From nonconvex optimization to strict saddle optimization

Clément Royer

Mathematics, Artificial Intelligence and Applications

February 23, 2025

Foreword: Nonconvex optimization

Nonconvex?

- Many data science problems are convex: linear classification, logistic regression,...
- **Nonconvex** instances: Deep learning, matrix/tensor optimization, robust statistics.

Foreword: Nonconvex optimization

Nonconvex?

- Many data science problems are convex: linear classification, logistic regression,...
- **Nonconvex** instances: Deep learning, matrix/tensor optimization, robust statistics.

Strict saddle?

- Those problems often come with nice structure;
- Guarantees to find global optima using local algorithms.

Foreword: Nonconvex optimization

Nonconvex?

- Many data science problems are convex: linear classification, logistic regression,...
- **Nonconvex** instances: Deep learning, matrix/tensor optimization, robust statistics.

Strict saddle?

- Those problems often come with nice structure;
- Guarantees to find global optima using local algorithms.

Optimization?

- Provably convergent algorithms for nonconvex problems.
- Provably fast algorithms (in a complexity sense).

The matrix completion example

Matrix completion

$$\min_{X \in \mathbb{R}^{n \times m}, \mathrm{rank}(X)=r} \|\mathcal{P}_\Omega(X-M)\|_F^2, \quad M \in \mathbb{R}^{n \times m},\ \Omega \subset [n] \times [m].$$

- Ω: Set of entries drawn i.i.d. with probability p.
- $M = U_* V_*^{\mathrm{T}}$, $U_* \in \mathbb{R}^{n \times r}$, $V_* \in \mathbb{R}^{m \times r}$.
- Convex objective in X.

The matrix completion example

Matrix completion

$$\min_{X \in \mathbb{R}^{n \times m}, \mathrm{rank}(X)=r} \|\mathcal{P}_\Omega(X - M)\|_F^2, \quad M \in \mathbb{R}^{n \times m}, \ \Omega \subset [n] \times [m].$$

- Ω: Set of entries drawn i.i.d. with probability p.
- $M = U_* V_*^{\mathrm{T}}$, $U_* \in \mathbb{R}^{n \times r}$, $V_* \in \mathbb{R}^{m \times r}$.
- Convex objective in X.

Nonconvex factored reformulation (Burer & Monteiro, '03)

$$\min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} \left\|\mathcal{P}_\Omega(U V^{\mathrm{T}} - M)\right\|_F^2,$$

- **Nonconvex problem** in U and V...
- but **global minima** can be characterized.

Numerical illustration

Matrix problem

$$\min_{U,V} \frac{1}{2} \left\| P_\Omega(UV^\top - M) \right\|_F^2,$$

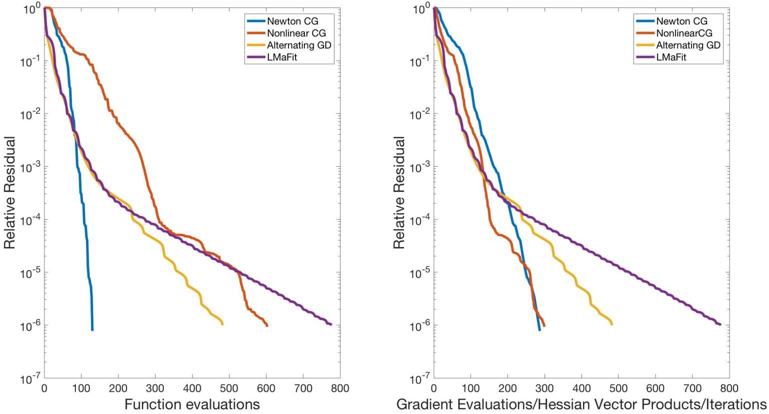
with $M \in \mathbb{R}^{m \times n},\ U \in \mathbb{R}^{m \times r},\ V \in \mathbb{R}^{n \times r}$, $|\Omega| \approx 15\% \times mn$.

- Synthetic data: $(n, m) = (500, 499)$.

Comparison: A second-order method VS first-order ones

- Newton-CG (us);
- Nonlinear CG (first-order method);
- Dedicated solvers (Alternating methods):
  - Alternated gradient descent (Tanner and Wei 2016);
  - LMaFit (Wen et al. 2012).

