Optimisation à points selles stricts

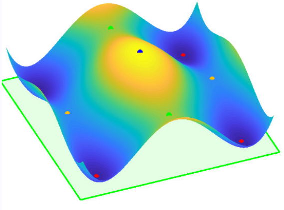
Panorama et opportunités

Clément Royer

ENSTA, 16 février 2026

Dauphine | PSL UNIVERSITÉ PARIS

PR[AI]RIE PaRis Artificial Intelligence Research InstitutE



Top: Pic du Midi d'Ossau (Royer '20), Bottom: Nonconvex landscape (Wright & Ma '22)

Nonconvex optimization (formally)

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$$

$f \in \mathcal{C}^2$, bounded below and **nonconvex**.

Nonconvex optimization (formally)

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$$

$f \in \mathcal{C}^2$, bounded below and **nonconvex**.

$\boldsymbol{x}^* \in \mathbb{R}^n$ is

- a global minimum of f if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \ \forall \boldsymbol{x} \in \mathbb{R}^n$
  $\rightarrow$ NP-hard to find/check! (Murti and Kabadi '87)

Nonconvex optimization (formally)

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\boldsymbol{x})$$

$f \in \mathcal{C}^2$, bounded below and **nonconvex**.

$\boldsymbol{x}^* \in \mathbb{R}^n$ is

- a global minimum of f if $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})\ \forall \boldsymbol{x} \in \mathbb{R}^n$
  $\rightarrow$ NP-hard to find/check! (Murti and Kabadi '87)
- a local minimum of f if $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})\ \forall \boldsymbol{x}$ close enough to $\boldsymbol{x}^*$
  $\rightarrow$ NP-hard to find/check!

Nonconvex optimization (formally)

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\boldsymbol{x})$$

$f \in \mathcal{C}^2$, bounded below and **nonconvex**.

$\boldsymbol{x}^* \in \mathbb{R}^n$ is

- a global minimum of f if $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})\ \forall \boldsymbol{x} \in \mathbb{R}^n$
  $\rightarrow$ NP-hard to find/check! (Murti and Kabadi '87)
- a local minimum of f if $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})\ \forall \boldsymbol{x}$ close enough to $\boldsymbol{x}^*$
  $\rightarrow$ NP-hard to find/check!
- a first-order stationary point of f if $\|\nabla f(\boldsymbol{x}^*)\| = 0$
  $\rightarrow$ PLS-complete to find (Hollender and Zampetakis '25).

Nonconvex optimization (formally)

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$$

$f \in \mathcal{C}^2$, bounded below and **nonconvex**.

$\boldsymbol{x}^* \in \mathbb{R}^n$ is

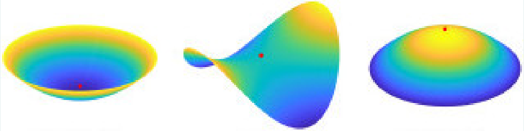
- a global minimum of f if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ $\forall \boldsymbol{x} \in \mathbb{R}^n$
  $\rightarrow$ NP-hard to find/check! (Murti and Kabadi '87)
- a local minimum of f if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ $\forall \boldsymbol{x}$ close enough to $\boldsymbol{x}^*$
  $\rightarrow$ NP-hard to find/check!
- a first-order stationary point of f if $\|\nabla f(\boldsymbol{x}^*)\| = 0$
  $\rightarrow$ PLS-complete to find (Hollender and Zampetakis '25).
- a second-order stationary point if $\|\nabla f(\boldsymbol{x}^*)\| = 0$ and $\nabla^2 f(\boldsymbol{x}^*) \succeq 0$
  $\rightarrow$ Already NP-hard for quartic polynomials.

Approximately solving $\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$

Goal: Reach $(\epsilon_{\mathrm{g}}, \epsilon_{\mathrm{H}})$-points

$$\|\nabla f(\boldsymbol{x})\| \le \epsilon_{\mathrm{g}} \quad \text{and} \quad \nabla^2 f(\boldsymbol{x}) \succeq -\epsilon_{\mathrm{H}}\, \boldsymbol{I}.$$

- $\epsilon_{\mathrm{g}} = \epsilon_{\mathrm{H}} = 0$: Second-order stationary point.
- $\epsilon_{\mathrm{g}} = 0$ but second condition false: Saddle points or maxima.

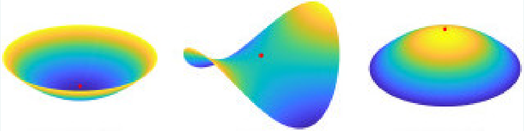


Approximately solving $\text{minimize}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$

Goal: Reach (ϵ_g, ϵ_H)-points

$$\|\nabla f(\mathbf{x})\| \le \epsilon_g \quad \text{and} \quad \nabla^2 f(\mathbf{x}) \succeq -\epsilon_H \, \mathbf{I}.$$

- $\epsilon_g = \epsilon_H = 0$: Second-order stationary point.
- $\epsilon_g = 0$ but second condition false: Saddle points or maxima.



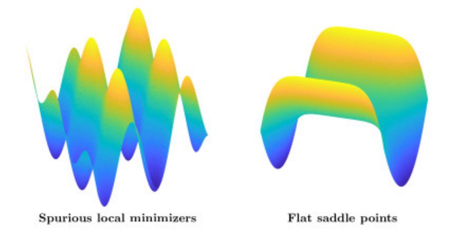
Complexity analysis

Cost of an algorithm to reach an (ϵ_g, ϵ_H)-point:

- Faster means lower cost in terms of ϵ_g/ϵ_H.
- Quality of (ϵ_g, ϵ_H)-points depend on the function landscape.

Bad nonconvex instances

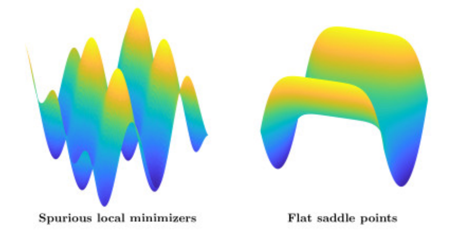
- Local, non-global minima.
- High-order saddle points.



Nonconvex optimization: A matter of landscape

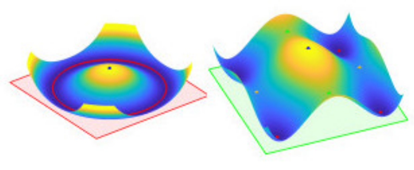
Bad nonconvex instances

- Local, non-global minima.
- High-order saddle points.



Nice instances

- All local minima are global.
- Strict (non-flat) saddle points.



Figures: J. Wright and Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models*, 2022.

What this talk is about

Nonconvex optimization

- Go beyond the (hard) general setting.
- Analyze algorithms from complexity viewpoint.

What this talk is about

Nonconvex optimization

- Go beyond the (hard) general setting.
- Analyze algorithms from complexity viewpoint.

Strict saddle problems

- A class of nonconvex problems with good landscape.
- Examples in data science.

What this talk is about

Nonconvex optimization

- Go beyond the (hard) general setting.
- Analyze algorithms from complexity viewpoint.

