Smoothness in nonsmooth optimization

(Newtonian ideas for partly smooth equations)

Adrian Lewis

Joint work with: D. Drusvyatskiy, X.Y. Han, A. Ioffe,
J. Liang, M.L. Overton, T. Tian, C. Wylie

September 2020

ORIE Cornell One World Optimization Seminar

Introduction: three questions

Question 1: Why is smoothness often inherent in nonsmooth optimization and equations?

Introduction: three questions

Question 1: Why is smoothness often inherent in nonsmooth optimization and equations?

Question 2: Can Newton methods use this smoothness?

Introduction: three questions

Question 1: Why is smoothness often inherent in nonsmooth optimization and equations?

Question 2: Can Newton methods use this smoothness?

Example: $\min_Q f$ becomes

$$-\nabla f(x) \in N_Q(x).$$

Projected gradient methods

$$x \leftarrow \operatorname{Proj}_Q\big(x - \gamma \nabla f(x)\big)$$

identify smoothness in Q.

Newtonian acceleration?

Introduction: three questions

Question 1: Why is smoothness often inherent in nonsmooth optimization and equations?

Question 2: Can Newton methods use this smoothness?

Example: $\min_Q f$ becomes

$$-\nabla f(x) \in N_Q(x).$$

Projected gradient methods

$$x \leftarrow \mathrm{Proj}_Q \big(x - \gamma \nabla f(x)\big)$$

identify smoothness in Q.

Newtonian acceleration?

Question 3: Superlinear convergence for black box nonsmooth optimization?

Introduction: three questions

Question 1: Why is smoothness often inherent in nonsmooth optimization and equations?

Question 2: Can Newton methods use this smoothness?

Example: $\min_Q f$ becomes

$$-\nabla f(x) \in N_Q(x).$$

Projected gradient methods

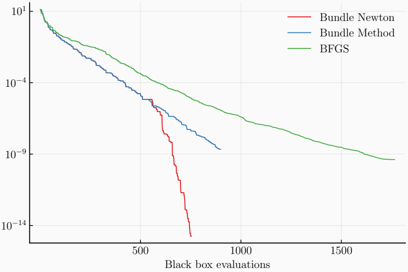
$$x \leftarrow \mathrm{Proj}_Q\left(x - \gamma \nabla f(x)\right)$$

identify smoothness in Q.

Newtonian acceleration?

Question 3: Superlinear convergence for black box nonsmooth optimization?

Example: Eigenvalue optimization.



Inherent structure: an example

Inherent structure: an example

The numerical radius of an n-by-n complex matrix A,

$$\rho(A) \;=\; \max_{\|u\|=1} |u^* A u|,$$

satisfies the "power inequality" (Berger '65): for $k = 1, 2, \ldots$

$$\frac{1}{2}\|A^k\|_2 \;\le \rho(A^k) \;\le \big(\rho(A)\big)^k,$$

and so controls transient stability in dynamics $x \leftarrow Ax$.

Inherent structure: an example

The numerical radius of an n-by-n complex matrix A,

$$\rho(A) \;=\; \max_{\|u\|=1} |u^* A u|,$$

satisfies the "power inequality" (Berger '65): for $k = 1, 2, \ldots$

$$\frac{1}{2}\|A^k\|_2 \;\leq \rho(A^k) \;\leq \big(\rho(A)\big)^k,$$

and so controls transient stability in dynamics $x \leftarrow Ax$.

Optimizing ρ often results in unusual matrices. . .

Inherent structure: an example

The numerical radius of an n-by-n complex matrix A,

$$\rho(A) \;=\; \max_{\|u\|=1} |u^* A u|,$$

satisfies the "power inequality" (Berger '65): for $k = 1, 2, \ldots$

$$\frac{1}{2}\|A^k\|_2 \;\leq \rho(A^k) \;\leq \big(\rho(A)\big)^k,$$

and so controls transient stability in dynamics $x \leftarrow Ax$.

Optimizing ρ often results in unusual matrices...

Example: For random matrices Y, proximal matrices A minimizing

$$\rho(A) + \lambda\|A - Y\|^2$$

often have disc fields of values $\{u^* A u : \|u\| = 1\}$.

Random proximal points (via `cvx`) are often disk matrices

Random proximal points (via `cvx`) are often disk matrices

$1 - \frac{\text{inner radius}}{\rho(A)}$

(via `chebfun`)

algebraic deviation
from disk

distance to
singularity

distance to null
Jordan block

Random proximal points (via cvx) are often disk matrices

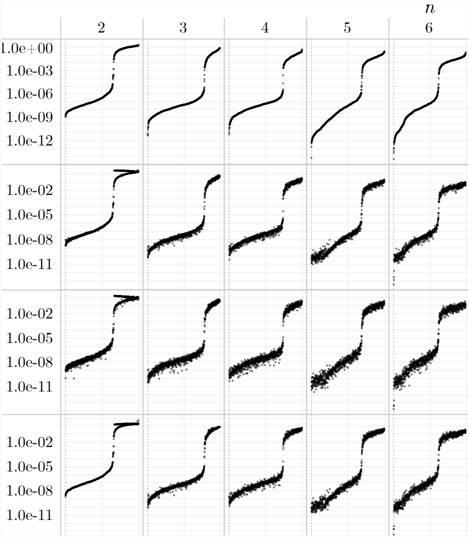
$$1 - \frac{\text{inner radius}}{\rho(A)}$$

(via chebfun)

algebraic deviation from disk

distance to singularity

distance to null Jordan block



Random proximal points (via cvx) are often disk matrices

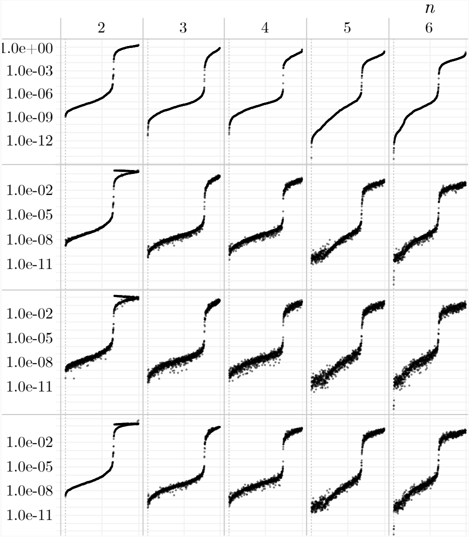
$1 - \dfrac{\text{inner radius}}{\rho(A)}$

(via chebfun)

algebraic deviation from disk

distance to singularity

distance to null Jordan block



Why. . . ? (Disk matrices comprise a small set, of codimension $2n$).

Nonsmooth optimization usually involves structured sets. Random instances are solved on manifolds of solutions under perturbation.

Partial smoothness and active sets

Nonsmooth optimization usually involves structured sets. Random instances are solved on manifolds of solutions under perturbation.

- $\{\text{smooth } g_i(x) \leq 0\}$ relative to active set $\{x : g_j(x) = 0\}$

Partial smoothness and active sets L '02

Nonsmooth optimization usually involves structured sets. Random instances are solved on manifolds of solutions under perturbation.

