Riemannian trust-region methods for strict saddle functions with complexity guarantees

Florentin Goyens
joint with
Clément Royer at Paris-Dauphine University

ISMP Montreal
23 July, 2024

Problem (P)

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

- $\mathcal{M}$ is a (smooth) Riemannian manifold
- $f \colon \mathcal{M} \to \mathbb{R}$ is smooth and nonconvex

Applications: unconstrained optimization ($\mathcal{M} = \mathbb{R}^n, \mathbb{R}^{m \times n} \dots$), orthogonality constraints, fixed-rank constraints, ...

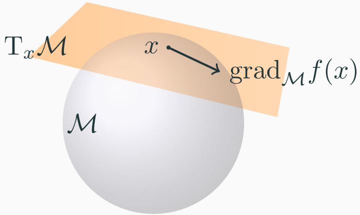
How many iterations of an optimization algorithm are required in the worst-case to reach an approximate solution of (P) from an arbitrary initial $x_0 \in \mathcal{M}$?

How many iterations of an optimization algorithm are required in the worst-case to reach an approximate solution of (P) from an arbitrary initial $x_0 \in \mathcal{M}$?

Answer We answer this question for strict saddle functions for the Riemannian trust-region algorithm (exact and inexact versions).

Optimization on Manifolds

Minimize $f\colon \mathcal{M} \to \mathbb{R}$ where the feasible set $\mathcal{M}$ is a Riemannian manifold.



$\mathrm{D}f(x)[\Delta] = \langle \mathrm{grad} f(x), \Delta \rangle$ with $\mathrm{grad}_{\mathcal{M}} f(x) \in \mathrm{T}_x\mathcal{M}$ and
$\mathrm{D}^2 f(x)[\Delta, \Delta] = \langle \mathrm{Hess} f(x)[\Delta], \Delta \rangle$ with $\mathrm{Hess} f(x)\colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$.

- Produces feasible sequence of iterates $x_0, x_1, x_2 \cdots \in \mathcal{M}$
- Requires $x_0 \in \mathcal{M}$ and retraction map $\mathrm{R}_x\colon \mathrm{T}_x\mathcal{M} \to \mathcal{M}$

Target points of optimization algorithms

First-order critical points

$$x \in \mathcal{M} \qquad \text{and} \qquad \operatorname{grad} f(x) = 0,$$

Second-order critical points

$$x \in \mathcal{M}, \qquad \operatorname{grad} f(x) = 0, \qquad \text{and} \qquad \operatorname{Hess} f(x) \succeq 0.$$

Target points of optimization algorithms

First-order critical points

$$x \in \mathcal{M} \qquad \text{and} \qquad \operatorname{grad} f(x) = 0,$$

Second-order critical points

$$x \in \mathcal{M}, \qquad \operatorname{grad} f(x) = 0, \qquad \text{and} \qquad \operatorname{Hess} f(x) \succeq 0.$$

Motivation Numerous applications have *benign* nonconvexity:

Second-Order Critical Point $\Longrightarrow$ global optimality

Target points of optimization algorithms

Second-order critical points

$$x \in \mathcal{M}, \qquad \operatorname{grad} f(x) = 0, \qquad \text{and} \qquad \operatorname{Hess} f(x) \succeq 0.$$

Their approximate version ε-SOCP

$$x \in \mathcal{M}, \qquad \|\operatorname{grad} f(x)\| \leq \varepsilon_1, \qquad \text{and} \qquad \operatorname{Hess} f(x) \succeq -\varepsilon_2 \mathrm{Id}.$$

Target points of optimization algorithms

Second-order critical points

$$x \in \mathcal{M}, \qquad \operatorname{grad} f(x) = 0, \qquad \text{and} \qquad \operatorname{Hess} f(x) \succeq 0.$$

