Avoiding saddle points in nonsmooth optimization

Damek Davis
School of Operations Research and Information Engineering
Cornell University

Joint with L. Jiang (Cornell) and D. Drusvyatskiy (U. Washington)

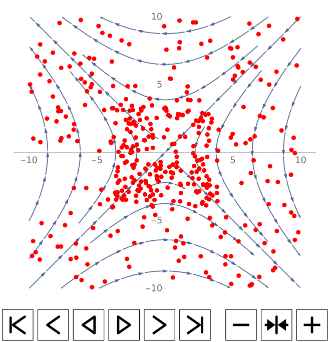
One World Optimization Seminar
Nov 2021

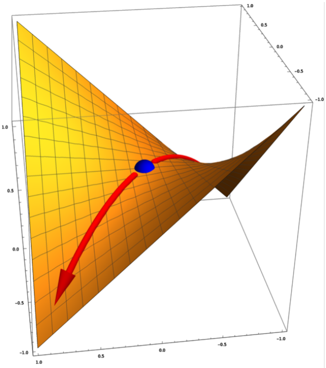
Saddle point avoidance

Recent Realization:

Simple algorithms for minimizing C^2 functions avoid all strict saddle points, when randomly initialized.[1]

- Simple algorithms: Gradient descent (GD), coordinate descent....
- *Strict* saddle points: Critical points that have negative curvature.





[1] Lee-Simchowitz-Jordan-Recht '16

Saddle point avoidance

Recent Realization:

Simple algorithms for minimizing C^2 functions avoid all strict saddle points, when randomly initialized.[2]

- Simple algorithms: Gradient descent (GD), coordinate descent....
- *Strict* saddle points: Critical points that have negative curvature.

Motivation:

For a wealth of estimation and learning problems, all spurious critical points are strict saddles and therefore avoidable!

(Sun-Qu-Wright '15-'18, Ge-Lee-Ma '16, Bhojanapalli-Neyshabur-Srebro '16, Ge-Jin-Zheng '17...)

[2] Lee-Simchowitz-Jordan-Recht '16

Saddle point avoidance

Recent Realization:

Simple algorithms for minimizing C^2 functions avoid all strict saddle points, when randomly initialized.[2]

- Simple algorithms: Gradient descent (GD), coordinate descent....
- *Strict* saddle points: Critical points that have negative curvature.

Motivation:

For a wealth of estimation and learning problems, all spurious critical points are strict saddles and therefore avoidable!

(Sun-Qu-Wright '15-'18, Ge-Lee-Ma '16, Bhojanapalli-Neyshabur-Srebro '16, Ge-Jin-Zheng '17. . .)

This talk:

Do first-order methods avoid "strict saddles" of nonsmooth functions?

[2] Lee-Simchowitz-Jordan-Recht '16

Weak convexity: an amenable problem class

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ F(x)$$

Running assumption: weak convexity

$$F(\cdot)+\frac{\rho}{2}\|\cdot\|^2 \qquad \text{is convex.}$$

Weak convexity: an amenable problem class

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; F(x)$$

Running assumption: weak convexity

$$F(\cdot) + \frac{\rho}{2}\|\cdot\|^2 \qquad \text{is convex.}$$

Main example: $\boxed{(\text{convex}) \circ (\text{smooth})}$

$$h(c(x))$$

h is convex and L-Lipschitz; c is smooth with ℓ-Lipschitz Jacobian ($\rho = L\ell$)

(Fletcher '80, Powell '83, Burke '85, Wright '90, Lewis-Wright '08, Cartis-Gould-Toint '11,...)

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

Examples: Matrix completion, robust PCA, phase retrieval...

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

Examples: Matrix completion, robust PCA, phase retrieval...

Natural Nonconvex Penalty Formulation:[3]

$$\min_{M \in \mathbb{R}^{d \times d}} \; |||\mathcal{A}(M) - b||| \qquad \text{subject to: } M \text{ is rank} \leq r$$

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

Examples: Matrix completion, robust PCA, phase retrieval...

Natural Nonconvex Penalty Formulation:[3]

$$M = XX^T \qquad X \in \mathbb{R}^{d \times r}$$

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

Examples: Matrix completion, robust PCA, phase retrieval...

Natural Nonconvex Penalty Formulation:[3]

$$\min_{X \in \mathbb{R}^{d \times r}} \; h(c(X)) := |||\mathcal{A}(XX^\top) - b|||$$

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

Examples: Matrix completion, robust PCA, phase retrieval...

Natural Nonconvex Penalty Formulation:[3]

$$\min_{X \in \mathbb{R}^{d \times r}} \; h(c(X)) := |||\mathcal{A}(XX^\top) - b|||$$

Question: Is there a natural norm $|||\cdot|||$ that enables recovery?

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

Examples: Matrix completion, robust PCA, phase retrieval...

Natural Nonconvex Penalty Formulation:[3]

$$\min_{X \in \mathbb{R}^{d \times r}} \; h(c(X)) := |||\mathcal{A}(XX^\top) - b|||$$

Question: Is there a natural norm $|||\cdot|||$ that enables recovery?

Typical norms[4]**:** $|||\cdot||| = \frac{1}{\sqrt{m}}\|\cdot\|_2$ and $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

Examples: Matrix completion, robust PCA, phase retrieval...

Natural Nonconvex Penalty Formulation:[3]

$$\min_{X \in \mathbb{R}^{d \times r}} \; h(c(X)) := |||\mathcal{A}(XX^\top) - b|||$$

Question: Is there a natural norm $|||\cdot|||$ that enables recovery?

Typical norms[4]**:** $|||\cdot||| = \frac{1}{\sqrt{m}}\|\cdot\|_2$ and $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$

- ℓ_2: Gaussian A_i/Gaussian noise, leads to smooth problems.

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

Example: Low-rank Matrix Estimation

Set-up: Fix rank r matrix $M_\sharp \succeq 0$ and observe measurements

$$\langle A_i, M_\sharp \rangle \approx b_i \qquad \forall i = 1, \ldots, m.$$

Goal: Recover $M_\sharp$ from b_i

Examples: Matrix completion, robust PCA, phase retrieval...

Natural Nonconvex Penalty Formulation:[3]

$$\min_{X \in \mathbb{R}^{d \times r}} \; h(c(X)) := |||\mathcal{A}(XX^\top) - b|||$$

Question: Is there a natural norm $|||\cdot|||$ that enables recovery?

Typical norms[4]**:** $|||\cdot||| = \frac{1}{\sqrt{m}}\|\cdot\|_2$ and $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$

- ℓ_2: Gaussian A_i/Gaussian noise, leads to smooth problems.
- ℓ_1: structured A_i/sparse corruption, leads to nonsmooth problems.

[3] Burer-Monteiro '01

[4] Candes-Tao '05, Chen-Chi-Goldsmith '13

First-order methods for nonsmooth problems

Common iterative methods take form

$$x_{t+1} = \arg\min_{y} F_{x_t}(y)$$

where $F_{x_t} =$ nonsmooth strongly convex model of F.

