Avoiding saddle points in nonsmooth optimization

Damek Davis
School of Operations Research and Information Engineering
Cornell University

Joint with M. Diaz (Caltech/JHU) and D. Drusvyatskiy (U. Washington)

SIAM Optimization Conference 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.*

- **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:

*Are “strict saddles” problematic for nonsmooth minimization?*

---

$^2$Lee-Simchowitz-Jordan-Recht '16
Recipe for smooth functions

Fixed point iteration

\[ x_{t+1} = T(x_t) \]
Recipe for smooth functions

Fixed point iteration

\[ x_{t+1} = T(x_t) \]

[Grad descent is \( T = I - \eta \nabla F \)]
Recipe for smooth functions

Fixed point iteration

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

Recipe:

- **Strict saddles** \( \bar{x} \) are **unstable** fixed points:
  \[ \nabla T(\bar{x}) \text{ has EigVal of magnitude } > 1 \]

- Classical **center-stable manifold theorem** implies
  \[ W := \left\{ x \in \mathbb{R}^d : \lim_{k \to \infty} T^k(x) \text{ is unstable} \right\} \text{ has Lebesgue measure zero.} \]

- Since random init will not land in \( W \), algorithm avoids **strict saddles**
Recipe for smooth functions

Fixed point iteration

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

Recipe:

- **Strict saddles** \( \bar{x} \) are **unstable** fixed points:

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

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

  \[ W := \left\{ x \in \mathbb{R}^d : \lim_{k \to \infty} T^k(x) \text{ is unstable} \right\} \text{ has Lebesgue measure zero.} \]

- Since **random init** will not land in \( W \), algorithm avoids **strict saddles**

**Limitation:** Proof for **GD** requires \( F \) to be \( C^2 \). **Can this be relaxed?**
Negative curvature is not enough even for $C^1$ functions

(a) $C^1$ loss $F$

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

\[
F(x, y) = \text{Moreau}\{(|x| + |y|)^2 - 2x^2\}
\]
Negative curvature is not enough even for $C^1$ functions

(a) $C^1$ loss $F$

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

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

Highly Unstable: small linear tilts do not exhibit this behavior!
Negative curvature + structured nonsmoothness

**An extra ingredient:** nonsmoothness manifests in structured way

![Graphical representation](image)

(a) A nonsmooth loss $F$

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

Critical point lies on $C^2$-smooth "active manifold $\mathcal{M}$" ($y$-axis):

- $F$ varies $C^2$-smoothly along $\mathcal{M}$ and 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...)

5 / 14
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. $F$ decreases quadratically along some direction $v \in \mathcal{T}_\mathcal{M}(\bar{x})$:

\[
d_2\frac{d}{dt}(F_{\mathcal{T}}(t)) - t = 0 < 0\text{ for some } C^2\text{ curve } \mu \in \mathcal{T}_\mathcal{M}(\bar{x}) \text{ with } \mu(0) = 0 \text{ and } \dot{\mu}(0) = v.
\]
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. $F$ decreases quadratically along some direction $v \in \mathcal{T}_\mathcal{M}(\bar{x})$:

$$\frac{d^2}{dt^2} (F \circ \gamma)(t) \bigg|_{t=0} < 0$$

for some $C^2$ curve $\gamma \subset \mathcal{M}$ with $\gamma(0) = 0$ and $\dot{\gamma}(0) = v$. 

Although it may seem stringent, this property is generic: Theorem (Drusvyatskiy-Ioeee-Lewis '16, D-Drusvyatskiy '19) If $F$ is semi-algebraic, then for full Lebesgue measure set of perturbations $v \in \mathcal{R}^d$ every critical point of $F_v(x) = F(x)$ is either an active strict saddle or a local minimizer.
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. $F$ decreases quadratically along some direction $v \in T_{\mathcal{M}}(\bar{x})$:

$$\left. \frac{d^2}{dt^2} (F \circ \gamma)(t) \right|_{t=0} < 0$$

for some $C^2$ curve $\gamma \subset \mathcal{M}$ with $\gamma(0) = 0$ and $\dot{\gamma}(0) = v$.

Although it may seem stringent, this property is generic:

**Theorem (Drusvyatskiy-Ioffe-Lewis ’16, D-Drusvyatskiy ’19)**

If $F$ is semi-algebraic, 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.
Avoiding active strict saddles

**Question:** Do simple iterative methods avoid active strict saddles?

**Common iterative methods** take form

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

for simpler *nonsmooth strongly convex models* \( F_x \) of \( F \).

