Riemannian Implicit Optimism with Applications to Min-Max Problems

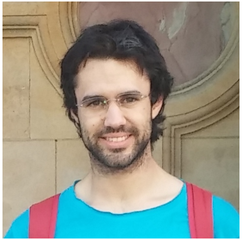
Christophe Roux

21.5.2025

Collaborators & References

Christophe Roux*, David Martínez-Rubio* , Sebastian Pokutta (2024). *Implicit Riemannian Optimism with Applications to Min-Max Problems*. arXiv:2403.10429

Christophe Roux
Zuse Institute Berlin



David Martínez-Rubio
Universidad Carlos III de Madrid

Sebastian Pokutta
Zuse Institute Berlin

Online Optimization

Online Optimization

Online Setup:

For rounds $t = 1, 2, \ldots, T$:

- Learner selects $x_t \in \mathcal{X}$ $\triangleright$*before observing* ℓ_t
- Environment reveals ℓ_t $\triangleright$*possibly adversarial*
- Learner pays $\ell_t(x_t)$

Goal: Minimize cumulative loss $\sum_{t=1}^{T} \ell_t(x_t)$

Regret: For a *fixed* comparator $u \in \mathcal{X}$,

$$R_T(u) = \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(u)$$

Online Gradient Descent:

$$\begin{aligned} x_{t+1} &= \arg\min_{z \in \mathcal{X}} \{\langle \nabla \ell_t(x_t), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2\} \\ &= P_X(x_t - \eta \nabla \ell_t(x_t)) \end{aligned}$$

Standard Results:

- convex, Lipschitz losses ℓ_t
- compact, convex $\mathcal{X}$

$\rightarrow R_T(u) = \Theta(\sqrt{T})$

Book recommendation: *A Modern Introduction to Online Learning*, Francesco Orabona

Online Optimization

Online Setup:
For rounds $t = 1, 2, \ldots, T$:

- Learner selects $x_t \in \mathcal{X}$ ▷*before observing* ℓ_t
- Environment reveals ℓ_t ▷*possibly adversarial*
- Learner pays $\ell_t(x_t)$

Goal: Minimize cumulative loss $\sum_{t=1}^{T} \ell_t(x_t)$

Regret: For a *fixed* comparator $u \in \mathcal{X}$,

$$R_T(u) = \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(u)$$

Online Gradient Descent:

$$\begin{aligned} x_{t+1} &= \arg\min_{z \in \mathcal{X}} \{\langle \nabla \ell_t(x_t), z\rangle + \frac{1}{2\eta}\|z - x_t\|^2\} \\ &= P_X(x_t - \eta \nabla \ell_t(x_t)) \end{aligned}$$

Standard Results:

- convex, Lipschitz losses ℓ_t
- compact, convex $\mathcal{X}$

$\to R_T(u) = \Theta(\sqrt{T})$

Book recommendation: *A Modern Introduction to Online Learning*, Francesco Orabona

Online Optimization

Online Setup:
For rounds $t = 1, 2, \ldots, T$:

- Learner selects $x_t \in \mathcal{X}$ $\triangleright$*before observing* ℓ_t
- Environment reveals ℓ_t $\triangleright$*possibly adversarial*
- Learner pays $\ell_t(x_t)$

Goal: Minimize cumulative loss $\sum_{t=1}^{T} \ell_t(x_t)$

Regret: For a *fixed* comparator $u \in \mathcal{X}$,

$$R_T(u) = \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(u)$$

Online Gradient Descent:

$$\begin{aligned} x_{t+1} &= \arg\min_{z \in \mathcal{X}} \{ \langle \nabla \ell_t(x_t), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2 \} \\ &= P_X(x_t - \eta \nabla \ell_t(x_t)) \end{aligned}$$

Standard Results:

- convex, Lipschitz losses ℓ_t
- compact, convex $\mathcal{X}$

$\rightarrow R_T(u) = \Theta(\sqrt{T})$

Book recommendation: *A Modern Introduction to Online Learning*, Francesco Orabona

Online Optimization

Online Setup:

For rounds $t = 1, 2, \dots, T$:

- Learner selects $x_t \in \mathcal{X}$ $\triangleright$ *before observing* ℓ_t
- Environment reveals ℓ_t $\triangleright$ *possibly adversarial*
- Learner pays $\ell_t(x_t)$