Strict saddle problems

- A class of nonconvex problems with good landscape.
- Examples in data science.

Trust-region methods

- Standard framework for nonconvex optimization.
- Theory and implementation revisited in the strict saddle setting.

Outline

Sommaire

Benign landscape property

Definition

$f : \mathbb{R}^n \times \mathbb{R}\ \mathcal{C}^2$ has **benign landscape** if

$$\|\nabla f(\boldsymbol{x})\| = 0 \text{ and } \nabla^2 f(\boldsymbol{x}) \succeq 0 \quad \Leftrightarrow \quad \boldsymbol{x} \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}).$$

- Every local minimum is global.
- All saddle points satisfy $\nabla^2 f(\boldsymbol{x}) \not\succeq 0 \rightarrow$ Strict saddle points.

Benign landscape property

Definition

$f : \mathbb{R}^n \times \mathbb{R}\ \mathcal{C}^2$ has **benign landscape** if

$$\|\nabla f(\boldsymbol{x})\| = 0 \text{ and } \nabla^2 f(\boldsymbol{x}) \succeq 0 \quad \Leftrightarrow \quad \boldsymbol{x} \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}).$$

- Every local minimum is global.
- All saddle points satisfy $\nabla^2 f(\boldsymbol{x}) \not\succeq 0 \rightarrow$ Strict saddle points.

Typical sources of benign nonconvex landscape

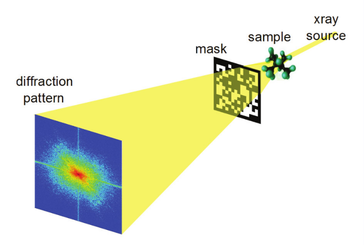
- Nonconvex formulations with symmetries.
  $\rightarrow$ Multiple (redundant) solutions, easy to go from one to another.
- Overparameterized models.
  $\rightarrow$ Introduces saddle points but those are strict.

Example 1: Phase retrieval (thanks to Irène Waldspurger)

Phase retrieval problem

$$\begin{array}{ll} \text{find} & \mathbf{x} \in \mathbb{C}^n \\ \text{such that} & |\langle \mathbf{x}, \mathbf{a}_i \rangle| = b_i, \quad i = 1, \dots, m. \end{array}$$

- $b_i = |\langle \mathbf{x}^*, \mathbf{a}_i \rangle|$: Complex moduli.
- $\mathbf{a}_i$: Measurement vectors (Gaussian, wavelets, etc).



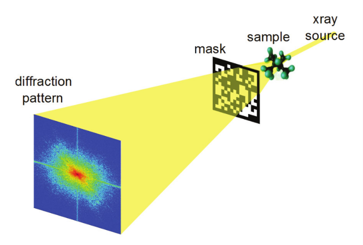
(Picture from Candès et al '15)

Example 1: Phase retrieval (thanks to Irène Waldspurger)

Phase retrieval problem

$$\begin{array}{ll} \text{find} & \mathbf{x} \in \mathbb{C}^n \\ \text{such that} & |\langle \mathbf{x}, \mathbf{a}_i \rangle| = b_i, \quad i = 1, \dots, m. \end{array}$$

- $b_i = |\langle \mathbf{x}^*, \mathbf{a}_i \rangle|$: Complex moduli.
- $\mathbf{a}_i$: Measurement vectors (Gaussian, wavelets, etc).



(Picture from Candès et al '15)

Gives rise to convex and **nonconvex** optimization problems!

Example 1: Phase retrieval ('ed)

Recall phase retrieval: Find $\boldsymbol{x} \in \mathbb{C}^n$ such that $|\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle| = b_i \ \forall i$.

PhaseLift SDP relaxation (Candès et al. '11)

$$
\begin{array}{ll}
\min_{\boldsymbol{X} \in \mathbb{C}^{n \times n}} & \mathrm{trace}(\boldsymbol{X}) \\
\text{subject to} & \langle \boldsymbol{X}, \boldsymbol{a}_i \boldsymbol{a}_i^* \rangle = b_i \ \forall i \\
& \boldsymbol{X} \succeq 0
\end{array}
$$

$\rightarrow$ Convex problem, can be solved in polynomial time.

$\rightarrow$ In practice, cost of matrix linear algebra prohibitive.

Example 1: Phase retrieval ('ed)

Recall phase retrieval: Find $\boldsymbol{x} \in \mathbb{C}^n$ such that $|\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle| = b_i \ \forall i$.

PhaseLift SDP relaxation (Candès et al. '11)

$$
\begin{array}{ll}
\min_{\boldsymbol{X} \in \mathbb{C}^{n \times n}} & \mathrm{trace}(\boldsymbol{X}) \\
\text{subject to} & \langle \boldsymbol{X}, \boldsymbol{a}_i \boldsymbol{a}_i^* \rangle = b_i \ \forall i \\
& \boldsymbol{X} \succeq 0
\end{array}
$$

→ Convex problem, can be solved in polynomial time.

→ In practice, cost of matrix linear algebra prohibitive.

Factored PhaseLift formulation

$$
\min_{\boldsymbol{U} \in \mathbb{C}^{n \times r}} f(\boldsymbol{U}) := \frac{1}{2} \| \mathcal{A}(\boldsymbol{U}\, \boldsymbol{U}^{\mathrm{T}}) - \boldsymbol{b} \|_F^2, \qquad \mathcal{A}(\boldsymbol{X}) = [\langle \boldsymbol{X}, \boldsymbol{a}_i\, \boldsymbol{a}_i^* \rangle]_i \,.
$$

→ Burer-Monteiro factorization (Burer and Monteiro '03).

→ Benign landscape for relatively small r (McRae '24).

Example 2: Linear neural networks

$$\underset{\boldsymbol{W}_1,\dots,\boldsymbol{W}_L}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}_L\boldsymbol{W}_{L-1}\cdots\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{A}-\boldsymbol{B}\|_F^2$$

- $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1}\times d_i}$.
- $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_0}$, $\boldsymbol{B} \in \mathbb{R}^{d_{L+1}\times d_0}$.
- $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{ij}^2$.

Example 2: Linear neural networks

$$\underset{\boldsymbol{W}_1,\dots,\boldsymbol{W}_L}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}_L\boldsymbol{W}_{L-1}\cdots\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{A}-\boldsymbol{B}\|_F^2$$

- $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1}\times d_i}$.
- $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_0}$, $\boldsymbol{B} \in \mathbb{R}^{d_{L+1}\times d_0}$.
- $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{ij}^2$.

- Also called deep matrix factorization.
- Initially used to better understand neural networks.
- Numerous landscape results, especially between 2016-2022.

Example 2: linear networks ('ed)

Case $L = 1$ (One-layer)

$$\operatorname*{minimize}_{\boldsymbol{W}_1} \frac{1}{2}\|\boldsymbol{W}_1\boldsymbol{A} - \boldsymbol{B}\|_F^2$$

- Convex problem!
- Explicit form of a solution.

Example 2: linear networks ('ed)

Case $L = 1$ (One-layer)

$$\underset{\boldsymbol{W}_1}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}_1 \boldsymbol{A} - \boldsymbol{B}\|_F^2$$

- Convex problem!
- Explicit form of a solution.