**Examples:**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Objective ( F )</th>
<th>Update function ( F_x(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prox-point</td>
<td>( r(x) )</td>
<td>( r(y) + \frac{1}{2\eta} | y - x |^2 )</td>
</tr>
<tr>
<td>Prox-gradient</td>
<td>( f(x) + r(x) )</td>
<td>( f(x) + \langle \nabla f(x), y - x \rangle + r(y) + \frac{1}{2\eta} | y - x |^2 )</td>
</tr>
<tr>
<td>Prox-linear</td>
<td>( h(c(x)) + r(x) )</td>
<td>( h(c(x) + \nabla c(x)(y - x)) + r(y) + \frac{1}{2\eta} | y - x |^2 )</td>
</tr>
</tbody>
</table>

*Table:* \( h \) is convex and Lipschitz, \( r \) is *weakly convex*\(^3\), and \( f \) and \( c \) are \( C^2 \)-smooth.

\(^3\)The function \( x \mapsto r(x) + \frac{\eta}{2} \| x \|^2 \) is convex.
Avoiding active strict saddles

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

**Theorem:** (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.

**Corollary:** (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?
Avoiding active strict saddles

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

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

**Example (Prox-point Method):**

By def’n of active manifold $\mathcal{M}$, $S$ maps nbhd of $\bar{x}$ into $\mathcal{M}$, so

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

\footnote{Lemaréchal-Sagastizábal '97}
Avoiding active strict saddles

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

**Example (Prox-point Method):**
By def’n of active manifold $\mathcal{M}$, $S$ maps nbhd of $\bar{x}$ into $\mathcal{M}$, so

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

Then

1. Weak convexity + classical perturbation theory $\implies$ $S$ is $C^1$ near $\bar{x}$.\(^4\)
2. Some computation shows $\nabla S(\bar{x})$ has real EigVal strictly greater than 1

\(^4\)Lemaréchal-Sagastizábal ’97
Avoiding active strict saddles

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

Example (Prox-point Method):

By def’n of active manifold $\mathcal{M}$, $S$ maps nbhd of $\bar{x}$ into $\mathcal{M}$, so

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

Then

1. Weak convexity + classical perturbation theory $\implies S$ is $C^1$ near $\bar{x}$.\(^4\)
2. Some computation shows $\nabla S(\bar{x})$ has real EigVal strictly greater than 1

Calculation is more interesting/surprising for prox-gradient and prox-linear.

\(^4\)Lemaréchal-Sagastizábal ’97
Complexity

**Question**: What is complexity of active strict saddle avoidance?
Complexity

**Question:** What is complexity of active strict saddle avoidance?

**Theorem:** (Du-Jin-Lee-Jordan-Poczos-Singh '17)

GD may take exponential time to avoid saddle points.

\[ x_{t+1} = x_t + \frac{1}{\lambda} (\nabla F(x_t) + u_t) \]

where \( u_t \geq \text{Unif}(rB) \).

**Idea:** Generalize perturbed methods to nonsmooth losses?

**Problem:** Proof requires \( C^2, 1 \)-smooth function.
**Complexity**

**Question:** What is complexity of active strict saddle avoidance?

**Theorem:** (Du-Jin-Lee-Jordan-Poczos-Singh ’17)

GD may take exponential time to avoid saddle points.

**Theorem:** (Jin-Netrapalli-Ge-Kakade-Jordan ’19)

Perturbed GD avoids saddle points in time poly\((1/\varepsilon)\).

\[
x_{t+1} = x_t - \alpha_t (\nabla F(x_t) + u_t) \quad \text{where} \quad u_t \sim \text{Unif}(rB).
\]
Complexity

**Question:** What is complexity of active strict saddle avoidance?

**Theorem:** (Du-Jin-Lee-Jordan-Poczos-Singh ’17)

GD may take exponential time to avoid saddle points.

