Nonsmooth, Nonconvex Optimization
Algorithms and Examples

Michael L. Overton
Courant Institute of Mathematical Sciences
New York University

Convex and Nonsmooth Optimization Class, Spring 2020, Lecture 13

Based on my research work with
Jim Burke (Washington), Adrian Lewis (Cornell) and others

Introduction

Nonsmooth, Nonconvex Optimization

Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

Nonsmooth, Nonconvex Optimization

Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

- Continuous

Nonsmooth, Nonconvex Optimization

Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

- Continuous
- Not differentiable everywhere, in particular often not differentiable at local minimizers

Nonsmooth, Nonconvex Optimization

Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

- Continuous
- Not differentiable everywhere, in particular often not differentiable at local minimizers
- Not convex

Nonsmooth, Nonconvex Optimization

Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

- Continuous
- Not differentiable everywhere, in particular often not differentiable at local minimizers
- Not convex
- Usually, but not always, locally Lipschitz: for all x there exists L_x s.t. $|f(y) - f(z)| \leq L_x \|y - z\|$ for all y, z near x

Nonsmooth, Nonconvex Optimization

Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

- Continuous
- Not differentiable everywhere, in particular often not differentiable at local minimizers
- Not convex
- Usually, but not always, locally Lipschitz: for all x there exists L_x s.t. $|f(y) - f(z)| \leq L_x \|y - z\|$ for all y, z near x
- Otherwise, no structure assumed

Nonsmooth, Nonconvex Optimization

Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

- Continuous
- Not differentiable everywhere, in particular often not differentiable at local minimizers
- Not convex
- Usually, but not always, locally Lipschitz: for all x there exists L_x s.t. $|f(y) - f(z)| \leq L_x \|y - z\|$ for all y, z near x
- Otherwise, no structure assumed

Lots of interesting applications

Nonsmooth, Nonconvex Optimization

Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

- Continuous
- Not differentiable everywhere, in particular often not differentiable at local minimizers
- Not convex
- Usually, but not always, locally Lipschitz: for all x there exists L_x s.t. $|f(y) - f(z)| \leq L_x \|y - z\|$ for all y, z near x
- Otherwise, no structure assumed

Lots of interesting applications

Any locally Lipschitz function is differentiable almost everywhere on its domain. So, whp, can evaluate gradient at any given point.

Nonsmooth, Nonconvex Optimization

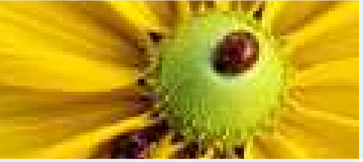
Problem: find x that locally minimizes f, where $f : \mathbb{R}^n \to \mathbb{R}$ is

- Continuous
- Not differentiable everywhere, in particular often not differentiable at local minimizers
- Not convex
- Usually, but not always, locally Lipschitz: for all x there exists L_x s.t. $|f(y) - f(z)| \leq L_x \|y - z\|$ for all y, z near x
- Otherwise, no structure assumed

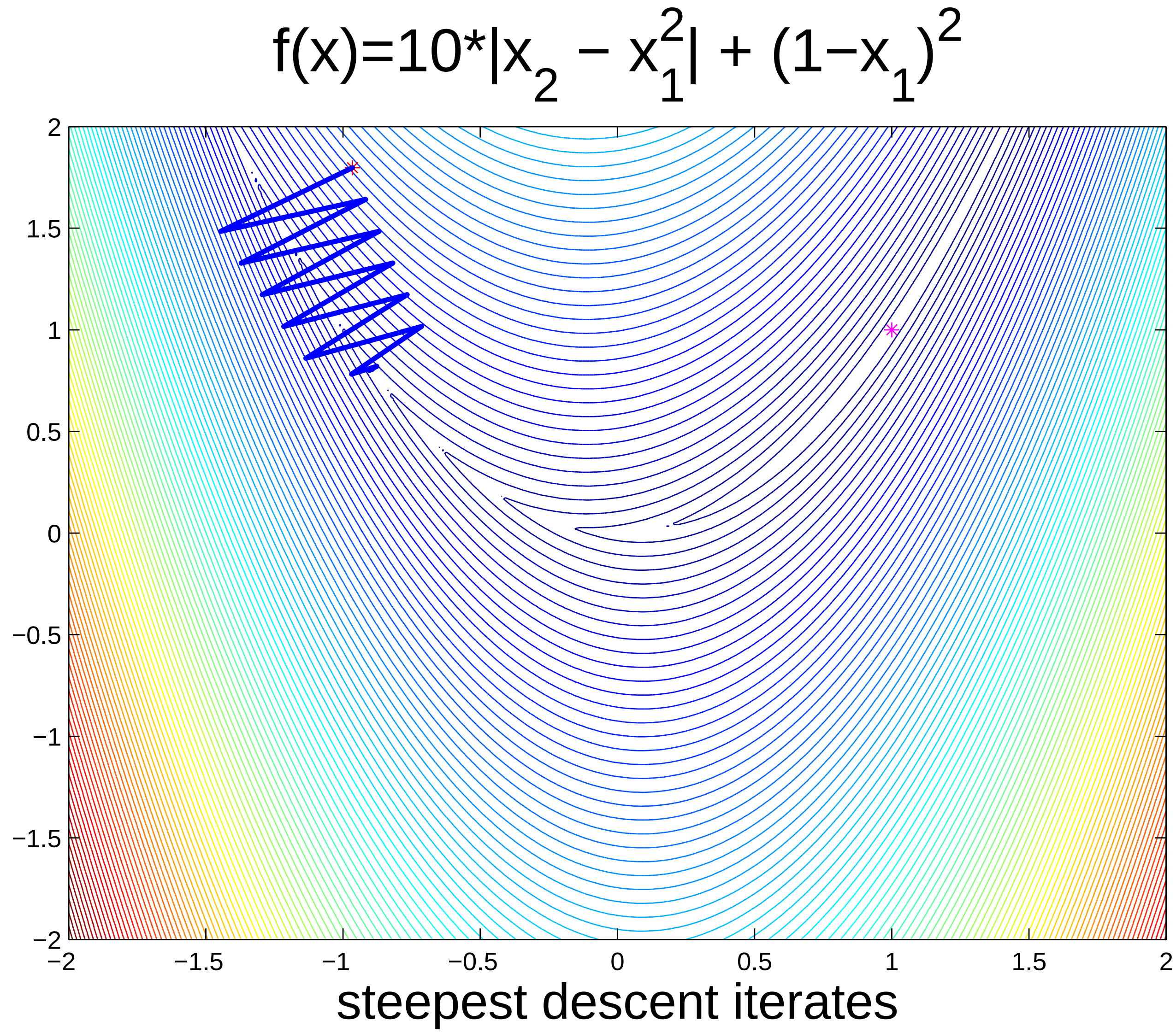
Lots of interesting applications

Any locally Lipschitz function is differentiable almost everywhere on its domain. So, whp, can evaluate gradient at any given point.

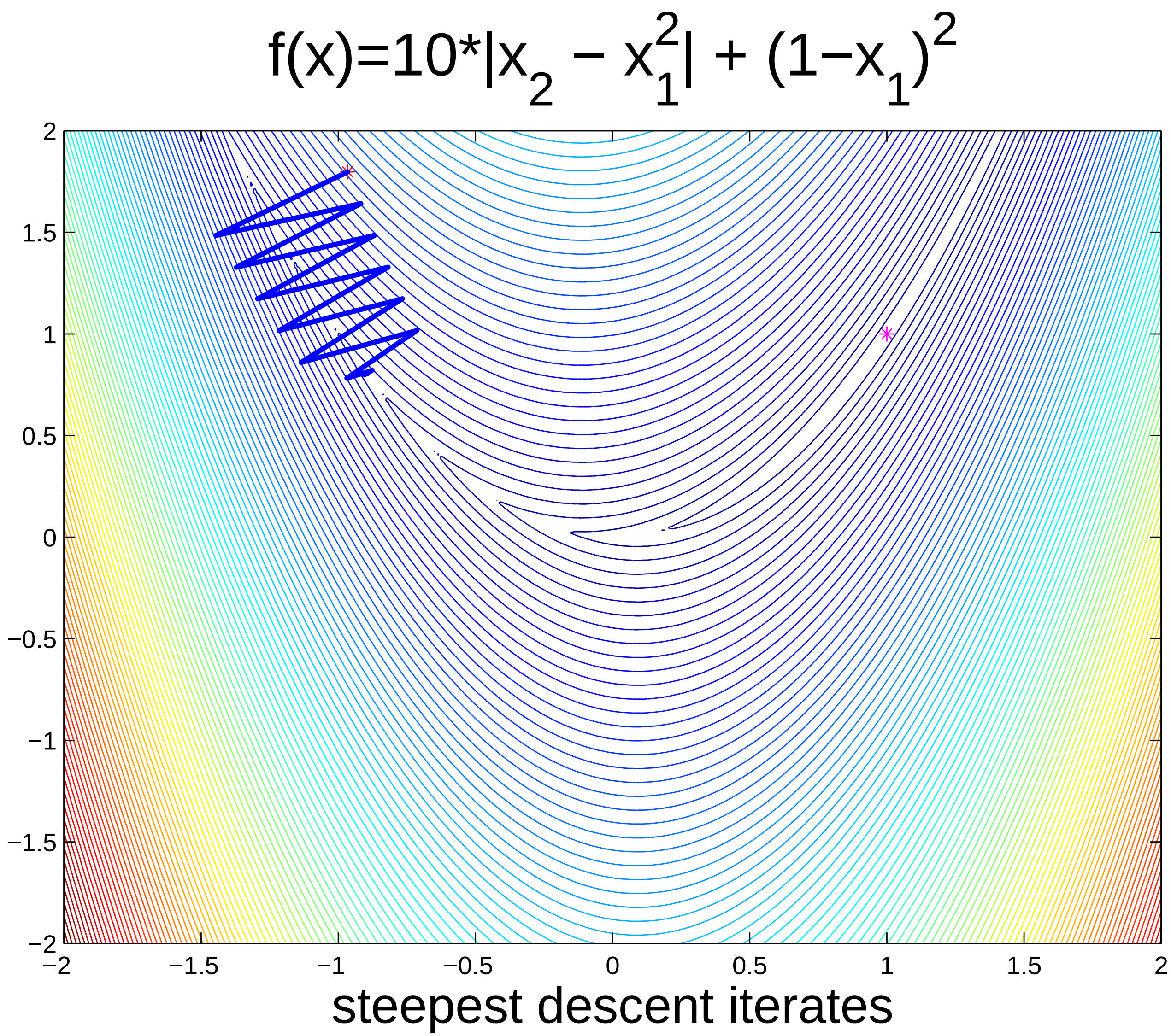
What happens if we simply use gradient descent (steepest descent) with a standard line search?



A Simple Nonconvex Example



A Simple Nonconvex Example



On this example, iterates invariably converge to a nonstationary point

Failure of Gradient Descent in Nonsmooth Case

Known for decades that gradient descent may converge to nonstationary points when f is nonsmooth, even if it is convex.

Failure of Gradient Descent in Nonsmooth Case

Known for decades that gradient descent may converge to nonstationary points when f is nonsmooth, even if it is convex.

- V.F. Dem'janov and V.N. Malozemov, 1970
- P. Wolfe, 1975
- J.-B. Hiriart-Urruty and C. Lemaréchal, 1993

But these are all examples cooked up to defeat exact line searches from a specific starting point.

Failure of Gradient Descent in Nonsmooth Case

Known for decades that gradient descent may converge to nonstationary points when f is nonsmooth, even if it is convex.

- V.F. Dem'janov and V.N. Malozemov, 1970
- P. Wolfe, 1975
- J.-B. Hiriart-Urruty and C. Lemaréchal, 1993

But these are all examples cooked up to defeat exact line searches from a specific starting point.

Failure can be avoided by using sufficiently short steplengths (N.Z. Shor, 1970s), but this is slow.

Armijo-Wolfe Line Search

Given x with f differentiable at x and d with $\nabla f(x)^T d < 0$, and parameters $0 < c_1 < c_2 < 1$, find steplength t so that

Armijo-Wolfe Line Search

Given x with f differentiable at x and d with $\nabla f(x)^T d < 0$, and parameters $0 < c_1 < c_2 < 1$, find steplength t so that

- sufficient decrease in function value: $f(x + td) < f(x) + c_1 t \nabla f(x)^T d$ (L. Armijo, 1966)

Armijo-Wolfe Line Search

Given x with f differentiable at x and d with $\nabla f(x)^T d < 0$, and parameters $0 < c_1 < c_2 < 1$, find steplength t so that

- sufficient decrease in function value: $f(x + td) < f(x) + c_1 t \nabla f(x)^T d$ (L. Armijo, 1966)
- sufficient increase in directional derivative: f is differentiable at $x + td$ and $\nabla f(x + td)^T d > c_2 \nabla f(x)^T d$ (P. Wolfe, 1969)

Armijo-Wolfe Line Search

Given x with f differentiable at x and d with $\nabla f(x)^T d < 0$, and parameters $0 < c_1 < c_2 < 1$, find steplength t so that

- sufficient decrease in function value: $f(x+td) < f(x) + c_1 t \nabla f(x)^T d$ (L. Armijo, 1966)
- sufficient increase in directional derivative: f is differentiable at $x + td$ and $\nabla f(x+td)^T d > c_2 \nabla f(x)^T d$ (P. Wolfe, 1969)

Assuming $\inf_t f(x+td)$ is bounded below,

- the Armijo condition holds for sufficiently small t as long as f is continuous
- the Wolfe condition holds for sufficiently large t as long as f is differentiable
- the intervals where each holds overlap

so combining the two conditions leads to a convenient, convergent bracketing line search (M.J.D. Powell, 1976)

Armijo-Wolfe Line Search

Given x with f differentiable at x and d with $\nabla f(x)^T d < 0$, and parameters $0 < c_1 < c_2 < 1$, find steplength t so that

- sufficient decrease in function value: $f(x + td) < f(x) + c_1 t \nabla f(x)^T d$ (L. Armijo, 1966)
- sufficient increase in directional derivative: f is differentiable at $x + td$ and $\nabla f(x + td)^T d > c_2 \nabla f(x)^T d$ (P. Wolfe, 1969)

Assuming $\inf_t f(x + td)$ is bounded below,

- the Armijo condition holds for sufficiently small t as long as f is continuous
- the Wolfe condition holds for sufficiently large t as long as f is differentiable
- the intervals where each holds overlap

so combining the two conditions leads to a convenient, convergent bracketing line search (M.J.D. Powell, 1976)

Extends to locally Lipschitz case (A.S. Lewis and M.L.O., 2013)

Armijo-Wolfe Line Search

Given x with f differentiable at x and d with $\nabla f(x)^T d < 0$, and parameters $0 < c_1 < c_2 < 1$, find steplength t so that

- sufficient decrease in function value: $f(x + td) < f(x) + c_1 t \nabla f(x)^T d$ (L. Armijo, 1966)
- sufficient increase in directional derivative: f is differentiable at $x + td$ and $\nabla f(x + td)^T d > c_2 \nabla f(x)^T d$ (P. Wolfe, 1969)

Assuming $\inf_t f(x + td)$ is bounded below,

- the Armijo condition holds for sufficiently small t as long as f is continuous
- the Wolfe condition holds for sufficiently large t as long as f is differentiable
- the intervals where each holds overlap

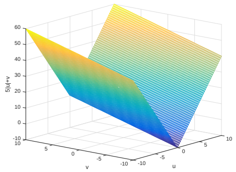
so combining the two conditions leads to a convenient, convergent bracketing line search (M.J.D. Powell, 1976)

Extends to locally Lipschitz case (A.S. Lewis and M.L.O., 2013)

Searching for "Armijo-Wolfe" on the web, we found Melissa Armijo-Wolfe's LinkedIn page!

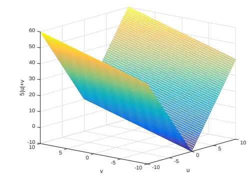
Failure of Gradient Method: Simple Convex Example

Let $f(x) = a|x_1| + x_2$, with $a \geq 1$. Although f is unbounded below, it is bounded below along any direction $d = -\nabla f(x)$.



Failure of Gradient Method: Simple Convex Example

Let $f(x) = a|x_1| + x_2$, with $a \geq 1$. Although f is unbounded below, it is bounded below along any direction $d = -\nabla f(x)$.



Theorem. Let $x^{(0)}$ satisfy $x_1^{(0)} \neq 0$ and define $x^{(k)} \in \mathbb{R}^2$ by

$$x^{(k+1)} = x^{(k)} + t_k d^{(k)} \text{ where } d^{(k)} = -\nabla f(x^{(k)})$$

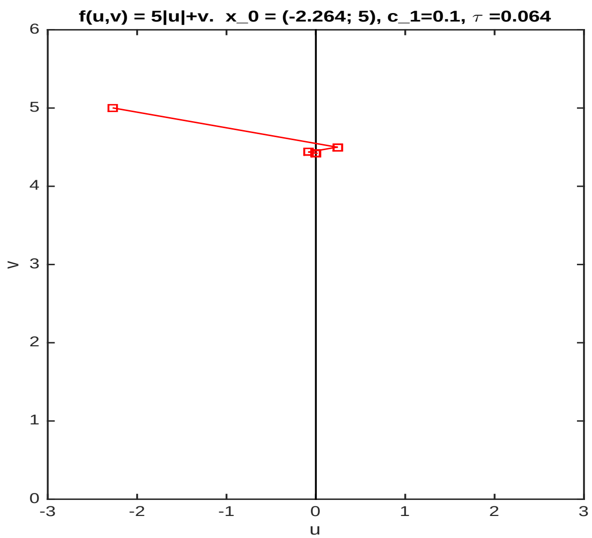
and t_k is *any* steplength satisfying the Armijo and Wolfe conditions with Armijo parameter c_1. If

$$c_1(a^2 + 1) > 1$$

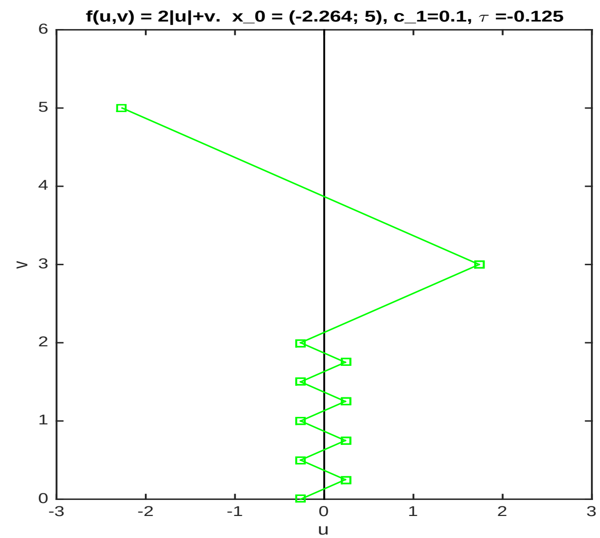
then $x^{(k)}$ converges to $\bar{x}$ with $\bar{x}_1 = 0$, even though f is unbounded below.

Azam Asl and M.L.O., 2017

Illustration of Failure and Success



$$a = 5, c_1 = 0.1$$
$$x^{(k)} \to \bar{x}$$



$$a = 2, c_1 = 0.1$$
$$f(x^{(k)}) \downarrow -\infty$$

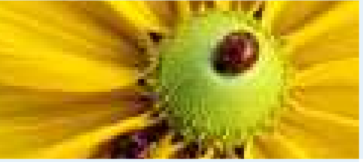
Methods Suitable for Nonsmooth Functions

Exploit the gradient information obtained at several points, not just at one point:

Methods Suitable for Nonsmooth Functions

Exploit the gradient information obtained at several points, not just at one point:

- Bundle methods (C. Lemaréchal, K.C. Kiwiel, 1980s –) extensive practical use and theoretical analysis, but complicated in nonconvex case



Methods Suitable for Nonsmooth Functions

Exploit the gradient information obtained at several points, not just at one point:

- Bundle methods (C. Lemaréchal, K.C. Kiwiel, 1980s –) extensive practical use and theoretical analysis, but complicated in nonconvex case
- Gradient sampling: an easily stated method with nice convergence theory (J.V. Burke, A.S. Lewis, M.L.O., 2005; K.C. Kiwiel, 2007), but computationally intensive

Methods Suitable for Nonsmooth Functions

Exploit the gradient information obtained at several points, not just at one point:

- Bundle methods (C. Lemaréchal, K.C. Kiwiel, 1980s –) extensive practical use and theoretical analysis, but complicated in nonconvex case
- Gradient sampling: an easily stated method with nice convergence theory (J.V. Burke, A.S. Lewis, M.L.O., 2005; K.C. Kiwiel, 2007), but computationally intensive
- BFGS: traditional workhorse for smooth optimization, works amazingly well for nonsmooth optimization too, but very limited convergence theory

Gradient Sampling

The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

The Gradient Sampling Method

Fix sample size $m \geq n + 1$, line search parameter $\beta \in (0, 1)$, reduction factors $\mu \in (0, 1)$ and $\theta \in (0, 1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

Repeat (outer loop)

The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

Repeat (outer loop)

- Repeat (inner loop: gradient sampling with fixed ϵ):

The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

Repeat (outer loop)

- Repeat (inner loop: gradient sampling with fixed ϵ):
  - Set $G = \{\nabla f(x), \nabla f(x+\epsilon u_1), \ldots, \nabla f(x+\epsilon u_m)\}$, sampling $u_1, \cdots, u_m$ uniformly from the unit ball

The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

Repeat (outer loop)

- Repeat (inner loop: gradient sampling with fixed ϵ):
  - Set $G = \{\nabla f(x), \nabla f(x+\epsilon u_1), \ldots, \nabla f(x+\epsilon u_m)\}$, sampling $u_1, \cdots, u_m$ uniformly from the unit ball
  - Set $g = \arg\min\{\|g\| : g \in \mathrm{conv}(G)\}$

The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

Repeat (outer loop)

- Repeat (inner loop: gradient sampling with fixed ϵ):
  - Set $G = \{\nabla f(x), \nabla f(x+\epsilon u_1), \ldots, \nabla f(x+\epsilon u_m)\}$, sampling $u_1, \cdots, u_m$ uniformly from the unit ball
  - Set $g = \arg\min\{\|g\| : g \in \mathrm{conv}(G)\}$
  - If $\|g\| \leq \tau$, break out of loop.

The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

Repeat (outer loop)

- Repeat (inner loop: gradient sampling with fixed ϵ):
  - Set $G = \{\nabla f(x), \nabla f(x+\epsilon u_1), \ldots, \nabla f(x+\epsilon u_m)\}$, sampling $u_1, \cdots, u_m$ uniformly from the unit ball
  - Set $g = \arg\min\{\|g\| : g \in \operatorname{conv}(G)\}$
  - If $\|g\| \leq \tau$, break out of loop.
  - Backtracking line search: set $d = -g$ and replace x by $x + td$, with $t \in \{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ and $f(x+td) < f(x) - \beta t \|g\|$

The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

Repeat (outer loop)

- Repeat (inner loop: gradient sampling with fixed ϵ):
  - Set $G = \{\nabla f(x), \nabla f(x+\epsilon u_1), \ldots, \nabla f(x+\epsilon u_m)\}$, sampling $u_1, \cdots, u_m$ uniformly from the unit ball
  - Set $g = \arg\min\{\|g\| : g \in \mathrm{conv}(G)\}$
  - If $\|g\| \leq \tau$, break out of loop.
  - Backtracking line search: set $d = -g$ and replace x by $x + td$, with $t \in \{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ and $f(x+td) < f(x) - \beta t\|g\|$
  - If f is not differentiable at $x+td$, replace $x+td$ by a nearby point where f is differentiable.[1]

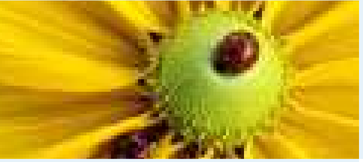
The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

Repeat (outer loop)

- Repeat (inner loop: gradient sampling with fixed ϵ):
  - Set $G = \{\nabla f(x), \nabla f(x+\epsilon u_1), \ldots, \nabla f(x+\epsilon u_m)\}$, sampling $u_1, \cdots, u_m$ uniformly from the unit ball
  - Set $g = \arg\min\{\|g\| : g \in \mathrm{conv}(G)\}$
  - If $\|g\| \leq \tau$, break out of loop.
  - Backtracking line search: set $d = -g$ and replace x by $x + td$, with $t \in \{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ and $f(x+td) < f(x) - \beta t\|g\|$
  - If f is not differentiable at $x + td$, replace $x + td$ by a nearby point where f is differentiable.[1]
- New phase: set $\epsilon = \mu\epsilon$ and $\tau = \theta\tau$.



The Gradient Sampling Method

Fix sample size $m \geq n+1$, line search parameter $\beta \in (0,1)$, reduction factors $\mu \in (0,1)$ and $\theta \in (0,1)$.

Initialize sampling radius $\epsilon > 0$, tolerance $\tau > 0$, iterate x.

