T}$ yielding $\Delta q(d_k; X_k) \leq \frac{1}{2} \epsilon^2$, contradicting $(\star)$.
- If $\overline{x}$ is not $\epsilon$-stationary, then w.p.1 we will obtain a subsequence with $\alpha_k$ bounded away from zero, contradicting $\alpha_k \to 0$.

Thus, w.p.1, $\epsilon \to 0$ and any cluster point $\overline{x}$ is stationary for $f$. 
Outline

Motivations

Gradient Sampling (GS)

Adaptive GS

SQP-GS

Future Work
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 enhancements:

▶ Adaptive sampling; only $O(1)$ gradients per iteration (Kiwiel (2010))
▶ Warm-started subproblem solves
▶ “Hessian” approximations for quadratic term
Adaptive Gradient Sampling (AGS)

At $x_k$, we had:

$$X_k := \{x_{k0}, x_{k1}, \ldots, x_{kp}\} \quad \text{(sample points)}$$

$$G_k := [g_{k0} \quad g_{k1} \quad \cdots \quad g_{kp}] \quad \text{(sample gradients)}$$

At $x_{k+1}$, we

- maintain sample points still within radius $\epsilon$; (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

Recall the primal-dual pair of GS subproblems:

\[
\begin{align*}
\max_{z,d} & \quad z + \frac{1}{2}d^T d \\
\text{s.t.} & \quad f(x_k)e + G_k^T d \leq ze
\end{align*}
\]

\[
\begin{align*}
\max_{\lambda} & \quad f(x_k) - \frac{1}{2}\lambda^T G_k^T G_k \lambda \\
\text{s.t.} & \quad e^T \lambda = 1, \quad \lambda \geq 0
\end{align*}
\]
Primal-dual pair of subproblems (variable-metric)

Recall the primal-dual pair of GS subproblems:

\[
\begin{align*}
& \max_{z,d} z + \frac{1}{2} d^T d \\
& \text{s.t. } f(x_k)e + G_k^T d \leq ze
\end{align*}
\]

\[
\begin{align*}
& \max_{\lambda} f(x_k) - \frac{1}{2} \lambda^T G_k^T G_k \lambda \\
& \text{s.t. } e^T \lambda = 1, \lambda \geq 0
\end{align*}
\]

Introduce second order terms with “Hessian” approximations:

\[
\begin{align*}
& \max_{z,d} z + \frac{1}{2} d^T H_k d \\
& \text{s.t. } f(x_k)e + G_k^T d \leq ze
\end{align*}
\]

\[
\begin{align*}
& \max_{\lambda} f(x_k) - \frac{1}{2} \lambda^T G_k^T W_k G_k \lambda \\
& \text{s.t. } e^T \lambda = 1, \lambda \geq 0
\end{align*}
\]

How should \( H_k \) (or \( W_k \)) be chosen?
Consider the model

\[ q(d; x_{k+1}, H_{k+1}) = f(x_{k+1}) + \nabla f(x_{k+1})^T d + \frac{1}{2} d^T H_{k+1} d. \]

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:

- Is BFGS effective within GS?
- Are we making the best use of info?
- Ill-conditioning: Bad or good?
Quasi-Newton updating (AGS-LBFGS)

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

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

With at most $p$ points in the sample set, this is an L-BFGS-type approach.
Suppose we also have function values at the sample points.

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

$$q(d; X_k, H_k) = f(x_k) + \max_{x \in X_k} \{\nabla f(x)^T d\} + \frac{1}{2} d^T H_k d.$$ 

- If $q(d_{ki}; X_k, H_k) < f(x_{ki})$, then “lift” $d_{ki}^T H_k d_{ki}$ so that $q(d_{ki}; X_k, H_k) = f(x_{ki})$.

- Updates we use have the form $H_k \leftarrow M_{ki}^T H_k M_{ki}$ where

$$M_{ki} = \left( I + \frac{\gamma}{d_{ki}^T d_{ki}} d_{ki} d_{ki}^T \right).$$
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).)
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 $\epsilon$

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

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

Motivations

Gradient Sampling (GS)

Adaptive GS

SQP-GS

Future Work
Nonlinear constrained optimization

Consider constrained optimization problems of the form:

\[
\begin{align*}
\min_x f(x) & \quad \text{(smooth)} \\
\text{s.t. } c_{\mathcal{E}}(x) &= 0 & \text{(smooth)} \\
& \quad \text{(smooth)} \\
& c_{\mathcal{I}}(x) \leq 0 & \text{(smooth)}
\end{align*}
\]

- 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:

\[ \min_x f(x) \quad \text{(non)smooth} \]

subject to:

\[ c_{\mathcal{C}}(x) = 0 \quad \text{(smooth)} \]
\[ c_{\mathcal{E}}(x) = 0 \quad \text{(nonsmooth)} \]
\[ c_{\mathcal{I}}(x) \leq 0 \quad \text{(smooth)} \]
\[ c_{\mathcal{I}}(x) \leq 0 \quad \text{(nonsmooth)} \]