Case $L = 2$ (two-layer network)

$$\underset{\substack{\boldsymbol{W}_1 \in \mathbb{R}^{d_2 \times d_1} \\ \boldsymbol{W}_2 \in \mathbb{R}^{d_3 \times d_2}}}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}_2 \boldsymbol{W}_1 \boldsymbol{A} - \boldsymbol{B}\|_F^2$$

- If $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ full rank, all local minima are global (optimum is 0 when $d_2 \geq \max\{d_1, d_3\}$).
- All saddle points are strict.

Example 2: Linear networks ('ed)

Deep linear networks

$$\underset{\boldsymbol{W}_1,\dots,\boldsymbol{W}_L}{\text{minimize}} \frac{1}{2} \|\boldsymbol{W}_L \boldsymbol{W}_{L-1} \cdots \boldsymbol{W}_2 \boldsymbol{W}_1 \boldsymbol{A} - \boldsymbol{B}\|_F^2$$

- $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1} \times d_i}$.
- $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_0}$, $\boldsymbol{B} \in \mathbb{R}^{d_{L+1} \times d_0}$.
- $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{ij}^2$.

Example 2: Linear networks ('ed)

Deep linear networks

$$\underset{\boldsymbol{W}_1,\dots,\boldsymbol{W}_L}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}_L\boldsymbol{W}_{L-1}\cdots\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{A}-\boldsymbol{B}\|_F^2$$

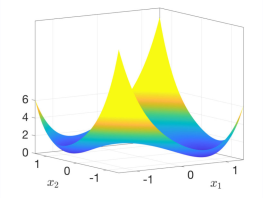
- $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1}\times d_i}$.
- $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_0}$, $\boldsymbol{B} \in \mathbb{R}^{d_{L+1}\times d_0}$.
- $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{ij}^2$.

- Full characterization of the landscape possible (Kawaguchi '16, Achour et al '24).
- IF all dimensions are equal and $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ full rank, all local minima are global!

Other examples (pictures from Chi et al '19; Wright and Ma '22)

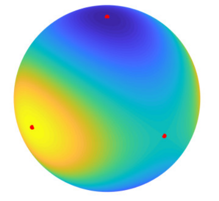
Rank-1 approximation

$$\operatorname*{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{4}\|\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{M}\|_F^2$$



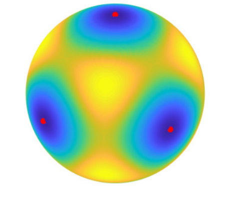
Minimum Eigenvalue

$$\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^{\mathrm{T}} A \boldsymbol{x}$$



Tensor optimization

$$\min_{\|\boldsymbol{x}\|=1} T(\boldsymbol{x},\boldsymbol{x},\boldsymbol{x},\boldsymbol{x})$$



For more: https://sunju.org/research/nonconvex/

Defining "good nonconvexity" mathematically

What is a "nice" nonconvex problem?

- All local minima are global?
- All saddle points can be escaped?
- Algorithms work well?
- All of the above?

Defining "good nonconvexity" mathematically

What is a "nice" nonconvex problem?

- All local minima are global?
- All saddle points can be escaped?
- Algorithms work well?
- All of the above?

Numerous existing results

- PŁ,KŁ, quadratic growth, error bound (Rebjock and Boumal '24);
- Regional complexity analysis (Curtis and Robinson '21);
- Phase complexity (Cartis, Gould, Toint '22).

Defining “good nonconvexity” mathematically

What is a “nice” nonconvex problem?

- All local minima are global?
- All saddle points can be escaped?
- Algorithms work well?
- All of the above?

Numerous existing results

- PŁ,KŁ, quadratic growth, error bound (Rebjock and Boumal ’24);
- Regional complexity analysis (Curtis and Robinson ’21);
- Phase complexity (Cartis, Gould, Toint ’22).

Our focus: Strict saddle property

- Parametric definition.
- Allows for complexity analysis.

Starting point: Strongly convex functions

- $f : \mathbb{R}^n \to \mathbb{R}$ $\mathcal{C}^2$ is γ-strongly convex if

$$\nabla^2 f(\boldsymbol{x}) \succeq \gamma I.$$

- Unique minimum $\boldsymbol{x}^* \in \mathbb{R}^n$.

One consequence of γ-strong convexity

For any $\alpha > 0$,

$$\|\nabla f(\boldsymbol{x})\| \leq \alpha \quad \Rightarrow \quad \|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \frac{2\alpha}{\gamma}.$$

Strict saddle functions

Definition (Ge et al '17; Goyens, R. '24)

$f : \mathbb{R}^n \to \mathbb{R}$ $\mathcal{C}^2$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle with $\alpha, \beta, \gamma, \delta > 0$ if for any $\boldsymbol{x} \in \mathbb{R}^n$, one of these properties holds:

1. $\|\nabla f(\boldsymbol{x})\| \geq \alpha$;
2. $\lambda_{\min}\left(\nabla^2 f(\boldsymbol{x})\right) \leq -\beta$;
3. There exists $\boldsymbol{x}^*$ local minimum of f such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \delta \quad \text{and} \quad \nabla^2 f(\boldsymbol{y}) \succeq \gamma \boldsymbol{I} \succ 0 \quad \forall \boldsymbol{y},\ \|\boldsymbol{y} - \boldsymbol{x}^*\| \leq 2\delta.$$

Strict saddle functions

Definition (Ge et al '17; Goyens, R. '24)

$f : \mathbb{R}^n \to \mathbb{R}$ $\mathcal{C}^2$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle with $\alpha, \beta, \gamma, \delta > 0$ if for any $\boldsymbol{x} \in \mathbb{R}^n$, one of these properties holds:

1. $\|\nabla f(\boldsymbol{x})\| \geq \alpha$;
2. $\lambda_{\min}\left(\nabla^2 f(\boldsymbol{x})\right) \leq -\beta$;
3. There exists $\boldsymbol{x}^*$ local minimum of f such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \delta \quad \text{and} \quad \nabla^2 f(\boldsymbol{y}) \succeq \gamma I \succ 0 \quad \forall \boldsymbol{y},\ \|\boldsymbol{y} - \boldsymbol{x}^*\| \leq 2\delta.$$

Interpretation: 3 regions in the space

1. Large gradient.
2. Negative curvature for Hessian.
3. Near minimum+strong convexity.

Simple strict saddle functions

Strict saddle: $\{\|\nabla f(x)\| \geq \alpha\}$ or $\{\nabla^2 f(x) \not\succeq -\beta I\}$ or $\{\|x - x^*\| \leq \delta$ and $\nabla^2 f(y) \succeq \gamma I \ \forall y: \ \|y - x^*\| \leq 2\delta\}$.

Strongly convex⇒Strict saddle!

If f is γ-strongly convex, then it is $\left(\alpha, \beta, \gamma, \frac{2\alpha}{\gamma}\right)$-strict saddle for any positive α and β.