Goal: Minimize cumulative loss $\sum_{t=1}^{T} \ell_t(x_t)$

Regret: For a *fixed* comparator $u \in \mathcal{X}$,

$$R_T(u) = \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(u)$$

Online Gradient Descent:

$$\begin{aligned} x_{t+1} &= \arg\min_{z \in \mathcal{X}} \{\langle \nabla \ell_t(x_t), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2\} \\ &= P_X(x_t - \eta \nabla \ell_t(x_t)) \end{aligned}$$

Standard Results:

- convex, Lipschitz losses ℓ_t
- compact, convex $\mathcal{X}$

$\rightarrow R_T(u) = \Theta(\sqrt{T})$

Book recommendation: *A Modern Introduction to Online Learning*, Francesco Orabona

Online Optimization: Optimistic Methods

Can we do better?
Predictable environment, i.e., not *fully* adversarial:
$\rightarrow$ Use a hint $\tilde{\ell}_t \approx \ell_t$ to improve regret

Optimistic Online Setup:
For rounds $t = 1, 2, \ldots, T$:

- Learner chooses hint $\tilde{\ell}_t$
- Learner selects $\tilde{x}_t \in \mathcal{X}$ $\quad\triangleright$ *using the hint*
- Environment reveals ℓ_t
- Learner pays $\ell_t(\tilde{x}_t)$

Optimistic Online Gradient Descent:

$$x_{t+1} = \arg\min_{z \in \mathcal{X}} \langle \nabla \ell_t(x_t), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2$$

$$\tilde{x}_{t+1} = \arg\min_{z \in \mathcal{X}} \langle \nabla \tilde{\ell}_t(x_{t+1}), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2$$

Results:
$R_T(u) = \mathcal{O}(D^2/\eta + \eta V_T)$

- $D \stackrel{\text{def}}{=} \text{diam}(\mathcal{X})$
- η: step size
- $V_T \stackrel{\text{def}}{=} \sum_{t=1}^T \|\nabla \ell_t(x_t) - \nabla \tilde{\ell}_t(x_t)\|_*^2$
- Good hints: $V_T = o(T)$
  $\rightarrow$ regret improves beyond $\mathcal{O}(\sqrt{T})$

Online Optimization: Optimistic Methods

Can we do better?
Predictable environment, i.e., not *fully* adversarial:
$\rightarrow$ Use a hint $\tilde{\ell}_t \approx \ell_t$ to improve regret

Optimistic Online Setup:
For rounds $t = 1, 2, \ldots, T$:

- Learner chooses hint $\tilde{\ell}_t$
- Learner selects $\tilde{x}_t \in \mathcal{X}$ $\quad \triangleright$ *using the hint*
- Environment reveals ℓ_t
- Learner pays $\ell_t(\tilde{x}_t)$

Optimistic Online Gradient Descent:

$$x_{t+1} = \arg\min_{z \in \mathcal{X}} \langle \nabla \ell_t(x_t), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2$$

$$\tilde{x}_{t+1} = \arg\min_{z \in \mathcal{X}} \langle \nabla \tilde{\ell}_t(x_{t+1}), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2$$

Results:
$R_T(u) = \mathcal{O}(D^2/\eta + \eta V_T)$

- $D \overset{\text{def}}{=} \operatorname{diam}(\mathcal{X})$
- η: step size
- $V_T \overset{\text{def}}{=} \sum_{t=1}^{T} \|\nabla \ell_t(x_t) - \nabla \tilde{\ell}_t(x_t)\|_*^2$
- Good hints: $V_T = o(T)$
  $\rightarrow$ regret improves beyond $\mathcal{O}(\sqrt{T})$

Online Optimization: Optimistic Methods

Can we do better?
Predictable environment, i.e., not *fully* adversarial:
$\rightarrow$ Use a hint $\tilde{\ell}_t \approx \ell_t$ to improve regret

Optimistic Online Setup:
For rounds $t = 1, 2, \ldots, T$:

- Learner chooses hint $\tilde{\ell}_t$
- Learner selects $\tilde{x}_t \in \mathcal{X}$ $\triangleright$ *using the hint*
- Environment reveals ℓ_t
- Learner pays $\ell_t(\tilde{x}_t)$