- $\{\text{smooth } g_i(x) \leq 0\}$ relative to active set $\{x : g_j(x) = 0\}$
- PSD matrices $\mathbf{S}^n_+$ relative to $\{X \in \mathbf{S}^n_+ : \text{rank}(X) = k\}$

Partial smoothness and active sets

Nonsmooth optimization usually involves structured sets. Random instances are solved on manifolds of solutions under perturbation.

- $\{\text{smooth } g_i(x) \leq 0\}$ relative to active set $\{x : g_j(x) = 0\}$
- PSD matrices $\mathbf{S}^n_+$ relative to $\{X \in \mathbf{S}^n_+ : \text{rank}(X) = k\}$
- smooth $f + ||\cdot||_1$ relative to fixed sparsity pattern

Partial smoothness and active sets

Nonsmooth optimization usually involves structured sets. Random instances are solved on manifolds of solutions under perturbation.

- $\{\text{smooth } g_i(x) \le 0\}$ relative to active set $\{x : g_j(x) = 0\}$
- PSD matrices $\mathbf{S}^n_+$ relative to $\{X \in \mathbf{S}^n_+ : \text{rank}(X) = k\}$
- smooth $f + ||\cdot||_1$ relative to fixed sparsity pattern
- numerical radius ρ relative to disk matrices (L-Overton '20, Han-L '20)

Partial smoothness and active sets L '02

Nonsmooth optimization usually involves structured sets. Random instances are solved on manifolds of solutions under perturbation.

- $\{\text{smooth } g_i(x) \leq 0\}$ relative to active set $\{x : g_j(x) = 0\}$
- PSD matrices $\mathbf{S}^n_+$ relative to $\{X \in \mathbf{S}^n_+ : \mathrm{rank}(X) = k\}$
- smooth $f + ||\cdot||_1$ relative to fixed sparsity pattern
- numerical radius ρ relative to disk matrices
  (L-Overton '20, Han-L '20)

History: generalized "active constraints" in nonlinear programming (Burke-Moré '88), "identifiable surfaces" (Wright '93), "$\mathcal{VU}$ decomposition" (Mifflin-Sagastizábal '00)...

Partial smoothness and algorithms

Partial smoothness and algorithms

Diverse first-order methods identify the manifold (L-Hare '04...), which drives the local convergence.

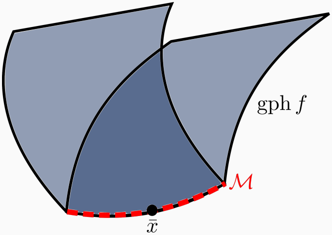
Partial smoothness and algorithms

Diverse first-order methods identify the manifold (L-Hare '04...), which drives the local convergence.

Example: l_1 regularization for sparsity. Proximal gradient iterates settle on a sparsity pattern (Hale-Yin-Zhang '08).

Partial smoothness and algorithms

Diverse first-order methods identify the manifold (L-Hare '04...), which drives the local convergence.



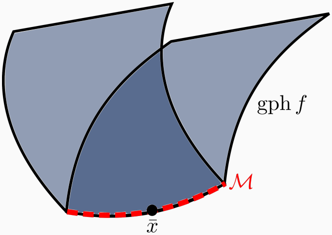
Example: l_1 regularization for sparsity. Proximal gradient iterates settle on a sparsity pattern (Hale-Yin-Zhang '08).

$$\min_x f(x) = h(x) + \lambda \|x\|_*$$

Partial smoothness and algorithms

Diverse first-order methods identify the manifold (L-Hare '04...), which drives the local convergence.

Example: l_1 regularization for sparsity. Proximal gradient iterates settle on a sparsity pattern (Hale-Yin-Zhang '08).



$$\min_x f(x) = h(x) + \lambda \|x\|_*$$

Example: Nuclear norm regularization for low-rank optimization. Proximal gradient (singular value thresholding) iterates settle on a fixed-rank manifold, then converge linearly to the solution.

(Liang-Fadili-Peyré '18)

Generalized equations

Generalized equations

Shift focus from optimization to optimality conditions:

$$x \text{ minimizes } f \quad \Rightarrow \quad 0 \in \partial f(x).$$

Generalized equations

Shift focus from optimization to optimality conditions:

$$x \text{ minimizes } f \quad \Rightarrow \quad 0 \in \partial f(x).$$

Generalize: $0 \in \Phi(u)$ for set-valued operator Φ on $\mathbf{R}^n$.

Generalized equations

Shift focus from optimization to optimality conditions:

$$x \text{ minimizes } f \quad \Rightarrow \quad 0 \in \partial f(x).$$

Generalize: $0 \in \Phi(u)$ for set-valued operator Φ on $\mathbf{R}^n$.

- Variational inequalities
  Find $x \in Q$ so $F(x)^T(z-x) \geq 0$ for all $z \in Q$: equivalently,

$$0 \in F(x) + N_Q(x).$$

Generalized equations

Shift focus from optimization to optimality conditions:

$$x \text{ minimizes } f \quad \Rightarrow \quad 0 \in \partial f(x).$$

Generalize: $0 \in \Phi(u)$ for set-valued operator Φ on $\mathbf{R}^n$.

- Variational inequalities
  Find $x \in Q$ so $F(x)^T(z - x) \geq 0$ for all $z \in Q$: equivalently,
  $$0 \in F(x) + N_Q(x).$$

- Composite optimization
  $\min_x h\big(c(x)\big)$ for convex h on $\mathbf{R}^m$ and smooth c into $\mathbf{R}^m$.
  Stationarity:
  $$0 \;\in\; \begin{pmatrix} \nabla c(x)^T y \\ -y \end{pmatrix} + \begin{pmatrix} 0 \\ \partial h(c(x)) \end{pmatrix}$$

Partly smooth generalized equations (L-Liang '18)

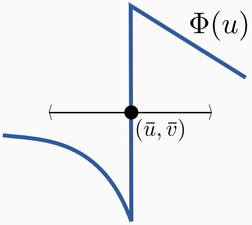
A set-valued operator Φ is partly smooth

at a solution $\bar{u}$ for given data $\bar{v}$

Partly smooth generalized equations (L-Liang '18)

A set-valued operator Φ is partly smooth at a solution $\bar{u}$ for given data $\bar{v}$ if

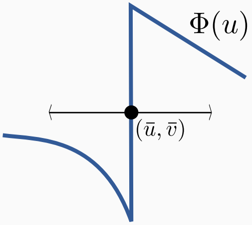
- $\operatorname{gph}\Phi = \{(u, v) : v \in \Phi(u)\}$ is a manifold around $(\bar{u}, \bar{v})$.
- $\operatorname{proj} : \operatorname{gph}\Phi \to \mathbf{R}^n : (u, v) \mapsto u$ is constant rank around $(\bar{u}, \bar{v})$...
  (i.e. the projection of the graph's tangent space has locally constant dimension).



Partly smooth generalized equations (L-Liang '18)

A set-valued operator Φ is partly smooth at a solution $\bar{u}$ for given data $\bar{v}$ if

- $\operatorname{gph}\Phi = \{(u, v) : v \in \Phi(u)\}$ is a manifold around $(\bar{u}, \bar{v})$.
- $\operatorname{proj} : \operatorname{gph}\Phi \to \mathbf{R}^n : (u, v) \mapsto u$ is constant rank around $(\bar{u}, \bar{v})$...
(i.e. the projection of the graph's tangent space has locally constant dimension).



