Worst-Case Complexity for the Minimization of Strict Saddle Functions on Manifolds

Florentin Goyens
joint with
Clément Royer

SIOPT, Seattle
June 2, 2023

Problem (P)

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

where $f\colon \mathcal{M} \to \mathbb{R}$ is smooth and nonconvex.

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

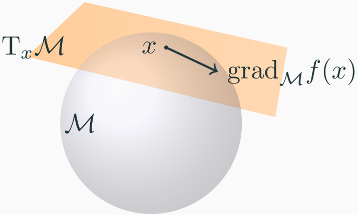
How many iterations of an optimization algorithm are required in the worst-case to reach an approximate solution of (P) from an arbitrary initial $x_0 \in \mathcal{M}$?

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

Background: Complexity without the strict saddle assumption

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

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

Optimality conditions for Riemannian optimization algorithms

First-order critical points

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

Second-order critical points

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

Their approximate version:

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

Quick example: Complexity of gradient descent

Gradient descent satisfies

$$f(x_k) - f(x_{k+1}) \geq c \cdot \|\mathrm{grad} f(x_k)\|^2 \text{ for all } k \geq 0.$$

As long as the algorithm has not converged $\|\mathrm{grad} f(x_k)\| \geq \varepsilon$,

$$\begin{cases} f(x_0) - f(x_1) \geq c \cdot \varepsilon^2 \\ f(x_1) - f(x_2) \geq c \cdot \varepsilon^2 \\ \qquad \vdots \\ f(x_{N-1}) - f(x_N) \geq c \cdot \varepsilon^2 \end{cases} \implies N \leq \frac{f(x_0) - f(x_N)}{c\varepsilon^2} = \mathcal{O}(\varepsilon^{-2}).$$

Complexity of Riemannian algorithms

?: 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 $\mathcal{O}(\varepsilon_1^{-2})$ 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 $\mathcal{O}\left(\max\left(\varepsilon_1^{-2}, \varepsilon_2^{-3}\right)\right)$ iterations.

Riemannian trust-region (RTR)

Algorithm 1 Riemannian trust-region (RTR)

Given: Tolerance $\varepsilon_g > 0$, $x_0 \in \mathcal{M}$, trust-region radius $\Delta_0 > 0$, $\bar{\Delta} > 0$, constants $0 < \eta_1 < \eta_2 < 1$ and $0 < \tau_1 < 1 < \tau_2$.
for $k = 1, 2, \ldots$ **do**
  Find step $s_k \in \mathrm{T}_{x_k}\mathcal{M}$ which minimizes (approximately)

$$m_k(s) = f(x_k) + \langle s, \operatorname{grad} f(x_k)\rangle + \frac{1}{2}\langle s, H_k(s)\rangle \text{ subjet to } \|s\| \leq \Delta_k. \tag{4.5}$$

  Compute $\rho = \dfrac{f(x_k) - f \circ R_{x_k}(s_k)}{m_k(0) - m_k(s_k)}$ and apply $x_{k+1} = \begin{cases} x_k & \text{if } \rho < \eta_1 \\ R_{x_k}(s_k) & \text{if } \eta_1 \leq \rho \end{cases}$

$$\Delta_{k+1} = \begin{cases} \tau_1 \Delta_k & \text{if } \rho < \eta_1 & \text{[unsuccessful]} \\ \Delta_k & \text{if } \eta_1 \leq \rho \leq \eta_2 & \text{[successful]} \\ \tau_2 \Delta_k & \text{if } \rho > \eta_2 & \text{[very successful]} \end{cases} \tag{4.6}$$

  $k \leftarrow k + 1$
end for

- $g_k = \operatorname{grad} f(x_k)$ and $H_k = \nabla^2 (f \circ R_{x_k}) = \operatorname{Hess} f(x_k)$ (second-order accurate model and second-order retraction)

Riemannian trust-region (RTR)

Algorithm 1 Riemannian trust-region (RTR)

Given: Tolerance $\varepsilon_g > 0$, $x_0 \in \mathcal{M}$, trust-region radius $\Delta_0 > 0$, $\bar{\Delta} > 0$, constants $0 < \eta_1 < \eta_2 < 1$ and $0 < \tau_1 < 1 < \tau_2$.