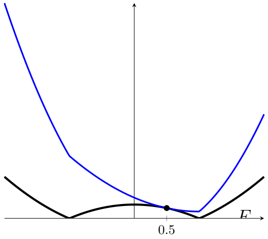
First-order methods for nonsmooth problems

Common iterative methods take form

$$x_{t+1} = \arg\min_{y} F_{x_t}(y)$$

where $F_{x_t} =$ nonsmooth strongly convex model of F.

Example: Proximal point



$$F_{x_t}(y) = F(y) + \frac{1}{2\eta}\|y - x_t\|^2$$

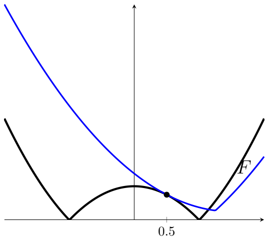
First-order methods for nonsmooth problems

Common iterative methods take form

$$x_{t+1} = \arg\min_y F_{x_t}(y)$$

where $F_{x_t} =$ nonsmooth strongly convex model of F.

Example: Proximal linear (for $F = h \circ c$)



$$F_{x_t}(y) = h(c(x_t) + \nabla c(x_t)(y - x_t)) + \frac{1}{2\eta}\|y - x_t\|^2$$

First-order methods for nonsmooth problems

Common iterative methods take form

$$x_{t+1} = \arg\min_{y} F_{x_t}(y)$$

where $F_{x_t} =$ nonsmooth strongly convex model of F.

Example:

Algorithm	Objective F	Update function $F_x(y)$
Prox-point	$F(x)$	$F(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-linear	$h(c(x)) + r(x)$	$h(c(x) + \nabla c(x)(y - x)) + r(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-gradient	$f(x) + r(x)$	$f(x) + \langle \nabla f(x), y - x\rangle + r(y) + \frac{1}{2\eta}\|y - x\|^2$

Table: h is convex and Lipschitz, r is weakly convex, and f and c are C^2-smooth.

Q: What is an avoidable saddle point in nonsmooth optimization?[5]

[5] (D-Drusvyatskiy '19)

Q: What is an avoidable saddle point in nonsmooth optimization?[5]

Recall C^2 case: A strict saddle is critical point with negative curvature:

$$\nabla F(x) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(x)) < 0$$

[5] (D-Drusvyatskiy '19)

Q: What is an avoidable saddle point in nonsmooth optimization?[5]

Recall C^2 case: A strict saddle is critical point with negative curvature:

$$\nabla F(x) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(x)) < 0$$

Generalization Attempt: A strict saddle is critical point such that

[5] (D-Drusvyatskiy '19)

Q: What is an avoidable saddle point in nonsmooth optimization?[5]

Recall C^2 case: A strict saddle is critical point with negative curvature:

$$\nabla F(x) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(x)) < 0$$

Generalization Attempt: A strict saddle is critical point such that

- There exists direction v s.t.

$$g(t) := F(x + tv) \text{ is } C^2.$$

[5](D-Drusvyatskiy '19)

Q: What is an avoidable saddle point in nonsmooth optimization?[5]

Recall C^2 case: A strict saddle is critical point with negative curvature:

$$\nabla F(x) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(x)) < 0$$

Generalization Attempt: A strict saddle is critical point such that

- There exists direction v s.t.

$$g(t) := F(x + tv) \text{ is } C^2.$$

- Function g has negative curvature:

$$g''(0) < 0.$$

[5] (D-Drusvyatskiy '19)

Q: What is an avoidable saddle point in nonsmooth optimization?[5]

Recall C^2 case: A strict saddle is critical point with negative curvature:

$$\nabla F(x) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(x)) < 0$$

Generalization Attempt: A strict saddle is critical point such that

- There exists direction v s.t.

$$g(t) := F(x + tv) \text{ is } C^2.$$

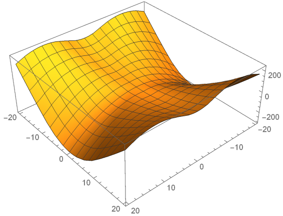
- Function g has negative curvature:

$$g''(0) < 0.$$

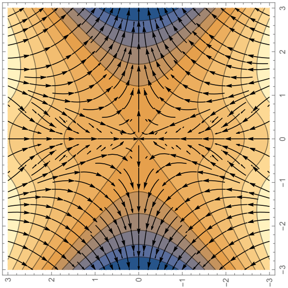
Equivalent when F is C^2.

[5] (D-Drusvyatskiy '19)

Negative curvature is not enough even for C^1 functions



(a) C^1loss F

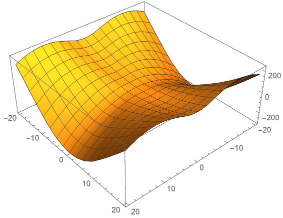


(b) Flow $\dot{\gamma} = -\nabla F(\gamma)$

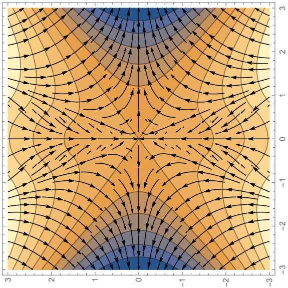
$$F(x, y) = \textbf{Moreau}\{(|x| + |y|)^2 - 2y^2\}$$

Negative curvature: $F(0, y) = -\alpha y^2$

Negative curvature is not enough even for C^1 functions



(a) C^1loss F



(b) Flow $\dot{\gamma} = -\nabla F(\gamma)$

$$F(x, y) = \textbf{Moreau}\{(|x| + |y|)^2 - 2y^2\}$$

Negative curvature: $F(0, y) = -\alpha y^2$

Problem: do not reach y axis fast enough to benefit from curvature!

An extra ingredient: sharpness

Idea: Require F to grow **sharply** away from axis:

$$\inf\{\|\nabla F(x,y)\|: \text{ for } (x,y) \text{ off of } y \text{ axis}\} > 0$$

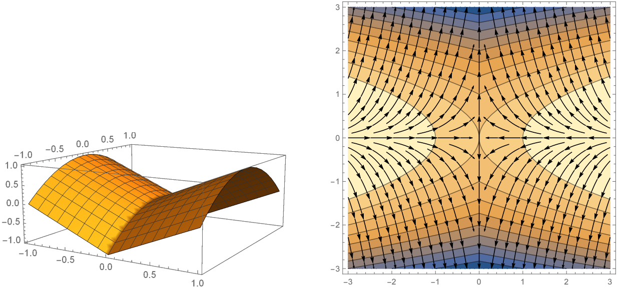
Benefit: Ensures grad. flow aims towards axis with (at least) constant speed.

An extra ingredient: sharpness

Idea: Require F to grow **sharply** away from axis:

$$\inf\{\|\nabla F(x,y)\|: \text{ for } (x,y) \text{ off of } y \text{ axis}\} > 0$$

Benefit: Ensures grad. flow aims towards axis with (at least) constant speed.