Simple strict saddle functions

Strict saddle: $\{\|\nabla f(x)\| \geq \alpha\}$ or $\{\nabla^2 f(x) \not\succeq -\beta I\}$ or $\{\|x - x^*\| \leq \delta$ and $\nabla^2 f(y) \succeq \gamma I \ \forall y : \ \|y - x^*\| \leq 2\delta\}$.

Strongly convex⇒Strict saddle!

If f is γ-strongly convex, then it is $\left(\alpha, \beta, \gamma, \frac{2\alpha}{\gamma}\right)$-strict saddle for any positive α and β.

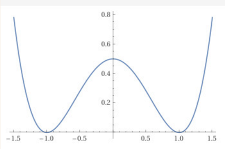
Strict saddle: $\{\|\nabla f(x)\| \geq \alpha\}$ **or** $\{\nabla^2 f(x) \not\succeq -\beta I\}$ **or** $\{\|x - x^*\| \leq \delta$ **and** $\nabla^2 f(y) \succeq \gamma I \ \forall y : \ \|y - x^*\| \leq 2\delta\}$.

Strongly convex⇒Strict saddle!

If f is γ-strongly convex, then it is $\left(\alpha, \beta, \gamma, \frac{2\alpha}{\gamma}\right)$-strict saddle for any positive α and β.

A "truly" strict saddle function

$$f : x \in \mathbb{R} \mapsto \frac{1}{2}(x^2 - 1)^2$$



Example: Matrix completion

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} f(\boldsymbol{U},\boldsymbol{V}) := \left\|\mathcal{P}_\Omega(\boldsymbol{U}\boldsymbol{V}^\mathrm{T} - \boldsymbol{M})\right\|_F^2, \qquad \Omega \subset [n]\times[m].$$

Assume Nice structure for $\boldsymbol{M}$ (incoherence), probability of sampling $\boldsymbol{M}_{ij}$ large enough.

Example: Matrix completion

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} f(\boldsymbol{U},\boldsymbol{V}) := \left\|\mathcal{P}_\Omega(\boldsymbol{U}\boldsymbol{V}^\mathrm{T}-\boldsymbol{M})\right\|_F^2, \qquad \Omega \subset [n]\times[m].$$

Assume Nice structure for $\boldsymbol{M}$ (incoherence), probability of sampling $\boldsymbol{M}_{ij}$ large enough.

Theorem (Ge et al. '17)

Let $(\boldsymbol{U},\boldsymbol{V}) \in \mathbb{R}^{n\times r}\times\mathbb{R}^{m\times r}$. Then, there exists $\alpha > 0$ such that one of these cases occur

1. $\|\nabla f(\boldsymbol{U},\boldsymbol{V})\| \geq \alpha$
2. The Hessian at U, V has negative curvature, i.e.

$$\lambda_{\min}\left(\nabla^2 f(\boldsymbol{U},\boldsymbol{V})\right) < -\mathcal{O}(\sigma_{\min}(M))$$

3. $(\boldsymbol{U},\boldsymbol{V})$ is at distance at most $\mathcal{O}(\frac{\alpha}{\sigma_{\min}(\boldsymbol{M})})$ from a global minimum.

Another example from phase retrieval

Nonconvex vector formulation (Sun et al '18)

$$\underset{\boldsymbol{x}\in\mathbb{C}^n}{\text{minimize}} \;\; f(\boldsymbol{x}) = \frac{1}{2m}\sum_{i=1}^{m}(b_i^2 - |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2)^2$$

with $\{\boldsymbol{a}_i\}$ Gaussian, $m = \mathcal{O}(n \log^3(n))$.

Another example from phase retrieval

Nonconvex vector formulation (Sun et al '18)

$$\underset{\boldsymbol{x} \in \mathbb{C}^n}{\text{minimize}} \;\; f(\boldsymbol{x}) = \frac{1}{2m} \sum_{i=1}^{m} \left(b_i^2 - |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2\right)^2$$

with $\{\boldsymbol{a}_i\}$ Gaussian, $m = \mathcal{O}(n \log^3(n))$.

Another example from phase retrieval

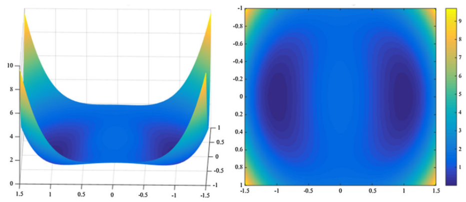
Nonconvex vector formulation (Sun et al '18)

$$\underset{\boldsymbol{x}\in\mathbb{C}^n}{\text{minimize}} \;\; f(\boldsymbol{x}) = \frac{1}{2m}\sum_{i=1}^{m}\left(b_i^2 - |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2\right)^2$$

with $\{\boldsymbol{a}_i\}$ Gaussian, $m = \mathcal{O}(n\log^3(n))$.

→ For some universal $c > 0$, f is $\left(\frac{c}{n\log(m)}, c, c, \frac{c}{n\log(m)}\right)$-strict saddle.

→ Generalized definition of strict saddle (more later).



Sommaire

Problem setup

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\boldsymbol{x})$$

$f \in \mathcal{C}^2$ bounded below and nonconvex.

Problem setup

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\boldsymbol{x})$$

$f \in \mathcal{C}^2$ bounded below and nonconvex.

Goal: Reach $(\epsilon_{\mathrm{g}}, \epsilon_{\mathrm{H}})$-point

$$\|\nabla f(\boldsymbol{x})\| \le \epsilon_{\mathrm{g}} \quad \text{and} \quad \nabla^2 f(\boldsymbol{x}) \succeq -\epsilon_{\mathrm{H}}\, \boldsymbol{I}.$$

- General nonconvex case: Close to second-order stationary point.
- Strict saddle setting: Close to a global minimum!

Algorithms for strict saddle problems

- **Gradient descent** Can be shown to escape strict saddle points almost surely (Lee et al '19).
- **Perturbed gradient descent** does so in polynomial time (Jin et al '17, Ma et al '25).
- **Line-search algorithms** can also be analyzed in this way (O'Neill and Wright '23).

Algorithms for strict saddle problems

- **Gradient descent** Can be shown to escape strict saddle points almost surely (Lee et al '19).
- **Perturbed gradient descent** does so in polynomial time (Jin et al '17, Ma et al '25).
- **Line-search algorithms** can also be analyzed in this way (O'Neill and Wright '23).

Our approach

- Use **trust-region** algorithms, originally developed for nonconvex problems.
- Exploit their applicability to manifold optimization.

Setup: Manifold optimization

Main problem

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathcal{M}} f(\boldsymbol{x}),$$

- f $\mathcal{C}^2$ nonconvex
- $\mathcal{M}$ Riemannian manifold ($\mathbb{C}^{n \times m}$, $\mathbb{R}^{n \times m}$, sphere,...).

Setup: Manifold optimization

Main problem

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathcal{M}} f(\boldsymbol{x}),$$

- f $\mathcal{C}^2$ nonconvex
- $\mathcal{M}$ Riemannian manifold ($\mathbb{C}^{n \times m}$, $\mathbb{R}^{n \times m}$, sphere,...).