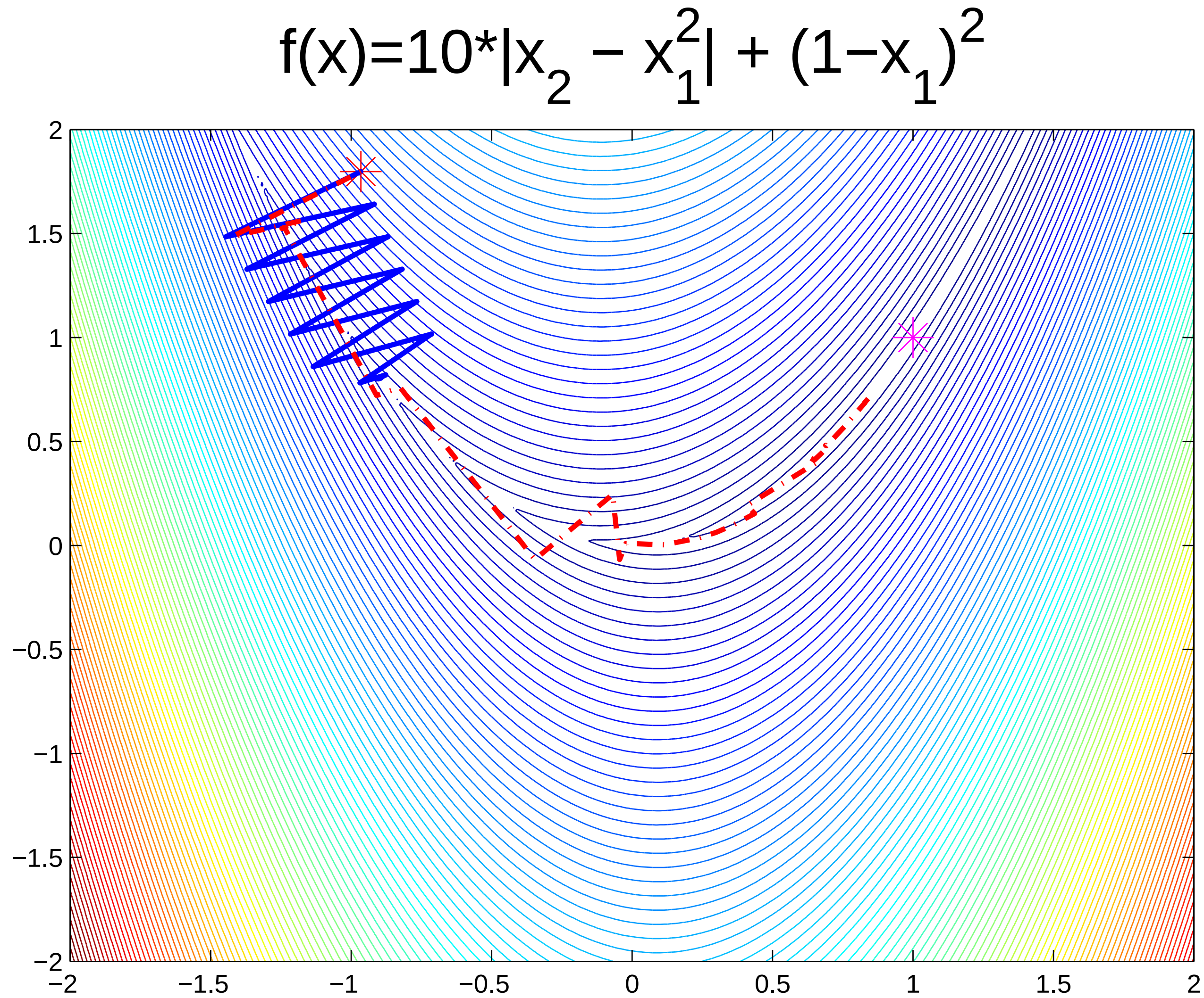
Repeat (outer loop)

- Repeat (inner loop: gradient sampling with fixed ϵ):
  - Set $G = \{\nabla f(x), \nabla f(x+\epsilon u_1), \ldots, \nabla f(x+\epsilon u_m)\}$, sampling $u_1, \cdots, u_m$ uniformly from the unit ball
  - Set $g = \arg\min\{\|g\| : g \in \operatorname{conv}(G)\}$
  - If $\|g\| \leq \tau$, break out of loop.
  - Backtracking line search: set $d = -g$ and replace x by $x + td$, with $t \in \{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ and $f(x+td) < f(x) - \beta t \|g\|$
  - If f is not differentiable at $x + td$, replace $x + td$ by a nearby point where f is differentiable.[1]
- New phase: set $\epsilon = \mu\epsilon$ and $\tau = \theta\tau$.

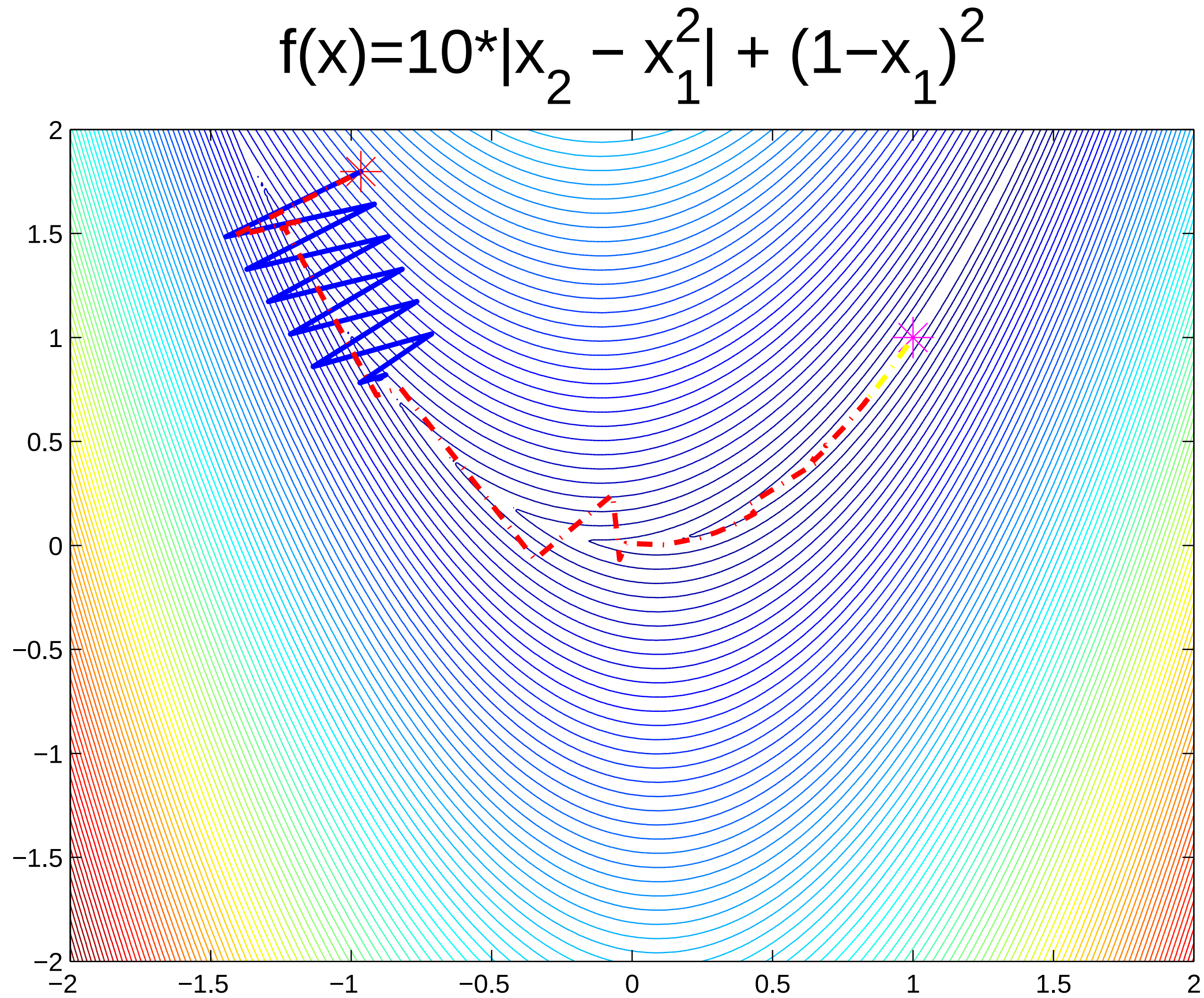
J.V. Burke, A.S. Lewis and M.L.O., SIOPT, 2005.

[1] Needed in theory, but typically not in practice.

With First Phase of Gradient Sampling



With Second Phase of Gradient Sampling



The Clarke Subdifferential

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, and let $D = \{x \in \mathbb{R}^n \ :\ f \text{ is differentiable at } x\}$.

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, and let $D = \{x \in \mathbb{R}^n \ : \ f \text{ is differentiable at } x\}$.

Rademacher's Theorem: $\mathbb{R}^n \backslash D$ has measure zero.

The Clarke Subdifferential

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, and let $D = \{x \in \mathbb{R}^n \ :\ f \text{ is differentiable at } x\}$.

Rademacher's Theorem: $\mathbb{R}^n \backslash D$ has measure zero.

The Clarke subdifferential of f at $\bar{x}$ is

$$\partial^C f(\bar{x}) = \operatorname{conv}\left\{ \lim_{x \to \bar{x}, x \in D} \nabla f(x) \right\}.$$

The Clarke Subdifferential

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, and let $D = \{x \in \mathbb{R}^n \ : \ f \text{ is differentiable at } x\}$.

Rademacher's Theorem: $\mathbb{R}^n \backslash D$ has measure zero.

The Clarke subdifferential of f at $\bar{x}$ is

$$\partial^C f(\bar{x}) = \text{conv}\left\{ \lim_{x \to \bar{x}, x \in D} \nabla f(x) \right\}.$$

F.H. Clarke, 1973 (he used the name "generalized gradient").

The Clarke Subdifferential

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, and
let $D = \{x \in \mathbb{R}^n \ : \ f \text{ is differentiable at } x\}$.

Rademacher's Theorem: $\mathbb{R}^n \backslash D$ has measure zero.

The Clarke subdifferential of f at $\bar{x}$ is

$$\partial^C f(\bar{x}) = \operatorname{conv}\left\{ \lim_{x \to \bar{x}, x \in D} \nabla f(x) \right\}.$$

F.H. Clarke, 1973 (he used the name "generalized gradient").

If f is continuously differentiable at $\bar{x}$, then $\partial^C f(\bar{x}) = \{\nabla f(\bar{x})\}$.

The Clarke Subdifferential

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, and let $D = \{x \in \mathbb{R}^n \ : \ f \text{ is differentiable at } x\}$.

Rademacher's Theorem: $\mathbb{R}^n \backslash D$ has measure zero.

The Clarke subdifferential of f at $\bar{x}$ is

$$\partial^C f(\bar{x}) = \text{conv}\left\{ \lim_{x \to \bar{x}, x \in D} \nabla f(x) \right\}.$$

F.H. Clarke, 1973 (he used the name "generalized gradient").

If f is continuously differentiable at $\bar{x}$, then $\partial^C f(\bar{x}) = \{\nabla f(\bar{x})\}$.

If f is convex, $\partial^C f$ is the subdifferential of convex analysis.

The Clarke Subdifferential

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, and
let $D = \{x \in \mathbb{R}^n \ : \ f \text{ is differentiable at } x\}$.

Rademacher's Theorem: $\mathbb{R}^n \backslash D$ has measure zero.

The Clarke subdifferential of f at $\bar{x}$ is

$$\partial^C f(\bar{x}) = \operatorname{conv}\left\{ \lim_{x \to \bar{x}, x \in D} \nabla f(x) \right\}.$$

F.H. Clarke, 1973 (he used the name "generalized gradient").

If f is continuously differentiable at $\bar{x}$, then $\partial^C f(\bar{x}) = \{\nabla f(\bar{x})\}$.

If f is convex, $\partial^C f$ is the subdifferential of convex analysis.

We say $\bar{x}$ is Clarke stationary for f if $0 \in \partial^C f(\bar{x})$.

The Clarke Subdifferential

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, and let $D = \{x \in \mathbb{R}^n \ : \ f \text{ is differentiable at } x\}$.

Rademacher's Theorem: $\mathbb{R}^n \backslash D$ has measure zero.

The Clarke subdifferential of f at $\bar{x}$ is

$$\partial^C f(\bar{x}) = \operatorname{conv}\left\{ \lim_{x \to \bar{x}, x \in D} \nabla f(x) \right\}.$$

F.H. Clarke, 1973 (he used the name "generalized gradient").

If f is continuously differentiable at $\bar{x}$, then $\partial^C f(\bar{x}) = \{\nabla f(\bar{x})\}$.

If f is convex, $\partial^C f$ is the subdifferential of convex analysis.

We say $\bar{x}$ is Clarke stationary for f if $0 \in \partial^C f(\bar{x})$.

Key point: the convex hull of the set G generated by Gradient Sampling is a surrogate for $\partial^C f$.

Example

Let

$$f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2$$

Example

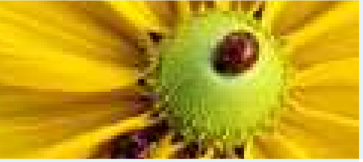
Let

$$f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2$$

For x with $x_2 \neq x_1^2$, f is differentiable with gradient

$$\nabla f(x) = 10\,\mathrm{sgn}\{x_2 - x_1^2\} \begin{bmatrix} -2x_1 \\ 1 \end{bmatrix} + \begin{bmatrix} -2(1 - x_1) \\ 0 \end{bmatrix}$$

so $\partial^C f(x) = \{\nabla f(x)\}$.



Example

Let

$$f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2$$

For x with $x_2 \neq x_1^2$, f is differentiable with gradient

$$\nabla f(x) = 10\,\mathrm{sgn}\{x_2 - x_1^2\} \begin{bmatrix} -2x_1 \\ 1 \end{bmatrix} + \begin{bmatrix} -2(1 - x_1) \\ 0 \end{bmatrix}$$

so $\partial^C f(x) = \{\nabla f(x)\}$. For x with $x_2 = x_1^2$, there are two limiting gradients, namely

$$\pm 10 \begin{bmatrix} -2x_1 \\ 1 \end{bmatrix} + \begin{bmatrix} -2(1 - x_1) \\ 0 \end{bmatrix}$$

so $\partial^C f(x)$ consists of the convex hull of these two vectors.

Example

Let

$$f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2$$

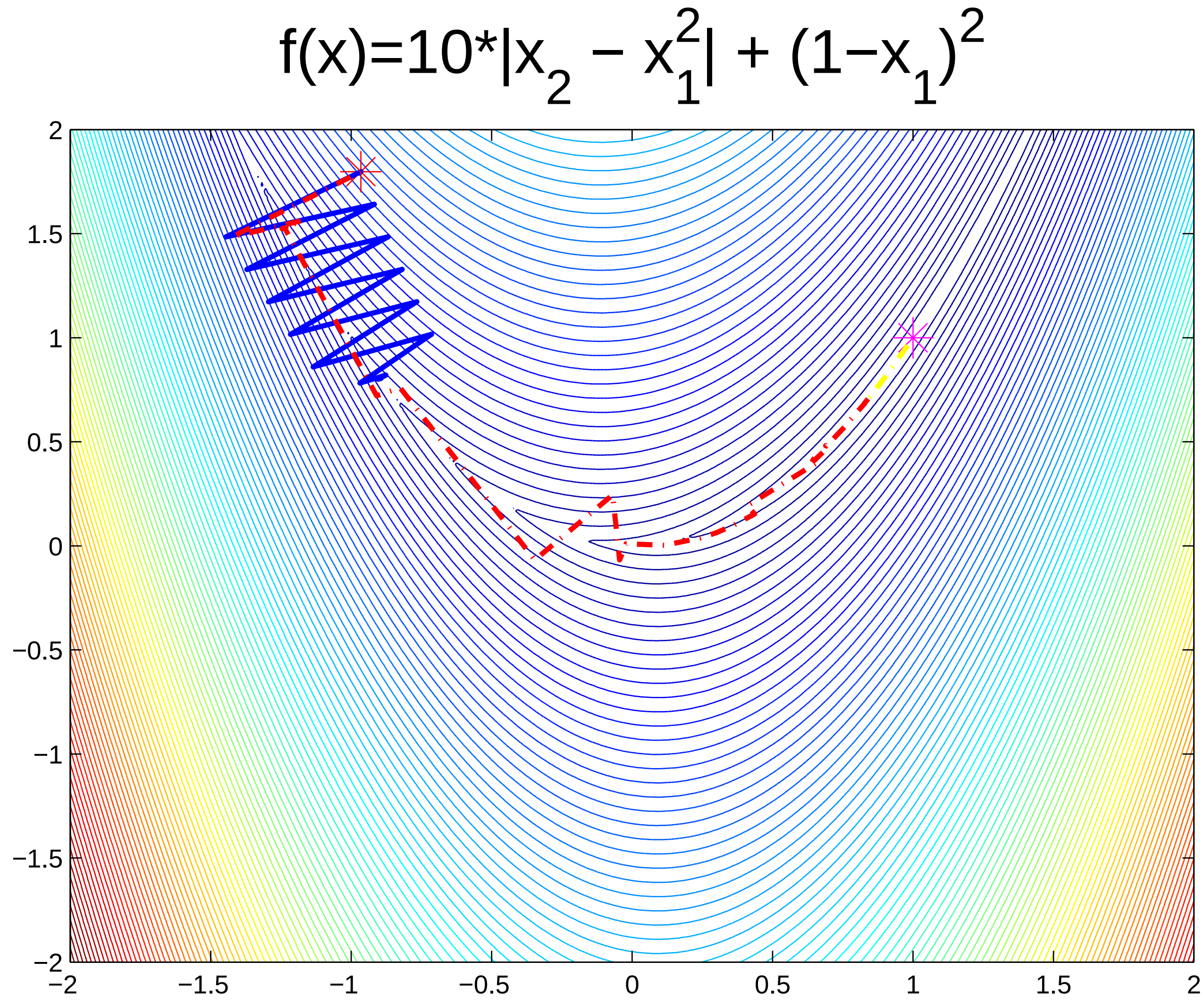
For x with $x_2 \neq x_1^2$, f is differentiable with gradient

$$\nabla f(x) = 10\,\mathrm{sgn}\{x_2 - x_1^2\} \begin{bmatrix} -2x_1 \\ 1 \end{bmatrix} + \begin{bmatrix} -2(1 - x_1) \\ 0 \end{bmatrix}$$

so $\partial^C f(x) = \{\nabla f(x)\}$. For x with $x_2 = x_1^2$, there are two limiting gradients, namely

$$\pm 10 \begin{bmatrix} -2x_1 \\ 1 \end{bmatrix} + \begin{bmatrix} -2(1 - x_1) \\ 0 \end{bmatrix}$$

so $\partial^C f(x)$ consists of the convex hull of these two vectors. The unique x for which $0 \in \partial^C f(x)$ is $x = [1;\ 1]^T$, so this is the unique Clarke stationary point of f (it follows that it is the global minimizer).


f(x)=10*|x_2 − x_1^2| + (1−x_1)^2
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2

Lemma. Let C be a compact convex set and $\|\cdot\| = \|\cdot\|_2$. Then

$$-\mathrm{dist}(0, C) = \min_{\|d\| \le 1} \max_{g \in C} g^T d$$

Grad. Samp.: A Stabilized Steepest Descent Method

Lemma. Let C be a compact convex set and $\|\cdot\| = \|\cdot\|_2$. Then

$$-\mathrm{dist}(0, C) = \min_{\|d\| \le 1} \max_{g \in C} g^T d$$

Proof.

$$\begin{aligned}
-\mathrm{dist}(0, C) &= -\min_{g \in C} \|g\| \\
&= -\min_{g \in C} \max_{\|\tilde{d}\| \le 1} g^T \tilde{d} \text{ (Cauchy-Schwartz)} \\
&= -\max_{\|\tilde{d}\| \le 1} \min_{g \in C} g^T \tilde{d} \text{ (min-max = max-min theorem)} \\
&= -\max_{\|d\| \le 1} \min_{g \in C} g^T (-d) \text{ (substitute } d = -\tilde{d}) \\
&= \min_{\|d\| \le 1} \max_{g \in C} g^T d \text{ (change sign)}
\end{aligned}$$

Note: the distance is nonnegative, and zero iff $0 \in C$.

Grad. Samp.: A Stabilized Steepest Descent Method

Lemma. Let C be a compact convex set and $\|\cdot\| = \|\cdot\|_2$. Then

$$-\mathrm{dist}(0, C) = \min_{\|d\| \leq 1} \max_{g \in C} g^T d$$

Proof.

$$\begin{aligned}
-\mathrm{dist}(0, C) &= -\min_{g \in C} \|g\| \\
&= -\min_{g \in C} \max_{\|\tilde{d}\| \leq 1} g^T \tilde{d} \text{ (Cauchy-Schwartz)} \\
&= -\max_{\|\tilde{d}\| \leq 1} \min_{g \in C} g^T \tilde{d} \text{ (min-max = max-min theorem)} \\
&= -\max_{\|d\| \leq 1} \min_{g \in C} g^T (-d) \text{ (substitute } d = -\tilde{d}) \\
&= \min_{\|d\| \leq 1} \max_{g \in C} g^T d \text{ (change sign)}
\end{aligned}$$

Note: the distance is nonnegative, and zero iff $0 \in C$.
Otherwise, equality is attained by $g = \Pi_C(0)$, $d = -g/|g\|$.

Grad. Samp.: A Stabilized Steepest Descent Method

Lemma. Let C be a compact convex set and $\|\cdot\| = \|\cdot\|_2$. Then

$$-\mathrm{dist}(0, C) = \min_{\|d\| \le 1} \max_{g \in C} g^T d$$

Proof.

$$\begin{aligned}
-\mathrm{dist}(0, C) &= -\min_{g \in C} \|g\| \\
&= -\min_{g \in C} \max_{\|\tilde{d}\| \le 1} g^T \tilde{d} \text{ (Cauchy-Schwartz)} \\
&= -\max_{\|\tilde{d}\| \le 1} \min_{g \in C} g^T \tilde{d} \text{ (min-max = max-min theorem)} \\
&= -\max_{\|d\| \le 1} \min_{g \in C} g^T (-d) \text{ (substitute } d = -\tilde{d}) \\
&= \min_{\|d\| \le 1} \max_{g \in C} g^T d \text{ (change sign)}
\end{aligned}$$

Note: the distance is nonnegative, and zero iff $0 \in C$.
Otherwise, equality is attained by $g = \Pi_C(0)$, $d = -g/|g\|$.
Ordinary steepest descent: $C = \{\nabla f(x)\}$.

Grad. Samp.: A Stabilized Steepest Descent Method

Lemma. Let C be a compact convex set and $\|\cdot\| = \|\cdot\|_2$. Then

$$-\mathrm{dist}(0, C) = \min_{\|d\| \le 1} \max_{g \in C} g^T d$$

Proof.

$$\begin{aligned}
-\mathrm{dist}(0, C) &= -\min_{g \in C} \|g\| \\
&= -\min_{g \in C} \max_{\|\tilde{d}\| \le 1} g^T \tilde{d} \text{ (Cauchy-Schwartz)} \\
&= -\max_{\|\tilde{d}\| \le 1} \min_{g \in C} g^T \tilde{d} \text{ (min-max = max-min theorem)} \\
&= -\max_{\|d\| \le 1} \min_{g \in C} g^T (-d) \text{ (substitute } d = -\tilde{d}) \\
&= \min_{\|d\| \le 1} \max_{g \in C} g^T d \text{ (change sign)}
\end{aligned}$$

Note: the distance is nonnegative, and zero iff $0 \in C$.
Otherwise, equality is attained by $g = \Pi_C(0)$, $d = -g/|g\|$.
Ordinary steepest descent: $C = \{\nabla f(x)\}$.
Gradient sampling: $C = \mathrm{conv}(G)$

$$= \mathrm{conv}(\{\nabla f(x), \nabla f(x + \epsilon u_1), \ldots, \nabla f(x + \epsilon u_m)\}).$$

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz
- is cont. differentiable on an open full-measure subset of $\mathbb{R}^n$

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz
- is cont. differentiable on an open full-measure subset of $\mathbb{R}^n$
- has bounded level sets

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz
- is cont. differentiable on an open full-measure subset of $\mathbb{R}^n$
- has bounded level sets

Then, with probability one, f is differentiable at the sampled points, the line search always terminates, and if the sequence of iterates $\{x\}$ converges to some point $\bar{x}$, then, with probability 1

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz
- is cont. differentiable on an open full-measure subset of $\mathbb{R}^n$
- has bounded level sets

Then, with probability one, f is differentiable at the sampled points, the line search always terminates, and if the sequence of iterates $\{x\}$ converges to some point $\bar{x}$, then, with probability 1

- the inner loop always terminates, so the sequences of sampling radii $\{\epsilon\}$ and tolerances $\{\tau\}$ converge to zero, and

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz
- is cont. differentiable on an open full-measure subset of $\mathbb{R}^n$
- has bounded level sets

Then, with probability one, f is differentiable at the sampled points, the line search always terminates, and if the sequence of iterates $\{x\}$ converges to some point $\bar{x}$, then, with probability 1

- the inner loop always terminates, so the sequences of sampling radii $\{\epsilon\}$ and tolerances $\{\tau\}$ converge to zero, and
- $\bar{x}$ is Clarke stationary for f, i.e., $0 \in \partial^C f(\bar{x})$.

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz
- is cont. differentiable on an open full-measure subset of $\mathbb{R}^n$
- has bounded level sets

Then, with probability one, f is differentiable at the sampled points, the line search always terminates, and if the sequence of iterates $\{x\}$ converges to some point $\bar{x}$, then, with probability 1

- the inner loop always terminates, so the sequences of sampling radii $\{\epsilon\}$ and tolerances $\{\tau\}$ converge to zero, and
- $\bar{x}$ is Clarke stationary for f, i.e., $0 \in \partial^C f(\bar{x})$.

J.V. Burke, A.S. Lewis and M.L.O., SIOPT, 2005.

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz
- is cont. differentiable on an open full-measure subset of $\mathbb{R}^n$
- has bounded level sets

Then, with probability one, f is differentiable at the sampled points, the line search always terminates, and if the sequence of iterates $\{x\}$ converges to some point $\bar{x}$, then, with probability 1

- the inner loop always terminates, so the sequences of sampling radii $\{\epsilon\}$ and tolerances $\{\tau\}$ converge to zero, and
- $\bar{x}$ is Clarke stationary for f, i.e., $0 \in \partial^C f(\bar{x})$.

J.V. Burke, A.S. Lewis and M.L.O., SIOPT, 2005.

Drop the assumption that f has bounded level sets. Then, wp 1, either the sequence $\{f(x)\} \to -\infty$, or every cluster point of the sequence of iterates $\{x\}$ is Clarke stationary.

Convergence of Gradient Sampling Method

Theorem. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$

- is locally Lipschitz
- is cont. differentiable on an open full-measure subset of $\mathbb{R}^n$
- has bounded level sets

Then, with probability one, f is differentiable at the sampled points, the line search always terminates, and if the sequence of iterates $\{x\}$ converges to some point $\bar{x}$, then, with probability 1

- the inner loop always terminates, so the sequences of sampling radii $\{\epsilon\}$ and tolerances $\{\tau\}$ converge to zero, and
- $\bar{x}$ is Clarke stationary for f, i.e., $0 \in \partial^C f(\bar{x})$.

J.V. Burke, A.S. Lewis and M.L.O., SIOPT, 2005.

Drop the assumption that f has bounded level sets. Then, wp 1, either the sequence $\{f(x)\} \to -\infty$, or every cluster point of the sequence of iterates $\{x\}$ is Clarke stationary.

K.C. Kiwiel, SIOPT, 2007.

Extensions

A more efficient version: Adaptive Gradient Sampling (F.E. Curtis and X. Que, 2013).

Extensions

A more efficient version: Adaptive Gradient Sampling (F.E. Curtis and X. Que, 2013).

Problems with Nonsmooth Constraints

$$\begin{aligned} \min\ & f(x) \\ \text{subject to } & c_i(x) \leq 0, \quad i = 1, \ldots, p \end{aligned}$$

Extensions

A more efficient version: Adaptive Gradient Sampling (F.E. Curtis and X. Que, 2013).

Problems with Nonsmooth Constraints

$$\begin{aligned} \min\ & f(x) \\ \text{subject to } & c_i(x) \le 0, \quad i = 1, \ldots, p \end{aligned}$$

where f and $c_1, \ldots, c_p$ are locally Lipschitz but may not be differentiable at local minimizers.

Extensions

A more efficient version: Adaptive Gradient Sampling (F.E. Curtis and X. Que, 2013).

Problems with Nonsmooth Constraints

$$\begin{aligned} \min\ & f(x) \\ \text{subject to } & c_i(x) \le 0, \quad i = 1, \ldots, p \end{aligned}$$

where f and $c_1, \ldots, c_p$ are locally Lipschitz but may not be differentiable at local minimizers.

A successive quadratic programming gradient sampling method with convergence theory.