Optimistic Online Gradient Descent:

$$x_{t+1} = \underset{z \in \mathcal{X}}{\arg\min} \langle \nabla \ell_t(x_t), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2$$

$$\tilde{x}_{t+1} = \underset{z \in \mathcal{X}}{\arg\min} \langle \nabla \tilde{\ell}_t(x_{t+1}), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2$$

Results:
$R_T(u) = \mathcal{O}(D^2/\eta + \eta V_T)$

- $D \overset{\text{def}}{=} \operatorname{diam}(\mathcal{X})$
- η: step size
- $V_T \overset{\text{def}}{=} \sum_{t=1}^T \|\nabla \ell_t(x_t) - \nabla \tilde{\ell}_t(x_t)\|_*^2$
- Good hints: $V_T = o(T)$
  $\rightarrow$ regret improves beyond $\mathcal{O}(\sqrt{T})$

Online Optimization: Optimistic Methods

Can we do better?
Predictable environment, i.e., not *fully* adversarial:
$\rightarrow$ Use a hint $\tilde{\ell}_t \approx \ell_t$ to improve regret

Optimistic Online Setup:
For rounds $t = 1, 2, \ldots, T$:

- Learner chooses hint $\tilde{\ell}_t$
- Learner selects $\tilde{x}_t \in \mathcal{X}$ $\quad \triangleright$ *using the hint*
- Environment reveals ℓ_t
- Learner pays $\ell_t(\tilde{x}_t)$

Optimistic Online Gradient Descent:

$$x_{t+1} = \underset{z \in \mathcal{X}}{\arg\min} \langle \nabla \ell_t(x_t), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2$$

$$\tilde{x}_{t+1} = \underset{z \in \mathcal{X}}{\arg\min} \langle \nabla \tilde{\ell}_t(x_{t+1}), z \rangle + \frac{1}{2\eta} \|z - x_t\|^2$$

Results:
$R_T(u) = \mathcal{O}(D^2/\eta + \eta V_T)$

- $D \stackrel{\text{def}}{=} \operatorname{diam}(\mathcal{X})$
- η: step size
- $V_T \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|\nabla \ell_t(x_t) - \nabla \tilde{\ell}_t(x_t)\|_*^2$
- Good hints: $V_T = o(T)$
  $\rightarrow$ regret improves beyond $\mathcal{O}(\sqrt{T})$

Riemannian optimization

Riemannian optimization

Problem:

Given a function $f : \mathcal{M} \to \mathbb{R}$, solve

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

where $\mathcal{M}$ is a Riemannian manifold.

Assumptions for this talk:

- Hadamard manifolds $\mathcal{H}$
  - Sectional curvature in $[\kappa_{\min}, 0]$
  - Uniquely geodesic (one shortest path)
- First-order methods: Oracles $\{f, \nabla f\}$
- Access to $\{\mathrm{Exp}, \mathrm{Exp}^{-1}\}$

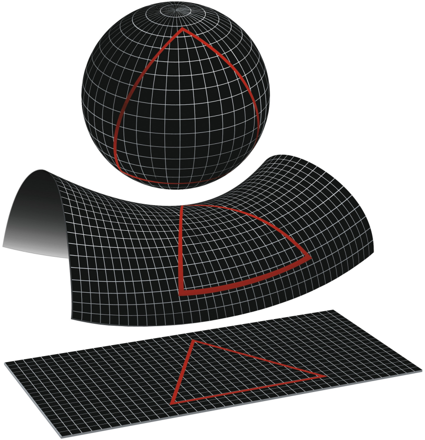


Image credit: NASA/WMAP Science Team

Why Riemannian Optimization?