(a) A nonsmooth loss F (b) Flow $\dot{\gamma} \in -\partial F(\gamma)$

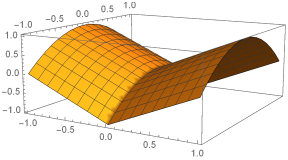
Negative curvature: $F(0, y) = -\alpha y^2$

An extra ingredient: sharpness

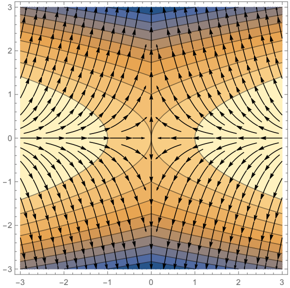
Idea: Require F to grow **sharply** away from axis:

$$\inf\{\|\nabla F(x,y)\| : \text{ for } (x,y) \text{ off of } y \text{ axis}\} > 0$$

Benefit: Ensures grad. flow aims towards axis with (at least) constant speed.



(a) A nonsmooth loss F



(b) Flow $\dot{\gamma} \in -\partial F(\gamma)$

Negative curvature: $F(0, y) = -\alpha y^2$

Question: How to generalize?

The active manifold

Idea: Replace axis with "active manifold" of smoothness.

(Wright '93, Lemaréchal-Oustry-Sagastizábal '96, Bonnans-Shapiro '00, Lewis '03, Drusvyatskiy-Lewis '14...)

The active manifold

Idea: Replace axis with "active manifold" of smoothness.

Defn: Critical point lies on C^2-smooth "active manifold $\mathcal{M}$":

1. F varies C^2-**smoothly** along $\mathcal{M}$.

(Wright '93, Lemaréchal-Oustry-Sagastizábal '96, Bonnans-Shapiro '00, Lewis '03, Drusvyatskiy-Lewis '14...)

The active manifold

Idea: Replace axis with "active manifold" of smoothness.

Defn: Critical point lies on C^2-smooth "active manifold $\mathcal{M}$":

1. F varies C^2-**smoothly** along $\mathcal{M}$.
2. F grows **sharply** normal to $\mathcal{M}$:

$$\inf\{\|v\| \colon v \in \partial F(z) \colon z \in U \backslash \mathcal{M}\} > 0.$$

(Wright '93, Lemaréchal-Oustry-Sagastizábal '96, Bonnans-Shapiro '00, Lewis '03, Drusvyatskiy-Lewis '14...)

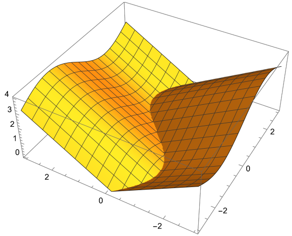
The active manifold

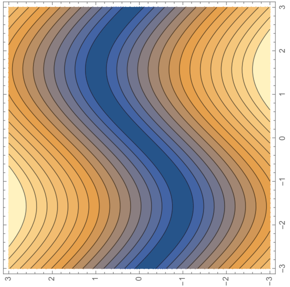
Idea: Replace axis with "active manifold" of smoothness.

Defn: Critical point lies on C^2-smooth "active manifold $\mathcal{M}$":

1. F varies C^2-**smoothly** along $\mathcal{M}$.
2. F grows **sharply** normal to $\mathcal{M}$:

$$\inf\{\|v\| : v \in \partial F(z) : z \in U \setminus \mathcal{M}\} > 0.$$





(Wright '93, Lemaréchal-Oustry-Sagastizábal '96, Bonnans-Shapiro '00, Lewis '03, Drusvyatskiy-Lewis '14. . .)

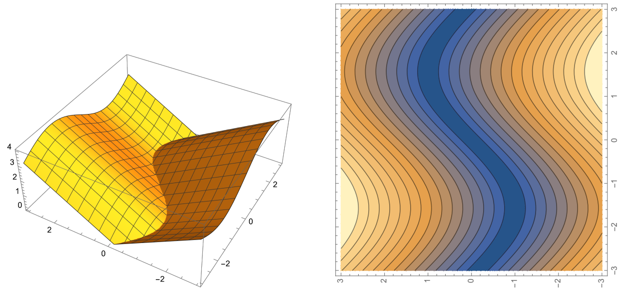
The active manifold

Idea: Replace axis with "active manifold" of smoothness.

Defn: Critical point lies on C^2-smooth "active manifold $\mathcal{M}$":

1. F varies C^2-**smoothly** along $\mathcal{M}$.
2. F grows **sharply** normal to $\mathcal{M}$:

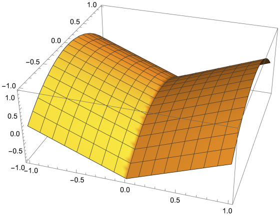
$$\inf\{\|v\| : v \in \partial F(z) : z \in U \backslash \mathcal{M}\} > 0.$$



Question: What about curvature?

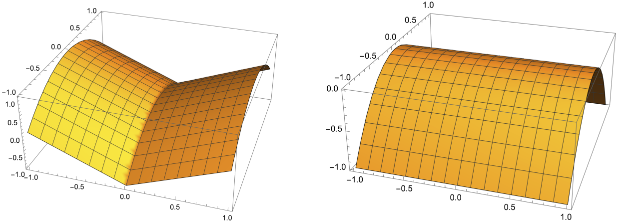
(Wright '93, Lemaréchal-Oustry-Sagastizábal '96, Bonnans-Shapiro '00, Lewis '03, Drusvyatskiy-Lewis '14...)

Putting it all together: the active strict saddle property



(a) A nonsmooth loss F

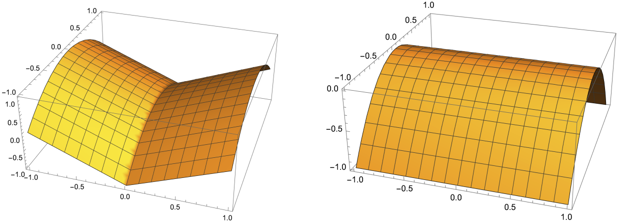
Putting it all together: the active strict saddle property



(a) A nonsmooth loss F (b) Smooth extension $F \circ P_{\mathcal{M}}$

Putting it all together: the active strict saddle property

Defn: (D-Drusvyatskiy '19) a critical point $\bar{x}$ of F is an active strict saddle if

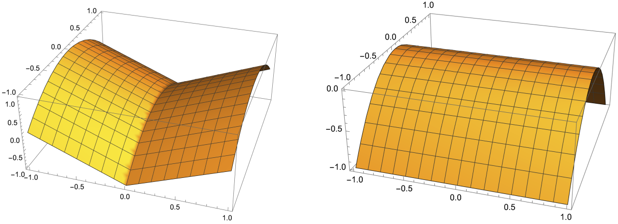


(a) A nonsmooth loss F (b) Smooth extension $F \circ P_{\mathcal{M}}$

Putting it all together: the active strict saddle property

Defn: (D-Drusvyatskiy '19) a critical point $\bar{x}$ of F is an active strict saddle if