Matrix completion (synthetic data, rank 15)



Takeaways from the example

The example

- Particular structure (linked to derivatives).
- Favorable case for second-order schemes.

Our questions

- Can we characterize nice problem structure?
- Can we build an algorithm for such structure?

Outline

Outline

A class of manifold optimization problems

Problem: $\min_{x \in \mathcal{M}} f(x)$, $\mathcal{M}$ Riemannian manifold.

Examples

- Vector spaces: $\mathbb{R}^n$, $\mathbb{C}^n$, $\mathbb{S}^{n-1}$.
- Matrices: $\mathbb{R}^{n \times m}$, Grassmann (subspaces), Stiefel (orthogonal matrices).

A class of manifold optimization problems

Problem: $\min_{x \in \mathcal{M}} f(x)$, $\mathcal{M}$ Riemannian manifold.

Examples

- Vector spaces: $\mathbb{R}^n$, $\mathbb{C}^n$, $\mathbb{S}^{n-1}$.
- Matrices: $\mathbb{R}^{n \times m}$, Grassmann (subspaces), Stiefel (orthogonal matrices).

Notations and conventions

- Riemannian displacements:
  - Moves defined over tangent spaces $\mathcal{T}_x^{\mathcal{M}} \equiv \mathbb{R}^m$.
  - Retraction that "projects" back onto the manifold.
  - Norms and inner products ($\|\cdot\|^2 = <\cdot,\cdot>$ here for simplicity).

A class of manifold optimization problems

Problem: $\min_{x \in \mathcal{M}} f(x)$, $\mathcal{M}$ Riemannian manifold.

Examples

- Vector spaces: $\mathbb{R}^n$, $\mathbb{C}^n$, $\mathbb{S}^{n-1}$.
- Matrices: $\mathbb{R}^{n \times m}$, Grassmann (subspaces), Stiefel (orthogonal matrices).

Notations and conventions

- Riemannian displacements:
  - Moves defined over tangent spaces $\mathcal{T}_x^{\mathcal{M}} \equiv \mathbb{R}^m$.
  - Retraction that "projects" back onto the manifold.
  - Norms and inner products ($\|\cdot\|^2 = <\cdot,\cdot>$ here for simplicity).
- Riemannian derivatives:
  - Counterparts of gradient and Hessian in Euclidean setting.
  - Riemannian gradient $g(\cdot) = g_{f,\mathcal{M}}(\cdot)$ seen as a vector.
  - Riemannian Hessian $\mathcal{H}(\cdot) = H_{f,\mathcal{M}}(\cdot)$ seen as a matrix.

A class of manifold optimization problems

Problem: $\min_{x \in \mathcal{M}} f(x)$, $\mathcal{M}$ Riemannian manifold.

Examples

- Vector spaces: $\mathbb{R}^n$, $\mathbb{C}^n$, $\mathbb{S}^{n-1}$.
- Matrices: $\mathbb{R}^{n \times m}$, Grassmann (subspaces), Stiefel (orthogonal matrices).

Notations and conventions

- Riemannian displacements:
  - Moves defined over tangent spaces $\mathcal{T}_x^{\mathcal{M}} \equiv \mathbb{R}^m$.
  - Retraction that "projects" back onto the manifold.
  - Norms and inner products ($\| \cdot \|^2 = < \cdot, \cdot >$ here for simplicity).
- Riemannian derivatives:
  - Counterparts of gradient and Hessian in Euclidean setting.
  - Riemannian gradient $g(\cdot) = g_{f,\mathcal{M}}(\cdot)$ seen as a vector.
  - Riemannian Hessian $\mathcal{H}(\cdot) = H_{f,\mathcal{M}}(\cdot)$ seen as a matrix.

Many formulas are available in modern toolboxes (`Manopt`).

General problem and definitions

$$\min_{x \in \mathcal{M}} f(x)$$

- $f \in \mathcal{C}^2$ bounded below and nonconvex.
- $\mathcal{M}$ Riemannian manifold.