Their approximate version ε-SOCP

$$x \in \mathcal{M}, \qquad \|\operatorname{grad} f(x)\| \leq \varepsilon_1, \qquad \text{and} \qquad \operatorname{Hess} f(x) \succeq -\varepsilon_2 \mathrm{Id}.$$

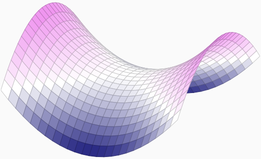
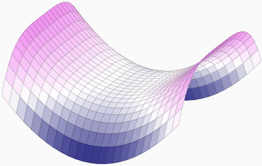
In this work, the landscape allows to show convergence to approximate minimizers

$$x \in \mathcal{M}, \qquad \|\operatorname{grad} f(x)\| \leq \varepsilon_1, \qquad \text{and} \qquad \operatorname{Hess} f(x) \succeq \gamma \mathrm{Id}$$

where γ is a local strong convexity constant

To find SOCP, avoid strict saddle points

Saddle point: $\operatorname{grad} f(x) = 0$ but $x \in \mathcal{M}$ is not a local minimizer.

Strict saddle point	Spurious SOCP
$f(x, y) = x^2 - y^2$	$f(x, y) = 10x^2 - y^4$
	
$\nabla^2 f(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$	$\nabla^2 f(0,0) = \begin{pmatrix} 20 & 0 \\ 0 & 0 \end{pmatrix}$
Indefinite	Positive semi-definite

To find SOCP, avoid strict saddle points

If all saddle points of f are strict, f is called a strict saddle function.

Proposition
If f is strict saddle,

$$\textit{Second-Order Critical Point} \implies \textit{(local) minimum}$$

To find SOCPs, algorithms must avoid strict saddle points

Algorithms that find ε-SOCP

Two strategies to provably avoid strict saddle points

First strategy: Randomization (Jin et al., 2017)

$$x_{k+1} = x_k - \alpha_k \mathrm{grad} f(x_k) + \xi_k$$

Second strategy: Negative curvature of the Hessian

$$\langle d, \mathrm{Hess} f(x) d \rangle < 0 \rightsquigarrow f(x + d) < f(x)$$

- Second-order methods naturally use curvature of the Hessian
- Easy to implement and simple proofs
- Deterministic results

Riemannian trust-region (RTR)

- Globally convergent variant of Newton's method

Algorithm 2 RTR with exact subproblem minimization

1: **for** $k = 1, 2, \ldots$ **do**

2: Compute s_k as a solution to the trust-region subproblem

$$s_k \in \underset{s \in \mathrm{T}_{x_k}\mathcal{M}}{\arg\min}\, f(x_k) + \langle \operatorname{grad} f(x_k), s_k \rangle + \frac{1}{2} \langle s_k, H_k s_k \rangle \ \text{ subject to } \ \|s\| \leq \Delta_k,$$

3: Use step s_k if f decreases sufficiently

4: Update trust-region radius Δ_k as needed

5: **end for**

Riemannian trust-region (RTR)

- Globally convergent variant of Newton's method

Algorithm 2 RTR with exact subproblem minimization

1: **for** $k = 1, 2, \ldots$ **do**
2: Compute s_k as a solution to the trust-region subproblem

$$s_k \in \operatorname*{arg\,min}_{s \in \mathrm{T}_{x_k}\mathcal{M}} f(x_k) + \langle \operatorname{grad} f(x_k), s_k \rangle + \frac{1}{2} \langle s_k, H_k s_k \rangle \text{ subject to } \|s\| \leq \Delta_k,$$

3: Use step s_k if f decreases sufficiently
4: Update trust-region radius Δ_k as needed
5: **end for**

- This algorithm provably returns an ε-SOCP
- We analyze this well-known algorithm for strict saddle functions
- Similar results for Newton + negative curvature steps

Complexity guarantees for generic nonconvex optimization

(without the strict saddle assumption)

How many iterations does it take to guarantee ε-SOCP in the worst-case ?