Extensions

A more efficient version: Adaptive Gradient Sampling (F.E. Curtis and X. Que, 2013).

Problems with Nonsmooth Constraints

$$\begin{aligned} \min \ & f(x) \\ \text{subject to } & c_i(x) \le 0, \quad i = 1, \ldots, p \end{aligned}$$

where f and $c_1, \ldots, c_p$ are locally Lipschitz but may not be differentiable at local minimizers.

A successive quadratic programming gradient sampling method with convergence theory.

F.E. Curtis and M.L.O., SIOPT, 2012.

Some Gradient Sampling Success Stories

- Non-Lipschitz eigenvalue optimization for non-normal matrices (J.V. Burke, A.S. Lewis and M.L.O., 2002 –)

Some Gradient Sampling Success Stories

- Non-Lipschitz eigenvalue optimization for non-normal matrices (J.V. Burke, A.S. Lewis and M.L.O., 2002 –)
- Design of fixed-order controllers for linear dynamical systems with input and output (D. Henrion and M.L.O., 2006, and many subsequent users of our HIFOO (H-Infinity Fixed Order Optimization) toolbox)

Some Gradient Sampling Success Stories

Introduction

Gradient Sampling
The Gradient Sampling Method
With First Phase of Gradient Sampling
With Second Phase of Gradient Sampling
The Clarke Subdifferential
Example
Note that $0 \in \partial^C f(x)$ at $x = [1;\ 1]^T$
Grad. Samp.: A Stabilized Steepest Descent Method
Convergence of Gradient Sampling Method
Extensions
Some Gradient Sampling Success Stories

Quasi-Newton Methods

A Difficult Nonconvex Problem from Nesterov

Limited Memory Methods

- Non-Lipschitz eigenvalue optimization for non-normal matrices (J.V. Burke, A.S. Lewis and M.L.O., 2002 –)
- Design of fixed-order controllers for linear dynamical systems with input and output (D. Henrion and M.L.O., 2006, and many subsequent users of our HIFOO (H-Infinity Fixed Order Optimization) toolbox)
- Design of path planning for robots: avoids “chattering” that otherwise arises from nonsmoothness (I. Mitchell et al, 2017)

Quasi-Newton Methods

Bill Davidon

W. Davidon, a physicist at Argonne, had the breakthrough idea in 1959: since it's too expensive to compute and factor the Hessian $\nabla^2 f(x)$ at every iteration, update an approximation to its inverse using information from gradient differences, and multiply this onto the negative gradient to approximate Newton's method.

Bill Davidon

W. Davidon, a physicist at Argonne, had the breakthrough idea in 1959: since it's too expensive to compute and factor the Hessian $\nabla^2 f(x)$ at every iteration, update an approximation to its inverse using information from gradient differences, and multiply this onto the negative gradient to approximate Newton's method.

Each inverse Hessian approximation differs from the previous one by a rank-two correction.

Bill Davidon

W. Davidon, a physicist at Argonne, had the breakthrough idea in 1959: since it's too expensive to compute and factor the Hessian $\nabla^2 f(x)$ at every iteration, update an approximation to its inverse using information from gradient differences, and multiply this onto the negative gradient to approximate Newton's method.

Each inverse Hessian approximation differs from the previous one by a rank-two correction.

Ahead of its time: the paper was rejected by the physics journals, but published 30 years later in the first issue of SIAM J. Optimization.

Bill Davidon

W. Davidon, a physicist at Argonne, had the breakthrough idea in 1959: since it's too expensive to compute and factor the Hessian $\nabla^2 f(x)$ at every iteration, update an approximation to its inverse using information from gradient differences, and multiply this onto the negative gradient to approximate Newton's method.

Each inverse Hessian approximation differs from the previous one by a rank-two correction.

Ahead of its time: the paper was rejected by the physics journals, but published 30 years later in the first issue of SIAM J. Optimization.

Davidon was a well known active anti-war protester during the Vietnam War. In December 2013, it was revealed that he was the mastermind behind the break-in at the FBI office in Media, PA, on March 8, 1971, during the Muhammad Ali - Joe Frazier world heavyweight boxing championship.

Fletcher and Powell

In 1963, R. Fletcher and M.J.D. Powell improved Davidon's method and established convergence for convex quadratic functions.

Fletcher and Powell

In 1963, R. Fletcher and M.J.D. Powell improved Davidon’s method and established convergence for convex quadratic functions.

They applied it to solve problems in 100 variables: a lot at the time.

Fletcher and Powell

In 1963, R. Fletcher and M.J.D. Powell improved Davidon's method and established convergence for convex quadratic functions.

They applied it to solve problems in 100 variables: a lot at the time.

The method became known as the DFP method.

Fletcher and Powell

In 1963, R. Fletcher and M.J.D. Powell improved Davidon's method and established convergence for convex quadratic functions.

They applied it to solve problems in 100 variables: a lot at the time.

The method became known as the DFP method.

Davidon, Fletcher and Powell all died during 2013–2016.

BFGS

In 1970, C.G. Broyden, R. Fletcher, D. Goldfarb and D. Shanno all independently proposed the BFGS method, which is a kind of dual of the DFP method. It was soon recognized that this was a remarkably effective method for smooth optimization.

BFGS

In 1970, C.G. Broyden, R. Fletcher, D. Goldfarb and D. Shanno all independently proposed the BFGS method, which is a kind of dual of the DFP method. It was soon recognized that this was a remarkably effective method for smooth optimization.

In 1973, C.G. Broyden, J.E. Dennis and J.J. Moré proved generic local superlinear convergence of BFGS and DFP and other quasi-Newton methods.

BFGS

In 1970, C.G. Broyden, R. Fletcher, D. Goldfarb and D. Shanno all independently proposed the BFGS method, which is a kind of dual of the DFP method. It was soon recognized that this was a remarkably effective method for smooth optimization.

In 1973, C.G. Broyden, J.E. Dennis and J.J. Moré proved generic local superlinear convergence of BFGS and DFP and other quasi-Newton methods.

In 1976, M.J.D. Powell established convergence of BFGS with an inexact Armijo-Wolfe line search for a general class of smooth convex functions for BFGS. In 1987, this was extended by R.H. Byrd, J. Nocedal and Y.-X. Yuan to include the whole "Broyden" class of methods interpolating BFGS and DFP: *except* for the DFP end point.

BFGS

In 1970, C.G. Broyden, R. Fletcher, D. Goldfarb and D. Shanno all independently proposed the BFGS method, which is a kind of dual of the DFP method. It was soon recognized that this was a remarkably effective method for smooth optimization.

In 1973, C.G. Broyden, J.E. Dennis and J.J. Moré proved generic local superlinear convergence of BFGS and DFP and other quasi-Newton methods.

In 1976, M.J.D. Powell established convergence of BFGS with an inexact Armijo-Wolfe line search for a general class of smooth convex functions for BFGS. In 1987, this was extended by R.H. Byrd, J. Nocedal and Y.-X. Yuan to include the whole "Broyden" class of methods interpolating BFGS and DFP: *except* for the DFP end point.

Pathological counterexamples to convergence in the smooth, nonconvex case are known to exist (Y.-H. Dai, 2002, 2013; W. Mascarenhas 2004), but it is widely accepted that the method works well in practice in the smooth, nonconvex case.

The BFGS Method ("Full" Version)

Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

The BFGS Method ("Full" Version)

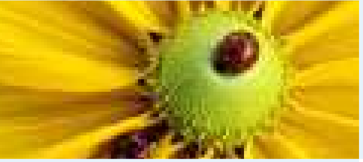
Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

Repeat

Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

Repeat

- Set $d = -H\nabla f(x)$.



The BFGS Method ("Full" Version)

Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

Repeat

- Set $d = -H\nabla f(x)$.
- Obtain t from Armijo-Wolfe line search

The BFGS Method ("Full" Version)

Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

Repeat

- Set $d = -H\nabla f(x)$.
- Obtain t from Armijo-Wolfe line search
- Set $s = td$, $y = \nabla f(x + td) - \nabla f(x)$

The BFGS Method ("Full" Version)

Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

Repeat

- Set $d = -H\nabla f(x)$.
- Obtain t from Armijo-Wolfe line search
- Set $s = td$, $y = \nabla f(x + td) - \nabla f(x)$
- Replace x by $x + td$

The BFGS Method ("Full" Version)

Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

Repeat

- Set $d = -H\nabla f(x)$.
- Obtain t from Armijo-Wolfe line search
- Set $s = td$, $y = \nabla f(x + td) - \nabla f(x)$
- Replace x by $x + td$
- Replace H by $VHV^T + \frac{1}{s^T y} ss^T$, where $V = I - \frac{1}{s^T y} sy^T$

The BFGS Method ("Full" Version)

Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

Repeat

- Set $d = -H\nabla f(x)$.
- Obtain t from Armijo-Wolfe line search
- Set $s = td$, $y = \nabla f(x + td) - \nabla f(x)$
- Replace x by $x + td$
- Replace H by $VHV^T + \frac{1}{s^Ty}ss^T$, where $V = I - \frac{1}{s^Ty}sy^T$

Note that H can be computed in $O(n^2)$ operations since V is a rank one perturbation of the identity

The BFGS Method ("Full" Version)

Initialize iterate x and positive-definite symmetric matrix H (which is supposed to approximate the *inverse* Hessian of f)

Repeat

- Set $d = -H\nabla f(x)$.
- Obtain t from Armijo-Wolfe line search
- Set $s = td$, $y = \nabla f(x + td) - \nabla f(x)$
- Replace x by $x + td$
- Replace H by $VHV^T + \frac{1}{s^T y} ss^T$, where $V = I - \frac{1}{s^T y} sy^T$

Note that H can be computed in $O(n^2)$ operations since V is a rank one perturbation of the identity
The Wolfe condition guarantees that $s^T y > 0$ and hence that the new H is positive definite.

BFGS for Nonsmooth Optimization

In 1982, C. Lemaréchal observed that quasi-Newton methods can be effective for nonsmooth optimization, but dismissed them as there was no theory behind them and no good way to terminate them.

BFGS for Nonsmooth Optimization

In 1982, C. Lemaréchal observed that quasi-Newton methods can be effective for nonsmooth optimization, but dismissed them as there was no theory behind them and no good way to terminate them.

Otherwise, there is not much in the literature on the subject until A.S. Lewis and M.L.O. (Math. Prog., 2013): we address both issues in detail, but our convergence results are limited to special cases.

BFGS for Nonsmooth Optimization

In 1982, C. Lemaréchal observed that quasi-Newton methods can be effective for nonsmooth optimization, but dismissed them as there was no theory behind them and no good way to terminate them.

Otherwise, there is not much in the literature on the subject until A.S. Lewis and M.L.O. (Math. Prog., 2013): we address both issues in detail, but our convergence results are limited to special cases.

Key point: use an Armijo-Wolfe line search. Do not insist on reducing the magnitude of the directional derivative along the line!

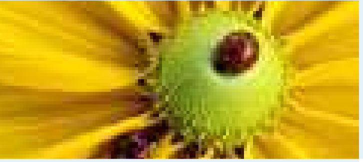
BFGS for Nonsmooth Optimization

In 1982, C. Lemaréchal observed that quasi-Newton methods can be effective for nonsmooth optimization, but dismissed them as there was no theory behind them and no good way to terminate them.

Otherwise, there is not much in the literature on the subject until A.S. Lewis and M.L.O. (Math. Prog., 2013): we address both issues in detail, but our convergence results are limited to special cases.

Key point: use an Armijo-Wolfe line search. Do not insist on reducing the magnitude of the directional derivative along the line!

In the nonsmooth case, BFGS builds a very ill-conditioned inverse "Hessian" approximation, with some tiny eigenvalues converging to zero, corresponding to "infinitely large" curvature in the directions defined by the associated eigenvectors.



BFGS for Nonsmooth Optimization

In 1982, C. Lemaréchal observed that quasi-Newton methods can be effective for nonsmooth optimization, but dismissed them as there was no theory behind them and no good way to terminate them.

Otherwise, there is not much in the literature on the subject until A.S. Lewis and M.L.O. (Math. Prog., 2013): we address both issues in detail, but our convergence results are limited to special cases.

Key point: use an Armijo-Wolfe line search. Do not insist on reducing the magnitude of the directional derivative along the line!

In the nonsmooth case, BFGS builds a very ill-conditioned inverse "Hessian" approximation, with some tiny eigenvalues converging to zero, corresponding to "infinitely large" curvature in the directions defined by the associated eigenvectors.

Remarkably, the condition number of the inverse Hessian approximation typically reaches 10^{16} before the method breaks down.

BFGS for Nonsmooth Optimization

In 1982, C. Lemaréchal observed that quasi-Newton methods can be effective for nonsmooth optimization, but dismissed them as there was no theory behind them and no good way to terminate them.

Otherwise, there is not much in the literature on the subject until A.S. Lewis and M.L.O. (Math. Prog., 2013): we address both issues in detail, but our convergence results are limited to special cases.

Key point: use an Armijo-Wolfe line search. Do not insist on reducing the magnitude of the directional derivative along the line!

In the nonsmooth case, BFGS builds a very ill-conditioned inverse "Hessian" approximation, with some tiny eigenvalues converging to zero, corresponding to "infinitely large" curvature in the directions defined by the associated eigenvectors.

Remarkably, the condition number of the inverse Hessian approximation typically reaches 10^{16} before the method breaks down.

We have never seen convergence to non-stationary points that cannot be explained by numerical difficulties.

BFGS for Nonsmooth Optimization

In 1982, C. Lemaréchal observed that quasi-Newton methods can be effective for nonsmooth optimization, but dismissed them as there was no theory behind them and no good way to terminate them.

Otherwise, there is not much in the literature on the subject until A.S. Lewis and M.L.O. (Math. Prog., 2013): we address both issues in detail, but our convergence results are limited to special cases.

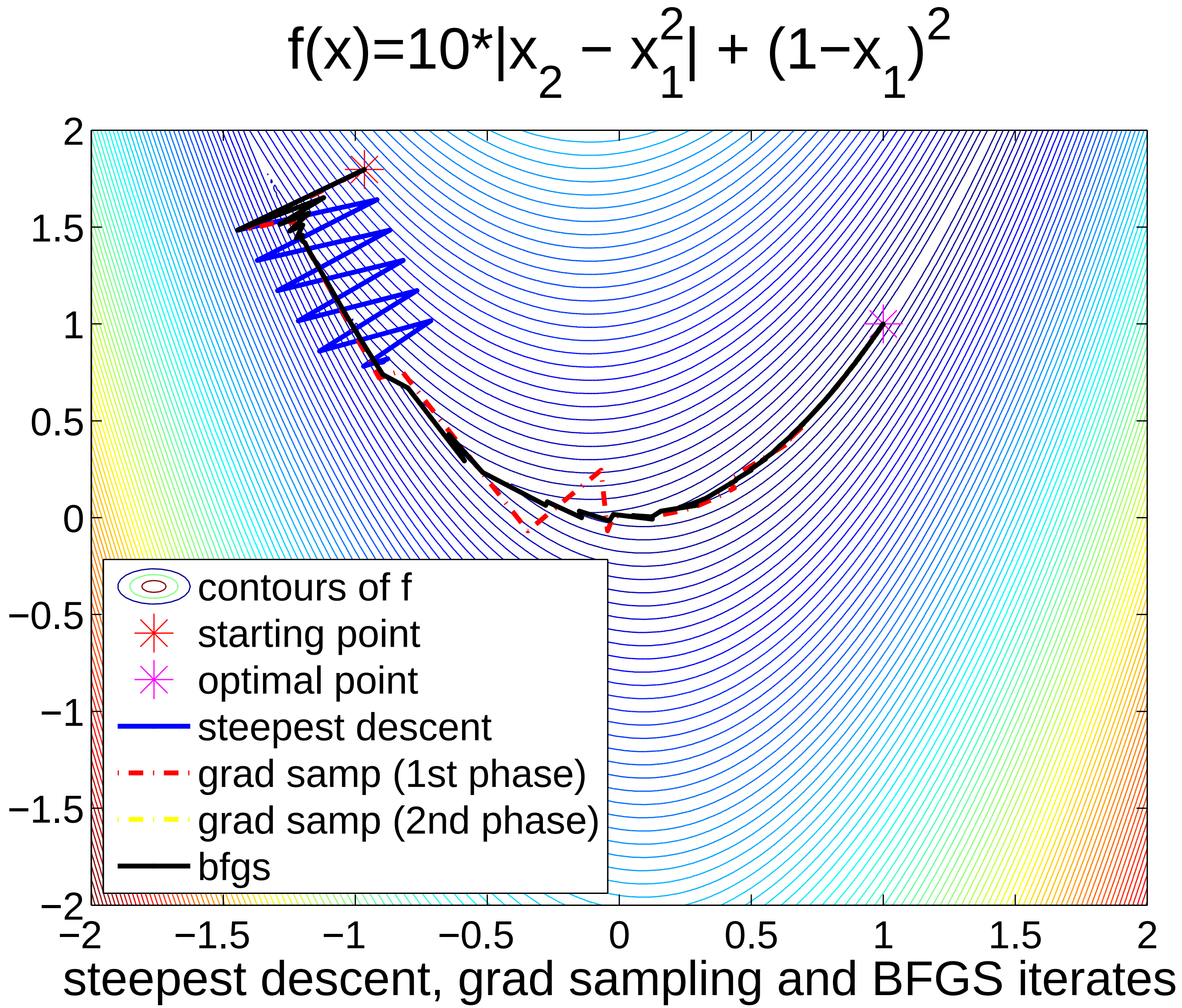
Key point: use an Armijo-Wolfe line search. Do not insist on reducing the magnitude of the directional derivative along the line!

In the nonsmooth case, BFGS builds a very ill-conditioned inverse "Hessian" approximation, with some tiny eigenvalues converging to zero, corresponding to "infinitely large" curvature in the directions defined by the associated eigenvectors.

Remarkably, the condition number of the inverse Hessian approximation typically reaches 10^{16} before the method breaks down.

We have never seen convergence to non-stationary points that cannot be explained by numerical difficulties.

Convergence rate of BFGS is typically linear (not superlinear) in the nonsmooth case.


$f(x)=10*|x_2 - x_1^2| + (1-x_1)^2$
contours of f
starting point
optimal point
steepest descent
grad samp (1st phase)
grad samp (2nd phase)
bfgs
steepest descent, grad sampling and BFGS iterates
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2

Example: Minimizing a Product of Eigenvalues

Let S^N denote the space of real symmetric $N \times N$ matrices, and

$$\lambda_1(X) \geq \lambda_2(X) \geq \cdots \lambda_N(X)$$

denote the eigenvalues of $X \in S^N$.

Example: Minimizing a Product of Eigenvalues

Let S^N denote the space of real symmetric $N \times N$ matrices, and

$$\lambda_1(X) \geq \lambda_2(X) \geq \cdots \lambda_N(X)$$

denote the eigenvalues of $X \in S^N$. We wish to minimize

$$f(X) = \log \prod_{i=1}^{N/2} \lambda_i(A \circ X)$$

where $A \in S^N$ is fixed and $\circ$ is the Hadamard (componentwise) matrix product, subject to the constraints that X is positive semidefinite and has diagonal entries equal to 1.

Example: Minimizing a Product of Eigenvalues

Let S^N denote the space of real symmetric $N \times N$ matrices, and

$$\lambda_1(X) \geq \lambda_2(X) \geq \cdots \lambda_N(X)$$

denote the eigenvalues of $X \in S^N$. We wish to minimize

$$f(X) = \log \prod_{i=1}^{N/2} \lambda_i(A \circ X)$$

where $A \in S^N$ is fixed and $\circ$ is the Hadamard (componentwise) matrix product, subject to the constraints that X is positive semidefinite and has diagonal entries equal to 1.

If we replace $\prod$ by $\sum$ we would have a semidefinite program.

Example: Minimizing a Product of Eigenvalues

Let S^N denote the space of real symmetric $N \times N$ matrices, and

$$\lambda_1(X) \geq \lambda_2(X) \geq \cdots \lambda_N(X)$$

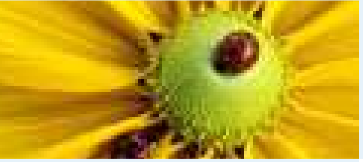
denote the eigenvalues of $X \in S^N$. We wish to minimize

$$f(X) = \log \prod_{i=1}^{N/2} \lambda_i(A \circ X)$$

where $A \in S^N$ is fixed and $\circ$ is the Hadamard (componentwise) matrix product, subject to the constraints that X is positive semidefinite and has diagonal entries equal to 1.

If we replace $\prod$ by $\sum$ we would have a semidefinite program.

Since f is not convex, may as well replace X by YY^T where $Y \in \mathbb{R}^{N \times N}$: eliminates psd constraint, and then also easy to eliminate diagonal constraint.



Example: Minimizing a Product of Eigenvalues

Let S^N denote the space of real symmetric $N \times N$ matrices, and

$$\lambda_1(X) \geq \lambda_2(X) \geq \cdots \lambda_N(X)$$

denote the eigenvalues of $X \in S^N$. We wish to minimize

$$f(X) = \log \prod_{i=1}^{N/2} \lambda_i(A \circ X)$$

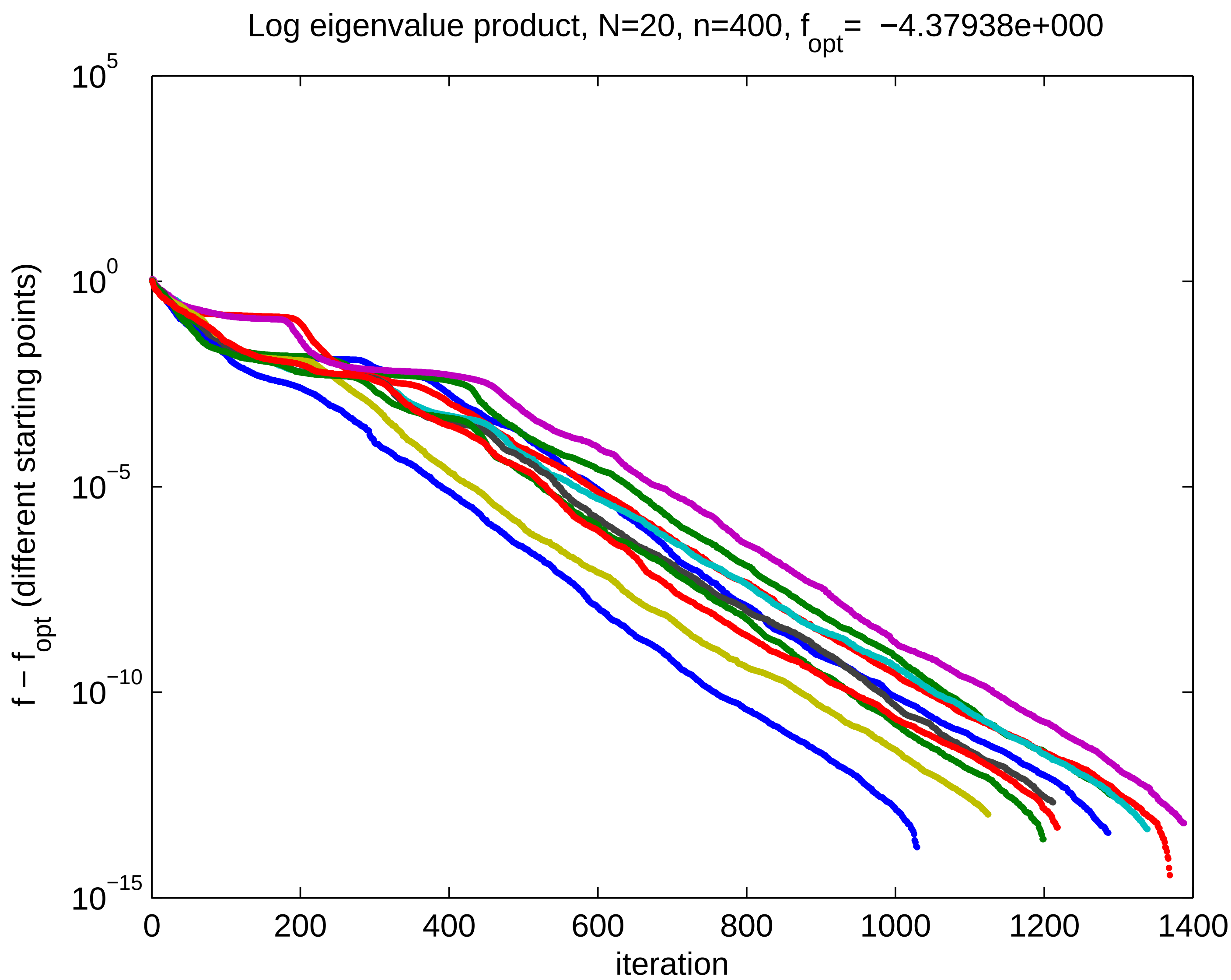
where $A \in S^N$ is fixed and $\circ$ is the Hadamard (componentwise) matrix product, subject to the constraints that X is positive semidefinite and has diagonal entries equal to 1.

If we replace $\prod$ by $\sum$ we would have a semidefinite program.

Since f is not convex, may as well replace X by YY^T where $Y \in \mathbb{R}^{N \times N}$: eliminates psd constraint, and then also easy to eliminate diagonal constraint.