Asymptotic solvers then identify the active manifold

$$\mathcal{M} \;=\; \operatorname{proj}\big(\operatorname{gph}\Phi \text{ around } (\bar{u}, \bar{v})\big),$$

since $v_k \in \Phi(u_k)$ with $(u_k, v_k) \to (\bar{u}, \bar{v})$ implies $u_k \in \mathcal{M}$ eventually.

Basic example: partly smooth sets

For closed convex (or "prox-regular") $S \subset \mathbf{R}^n$, suppose $\bar{x}$ solves

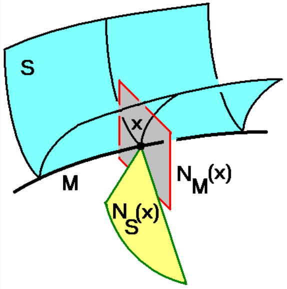
$$\min_{x \in S} \langle \bar{y}, x \rangle \qquad \text{and hence} \qquad \bar{y} \in N_S(x).$$

Basic example: partly smooth sets

For closed convex (or "prox-regular") $S \subset \mathbf{R}^n$, suppose $\bar{x}$ solves

$$\min_{x \in S} \langle \bar{y}, x \rangle \qquad \text{and hence} \qquad \bar{y} \in N_S(x).$$

If S contains a ridge manifold $\mathcal{M}$ (the normal cone $N_S(x)$ depends on $x \in \mathcal{M}$ continuously and spans $N_{\mathcal{M}}(x)$), and nondegeneracy holds $\big(\bar{y} \in \mathrm{ri}\big(N_S(\bar{x})\big)\big)$, then the operator N_S is partly smooth at $\bar{x}$ for $\bar{y}$, with active manifold $\mathcal{M}$.

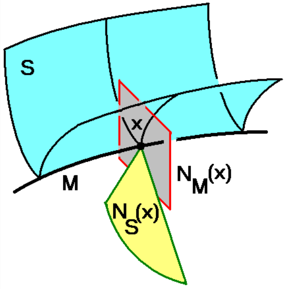


Basic example: partly smooth sets

For closed convex (or "prox-regular") $S \subset \mathbf{R}^n$, suppose $\bar{x}$ solves

$$\min_{x \in S} \langle \bar{y}, x \rangle \qquad \text{and hence} \qquad \bar{y} \in N_S(x).$$

If S contains a ridge manifold $\mathcal{M}$ (the normal cone $N_S(x)$ depends on $x \in \mathcal{M}$ continuously and spans $N_{\mathcal{M}}(x)$), and nondegeneracy holds $\Big(\bar{y} \in \mathrm{ri}\big(N_S(\bar{x})\big)\Big)$, then the operator N_S is partly smooth at $\bar{x}$ for $\bar{y}$, with active manifold $\mathcal{M}$.



So if S is convex and $\bar{x}$ is unique, projected gradient iterations $x \leftarrow \mathrm{proj}_Q(x - \alpha\bar{y})$ converge to $\bar{x}$ (if α small) and identify $\mathcal{M}$.

Example: max functions of degree k

Example: max functions of degree k

Given a decomposition

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

using smooth components f_i,

Example: max functions of degree k

Given a decomposition

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

using smooth components f_i, call $\bar{x}$ a strictly active critical point when the values $f_i(\bar{x})$ are all equal, and the system

$$\sum_i \lambda_i \;=\; 1, \qquad \sum_i \lambda_i \nabla f_i(\bar{x}) \;=\; 0$$

has a unique solution, which furthermore has each $\lambda_i > 0$.

Example: max functions of degree k

Given a decomposition

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

using smooth components f_i, call $\bar{x}$ a strictly active critical point when the values $f_i(\bar{x})$ are all equal, and the system

$$\sum_i \lambda_i \;=\; 1, \qquad \sum_i \lambda_i \nabla f_i(\bar{x}) \;=\; 0$$

has a unique solution, which furthermore has each $\lambda_i > 0$. Then

$$x \;\mapsto\; \partial f(x) \;=\; \operatorname{conv}\{\nabla f_i(x) : f_i(x) = f(x)\}$$

is partly smooth at $\bar{x}$ for 0 relative to the active manifold $\mathcal{M}$ of points where each f_i has equal value.

Sard-type behavior: partial smoothness is common

Sard-type behavior: partial smoothness is common

Sard's Theorem: almost no values of smooth operators are critical.

Sard-type behavior: partial smoothness is common

Sard's Theorem: almost no values of smooth operators are critical.

What about set-valued operators and generalized equations on $\mathbf{R}^n$?

Sard-type behavior: partial smoothness is common

Sard's Theorem: almost no values of smooth operators are critical.

What about set-valued operators and generalized equations on $\mathbf{R}^n$?

Consider a semi-algebraic operator Φ with n-dimensional graph:

$$\operatorname{gph}\Phi \;=\; \bigcup_{i=1}^{q}\bigcap_{j=1}^{r}\Big\{(x,y)\in\mathbf{R}^{2n} : p_{ij}(x,y)\;\left\{\begin{matrix}\le\\ \text{or}\\ <\end{matrix}\right\}\;0\Big\}$$

for polynomials p_{ij}.

Sard-type behavior: partial smoothness is common

Sard's Theorem: almost no values of smooth operators are critical.
What about set-valued operators and generalized equations on $\mathbf{R}^n$?

Consider a semi-algebraic operator Φ with n-dimensional graph:

$$\operatorname{gph}\Phi \;=\; \bigcup_{i=1}^{q}\bigcap_{j=1}^{r}\left\{(x,y)\in\mathbf{R}^{2n} : p_{ij}(x,y)\;\left\{\begin{array}{c}\leq\\ \text{or}\\ <\end{array}\right\}\;0\right\}$$

for polynomials p_{ij}. Eg: Subdifferentials, monotone operators...

Sard-type behavior: partial smoothness is common

Sard's Theorem: almost no values of smooth operators are critical.
What about set-valued operators and generalized equations on $\mathbf{R}^n$?

Consider a semi-algebraic operator Φ with n-dimensional graph:

$$\operatorname{gph}\Phi \;=\; \bigcup_{i=1}^{q}\bigcap_{j=1}^{r}\left\{(x,y)\in\mathbf{R}^{2n} : p_{ij}(x,y) \begin{Bmatrix} \leq \\ \text{or} \\ < \end{Bmatrix} 0\right\}$$

for polynomials p_{ij}. Eg: Subdifferentials, monotone operators...

Theorem Around generic data y, there are smooth maps G_i so

$$\Phi^{-1} \;=\; \{G_1,\ldots,G_k\} \qquad \text{(possibly empty).}$$

Sard-type behavior: partial smoothness is common

Sard's Theorem: almost no values of smooth operators are critical.
What about set-valued operators and generalized equations on $\mathbf{R}^n$?

Consider a semi-algebraic operator Φ with n-dimensional graph:

$$\operatorname{gph}\Phi \;=\; \bigcup_{i=1}^{q}\bigcap_{j=1}^{r}\left\{(x,y)\in\mathbf{R}^{2n} : p_{ij}(x,y)\,\left\{\begin{array}{c}\leq\\ \text{or}\\ <\end{array}\right\}\,0\right\}$$

for polynomials p_{ij}. Eg: Subdifferentials, monotone operators...