- 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:

\[
\min_x f(x) \quad \text{(smooth)}
\]
\[
\text{s.t. } c(x) \leq 0 \quad \text{(smooth)}
\]

At \(x_k\), solve the SQP subproblem

\[
\min_d f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T H_k d
\]
\[
\text{s.t. } c(x_k) + \nabla c(x_k)^T d \leq 0
\]

to compute the search direction \(d_k\).
Inconsistent linearizations of the constraints

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

$$\begin{align*}
\min_{d,s} \ & \rho(f(x_k) + \nabla f(x_k)^T d) + e^T s + \frac{1}{2} d^T H_k d \\
\text{s.t.} \ & c(x_k) + \nabla c(x_k)^T d \leq s, \ s \geq 0,
\end{align*}$$

Solving the (P)SQP subproblem is equivalent to minimizing

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

We perform a line search on the exact penalty function

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

to promote global convergence.
SQP method

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

- Solve the SQP subproblem

$$\min_{d,s} \rho(f(x_k) + \nabla f(x_k)^T d) + e^T s + \frac{1}{2} d^T H_k d$$

s.t. $c(x_k) + \nabla c(x_k)^T d \leq s, \quad s \geq 0$

- 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 $x_{k+1} \leftarrow x_k + \alpha_k d_k$. 
Consider constrained optimization problems of the form

\[ \min_{x} f(x) \quad \text{(nonsmooth, locally Lipschitz)} \]

s.t. \( c(x) \leq 0 \quad \text{(nonsmooth, locally Lipschitz)} \)

We may consider applying an unconstrained technique (e.g., AGS) directly to

\[ \min_{x} \phi_{\rho}(x), \]

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

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

\[
\min_{z,d,s} \rho z + e^T s + \frac{1}{2} d^T H_k d \\
\text{s.t. } f(x_k) + \nabla f(x_k)^T d \leq z \\
\quad c(x_k) + \nabla c(x_k)^T d \leq s, \ s \geq 0.
\]

- The AGS subproblem (for a nonsmooth objective) is

\[
\min_{z,d} z + \frac{1}{2} d^T H_k d \\
\text{s.t. } f(x_k) + \nabla f(x)^T d \leq z, \text{ for } x \in X_k.
\]
SQP and GS

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

\[
\min_{z,d,s} \rho z + e^T s + \frac{1}{2} d^T H_k d \\
\text{s.t. } f(x_k) + \nabla f(x_k)^T d \leq z \\
c(x_k) + \nabla c(x_k)^T d \leq s, \ s \geq 0.
\]

- The AGS subproblem (for a nonsmooth objective) is

\[
\min_{z,d} z + \frac{1}{2} d^T H_k d \\
\text{s.t. } f(x_k) + \nabla f(x)^T d \leq z, \ \text{for } x \in X_k.
\]

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

\[
\min_{z,d,s} \rho z + e^T s + \frac{1}{2} d^T H_k d \\
\text{s.t. } f(x_k) + \nabla f(x)^T d \leq z, \ \text{for } x \in X_k^f \\
c^i(x_k) + \nabla c^i(x)^T d \leq s^i, \ s^i \geq 0, \ \text{for } x \in X_k^{c^i}, \ i = 1, \ldots, m
\]
SQP-GS in more detail

- The SQP-GS subproblem is

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

where \(X_k\) is composed of

\[
X_k^f = \{x_k, x_k^f_1, \ldots, x_k^f_{kp}\} \subset \mathbb{B}_{\epsilon}(x_k) \cap D^f
\]
and \(X_k^{c^i} = \{x_k, x_k^{c^i}_1, \ldots, x_k^{c^i}_{kp}\} \subset \mathbb{B}_{\epsilon}(x_k) \cap D^{c^i} \quad \text{for } i = 1, \ldots, m.

- This is equivalent to minimizing

\[
q_{\rho}(d; X_k, H_k) := \\
\rho \max_{x \in X_k^f} (f(x_k) + \nabla f(x)^T d) + \sum_{x \in X_k^{c^i}} \max \{c^i(x_k) + \nabla c^i(x)^T d, 0\} + \frac{1}{2} d^T H_k d.
\]
SQP-GS illustration

\[
\min_{x} 10|x_2 - x_1^2| + (1 - x_1)^2 \quad \text{s.t.} \quad \max\{\sqrt{2}x_1, 2x_2\} - 1 \leq 0 \quad \text{at} \quad x_k = \left(\frac{2}{5}, \frac{3}{10}\right).
\]
SQP-GS illustration