Goal: Approximate stationary points

$$\|g(\boldsymbol{x}_k)\| \le \epsilon_{\mathrm{g}} \quad \text{and} \quad \lambda_{\min}\left(\mathcal{H}(\boldsymbol{x}_k)\right) \ge -\epsilon_{\mathrm{H}}.$$

- $g(\cdot)$ Riemannian gradient.
- $\mathcal{H}(\cdot)$ Riemannian Hessian.

Strict saddle property on manifold $\mathcal{M}$

Definition (Goyens, R. '24)

$f : \mathbb{R}^n \to \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle if for any $\boldsymbol{x} \in \mathbb{R}^n$, one of these properties holds:

1. $\|g(\boldsymbol{x})\| \geq \alpha$;
2. $\lambda_{\min}\left(\mathcal{H}(\boldsymbol{x})\right) \leq -\beta$;
3. There exists $\boldsymbol{x}^*$ local minimum of f such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \delta \quad \text{and} \quad \lambda_{\min}(\mathcal{H}(\boldsymbol{y})) \geq \gamma \quad \forall \boldsymbol{y},\ \|\boldsymbol{y} - \boldsymbol{x}^*\| \leq 2\delta.$$

Strict saddle property on manifold $\mathcal{M}$

Definition (Goyens, R. '24)

$f : \mathbb{R}^n \to \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle if for any $\boldsymbol{x} \in \mathbb{R}^n$, one of these properties holds:

1. $\|g(\boldsymbol{x})\| \geq \alpha$;
2. $\lambda_{\min}\left(\mathcal{H}(\boldsymbol{x})\right) \leq -\beta$;
3. There exists $\boldsymbol{x}^*$ local minimum of f such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \delta \quad \text{and} \quad \lambda_{\min}(\mathcal{H}(\boldsymbol{y})) \geq \gamma \quad \forall \boldsymbol{y},\ \|\boldsymbol{y} - \boldsymbol{x}^*\| \leq 2\delta.$$

Interpretation: 3 regions in the space

1. Large Riemannian gradient g.
2. Negative curvature for Riemannian Hessian $\mathcal{H}$.
3. Near minimum+geodesic strong convexity.

N.B. Already studied for special problem classes (Pumir et al '18, Sun et al '16-'18).

Trust region for $\min_{\mathbf{x} \in \mathcal{M}} f(\mathbf{x})$

Inputs: $\mathbf{x}_0 \in \mathcal{M}$, $\Delta_0 > 0$, $\eta > 0$.

For k$=0, 1, 2, \dots$

1. Define $m_k(\mathbf{x}_k + \mathbf{s}) := \langle g(\mathbf{x}_k), \mathbf{s} \rangle + \frac{1}{2} \langle \mathbf{s}, \mathcal{H}(\mathbf{x}_k)\mathbf{s} \rangle$ and compute
$$\mathbf{s}_k \in \underset{\substack{\mathbf{s} \in \mathcal{T}^{\mathcal{M}}_{\mathbf{x}_k} \\ \|\mathbf{s}\| \le \Delta_k}}{\operatorname{argmin}} m_k(\mathbf{x}_k + \mathbf{s}).$$
2. Define $\mathbf{x}_k^{\mathcal{M}}$ as the retraction of $\mathbf{x}_k + \mathbf{s}_k$ onto $\mathcal{M}$.
3. Compute $\rho_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k^{\mathcal{M}})}{m_k(\mathbf{x}_k) - m_k(\mathbf{x}_k^{\mathcal{M}})}$.
4. If $\rho_k \ge \eta$, set $\mathbf{x}_{k+1} = \mathbf{x}_k^{\mathcal{M}}$ and $\Delta_{k+1} = 2\Delta_k$.
5. Otherwise, set $\mathbf{x}_{k+1} = \mathbf{x}_k$ and $\Delta_{k+1} = 0.5\Delta_k$.

Complexity results for trust region

Goal: Compute $\boldsymbol{x}_K$ such that $\|g(\boldsymbol{x}_K)\| \le \epsilon_{\mathrm{g}}$ and $\lambda_{\min}(\mathcal{H}(\boldsymbol{x}_K)) \ge -\epsilon_{\mathrm{H}}$.

For general nonconvex f (Boumal et al '19)

$$K = \mathcal{O}\left(\max\{\epsilon_{\mathrm{g}}^{-2}, \epsilon_{\mathrm{H}}^{-3}\}\right).$$

Complexity results for trust region

Goal: Compute $\mathbf{x}_K$ such that $\|g(\mathbf{x}_K)\| \leq \epsilon_{\mathrm{g}}$ and $\lambda_{\min}(\mathcal{H}(\mathbf{x}_K)) \geq -\epsilon_{\mathrm{H}}$.

For general nonconvex f (Boumal et al '19)

$$K = \mathcal{O}\left(\max\{\epsilon_{\mathrm{g}}^{-2}, \epsilon_{\mathrm{H}}^{-3}\}\right).$$

For strict saddle f (Goyens and R. '24)

If f is $(\alpha, \beta, \gamma, \delta)$-strict saddle, $K = K_f + K_\epsilon$, with

$$K_f = \mathcal{O}\left(\max\left\{\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}\right\}\right)$$

$$K_\epsilon = \log\log\left[\mathcal{O}\left(\gamma\epsilon_{\mathrm{g}}^{-1}\right)\right].$$

Complexity results for trust region

Goal: Compute $\mathbf{x}_K$ such that $\|g(\mathbf{x}_K)\| \le \epsilon_{\mathrm{g}}$ and $\lambda_{\min}(\mathcal{H}(\mathbf{x}_K)) \ge -\epsilon_{\mathrm{H}}$.

For general nonconvex f (Boumal et al '19)

$$K = \mathcal{O}\left(\max\{\epsilon_{\mathrm{g}}^{-2}, \epsilon_{\mathrm{H}}^{-3}\}\right).$$

For strict saddle f (Goyens and R. '24)

If f is $(\alpha,\beta,\gamma,\delta)$-strict saddle, $K = K_f + K_\epsilon$, with

$$K_f = \mathcal{O}\left(\max\left\{\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}\right\}\right)$$

$$K_\epsilon = \log\log\left[\mathcal{O}\left(\gamma\epsilon_{\mathrm{g}}^{-1}\right)\right].$$

- $\log\log$ dependency in ϵ_{g} (none on ϵ_{H})!
- Complexity depends more on landscape parameters!

Adding inexact steps

What we have so far

- Trust-region method with good complexity.
- Requires exact step computation.
- Agnostic to the function landscape.

Adding inexact steps

What we have so far

- Trust-region method with good complexity.
- Requires exact step computation.
- Agnostic to the function landscape.

Moving further with inexactness

- Practical trust-region use inexact steps.
- Never store the Hessian matrix ("matrix-free")!

Can we build an inexact variant that leverages the strict saddle property?