Theorem Around generic data y, there are smooth maps G_i so

$$\Phi^{-1} \;=\; \{G_1,\ldots,G_k\} \qquad \text{(possibly empty).}$$

Φ is partly smooth for y at each solution $x_i = G_i(y)$,

Sard-type behavior: partial smoothness is common

Sard's Theorem: almost no values of smooth operators are critical.
What about set-valued operators and generalized equations on $\mathbf{R}^n$?

Consider a semi-algebraic operator Φ with n-dimensional graph:

$$\mathrm{gph}\,\Phi \;=\; \bigcup_{i=1}^{q}\bigcap_{j=1}^{r}\left\{(x,y)\in\mathbf{R}^{2n} : p_{ij}(x,y)\,\left\{\begin{matrix}\leq\\ \text{or}\\ <\end{matrix}\right\}\,0\right\}$$

for polynomials p_{ij}. Eg: Subdifferentials, monotone operators...

Theorem Around generic data y, there are smooth maps G_i so

$$\Phi^{-1} \;=\; \{G_1,\ldots,G_k\} \qquad \text{(possibly empty).}$$

Φ is partly smooth for y at each solution $x_i = G_i(y)$, and

$$\mathrm{gph}\,\Phi \quad \text{intersects} \quad (\mathbf{R}^n\times\{y\}) \quad \text{transversally at } (x_i,y).$$

(Ioffe '07, Bolte...'11, Drusvyatskiy...'16, Lee...'19, L-Tian)

Semi-Newton iterations for $0 \in \Phi(u)$ (L-Wylie '20)

Recast as set intersection: find a point $z = (u, 0)$ where

$$X = \operatorname{gph} \Phi \qquad \text{intersects} \qquad Y = \mathbf{R}^n \times \{0\}.$$

Semi-Newton iterations for $0 \in \Phi(u)$ (L-Wylie '20)

Recast as set intersection: find a point $z = (u, 0)$ where

$$X = \operatorname{gph} \Phi \qquad \text{intersects} \qquad Y = \mathbf{R}^n \times \{0\}.$$

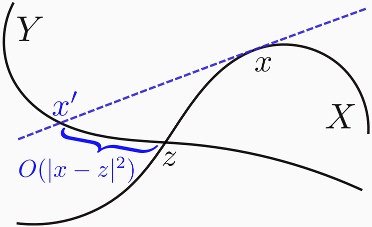
Assume transversality: $N_X(z) \cap N_Y(z) = \{0\}$.

Semi-Newton iterations for $0 \in \Phi(u)$ (L-Wylie '20)

Recast as set intersection: find a point $z = (u, 0)$ where

$$X = \operatorname{gph} \Phi \qquad \text{intersects} \qquad Y = \mathbf{R}^n \times \{0\}.$$

Assume transversality: $N_X(z) \cap N_Y(z) = \{0\}$.



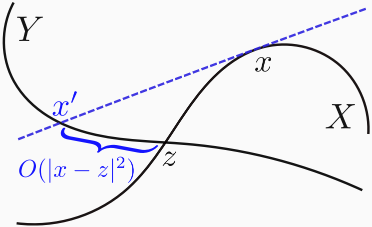
Step 1: Linearize X;
intersect with Y.

Semi-Newton iterations for $0 \in \Phi(u)$ (L-Wylie '20)

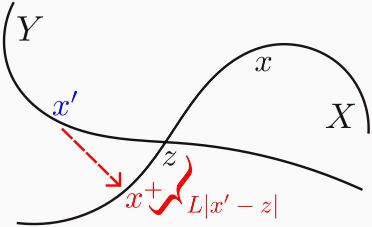
Recast as set intersection: find a point $z = (u, 0)$ where

$$X = \operatorname{gph} \Phi \qquad \text{intersects} \qquad Y = \mathbf{R}^n \times \{0\}.$$

Assume transversality: $N_X(z) \cap N_Y(z) = \{0\}$.



Step 1: Linearize X; intersect with Y.



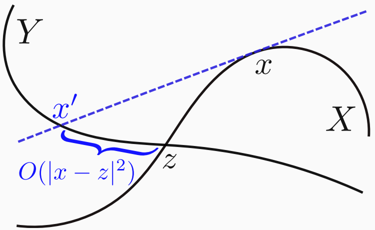
Step 2: Restore to X via a Lipschitz map fixing z.

Semi-Newton iterations for $0 \in \Phi(u)$ (L-Wylie '20)

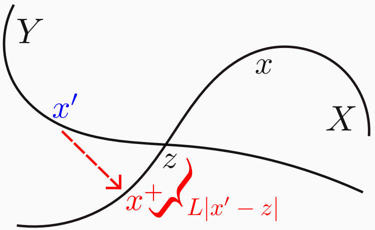
Recast as set intersection: find a point $z = (u, 0)$ where

$$X = \operatorname{gph} \Phi \qquad \text{intersects} \qquad Y = \mathbf{R}^n \times \{0\}.$$

Assume transversality: $N_X(z) \cap N_Y(z) = \{0\}$.



Step 1: Linearize X; intersect with Y.



Step 2: Restore to X via a Lipschitz map fixing z.

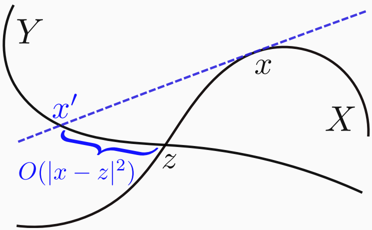
Linearize around $v \in \Phi(u)$; solve for u'

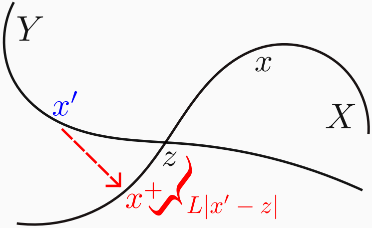
Semi-Newton iterations for $0 \in \Phi(u)$ (L-Wylie '20)

Recast as set intersection: find a point $z = (u, 0)$ where

$$X = \operatorname{gph} \Phi \qquad \text{intersects} \qquad Y = \mathbf{R}^n \times \{0\}.$$

Assume transversality: $N_X(z) \cap N_Y(z) = \{0\}$.





Step 1: Linearize X; intersect with Y.

Linearize around $v \in \Phi(u)$; solve for u'

Step 2: Restore to X via a Lipschitz map fixing z.

$u^+ = \mathrm{Proj}_{\mathcal{M}}(u')$; $\quad v^+ = \mathrm{Proj}_{\Phi(u^+)}(0)$

Fast black box nonsmooth optimization

Newtonian methods for partly smooth optimization $0 \in \partial f(x)$ are interesting, but typically need structural knowledge of ∂f.

Fast black box nonsmooth optimization

Newtonian methods for partly smooth optimization $0 \in \partial f(x)$ are interesting, but typically need structural knowledge of ∂f.

Classical special case: sequential quadratic programming.

Fast black box nonsmooth optimization

Newtonian methods for partly smooth optimization $0 \in \partial f(x)$ are interesting, but typically need structural knowledge of ∂f.

Classical special case: sequential quadratic programming.