1. F admits active manifold $\mathcal{M}$ containing $\bar{x}$.



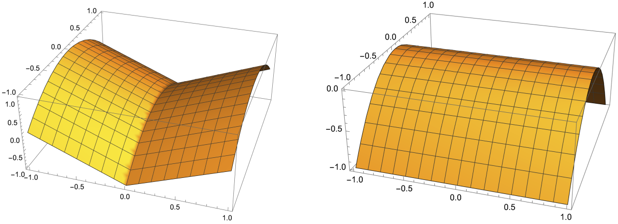
(a) A nonsmooth loss F (b) Smooth extension $F \circ P_{\mathcal{M}}$

Putting it all together: the active strict saddle property

Defn: (D-Drusvyatskiy '19) a critical point $\bar{x}$ of F is an active strict saddle if

1. F admits active manifold $\mathcal{M}$ containing $\bar{x}$.
2. The smooth extension $F \circ P_{\mathcal{M}}$ has a strict saddle point at $\bar{x}$:

$$\lambda_{\min}(\nabla^2(F \circ P_{\mathcal{M}})(\bar{x})) < 0.$$



(a) A nonsmooth loss F (b) Smooth extension $F \circ P_{\mathcal{M}}$

Putting it all together: the active strict saddle property

Although it may seem stringent, this property is **generic**:

Theorem (Drusvyatskiy-Ioffe-Lewis '16, D-Drusvyatskiy '19)

If F is **semi-algebraic** *and weakly convex, then for full Lebesgue measure set of perturbations $v \in \mathbb{R}^d$ every critical point of*

$$F_v(x) = F(x) - \langle v, x \rangle$$

is either an active strict saddle or a **local minimizer**.

Putting it all together: the active strict saddle property

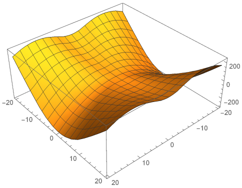
Although it may seem stringent, this property is **generic**:

Theorem (Drusvyatskiy-Ioffe-Lewis '16, D-Drusvyatskiy '19)

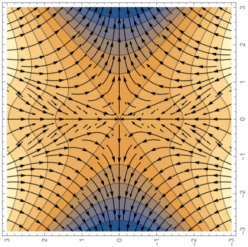
If F is **semi-algebraic** *and weakly convex, then for full Lebesgue measure set of perturbations $v \in \mathbb{R}^d$ every critical point of*

$$F_v(x) = F(x) - \langle v, x \rangle$$

is either an active strict saddle or a **local minimizer**.



(a) C^1loss F



(b) Flow $\dot{\gamma} = -\nabla F(\gamma)$

Example is Highly Unstable: small linear tilts do not exhibit this behavior!

Question: Do the three proximal methods avoid active strict saddles?

[6] For the algorithms considered thus far, critical points are fixed points of the iteration.

Question: Do the three proximal methods avoid active strict saddles?

Strategy: Borrow "stable manifold theorem" argument from smooth setting!

[6] For the algorithms considered thus far, critical points are fixed points of the iteration.

Question: Do the three proximal methods avoid active strict saddles?

Strategy: Borrow "stable manifold theorem" argument from smooth setting!

Key: view algorithms

$$x_{t+1} = \arg\min_y F_{x_t}(y),$$

as fixed-point iteration of well-behaved operator T.[6]

[6] For the algorithms considered thus far, critical points are fixed points of the iteration.

Recipe for smooth functions

Fixed point iteration

$$x_{t+1} = T(x_t)$$

Recipe for smooth functions

Fixed point iteration

$$x_{t+1} = T(x_t) \qquad [\text{Grad descent is } T = I - \eta \nabla F]$$

Recipe for smooth functions

Fixed point iteration

$$x_{t+1} = T(x_t) \qquad\qquad [\text{Grad descent is } T = I - \eta\nabla F]$$

Recipe:

- Strict saddles $\bar{x}$

$$\nabla F(\bar{x}) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(\bar{x})) < 0$$

are **unstable** fixed points:

$$\nabla T(\bar{x}) \text{ has EigVal of magnitude } > 1$$

Recipe for smooth functions

Fixed point iteration

$$x_{t+1} = T(x_t) \qquad\qquad [\text{Grad descent is } T = I - \eta \nabla F]$$

Recipe:

- Strict saddles $\bar{x}$

  $$\nabla F(\bar{x}) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(\bar{x})) < 0$$

  are **unstable** fixed points:

  $$\nabla T(\bar{x}) \text{ has EigVal of magnitude } > 1$$

- Classical **center-stable manifold theorem** implies

  $$W := \left\{ x : \lim_{k \to \infty} T^k(x) \text{ is unstable} \right\} \quad \text{has Lebesgue measure zero.}$$

Recipe for smooth functions

Fixed point iteration

$$x_{t+1} = T(x_t) \qquad \text{[Grad descent is } T = I - \eta \nabla F\text{]}$$

Recipe:

- Strict saddles $\bar{x}$

  $$\nabla F(\bar{x}) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(\bar{x})) < 0$$

  are **unstable** fixed points:

  $$\nabla T(\bar{x}) \text{ has EigVal of magnitude } > 1$$

- Classical **center-stable manifold theorem** implies

  $$W := \left\{ x : \lim_{k \to \infty} T^k(x) \text{ is unstable} \right\} \qquad \text{has Lebesgue measure zero.}$$

- Since random init will not land in W, algorithm avoids strict saddles

Recipe for smooth functions

Fixed point iteration

$$x_{t+1} = T(x_t) \qquad\qquad [\text{Grad descent is } T = I - \eta\nabla F]$$

Recipe:

- Strict saddles $\bar{x}$

  $$\nabla F(\bar{x}) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 F(\bar{x})) < 0$$

  are **unstable** fixed points:

  $$\nabla T(\bar{x}) \text{ has EigVal of magnitude } > 1$$

- Classical **center-stable manifold theorem** implies

  $$W := \left\{ x \colon \lim_{k\to\infty} T^k(x) \text{ is unstable} \right\} \qquad \text{has Lebesgue measure zero.}$$

- Since random init will not land in W, algorithm avoids strict saddles

Important: Argument requires that T is local diffeomorphism.

Beyond gradient descent

To apply argument, need

1. **Local Smoothness:** The update mapping

$$S(x) = \arg\min_y F_x(y),$$

is a local C^1 diffeomorphism near active strict saddle points.

Beyond gradient descent

To apply argument, need

1. **Local Smoothness:** The update mapping

$$S(x) = \arg\min_y F_x(y),$$

   is a local C^1 diffeomorphism near active strict saddle points.

2. **Unstable:** Active strict saddle points $\bar{x}$ are unstable:

$$\nabla S(\bar{x}) \text{ has EigVal of magnitude } > 1.$$

Beyond gradient descent

To apply argument, need

1. **Local Smoothness:** The update mapping

$$S(x) = \arg\min_{y} F_x(y),$$

   is a local C^1 diffeomorphism near active strict saddle points.