Quick example: Complexity of gradient descent

Theorem (informal)

Gradient descent produces a point with $\|\mathrm{grad} f(x)\| \leq \varepsilon$ in at most

$$\frac{f(x_0) - f(x^*)}{c\varepsilon^2} = \mathcal{O}(1/\varepsilon^2)$$

iterations.

Complexity of Riemannian algorithms

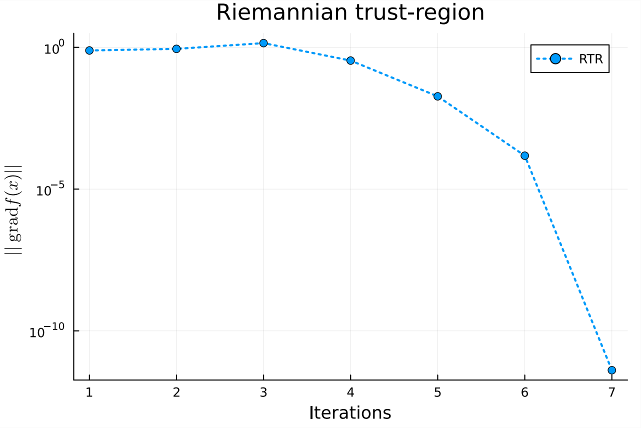
Boumal et al. (2019): Worst-case rates of optimization algorithms on manifolds are identical to the unconstrained case with respect to ε.

- Riemannian gradient descent produces a point $x \in \mathcal{M}$ that satisfies $\|\mathrm{grad}_{\mathcal{M}} f(x)\| \leq \varepsilon_1$ in at most $\mathcal{O}(1/\varepsilon_1^2)$ iterations
- Second-order Riemannian trust-region produces a point $x \in \mathcal{M}$ that satisfies $\|\mathrm{grad}_{\mathcal{M}} f(x)\| \leq \varepsilon_1$ and $\mathrm{Hess}_{\mathcal{M}} f(x) \succeq -\varepsilon_2 \mathrm{Id}$ in at most

$$\mathcal{O}\left(\frac{1}{\min\left(\varepsilon_1^2, \varepsilon_2^3\right)}\right)$$

iterations

$\mathcal{O}(\varepsilon^{-3})$: pessimistic worst-case bound which does not reflect the practical behaviour



- We improve this result in the case of **strict saddle functions**
- Our results are similar to strongly convex optimization results

Complexity guarantees with the strict saddle assumption

Strict saddle functions

All saddle points of f are strict

$$\Longleftrightarrow$$

$\|\mathrm{grad} f(x)\|$ is small

$\lambda_{\min}(\mathrm{Hess} f(x)) < 0$

x is near a (local) minimizer

Strict saddle functions

All saddle points of f are strict

$$\iff$$

$\|\mathrm{grad} f(x)\|$ is small

$\lambda_{\min}(\mathrm{Hess} f(x)) < 0$

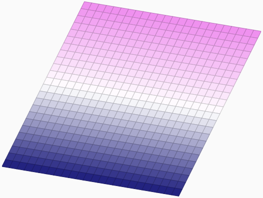
x is near a (local) minimizer

Our contribution: assumes that minimizers are isolated with local strong convexity

Landscape parameters: $(\alpha, \beta, \gamma, \delta) > 0$

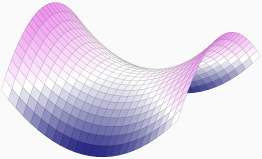
$$\mathcal{M} = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3$$

