Convergence and Trade-Offs in Riemannian Gradient Descent and Proximal Point

Technische Universität Berlin, Zuse Institute Berlin, Carlos III University

Our Riemannian Optimization Setting

Function $f : \mathcal{M} \to \mathbb{R}$

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

Smoothness and (possibly μ-strong) geodesic convexity:

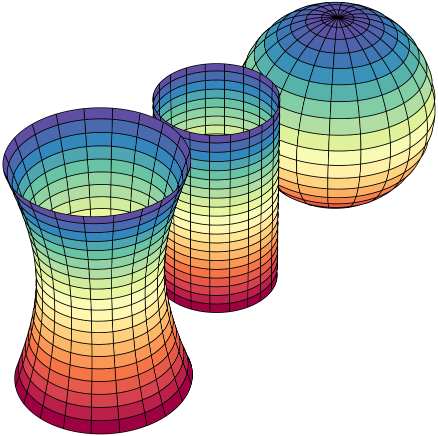
$$\mu \preccurlyeq \nabla^2 f(x) \preccurlyeq L.$$

Riemannian manifold $\mathcal{M}$:

- Uniquely geodesic.
- Geodesically convex.
- Sectional curvature in $[\kappa_{\min}, \kappa_{\max}]$.

First-order methods

Access to an oracle $x \mapsto \{f(x), \nabla f(x)\}$.



Why?

- Constrained problems to unconstrained ones on a manifold.
- Euclidean non-convex problems can be geodesically convex on a manifold with the right metric.

Applications:

- **Fixed-rank matrices:** Low-rank matrix factorization.
- **SPD matrices:** Gaussian mixtures, covariance estimation, operator scaling.
- **Stiefel manifold (orthonormal matrices)**: Sparse PCA, DNNs with orthogonality constraints.
- **Sphere:** PCA.

Some assumptions in Riemannian optimization to improve on

- Assume iterates remain bounded and then show convergence.
  $\rightarrow$ Compatible with the algorithm **diverging**!

Some assumptions in Riemannian optimization to improve on

- Assume iterates remain bounded and then show convergence.
  $\rightarrow$ Compatible with the algorithm **diverging**!
- **Less nice:** The algorithm knows the bound, uses its value and the **iterates depend on it**.
  $\rightarrow$ Many circular arguments!

Some assumptions in Riemannian optimization to improve on

- Assume iterates remain bounded and then show convergence.
  $\rightarrow$ Compatible with the algorithm **diverging**!
- **Less nice:** The algorithm knows the bound, uses its value and the **iterates depend on it**.
  $\rightarrow$ Many circular arguments!
- **Tricky:** *Assume L-smoothness and μ-strong convexity globally...*

Some assumptions in Riemannian optimization to improve on

- Assume iterates remain bounded and then show convergence.
  $\rightarrow$ Compatible with the algorithm **diverging**!
- **Less nice:** The algorithm knows the bound, uses its value and the **iterates depend on it**.
  $\rightarrow$ Many circular arguments!
- **Tricky:** *Assume L-smoothness and μ-strong convexity globally...*
  $\rightarrow$ Impossible for manifolds with curvature $\leq c < 0$. E.g. In $B(0, R) \subset \mathcal{H}^d$ it's $\frac{L}{\mu} = \Omega(R+1)$.

Some assumptions in Riemannian optimization to improve on

- Assume iterates remain bounded and then show convergence.
  $\rightarrow$ Compatible with the algorithm **diverging**!
- **Less nice:** The algorithm knows the bound, uses its value and the **iterates depend on it**.
  $\rightarrow$ Many circular arguments!
- **Tricky:** *Assume L-smoothness and μ-strong convexity globally...*
  $\rightarrow$ Impossible for manifolds with curvature $\leq c < 0$. E.g. In $B(0, R) \subset \mathcal{H}^d$ it's $\frac{L}{\mu} = \Omega(R+1)$.
- *...or in a local region without guaranteeing iterates stay in it.*
  $\rightarrow$ **Need for ensuring quantified bounded iterates!**.