2. **Unstable:** Active strict saddle points $\bar{x}$ are unstable:

$$\nabla S(\bar{x}) \text{ has EigVal of magnitude } > 1.$$

Focus on Local Smoothness, since other calculation complex.

Local smoothness

Surprising: Function F is nonsmooth, yet S is C^1 around strict saddles. Why?

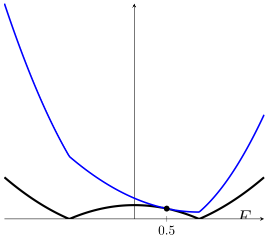
[7] Lemaréchal-Sagastizábal '97

Local smoothness

Surprising: Function F is nonsmooth, yet S is C^1 around strict saddles. Why?

$$\text{Sharpness} \implies \text{Identification}$$
$$S(x) \in \mathcal{M} \text{ near } \bar{x}!$$

Example: Prox-point



$$F_{x_t}(y) = F(y) + \frac{1}{2\eta}\|y - x_t\|^2$$

[7] Lemaréchal-Sagastizábal '97

Local smoothness

Surprising: Function F is nonsmooth, yet S is C^1 around strict saddles. Why?

$$\text{Sharpness} \implies \text{Identification}$$
$$S(x) \in \mathcal{M} \text{ near } \bar{x}!$$

Important: Do not need to know $\mathcal{M}$!

[7] Lemaréchal-Sagastizábal '97

Local smoothness

Surprising: Function F is nonsmooth, yet S is C^1 around strict saddles. Why?

$$\text{Sharpness} \implies \text{Identification}$$
$$S(x) \in \mathcal{M} \text{ near } \bar{x}!$$

Important: Do not need to know $\mathcal{M}$!

Consequence (Prox-point Method):

$$S(x) = \arg\min_{y} F(y) + \frac{1}{2\eta}\|y - x\|^2 = \arg\min_{y \in \mathcal{M}} F(y) + \frac{1}{2\eta}\|y - x\|^2.$$

[7] Lemaréchal-Sagastizábal '97

Local smoothness

Surprising: Function F is nonsmooth, yet S is C^1 around strict saddles. Why?

$$\text{Sharpness} \implies \text{Identification}$$
$$S(x) \in \mathcal{M} \text{ near } \bar{x}!$$

Important: Do not need to know $\mathcal{M}$!

Consequence (Prox-point Method):

$$S(x) = \operatorname*{arg\,min}_{y} F(y) + \frac{1}{2\eta}\|y - x\|^2 = \operatorname*{arg\,min}_{y \in \mathcal{M}} F(y) + \frac{1}{2\eta}\|y - x\|^2.$$

$\implies$ minimizing smooth function over smooth manifold!

[7] Lemaréchal-Sagastizábal '97

Local smoothness

Surprising: Function F is nonsmooth, yet S is C^1 around strict saddles. Why?

$$\text{Sharpness} \implies \text{Identification}$$
$$S(x) \in \mathcal{M} \text{ near } \bar{x}!$$

Important: Do not need to know $\mathcal{M}$!

Consequence (Prox-point Method):

$$S(x) = \arg\min_{y} F(y) + \frac{1}{2\eta}\|y - x\|^2 = \arg\min_{y \in \mathcal{M}} F(y) + \frac{1}{2\eta}\|y - x\|^2.$$

$\implies$ minimizing smooth function over smooth manifold!

Then Weak convexity + classical perturbation theory $\implies$ S is C^1 near $\bar{x}$.[7]

[7] Lemaréchal-Sagastizábal '97

Avoiding active strict saddles

Proof extends to the three methods:

Algorithm	Objective F	Update function $F_x(y)$
Prox-point	$F(x)$	$F(y) + \frac{1}{2\eta}\|y-x\|^2$
Prox-linear	$h(c(x)) + r(x)$	$h(c(x) + \nabla c(x)(y-x)) + r(y) + \frac{1}{2\eta}\|y-x\|^2$
Prox-gradient	$f(x) + r(x)$	$f(x) + \langle \nabla f(x), y-x \rangle + r(y) + \frac{1}{2\eta}\|y-x\|^2$

Table: h is convex and Lipschitz, r is weakly convex, and f and c are C^2-smooth.

Avoiding active strict saddles

Theorem: (Local smoothness, D-Drusvyatskiy '19)

Around each active strict saddle $\bar{x}$ of F, the iteration mapping

$$S(x) = \arg\min_y F_x(y),$$

is C^1 and the Jacobian $\nabla S(\bar{x})$ has a real EigVal strictly greater than 1

Proof more interesting/surprising for prox-gradient and prox-linear.

Avoiding active strict saddles

Problem: S may not be Local diffeomorphism

Avoiding active strict saddles

Problem: S may not be Local diffeomorphism

Easy solution: Add damping

$$T = (1 - \lambda)I + \lambda S.$$

Avoiding active strict saddles

Corollary: (Random initialization, D-Drusvyatskiy '19)

Randomly initialized three methods with small damping

$$x_{t+1} = (1-\lambda)x_t + \lambda S(x_t),$$

locally escape active strict saddles.

Globalization:

- Results hold globally when S is Lipschitz (prox-point, prox-gradient)
- **Open Problem:** Is prox-linear update globally Lipschitz?

Beyond proximal methods

Limitation of result: Only applies to three "proximal methods."

Algorithm	Objective F	Update function $F_x(y)$
Prox-point	$F(x)$	$F(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-linear	$h(c(x)) + r(x)$	$h(c(x) + \nabla c(x)(y - x)) + r(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-gradient	$f(x) + r(x)$	$f(x) + \langle \nabla f(x), y - x \rangle + r(y) + \frac{1}{2\eta}\|y - x\|^2$

Table: h is convex and Lipschitz, r is weakly convex, and f and c are C^2-smooth.

Beyond proximal methods

Limitation of result: Only applies to three "proximal methods."

Algorithm	Objective F	Update function $F_x(y)$
Prox-point	$F(x)$	$F(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-linear	$h(c(x)) + r(x)$	$h(c(x) + \nabla c(x)(y - x)) + r(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-gradient	$f(x) + r(x)$	$f(x) + \langle \nabla f(x), y - x \rangle + r(y) + \frac{1}{2\eta}\|y - x\|^2$

Table: h is convex and Lipschitz, r is weakly convex, and f and c are C^2-smooth.

Drawbacks:

1. **Numerical Difficulties:** need exact solutions to subproblems.

Beyond proximal methods

Limitation of result: Only applies to three "proximal methods."

Algorithm	Objective F	Update function $F_x(y)$
Prox-point	$F(x)$	$F(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-linear	$h(c(x)) + r(x)$	$h(c(x) + \nabla c(x)(y - x)) + r(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-gradient	$f(x) + r(x)$	$f(x) + \langle \nabla f(x), y - x \rangle + r(y) + \frac{1}{2\eta}\|y - x\|^2$

Table: h is convex and Lipschitz, r is weakly convex, and f and c are C^2-smooth.