More generally, semismooth Newton methods: Klatte-Kummer '02, Facchinei-Pang '03, Izmailov-Solodov '14, Gfrerer-Outrata '19.

Fast black box nonsmooth optimization

Newtonian methods for partly smooth optimization $0 \in \partial f(x)$ are interesting, but typically need structural knowledge of ∂f.

Classical special case: sequential quadratic programming.

More generally, semismooth Newton methods: Klatte-Kummer '02, Facchinei-Pang '03, Izmailov-Solodov '14, Gfrerer-Outrata '19.

With just an oracle for **linear** approximations to convex f at input points, bundle methods are appealing (Sagastizábal '18 ICM).

Fast black box nonsmooth optimization

Newtonian methods for partly smooth optimization $0 \in \partial f(x)$ are interesting, but typically need structural knowledge of ∂f.

Classical special case: sequential quadratic programming.

More generally, semismooth Newton methods: Klatte-Kummer '02, Facchinei-Pang '03, Izmailov-Solodov '14, Gfrerer-Outrata '19.

With just an oracle for **linear** approximations to convex f at input points, bundle methods are appealing (Sagastizábal '18 ICM).

- "Null" steps enhance a cutting plane model.
- "Serious" steps sufficiently decrease the objective
- Partial smoothness ("$\mathcal{V}U$") can accelerate the serious steps.

Fast black box nonsmooth optimization

Newtonian methods for partly smooth optimization $0 \in \partial f(x)$ are interesting, but typically need structural knowledge of ∂f.

Classical special case: sequential quadratic programming.

More generally, semismooth Newton methods: Klatte-Kummer '02, Facchinei-Pang '03, Izmailov-Solodov '14, Gfrerer-Outrata '19.

With just an oracle for **linear** approximations to convex f at input points, bundle methods are appealing (Sagastizábal '18 ICM).

- "Null" steps enhance a cutting plane model.
- "Serious" steps sufficiently decrease the objective
- Partial smoothness ("$\mathcal{V}U$") can accelerate the serious steps.

But can we reduce oracle calls using **quadratic** approximations?

Second-order oracles for nonsmooth optimization

Second-order oracles for nonsmooth optimization

Convex $f\colon \mathbf{R}^n \to \mathbf{R}$ are twice differentiable off a negligible set N.

Second-order oracles for nonsmooth optimization

Convex $f\colon \mathbf{R}^n \to \mathbf{R}$ are twice differentiable off a negligible set N.

Black-box methods (bundle, BFGS) typically never encounter N.

Second-order oracles for nonsmooth optimization

Convex $f : \mathbf{R}^n \to \mathbf{R}$ are twice differentiable off a negligible set N.

Black-box methods (bundle, BFGS) typically never encounter N.

What if an oracle returns $f(x), \nabla f(x), \nabla^2 f(x)$ for $x \notin N$?

Second-order oracles for nonsmooth optimization

Convex $f : \mathbf{R}^n \to \mathbf{R}$ are twice differentiable off a negligible set N.

Black-box methods (bundle, BFGS) typically never encounter N.
What if an oracle returns $f(x), \nabla f(x), \nabla^2 f(x)$ for $x \notin N$?

Aim: find a bundle S of k reference points, with small diameter

$$\max\{\|s - s'\| : s, s' \in S\}$$

and small optimality measure

$$\mathrm{dist}\Big(0, \mathrm{conv}\big(\nabla f(S)\big)\Big).$$

Second-order oracles for nonsmooth optimization

Convex $f : \mathbf{R}^n \to \mathbf{R}$ are twice differentiable off a negligible set N.

Black-box methods (bundle, BFGS) typically never encounter N.
What if an oracle returns $f(x), \nabla f(x), \nabla^2 f(x)$ for $x \notin N$?

Aim: find a bundle S of k reference points, with small diameter

$$\max\{\|s - s'\| : s, s' \in S\}$$

and small optimality measure

$$\operatorname{dist}\Big(0, \operatorname{conv}\big(\nabla f(S)\big)\Big).$$

Intuition: if the bundle size k is large enough,

$$\lim_{S \to \{x\}} \operatorname{conv}\big(\nabla f(S)\big) \;=\; \partial f(x).$$

A k-bundle Newton method (L-Wylie '19)

A k-bundle Newton method (L-Wylie '19)

For each of the k current reference points $s \in S$, use the oracle to form the linear and quadratic approximations

$$\begin{aligned} l_s(x) &= f(s) + \nabla f(s)^T (x - s) \\ q_s(x) &= l_s(x) + \frac{1}{2}(x - s)^T \nabla^2 f(s)(x - s). \end{aligned}$$

A k-bundle Newton method (L-Wylie '19)

For each of the k current reference points $s \in S$, use the oracle to form the linear and quadratic approximations

$$\begin{aligned} l_s(x) &= f(s) + \nabla f(s)^T (x - s) \\ q_s(x) &= l_s(x) + \frac{1}{2}(x - s)^T \nabla^2 f(s)(x - s). \end{aligned}$$

- Choose bundle weights λ_s solving

$$\min \Big\{ \Big\| \sum_{s \in S} \lambda_s \nabla f(s) \Big\| : \lambda \geq 0, \ \sum_s \lambda_s = 1 \Big\}.$$

A k-bundle Newton method (L-Wylie '19)

For each of the k current reference points $s \in S$, use the oracle to form the linear and quadratic approximations

$$\begin{aligned} l_s(x) &= f(s) + \nabla f(s)^T (x - s) \\ q_s(x) &= l_s(x) + \frac{1}{2}(x - s)^T \nabla^2 f(s)(x - s). \end{aligned}$$

- Choose bundle weights λ_s solving

$$\min\Big\{\Big\|\sum_{s \in S} \lambda_s \nabla f(s)\Big\| : \lambda \geq 0,\ \sum_s \lambda_s = 1\Big\}.$$

- Choose a new reference point x solving

$$\min\Big\{\sum_s \lambda_s q_s(x) : l_s(s) \text{ equal for all } s \in S\Big\}.$$

A k-bundle Newton method (L-Wylie '19)

For each of the k current reference points $s \in S$, use the oracle to form the linear and quadratic approximations

$$\begin{aligned} l_s(x) &= f(s) + \nabla f(s)^T (x - s) \\ q_s(x) &= l_s(x) + \frac{1}{2}(x - s)^T \nabla^2 f(s)(x - s). \end{aligned}$$

- Choose bundle weights λ_s solving

$$\min \Big\{ \Big\| \sum_{s \in S} \lambda_s \nabla f(s) \Big\| : \lambda \geq 0,\ \sum_s \lambda_s = 1 \Big\}.$$

- Choose a new reference point x solving

$$\min \Big\{ \sum_s \lambda_s q_s(x) : l_s(s) \text{ equal for all } s \in S \Big\}.$$

- Replace $s \in S$ minimizing $\|\nabla f(s) - \nabla f(x)\|$ with x.

Motivating the bundle Newton method

Given a current bundle of reference points $s \in S$, model

$$f(x) \;\approx\; \max_{s \in S} f_s(x):$$

unknown smooth component f_s matches f to 2nd order around s.

Motivating the bundle Newton method

Given a current bundle of reference points $s \in S$, model

$$f(x) \;\approx\; \max_{s \in S} f_s(x):$$

unknown smooth component f_s matches f to 2nd order around s.