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

Find step $s_k \in \mathrm{T}_{x_k}\mathcal{M}$ which minimizes (approximately)

$$m_k(s) = f(x_k) + \langle s, \mathrm{grad} f(x_k) \rangle + \frac{1}{2} \langle s, H_k(s) \rangle \text{ subjet to } \|s\| \leq \Delta_k. \tag{4.5}$$

Compute $\rho = \dfrac{f(x_k) - f \circ R_{x_k}(s_k)}{m_k(0) - m_k(s_k)}$ and apply $x_{k+1} = \begin{cases} x_k & \text{if } \rho < \eta_1 \\ R_{x_k}(s_k) & \text{if } \eta_1 \leq \rho \end{cases}$

$$\Delta_{k+1} = \begin{cases} \tau_1 \Delta_k & \text{if } \rho < \eta_1 & \text{[unsuccessful]} \\ \Delta_k & \text{if } \eta_1 \leq \rho \leq \eta_2 & \text{[successful]} \\ \tau_2 \Delta_k & \text{if } \rho > \eta_2 & \text{[very successful]} \end{cases} \tag{4.6}$$

$k \leftarrow k + 1$

end for

- $g_k = \mathrm{grad} f(x_k)$ and $H_k = \nabla^2 (f \circ R_{x_k}) = \mathrm{Hess} f(x_k)$ (second-order accurate model and second-order retraction)
- We adapt this algorithm to **strict saddle functions**

Part 2: Complexity with the strict saddle assumption

Saddle points make life difficult

Definition
If $\operatorname{grad} f(x) = 0$ but $x \in \mathcal{M}$ is not a local minimum, then x is a saddle point.

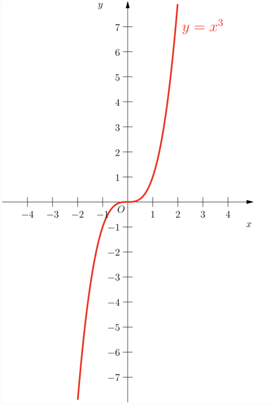


Figure 1: Saddle point with $\nabla^2 f(x) = 0$

- If $\lambda_{\min}\left(\nabla^2 f(x)\right) < 0$, then x is a **strict saddle point**
- Strict saddle points can provably be escaped by algorithms!

Strict saddle functions on manifolds

Definition (Robust strict saddle functions on manifolds)
There exists positive constants $\alpha, \beta, \gamma, \delta$ such that, at any point $x \in \mathcal{M}$, at least one of the following holds:

1. $\|\mathrm{grad} f(x)\| \geq \alpha$ (large gradient);
2. $\lambda_{\min}(\mathrm{Hess} f(x)) \leq -\beta$ (negative curvature of the Hessian);
3. there exists a local minimum x^* such that x belongs to the set $S = \{y \in \mathcal{M} \colon \mathrm{dist}(x^*, y) \leq 2\delta\}$ which is geodesically convex, with $\lambda_{\min}(\mathrm{Hess} f(y)) \geq \gamma$ at every $y \in S$ (local geodesic strong convexity).

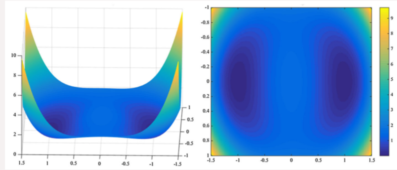
Strict saddle functions appear in many problems of interest! (a.k.a. benign non-convexity)

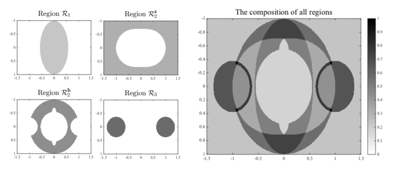
Examples of strict saddle problems

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 $(c, c/(n \log m), c, c/(n \log m))$ strict saddle function for some constant c (**??**).