General problem and definitions

$$\min_{x \in \mathcal{M}} f(x)$$

- $f \in \mathcal{C}^2$ bounded below and nonconvex.
- $\mathcal{M}$ Riemannian manifold.

Goal: Reach an ϵ-stationary point

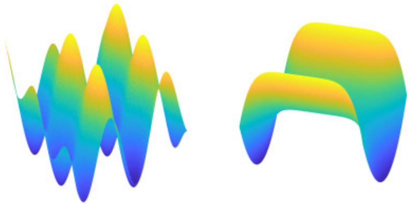
$$\|g(x)\| \leq \epsilon \quad \text{and} \quad \lambda_{\min}\left(\mathcal{H}(x)\right) \geq -\epsilon^{1/2}.$$

- For convex functions: Second condition always true$\Rightarrow$ Close to a global minimum!
- For nonconvex functions: ?

Pathological cases

ϵ-stationary points can be close to

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



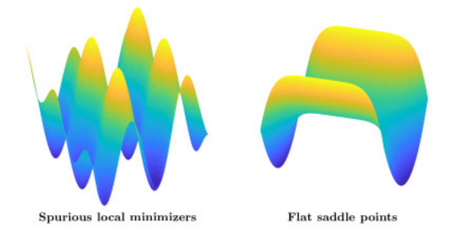
Spurious local minimizers

Flat saddle points

Pathological cases

ϵ-stationary points can be close to

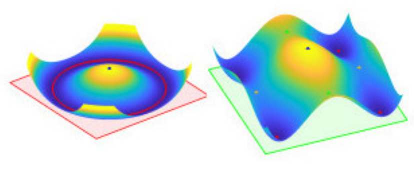
- Local, non-global minima.
- High-order saddle points.



Nice instances

ϵ-stationary points are close to

- Strict (non-flat) saddle points
- Global minima.



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

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

Definition

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

1. $\|g(x)\| \geq \alpha$;
2. $\lambda_{\min}\left(\mathcal{H}(x)\right) \leq -\beta$;
3. There exists x^* local minimum of f such that $d(x, x^*) \leq \delta$ and $\lambda_{\min}\left(\mathcal{H}(y)\right) \geq \gamma$ for all $\{y \in \mathcal{M} : d(x, x^*) \leq 2\delta\}$.

$d(\cdot, \cdot)$: Riemannian distance.

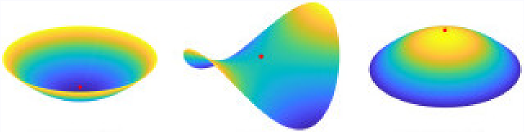
Strict saddle property on manifold $\mathcal{M}$

Definition

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

1. $\|g(x)\| \geq \alpha$;
2. $\lambda_{\min}(\mathcal{H}(x)) \leq -\beta$;
3. There exists x^* local minimum of f such that $d(x, x^*) \leq \delta$ and $\lambda_{\min}(\mathcal{H}(y)) \geq \gamma$ for all $\{y \in \mathcal{M} : d(x, x^*) \leq 2\delta\}$.

$d(\cdot, \cdot)$: Riemannian distance.



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

Definition

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

1. $\|g(x)\| \geq \alpha$;
2. $\lambda_{\min}\left(\mathcal{H}(x)\right) \leq -\beta$;
3. There exists x^* local minimum of f such that $d(x, x^*) \leq \delta$ and $\lambda_{\min}\left(\mathcal{H}(y)\right) \geq \gamma$ for all $\{y \in \mathcal{M} : d(x, x^*) \leq 2\delta\}$.

$d(\cdot, \cdot)$: Riemannian distance.

Interpretation: 3 regions in the space

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

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

Example: Matrix completion

$$\min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} f(U,V) := \left\| \mathcal{P}_\Omega(U\,V^{\mathrm{T}} - M) \right\|_F^2,$$

Assumptions

- Probability of sampling entries large enough.
- M has favorable structure (incoherence).