Optimize model via sequential quadratic programming steps on

$$\begin{cases} \text{minimize} & t \\ \text{subject to} & f_s(x) \le t \qquad (s \in S). \end{cases}$$

Motivating the bundle Newton method

Given a current bundle of reference points $s \in S$, model

$$f(x) \;\approx\; \max_{s \in S} f_s(x):$$

unknown smooth component f_s matches f to 2nd order around s.

Optimize model via sequential quadratic programming steps on

$$\begin{cases} \text{minimize} & t \\ \text{subject to} & f_s(x) \le t \qquad (s \in S). \end{cases}$$

- Estimate multipliers via least squares.

Motivating the bundle Newton method

Given a current bundle of reference points $s \in S$, model

$$f(x) \;\approx\; \max_{s \in S} f_s(x):$$

unknown smooth component f_s matches f to 2nd order around s.

Optimize model via sequential quadratic programming steps on

$$\begin{cases} \text{minimize} & t \\ \text{subject to} & f_s(x) \le t \qquad (s \in S). \end{cases}$$

- Estimate multipliers via least squares.
- Minimize the quadratic approximation of the Lagrangian. . .

Motivating the bundle Newton method

Given a current bundle of reference points $s \in S$, model

$$f(x) \;\approx\; \max_{s \in S} f_s(x) :$$

unknown smooth component f_s matches f to 2nd order around s.

Optimize model via sequential quadratic programming steps on

$$\begin{cases} \text{minimize} & t \\ \text{subject to} & f_s(x) \leq t \qquad (s \in S). \end{cases}$$

- Estimate multipliers via least squares.
- Minimize the quadratic approximation of the Lagrangian...
- ...subject to the linearized constraints (assumed all active).

Motivating the bundle Newton method

Given a current bundle of reference points $s \in S$, model

$$f(x) \;\approx\; \max_{s \in S} f_s(x):$$

unknown smooth component f_s matches f to 2nd order around s.

Optimize model via sequential quadratic programming steps on

$$\begin{cases} \text{minimize} & t \\ \text{subject to} & f_s(x) \le t \qquad (s \in S). \end{cases}$$

- Estimate multipliers via least squares.
- Minimize the quadratic approximation of the Lagrangian...
- ...subject to the linearized constraints (assumed all active).

New point x improves model's most closely matching component.

Fast convergence on max functions

Theorem If f decomposes as max function of degree k,

$$f(x) \; = \; \max_{i=1,\ldots,k} f_i(x),$$

where each component f_i is smooth

Fast convergence on max functions

Theorem If f decomposes as max function of degree k,

$$f(x) \ = \ \max_{i=1,\dots,k} f_i(x),$$

where each component f_i is smooth and strongly convex

Fast convergence on max functions

Theorem If f decomposes as max function of degree k,

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

where each component f_i is smooth and strongly convex around a strictly active critical point $\bar{x}$,

Fast convergence on max functions

Theorem If f decomposes as max function of degree k,

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

where each component f_i is smooth and strongly convex around a strictly active critical point $\bar{x}$, and initial $S = \{s_1, \ldots, s_k\}$ is a full bundle near $\bar{x}$, meaning

$$f_i(s_i) > f_j(s_i) \quad \text{whenever } i \neq j,$$

Fast convergence on max functions

Theorem If f decomposes as max function of degree k,

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

where each component f_i is smooth and strongly convex
around a strictly active critical point $\bar{x}$, and
initial $S = \{s_1, \ldots, s_k\}$ is a full bundle near $\bar{x}$, meaning

$$f_i(s_i) > f_j(s_i) \quad \text{whenever } i \neq j,$$

then k-bundle Newton converges k-step quadratically to $\bar{x}$.

Fast convergence on max functions

Theorem If f decomposes as max function of degree k,

$$f(x) \; = \; \max_{i=1,\ldots,k} f_i(x),$$

where each component f_i is smooth and strongly convex around a strictly active critical point $\bar{x}$, and initial $S = \{s_1, \ldots, s_k\}$ is a full bundle near $\bar{x}$, meaning

$$f_i(s_i) > f_j(s_i) \quad \text{whenever } i \neq j,$$

then k-bundle Newton converges k-step quadratically to $\bar{x}$.

Note: The required bundle size k and the partly smooth geometry of the active manifold $\mathcal{M}$ are related:

$$k + \dim \mathcal{M} = n + 1.$$

Fast convergence on max functions

Theorem If f decomposes as max function of degree k,

$$f(x) = \max_{i=1,\ldots,k} f_i(x),$$

where each component f_i is smooth and strongly convex around a strictly active critical point $\bar{x}$, and initial $S = \{s_1, \ldots, s_k\}$ is a full bundle near $\bar{x}$, meaning

$$f_i(s_i) > f_j(s_i) \quad \text{whenever } i \neq j,$$

then k-bundle Newton converges k-step quadratically to $\bar{x}$.

Note: The required bundle size k and the partly smooth geometry of the active manifold $\mathcal{M}$ are related:

$$k + \dim \mathcal{M} = n + 1.$$

The classical Newton method has $k = 1$.

A simple max function example

$$f(x, y) = 2x^2 + y^2 + |x^2 - y|$$

A simple max function example

$f(x,y) = 2x^2 + y^2 + |x^2 - y|$

Objective value against
oracle calls.

A simple max function example

$f(x, y) = 2x^2 + y^2 + |x^2 - y|$

Objective value against
oracle calls.

- Naive proximal bundle $*$

A simple max function example

$f(x, y) = 2x^2 + y^2 + |x^2 - y|$

Objective value against
oracle calls.

- Naive proximal bundle $*$
- Nonsmooth BFGS $\circ$

A simple max function example

$f(x, y) = 2x^2 + y^2 + |x^2 - y|$

Objective value against
oracle calls.

- Naive proximal bundle $*$
- Nonsmooth BFGS $\circ$
- 2-bundle Newton

A simple max function example

$f(x, y) = 2x^2 + y^2 + |x^2 - y|$

Objective value against
oracle calls.

- Naive proximal bundle $*$
- Nonsmooth BFGS $\circ$
- 2-bundle Newton
  (initiated from
  proximal bundle)

A simple max function example

$f(x, y) = 2x^2 + y^2 + |x^2 - y|$

Objective value against oracle calls.

- Naive proximal bundle $*$
- Nonsmooth BFGS $\circ$
- 2-bundle Newton (initiated from proximal bundle)
  - objective **X**

A simple max function example

$f(x, y) = 2x^2 + y^2 + |x^2 - y|$

Objective value against oracle calls.

- Naive proximal bundle $*$
- Nonsmooth BFGS $\circ$
- 2-bundle Newton (initiated from proximal bundle)
  - objective **X**
  - optimality measure $\triangle$

A simple max function example

$f(x, y) = 2x^2 + y^2 + |x^2 - y|$

Objective value against oracle calls.

- Naive proximal bundle $*$
- Nonsmooth BFGS $\circ$
- 2-bundle Newton (initiated from proximal bundle)
  - objective **X**
  - optimality measure $\triangle$
  - bundle diameter $\diamond$

A simple max function example

$$f(x,y) = 2x^2 + y^2 + |x^2 - y|$$

Objective value against oracle calls.