Region $\mathcal{R}_1$



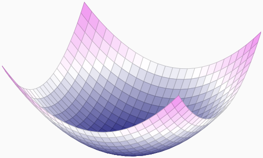
$\|\mathrm{grad} f(x)\| \geq \alpha$

Region $\mathcal{R}_2$



$\lambda_{\min}(\mathrm{Hess} f(x)) \leq -\beta$

Region $\mathcal{R}_3$



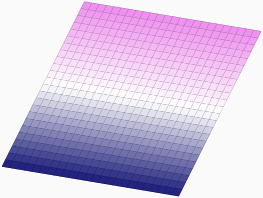
$\mathrm{Hess} f(x) \succeq \gamma \mathrm{Id}$

g-convex in $\mathrm{dist}(x, x^*) \leq \delta$

Landscape parameters: $(\alpha, \beta, \gamma, \delta) > 0$

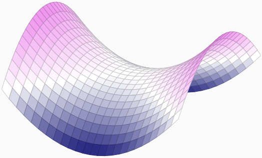
$$\mathcal{M} = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3$$

Region $\mathcal{R}_1$



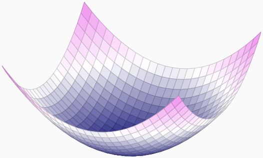
$\|\mathrm{grad} f(x)\| \geq \alpha$

Region $\mathcal{R}_2$



$\lambda_{\min}(\mathrm{Hess} f(x)) \leq -\beta$

Region $\mathcal{R}_3$



$\mathrm{Hess} f(x) \succeq \gamma \mathrm{Id}$

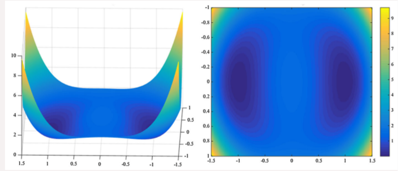
g-convex in $\mathrm{dist}(x, x^*) \leq \delta$

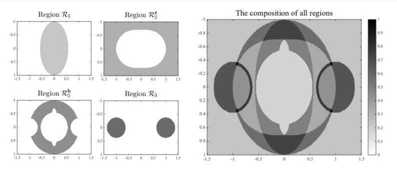
Near critical points, the smallest eigenvalue of $\mathrm{Hess} f(x)$ is bounded away from zero

Phase Retrieval: Recover $x \in \mathbb{C}^n$ from $b = |Ax| \in \mathbb{R}^m$ for some $A\colon \mathbb{C}^n \to \mathbb{C}^m$ with $m \geq 4n$. A natural formulation is

$$\min_{z \in \mathbb{C}^n} \frac{1}{4m} \sum_{k=1}^{m} (|a_k^* z|^2 - b_k^2)^2$$

which is a $(c/(n \log m), c, c, c/(n \log m))$ strict saddle function for some constant c (Sun et al., 2015, 2018).





Examples of strict saddle problems

Strict saddle functions appear in many other applications, such as:

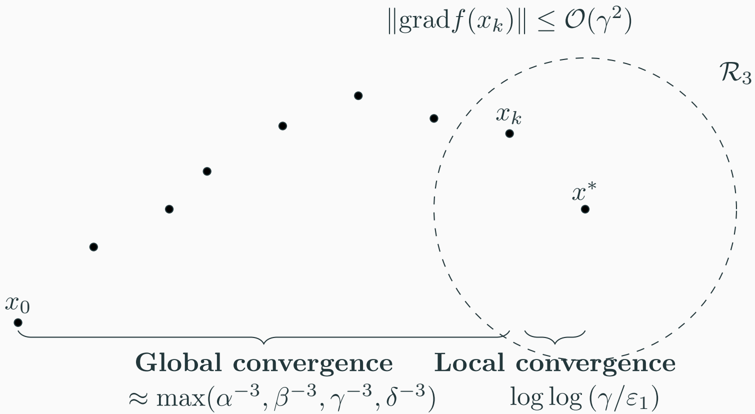
- Rayley quotient for eigenvalues (Sun et al., 2015)
- Burer-Monteiro Decomposition (Boumal et al., 2020; Luo and Trillos, 2022)
- Neural networks (El Mehdi Achour and Gerchinovitz, 2021; Ubl et al., 2022)
- Dictionary Learning (Sun et al., 2017; Qu et al., 2019)
- Matrix completion (Ge et al., 2016; Li and Tang, 2017)
- For more, see `https://sunju.org/research/nonconvex/`