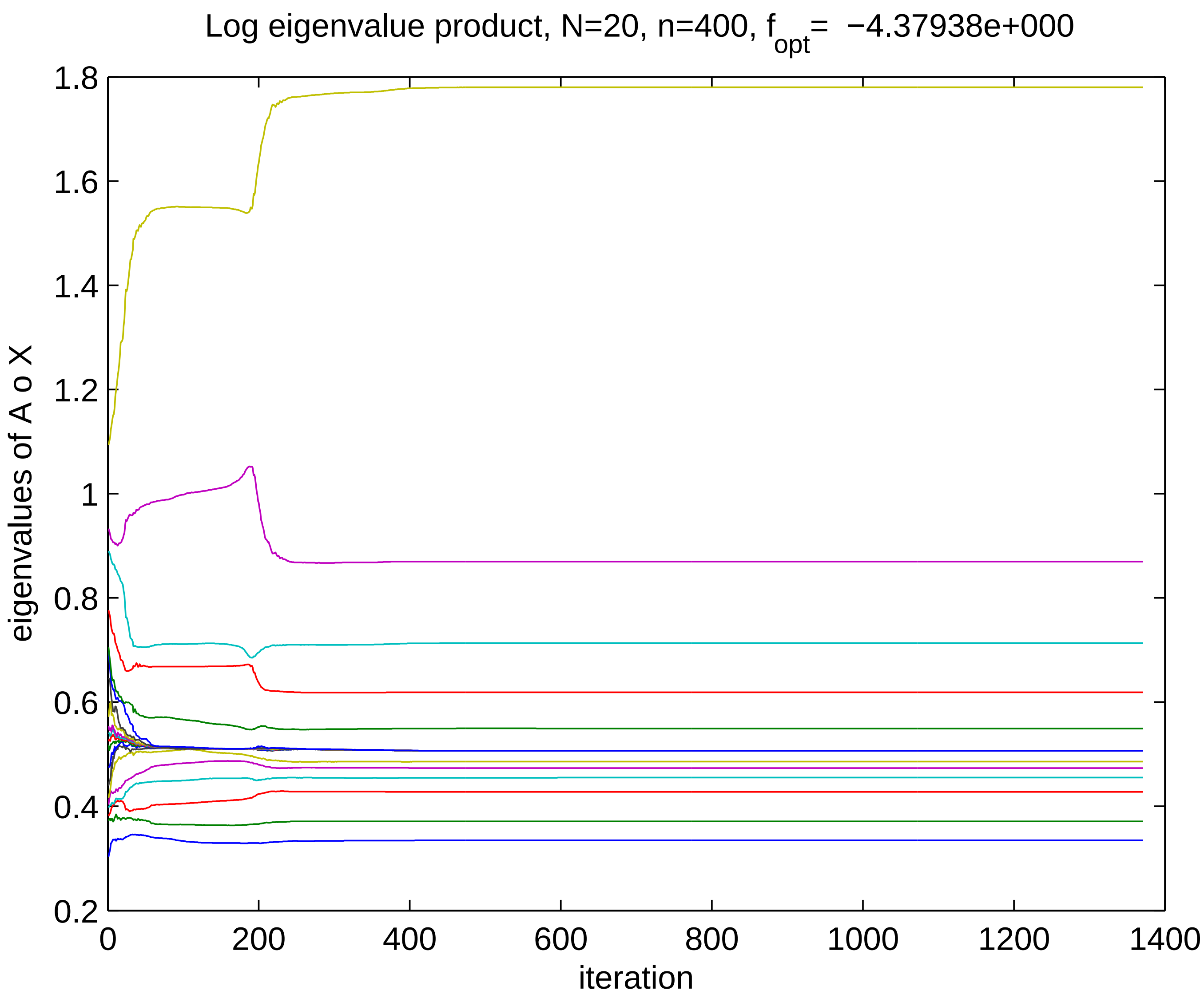
Application: entropy minimization in an environmental application (K.M. Anstreicher and J. Lee, 2004)

BFGS from 10 Randomly Generated Starting Points

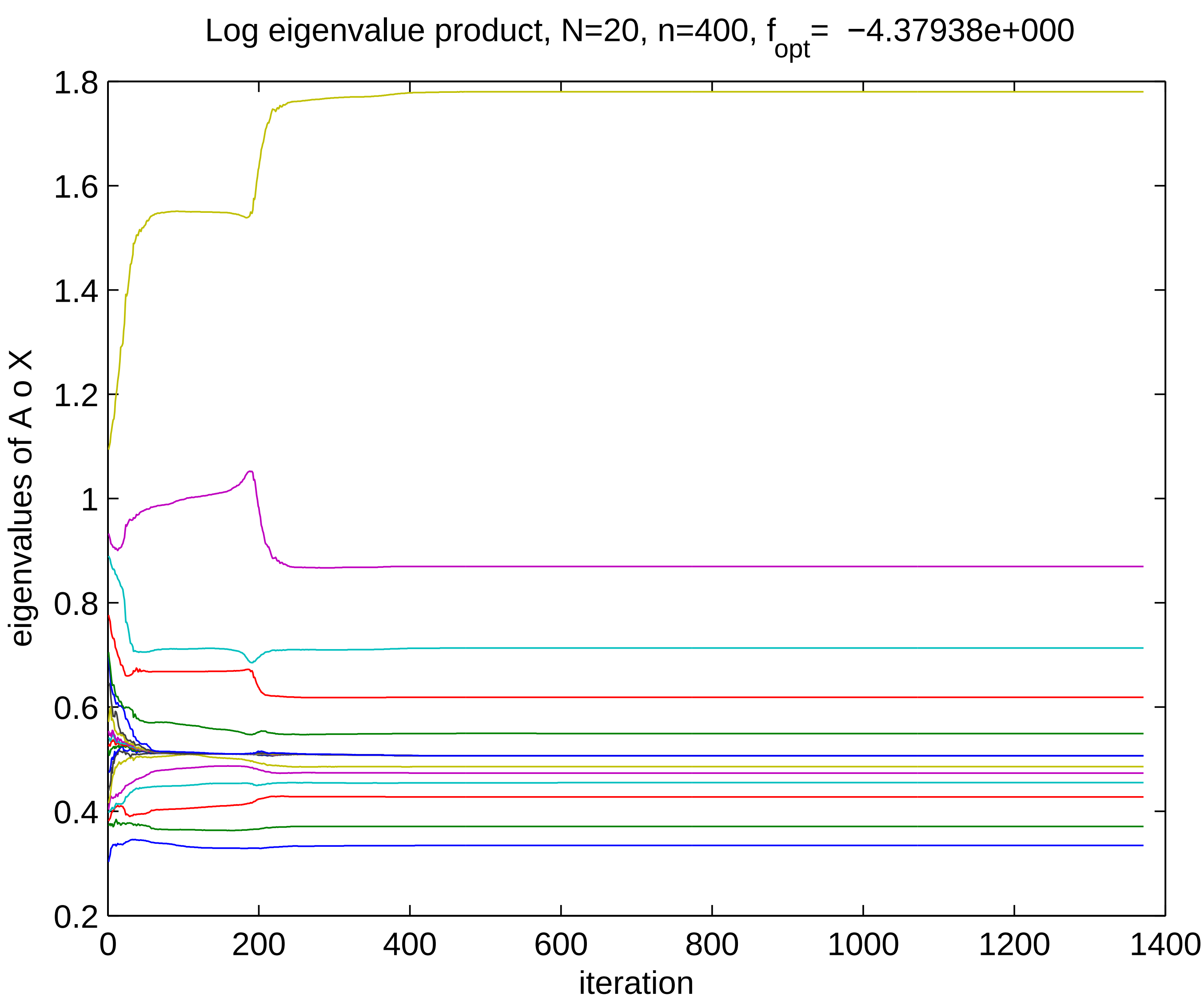


$f - f_{\text{opt}}$, where f_{opt} is least value of f found over all runs

Evolution of Eigenvalues of $A \circ X$

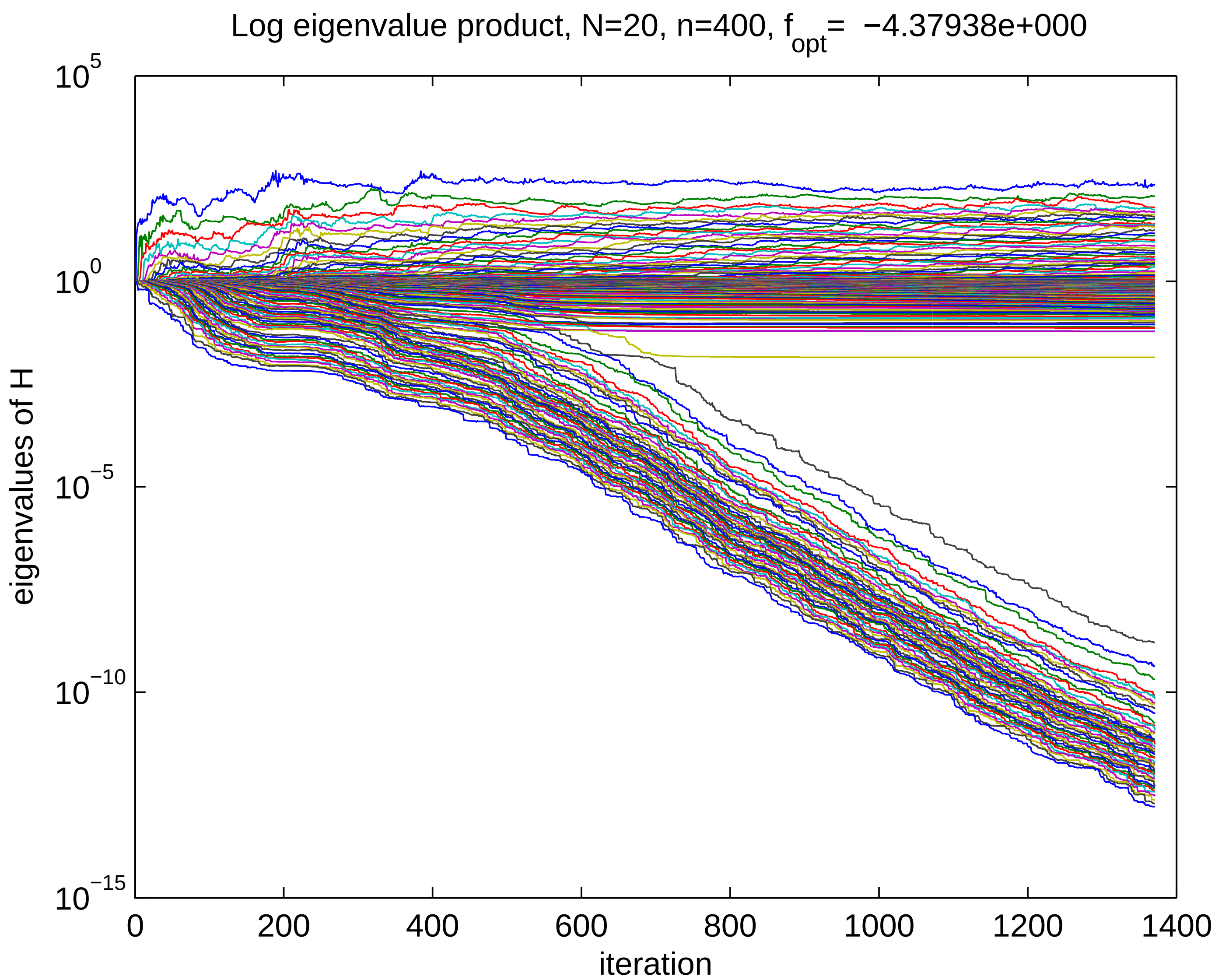


Evolution of Eigenvalues of $A \circ X$

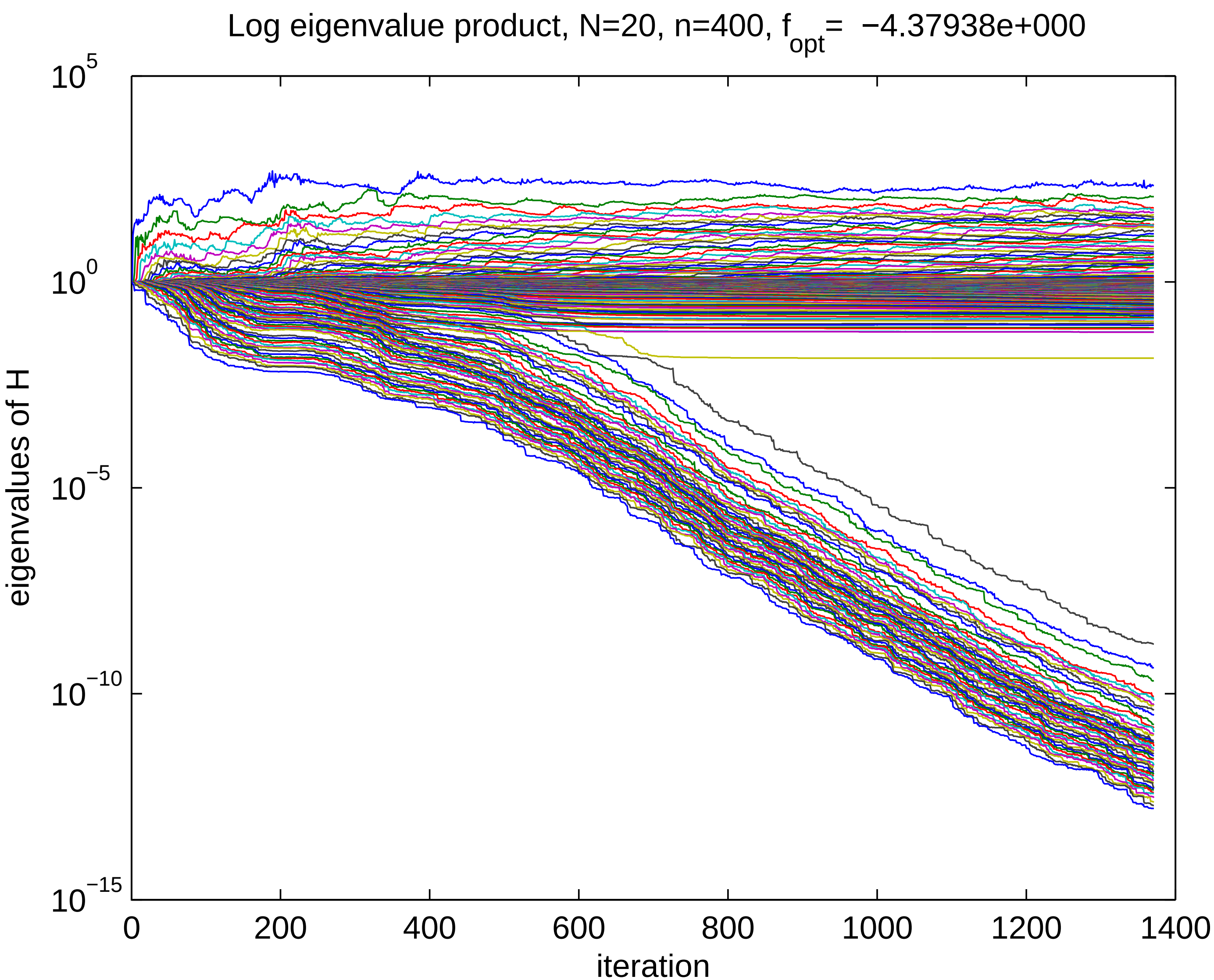


Note that $\lambda_6(X), \ldots, \lambda_{14}(X)$ coalesce

Evolution of Eigenvalues of H



Evolution of Eigenvalues of H



44 eigenvalues of H converge to zero...why???

Regularity

A locally Lipschitz, directionally differentiable function f is *regular* (Clarke 1970s) near a point x when its directional derivative $f'(\cdot\,;d)$ is upper semicontinuous there for every fixed direction d.

Regularity

A locally Lipschitz, directionally differentiable function f is *regular* (Clarke 1970s) near a point x when its directional derivative $f'(\,\cdot\,; d)$ is upper semicontinuous there for every fixed direction d.

In this case $0 \in \partial^C f(x)$ is equivalent to the first-order optimality condition $f'(x, d) \geq 0$ for all directions d.

Regularity

A locally Lipschitz, directionally differentiable function f is *regular* (Clarke 1970s) near a point x when its directional derivative $f'(\,\cdot\,; d)$ is upper semicontinuous there for every fixed direction d.

In this case $0 \in \partial^C f(x)$ is equivalent to the first-order optimality condition $f'(x, d) \geq 0$ for all directions d.

- All convex functions are regular

Regularity

A locally Lipschitz, directionally differentiable function f is *regular* (Clarke 1970s) near a point x when its directional derivative $f'(\,\cdot\,; d)$ is upper semicontinuous there for every fixed direction d.

In this case $0 \in \partial^C f(x)$ is equivalent to the first-order optimality condition $f'(x, d) \geq 0$ for all directions d.

- All convex functions are regular
- All smooth (C^1) functions are regular

Regularity

A locally Lipschitz, directionally differentiable function f is *regular* (Clarke 1970s) near a point x when its directional derivative $f'(\cdot\,; d)$ is upper semicontinuous there for every fixed direction d.

In this case $0 \in \partial^C f(x)$ is equivalent to the first-order optimality condition $f'(x, d) \geq 0$ for all directions d.

- All convex functions are regular
- All smooth (C^1) functions are regular
- Nonsmooth concave functions are not regular
  Example: $f(x) = -|x|$

Note: this is a somewhat simpler definition of regularity than the one in Lecture 12, but it is less precise: it defines regularity in a neighborhood, not at a point.

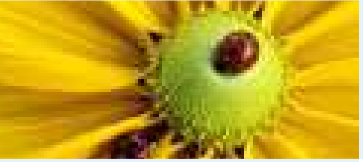
Partly Smooth Functions

A regular function f is *partly smooth* at x relative to a manifold $\mathcal{M}$ containing x (A.S. Lewis 2003) if

Partly Smooth Functions

A regular function f is *partly smooth* at x relative to a manifold $\mathcal{M}$ containing x (A.S. Lewis 2003) if

- its restriction to $\mathcal{M}$ is twice continuously differentiable near x



Partly Smooth Functions

A regular function f is *partly smooth* at x relative to a manifold $\mathcal{M}$ containing x (A.S. Lewis 2003) if

- its restriction to $\mathcal{M}$ is twice continuously differentiable near x
- the subdifferential ∂f is continuous on $\mathcal{M}$ near x

Partly Smooth Functions

A regular function f is *partly smooth* at x relative to a manifold $\mathcal{M}$ containing x (A.S. Lewis 2003) if

- its restriction to $\mathcal{M}$ is twice continuously differentiable near x
- the subdifferential ∂f is continuous on $\mathcal{M}$ near x
- par $\partial f(x)$, the subspace parallel to the affine hull of the subdifferential of f at x, is exactly the subspace normal to $\mathcal{M}$ at x.

Partly Smooth Functions

A regular function f is *partly smooth* at x relative to a manifold $\mathcal{M}$ containing x (A.S. Lewis 2003) if

- its restriction to $\mathcal{M}$ is twice continuously differentiable near x
- the subdifferential ∂f is continuous on $\mathcal{M}$ near x
- par $\partial f(x)$, the subspace parallel to the affine hull of the subdifferential of f at x, is exactly the subspace normal to $\mathcal{M}$ at x.

We refer to par $\partial f(x)$ as the V*-space* for f at x (with respect to $\mathcal{M}$), and to its orthogonal complement, the subspace tangent to $\mathcal{M}$ at x, as the U*-space* for f at x.

Partly Smooth Functions

A regular function f is *partly smooth* at x relative to a manifold $\mathcal{M}$ containing x (A.S. Lewis 2003) if

- its restriction to $\mathcal{M}$ is twice continuously differentiable near x
- the subdifferential ∂f is continuous on $\mathcal{M}$ near x
- $\operatorname{par} \partial f(x)$, the subspace parallel to the affine hull of the subdifferential of f at x, is exactly the subspace normal to $\mathcal{M}$ at x.

We refer to $\operatorname{par} \partial f(x)$ as the V*-space* for f at x (with respect to $\mathcal{M}$), and to its orthogonal complement, the subspace tangent to $\mathcal{M}$ at x, as the U*-space* for f at x.

When we refer to the V-space and U-space without reference to a point x, we mean at a minimizer.

Partly Smooth Functions

A regular function f is *partly smooth* at x relative to a manifold $\mathcal{M}$ containing x (A.S. Lewis 2003) if

- its restriction to $\mathcal{M}$ is twice continuously differentiable near x
- the subdifferential ∂f is continuous on $\mathcal{M}$ near x
- $\operatorname{par} \partial f(x)$, the subspace parallel to the affine hull of the subdifferential of f at x, is exactly the subspace normal to $\mathcal{M}$ at x.

We refer to $\operatorname{par} \partial f(x)$ as the V*-space* for f at x (with respect to $\mathcal{M}$), and to its orthogonal complement, the subspace tangent to $\mathcal{M}$ at x, as the U*-space* for f at x.

When we refer to the V-space and U-space without reference to a point x, we mean at a minimizer.

For nonzero y in the V-space, the mapping $t \mapsto f(x + ty)$ is necessarily nonsmooth at $t = 0$, while for nonzero y in the U-space, $t \mapsto f(x + ty)$ is differentiable at $t = 0$ as long as f is locally Lipschitz.

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \le x_2 \le 1\}$

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \le x_2 \le 1\}$
- $V = \text{par}\, \partial f(\bar{x}) = \{x : x_1 = 0, x_2 \in \mathbb{R}\}$

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \leq x_2 \leq 1\}$
- $V = \text{par}\, \partial f(\bar{x}) = \{x : x_1 = 0, x_2 \in \mathbb{R}\}$
- $U = V^\perp = \{x : x_1 \in \mathbb{R}, x_2 = 0\}$

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \leq x_2 \leq 1\}$
- $V = \text{par}\, \partial f(\bar{x}) = \{x : x_1 = 0, x_2 \in \mathbb{R}\}$
- $U = V^\perp = \{x : x_1 \in \mathbb{R}, x_2 = 0\}$

Example: $f(x) = \|x\|_2 = \sqrt{x_1^2 + x_2^2}$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$?

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \le x_2 \le 1\}$
- $V = \text{par}\, \partial f(\bar{x}) = \{x : x_1 = 0, x_2 \in \mathbb{R}\}$
- $U = V^{\perp} = \{x : x_1 \in \mathbb{R}, x_2 = 0\}$

Example: $f(x) = \|x\|_2 = \sqrt{x_1^2 + x_2^2}$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_1 = x_2 = 0\}$

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \leq x_2 \leq 1\}$
- $V = \text{par}\,\partial f(\bar{x}) = \{x : x_1 = 0, x_2 \in \mathbb{R}\}$
- $U = V^\perp = \{x : x_1 \in \mathbb{R}, x_2 = 0\}$

Example: $f(x) = \|x\|_2 = \sqrt{x_1^2 + x_2^2}$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_1 = x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : \|x\|_2 \leq 1\}$

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \leq x_2 \leq 1\}$
- $V = \text{par}\, \partial f(\bar{x}) = \{x : x_1 = 0, x_2 \in \mathbb{R}\}$
- $U = V^\perp = \{x : x_1 \in \mathbb{R}, x_2 = 0\}$

Example: $f(x) = \|x\|_2 = \sqrt{x_1^2 + x_2^2}$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_1 = x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : \|x\|_2 \leq 1\}$
- $V = \text{par}\, \partial f(\bar{x}) = \mathbb{R}^2$

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \leq x_2 \leq 1\}$
- $V = \text{par}\, \partial f(\bar{x}) = \{x : x_1 = 0, x_2 \in \mathbb{R}\}$
- $U = V^\perp = \{x : x_1 \in \mathbb{R}, x_2 = 0\}$

Example: $f(x) = \|x\|_2 = \sqrt{x_1^2 + x_2^2}$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_1 = x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : \|x\|_2 \leq 1\}$
- $V = \text{par}\, \partial f(\bar{x}) = \mathbb{R}^2$
- $U = V^\perp = \{0\}$

Partly Smooth Functions: Some Examples, $n = 2$

Example: $f(x) = \|x\|_1 = |x_1| + |x_2|$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : x_1 = 1, -1 \le x_2 \le 1\}$
- $V = \text{par}\,\partial f(\bar{x}) = \{x : x_1 = 0, x_2 \in \mathbb{R}\}$
- $U = V^\perp = \{x : x_1 \in \mathbb{R}, x_2 = 0\}$

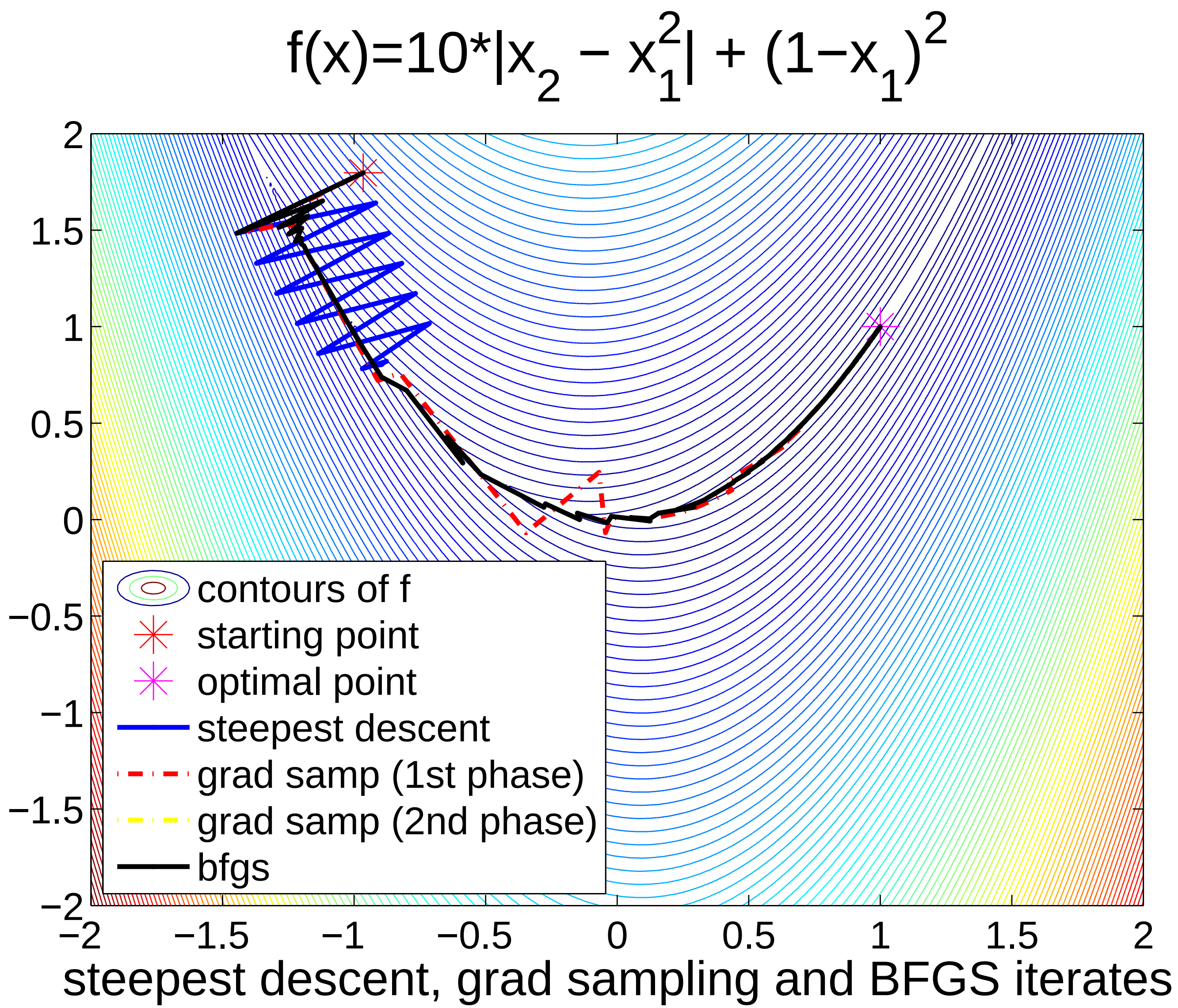
Example: $f(x) = \|x\|_2 = \sqrt{x_1^2 + x_2^2}$.

Question: What is $\mathcal{M}$ and what are the U and V spaces at $\bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$?

- $\mathcal{M} = \{x : x_1 = x_2 = 0\}$
- $\partial f(\bar{x}) = \{x : \|x\|_2 \le 1\}$
- $V = \text{par}\,\partial f(\bar{x}) = \mathbb{R}^2$
- $U = V^\perp = \{0\}$

Example: $f(x) = 10|x_2 - x_1^2| + (1 - x_1)^2$: same example as earlier...

Same Example Again



Why Did 44 Eigenvalues of H Converge to Zero?