Examples of strict saddle problems

Strict saddle properties have been investigated in many other applications, such as:

- Rayley quotient for eigenvalues (?)
- Burer-Monteiro Decomposition (??)
- Neural networks (??)
- Dictionary Learning (??)
- Matrix completion (??)
- For more, see `https://sunju.org/research/nonconvex/`

Our landscape-aware algorithm for strict saddle functions

Take a step that is appropriate for the local landscape

1. If $\|\mathrm{grad} f(x)\| \geq \alpha$, take gradient step
2. If $\lambda_{\min}(\mathrm{Hess} f(x)) \preceq -\beta \mathrm{Id}$, take negative curvature step
3. If $\mathrm{Hess} f(x) \succeq \gamma \mathrm{Id}$, take (regularized) Newton step.

We embed these 3 steps in a single trust-region method

Algorithm 2 Exact strict saddle RTR algorithm

Given: Tolerance $\varepsilon_g > 0$, Constants α, β of the strict saddle function, $x_0 \in \mathcal{M}$, trust-region radius $\Delta_0 > 0$, $\bar{\Delta} > 0$, constants $0 < \eta_1 < \eta_2 < 1$ and $0 < \tau_1 < 1 < \tau_2$.

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

  if $\|\mathrm{grad} f(x_k)\| \geq \alpha$ **then**

    Compute the Cauchy point: $s_k = \arg\min_{s \in \mathrm{T}_{x_k}\mathcal{M}} \langle s, g_k \rangle$ subjet to $\|s\| = \Delta_k$.

  else if $\lambda_{\min}(\mathrm{Hess} f(x_k)) \leq -\beta$ **then**

    Compute s_k as the eigenstep, satisfying

$$\|s_k\| = \Delta_k, \qquad \langle s_k, g_k \rangle \leq 0 \qquad \text{and} \qquad \langle s_k, H_k s_k \rangle \leq -\beta \|s_k\|^2 .$$

  else $\triangleright H_k \succeq \gamma\, \mathrm{Id}$

    Compute s_k as the exact solution to

$$\min_{s \in \mathrm{T}_{x_k}\mathcal{M}} f(x_k) + \langle s, \mathrm{grad} f(x_k) \rangle + \frac{1}{2} \langle s, H_k s \rangle \ \text{ subjet to } \ \|s\| \leq \Delta_k.$$

  end if

  Compute $\rho = \dfrac{f(x_k) - f \circ R_{x_k}(s_k)}{m_k(0) - m_k(s_k)}$ and apply $x_{k+1} = \begin{cases} x_k & \text{if } \rho < \eta_1 \\ R_{x_k}(s_k) & \text{if } \eta_1 \leq \rho \end{cases}$

$$\Delta_{k+1} = \begin{cases} \tau_1 \Delta_k & \text{if } \rho < \eta_1 & \text{[unsuccessful]} \\ \Delta_k & \text{if } \eta_1 \leq \rho \leq \eta_2 & \text{[successful]} \\ \tau_2 \Delta_k & \text{if } \rho > \eta_2 & \text{[very successful]} \end{cases}$$

  $k \leftarrow k + 1$

end for

Complexity of Riemannian trust-region for strict saddle

Theorem [G. and Royer]
Let $f\colon \mathcal{M} \to \mathbb{R}$ be a $(\alpha, \beta, \gamma, \delta)$-strict saddle function on the manifold $\mathcal{M}$. If the pullback $f \circ R$ is twice Lipschitz continuously differentiable and R second-order, for any $x_0 \in \mathcal{M}$ the strict saddle Riemannian trust-region algorithm finds a point $x \in \mathcal{M}$ such that $\|\mathrm{grad} f(x)\| \leq \varepsilon$ and $\mathrm{Hess} f(x) \succeq \gamma \mathrm{Id}$ in at most

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

iterations.