Use geometric properties (Euclidean constrained $\Rightarrow$ Riemannian unconstrained)

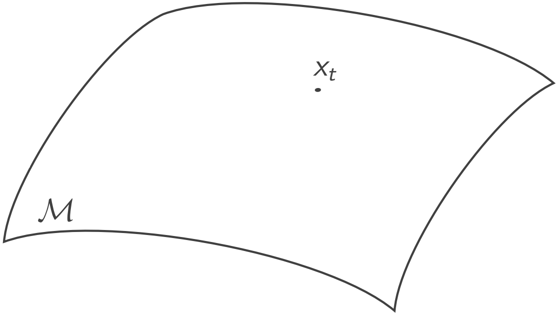
- Fitting Gaussian mixture models: SPD matrices
- DNNs with orthogonality constraints: Stiefel manifold
- Hyperbolic embeddings: Hyperbolic space
- Low-rank matrix factorization: Fixed-rank matrices

Improves problem structure:

- Non-convex *Euclidean* problems can become *geodesically* convex (g-convex) on a manifold
- Example: Operator scaling [Allen-Zhu et al., 2018]

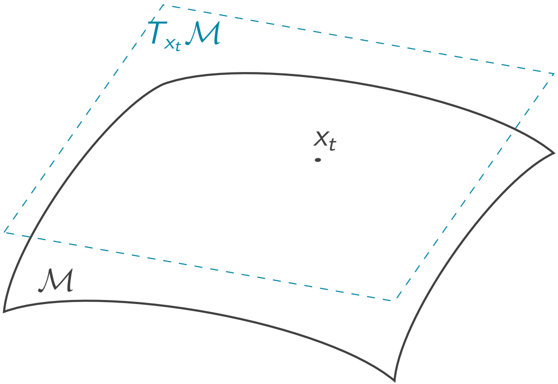
Riemannian Gradient Descent - Key Concepts

- Point x_t on manifold $\mathcal{M}$
  point on curved surface
- Tangent Space: $\mathrm{T}_{x_t}\mathcal{M}$
  local linear approximation
- Riemannian Gradient: $\nabla f(x_t) \in \mathrm{T}_{x_t}\mathcal{M}$
  Gradient projected onto tangent space
- Exponential Map: $x_{t+1} = \mathrm{Exp}_{x_t}(-\eta \nabla f(x_t))$
  move along geodesic
- Log. Map: $\mathrm{Exp}_{x_t}^{-1}(x_{t+1}) = -\eta \nabla f(x_t)$
  vector from x_t to x_{t+1}



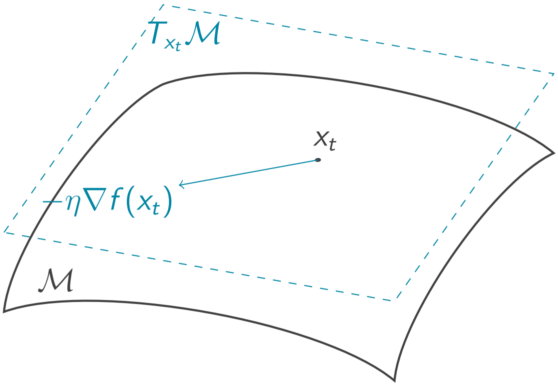
Riemannian Gradient Descent - Key Concepts

- Point x_t on manifold $\mathcal{M}$
  point on curved surface
- Tangent Space: $\mathrm{T}_{x_t}\mathcal{M}$
  local linear approximation
- Riemannian Gradient: $\nabla f(x_t) \in \mathrm{T}_{x_t}\mathcal{M}$
  Gradient projected onto tangent space
- Exponential Map: $x_{t+1} = \mathrm{Exp}_{x_t}(-\eta \nabla f(x_t))$
  move along geodesic
- Log. Map: $\mathrm{Exp}_{x_t}^{-1}(x_{t+1}) = -\eta \nabla f(x_t)$
  vector from x_t to x_{t+1}



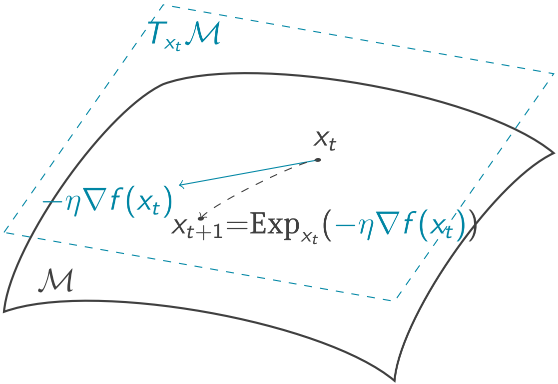
Riemannian Gradient Descent - Key Concepts