Theorem (Ge et al. '17)

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

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

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

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

A special case

Phase retrieval (Sun et al '18)

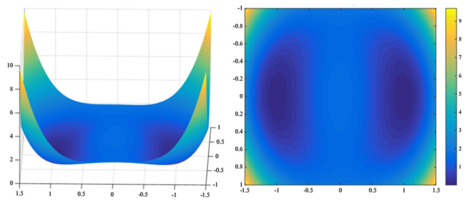
- Given $A = [a_i]_{i=1}^m \in \mathbb{C}^{m \times n}$, $b \in \mathbb{R}^m$, find $x \in \mathbb{C}^n$ such that $|a_i^* x| = b_i \quad \forall i = 1, \dots, n$.
- Assumptions: $\{a_i\}$ Gaussian, $m = \mathcal{O}(n \log^3(n))$.
- Nonconvex formulation: $\min_{x \in \mathbb{C}^n} f(x) = \frac{1}{2m} \sum_{i=1}^m (b_i^2 - |a_i^* x|^2)^2$.

A special case

Phase retrieval (Sun et al '18)

- Given $A = [a_i]_{i=1}^m \in \mathbb{C}^{m \times n}$, $b \in \mathbb{R}^m$, find $x \in \mathbb{C}^n$ such that $|a_i^* x| = b_i \quad \forall i = 1, \dots, n$.
- Assumptions: $\{a_i\}$ Gaussian, $m = \mathcal{O}(n \log^3(n))$.
- Nonconvex formulation: $\min_{x \in \mathbb{C}^n} f(x) = \frac{1}{2m} \sum_{i=1}^m (b_i^2 - |a_i^* x|^2)^2$.

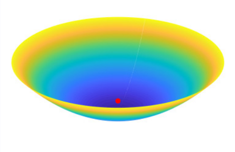
There exists $c > 0$ such that f is $\left(\frac{c}{n \log(m)}, c, c, \frac{c}{n \log(m)}\right)$-strict saddle.



Other examples (pictures from Wright, Ma '22)

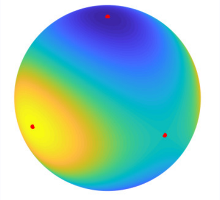
Strongly convex functions (!)

$$\lambda_{\min}(\mathcal{H}(x)) \geq \gamma \ \forall x.$$



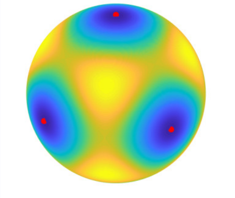
Minimum Eigenvalue

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



Tensor optimization

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



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

Outline

What we want: Develop a method that explicitly uses the strict saddle nature of the problem.

Solving strict-saddle problems

What we want: Develop a method that explicitly uses the strict saddle nature of the problem.

Our friends at work (O'Neill and Wright '23)

- Line-search approach for strict saddle functions
- Focus on factored formulations low-rank matrix problems.

What we want: Develop a method that explicitly uses the strict saddle nature of the problem.

Our friends at work (O'Neill and Wright '23)

- Line-search approach for strict saddle functions
- Focus on factored formulations low-rank matrix problems.

How we want to stand out

Apply to **any strict saddle function**.

- Newton-type steps.
- Trust-region framework.
- General manifold constraints.

Trust-region algorithm

Inputs: $x_0 \in \mathcal{M}$, $\Delta_0 > 0$, $\eta > 0$.

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

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

Trust-region algorithm

Inputs: $x_0 \in \mathcal{M}$, $\Delta_0 > 0$, $\eta > 0$.

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

1. Define $m_k(x_k + s) :=< g(x_k), s > +\frac{1}{2} < s, \mathcal{H}(x_k)s >$ and compute

$$s_k \in \operatorname*{argmin}_{\substack{s \in \mathcal{T}^{\mathcal{M}}_{x_k} \\ \|s\| \le \Delta_k}} m_k(x_k + s).$$

2. Define $x_k^{\mathcal{M}}$ as the retraction of $x_k + s_k$ onto $\mathcal{M}$.
3. Compute $\rho_k = \frac{f(x_k) - f(x_k^{\mathcal{M}})}{m_k(x_k) - m_k(x_k^{\mathcal{M}})}$.
4. If $\rho_k \ge \eta$, set $x_{k+1} = x_k^{\mathcal{M}}$ and $\Delta_{k+1} = 2\Delta_k$.
5. Otherwise, set $x_{k+1} = x_k$ and $\Delta_{k+1} = 0.5\Delta_k$.