Drawbacks:

1. **Numerical Difficulties:** need exact solutions to subproblems.
2. Decomposable **structure** not always available.

Beyond proximal methods

Limitation of result: Only applies to three "proximal methods."

Algorithm	Objective F	Update function $F_x(y)$
Prox-point	$F(x)$	$F(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-linear	$h(c(x)) + r(x)$	$h(c(x) + \nabla c(x)(y - x)) + r(y) + \frac{1}{2\eta}\|y - x\|^2$
Prox-gradient	$f(x) + r(x)$	$f(x) + \langle \nabla f(x), y - x \rangle + r(y) + \frac{1}{2\eta}\|y - x\|^2$

Table: h is convex and Lipschitz, r is weakly convex, and f and c are C^2-smooth.

Drawbacks:

1. **Numerical Difficulties:** need exact solutions to subproblems.
2. Decomposable **structure** not always available.

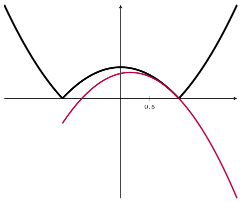
Alternative: subgradient method

The subdifferential of a weakly convex function

Fact: For any $F\colon \mathbb{R}^d \to \mathbb{R}$, have equivalence:

- F is ρ-weakly convex
- Subgradient inequality: $\forall x \exists v_x$ satisfying

$$F(y) \geq F(x) + \langle v_x, y - x\rangle - \frac{\rho}{2}\|y - x\|^2$$

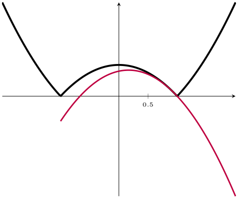


The subdifferential of a weakly convex function

Fact: For any $F\colon \mathbb{R}^d \to \mathbb{R}$, have equivalence:

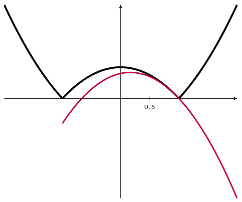
- F is ρ-weakly convex
- Subgradient inequality: $\forall x \exists v_x$ satisfying

$$F(y) \geq F(x) + \langle v_x, y - x\rangle - \frac{\rho}{2}\|y - x\|^2$$



Subdifferential: $\partial F(x) := \{v_x\}$

The subdifferential of a weakly convex function



Fact: For any $F \colon \mathbb{R}^d \to \mathbb{R}$, have equivalence:

- F is ρ-weakly convex
- Subgradient inequality: $\forall x \exists v_x$ satisfying

$$F(y) \geq F(x) + \langle v_x, y - x \rangle - \frac{\rho}{2}\|y - x\|^2$$

Subdifferential: $\partial F(x) := \{v_x\}$

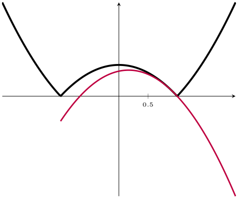
Calculus: $\boxed{\partial (h \circ c)(x) := \nabla c(x)^T \partial h(c(x))}$

The subdifferential of a weakly convex function

Fact: For any $F\colon \mathbb{R}^d \to \mathbb{R}$, have equivalence:

- F is ρ-weakly convex
- Subgradient inequality: $\forall x \exists v_x$ satisfying

$$F(y) \geq F(x) + \langle v_x, y - x\rangle - \frac{\rho}{2}\|y - x\|^2$$



Subdifferential: $\partial F(x) := \{v_x\}$

Calculus: $\boxed{\partial(h \circ c)(x) := \nabla c(x)^T \partial h(c(x))}$

Fermat's rule: If $\bar{x}$ is a local minimizer of F then

$$0 \in \partial F(\bar{x}).$$

Alternative: the subgradient method

Idea: At time t

[8] Bolte-Pauwels '19-'20

Alternative: the subgradient method

Idea: At time t

1. **"Linearize F:"** choose $v_t \in \partial F(x_t)$ and form

$$F_{x_t,\alpha_t}(y) = F(x_t) + \langle v_t, y - x_t \rangle + \frac{1}{2\alpha_t}\|y - x_t\|^2.$$

[8] Bolte-Pauwels '19-'20

Alternative: the subgradient method

Idea: At time t

1. **"Linearize F:"** choose $v_t \in \partial F(x_t)$ and form

$$F_{x_t,\alpha_t}(y) = F(x_t) + \langle v_t, y - x_t \rangle + \frac{1}{2\alpha_t}\|y - x_t\|^2.$$

2. Next iterate minimizes:

$$\begin{aligned} x_{t+1} &= \arg\min_y F_{x_t,\alpha_t}(y) \\ &= x_t - \alpha_t v_t. \end{aligned}$$

[8] Bolte-Pauwels '19-'20

Alternative: the subgradient method

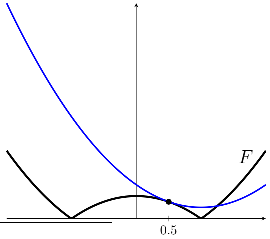
Idea: At time t

1. **"Linearize F:"** choose $v_t \in \partial F(x_t)$ and form

$$F_{x_t,\alpha_t}(y) = F(x_t) + \langle v_t, y - x_t \rangle + \frac{1}{2\alpha_t}\|y - x_t\|^2.$$

2. Next iterate minimizes:

$$\begin{aligned} x_{t+1} &= \arg\min_y F_{x_t,\alpha_t}(y) \\ &= x_t - \alpha_t v_t. \end{aligned}$$



[8] Bolte-Pauwels '19-'20

Alternative: the subgradient method

Idea: At time t

1. **"Linearize F:"** choose $v_t \in \partial F(x_t)$ and form

$$F_{x_t,\alpha_t}(y) = F(x_t) + \langle v_t, y - x_t \rangle + \frac{1}{2\alpha_t}\|y - x_t\|^2.$$

2. Next iterate minimizes:

$$\begin{aligned} x_{t+1} &= \operatorname*{arg\,min}_y F_{x_t,\alpha_t}(y) \\ &= x_t - \alpha_t v_t. \end{aligned}$$

Benefits:

1. Computable with extensive calculus: $\partial(h \circ c)(x) := \nabla c(x)^T \partial h(c(x))$

[8] Bolte-Pauwels '19-'20

Alternative: the subgradient method

Idea: At time t

1. **"Linearize F:"** choose $v_t \in \partial F(x_t)$ and form

$$F_{x_t,\alpha_t}(y) = F(x_t) + \langle v_t, y - x_t \rangle + \frac{1}{2\alpha_t}\|y - x_t\|^2.$$

2. Next iterate minimizes:

$$\begin{aligned} x_{t+1} &= \arg\min_y F_{x_t,\alpha_t}(y) \\ &= x_t - \alpha_t v_t. \end{aligned}$$

Benefits:

1. Computable with extensive calculus: $\partial(h \circ c)(x) := \nabla c(x)^T \partial h(c(x))$
2. Can often replace v_t with result of auto-differentiation procedure.[8]

[8] Bolte-Pauwels '19-'20

Extension: Subgradient method

Question: Does subgradient method avoid active strict saddle points?

$$x_{t+1} \in x_t - \alpha_t \partial F(x_t)$$

[9] D-Drusvyatskiy-Jiang '21

Extension: Subgradient method

Question: Does subgradient method avoid active strict saddle points?

$$x_{t+1} \in x_t - \alpha_t \partial F(x_t)$$

Difficulties:

- Identification fails: $x_t \notin \mathcal{M}$.
- Unclear how to leverage smoothness on the manifold.