- Point x_t on manifold $\mathcal{M}$
  point on curved surface
- Tangent Space: $\mathrm{T}_{x_t}\mathcal{M}$
  local linear approximation
- Riemannian Gradient: $\nabla f(x_t) \in \mathrm{T}_{x_t}\mathcal{M}$
  Gradient projected onto tangent space
- Exponential Map: $x_{t+1} = \mathrm{Exp}_{x_t}(-\eta \nabla f(x_t))$
  move along geodesic
- Log. Map: $\mathrm{Exp}_{x_t}^{-1}(x_{t+1}) = -\eta \nabla f(x_t)$
  vector from x_t to x_{t+1}



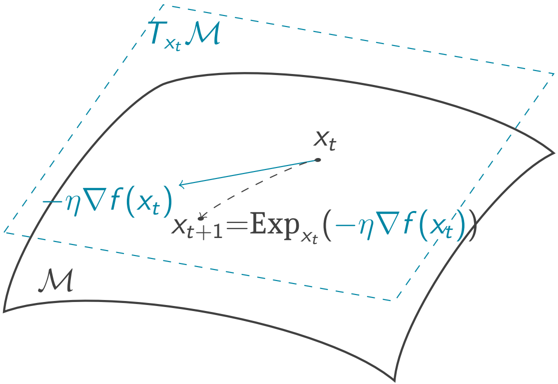
Riemannian Gradient Descent - Key Concepts

- Point x_t on manifold $\mathcal{M}$
  point on curved surface
- Tangent Space: $\mathrm{T}_{x_t}\mathcal{M}$
  local linear approximation
- Riemannian Gradient: $\nabla f(x_t) \in \mathrm{T}_{x_t}\mathcal{M}$
  Gradient projected onto tangent space
- Exponential Map: $x_{t+1} = \mathrm{Exp}_{x_t}(-\eta \nabla f(x_t))$
  move along geodesic
- Log. Map: $\mathrm{Exp}_{x_t}^{-1}(x_{t+1}) = -\eta \nabla f(x_t)$
  vector from x_t to x_{t+1}



Riemannian Gradient Descent - Key Concepts

- Point x_t on manifold $\mathcal{M}$
  point on curved surface
- Tangent Space: $\mathrm{T}_{x_t}\mathcal{M}$
  local linear approximation
- Riemannian Gradient: $\nabla f(x_t) \in \mathrm{T}_{x_t}\mathcal{M}$
  Gradient projected onto tangent space
- Exponential Map: $x_{t+1} = \mathrm{Exp}_{x_t}(-\eta\nabla f(x_t))$
  move along geodesic
- Log. Map: $\mathrm{Exp}_{x_t}^{-1}(x_{t+1}) = -\eta\nabla f(x_t)$
  vector from x_t to x_{t+1}



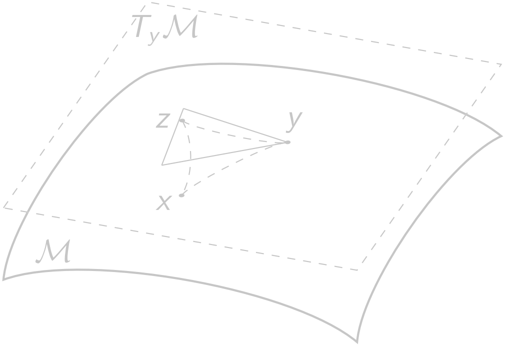
Riemannian Cosine Inequality

Euclidean cosine equality:

$$2\langle x - y, z - y\rangle = \|x - y\|^2 + \|z - y\|^2 - \|x - z\|^2$$

Riemannian Cosine Inequality: $D = \mathrm{Diam}(\triangle xyz)$, $\zeta_D = \Theta(1 + D\sqrt{|\kappa_{\min}|})$

$$2\langle \mathrm{Exp}_y^{-1}(x), \mathrm{Exp}_y^{-1}(z)\rangle_y \leq \zeta_D d(y,x)^2 + d(y,z)^2 - d(x,z)^2$$



Interpretation:

- (Informal) ζ_D measures the deformation caused by the non-linearity of the manifold
- For Hadamard manifolds: $d(x,z) \leq d_y(x,z)$
- The squared Riemannian distance function $\frac{1}{2}d(\cdot, y)^2$ is ζ_D-smooth