- Suboptimal guarantees for generic, nonconvex f.

Trust-region algorithm

Inputs: $x_0 \in \mathcal{M}$, $\Delta_0 > 0$, $\eta > 0$.

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

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

- Suboptimal guarantees for generic, nonconvex f.
- Improved guarantees for strict saddle f!

What happens if the function is strict saddle?

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

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

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

What happens if the function is strict saddle?

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

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

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

One step per strict saddle case

1. $\|g(x_k)\| \geq \alpha$: Descent/Cauchy step.
2. $\lambda_{\min}(\mathcal{H}(x_k)) \leq -\beta$: Negative curvature step.
3. Otherwise: TR-Newton step **without regularization**.

Analysis

Trust-region radius bound

For all iterations, $\Delta_k \geq \mathcal{O}\left(\min\{\alpha, \beta, \gamma\}\right)$.

Trust-region radius bound

For all iterations, $\Delta_k \geq \mathcal{O}\left(\min\{\alpha, \beta, \gamma\}\right)$.

Decrease guarantees for successful iterations

1. If $\|g(x_k)\| \geq \alpha$,

$$f(x_k) - f(x_{k+1}) \geq \mathcal{O}\left(\min\{\alpha^2, \alpha\Delta_k\}\right).$$

Trust-region radius bound

For all iterations, $\Delta_k \geq \mathcal{O}\left(\min\{\alpha, \beta, \gamma\}\right)$.

Decrease guarantees for successful iterations

1. If $\|g(x_k)\| \geq \alpha$,

$$f(x_k) - f(x_{k+1}) \geq \mathcal{O}\left(\min\{\alpha^2, \alpha\Delta_k\}\right).$$

2. If $\lambda_{\min}(\mathcal{H}(x_k)) \leq -\beta$:

$$f(x_k) - f(x_{k+1}) \geq \mathcal{O}\left(\beta\Delta_k^2\right).$$

Analysis

Trust-region radius bound

For all iterations, $\Delta_k \geq \mathcal{O}\left(\min\{\alpha, \beta, \gamma\}\right)$.

Decrease guarantees for successful iterations

1. If $\|g(x_k)\| \geq \alpha$,

$$f(x_k) - f(x_{k+1}) \geq \mathcal{O}\left(\min\{\alpha^2, \alpha\Delta_k\}\right).$$

2. If $\lambda_{\min}(\mathcal{H}(x_k)) \leq -\beta$:

$$f(x_k) - f(x_{k+1}) \geq \mathcal{O}\left(\beta\Delta_k^2\right).$$

3. Otherwise, either

$$f(x_k) - f(x_{k+1}) \geq \mathcal{O}\left(\min\left\{\gamma\Delta_k^2, \gamma\|g(x_{k+1})\|\right\}\right)$$

or $\|g(x_k)\| \leq \mathcal{O}(\min\{\delta\gamma, \gamma^2\})$ and we enter a local convergence phase.

Complexity of strict saddle Newton

Goal: Compute x_k such that $\|g(x_k)\| \leq \epsilon$ and $\lambda_{\min}\left(\mathcal{H}(x_k)\right) \geq -\epsilon^{1/2}$.

Iteration complexity (Goyens and R., '23)

Suppose $\epsilon < \min\{\alpha, \beta, \gamma\}^2 < 1$.
The method reaches an $(\epsilon, \epsilon^{1/2})$-point in at most

$$\mathcal{O}\left(\max\left\{\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}\right\}\right) + \log\log\left[\mathcal{O}\left(\gamma\epsilon^{-1}\right)\right]$$

iterations.

Complexity of strict saddle Newton

Goal: Compute x_k such that $\|g(x_k)\| \leq \epsilon$ and $\lambda_{\min}\left(\mathcal{H}(x_k)\right) \geq -\epsilon^{1/2}$.

Iteration complexity (Goyens and R., '23)

Suppose $\epsilon < \min\{\alpha, \beta, \gamma\}^2 < 1$.
The method reaches an $(\epsilon, \epsilon^{1/2})$-point in at most

$$\mathcal{O}\left(\max\left\{\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}\right\}\right) + \log\log\left[\mathcal{O}\left(\gamma\epsilon^{-1}\right)\right]$$

iterations.