Our recent work[9] overcomes these difficulties.

[9] D-Drusvyatskiy-Jiang '21

Extension: Subgradient method

Question: Does subgradient method avoid active strict saddle points?

$$x_{t+1} \in x_t - \alpha_t \partial F(x_t)$$

Difficulties:

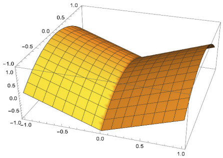
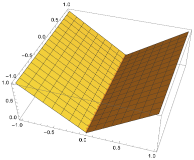
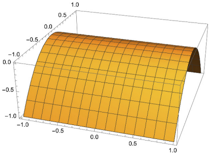
- Identification fails: $x_t \notin \mathcal{M}$.
- Unclear how to leverage smoothness on the manifold.

Our recent work[9] overcomes these difficulties.

Key: "orthogonal decomposition" of trajectory.

[9] D-Drusvyatskiy-Jiang '21

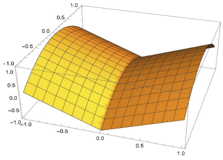
$\mathcal{VU}$ decomposition[10]

 $=$  $+$ 

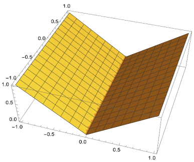
$F \qquad\qquad F_{\mathcal{V}} \qquad\qquad F_{\mathcal{U}}$

[10] Mifflin-Sagastizábal '05

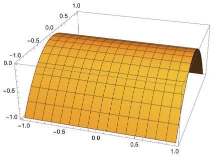
$\mathcal{VU}$ decomposition[10]



$=$



$+$

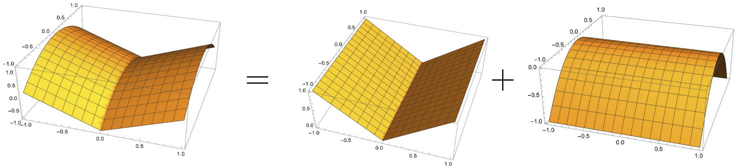


$F \qquad F_{\mathcal{V}} \qquad F_{\mathcal{U}}$

Decompose trajectory:

[10] Mifflin-Sagastizábal '05

$\mathcal{V}\mathcal{U}$ decomposition[10]



F $\qquad\qquad$ $F_{\mathcal{V}}$ $\qquad\qquad$ $F_{\mathcal{U}}$

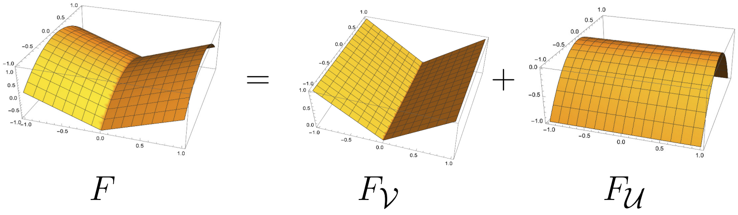
Decompose trajectory:

1. Tangent directions:

$$P_{\mathcal{M}}(x_{t+1}) \approx P_{\mathcal{M}}(x_t) - \alpha_t \nabla F_{\mathcal{U}}(x_t)$$

[10] Mifflin-Sagastizábal '05

$\mathcal{VU}$ decomposition[10]



Decompose trajectory:

1. Tangent directions:

$$P_{\mathcal{M}}(x_{t+1}) \approx P_{\mathcal{M}}(x_t) - \alpha_t \nabla F_{\mathcal{U}}(x_t)$$

2. Normal directions:

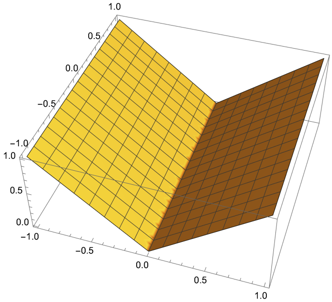
$$x_{t+1} - P_{\mathcal{M}}(x_{t+1}) \approx x_t - P_{\mathcal{M}}(x_t) - \alpha_t \widetilde{\nabla} F_{\mathcal{V}}(x_t)$$

[10] Mifflin-Sagastizábal '05

The two regularity assumptions

1. **Aiming:** Negative subgradients aim towards manifold:

$$\text{Sharpness} \implies \langle \widetilde{\nabla} F_{\mathcal{V}}(x_t), x_t - P_{\mathcal{M}}(x_t) \rangle \geq \mu \, \mathrm{dist}(x_t, \mathcal{M})$$



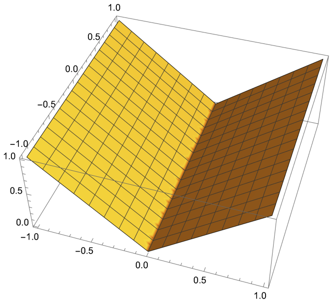
The two regularity assumptions

1. **Aiming:** Negative subgradients aim towards manifold:

$$\text{Sharpness} \implies \langle \widetilde{\nabla} F_{\mathcal{V}}(x_t), x_t - P_{\mathcal{M}}(x_t) \rangle \geq \mu \operatorname{dist}(x_t, \mathcal{M})$$

2. **Smooth in tangent directions:**

$$\|P_{\mathcal{T}_{\mathcal{M}}(y)} \tilde{\nabla} F_{\mathcal{V}}(x_t)\| \leq C \|x_t - y\| \qquad \text{for } y \in \mathcal{M}.$$



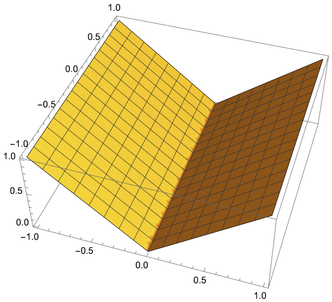
The two regularity assumptions

1. **Aiming:** Negative subgradients aim towards manifold:

$$\text{Sharpness} \implies \langle \widetilde{\nabla} F_{\mathcal{V}}(x_t), x_t - P_{\mathcal{M}}(x_t)\rangle \geq \mu \,\mathrm{dist}(x_t, \mathcal{M})$$

2. **Smooth in tangent directions:**

$$\|P_{\mathcal{T}_{\mathcal{M}}(y)} \tilde{\nabla} F_{\mathcal{V}}(x_t)\| \leq C\|x_t - y\| \qquad \text{for } y \in \mathcal{M}.$$



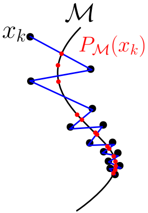
Prevalent: true generically for weakly convex semialgebraic problems.