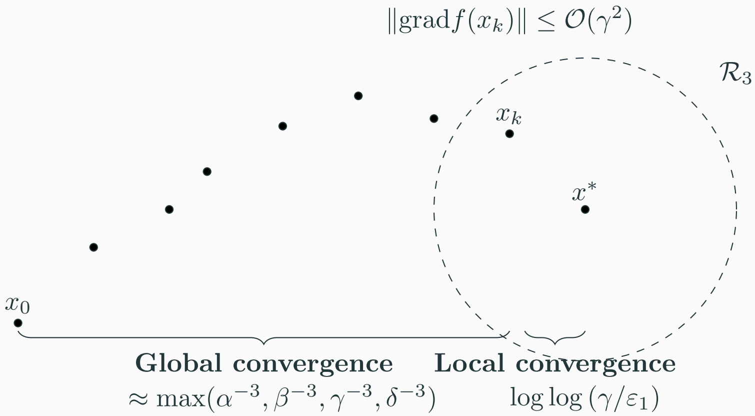
Complexity of Riemannian trust-region for strict saddle

Theorem [Goyens and Royer, 2024]
Let $f\colon \mathcal{M} \to \mathbb{R}$ be an $(\alpha, \beta, \gamma, \delta)$-strict saddle function and $\varepsilon_1 > 0$. Under typical smoothness assumptions, RTR with exact subproblem minimization finds $x \in \mathcal{M}$ such that $\|\mathrm{grad} f(x)\| \leq \varepsilon_1$ and $\mathrm{Hess} f(x) \succeq \gamma \mathrm{Id}$ in at most

$$\mathcal{O}\left(1/\min(\alpha^2\beta, \alpha\gamma^2, \beta^3, \beta\gamma^2, \gamma^3, \delta\gamma^2, \alpha^2\gamma, \beta^2\gamma) + \log\log\left(\gamma/\varepsilon_1\right)\right)$$

iterations.

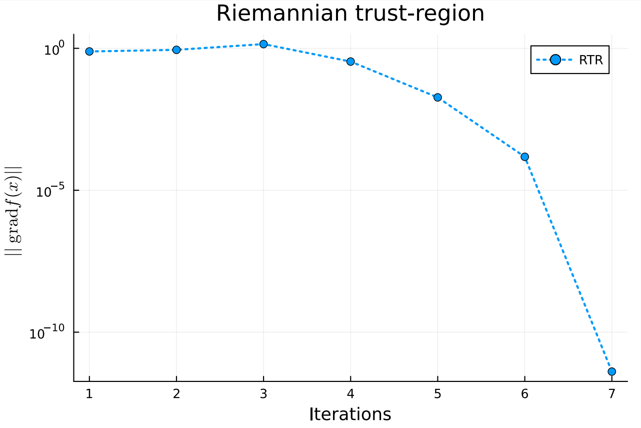

$\|\mathrm{grad} f(x_k)\| \leq \mathcal{O}(\gamma^2)$
$\mathcal{R}_3$
x_k
x^*
x_0
Global convergence
$\approx \max(\alpha^{-3}, \beta^{-3}, \gamma^{-3}, \delta^{-3})$
Local convergence
$\log \log (\gamma/\varepsilon_1)$



- Improvement from $\mathcal{O}(\varepsilon^{-3})$ to $\mathcal{O}\left(\log\log(\varepsilon^{-1})\right)$, closer to practical behaviour (fast local convergence)
- Complexity independent of ε_2 and practically independent of ε_1.