- Second term vanishes when $\epsilon \geq \max\{\alpha, \beta\}$.
- Otherwise $\log\log$ dependency in ϵ (from local phase)!

Illustration for phase retrieval

Phase retrieval (Sun et al '18)

$$\min_{x \in \mathbb{C}^n} \frac{1}{2m} \sum_{i=1}^{m} (b_i^2 - |a_i^* x|^2)^2.$$

If $\{a_i\}$ 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 $c > 0$.

Impact on the complexity

- For generic Newton, get $\mathcal{O}(\epsilon^{-3/2})$ complexity.
- For strict saddle Newton, we obtain

$$\tilde{\mathcal{O}}\left(n^2\right) + \log\log(\mathcal{O}(\epsilon^{-1}))\,.$$

Adding inexact steps

What we have so far

- Newton-type method with good complexity;
- Three kinds of steps;
- Require exact step computation.

Inexactness

- Solve linear systems;
- Compute negative curvature directions.

Inexact steps in trust region

Trust-region subproblem

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

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

Inexact steps in trust region

Trust-region subproblem

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

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

Changes (for complexity)

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

Our method: Capped conjugate gradient

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

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

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

1. A vector s such that

$$s^{\mathrm{T}} \mathcal{H}(x_k) s \leq -\frac{\beta}{2} \|s\|^2.$$

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

Minimum eigenvalue oracle (MEO)

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

1. A vector s such that

$$s^{\mathrm{T}} \mathcal{H}(x_k) s \leq -\frac{\beta}{2} \|s\|^2.$$

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

An example of MEO

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

Minimum eigenvalue oracle (MEO)

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

1. A vector s such that

$$s^{\mathrm{T}} \mathcal{H}(x_k) s \leq -\frac{\beta}{2} \|s\|^2.$$

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

An example of MEO

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

Strict saddle version: Identical, but we know that negative curvature exists!

Inexact algorithm for $\min_{x \in \mathcal{M}} f(x)$

Inputs: $x_0 \in \mathcal{M}$, $\Delta_0 > 0$, $\eta > 0$.

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

1. Define

$$m_k(x_k + s) = \begin{cases} < g(x_k), s > & \text{if } \|g(x_k)\| \geq \alpha \\ \\ < g(x_k), s > + \frac{1}{2} < s, \mathcal{H}(x_k) s > & \text{otherwise.} \end{cases}$$

2. Compute $s_k \approx \operatorname{argmin}_{\substack{s \in \mathcal{T}_{x_k}^{\mathcal{M}} \\ \|s\| \leq \Delta_k}} m_k(x_k + s)$ by CG(+MEO) when $\|g(x_k)\| < \alpha$.
3. Define $x_k^{\mathcal{M}}$ as the retraction of $x_k + s_k$ onto $\mathcal{M}$.
4. Compute $\rho_k = \frac{f(x_k) - f(x_k^{\mathcal{M}})}{m_k(x_k) - m_k(x_k^{\mathcal{M}})}$.
5. If $\rho_k \geq \eta$, set $x_{k+1} = x_k^{\mathcal{M}}$ and $\Delta_{k+1} = 2\Delta_k$.
6. Otherwise, set $x_{k+1} = x_k$ and $\Delta_{k+1} = 0.5\Delta_k$.

Operation complexity

Goal: Compute x_k such that $\|g(x_k)\| \leq \epsilon$ and $\mathcal{H}(x_k) \succeq -\epsilon^{1/2} I$.

Operation complexity (Goyens and R., '23)

Suppose $\epsilon < \min\{\alpha, \beta, \gamma\}^2 < 1$.
The method reaches an $(\epsilon, \epsilon^{1/2})$-point in

$$\begin{aligned} N_\epsilon &= \tilde{\mathcal{O}}\left(\min\left\{n, \max\{\beta^{-1/2}, \gamma^{-1/2}\}\right\}\right) \\ &\quad \times \left(\max\left\{\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}\right\} + \log\log\left[\mathcal{O}\left(\gamma\epsilon^{-1}\right)\right]\right) \end{aligned}$$

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

Operation complexity

Goal: Compute x_k such that $\|g(x_k)\| \leq \epsilon$ and $\mathcal{H}(x_k) \succeq -\epsilon^{1/2} I$.