Inexact steps in trust region

Trust-region subproblem

$$\underset{\boldsymbol{s} \in \mathcal{T}_{\boldsymbol{x}_k}^{\mathcal{M}}}{\text{minimize}} < g(\boldsymbol{x}_k), \boldsymbol{s} > + \frac{1}{2} < \boldsymbol{s}, \mathcal{H}(\boldsymbol{x}_k)\boldsymbol{s} > \quad \text{s.t.} \quad \|\boldsymbol{s}\| \leq \Delta_k.$$

- Apply conjugate gradient (CG) to the linear system $\mathcal{H}(\boldsymbol{x}_k)\boldsymbol{s} = -g(\boldsymbol{x}_k)$;
- Stop when residual $\|\mathcal{H}(\boldsymbol{x}_k)\boldsymbol{s} + g(\boldsymbol{x}_k)\|$ is small enough or the $\|\boldsymbol{s}\| = \Delta_k$;
- For $\mathcal{H}(\boldsymbol{x}_k) \not\succeq 0$: if **negative curvature** is encountered, take a negative curvature step such that $\|\boldsymbol{s}\| = \Delta_k$.

Inexact steps in trust region

Trust-region subproblem

$$\operatorname*{minimize}_{\boldsymbol{s} \in \mathcal{T}_{\boldsymbol{x}_k}^{\mathcal{M}}} < g(\boldsymbol{x}_k), \boldsymbol{s} > + \frac{1}{2} < \boldsymbol{s}, \mathcal{H}(\boldsymbol{x}_k)\boldsymbol{s} > \quad \text{s.t.} \quad \|\boldsymbol{s}\| \leq \Delta_k.$$

- Apply conjugate gradient (CG) to the linear system $\mathcal{H}(\boldsymbol{x}_k)\boldsymbol{s} = -g(\boldsymbol{x}_k)$;
- Stop when residual $\|\mathcal{H}(\boldsymbol{x}_k)\boldsymbol{s} + g(\boldsymbol{x}_k)\|$ is small enough or the $\|\boldsymbol{s}\| = \Delta_k$;
- For $\mathcal{H}(\boldsymbol{x}_k) \not\succeq 0$: if **negative curvature** is encountered, take a negative curvature step such that $\|\boldsymbol{s}\| = \Delta_k$.

Changes (for complexity)

- Add a cap on the number of CG iterations.
- Guarantee negative curvature detection.

Inexact steps for strict saddle functions

$f : \mathbb{R}^n \to \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle if for any $\mathbf{x} \in \mathbb{R}^n$, one of these properties holds:

1. $\|g(\mathbf{x})\| \geq \alpha$;
2. $\lambda_{\min}(\mathcal{H}(\mathbf{x})) \leq -\beta$;
3. There exists $\mathbf{x}^*$ local minimum of f such that

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \delta \quad \text{and} \quad \lambda_{\min}(\mathcal{H}(\mathbf{y})) \geq \gamma \quad \forall \mathbf{y},\ \|\mathbf{y} - \mathbf{x}^*\| \leq 2\delta.$$

Inexact steps for strict saddle functions

$f : \mathbb{R}^n \to \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle if for any $\boldsymbol{x} \in \mathbb{R}^n$, one of these properties holds:

1. $\|g(\boldsymbol{x})\| \geq \alpha$;
2. $\lambda_{\min}\left(\mathcal{H}(\boldsymbol{x})\right) \leq -\beta$;
3. There exists $\boldsymbol{x}^*$ local minimum of f such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \delta \quad \text{and} \quad \lambda_{\min}(\mathcal{H}(\boldsymbol{y})) \geq \gamma \quad \forall \boldsymbol{y},\ \|\boldsymbol{y} - \boldsymbol{x}^*\| \leq 2\delta.$$

One step per strict saddle case

1. $\|g(\boldsymbol{x}_k)\| \geq \alpha$: Cheap gradient step (along $-g(\boldsymbol{x}_k)$).

Inexact steps for strict saddle functions

$f : \mathbb{R}^n \to \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle if for any $\boldsymbol{x} \in \mathbb{R}^n$, one of these properties holds:

1. $\|g(\boldsymbol{x})\| \geq \alpha$;
2. $\lambda_{\min}\left(\mathcal{H}(\boldsymbol{x})\right) \leq -\beta$;
3. There exists $\boldsymbol{x}^*$ local minimum of f such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \delta \quad \text{and} \quad \lambda_{\min}(\mathcal{H}(\boldsymbol{y})) \geq \gamma \quad \forall \boldsymbol{y}, \ \|\boldsymbol{y} - \boldsymbol{x}^*\| \leq 2\delta.$$

One step per strict saddle case

1. $\|g(\boldsymbol{x}_k)\| \geq \alpha$: Cheap gradient step (along $-g(\boldsymbol{x}_k)$).
2. Otherwise try the last case (CG should work fast!)

Inexact steps for strict saddle functions

$f : \mathbb{R}^n \to \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle if for any $\boldsymbol{x} \in \mathbb{R}^n$, one of these properties holds:

1. $\|g(\boldsymbol{x})\| \geq \alpha$;
2. $\lambda_{\min}(\mathcal{H}(\boldsymbol{x})) \leq -\beta$;
3. There exists $\boldsymbol{x}^*$ local minimum of f such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \delta \quad \text{and} \quad \lambda_{\min}(\mathcal{H}(\boldsymbol{y})) \geq \gamma \quad \forall \boldsymbol{y},\ \|\boldsymbol{y} - \boldsymbol{x}^*\| \leq 2\delta.$$

One step per strict saddle case

1. $\|g(\boldsymbol{x}_k)\| \geq \alpha$: Cheap gradient step (along $-g(\boldsymbol{x}_k)$).
2. Otherwise try the last case (CG should work fast!)
3. If CG does not work fast, then $\lambda_{\min}(\mathcal{H}(\boldsymbol{x}_k)) \leq -\beta$: Negative curvature step.

Our method: Capped conjugate gradient

Goal:

$$\min_{\boldsymbol{s}\in\mathcal{T}_{\boldsymbol{x}_k}^{\mathcal{M}}} < g(\boldsymbol{x}_k), \boldsymbol{s} > + \frac{1}{2} < \boldsymbol{s}, (\mathcal{H}(\boldsymbol{x}_k) + 2\gamma I)\boldsymbol{s} > \quad \text{s.t.} \quad \|\boldsymbol{s}\| \leq \Delta.$$

Theorem (Curtis, Robinson, R., Wright '21)

Suppose that we run CG for at most $J^{CG} = \min\{n, \tilde{\mathcal{O}}(\gamma^{-1/2})\}$ iterations/Hessian-vector products. Then,

- Either we compute a good enough step using CG...
- ...or we find a negative curvature direction for H...
- ...or we know that it exists and we can call a minimum eigenvalue oracle to find it.

Our method: Capped conjugate gradient

Goal:

$\min_{s \in \mathcal{T}_{x_k}^{\mathcal{M}}} < g(x_k), s > + \frac{1}{2} < s, (\mathcal{H}(x_k) + 2\gamma I)s >$ s.t. $\|s\| \leq \Delta$.

Theorem (Curtis, Robinson, R., Wright '21)

Suppose that we run CG for at most $J^{CG} = \min\{n, \tilde{\mathcal{O}}(\gamma^{-1/2})\}$ iterations/Hessian-vector products. Then,

- Either we compute a good enough step using CG...
- ...or we find a negative curvature direction for H...
- ...or we know that it exists and we can call a minimum eigenvalue oracle to find it.