Practical behaviour of trust-region vs guarantees

	$\varepsilon = 10^{-3}$	$\varepsilon = 10^{-6}$
$\mathcal{O}(\varepsilon^{-3})$	10^9	10^{18}
$\log\log(\varepsilon^{-1})$	2	3



Complexity of Riemannian trust-region for strict saddle

Theorem [Goyens and Royer, 2024]
Let $f\colon \mathcal{M} \to \mathbb{R}$ be a $(\alpha, \beta, \gamma, \delta)$-strict saddle function. Under typical smoothness assumptions, RTR with exact subproblem minimization finds $x \in \mathcal{M}$ such that $\|\mathrm{grad} f(x)\| \leq \varepsilon_1$ and $\mathrm{Hess} f(x) \succeq \gamma \mathrm{Id}$ in at most

$$\mathcal{O}\left(1/\min(\alpha^2\beta, \alpha\gamma^2, \beta^3, \beta\gamma^2, \gamma^3, \delta\gamma^2, \alpha^2\gamma, \beta^2\gamma) + \log\log\left(\gamma/\varepsilon_1\right)\right)$$

iterations.

- Related work: (Sun et al., 2018) and (O'Neill and Wright, 2023)

Complexity guarantees mimic strongly convex optimization

(Boyd and Vandenberghe, 2004) For $f\colon \mathbb{R}^n \to \mathbb{R}$ that is γ-strongly convex over $\mathbb{R}^n$ with Lipschitz continuous Hessian, the Newton method with Armijo backtracking requires at most

$$\mathcal{O}\left(\gamma^{-5} + \log\log\left(\varepsilon^{-1}\right)\right)$$

iterations to find a point such that $\|\nabla f(x)\| \le \varepsilon$

Strict saddle RTR $\approx \mathcal{O}\left(\max(\alpha^{-3}, \beta^{-3}, \gamma^{-3}, \delta^{-3}) + \log\log\left(\gamma\varepsilon^{-1}\right)\right)$

Inexact minimization of the trust-region subproblems

Inexact Riemannian trust-region

Algorithm 2 RTR with exact subproblem minimization

1: **for** $k = 1, 2, \ldots$ **do**
2: Compute s_k as a solution to the trust-region subproblem

$$s_k \in \underset{s \in \mathrm{T}_{x_k}\mathcal{M}}{\arg\min}\, f(x_k) + \langle \operatorname{grad} f(x_k), s_k \rangle + \frac{1}{2} \langle s_k, H_k s_k \rangle \text{ subject to } \|s\| \leq \Delta_k,$$

3: Use step s_k if f decreases sufficiently
4: Update trust-region radius Δ_k as needed
5: **end for**

Approximately minimizing the model in the subproblems ?

- Vanilla tCG does not provably avoid strict saddle points
- Adaptation of tCG that exploits the strict saddle geometry
- Similar complexity result when α, β, γ are known
- Lanczos's method to estimate negative eigenvalues

Conclusion

Main points:

- Complexity result for the Riemannian trust-region with exact and inexact minimization of the subproblems on strict saddle functions
- The worst-case complexity depends on the landscape parameters $(\alpha, \beta, \gamma, \delta)$ instead of the problem accuracy ε
- Second-order method: benefits from the local quadratic convergence of Newton's method.
- Remaining challenges : non-isolated minimizers, estimation of landscape parameters

Thank you !

Goyens and Royer, 2024, *Riemannian trust-region methods for strict saddle functions with complexity guarantees*, `https://arxiv.org/abs/2402.07614`.