The eigenvalue product is *regular* and also *partly smooth* (in the sense of A.S. Lewis, 2003) with respect to the manifold of matrices with an eigenvalue with given multiplicity. This implies that *tangent* to this manifold (preserving the multiplicity to first-order) the function is *smooth* ("U-shaped") and *normal* to it, the function is *nonsmooth* ("V-shaped").

Why Did 44 Eigenvalues of H Converge to Zero?

The eigenvalue product is *regular* and also *partly smooth* (in the sense of A.S. Lewis, 2003) with respect to the manifold of matrices with an eigenvalue with given multiplicity. This implies that *tangent* to this manifold (preserving the multiplicity to first-order) the function is *smooth* ("U-shaped") and *normal* to it, the function is *nonsmooth* ("V-shaped").

Recall that at the computed minimizer,

$$\lambda_6(A \circ X) \approx \ldots \approx \lambda_{14}(A \circ X).$$

Matrix theory says that imposing multiplicity m on an eigenvalue a matrix $\in S^N$ is $\frac{m(m+1)}{2} - 1$ conditions, or 44 when $m = 9$, so the dimension of the V-space at this minimizer is 44.

Why Did 44 Eigenvalues of H Converge to Zero?

The eigenvalue product is *regular* and also *partly smooth* (in the sense of A.S. Lewis, 2003) with respect to the manifold of matrices with an eigenvalue with given multiplicity. This implies that *tangent* to this manifold (preserving the multiplicity to first-order) the function is *smooth* ("U-shaped") and *normal* to it, the function is *nonsmooth* ("V-shaped").

Recall that at the computed minimizer,

$$\lambda_6(A \circ X) \approx \ldots \approx \lambda_{14}(A \circ X).$$

Matrix theory says that imposing multiplicity m on an eigenvalue a matrix $\in S^N$ is $\frac{m(m+1)}{2} - 1$ conditions, or 44 when $m = 9$, so the dimension of the V-space at this minimizer is 44.

Tiny eigenvalues of H correspond to huge curvature, which corresponds to V-space directions.

Why Did 44 Eigenvalues of H Converge to Zero?

The eigenvalue product is *regular* and also *partly smooth* (in the sense of A.S. Lewis, 2003) with respect to the manifold of matrices with an eigenvalue with given multiplicity. This implies that *tangent* to this manifold (preserving the multiplicity to first-order) the function is *smooth* ("U-shaped") and *normal* to it, the function is *nonsmooth* ("V-shaped").

Recall that at the computed minimizer,

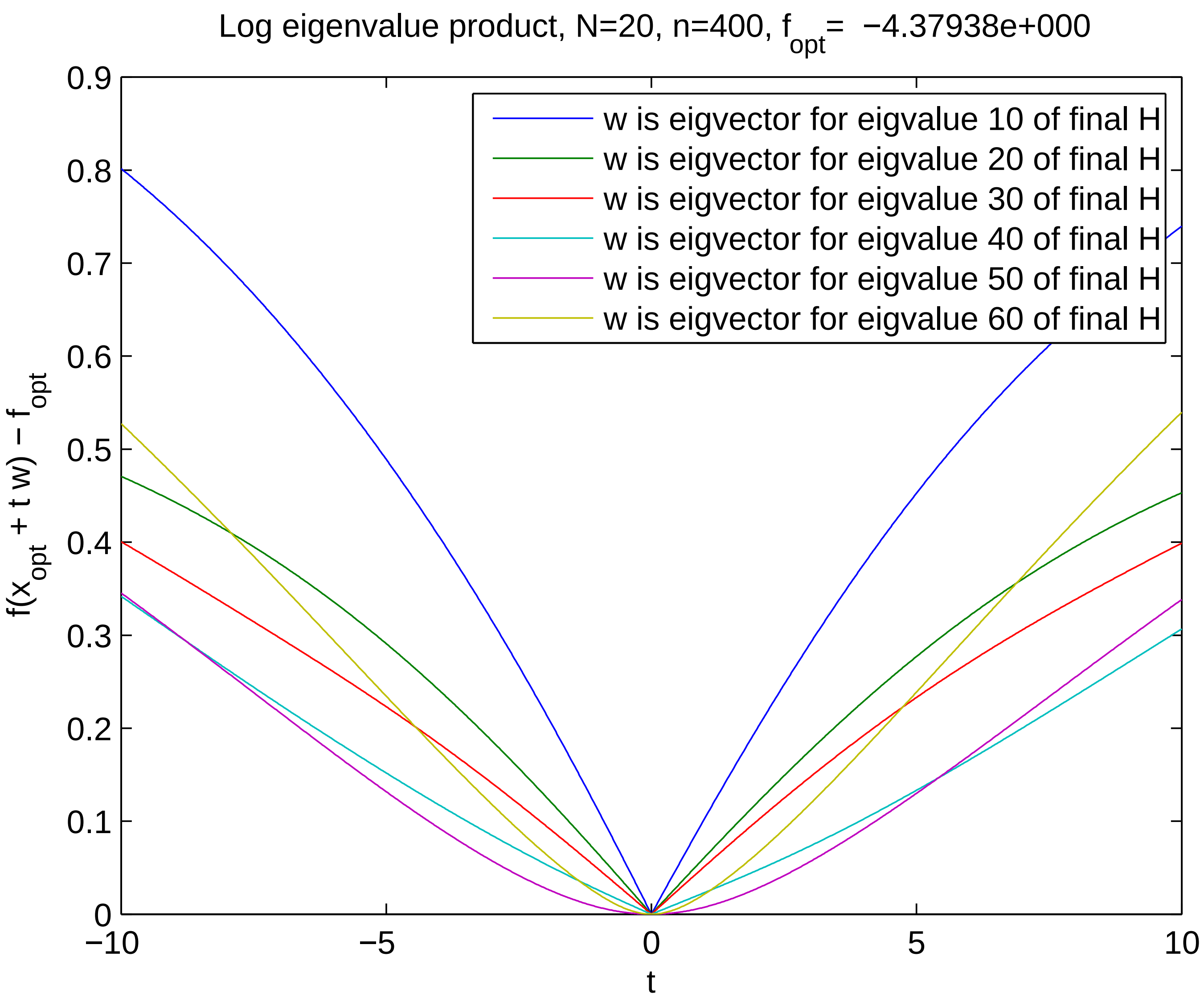
$$\lambda_6(A \circ X) \approx \ldots \approx \lambda_{14}(A \circ X).$$

Matrix theory says that imposing multiplicity m on an eigenvalue a matrix $\in S^N$ is $\frac{m(m+1)}{2} - 1$ conditions, or 44 when $m = 9$, so the dimension of the V-space at this minimizer is 44.

Tiny eigenvalues of H correspond to huge curvature, which corresponds to V-space directions.

Thus BFGS *automatically* detected the U and V space partitioning without knowing anything about the mathematical structure of f!

Variation of f from Minimizer, along EigVecs of H



Eigenvalues of H numbered *smallest to largest*

BFGS Theory for Special Nonsmooth Functions

Convergence results for BFGS with Armijo-Wolfe line search when f is nonsmooth are limited to very special cases.

- $f(x) = |x|$ (one variable!): sequence generated converging to 0 is related to a certain binary expansion of the starting point (A.S. Lewis and M.L.O., 2013)

BFGS Theory for Special Nonsmooth Functions

Convergence results for BFGS with Armijo-Wolfe line search when f is nonsmooth are limited to very special cases.

- $f(x) = |x|$ (one variable!): sequence generated converging to 0 is related to a certain binary expansion of the starting point (A.S. Lewis and M.L.O., 2013)
- $f(x) = |x_1| + x_2$: $f(x) \downarrow -\infty$ (A.S. Lewis and Shanshan Zhang, 2015)

BFGS Theory for Special Nonsmooth Functions

Convergence results for BFGS with Armijo-Wolfe line search when f is nonsmooth are limited to very special cases.

- $f(x) = |x|$ (one variable!): sequence generated converging to 0 is related to a certain binary expansion of the starting point (A.S. Lewis and M.L.O., 2013)
- $f(x) = |x_1| + x_2$: $f(x) \downarrow -\infty$ (A.S. Lewis and Shanshan Zhang, 2015)
- $f(x) = |x_1| + \sum_{i=2}^{n} x_i$: eventually a direction is identified on which f is unbounded below (Yuchen Xie and A. Waechter, 2017)

BFGS Theory for Special Nonsmooth Functions

Convergence results for BFGS with Armijo-Wolfe line search when f is nonsmooth are limited to very special cases.

- $f(x) = |x|$ (one variable!): sequence generated converging to 0 is related to a certain binary expansion of the starting point (A.S. Lewis and M.L.O., 2013)
- $f(x) = |x_1| + x_2$: $f(x) \downarrow -\infty$ (A.S. Lewis and Shanshan Zhang, 2015)
- $f(x) = |x_1| + \sum_{i=2}^{n} x_i$: eventually a direction is identified on which f is unbounded below (Yuchen Xie and A. Waechter, 2017)
- $f(x) = \sqrt{\sum_{i=1}^{n} x_i^2}$: iterates converge to $[0, \ldots, 0]$ (Jiayi Guo and A.S. Lewis, 2017) (proof based on Powell (1976))

Convergence results for BFGS with Armijo-Wolfe line search when f is nonsmooth are limited to very special cases.

- $f(x) = |x|$ (one variable!): sequence generated converging to 0 is related to a certain binary expansion of the starting point (A.S. Lewis and M.L.O., 2013)
- $f(x) = |x_1| + x_2$: $f(x) \downarrow -\infty$ (A.S. Lewis and Shanshan Zhang, 2015)
- $f(x) = |x_1| + \sum_{i=2}^n x_i$: eventually a direction is identified on which f is unbounded below (Yuchen Xie and A. Waechter, 2017)
- $f(x) = \sqrt{\sum_{i=1}^n x_i^2}$: iterates converge to $[0, \ldots, 0]$ (Jiayi Guo and A.S. Lewis, 2017) (proof based on Powell (1976))
- $f(x) = |x_1| + x_2^2$: remains open!

Challenge: General Nonsmooth Case

Challenge: General Nonsmooth Case

Assume f is locally Lipschitz with bounded level sets and is semi-algebraic (its graph is a finite union of sets each defined by a finite list of polynomial inequalities)

Challenge: General Nonsmooth Case

Assume f is locally Lipschitz with bounded level sets and is semi-algebraic (its graph is a finite union of sets each defined by a finite list of polynomial inequalities)

Assume the initial x and H are generated randomly (e.g. from normal and Wishart distributions)

Challenge: General Nonsmooth Case

Assume f is locally Lipschitz with bounded level sets and is semi-algebraic (its graph is a finite union of sets each defined by a finite list of polynomial inequalities)

Assume the initial x and H are generated randomly (e.g. from normal and Wishart distributions)

Prove or disprove that the following hold with probability one:

Challenge: General Nonsmooth Case

Assume f is locally Lipschitz with bounded level sets and is semi-algebraic (its graph is a finite union of sets each defined by a finite list of polynomial inequalities)

Assume the initial x and H are generated randomly (e.g. from normal and Wishart distributions)

Prove or disprove that the following hold with probability one:

1. BFGS generates an infinite sequence $\{x\}$ with f differentiable at all iterates

Challenge: General Nonsmooth Case

Assume f is locally Lipschitz with bounded level sets and is semi-algebraic (its graph is a finite union of sets each defined by a finite list of polynomial inequalities)

Assume the initial x and H are generated randomly (e.g. from normal and Wishart distributions)

Prove or disprove that the following hold with probability one:

1. BFGS generates an infinite sequence $\{x\}$ with f differentiable at all iterates
2. Any cluster point $\bar{x}$ is Clarke stationary

Challenge: General Nonsmooth Case

Assume f is locally Lipschitz with bounded level sets and is semi-algebraic (its graph is a finite union of sets each defined by a finite list of polynomial inequalities)

Assume the initial x and H are generated randomly (e.g. from normal and Wishart distributions)

Prove or disprove that the following hold with probability one:

1. BFGS generates an infinite sequence $\{x\}$ with f differentiable at all iterates
2. Any cluster point $\bar{x}$ is Clarke stationary
3. The sequence of function values generated (including all of the line search iterates) converges to $f(\bar{x})$ R-linearly

Challenge: General Nonsmooth Case

Assume f is locally Lipschitz with bounded level sets and is semi-algebraic (its graph is a finite union of sets each defined by a finite list of polynomial inequalities)

Assume the initial x and H are generated randomly (e.g. from normal and Wishart distributions)

Prove or disprove that the following hold with probability one:

1. BFGS generates an infinite sequence $\{x\}$ with f differentiable at all iterates
2. Any cluster point $\bar{x}$ is Clarke stationary
3. The sequence of function values generated (including all of the line search iterates) converges to $f(\bar{x})$ R-linearly
4. If $\{x\}$ converges to $\bar{x}$ where f is "partly smooth" w.r.t. a manifold $\mathcal{M}$ then the subspace defined by the eigenvectors corresponding to eigenvalues of H converging to zero converges to the "V-space" of f w.r.t. $\mathcal{M}$ at $\bar{x}$

Some BFGS Nonsmooth Success Stories

- Design of fixed-order controllers for linear dynamical systems with input and output (D. Henrion and M.L.O., 2006, and many subsequent users of our HIFOO (H-Infinity Fixed Order Optimization) toolbox)

Some BFGS Nonsmooth Success Stories

- Design of fixed-order controllers for linear dynamical systems with input and output (D. Henrion and M.L.O., 2006, and many subsequent users of our HIFOO (H-Infinity Fixed Order Optimization) toolbox)
- Shape optimization for spectral functions of Dirichlet-Laplacian operators (B. Osting, 2010)

- Design of fixed-order controllers for linear dynamical systems with input and output (D. Henrion and M.L.O., 2006, and many subsequent users of our HIFOO (H-Infinity Fixed Order Optimization) toolbox)
- Shape optimization for spectral functions of Dirichlet-Laplacian operators (B. Osting, 2010)
- Condition metric optimization (P. Boito and J. Dedieu, 2010)

Some BFGS Nonsmooth Success Stories

- Design of fixed-order controllers for linear dynamical systems with input and output (D. Henrion and M.L.O., 2006, and many subsequent users of our HIFOO (H-Infinity Fixed Order Optimization) toolbox)
- Shape optimization for spectral functions of Dirichlet-Laplacian operators (B. Osting, 2010)
- Condition metric optimization (P. Boito and J. Dedieu, 2010)

Software is available: HANSO

Extensions of BFGS for Nonsmooth Optimization

A combined BFGS-Gradient Sampling method with convergence theory (F.E. Curtis and X. Que, 2015)

Extensions of BFGS for Nonsmooth Optimization

A combined BFGS-Gradient Sampling method with convergence theory (F.E. Curtis and X. Que, 2015)

Constrained Problems

$$\min \; f(x)$$
$$\text{subject to } c_i(x) \leq 0, \quad i = 1, \ldots, p$$

where f and $c_1, \ldots, c_p$ are locally Lipschitz but may not be differentiable at local minimizers.

A combined BFGS-Gradient Sampling method with convergence theory (F.E. Curtis and X. Que, 2015)

Constrained Problems

$$\min \ f(x)$$
$$\text{subject to } c_i(x) \leq 0, \quad i = 1, \ldots, p$$

where f and $c_1, \ldots, c_p$ are locally Lipschitz but may not be differentiable at local minimizers.

A successive quadratic programming (SQP) BFGS method applied to challenging problems in static-output-feedback control design (F.E. Curtis, T. Mitchell and M.L.O., 2015).

Extensions of BFGS for Nonsmooth Optimization

A combined BFGS-Gradient Sampling method with convergence theory (F.E. Curtis and X. Que, 2015)

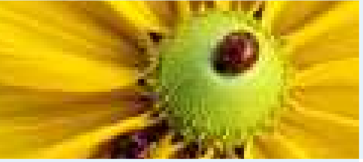
Constrained Problems

$$\begin{gathered}\min \; f(x) \\ \text{subject to } c_i(x) \le 0, \quad i = 1, \ldots, p\end{gathered}$$

where f and $c_1, \ldots, c_p$ are locally Lipschitz but may not be differentiable at local minimizers.

A successive quadratic programming (SQP) BFGS method applied to challenging problems in static-output-feedback control design (F.E. Curtis, T. Mitchell and M.L.O., 2015).

Although there are no theoretical results, it is much more efficient and effective than the SQP Gradient Sampling method which does have convergence results.



Extensions of BFGS for Nonsmooth Optimization

A combined BFGS-Gradient Sampling method with convergence theory (F.E. Curtis and X. Que, 2015)

Constrained Problems

$$\min \; f(x)$$
$$\text{subject to } c_i(x) \leq 0, \quad i = 1, \ldots, p$$

where f and $c_1, \ldots, c_p$ are locally Lipschitz but may not be differentiable at local minimizers.

A successive quadratic programming (SQP) BFGS method applied to challenging problems in static-output-feedback control design (F.E. Curtis, T. Mitchell and M.L.O., 2015).

Although there are no theoretical results, it is much more efficient and effective than the SQP Gradient Sampling method which does have convergence results.

Software is available: GRANSO

A Difficult Nonconvex Problem from Nesterov

An Aside: Chebyshev Polynomials

A sequence of orthogonal polynomials defined on $[-1, 1]$ by

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

An Aside: Chebyshev Polynomials

A sequence of orthogonal polynomials defined on $[-1, 1]$ by

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

So $T_2(x) = 2x^2 - 1$,

An Aside: Chebyshev Polynomials

A sequence of orthogonal polynomials defined on $[-1, 1]$ by

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

So $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3$, etc.

An Aside: Chebyshev Polynomials

A sequence of orthogonal polynomials defined on $[-1, 1]$ by

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

So $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3$, etc.

Important properties that can be proved easily include

An Aside: Chebyshev Polynomials

A sequence of orthogonal polynomials defined on $[-1, 1]$ by

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

So $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3$, etc.

Important properties that can be proved easily include

- $T_n(x) = \cos(n \cos^{-1}(x))$

An Aside: Chebyshev Polynomials

A sequence of orthogonal polynomials defined on $[-1, 1]$ by

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

So $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3$, etc.

Important properties that can be proved easily include

- $T_n(x) = \cos(n \cos^{-1}(x))$
- $T_m(T_n(x)) = T_{mn}(x)$

An Aside: Chebyshev Polynomials

A sequence of orthogonal polynomials defined on $[-1, 1]$ by

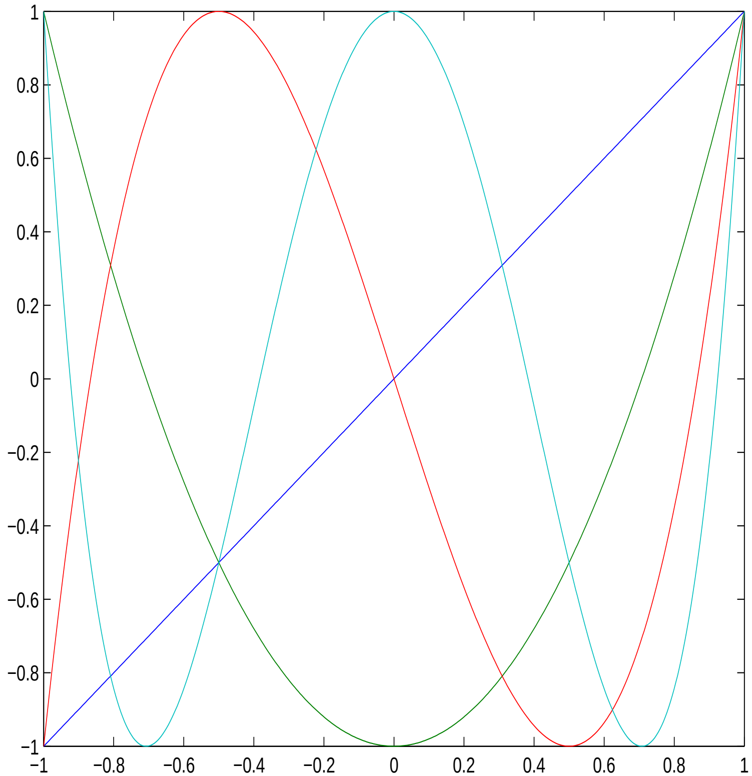
$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

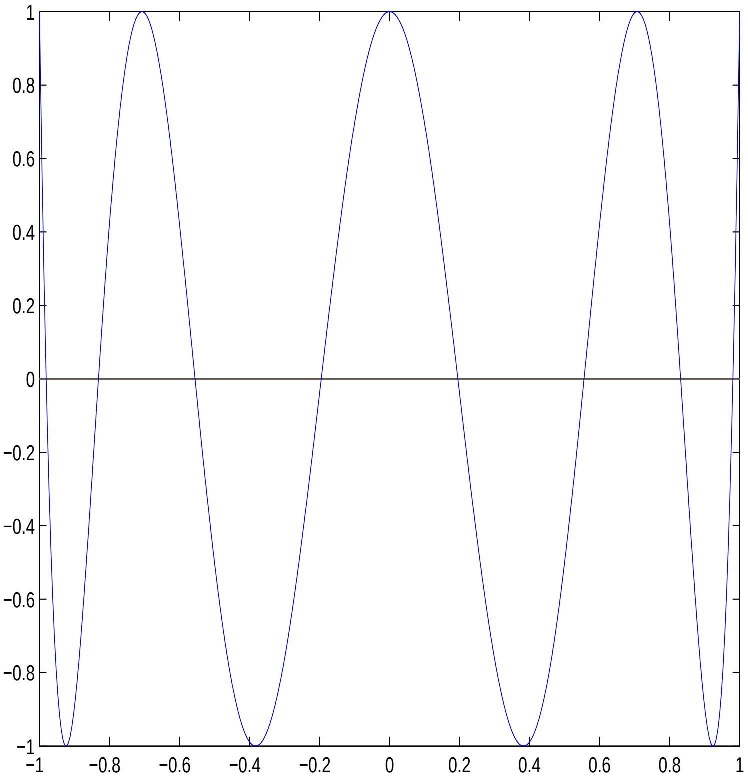
So $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3$, etc.

Important properties that can be proved easily include

- $T_n(x) = \cos(n \cos^{-1}(x))$
- $T_m(T_n(x)) = T_{mn}(x)$
- $\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} T_i(x) T_j(x) dx = 0$ if $i \neq j$

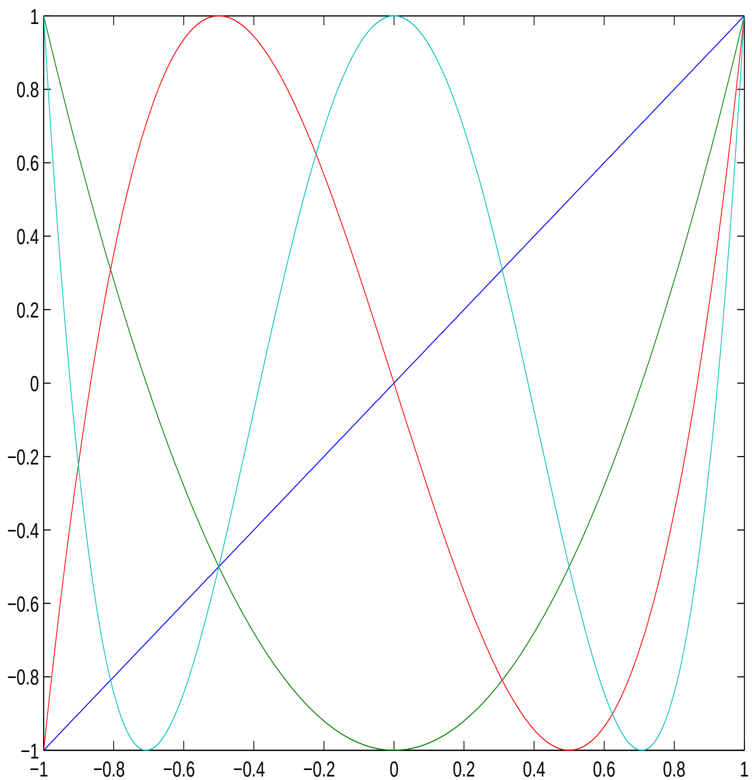
Plots of Chebyshev Polynomials

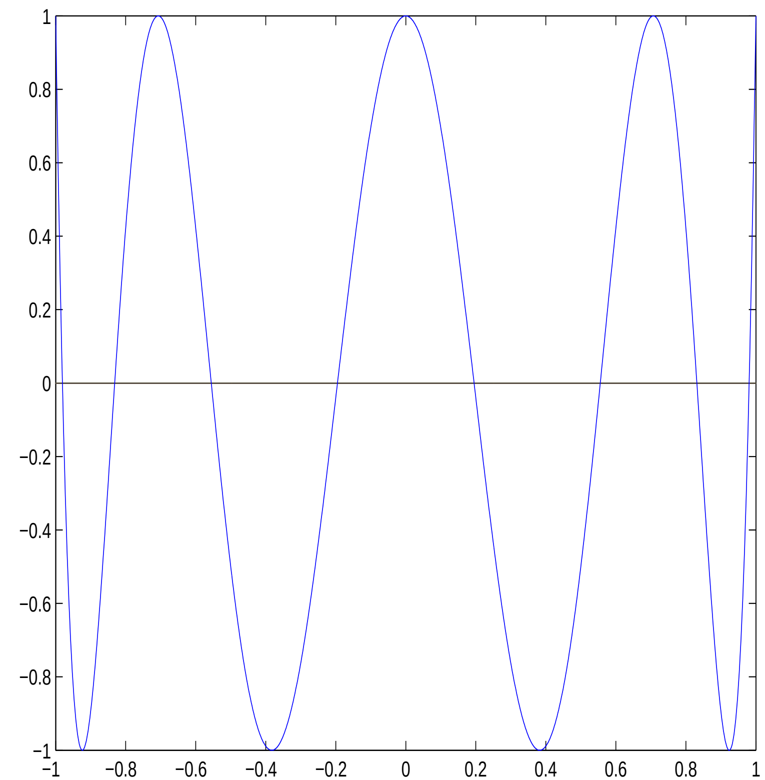




Left: Plots of $T_0(x), \ldots, T_4(x)$ Right: Plot of $T_8(x)$.

Plots of Chebyshev Polynomials





Left: Plots of $T_0(x), \ldots, T_4(x)$ Right: Plot of $T_8(x)$.

Question: How many extrema does $T_n(x)$ have in $[-1, 1]$?

Nesterov's Chebyshev-Rosenbrock Functions

Consider the function

$$N_p(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|^p, \quad \text{where } p \geq 1$$

Nesterov's Chebyshev-Rosenbrock Functions

Consider the function

$$N_p(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|^p, \quad \text{where } p \geq 1$$

The unique minimizer is $x^* = [1, 1, \ldots, 1]^T$ with $N_p(x^*) = 0$.

Nesterov's Chebyshev-Rosenbrock Functions

Consider the function

$$N_p(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|^p, \quad \text{where } p \geq 1$$

The unique minimizer is $x^* = [1, 1, \ldots, 1]^T$ with $N_p(x^*) = 0$.