Strict saddle setting

Suppose that $\|g(x_k)\| \leq \alpha$ and run CG for J^{CG} iterations. Then,

- Either the step is accurate enough
- or **we know that** $\lambda_{\min}(\mathcal{H}(x_k)) \leq -\beta I$ and we call a minimum eigenvalue oracle to find negative curvature.

Minimum eigenvalue oracle (MEO)

Given $\mathcal{H}(\boldsymbol{x}_k) \in \mathbb{R}^{n \times n}$, $\beta \in (0,1)$, *and* $\xi \in (0,1)$, output

1. A vector s such that

$$\boldsymbol{s}^{\mathrm{T}} \mathcal{H}(\boldsymbol{x}_k) \boldsymbol{s} \le -\frac{\beta}{2} \|\boldsymbol{s}\|^2.$$

2. **OR** a certificate that $\mathcal{H}(\boldsymbol{x}_k) \succ -\beta I$, valid with probability $1 - \xi$.

Minimum eigenvalue oracle (MEO)

Given $\mathcal{H}(\boldsymbol{x}_k) \in \mathbb{R}^{n \times n}$, $\beta \in (0,1)$, *and* $\xi \in (0,1)$, output

1. A vector s such that

$$\boldsymbol{s}^{\mathrm{T}} \mathcal{H}(\boldsymbol{x}_k) \boldsymbol{s} \le -\frac{\beta}{2} \|\boldsymbol{s}\|^2.$$

2. **OR** a certificate that $\mathcal{H}(\boldsymbol{x}_k) \succ -\beta I$, valid with probability $1 - \xi$.

An example of MEO

Run CG on $\mathcal{H}(\boldsymbol{x}_k)\boldsymbol{s} = \boldsymbol{b}$, $\boldsymbol{b}$ uniform on the unit sphere. produces output in $J^{MEO} = \min\{n, \tilde{\mathcal{O}}(\beta^{-1/2})\}$ iterations/Hessian-vector products!

Landscape-aware algorithm for $\text{minimize}_{\mathbf{x}\in\mathcal{M}} f(\mathbf{x})$

Inputs: $\mathbf{x}_0 \in \mathcal{M}$, $\Delta_0 > 0$, $\eta > 0$, α, β, γ.

For k=$0, 1, 2, \dots$

1. If $\|g(\mathbf{x}_k)\| \geq \alpha$, take a (cheap) gradient-type step:

$$\mathbf{s}_k = \underset{\substack{\mathbf{s}\in\mathcal{T}^{\mathcal{M}}_{\mathbf{x}_k} \\ \|\mathbf{s}\|\leq\Delta_k}}{\operatorname{argmin}} < g(\mathbf{x}_k), \mathbf{s} > .$$

2. Otherwise, apply truncated CG to

$$\underset{\substack{\mathbf{s}\in\mathcal{T}^{\mathcal{M}}_{\mathbf{x}_k} \\ \|\mathbf{s}\|\leq\Delta_k}}{\text{minimize}} < g(\mathbf{x}_k), \mathbf{s} > + \frac{1}{2} < \mathbf{s}, \mathcal{H}(\mathbf{x}_k) + 2\gamma I \mathbf{s} >$$

   If it terminates with $\mathbf{s}$, use $\mathbf{s}_k = \mathbf{s}$.

3. Otherwise, call MEO to find $\mathbf{s}_k$ such that $< \mathbf{s}, \mathcal{H}(\mathbf{x}_k)\mathbf{s} > \leq -\beta/2\|\mathbf{s}\|^2$.
4. Define $\mathbf{x}_k^{\mathcal{M}}$ as the retraction of $\mathbf{x}_k + \mathbf{s}_k$ onto $\mathcal{M}$.
5. Update $\mathbf{x}_{k+1}$ and Δ_{k+1} as before.

Goal: Compute $\boldsymbol{x}_K$ such that $\|g(\boldsymbol{x}_K)\| \le \epsilon_{\mathrm{g}}$ and $\mathcal{H}(\boldsymbol{x}_K) \succeq -\epsilon_{\mathrm{H}} I$.

Operation complexity (Goyens and R., '24)

The method reaches an $(\epsilon_{\mathrm{g}}, \epsilon_{\mathrm{H}})$-point in

$$N_{\epsilon_{\mathrm{g}}} = \tilde{\mathcal{O}}\left(\min\left\{n, \max\{\beta^{-1/2}, \gamma^{-1/2}\}\right\}\right) \times \left(\max\left\{\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}\right\} + \log\log\left[\mathcal{O}\left(\gamma\epsilon_{\mathrm{g}}^{-1}\right)\right]\right)$$

gradient/Hessian-vector products with probability $(1-\xi)^{N_{\epsilon_{\mathrm{g}}}}$.

Goal: Compute $\boldsymbol{x}_K$ such that $\|g(\boldsymbol{x}_K)\| \le \epsilon_{\mathrm{g}}$ and $\mathcal{H}(\boldsymbol{x}_K) \succeq -\epsilon_{\mathrm{H}} I$.

Operation complexity (Goyens and R., '24)

The method reaches an $(\epsilon_{\mathrm{g}}, \epsilon_{\mathrm{H}})$-point in

$$N_{\epsilon_{\mathrm{g}}} = \tilde{\mathcal{O}}\left(\min\left\{n, \max\{\beta^{-1/2}, \gamma^{-1/2}\}\right\}\right) \times \left(\max\left\{\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}\right\} + \log\log\left[\mathcal{O}\left(\gamma\epsilon_{\mathrm{g}}^{-1}\right)\right]\right)$$

gradient/Hessian-vector products with probability $(1-\xi)^{N_{\epsilon_{\mathrm{g}}}}$.

- Probability holds for second-order guarantee;
- Per-iteration cost does not depend on ϵ_{g}!

Illustration for phase retrieval

Phase retrieval (Sun et al '18)

$$\underset{\boldsymbol{x}\in\mathbb{C}^n}{\text{minimize}} \, \frac{1}{2m}\sum_{i=1}^{m}(b_i^2 - |\boldsymbol{a}_i^*\boldsymbol{x}|^2)^2.$$

If $\{\boldsymbol{a}_i\}$ are Gaussian and $m = \mathcal{O}(n\log^3(n))$, the objective is $(\frac{c}{n\log(m)}, c, c, \frac{c}{n\log(m)})$-strict saddle for some absolute constant $c > 0$.

Illustration for phase retrieval

Phase retrieval (Sun et al '18)

$$\operatorname*{minimize}_{\mathbf{x}\in\mathbb{C}^n} \frac{1}{2m}\sum_{i=1}^{m}(b_i^2 - |\mathbf{a}_i^*\mathbf{x}|^2)^2.$$

If $\{\mathbf{a}_i\}$ are Gaussian and $m = \mathcal{O}(n\log^3(n))$, the objective is $(\frac{c}{n\log(m)}, c, c, \frac{c}{n\log(m)})$-strict saddle for some absolute constant $c > 0$.