Operation complexity (Goyens and R., '23)

Suppose $\epsilon < \min\{\alpha,\beta,\gamma\}^2 < 1$.
The method reaches an $(\epsilon,\epsilon^{1/2})$-point in

$$\begin{aligned} N_\epsilon &= \tilde{\mathcal{O}}\left(\min\left\{n, \max\{\beta^{-1/2},\gamma^{-1/2}\}\right\}\right) \\ &\quad \times \left(\max\left\{\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}\right\} + \log\log\left[\mathcal{O}\left(\gamma\epsilon^{-1}\right)\right]\right) \end{aligned}$$

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

- Probability holds for second-order guarantee;
- Per-iteration cost does not depend on ϵ!

Illustration for phase retrieval

Phase retrieval (Sun et al '18)

$$\min_{x \in \mathbb{C}^n} \frac{1}{2m} \sum_{i=1}^{m} (b_i^2 - |a_i^* x|^2)^2.$$

If $\{a_i\}$ 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 $c > 0$.

Impact on the complexity

- For generic Newton, get $\mathcal{O}(\epsilon^{-7/4})$ complexity.
- For strict saddle Newton, we obtain

$$\tilde{\mathcal{O}}\left(n^{5/2}\right) + \tilde{\mathcal{O}}(n^{1/2}) \log\log(\mathcal{O}(\epsilon^{-1})).$$

Summary

Strict saddle optimization

- A wide class of nonconvex problems.
- Favorable landscape.
- Room for efficient algorithms!

Summary

Strict saddle optimization

- A wide class of nonconvex problems.
- Favorable landscape.
- Room for efficient algorithms!

Our proposal

- Trust-region framework (good for nonconvex).
- Inexact variant tailored to strict saddle problems.
- Ongoing implementation.

References

- S. Bhojanapalli, B. Neyshabur and N. Srebro, *Global optimality of local search for low-rank matrix recovery*, Neural Information Processing Systems, 2016.
- 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.
- 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.
- R. Ge and T. Ma, *On the optimization landscape of tensor decompositions*, Advances in Neural Information Processing Systems, 2017.
- F. Goyens and C. W. Royer. *Riemannian trust-region methods for strict saddle functions with complexity guarantees*, Mathematical Programming, 2024.
- M. O'Neill and S. J. Wright. *A line-search descent algorithm for strict saddle functions with complexity guarantees*, Journal of Machine Learning Research, 2023.
- J. Sun, Q. Qu and J. Wright. *A geometric analysis of phase retrieval*, Found. Comput. Math., 2018.
- J. Tanner and K. Wei. *Low rank matrix completion by alternating steepest descent methods*, Applied and Computational Harmonic Analysis, 2016.
- Z. Wen, W. Yin and Y. Zhang. *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*, Mathematical Programming, 2012.
- J. Wright and Y. Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*, Cambridge University Press, 2022.

References

- S. Bhojanapalli, B. Neyshabur and N. Srebro, *Global optimality of local search for low-rank matrix recovery*, Neural Information Processing Systems, 2016.
- 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.
- 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.
- R. Ge and T. Ma, *On the optimization landscape of tensor decompositions*, Advances in Neural Information Processing Systems, 2017.
- F. Goyens and C. W. Royer. *Riemannian trust-region methods for strict saddle functions with complexity guarantees*, Mathematical Programming, 2024.
- M. O'Neill and S. J. Wright. *A line-search descent algorithm for strict saddle functions with complexity guarantees*, Journal of Machine Learning Research, 2023.
- J. Sun, Q. Qu and J. Wright. *A geometric analysis of phase retrieval*, Found. Comput. Math., 2018.
- J. Tanner and K. Wei. *Low rank matrix completion by alternating steepest descent methods*, Applied and Computational Harmonic Analysis, 2016.
- Z. Wen, W. Yin and Y. Zhang. *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*, Mathematical Programming, 2012.
- J. Wright and Y. Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*, Cambridge University Press, 2022.

Thank you!

clement.royer@lamsade.dauphine.fr