Define $\hat{x} = [-1, 1, 1, \ldots, 1]^T$ with $N_p(\hat{x}) = 1$ and the manifold

$$\mathcal{M}_N = \{x : x_{i+1} = 2x_i^2 - 1, \quad i = 1, \ldots, n-1\}$$

Nesterov's Chebyshev-Rosenbrock Functions

Consider the function

$$N_p(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|^p, \quad \text{where } p \geq 1$$

The unique minimizer is $x^* = [1, 1, \ldots, 1]^T$ with $N_p(x^*) = 0$.

Define $\hat{x} = [-1, 1, 1, \ldots, 1]^T$ with $N_p(\hat{x}) = 1$ and the manifold

$$\mathcal{M}_N = \{x : x_{i+1} = 2x_i^2 - 1, \quad i = 1, \ldots, n-1\}$$

For $x \in \mathcal{M}_N$, e.g. $x = x^*$ or $x = \hat{x}$, the 2nd term of N_p is zero. Starting at $\hat{x}$, BFGS needs to approximately follow $\mathcal{M}_N$ to reach x^* (unless it "gets lucky").

Nesterov's Chebyshev-Rosenbrock Functions

Consider the function

$$N_p(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|^p, \quad \text{where } p \geq 1$$

The unique minimizer is $x^* = [1, 1, \ldots, 1]^T$ with $N_p(x^*) = 0$.

Define $\hat{x} = [-1, 1, 1, \ldots, 1]^T$ with $N_p(\hat{x}) = 1$ and the manifold

$$\mathcal{M}_N = \{x : x_{i+1} = 2x_i^2 - 1, \quad i = 1, \ldots, n-1\}$$

For $x \in \mathcal{M}_N$, e.g. $x = x^*$ or $x = \hat{x}$, the 2nd term of N_p is zero. Starting at $\hat{x}$, BFGS needs to approximately follow $\mathcal{M}_N$ to reach x^* (unless it "gets lucky").

When $p = 2$: N_2 is smooth but not convex. Starting at $\hat{x}$:

Nesterov's Chebyshev-Rosenbrock Functions

Consider the function

$$N_p(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|^p, \quad \text{where } p \geq 1$$

The unique minimizer is $x^* = [1, 1, \ldots, 1]^T$ with $N_p(x^*) = 0$.

Define $\hat{x} = [-1, 1, 1, \ldots, 1]^T$ with $N_p(\hat{x}) = 1$ and the manifold

$$\mathcal{M}_N = \{x : x_{i+1} = 2x_i^2 - 1, \quad i = 1, \ldots, n-1\}$$

For $x \in \mathcal{M}_N$, e.g. $x = x^*$ or $x = \hat{x}$, the 2nd term of N_p is zero. Starting at $\hat{x}$, BFGS needs to approximately follow $\mathcal{M}_N$ to reach x^* (unless it "gets lucky").

When $p = 2$: N_2 is smooth but not convex. Starting at $\hat{x}$:

- $n = 5$: BFGS needs 370 iterations to reduce N_2 below 10^{-15}

Nesterov's Chebyshev-Rosenbrock Functions

Consider the function

$$N_p(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|^p, \quad \text{where } p \geq 1$$

The unique minimizer is $x^* = [1, 1, \ldots, 1]^T$ with $N_p(x^*) = 0$.

Define $\hat{x} = [-1, 1, 1, \ldots, 1]^T$ with $N_p(\hat{x}) = 1$ and the manifold

$$\mathcal{M}_N = \{x : x_{i+1} = 2x_i^2 - 1, \quad i = 1, \ldots, n-1\}$$

For $x \in \mathcal{M}_N$, e.g. $x = x^*$ or $x = \hat{x}$, the 2nd term of N_p is zero. Starting at $\hat{x}$, BFGS needs to approximately follow $\mathcal{M}_N$ to reach x^* (unless it "gets lucky").

When $p = 2$: N_2 is smooth but not convex. Starting at $\hat{x}$:

- $n = 5$: BFGS needs 370 iterations to reduce N_2 below 10^{-15}
- $n = 10$: needs $\sim$ 50,000 iterations to reduce N_2 below 10^{-15}

even though N_2 is *smooth*!

Nesterov's Chebyshev-Rosenbrock Functions

Consider the function

$$N_p(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|^p, \quad \text{where } p \geq 1$$

The unique minimizer is $x^* = [1, 1, \ldots, 1]^T$ with $N_p(x^*) = 0$.

Define $\hat{x} = [-1, 1, 1, \ldots, 1]^T$ with $N_p(\hat{x}) = 1$ and the manifold

$$\mathcal{M}_N = \{x : x_{i+1} = 2x_i^2 - 1, \quad i = 1, \ldots, n-1\}$$

For $x \in \mathcal{M}_N$, e.g. $x = x^*$ or $x = \hat{x}$, the 2nd term of N_p is zero. Starting at $\hat{x}$, BFGS needs to approximately follow $\mathcal{M}_N$ to reach x^* (unless it "gets lucky").

When $p = 2$: N_2 is smooth but not convex. Starting at $\hat{x}$:

- $n = 5$: BFGS needs 370 iterations to reduce N_2 below 10^{-15}
- $n = 10$: needs $\sim$ 50,000 iterations to reduce N_2 below 10^{-15}

even though N_2 is *smooth*! . . . In the last few iterations, we observe superlinear convergence!

Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

$$x_{i+1} = 2x_i^2 - 1 = T_2(x_i) = T_2(T_2(x_{i-1}))$$

Why BFGS Takes So Many Iterations to Minimize N_2

Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

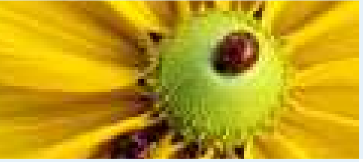
$$
\begin{aligned}
x_{i+1} = 2x_i^2 - 1 = T_2(x_i) &= T_2(T_2(x_{i-1})) \\
&= T_2(T_2(\ldots T_2(x_1)\ldots)) = T_{2^i}(x_1).
\end{aligned}
$$

Why BFGS Takes So Many Iterations to Minimize N_2

Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

$$\begin{aligned} x_{i+1} = 2x_i^2 - 1 &= T_2(x_i) = T_2(T_2(x_{i-1})) \\ &= T_2(T_2(\ldots T_2(x_1)\ldots)) = T_{2^i}(x_1). \end{aligned}$$

To move from $\hat{x}$ to x^* along the manifold $\mathcal{M}_N$ *exactly* requires



Why BFGS Takes So Many Iterations to Minimize N_2

Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

$$\begin{aligned} x_{i+1} &= 2x_i^2 - 1 = T_2(x_i) = T_2(T_2(x_{i-1})) \\ &= T_2(T_2(\ldots T_2(x_1)\ldots)) = T_{2^i}(x_1). \end{aligned}$$

To move from $\hat{x}$ to x^* along the manifold $\mathcal{M}_N$ *exactly* requires

- x_1 to change from -1 to 1

Why BFGS Takes So Many Iterations to Minimize N_2

Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

$$x_{i+1} = 2x_i^2 - 1 = T_2(x_i) = T_2(T_2(x_{i-1}))$$
$$= T_2(T_2(\ldots T_2(x_1)\ldots)) = T_{2^i}(x_1).$$

To move from $\hat{x}$ to x^* along the manifold $\mathcal{M}_N$ *exactly* requires

- x_1 to change from -1 to 1
- $x_2 = 2x_1^2 - 1$ to trace the graph of $T_2(x_1)$ on $[-1, 1]$

Why BFGS Takes So Many Iterations to Minimize N_2

Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

$$\begin{aligned} x_{i+1} &= 2x_i^2 - 1 = T_2(x_i) = T_2(T_2(x_{i-1})) \\ &= T_2(T_2(\ldots T_2(x_1)\ldots)) = T_{2^i}(x_1). \end{aligned}$$

To move from $\hat{x}$ to x^* along the manifold $\mathcal{M}_N$ *exactly* requires

- x_1 to change from -1 to 1
- $x_2 = 2x_1^2 - 1$ to trace the graph of $T_2(x_1)$ on $[-1, 1]$
- $x_3 = T_2(T_2(x))$ to trace the graph of $T_4(x_1)$ on $[-1, 1]$

Why BFGS Takes So Many Iterations to Minimize N_2

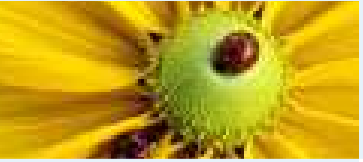
Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

$$
\begin{aligned}
x_{i+1} = 2x_i^2 - 1 = T_2(x_i) &= T_2(T_2(x_{i-1})) \\
&= T_2(T_2(\ldots T_2(x_1)\ldots)) = T_{2^i}(x_1).
\end{aligned}
$$

To move from $\hat{x}$ to x^* along the manifold $\mathcal{M}_N$ *exactly* requires

- x_1 to change from -1 to 1
- $x_2 = 2x_1^2 - 1$ to trace the graph of $T_2(x_1)$ on $[-1, 1]$
- $x_3 = T_2(T_2(x))$ to trace the graph of $T_4(x_1)$ on $[-1, 1]$
- $x_n = T_{2^{n-1}}(x)$ to trace the graph of $T_{2^{n-1}}(x_1)$ on $[-1, 1]$

which has $2^{n-1} - 1$ extrema in $(-1, 1)$.



Why BFGS Takes So Many Iterations to Minimize N_2

Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

$$\begin{aligned} x_{i+1} = 2x_i^2 - 1 = T_2(x_i) &= T_2(T_2(x_{i-1})) \\ &= T_2(T_2(\ldots T_2(x_1)\ldots)) = T_{2^i}(x_1). \end{aligned}$$

To move from $\hat{x}$ to x^* along the manifold $\mathcal{M}_N$ *exactly* requires

- x_1 to change from -1 to 1
- $x_2 = 2x_1^2 - 1$ to trace the graph of $T_2(x_1)$ on $[-1, 1]$
- $x_3 = T_2(T_2(x))$ to trace the graph of $T_4(x_1)$ on $[-1, 1]$
- $x_n = T_{2^{n-1}}(x)$ to trace the graph of $T_{2^{n-1}}(x_1)$ on $[-1, 1]$

which has $2^{n-1} - 1$ extrema in $(-1, 1)$.
Even though BFGS will *not* track the manifold $\mathcal{M}_N$ exactly, it will follow it approximately. So, since the manifold is highly oscillatory, BFGS must take relatively short steps to obtain reduction in N_2 in the line search, and hence *many* iterations!

Why BFGS Takes So Many Iterations to Minimize N_2

Let $T_i(x)$ denote the ith Chebyshev polynomial. For $x \in \mathcal{M}_N$,

$$\begin{aligned} x_{i+1} = 2x_i^2 - 1 = T_2(x_i) &= T_2(T_2(x_{i-1})) \\ &= T_2(T_2(\ldots T_2(x_1)\ldots)) = T_{2^i}(x_1). \end{aligned}$$

To move from $\hat{x}$ to x^* along the manifold $\mathcal{M}_N$ *exactly* requires

- x_1 to change from -1 to 1
- $x_2 = 2x_1^2 - 1$ to trace the graph of $T_2(x_1)$ on $[-1, 1]$
- $x_3 = T_2(T_2(x))$ to trace the graph of $T_4(x_1)$ on $[-1, 1]$
- $x_n = T_{2^{n-1}}(x)$ to trace the graph of $T_{2^{n-1}}(x_1)$ on $[-1, 1]$

which has $2^{n-1} - 1$ extrema in $(-1, 1)$.
Even though BFGS will *not* track the manifold $\mathcal{M}_N$ exactly, it will follow it approximately. So, since the manifold is highly oscillatory, BFGS must take relatively short steps to obtain reduction in N_2 in the line search, and hence *many* iterations!

Newton's method is not much faster, although it converges quadratically at the end.

First Nonsmooth Variant of Nesterov's Function

$$N_1(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|$$

First Nonsmooth Variant of Nesterov's Function

$$N_1(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|$$

N_1 is nonsmooth (though locally Lipschitz) as well as nonconvex. The second term is still zero on the manifold $\mathcal{M}_N$, but N_1 is not differentiable on $\mathcal{M}_N$.

First Nonsmooth Variant of Nesterov's Function

$$N_1(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|$$

N_1 is nonsmooth (though locally Lipschitz) as well as nonconvex. The second term is still zero on the manifold $\mathcal{M}_N$, but N_1 is not differentiable on $\mathcal{M}_N$.

However, N_1 is regular at $x \in \mathcal{M}_N$ and partly smooth at x w.r.t. $\mathcal{M}_N$.

First Nonsmooth Variant of Nesterov's Function

$$N_1(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|$$

N_1 is nonsmooth (though locally Lipschitz) as well as nonconvex. The second term is still zero on the manifold $\mathcal{M}_N$, but N_1 is not differentiable on $\mathcal{M}_N$.

However, N_1 is regular at $x \in \mathcal{M}_N$ and partly smooth at x w.r.t. $\mathcal{M}_N$.

We cannot initialize BFGS at $\hat{x}$, so starting at normally distributed random points:

First Nonsmooth Variant of Nesterov's Function

$$N_1(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|$$

N_1 is nonsmooth (though locally Lipschitz) as well as nonconvex. The second term is still zero on the manifold $\mathcal{M}_N$, but N_1 is not differentiable on $\mathcal{M}_N$.

However, N_1 is regular at $x \in \mathcal{M}_N$ and partly smooth at x w.r.t. $\mathcal{M}_N$.

We cannot initialize BFGS at $\hat{x}$, so starting at normally distributed random points:

- $n = 5$: BFGS reduces N_1 only to about 10^{-2} in 10,000 iterations

First Nonsmooth Variant of Nesterov's Function

$$N_1(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|$$

N_1 is nonsmooth (though locally Lipschitz) as well as nonconvex. The second term is still zero on the manifold $\mathcal{M}_N$, but N_1 is not differentiable on $\mathcal{M}_N$.

However, N_1 is regular at $x \in \mathcal{M}_N$ and partly smooth at x w.r.t. $\mathcal{M}_N$.

We cannot initialize BFGS at $\hat{x}$, so starting at normally distributed random points:

- $n = 5$: BFGS reduces N_1 only to about 10^{-2} in 10,000 iterations
- $n = 10$: BFGS reduces N_1 only to about 5×10^{-2} in 10,000 iterations

$$N_1(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|$$

N_1 is nonsmooth (though locally Lipschitz) as well as nonconvex. The second term is still zero on the manifold $\mathcal{M}_N$, but N_1 is not differentiable on $\mathcal{M}_N$.

However, N_1 is regular at $x \in \mathcal{M}_N$ and partly smooth at x w.r.t. $\mathcal{M}_N$.

We cannot initialize BFGS at $\hat{x}$, so starting at normally distributed random points:

- $n = 5$: BFGS reduces N_1 only to about 10^{-2} in 10,000 iterations
- $n = 10$: BFGS reduces N_1 only to about 5×10^{-2} in 10,000 iterations

The method appears to be converging, very slowly, but may be having numerical difficulties.

Second Nonsmooth Variant of Nesterov's Function

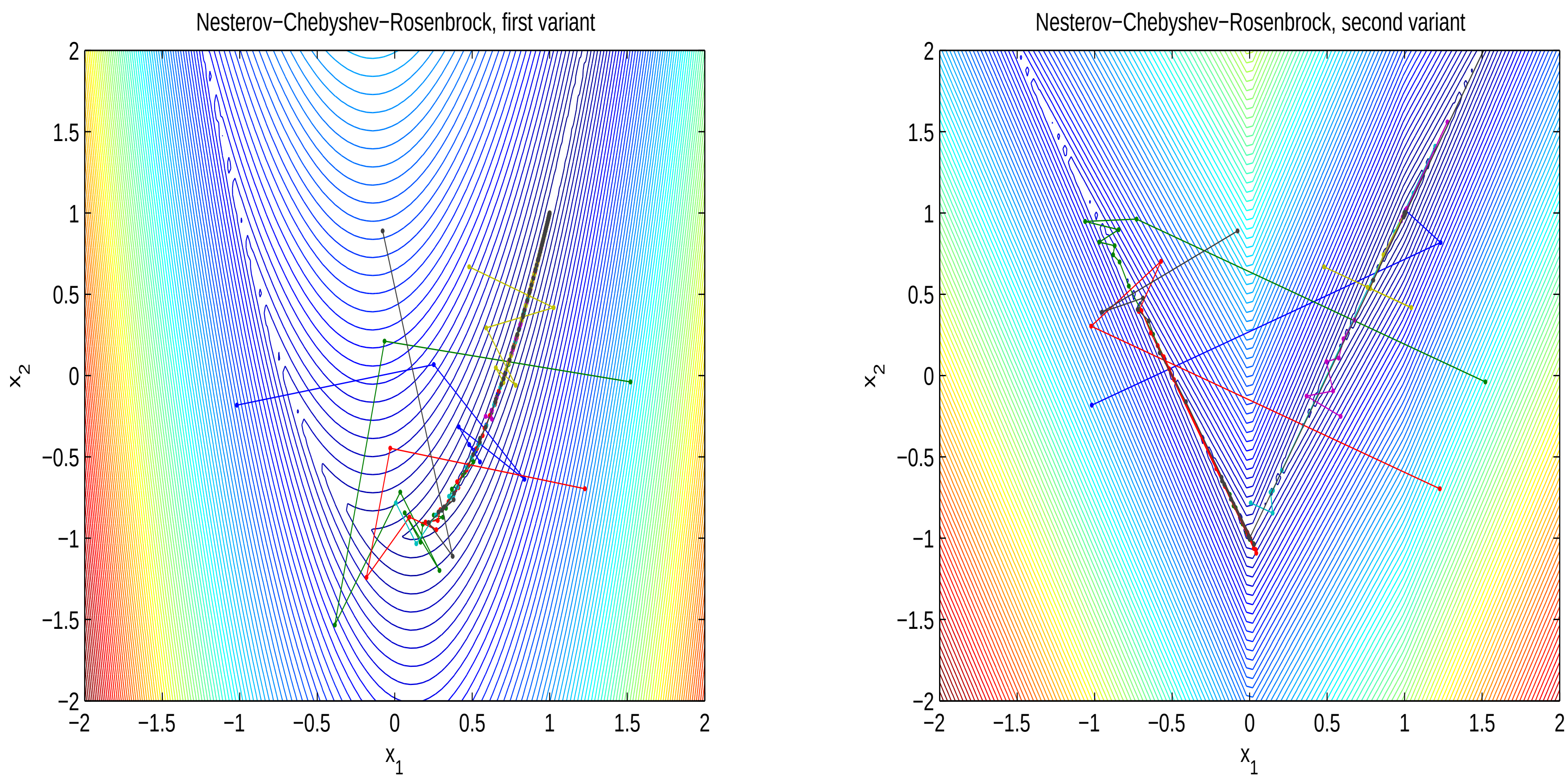
$$\widehat{N}_1(x) = \frac{1}{4}|x_1 - 1| + \sum_{i=1}^{n-1} |x_{i+1} - 2|x_i| + 1|.$$

Again, the unique global minimizer is x^*. The second term is zero on the set

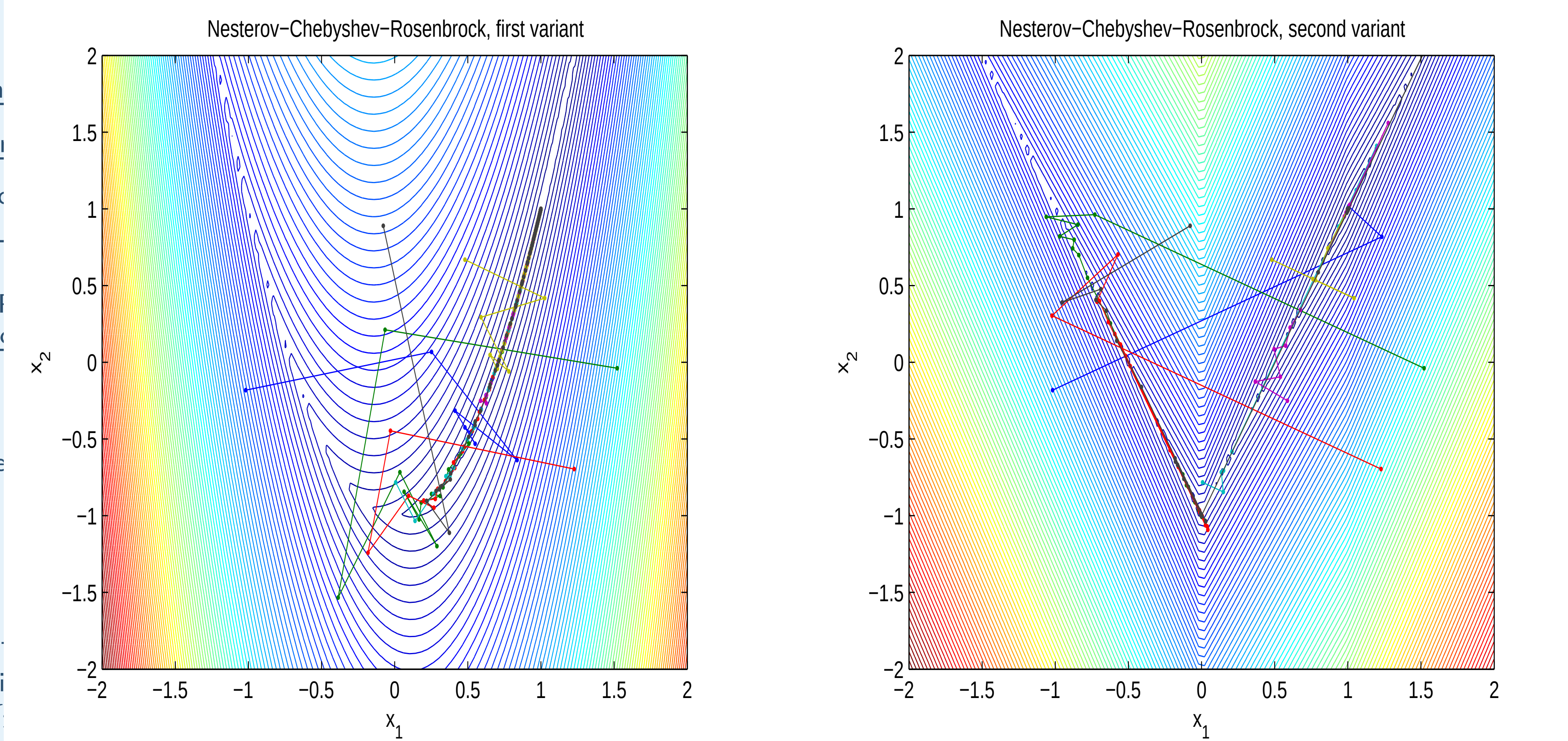
$$S = \{x : x_{i+1} = 2|x_i| - 1, \quad i = 1, \ldots, n-1\}$$

but S is not a manifold: it has "corners".

Contour Plots of the Nonsmooth Variants for $n = 2$



Contour plots of nonsmooth Chebyshev-Rosenbrock functions N_1 (left) and $\widehat{N}_1$ (right), with $n = 2$, with iterates generated by BFGS initialized at 7 different randomly generated points.



Contour plots of nonsmooth Chebyshev-Rosenbrock functions N_1 (left) and $\widehat{N}_1$ (right), with $n = 2$, with iterates generated by BFGS initialized at 7 different randomly generated points.
On the left, always get convergence to $x^* = [1, 1]^T$. On the right, most runs converge to $[1, 1]$ but some go to $x = [0, -1]^T$.

Properties of the Second Nonsmooth Variant $\widehat{N}_1$

When $n = 2$, the point $x = [0, -1]^T$ is Clarke stationary for the second nonsmooth variant $\widehat{N}_1$. We can see this because zero is in the convex hull of the gradient limits for $\widehat{N}_1$ at the point x.

Properties of the Second Nonsmooth Variant $\widehat{N}_1$

When $n = 2$, the point $x = [0, -1]^T$ is Clarke stationary for the second nonsmooth variant $\widehat{N}_1$. We can see this because zero is in the convex hull of the gradient limits for $\widehat{N}_1$ at the point x.

However, $x = [0, -1]^T$ is not a local minimizer, because $d = [1, 2]^T$ is a direction of linear descent: $\widehat{N}_1'(x, d) < 0$.

Properties of the Second Nonsmooth Variant $\widehat{N}_1$

When $n = 2$, the point $x = [0, -1]^T$ is Clarke stationary for the second nonsmooth variant $\widehat{N}_1$. We can see this because zero is in the convex hull of the gradient limits for $\widehat{N}_1$ at the point x.

However, $x = [0, -1]^T$ is not a local minimizer, because $d = [1, 2]^T$ is a direction of linear descent: $\widehat{N}_1'(x, d) < 0$.

These two properties mean that $\widehat{N}_1$ is *not regular* at $[0, -1]^T$.

Properties of the Second Nonsmooth Variant $\widehat{N}_1$

When $n = 2$, the point $x = [0, -1]^T$ is Clarke stationary for the second nonsmooth variant $\widehat{N}_1$. We can see this because zero is in the convex hull of the gradient limits for $\widehat{N}_1$ at the point x.

However, $x = [0, -1]^T$ is not a local minimizer, because $d = [1, 2]^T$ is a direction of linear descent: $\widehat{N}_1'(x, d) < 0$.

These two properties mean that $\widehat{N}_1$ is *not regular* at $[0, -1]^T$.

In fact, for $n \geq 2$:

- $\widehat{N}_1$ has 2^{n-1} Clarke stationary points

Properties of the Second Nonsmooth Variant $\widehat{N}_1$

When $n = 2$, the point $x = [0, -1]^T$ is Clarke stationary for the second nonsmooth variant $\widehat{N}_1$. We can see this because zero is in the convex hull of the gradient limits for $\widehat{N}_1$ at the point x.

However, $x = [0, -1]^T$ is not a local minimizer, because $d = [1, 2]^T$ is a direction of linear descent: $\widehat{N}_1'(x, d) < 0$.

These two properties mean that $\widehat{N}_1$ is *not regular* at $[0, -1]^T$.

In fact, for $n \geq 2$:

- $\widehat{N}_1$ has 2^{n-1} Clarke stationary points
- the only local minimizer is the global minimizer x^*

Properties of the Second Nonsmooth Variant $\widehat{N}_1$

When $n = 2$, the point $x = [0, -1]^T$ is Clarke stationary for the second nonsmooth variant $\widehat{N}_1$. We can see this because zero is in the convex hull of the gradient limits for $\widehat{N}_1$ at the point x.

However, $x = [0, -1]^T$ is not a local minimizer, because $d = [1, 2]^T$ is a direction of linear descent: $\widehat{N}_1'(x, d) < 0$.

These two properties mean that $\widehat{N}_1$ is *not regular* at $[0, -1]^T$.

In fact, for $n \geq 2$:

- $\widehat{N}_1$ has 2^{n-1} Clarke stationary points
- the only local minimizer is the global minimizer x^*
- x^* is the only stationary point in the sense of Mordukhovich (i.e., with $0 \in \partial N_1(x)$ where we defined ∂ in Lecture 12) (see also Rockafellar and Wets, *Variational Analysis*, 1998).