Complexity results

- Basic trust region:
  - $\mathcal{O}(\epsilon_g^{-3/2})$ iterations
  - $\tilde{\mathcal{O}}(\epsilon_g^{-7/4})$ operations.
- Using strict saddle:
  - $\mathcal{O}(n^2) + \log\log(\mathcal{O}(\epsilon_g^{-1}))$ iterations.
  - $\tilde{\mathcal{O}}\left(n^{5/2}\right) + \tilde{\mathcal{O}}(n^{1/2})\log\log(\mathcal{O}(\epsilon_g^{-1}))$ operations.

Outline

Summary

Nonconvex problems can be (relatively) easy!

- Instances in data science have favorable landscape.
- No bad local minima, non-minima points easy to escape from.

Summary

Nonconvex problems can be (relatively) easy!

- Instances in data science have favorable landscape.
- No bad local minima, non-minima points easy to escape from.

Strict saddle functions are nice

- Matrix completion, phase retrieval, etc, can be formulated as strict saddle optimization problems.
- Formal definition is parametric (akin to strong convexity).

Summary

Nonconvex problems can be (relatively) easy!

- Instances in data science have favorable landscape.
- No bad local minima, non-minima points easy to escape from.

Strict saddle functions are nice

- Matrix completion, phase retrieval, etc, can be formulated as strict saddle optimization problems.
- Formal definition is parametric (akin to strong convexity).

Nonconvex algorithms are even better with strict saddle

- Trust-region methods are one example.
- Can design landscape-aware versions!

OSSOBUCCO

On Strict Saddle Optimization

- *Strict saddle* today: Good landscape.
- More broadly: Parametric class of functions.

OSSOBUCCO

On Strict Saddle Optimization

- *Strict saddle* today: Good landscape.
- More broadly: Parametric class of functions.

Unified Classification

- Beyond our strict saddle definition, many variants.
- Goal: Review those structures and give examples.

OSSOBUCCO

On Strict Saddle Optimization

- *Strict saddle* today: Good landscape.
- More broadly: Parametric class of functions.

Benchmarking

- A number of examples in the literature.
- A library (like CUTEst) would be useful!

Unified Classification

- Beyond our strict saddle definition, many variants.
- Goal: Review those structures and give examples.

OSSOBUCCO

On Strict Saddle Optimization

- *Strict saddle* today: Good landscape.
- More broadly: Parametric class of functions.

Benchmarking

- A number of examples in the literature.
- A library (like CUTEst) would be useful!

Unified Classification

- Beyond our strict saddle definition, many variants.
- Goal: Review those structures and give examples.

COmplexity

- Fastest algorithms possible?
- Better complexity with strict saddle parameters?

References

- E. M. Achour, F. Malgouyres and S. Gerchinovitz, *The loss landscape of deep linear neural networks: a second-order analysis*, Journal of Machine Learning Research, 2024.
- N. Boumal, P.-A. Absil and C. Cartis, *Global rates of convergence for nonconvex optimization on manifolds*, IMA Journal of Numerical Analysis, 2018.
- S. Burer and R. D. C. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming, 2003.
- C. Cartis, N. I. M. Gould and Ph. L. Toint, *Evaluation complexity of algorithms for nonconvex optimization: Theory, computation and perspectives*, SIAM, 2022.
- P. Caucheteux, F. Goyens, C. W. Royer and I. Waldspurger, *Trust-region methods for Burer-Monteiro factorizations of phase retrieval problems*, working paper, 2025.
- Y. Chi, Y. M. Lu and Y. Chen, *Nonconvex optimization meets low-rank matrix factorization: An overview*, IEEE Transactions on Signal Processing, 2019.
- F. E. Curtis, D. P. Robinson, C. W. Royer and S. J. Wright. *Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization*, SIAM Journal on Optimization, 2021.
- R. Ge, C. Jin and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, International Conference on Machine Learning, 2017.
- F. Goyens and C. W. Royer, *Riemannian trust-region methods for strict saddle functions with complexity guarantees*, Mathematical Programming, 2024.
- C. Jin, P. Netrapalli, R. Ge, S. M. Kakade and M. I. Jordan, *How to escape saddle points efficiently*, International Conference on Machine Learning, 2017.
- J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan and B. Recht, *First-order methods almost always avoid strict saddle points*, Mathematical Programming, 2019.
- J. Ma, G. Liang and S. Fattahi, *Implicit regularization of infinitesimally-perturbed gradient descent toward low-dimensional solutions*, arXiv:2505.17304, 2025.
- A. D. McRae, *Phase retrieval and matrix sensing via benign and overparametrized nonconvex optimization*, arXiv:2505.02636, 2025.
- Q. Rebjock and N. Boumal, *Fast convergence to non-isolated minima: four equivalent conditions for C2 functions*, Mathematical Programming, 2025.
- J. Wright and Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models*, 2022.

References

- E. M. Achour, F. Malgouyres and S. Gerchinovitz, *The loss landscape of deep linear neural networks: a second-order analysis*, Journal of Machine Learning Research, 2024.
- N. Boumal, P.-A. Absil and C. Cartis, *Global rates of convergence for nonconvex optimization on manifolds*, IMA Journal of Numerical Analysis, 2018.
- S. Burer and R. D. C. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming, 2003.
- C. Cartis, N. I. M. Gould and Ph. L. Toint, *Evaluation complexity of algorithms for nonconvex optimization: Theory, computation and perspectives*, SIAM, 2022.
- P. Caucheteux, F. Goyens, C. W. Royer and I. Waldspurger, *Trust-region methods for Burer-Monteiro factorizations of phase retrieval problems*, working paper, 2025.
- Y. Chi, Y. M. Lu and Y. Chen, *Nonconvex optimization meets low-rank matrix factorization: An overview*, IEEE Transactions on Signal Processing, 2019.
- F. E. Curtis, D. P. Robinson, C. W. Royer and S. J. Wright. *Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization*, SIAM Journal on Optimization, 2021.
- R. Ge, C. Jin and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, International Conference on Machine Learning, 2017.
- F. Goyens and C. W. Royer, *Riemannian trust-region methods for strict saddle functions with complexity guarantees*, Mathematical Programming, 2024.
- C. Jin, P. Netrapalli, R. Ge, S. M. Kakade and M. I. Jordan, *How to escape saddle points efficiently*, International Conference on Machine Learning, 2017.
- J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan and B. Recht, *First-order methods almost always avoid strict saddle points*, Mathematical Programming, 2019.
- J. Ma, G. Liang and S. Fattahi, *Implicit regularization of infinitesimally-perturbed gradient descent toward low-dimensional solutions*, arXiv:2505.17304, 2025.
- A. D. McRae, *Phase retrieval and matrix sensing via benign and overparametrized nonconvex optimization*, arXiv:2505.02636, 2025.
- Q. Rebjock and N. Boumal, *Fast convergence to non-isolated minima: four equivalent conditions for C2 functions*, Mathematical Programming, 2025.
- J. Wright and Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models*, 2022.

Merci !

clement.royer@lamsade.dauphine.fr