The two pillars

The two pillars: For a wide class of problems

- Subgradient method quickly approaches the active manifold:

$$\mathrm{dist}(x_t, \mathcal{M}) = O(\alpha_t).$$



(a) Quickly approach manifold

The two pillars

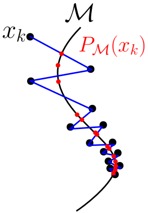
The two pillars: For a wide class of problems

- Subgradient method quickly approaches the active manifold:

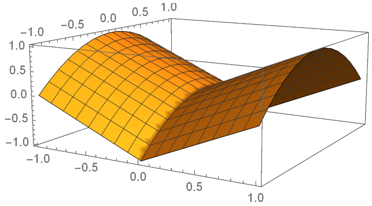
$$\operatorname{dist}(x_t, \mathcal{M}) = O(\alpha_t).$$

- The shadow $y_t = P_{\mathcal{M}}(x_t)$ forms inexact Riemannian gradient sequence:

$$y_{t+1} = y_t - \alpha_t \nabla_{\mathcal{M}} F(y_t) + O(\alpha_t \operatorname{dist}(x_t, \mathcal{M}) + \alpha_t^2).$$



(a) Quickly approach manifold



(b) “Smooth in tangent directions”

The two pillars

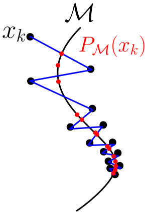
The two pillars: For a wide class of problems

- Subgradient method quickly approaches the active manifold:

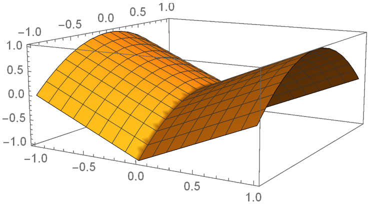
$$\mathrm{dist}(x_t, \mathcal{M}) = O(\alpha_t).$$

- The shadow $y_t = P_{\mathcal{M}}(x_t)$ forms inexact Riemannian gradient sequence:

$$y_{t+1} = y_t - \alpha_t \nabla_{\mathcal{M}} F(y_t) + O(\alpha_t \mathrm{dist}(x_t, \mathcal{M}) + \alpha_t^2).$$



(a) Quickly approach manifold



(b) "Smooth in tangent directions"

Conclusion: Get to the manifold quick enough to leverage smoothness of F!

Main result

Due to inexactness, must analyze "perturbed" subgradient method[11]:

$$x_{t+1} \in x_t - \alpha_t(\partial F(x_t) + \nu_t) \qquad \text{where } \nu_t \sim \mathsf{Unif}(B).$$

[11] D-Drusvyatskiy-Jiang '21

[12] Concurrent work: Bianchi-Hachem-Schechtman'21.

Main result

Due to inexactness, must analyze "perturbed" subgradient method[11]:

$$x_{t+1} \in x_t - \alpha_t(\partial F(x_t) + \nu_t) \qquad \text{where } \nu_t \sim \mathrm{Unif}(B).$$

$$\implies y_{t+1} = y_t - \alpha_t(\nabla_{\mathcal{M}} F(y_t) + \nu_t) + O(\alpha_t \mathrm{dist}(x_t, \mathcal{M}) + \alpha_t^2).$$

[11] D-Drusvyatskiy-Jiang '21

[12] Concurrent work: Bianchi-Hachem-Schechtman'21.

Main result

Due to inexactness, must analyze "perturbed" subgradient method[11]:

$$x_{t+1} \in x_t - \alpha_t(\partial F(x_t) + \nu_t) \qquad \text{where } \nu_t \sim \mathrm{Unif}(B).$$

Under mild conditions, we show

Theorem: (D-Drusvyatskiy-Jiang '19)[12]

Almost surely, x_t does not converge to an active strict saddle point.

[11] D-Drusvyatskiy-Jiang '21

[12] Concurrent work: Bianchi-Hachem-Schechtman'21.

Main result

Due to inexactness, must analyze "perturbed" subgradient method[11]:

$$x_{t+1} \in x_t - \alpha_t(\partial F(x_t) + \nu_t) \qquad \text{where } \nu_t \sim \mathrm{Unif}(B).$$

Under mild conditions, we show

Theorem: (D-Drusvyatskiy-Jiang '19)[12]

Almost surely, x_t does not converge to an active strict saddle point.

Corollary: (D-Drusvyatskiy-Jiang '19)

Perturbed subgradient method converges only to local minimizers of generic semialgebraic weakly convex functions.

[11] D-Drusvyatskiy-Jiang '21

[12] Concurrent work: Bianchi-Hachem-Schechtman'21.

Main result

Due to inexactness, must analyze "perturbed" subgradient method[11]:

$$x_{t+1} \in x_t - \alpha_t(\partial F(x_t) + \nu_t) \qquad \text{where } \nu_t \sim \mathrm{Unif}(B).$$

Under mild conditions, we show

Theorem: (D-Drusvyatskiy-Jiang '19)[12]

Almost surely, x_t does not converge to an active strict saddle point.

Corollary: (D-Drusvyatskiy-Jiang '19)

Perturbed subgradient method converges only to local minimizers of generic semialgebraic weakly convex functions.

Extensions.

1. **Algorithms:** Proximal/projected subgradient methods.

[11] D-Drusvyatskiy-Jiang '21

[12] Concurrent work: Bianchi-Hachem-Schechtman'21.

Main result

Due to inexactness, must analyze "perturbed" subgradient method[11]:

$$x_{t+1} \in x_t - \alpha_t(\partial F(x_t) + \nu_t) \qquad \text{where } \nu_t \sim \mathrm{Unif}(B).$$

Under mild conditions, we show

Theorem: (D-Drusvyatskiy-Jiang '19)[12]

Almost surely, x_t does not converge to an active strict saddle point.

Corollary: (D-Drusvyatskiy-Jiang '19)

Perturbed subgradient method converges only to local minimizers of generic semialgebraic weakly convex functions.

Extensions.

1. **Algorithms:** Proximal/projected subgradient methods.
2. **Beyond weak convexity:** Clarke regularity.

---

[11] D-Drusvyatskiy-Jiang '21

[12] Concurrent work: Bianchi-Hachem-Schechtman'21.

Thank you!

References

- Proximal methods avoid active strict saddles of weakly convex functions
  D, Drusvyatskiy. *Found. Comput. Math.* `arxiv.org/abs/1912.07146`.
- Subgradient methods near active manifolds: saddle point avoidance, local convergence, and asymptotic normality
  D, Drusvyatskiy, Jiang. `https://arxiv.org/abs/2108.11832`.