Complexity of Riemannian trust-region for strict saddle

Theorem [G. and Royer]
Let $f\colon \mathcal{M} \to \mathbb{R}$ be a $(\alpha, \beta, \gamma, \delta)$-strict saddle function on the manifold $\mathcal{M}$. If the pullback $f \circ R$ is twice Lipschitz continuously differentiable and R second-order, for any $x_0 \in \mathcal{M}$ the strict saddle Riemannian trust-region algorithm finds a point $x \in \mathcal{M}$ such that $\|\mathrm{grad} f(x)\| \leq \varepsilon$ and $\mathrm{Hess} f(x) \succeq \gamma \mathrm{Id}$ in at most

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

iterations.

- Complexity is with respect to strict saddle parameters and no longer depends on ε (up to log-log factor)
- Similar result in (**?**) for a first-order algorithm, we improve the complexity in the local phase

Ingredients of the proofs

- Cauchy step: $f(x_k) - f(x_{k+1}) \geq \mathcal{O}(\alpha^2)$
  $\longrightarrow$ follows from (Boumal, 2023) and $\|\text{grad} f(x_k)\| \geq \alpha$
- Eigenstep: $f(x_k) - f(x_{k+1}) \geq \mathcal{O}(\beta^3)$
  $\longrightarrow$ follows from (Boumal, 2023) and $\lambda_{\min}(\text{Hess} f(x_k)) \leq -\beta$

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 (**?**) to manifolds

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 (**?**) to manifolds
- Local phase: quadratic convergence in $\log\log$ steps
  $\longrightarrow$ quantifying when the local phase becomes a pure Newton method (g-convexity + ideas from Cartis and Shek) with quadratic convergence $\|\mathrm{grad} f(x_{k+1})\| \leq c \|\mathrm{grad} f(x_k)\|^2$ (**?**)

Complexity mimics convex optimization

Strict saddle RTR:

$$\mathcal{O}\left(\max(\alpha^{-2}\beta^{-1}, \alpha^{-2}\gamma^{-1}, \beta^{-3}, \gamma^{-3}, \gamma^{-2}\delta^{-1}) + \log\log\left(\gamma/\varepsilon\right)\right).$$

(**?**) For $f\colon \mathbb{R}^n \to \mathbb{R}$ that is γ-strongly convex over $\mathbb{R}^n$ and Lipschitz 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 $\|\nabla f(x)\| \leq \varepsilon$

The Steihaug-Toint approach: truncated CG

We don't want to compute $\lambda_{\min}(H_k)$ at every iteration to branch between cases 2 and 3:

$\Longrightarrow$ Approximate solutions of the trust-region subproblem

The Steihaug-Toint approach: truncated CG

We don't want to compute $\lambda_{\min}(H_k)$ at every iteration to branch between cases 2 and 3:

$\Longrightarrow$ Approximate solutions of the trust-region subproblem

- Apply conjugate gradient (CG) to the linear system $H_k s = -g_k$ as long as H_k appears γ-strongly convex in CG directions

The Steihaug-Toint approach: truncated CG

We don't want to compute $\lambda_{\min}(H_k)$ at every iteration to branch between cases 2 and 3:

$\Longrightarrow$ Approximate solutions of the trust-region subproblem

- Apply 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 $\|H_k s + g_k\|$ is small enough or $\|s\| = \Delta_k$.
- If $H_k \succeq \gamma I$, CG reaches a small residual in at most $\min(n, \tilde{\mathcal{O}}(\gamma^{-1/2}))$ matrix-vector products (**?**).
- When $H_k \not\succeq 0$, if curvature below γ is encountered in H_k, take a negative curvature step such that $\|s_k\| = \Delta_k$.
- 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 (**?**).

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

Conclusion

Main points:

- Landscape-aware second-order optimization algorithm for strict saddle functions
- The worst-case complexity depends on the landscape parameters $(\alpha, \beta, \gamma, \delta)$ instead of the problem accuracy ε
- Quadratic local convergence of second-order method.
- Estimation of the landscape parameters, see (**?**)

Questions

Thank you !

References

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