(M. Gürbüzbalaban and M.L.O., 2012)

Properties of the Second Nonsmooth Variant $\widehat{N}_1$

When $n = 2$, the point $x = [0, -1]^T$ is Clarke stationary for the second nonsmooth variant $\widehat{N}_1$. We can see this because zero is in the convex hull of the gradient limits for $\widehat{N}_1$ at the point x.

However, $x = [0, -1]^T$ is not a local minimizer, because $d = [1, 2]^T$ is a direction of linear descent: $\widehat{N}_1'(x, d) < 0$.

These two properties mean that $\widehat{N}_1$ is *not regular* at $[0, -1]^T$.

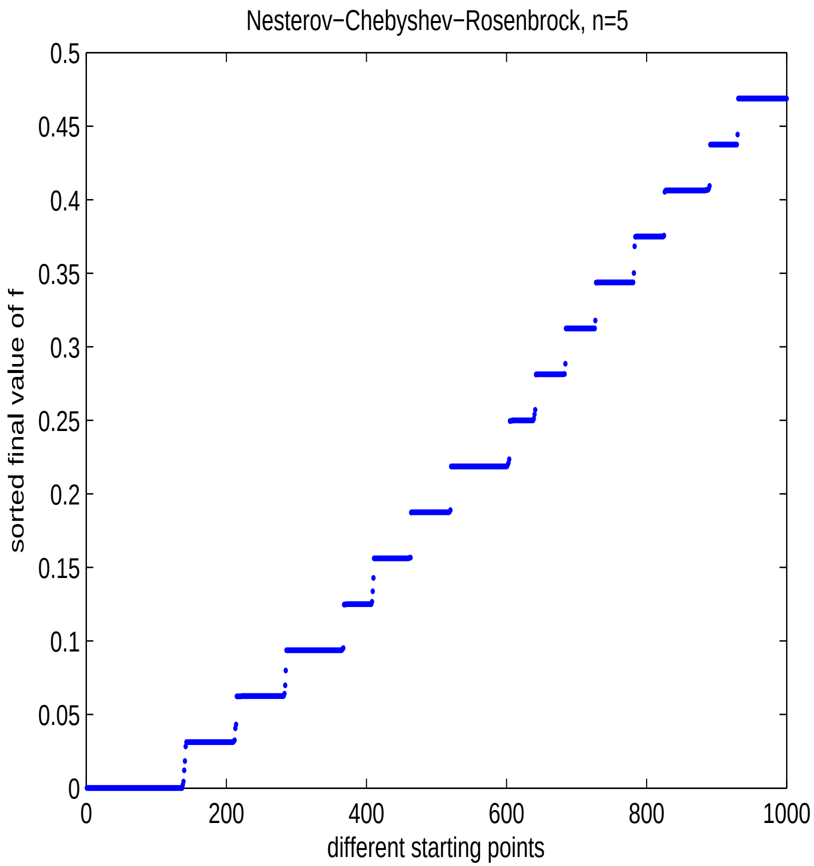
In fact, for $n \geq 2$:

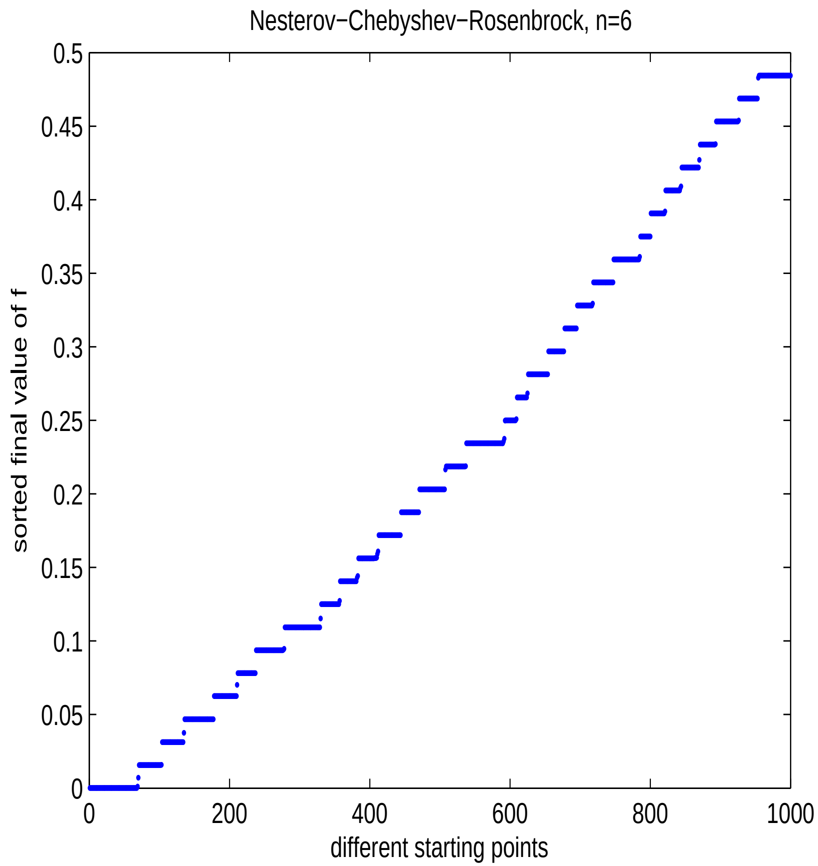
- $\widehat{N}_1$ has 2^{n-1} Clarke stationary points
- the only local minimizer is the global minimizer x^*
- x^* is the only stationary point in the sense of Mordukhovich (i.e., with $0 \in \partial N_1(x)$ where we defined ∂ in Lecture 12) (see also Rockafellar and Wets, *Variational Analysis*, 1998).

(M. Gürbüzbalaban and M.L.O., 2012)

Furthermore, starting from enough randomly generated starting points, BFGS finds all 2^{n-1} Clarke stationary points!

Behavior of BFGS on the Second Nonsmooth Variant





Left: *sorted* final values of $\widehat{N}_1$ for 1000 randomly generated starting points, when $n = 5$: BFGS finds all 16 Clarke stationary points. Right: same with $n = 6$: BFGS finds all 32 Clarke stationary points.

Convergence to Non-Locally-Minimizing Points

When f is *smooth*, convergence of methods such as BFGS to non-locally-minimizing stationary points or local maxima is *possible* but not likely, because of the line search, and such convergence will not be stable under perturbation.

Convergence to Non-Locally-Minimizing Points

When f is *smooth*, convergence of methods such as BFGS to non-locally-minimizing stationary points or local maxima is *possible* but not likely, because of the line search, and such convergence will not be stable under perturbation.

However, this kind of convergence is what we are seeing for the non-regular, non-smooth Nesterov Chebyshev-Rosenbrock example, and it *is* stable under perturbation. The same behavior occurs for gradient sampling or bundle methods.

Convergence to Non-Locally-Minimizing Points

When f is *smooth*, convergence of methods such as BFGS to non-locally-minimizing stationary points or local maxima is *possible* but not likely, because of the line search, and such convergence will not be stable under perturbation.

However, this kind of convergence is what we are seeing for the non-regular, non-smooth Nesterov Chebyshev-Rosenbrock example, and it *is* stable under perturbation. The same behavior occurs for gradient sampling or bundle methods.

Kiwiel (private communication): the Nesterov example is the first he had seen which causes his bundle code to have this behavior.

Convergence to Non-Locally-Minimizing Points

When f is *smooth*, convergence of methods such as BFGS to non-locally-minimizing stationary points or local maxima is *possible* but not likely, because of the line search, and such convergence will not be stable under perturbation.

However, this kind of convergence is what we are seeing for the non-regular, non-smooth Nesterov Chebyshev-Rosenbrock example, and it *is* stable under perturbation. The same behavior occurs for gradient sampling or bundle methods.

Kiwiel (private communication): the Nesterov example is the first he had seen which causes his bundle code to have this behavior.

Nonetheless, we don't know whether, in exact arithmetic, the methods would actually generate sequences converging to the nonminimizing Clarke stationary points. Experiments by Kaku (2011) suggest that the higher the precision used, the more likely BFGS is to eventually move away from such a point.

Limited Memory Methods

Limited Memory BFGS

"Full" BFGS requires storing an $n \times n$ matrix and doing matrix-vector multiplies, which is not possible when n is large.

Limited Memory BFGS

"Full" BFGS requires storing an $n \times n$ matrix and doing matrix-vector multiplies, which is not possible when n is large.

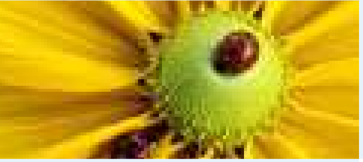
In the 1980s, J. Nocedal and others developed a "limited memory" version of BFGS, with $O(n)$ space and time requirements, which is very widely used for minimizing smooth functions in many variables. At the kth iteration, it applies only the most recent m rank-two updates, defined by

$$(s_j, y_j), \quad j = k-m, \ldots, k-1,$$

where

$$s_j = x_{j+1} - x_j, \quad y_{j+1} = \nabla f(x_{j+1}) - \nabla f(x_j)$$

to an initial inverse Hessian approximation $H_0^{(k)}$.



Limited Memory BFGS

"Full" BFGS requires storing an $n \times n$ matrix and doing matrix-vector multiplies, which is not possible when n is large.

In the 1980s, J. Nocedal and others developed a "limited memory" version of BFGS, with $O(n)$ space and time requirements, which is very widely used for minimizing smooth functions in many variables. At the kth iteration, it applies only the most recent m rank-two updates, defined by

$$(s_j, y_j), \quad j = k-m, \ldots, k-1,$$

where

$$s_j = x_{j+1} - x_j, \quad y_{j+1} = \nabla f(x_{j+1}) - \nabla f(x_j)$$

to an initial inverse Hessian approximation $H_0^{(k)}$.

There are two variants: with "scaling" ($H_0^{(k)} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} I$) and without scaling ($H_0^{(k)} = I$).

Limited Memory BFGS

"Full" BFGS requires storing an $n \times n$ matrix and doing matrix-vector multiplies, which is not possible when n is large.

In the 1980s, J. Nocedal and others developed a "limited memory" version of BFGS, with $O(n)$ space and time requirements, which is very widely used for minimizing smooth functions in many variables. At the kth iteration, it applies only the most recent m rank-two updates, defined by

$$(s_j, y_j), \quad j = k-m, \ldots, k-1,$$

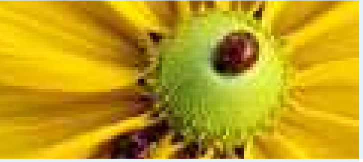
where

$$s_j = x_{j+1} - x_j, \quad y_{j+1} = \nabla f(x_{j+1}) - \nabla f(x_j)$$

to an initial inverse Hessian approximation $H_0^{(k)}$.

There are two variants: with "scaling" ($H_0^{(k)} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} I$) and without scaling ($H_0^{(k)} = I$).

The convergence rate of limited memory BFGS is linear, not superlinear, on smooth problems.



Limited Memory BFGS

"Full" BFGS requires storing an $n \times n$ matrix and doing matrix-vector multiplies, which is not possible when n is large.

In the 1980s, J. Nocedal and others developed a "limited memory" version of BFGS, with $O(n)$ space and time requirements, which is very widely used for minimizing smooth functions in many variables. At the kth iteration, it applies only the most recent m rank-two updates, defined by

$$(s_j, y_j), \quad j = k-m, \ldots, k-1,$$

where

$$s_j = x_{j+1} - x_j, \quad y_{j+1} = \nabla f(x_{j+1}) - \nabla f(x_j)$$

to an initial inverse Hessian approximation $H_0^{(k)}$.

There are two variants: with "scaling" ($H_0^{(k)} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} I$) and without scaling ($H_0^{(k)} = I$).

The convergence rate of limited memory BFGS is linear, not superlinear, on smooth problems.

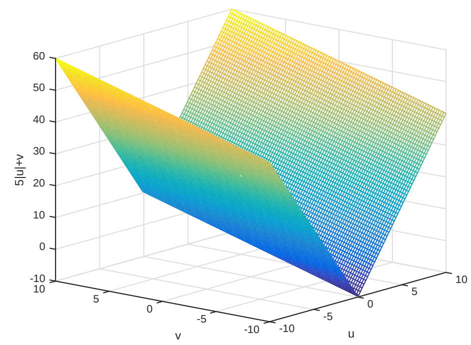
Question: how effective is it on nonsmooth problems?

A Nonsmooth Convex Function, Unbounded Below

Let's reconsider

$$f(x) = a|x_1| + x_2$$

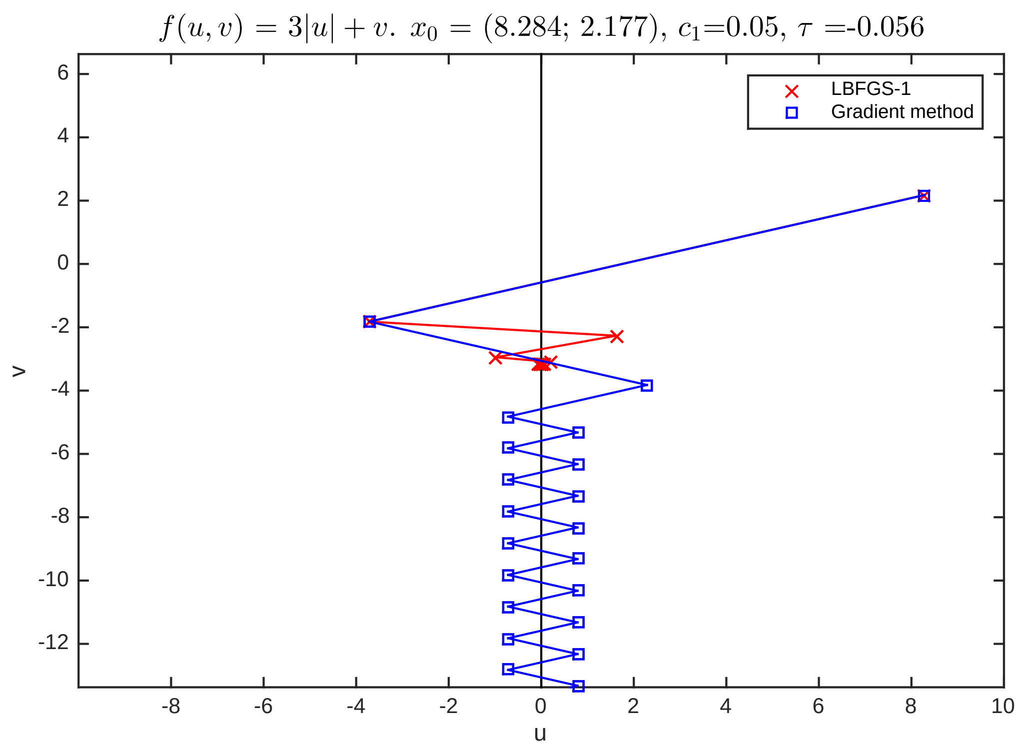
with $a \geq 1$.



Turns out that L-BFGS-1 (saving just one update) with scaling fails for *smaller* values of a than the critical value beyond which Gradient Descent fails!

L-BFGS-1 vs. Gradient Descent

Red: path of L-BFGS-1 with scaling, converges to non-stationary point.
Blue: path of the gradient method with same Armijo-Wolfe line search, generates $f(x) \downarrow -\infty$.



Convergence of the L-BFGS-1 Search Direction

Theorem. Let $d^{(k)}$ be the search direction generated by L-BFGS-1 with scaling applied to $f(x) = a|x_1| + \sum_{i=2}^n x_i$ using an Armijo-Wolfe line search. If $\sqrt{4(n-1)} \leq a$, then $\frac{|d^{(k)}|}{||d^{(k)}||}$ converges to some constant direction d. Furthermore, if

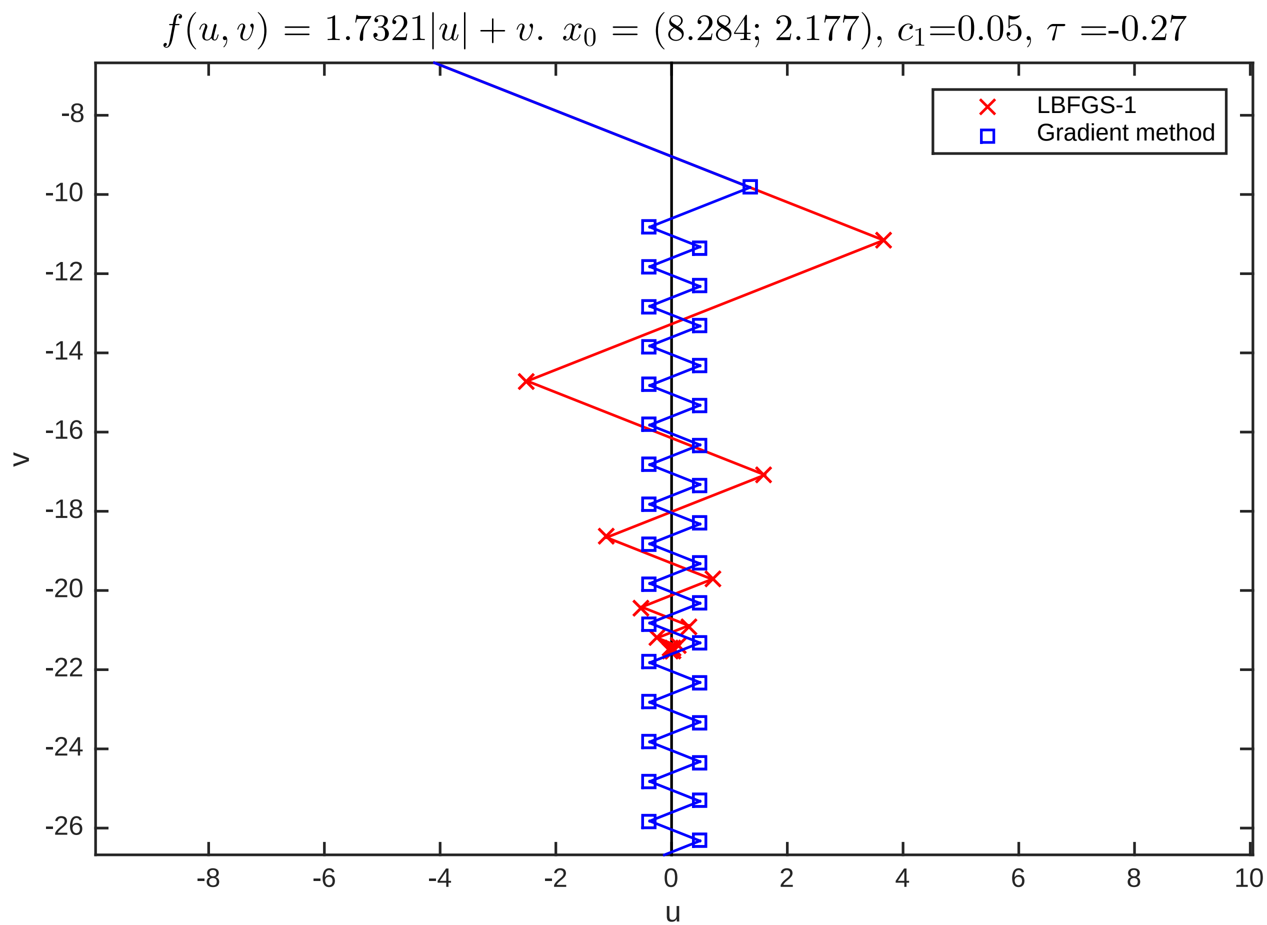
$$a(a + \sqrt{a^2 - 3(n-1)}) > (\frac{1}{c_1} - 1)(n-1),$$

where c_1 is the Armijo parameter, then the iterates $x^{(k)}$ converge to a non-stationary point.

Azam Asl, 2018.

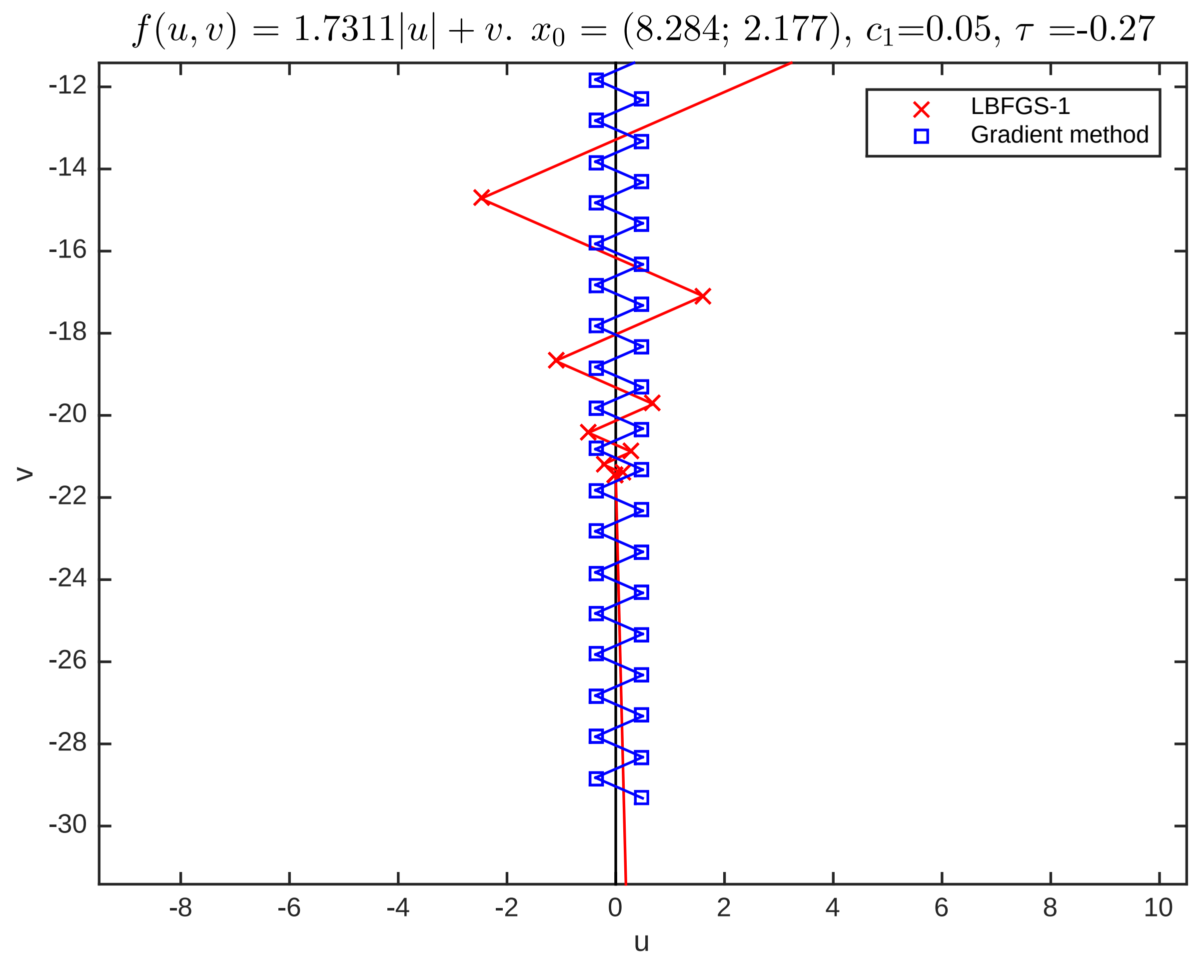
Experiment, with $n = 2$ and $a = \sqrt{3}$

In practice we observe that $\sqrt{3(n-1)} \leq a$ suffices for the method to fail, which is a weaker condition than the previous one. Below with $n = 2$ and $a = \sqrt{3}$ the method fails:



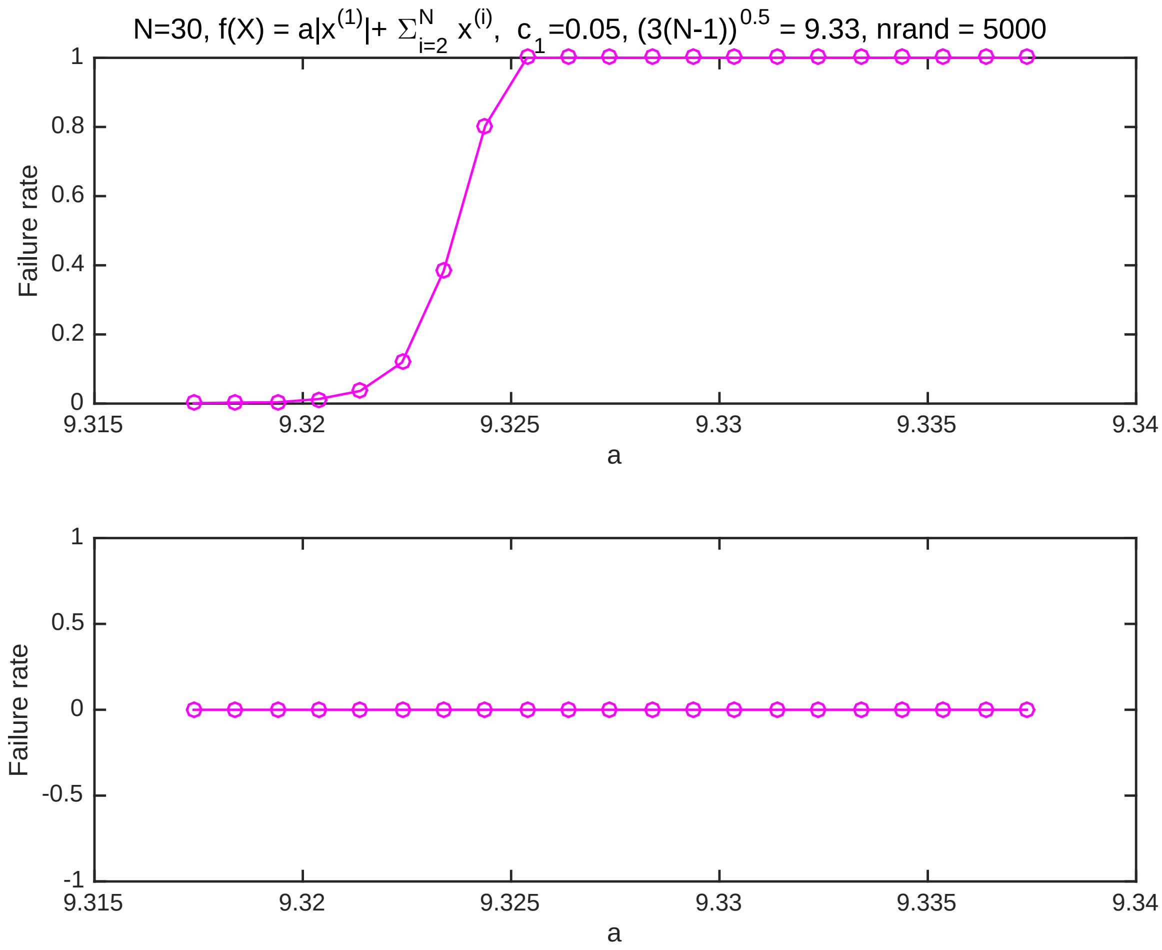
Experiment: slightly smaller a

But if we set $a = \sqrt{3} - 0.001$, it succeeds "at the last minute".