Examples from Prior Work

- $R \stackrel{\text{def}}{=} d(x_0, x^*)$
- $D \stackrel{\text{def}}{=} \max_{t\in[T]} d(x_t, x^*)$
- Geometric constants $\zeta_D = \Theta(D+1)$, $\delta_D \in (0,1]$. In a ball $B(x_0, \tilde{R})$, it is:

$$\nabla_x \left(\frac{1}{2} d(x, x_0)^2\right) = -\operatorname{Exp}_x^{-1}(x_0). \quad \text{and} \quad \delta_{\tilde{R}} \preccurlyeq \nabla^2 \left(\frac{1}{2} d(x, x_0)^2\right) \preccurlyeq \zeta_{\tilde{R}}$$

	convex	str. convex	D
Euclidean GD	$O(\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\frac{L}{\mu})$	R
RGD (Udr94)	-	$\widetilde{O}(\frac{L}{\mu})$	?
RGD (ZS16)	$O(\zeta_D \frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta_D + \frac{L}{\mu})$	?
RGD (MP23)	$O(\frac{LD^2}{\varepsilon})$	-	?

Examples from Prior Work

- $R \overset{\text{def}}{=} d(x_0, x^*)$
- $D \overset{\text{def}}{=} \max_{t\in[T]} d(x_t, x^*)$
- Geometric constants $\zeta_D = \Theta(D+1)$, $\delta_D \in (0,1]$. In a ball $B(x_0, \tilde{R})$, it is:

$$\nabla_x \left(\frac{1}{2} d(x, x_0)^2\right) = -\operatorname{Exp}_x^{-1}(x_0). \quad \text{and} \quad \delta_{\tilde{R}} \preccurlyeq \nabla^2 \left(\frac{1}{2} d(x, x_0)^2\right) \preccurlyeq \zeta_{\tilde{R}}$$

	convex	str. convex	D
Euclidean GD	$O(\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\frac{L}{\mu})$	R
RGD (Udr94)	-	$\widetilde{O}(\frac{L}{\mu})$	?
RGD (ZS16)	$O(\zeta_D \frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta_D + \frac{L}{\mu})$	?
RGD (MP23)	$O(\frac{LD^2}{\varepsilon})$	-	?

Examples from Prior Work

- $R \overset{\text{def}}{=} d(x_0, x^*)$
- $D \overset{\text{def}}{=} \max_{t\in[T]} d(x_t, x^*)$
- Geometric constants $\zeta_D = \Theta(D+1)$, $\delta_D \in (0,1]$. In a ball $B(x_0, \tilde{R})$, it is:

$$\nabla_x \left(\frac{1}{2} d(x, x_0)^2\right) = -\operatorname{Exp}_x^{-1}(x_0). \quad \text{and} \quad \delta_{\tilde{R}} \preccurlyeq \nabla^2 \left(\frac{1}{2} d(x, x_0)^2\right) \preccurlyeq \zeta_{\tilde{R}}$$

	convex	str. convex	D
Euclidean GD	$O(\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\frac{L}{\mu})$	R
RGD (Udr94)	-	$\widetilde{O}(\frac{L}{\mu})$	?
RGD (ZS16)	$O(\zeta_D \frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta_D + \frac{L}{\mu})$	?
RGD (MP23)	$O(\frac{LD^2}{\varepsilon})$	-	?

Our Riemannian Gradient Descent Results

Recall: $R := d(x_0, x^*)$, $D := \max_{t\in[T]} d(x_t, x^*)$

Riemannian Gradient Descent (RGD): $x_{t+1} \leftarrow \mathrm{Exp}_{x_t}(-\eta \nabla f(x_t))$

- For $\eta = 1/L$: Maximal distance to optimizer is at most $D = O(R\zeta_R)$
  - Hyperbolic space: $D = O(R)$. **And we match Euclidean rates!**
- Mirror-descent–style analysis. In hyperbolic space: maximal optimality gap at distance R is $O(\frac{LR^2}{\zeta_R})$.