- Naive proximal bundle $*$
- Nonsmooth BFGS $\circ$
- 2-bundle Newton (initiated from proximal bundle)
  - objective **X**
  - optimality measure $\triangle$
  - bundle diameter $\diamond$

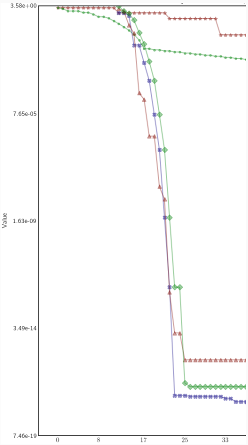
10^{-1}

10^{-5}

10^{-9}

10^{-13}

value



iterations 0 10 20 30

An eigenvalue optimization example

$$f(s,t,u,v) \;=\; \lambda_{\max}\begin{bmatrix} 0 & s & t \\ s & 1+u & v \\ t & v & 1-u \end{bmatrix}$$

An eigenvalue optimization example

$$f(s,t,u,v) \;=\; \lambda_{\max}\begin{bmatrix} 0 & s & t \\ s & 1{+}u & v \\ t & v & 1{-}u \end{bmatrix} - 1$$

Not a max function around its minimizer, zero.

An eigenvalue optimization example

$$f(s,t,u,v) \;=\; \lambda_{\max} \begin{bmatrix} 0 & s & t \\ s & 1+u & v \\ t & v & 1-u \end{bmatrix} - 1$$

Not a max function around its minimizer, zero. But...

...partly smooth relative to a 2-dimensional manifold.

An eigenvalue optimization example

$$f(s,t,u,v) \;=\; \lambda_{\max}\begin{bmatrix} 0 & s & t \\ s & 1+u & v \\ t & v & 1-u \end{bmatrix} - 1$$

Not a max function around its minimizer, zero. But...

...partly smooth relative to a 2-dimensional manifold.

- Proximal bundle $*$
- BFGS $\circ$
- 3-bundle Newton
  - objective **X**
  - optimality $\triangle$
  - diameter $\diamond$

An eigenvalue optimization example

$$f(s,t,u,v) \;=\; \lambda_{\max}\begin{bmatrix} 0 & s & t \\ s & 1+u & v \\ t & v & 1-u \end{bmatrix} - 1$$

Not a max function around its minimizer, zero. But…

…partly smooth relative to a 2-dimensional manifold.

- Proximal bundle $*$
- BFGS $\circ$
- 3-bundle Newton
  - objective **X**
  - optimality $\triangle$
  - diameter $\diamond$

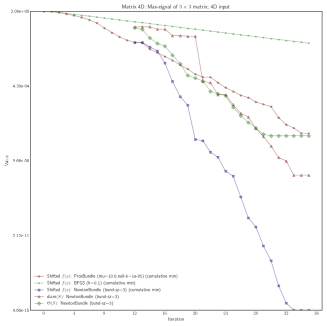
1

10^{-4}

10^{-8}

10^{-12}

value



iterations 0 10 20 30

The proof for max functions

Consider a maximum of smooth strongly convex components,

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

with a strictly active critical point $\bar{x}$. on the active manifold $\mathcal{M}$.

The proof for max functions

Consider a maximum of smooth strongly convex components,

$$f(x) \; = \; \max_{i=1,\dots,k} f_i(x),$$

with a strictly active critical point $\bar{x}$. on the active manifold $\mathcal{M}$.

Near x on the active manifold $\mathcal{M}$, full bundles $S = \{s_1, \dots, s_k\}$ (so $f_i(s_i) > f_j(s_i)$ for $i \neq j$) approximate the subdifferential:

$$\partial f(x) \; \approx \; \operatorname{conv} \nabla f(S),$$

The proof for max functions

Consider a maximum of smooth strongly convex components,

$$f(x) \ = \ \max_{i=1,\ldots,k} f_i(x),$$

with a strictly active critical point $\bar{x}$. on the active manifold $\mathcal{M}$.

Near x on the active manifold $\mathcal{M}$, full bundles $S = \{s_1, \ldots, s_k\}$ (so $f_i(s_i) > f_j(s_i)$ for $i \neq j$) approximate the subdifferential:

$$\partial f(x) \ \approx \ \operatorname{conv} \nabla f(S),$$

and the Hessians $\nabla^2 f(s_i)$ predict curvature on $\mathcal{M}$.

The proof for max functions

Consider a maximum of smooth strongly convex components,

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

with a strictly active critical point $\bar{x}$. on the active manifold $\mathcal{M}$.

Near x on the active manifold $\mathcal{M}$, full bundles $S = \{s_1, \ldots, s_k\}$ (so $f_i(s_i) > f_j(s_i)$ for $i \neq j$) approximate the subdifferential:

$$\partial f(x) \;\approx\; \operatorname{conv} \nabla f(S),$$

and the Hessians $\nabla^2 f(s_i)$ predict curvature on $\mathcal{M}$.

Partly smooth geometry then ensures $\hat{x} - \bar{x} = O(|\bar{x} - S|^2)$.

The proof for max functions

Consider a maximum of smooth strongly convex components,

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

with a strictly active critical point $\bar{x}$. on the active manifold $\mathcal{M}$.

Near x on the active manifold $\mathcal{M}$, full bundles $S = \{s_1, \ldots, s_k\}$ (so $f_i(s_i) > f_j(s_i)$ for $i \neq j$) approximate the subdifferential:

$$\partial f(x) \;\approx\; \operatorname{conv} \nabla f(S),$$

and the Hessians $\nabla^2 f(s_i)$ predict curvature on $\mathcal{M}$.

Partly smooth geometry then ensures $\hat{x} - \bar{x} = O(|\bar{x} - S|^2)$.

Updating $s_i \leftarrow \hat{x}$ keeps the bundle full, because $\bar{x}$ is strictly active.

The proof for max functions

Consider a maximum of smooth strongly convex components,

$$f(x) \;=\; \max_{i=1,\ldots,k} f_i(x),$$

with a strictly active critical point $\bar{x}$. on the active manifold $\mathcal{M}$.

Near x on the active manifold $\mathcal{M}$, full bundles $S = \{s_1, \ldots, s_k\}$ (so $f_i(s_i) > f_j(s_i)$ for $i \neq j$) approximate the subdifferential:

$$\partial f(x) \;\approx\; \operatorname{conv} \nabla f(S),$$

and the Hessians $\nabla^2 f(s_i)$ predict curvature on $\mathcal{M}$.

Partly smooth geometry then ensures $\hat{x} - \bar{x} = O(|\bar{x} - S|^2)$.

Updating $s_i \leftarrow \hat{x}$ keeps the bundle full, because $\bar{x}$ is strictly active.

Each reference point s_i updates within k steps, by strong convexity.

Finding an initial bundle

Black-box methods for finding a minimizer $\bar{x}$ for nonsmooth f, like

- bundle methods (Lemaréchal, Wolfe '70's)
- BFGS (L-Overton '13)
- gradient sampling (Burke-L-Overton '05)

Finding an initial bundle

Black-box methods for finding a minimizer $\bar{x}$ for nonsmooth f, like

- bundle methods (Lemaréchal, Wolfe '70's)
- BFGS (L-Overton '13)
- gradient sampling (Burke-L-Overton '05)

asymptotically generate subdifferential approximations:

$$\partial f(\bar{x}) \;\approx\; \mathrm{conv}\left(\nabla f(\Omega)\right)$$

for sets Ω of points near $\bar{x}$.