Experiments: Top, scaling on; Bottom, scaling off

$$n = 30, \ \sqrt{3(n-1)} = 9.327$$



Nesterov's Ill-Conditioned Nonsmooth Convex Function

Another Challenging Problem from Nesterov

Consider the nonsmooth, convex function

$$f(x) = \max\left(|x_1|, \quad \{|x_i - 2x_{i-1}|, i = 2, \ldots, n\}\right)$$

and set

$$\tilde{x}_1 = 1, \quad \tilde{x}_i = 2\tilde{x}_{i-1} + 1, i = 2, \ldots, n.$$

Then $f(\tilde{x}) = 1$ although $\|\tilde{x}\| \approx 2^n$.

Another Challenging Problem from Nesterov

Consider the nonsmooth, convex function

$$f(x) = \max\Big(|x_1|, \quad \{|x_i - 2x_{i-1}|, i = 2, \ldots, n\}\Big)$$

and set

$$\tilde{x}_1 = 1, \quad \tilde{x}_i = 2\tilde{x}_{i-1} + 1, i = 2, \ldots, n.$$

Then $f(\tilde{x}) = 1$ although $\|\tilde{x}\| \approx 2^n$.

Also $f([1, \ldots, 1]) = 1$, so level sets of f are very distorted.

Another Challenging Problem from Nesterov

Consider the nonsmooth, convex function

$$f(x) = \max\left(|x_1|, \quad \{|x_i - 2x_{i-1}|, i = 2, \ldots, n\}\right)$$

and set

$$\tilde{x}_1 = 1, \quad \tilde{x}_i = 2\tilde{x}_{i-1} + 1, i = 2, \ldots, n.$$

Then $f(\tilde{x}) = 1$ although $\|\tilde{x}\| \approx 2^n$.

Also $f([1, \ldots, 1]) = 1$, so level sets of f are very distorted.

The minimizer is 0 with $f(0) = 0$.

Another Challenging Problem from Nesterov

Consider the nonsmooth, convex function

$$f(x) = \max\Big(|x_1|, \quad \{|x_i - 2x_{i-1}|, i = 2, \ldots, n\}\Big)$$

and set

$$\tilde{x}_1 = 1, \quad \tilde{x}_i = 2\tilde{x}_{i-1} + 1, i = 2, \ldots, n.$$

Then $f(\tilde{x}) = 1$ although $\|\tilde{x}\| \approx 2^n$.

Also $f([1, \ldots, 1]) = 1$, so level sets of f are very distorted.

The minimizer is 0 with $f(0) = 0$.

Smoothed variant:

$$f_\mu(x) = \mathrm{softmax}\Big(\pm x_1, \pm\{x_i - 2x_{i-1}, i = 2, ..., n\}\Big)$$

with minimizer 0 with $f_\mu(0) = \mu \log(2n)$, where

$$\mathrm{softmax}(y) = \mu \log\left(\exp\left(\frac{y_1}{\mu}\right) + \ldots + \exp\left(\frac{y_{2n}}{\mu}\right)\right)$$

Another Challenging Problem from Nesterov

Consider the nonsmooth, convex function

$$f(x) = \max\Big(|x_1|, \quad \big\{|x_i - 2x_{i-1}|, i = 2, \ldots, n\big\}\Big)$$

and set

$$\tilde{x}_1 = 1, \quad \tilde{x}_i = 2\tilde{x}_{i-1} + 1, i = 2, \ldots, n.$$

Then $f(\tilde{x}) = 1$ although $\|\tilde{x}\| \approx 2^n$.

Also $f([1, \ldots, 1]) = 1$, so level sets of f are very distorted.

The minimizer is 0 with $f(0) = 0$.

Smoothed variant:

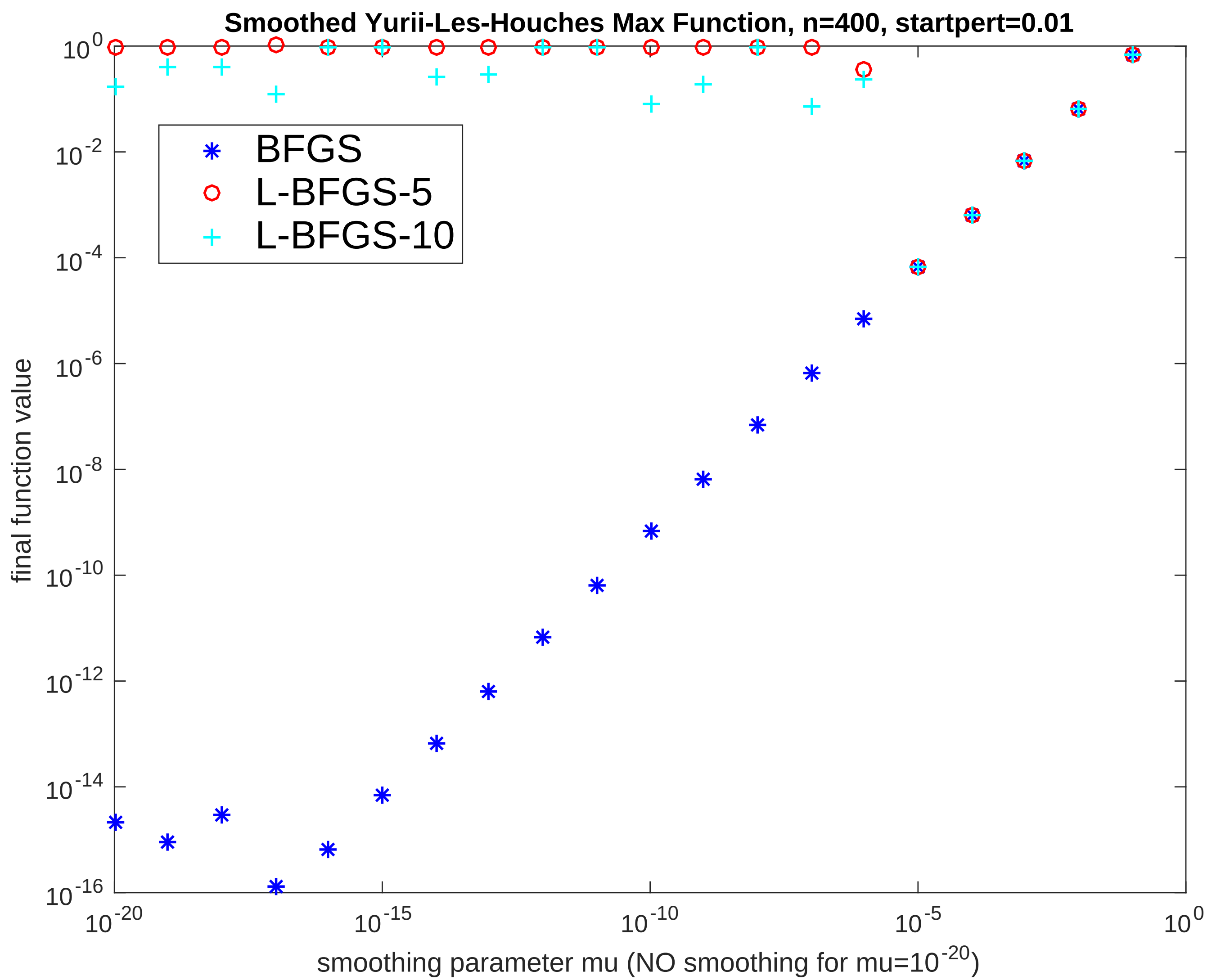
$$f_\mu(x) = \mathrm{softmax}\Big(\pm x_1, \pm\big\{x_i - 2x_{i-1}, i = 2, ..., n\big\}\Big)$$

with minimizer 0 with $f_\mu(0) = \mu \log(2n)$, where

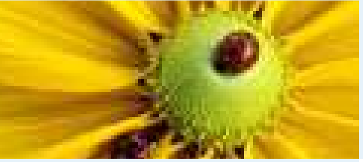
$$\mathrm{softmax}(y) = \mu \log\left(\exp\left(\frac{y_1}{\mu}\right) + \ldots + \exp\left(\frac{y_{2n}}{\mu}\right)\right)$$

As smoothing parameter $\mu \to 0$, function $f_\mu \to f$.

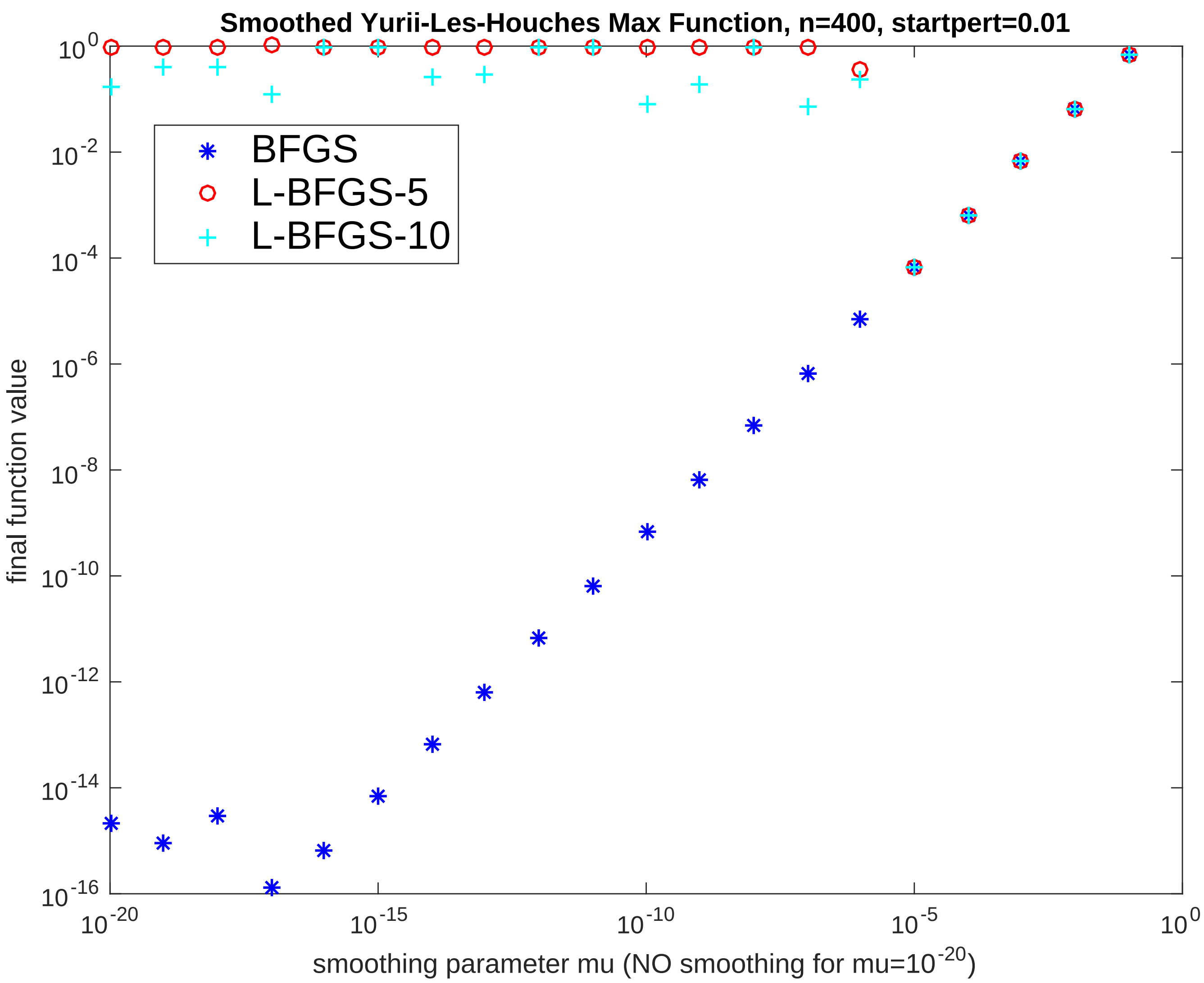
BFGS and Limited Memory BFGS for $n = 400$



BFGS finds accurate solution independent of conditioning as $\mu \to 0$



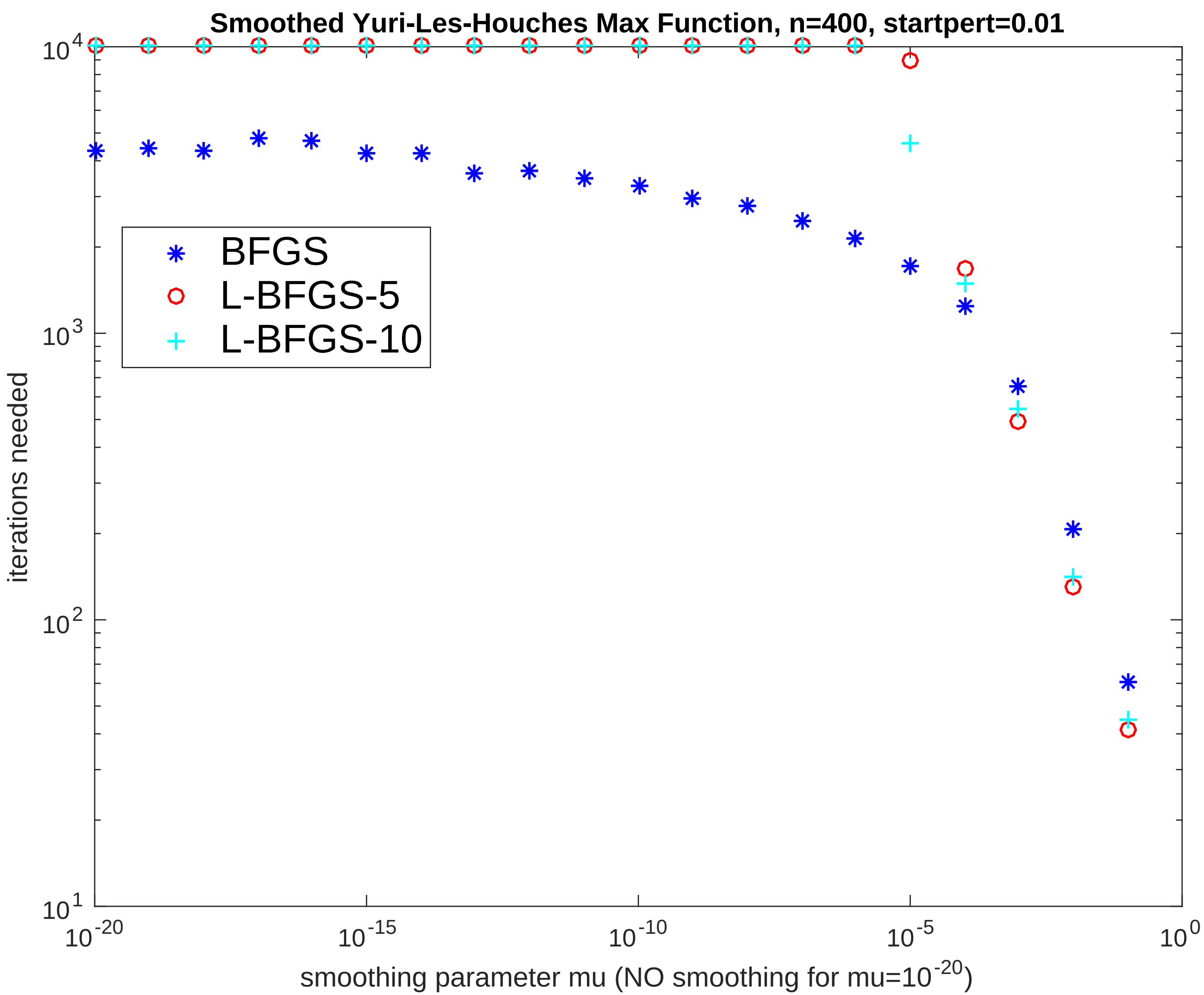
BFGS and Limited Memory BFGS for $n = 400$



BFGS finds accurate solution independent of conditioning as $\mu \to 0$

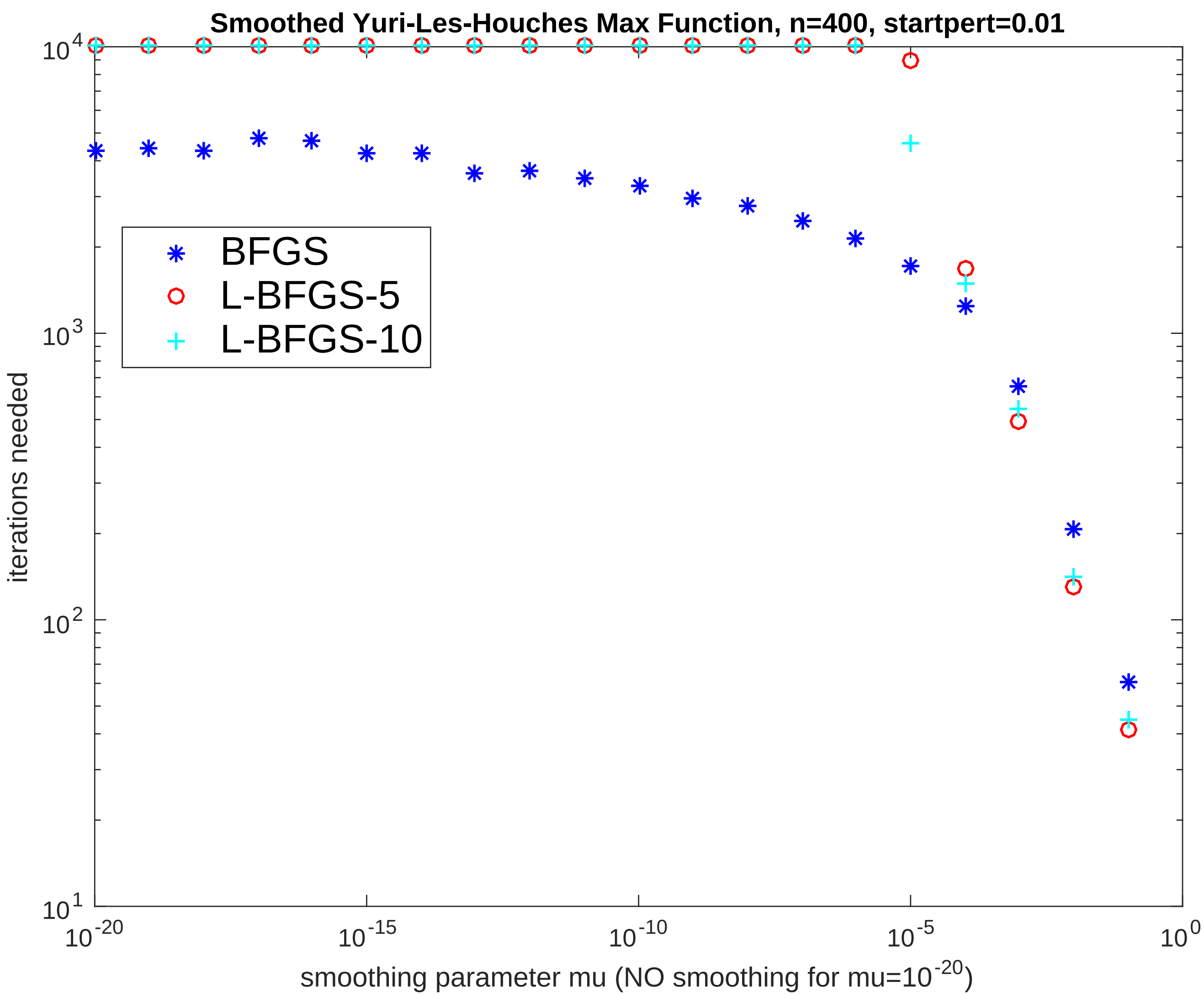
L-BFGS breaks down as $\mu \to 0$

Number of Iterations for $n = 400$



BFGS iterations is almost independent of the conditioning as $\mu \to 0$

Number of Iterations for $n = 400$



BFGS iterations is almost independent of the conditioning as $\mu \to 0$

L-BFGS iterations hits max iteration limit for small μ

Limited Effectiveness of Limited Memory BFGS

We have observed that that addition of nonsmoothness to a problem, convex or nonconvex, creates great difficulties for Limited Memory BFGS, with and without scaling, even when the dimension of the V-space is less than the size of the memory.

Limited Effectiveness of Limited Memory BFGS

We have observed that that addition of nonsmoothness to a problem, convex or nonconvex, creates great difficulties for Limited Memory BFGS, with and without scaling, even when the dimension of the V-space is less than the size of the memory.

Azam Asl's result establishes failure of L-BFGS-1 for a specific f when scaling is on; no such result is proved yet when scaling is off.

Limited Effectiveness of Limited Memory BFGS

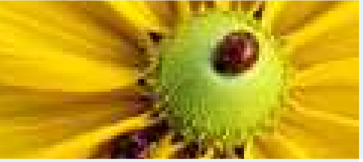
We have observed that that addition of nonsmoothness to a problem, convex or nonconvex, creates great difficulties for Limited Memory BFGS, with and without scaling, even when the dimension of the V-space is less than the size of the memory.

Azam Asl's result establishes failure of L-BFGS-1 for a specific f when scaling is on; no such result is proved yet when scaling is off.

We have also investigated Limited Memory Gradient Sampling which does not work well either.

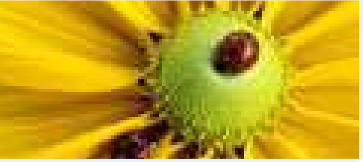
Other Ideas for Large Scale Nonsmooth Optimization

- Exploit structure! Lots of work on this has been done, e.g. using proximal point methods or ADMM (Alternating Direction Method of Multipliers)



Other Ideas for Large Scale Nonsmooth Optimization

- Exploit structure! Lots of work on this has been done, e.g. using proximal point methods or ADMM (Alternating Direction Method of Multipliers)
- Smoothing! Lots of work on this has been done too, most notably by Yu. Nesterov. Azam Asl has successfully applied L-BFGS to nonsmooth problems using Nesterov smoothing in her Ph.D. thesis, 2020.



Other Ideas for Large Scale Nonsmooth Optimization

- Exploit structure! Lots of work on this has been done, e.g. using proximal point methods or ADMM (Alternating Direction Method of Multipliers)
- Smoothing! Lots of work on this has been done too, most notably by Yu. Nesterov. Azam Asl has successfully applied L-BFGS to nonsmooth problems using Nesterov smoothing in her Ph.D. thesis, 2020.
- Automatic Differentiation (AD): (A. Griewank et. al.)

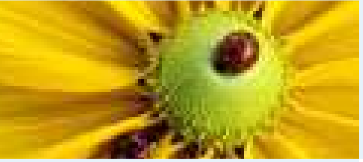
Other Ideas for Large Scale Nonsmooth Optimization

- Exploit structure! Lots of work on this has been done, e.g. using proximal point methods or ADMM (Alternating Direction Method of Multipliers)
- Smoothing! Lots of work on this has been done too, most notably by Yu. Nesterov. Azam Asl has successfully applied L-BFGS to nonsmooth problems using Nesterov smoothing in her Ph.D. thesis, 2020.
- Automatic Differentiation (AD): (A. Griewank et. al.)
- Stochastic Subgradient Method (D. Davis and D. Drusvyatskiy, 2018)

Concluding Remarks

Summary

Gradient descent frequently fails on nonsmooth problems.



Summary

Gradient descent frequently fails on nonsmooth problems.

Gradient Sampling is a simple method for nonsmooth, nonconvex optimization for which a convergence theory is known, but it is too expensive to use in most applications.

Summary

Gradient descent frequently fails on nonsmooth problems.

Gradient Sampling is a simple method for nonsmooth, nonconvex optimization for which a convergence theory is known, but it is too expensive to use in most applications.

BFGS — the full version — is remarkably effective on nonsmooth problems, but little theory is known.

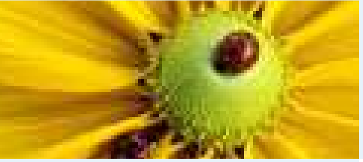
Summary

Gradient descent frequently fails on nonsmooth problems.

Gradient Sampling is a simple method for nonsmooth, nonconvex optimization for which a convergence theory is known, but it is too expensive to use in most applications.

BFGS — the full version — is remarkably effective on nonsmooth problems, but little theory is known.

Limited Memory BFGS is not so effective on nonsmooth problems.



Summary

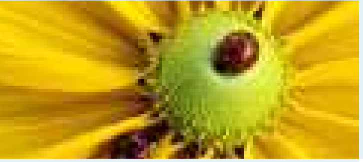
Gradient descent frequently fails on nonsmooth problems.

Gradient Sampling is a simple method for nonsmooth, nonconvex optimization for which a convergence theory is known, but it is too expensive to use in most applications.

BFGS — the full version — is remarkably effective on nonsmooth problems, but little theory is known.

Limited Memory BFGS is not so effective on nonsmooth problems.

Diabolical nonconvex problems such as Nesterov's Chebyshev-Rosenbrock problems can be very difficult, especially in the nonsmooth case.



Summary

Gradient descent frequently fails on nonsmooth problems.

Gradient Sampling is a simple method for nonsmooth, nonconvex optimization for which a convergence theory is known, but it is too expensive to use in most applications.

BFGS — the full version — is remarkably effective on nonsmooth problems, but little theory is known.

Limited Memory BFGS is not so effective on nonsmooth problems.

Diabolical nonconvex problems such as Nesterov's Chebyshev-Rosenbrock problems can be very difficult, especially in the nonsmooth case.

Our software, HANSO and GRANSO, is available (unconstrained and constrained) along with HIFOO (H-infinity fixed order optimization) for controller design, which has been used successfully in many applications.

Our Papers

J. V. Burke, A. S. Lewis and M. L. Overton, A Robust Gradient Sampling Method for Nonsmooth, Nonconvex Optimization, *SIAM J. Optimization*, 2005

M. Gürbüzbalaban and M. L. Overton, On Nesterov's Nonsmooth Chebyshev-Rosenbrock Functions, *SIAM J. Optimization*, 2012

A. S. Lewis and M. L. Overton, Nonsmooth Optimization via Quasi-Newton Methods, *Math. Programming*, 2013

F. E. Curtis, T. Mitchell and M.L. Overton, A BFGS-SQP Method for Nonsmooth, Nonconvex, Constrained Optimization and its Evaluation using Relative Minimization Profiles, *Optimization Methods and Software*, 2016

A. Asl and M. L. Overton, Analysis of the Gradient Method with an Armijo-Wolfe Line Search on a Class of Nonsmooth Convex Functions, *Optimization Methods and Software*, 2020

A. Asl and M. L. Overton, Analysis of Limited-Memory BFGS on a Class of Nonsmooth Convex Functions IMA J. Numer. Anal., 2020

J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton and L. Simões, Gradient Sampling Methods for Nonsmooth Optimization, in: Numerical Nonsmooth Optimization: State of the Art Algorithms (Springer, 2020).

Papers, software are available at **www.cs.nyu.edu/overton**.