Our Riemannian Gradient Descent Results

Recall: $R := d(x_0, x^*)$, $D := \max_{t\in[T]} d(x_t, x^*)$

Riemannian Gradient Descent (RGD): $x_{t+1} \leftarrow \mathrm{Exp}_{x_t}(-\eta\nabla f(x_t))$

- For $\eta = 1/L$: Maximal distance to optimizer is at most $D = O(R\zeta_R)$
  - Hyperbolic space: $D = O(R)$. **And we match Euclidean rates!**
- For $\eta = 1/(L\zeta_R)$, $D = R$. RGD is **quasi-nonexpansive**: $d(x_{t+1}, x^*) \leq d(x_t, x^*)$ for all t.
- Mirror-descent–style analysis. In hyperbolic space: maximal optimality gap at distance R is $O(\frac{LR^2}{\zeta_R})$.
- Polyak step-size type of analysis.

Our Riemannian Gradient Descent Results

Recall: $R := d(x_0, x^*)$, $D := \max_{t\in[T]} d(x_t, x^*)$

Riemannian Gradient Descent (RGD): $x_{t+1} \leftarrow \mathrm{Exp}_{x_t}(-\eta \nabla f(x_t))$

- For $\eta = 1/L$: Maximal distance to optimizer is at most $D = O(R\zeta_R)$
  - Hyperbolic space: $D = O(R)$. **And we match Euclidean rates!**
- For $\eta = 1/(L\zeta_R)$, $D = R$. RGD is **quasi-nonexpansive**: $d(x_{t+1}, x^*) \leq d(x_t, x^*)$ for all t.
- Mirror-descent–style analysis. In hyperbolic space: maximal optimality gap at distance R is $O(\frac{LR^2}{\zeta_R})$.
- Polyak step-size type of analysis.
- Convergence rates for Composite RGD:

$$x_{t+1} \leftarrow \operatorname*{arg\,min}_{x\in\mathcal{X}} \left\{ \langle \nabla f(x_t), \mathrm{Exp}_{x_t}^{-1}(x) \rangle + \frac{L}{2} d(x, x_t)^2 + g(x) \right\}.$$

Our Proximal Point Results

Recall: $R := d(x_0, x^*)$, $D := \max_{t\in[T]} d(x_t, x^*)$

Riemannian Proximal Point Algorithm: $\mathrm{prox}_\eta(x) \stackrel{\text{def}}{=} \arg\min_{y\in\mathcal{X}} \left\{ f(y) + \frac{1}{2\eta} d(y,x)^2 \right\}$

- Rates for general manifolds. Only Hadamard before.
- Moreau envelope is not g-convex in positive curvature but still we show $O(\frac{1}{T})$ convergence.

Our Proximal Point Results

Recall: $R := d(x_0, x^*)$, $D := \max_{t\in[T]} d(x_t, x^*)$

Riemannian Proximal Point Algorithm: $\mathrm{prox}_\eta(x) \stackrel{\text{def}}{=} \arg\min_{y\in\mathcal{X}} \left\{ f(y) + \frac{1}{2\eta} d(y,x)^2 \right\}$

- Rates for general manifolds. Only Hadamard before.
- Moreau envelope is not g-convex in positive curvature but still we show $O(\frac{1}{T})$ convergence.
- The prox operator in quasi-nonexpansive.
- Bounded iterates!

Our Proximal Point Results

Recall: $R := d(x_0, x^*)$, $D := \max_{t\in[T]} d(x_t, x^*)$

Riemannian Proximal Point Algorithm: $\mathrm{prox}_\eta(x) \stackrel{\mathrm{def}}{=} \arg\min_{y\in\mathcal{X}} \left\{ f(y) + \frac{1}{2\eta} d(y,x)^2 \right\}$

- Rates for general manifolds. Only Hadamard before.
- Moreau envelope is not g-convex in positive curvature but still we show $O(\frac{1}{T})$ convergence.
- The prox operator in quasi-nonexpansive.
- Bounded iterates!
- The Moreau envelope is (ζ_D/η)-smooth.
- Exploit the ζ_D smoothness of the squared distance.