Additional material

Ingredients of the proofs

- Cauchy step: $f(x_k) - f(x_{k+1}) \geq \mathcal{O}(\alpha^2)$
  $\longrightarrow$ follows from (Boumal, 2023) and $\|\mathrm{grad} f(x_k)\| \geq \alpha$
- Eigenstep: $f(x_k) - f(x_{k+1}) \geq \mathcal{O}(\beta^3)$
  $\longrightarrow$ follows from (Boumal, 2023) and $\lambda_{\min}(\mathrm{Hess} f(x_k)) \leq -\beta$
- Convex model step: $f(x_k) - f(x_{k+1}) \geq \mathcal{O}(\gamma^3)$
  $\longrightarrow$ Adaptation of (Curtis et al., 2021) to manifolds
- Local phase: quadratic convergence in $\log\log$ steps
  $\longrightarrow$ quantifying when the local phase becomes a pure Newton sequence (g-convexity + ideas from Cartis and Shek) with quadratic convergence $\|\mathrm{grad} f(x_{k+1})\| \leq c \, \|\mathrm{grad} f(x_k)\|^2$ (Absil et al., 2008)

Inexact minimization of the trust-region subproblems

truncated Conjugate Gradient algorithm:

run the classical CG algorithm on the quadratic

$$s \mapsto \langle \operatorname{grad} f(x_k), s \rangle + \frac{1}{2} \langle s, \operatorname{Hess} f(x_k) s \rangle$$

- if CG iterate has negative curvature $\langle y_j, H_k y_j \rangle < 0$, stop CG and use direction y_j to decrease the model
- if CG iterate leaves trust region $\|y_j\| > \Delta_k$, stop on boundary
- if CG residual is small enough, stop CG and return

truncated Conjugate Gradient algorithm

- Used in practice for large-scale problems, unfortunately **no existing results** of convergence to ε-Second Order Critical Point for traditional tCG

- We make minimal adjustements to tCG on strict saddle functions to obtain good complexity guarantees for convergence to ε-Second Order Critical Point

Inexact minimization of the subproblems

Assume α, β, γ are known

- Run truncated conjugate gradient (CG) to the linear system $H_k s = -g_k$ as long as H_k appears γ-strongly convex in CG directions
- Stop if the residual $\|H_k y_j + g_k\|$ is small enough or $\|y_j\| > \Delta_k$.
- If curvature below β is encountered in H_k, take a negative curvature step such that $\|s_k\| = \Delta_k$.
- If $H_k \succeq \gamma \mathrm{Id}$, CG reaches a small residual in at most $\min(n, \tilde{\mathcal{O}}(\gamma^{-1/2}))$ matrix-vector products (Royer et al., 2020).
- If $\lambda_{\min}(H_k) \leq -\beta$, the Lanczos method finds a direction of curvature $-\beta$ in at most $\min(n, \tilde{\mathcal{O}}(\ln(n/p)\beta^{-1/2}))$ matrix-vector products with probability p (Royer et al., 2020).

$\Longrightarrow$ similar complexity guarantees which count the total number of matrix-vector products

References

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, December 2008. ISBN 0-691-13298-4.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164.

Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. Deterministic Guarantees for Burer-Monteiro Factorizations of Smooth Semidefinite Programs. *Communications on Pure and Applied Mathematics*, 73(3):581–608, 2020.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0-521-83378-7.

Frank E. Curtis, Daniel P. Robinson, Clément W. Royer, and Stephen J. Wright. Trust-Region Newton-CG with Strong Second-Order Complexity Guarantees for Nonconvex Optimization. *SIAM Journal on Optimization*, 31(1):518–544, January 2021. doi: 10.1137/19M130563X.

François Malgouyres El Mehdi Achour and Sébastien Gerchinovitz. Global minimizers, strict and non-strict saddle points, and implicit regularization for deep linear neural networks. *arXiv preprint arXiv:2107.13289*, 2021.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix Completion has No Spurious Local Minimum. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2973–2981. Curran Associates, Inc., 2016.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

Q. Li and G. Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1235–1239, November 2017. doi: 10.1109/GlobalSIP.2017.8309158.

Yuetian Luo and Nicolas Garcia Trillos. Nonconvex Matrix Factorization is Geodesically Convex: Global Landscape Analysis for Fixed-rank Matrix Optimization From a Riemannian Perspective, November 2022.

Michael O'Neill and Stephen J Wright. A Line-Search Descent Algorithm for Strict Saddle Functions