**Theorem:** (Jin-Netrapalli-Ge-Kakade-Jordan ’19)

Perturbed GD avoids saddle points in time \( \text{poly}(1/\varepsilon) \).

\[
x_{t+1} = x_t - \alpha_t(\nabla F(x_t) + u_t) \quad \text{where} \quad u_t \sim \text{Unif}(rB).
\]

**Idea:** Generalize perturbed methods to nonsmooth losses?

**Problem:** Proof requires \( C^{2,1} \)-smooth function.
A smoothing approach: the Moreau envelope

Assumption: $F$ is $\rho$-weakly convex, i.e., $x \mapsto F(x) + \frac{\rho}{2} \|x\|^2$ is convex.
A smoothing approach: the Moreau envelope

**Assumption:** $F$ is $\rho$-weakly convex, i.e., $x \mapsto F(x) + \frac{\rho}{2} \|x\|^2$ is convex.

$$F_\lambda(x) := \inf_y \left\{ F(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$$

$$\nabla F_\lambda(x) = \lambda^{-1} (x - \text{prox}_{\lambda F}(x))$$

where $\text{prox}_{\lambda F}(x) = \arg\min_y \left\{ F(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$
A smoothing approach: the Moreau envelope

Assumption: $F$ is $\rho$-weakly convex, i.e., $x \mapsto F(x) + \frac{\rho}{2} \|x\|^2$ is convex.

$$F_\lambda(x) := \inf_y \left\{ F(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$$

$$\nabla F_\lambda(x) = \lambda^{-1} (x - \text{prox}_{\lambda F}(x))$$

where $\text{prox}_{\lambda F}(x) = \arg\min_y \left\{ F(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$

Properties: For $\lambda < \rho^{-1}$, $F_\lambda$ is $C^1$ and
1. $F_\lambda$ and $F$ share local minimizers;
2. If $F$ is $C^3$ along active manifolds, $F_\lambda$ is $C^3$ near active strict saddles and strict saddles of $F_\lambda \leftrightarrow$ active strict saddles of $F$
3. Estimate $\nabla F_\lambda$ by apxly. solving strongly convex problem “efficiently.”
A smoothing approach: the Moreau envelope

Assumption: $F$ is $\rho$-weakly convex, i.e., $x \mapsto F(x) + \frac{\rho}{2} \|x\|^2$ is convex.

$$F_\lambda(x) := \inf_y \left\{ F(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$$

$$\nabla F_\lambda(x) = \lambda^{-1} (x - \text{prox}_{\lambda F}(x))$$

where $\text{prox}_{\lambda F}(x) = \arg\min_y \left\{ F(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$

Properties: For $\lambda < \rho^{-1}$, $F_\lambda$ is $C^1$ and
1. $F_\lambda$ and $F$ share local minimizers;
2. If $F$ is $C^3$ along active manifolds, $F_\lambda$ is $C^3$ near active strict saddles and strict saddles of $F_\lambda$ $\longleftrightarrow$ active strict saddles of $F$;
3. Estimate $\nabla F_\lambda$ by apxly. solving strongly convex problem “efficiently.”

Approach: Perturbed inexact GD on Moreau

$$x_{t+1} \approx x_t - \alpha_t (\nabla F_\lambda(x_t) + u_t) \quad \text{where } u_t \sim \text{Unif}(rB).$$
Complexity?

**Theorem:** (D-Diaz-Drusvyatskiy '21)\(^5\)

Perturbed inexact GD on \(F_\lambda\) finds point \(x\) satisfying

\[
\|\nabla F_\lambda(x)\| \leq \epsilon_1 \quad \text{and} \quad \lambda_{\text{min}}(\nabla^2 F_\lambda(x)) \geq -\epsilon_2
\]

with complexity:

<table>
<thead>
<tr>
<th>Algorithm to Evaluate (\nabla F_\lambda(x))</th>
<th>Overall Algorithm Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prox-subgradient</td>
<td>(\tilde{O}(d \max{\epsilon_1^{-6}, \epsilon_2^{-6}, \epsilon_2^{-18}}))</td>
</tr>
<tr>
<td>Prox-gradient</td>
<td>(\tilde{O}(\max{\epsilon_1^{-2}, \epsilon_2^{-4}}))</td>
</tr>
<tr>
<td>Prox-linear</td>
<td>(\tilde{O}(\max{\epsilon_1^{-2}, \epsilon_2^{-4}}))</td>
</tr>
</tbody>
</table>

**Comments:**

1. Slow rate for prox-subgradient due to high accuracy eval. of \(\nabla F_\lambda\).

2. Prox-gradient/prox-linear rate matches perturbed GD.

3. Immediate extension to stochastic methods.

---

\(^5\)See also alternative approach of Huang '21.
Thank you!
References

- Proximal methods avoid active strict saddles of weakly convex functions

- Escaping strict saddle points of the Moreau envelope in nonsmooth optimization