Our Proximal Point Results

Recall: $R := d(x_0, x^*)$, $D := \max_{t\in[T]} d(x_t, x^*)$

Riemannian Proximal Point Algorithm: $\mathrm{prox}_\eta(x) \stackrel{\mathrm{def}}{=} \arg\min_{y\in\mathcal{X}} \left\{ f(y) + \frac{1}{2\eta} d(y,x)^2 \right\}$

- Rates for general manifolds. Only Hadamard before.
- The prox operator in quasi-nonexpansive.
- The Moreau envelope is (ζ_D/η)-smooth.
- An efficient inexact implementation for smooth functions.
- Moreau envelope is not g-convex in positive curvature but still we show $O(\frac{1}{T})$ convergence.
- Bounded iterates!
- Exploit the ζ_D smoothness of the squared distance.
- By RGD in $\widetilde{O}(\zeta_D)$ or by Composite RGD in $\widetilde{O}(1)$. Monteiro-Svaiter-like criterion for inexactness.

Result Overview and Trade-offs

Min				
Method	g-convex	μ-str. g-convex	D	Needs R?
$\text{RGD}_{L^{-1}}$	$O(\zeta_R^2 \frac{LR^2}{\varepsilon})$	$\widetilde{O}(\frac{L}{\mu})$	$O(R\zeta_R)$	No
$^{\diamondsuit}\text{RGD}_{L^{-1}}$	$O(\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\frac{L}{\mu})$	$O(R)$	No
$^{\dagger}\text{Red. RGD}_{L^{-1}}$	$\widetilde{O}(\zeta_R^2 + \frac{LR^2}{\varepsilon})$	–	$O(R\zeta_R)$	Yes
$\text{RGD}_{L^{-1}\zeta_R^{-1}}$	$O(\zeta_R \frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta_R \frac{L}{\mu})$	R	Yes
RIPPA-CRGD	$\widetilde{O}(\frac{LR^2}{\delta_{2R}\varepsilon})$	$\widetilde{O}(\frac{L}{\delta_{2R}\mu})$	$O(R)$	Yes
†RIPPA-PRGD	$O(\zeta_R^2 \frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta_R^2 \frac{L}{\mu})$	$O(R)$	Yes
Min-Max				
RIPPA-RGDA	$\widetilde{O}(\zeta_R^4 \frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta_R^4 \frac{L}{\mu})$	$O(R\zeta_R)$	No

$^{\diamondsuit}$Hyperbolic Space, †Hadamard manifolds.

Desiderata

- Best oracle complexity: $O(LR^2/\varepsilon)$ and $\widetilde{O}(L/\mu)$ in the convex and strongly convex setting.
- No knowlegde of R required to set the step-size.
- Efficiently computable iterations.
- Best bound on D: L & μ may grow with D and are not equal between rows.

Outlook

- Experimental results: No increase in distance observed.
- Is RGD with $\eta = \frac{1}{L}$ quasi-nonexpansive?
- Can we achieve the best of all worlds? I.e., best of our rates $O(LR^2/\varepsilon)$ and $\widetilde{O}(L/\mu)$, best bound $O(R)$ on iterates, efficiently implementable, no knowledge of R.

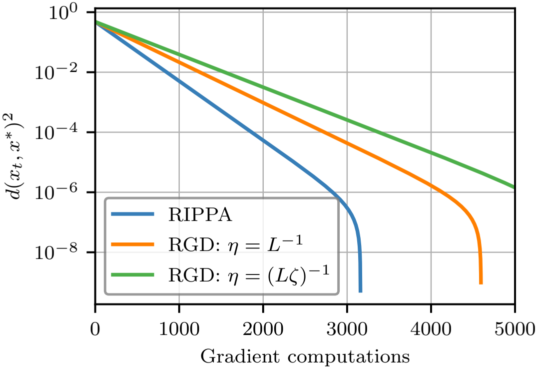


Figure: Karcher mean with $n = 1000$ centers in $\mathbb{S}_+^{100}$.