\[ \min_{x} 10|x_2 - x_1^2| + (1 - x_1)^2 \quad \text{s.t.} \quad \max\{\sqrt{2}x_1, 2x_2\} - 1 \leq 0 \quad \text{at} \quad x_k = \left(\frac{2}{5}, \frac{3}{10}\right). \]
$\min_{x} 10|x_2 - x_1^2| + (1 - x_1)^2 \ 	ext{s.t.} \ \max\{\sqrt{2}x_1, 2x_2\} - 1 \leq 0 \ \text{at} \ \ x_k = \left(\frac{2}{5}, \frac{3}{10}\right)$. 
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
  \[
  \min_{z, d, s} \rho z + e^T s + \frac{1}{2} d^T H_k d \\
  \text{s.t. } f(x_k) + \nabla f(x)^T d \leq z, \text{ for } x \in X_k^f \\
  c^i(x_k) + \nabla c^i(x)^T d \leq s^i, s^i \geq 0, \text{ for } x \in X_k^{c_i}, i = 1, \ldots, m
  \]
- 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 \( x_{k+1} \in D^f \cap D^{c_1} \cap \cdots \cap D^{c_m} \))
- If \( \Delta q_\rho(d_k; X_k, H_k) \leq \frac{1}{2} \epsilon^2 \), then reduce \( \epsilon \).
- If \( \epsilon \) has been reduced and \( x_k \) is not sufficiently feasible, then reduce \( \rho \).
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 $\phi_{\rho_*}$. Moreover, with $K$ defined as the infinite subsequence of iterates during which $\epsilon$ 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 $\phi_0$. 
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

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

\[
\min_x 10|x_1^2 - x_2| + (1 - x_1)^2 \quad \text{s.t. } \max\{\sqrt{2}x_1, 2x_2\} \leq 1.
\]
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 (no sampling)
Example 2: Entropy minimization

Find a $N \times N$ matrix $X$ that solves

$$\min_X \ln \left( \prod_{j=1}^{K} \lambda_j(A \circ X^T X) \right)$$

s.t. $\|X_j\| = 1, \; j = 1, \ldots, N$

where $\lambda_j(M)$ denotes the $j$th largest eigenvalue of $M$, $A$ is a real symmetric $N \times N$ matrix, $\circ$ denotes the Hadamard matrix product, and $X_j$ denotes the $j$th column of $X$. 
Example 2: Entropy minimization

<table>
<thead>
<tr>
<th>$N$</th>
<th>$K$</th>
<th>$n$</th>
<th>Objective</th>
<th>Infeasibility</th>
<th>Final $\epsilon$</th>
<th>Opt. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1.0000e+00</td>
<td>3.1752e-14</td>
<td>5.9605e-09</td>
<td>7.6722e-12</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>16</td>
<td>7.4630e-01</td>
<td>2.8441e-07</td>
<td>4.8828e-05</td>
<td>1.1938e-04</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>36</td>
<td>6.3359e-01</td>
<td>2.1149e-06</td>
<td>9.7656e-05</td>
<td>8.7263e-02</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>64</td>
<td>5.5832e-01</td>
<td>2.0492e-05</td>
<td>9.7656e-05</td>
<td>2.7521e-03</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>100</td>
<td>2.1841e-01</td>
<td>9.8364e-06</td>
<td>7.8125e-04</td>
<td>9.6041e-03</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>144</td>
<td>1.2265e-01</td>
<td>1.8341e-04</td>
<td>7.8125e-04</td>
<td>6.0492e-03</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>196</td>
<td>8.4650e-02</td>
<td>1.6692e-04</td>
<td>7.8125e-04</td>
<td>7.1461e-03</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>256</td>
<td>6.5051e-02</td>
<td>6.4628e-04</td>
<td>1.5625e-03</td>
<td>3.1596e-03</td>
</tr>
</tbody>
</table>
Example 3: $\ell_{0.5}$ norm minimization

Recover a sparse signal by solving

$$\min_{x} \|x\|_{0.5}$$

s.t. $Ax = b$

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

(Use $\ell_{0.5}$ norm as $\ell_1$ does not recover sparse solution.)
Example 3: $\ell_{0.5}$ norm minimization

$k = 1$
Example 3: $\ell_{0.5}$ norm minimization

$k = 10$
Example 3: $\ell_{0.5}$ norm minimization
Example 3: $\ell_{0.5}$ norm minimization
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:

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

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

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

\( A \in \mathbb{R}^{n \times n} \) 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)
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

The 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
Future work

- C++ implementation
- Tailored QO solver for constrained case
- Adaptive sampling in constrained case
- Special handling of partly smooth functions
- Merge with bundle techniques for convex problems
Thanks!

▶ FEC and X. Que, “An Adaptive Gradient Sampling Algorithm for Nonsmooth Optimization,” in 1st round of revision for *Optimization Methods and Software*.