Finding an initial bundle

Black-box methods for finding a minimizer $\bar{x}$ for nonsmooth f, like

- bundle methods (Lemaréchal, Wolfe '70's)
- BFGS (L-Overton '13)
- gradient sampling (Burke-L-Overton '05)

asymptotically generate subdifferential approximations:

$$\partial f(\bar{x}) \;\approx\; \text{conv}\big(\nabla f(\Omega)\big)$$

for sets Ω of points near $\bar{x}$. So, we could choose

$$k \;=\; \dim\Big(\text{affine span}\big(\nabla f(\Omega)\big)\Big) \qquad \text{(numerically)}$$

Finding an initial bundle

Black-box methods for finding a minimizer $\bar{x}$ for nonsmooth f, like

- bundle methods (Lemaréchal, Wolfe '70's)
- BFGS (L-Overton '13)
- gradient sampling (Burke-L-Overton '05)

asymptotically generate subdifferential approximations:

$$\partial f(\bar{x}) \;\approx\; \mathrm{conv}\big(\nabla f(\Omega)\big)$$

for sets Ω of points near $\bar{x}$. So, we could choose

$$k \;=\; \dim\Big(\text{affine span}\big(\nabla f(\Omega)\big)\Big) \qquad \text{(numerically)}$$

and initial $S \subset \Omega$ of size k with $\nabla f(S)$ affinely independent.

Example: eigenvalue optimization

For random 25-by-25 symmetric matrices, minimize $\lambda_{\max}$ for

$$A(x) \ = \ A_0 + x_1 A_1 + x_2 A_2 + \ldots + x_{50} A_{50}.$$

Example: eigenvalue optimization

For random 25-by-25 symmetric matrices, minimize λ_{max} for

$$A(x) \; = \; A_0 + x_1 A_1 + x_2 A_2 + \ldots + x_{50} A_{50}.$$

Active manifold, where $\lambda_{\mathsf{max}}\big(A(x)\big)$ has multiplicity 6, has dim 30.

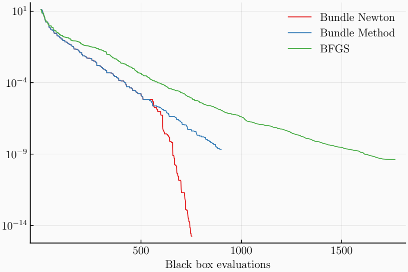
Example: eigenvalue optimization

For random 25-by-25 symmetric matrices, minimize $\lambda_{\max}$ for

$$A(x) \ = \ A_0 + x_1 A_1 + x_2 A_2 + \ldots + x_{50} A_{50}.$$

Active manifold, where $\lambda_{\max}\big(A(x)\big)$ has multiplicity 6, has dim 30.

$f - \min f$



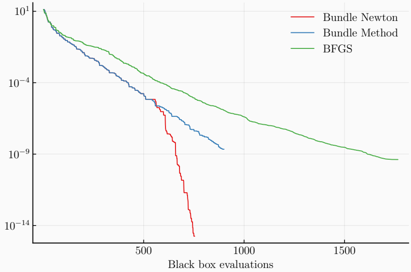
Example: eigenvalue optimization

For random 25-by-25 symmetric matrices, minimize $\lambda_{\max}$ for

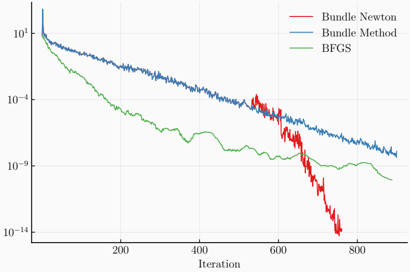
$$A(x) \;=\; A_0 + x_1 A_1 + x_2 A_2 + \ldots + x_{50} A_{50}.$$

Active manifold, where $\lambda_{\max}\big(A(x)\big)$ has multiplicity 6, has dim 30.

$f - \min f$



$\lambda_1(A(x)) - \lambda_6(A(x))$



Extensions. . .

Extensions...

- Avoiding Hessians...

Extensions. . .

- Avoiding Hessians. . .
    - . . . using automatic differentiation

Extensions. . .

- Avoiding Hessians. . .
    - . . . using automatic differentiation ✓

Extensions. . .

- Avoiding Hessians. . .
  - . . . using automatic differentiation ✓
  - . . . with a linearly convergent first-order analogue

Extensions. . .

- Avoiding Hessians. . .
    - . . . using automatic differentiation ✓
    - . . . with a linearly convergent first-order analogue ✓?

Extensions...

- Avoiding Hessians...
  - ...using automatic differentiation ✓
  - ...with a linearly convergent first-order analogue ✓?
- Extending the local convergence analysis...

Extensions. . .

- Avoiding Hessians. . .
    - . . . using automatic differentiation ✓
    - . . . with a linearly convergent first-order analogue ✓?
- Extending the local convergence analysis. . .
    - . . . to nonconvex max functions

Extensions. . .

- Avoiding Hessians. . .
    - . . . using automatic differentiation ✓
    - . . . with a linearly convergent first-order analogue ✓?
- Extending the local convergence analysis. . .
    - . . . to nonconvex max functions ✓

Extensions. . .

- Avoiding Hessians. . .
    - . . . using automatic differentiation ✓
    - . . . with a linearly convergent first-order analogue ✓?
- Extending the local convergence analysis. . .
    - . . . to nonconvex max functions ✓
    - . . . to partly smooth functions

Extensions...

- Avoiding Hessians...
  - ...using automatic differentiation ✓
  - ...with a linearly convergent first-order analogue ✓?
- Extending the local convergence analysis...
  - ...to nonconvex max functions ✓
  - ...to partly smooth functions ??

Extensions. . .

- Avoiding Hessians. . .
  - . . . using automatic differentiation ✓
  - . . . with a linearly convergent first-order analogue ✓?
- Extending the local convergence analysis. . .
  - . . . to nonconvex max functions ✓
  - . . . to partly smooth functions ??
- Globalizing the algorithm

Extensions. . .

- Avoiding Hessians. . .
  - . . . using automatic differentiation ✓
  - . . . with a linearly convergent first-order analogue ✓?
- Extending the local convergence analysis. . .
  - . . . to nonconvex max functions ✓
  - . . . to partly smooth functions ??
- Globalizing the algorithm ??

Take-away...

Partial smoothness is a simple differential-geometric idea that captures the generic interplay between smooth and nonsmooth geometry in concrete variational problems, illuminating the analysis and design of algorithms.

References

L and M.L. Overton, “Partial smoothness of the numerical radius at matrices whose fields of values are disks”, *SIMAX* 2020.

X.Y. Han and L, “Disk matrices and the proximal mapping for the numerical radius”, `arXiv:2004.14542`

L and J. Liang, “Partial smoothness and constant rank”, `arXiv:1807.03134`.

L and C.J.S. Wylie, “Active-set Newton methods and partial smoothness”, *MOR* 2020.

L and C.J.S. Wylie, “A simple Newton method for local nonsmooth optimization”, `arxiv:1907.11742`.