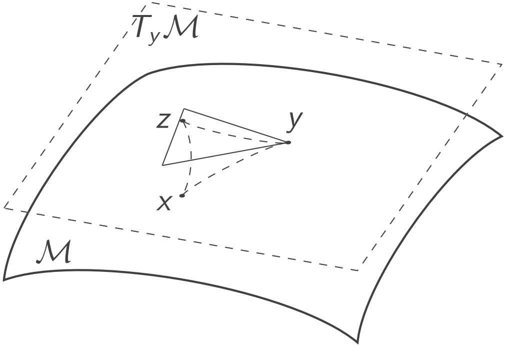
Riemannian Cosine Inequality

Euclidean cosine equality:

$$2\langle x-y, z-y\rangle = \|x-y\|^2 + \|z-y\|^2 - \|x-z\|^2$$

Riemannian Cosine Inequality: $D = \mathrm{Diam}(\triangle xyz)$, $\zeta_D = \Theta(1 + D\sqrt{|\kappa_{\min}|})$

$$2\langle \mathrm{Exp}_y^{-1}(x), \mathrm{Exp}_y^{-1}(z)\rangle_y \leq \zeta_D d(y,x)^2 + d(y,z)^2 - d(x,z)^2$$



Interpretation:

- (Informal) ζ_D measures the deformation caused by the non-linearity of the manifold
- For Hadamard manifolds: $d(x,z) \leq d_y(x,z)$
- The squared Riemannian distance function $\frac{1}{2}d(\cdot,y)^2$ is ζ_D-smooth

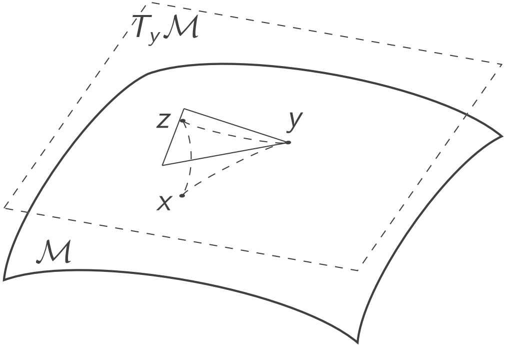
Riemannian Cosine Inequality

Euclidean cosine equality:

$$2\langle x - y, z - y\rangle = \|x - y\|^2 + \|z - y\|^2 - \|x - z\|^2$$

Riemannian Cosine Inequality: $D = \mathrm{Diam}(\triangle xyz)$, $\zeta_D = \Theta(1 + D\sqrt{|\kappa_{\min}|})$

$$2\langle \mathrm{Exp}_y^{-1}(x), \mathrm{Exp}_y^{-1}(z)\rangle_y \leq \zeta_D d(y,x)^2 + d(y,z)^2 - d(x,z)^2$$



Interpretation:

- (Informal) ζ_D measures the deformation caused by the non-linearity of the manifold
- For Hadamard manifolds: $d(x,z) \leq d_y(x,z)$
- The squared Riemannian distance function $\frac{1}{2}d(\cdot, y)^2$ is ζ_D-smooth

Riemannian Optimistic Online Optimization

Prior Work

Euclidean Regret Bound:

$R_T(u) = \mathcal{O}(\frac{D^2}{\eta} + \eta V_T)$

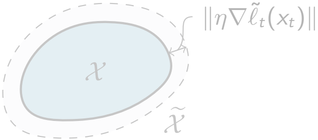
- $D \stackrel{\text{def}}{=} \operatorname{diam}(\mathcal{X})$
- η: step size
- $V_T \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|\nabla \ell_t(x_t) - \nabla \tilde{\ell}_t(x_t)\|^2$

[Wang et al., 2023]:

- General manifolds $\mathcal{M}$, no constraints
- $R_T(u) = \mathcal{O}(\frac{D^2}{\eta} + \eta \zeta_D^2 (V_T + G^2))$
- Recurrent relationship between η and D without assuming boundedness.
- *Assume* the iterates to lie in bounded domain with diameter D

[Hu et al., 2023]:

- Hadamard manifolds $\mathcal{H}$ + constraints $\mathcal{X}$
- Two sequences, only *one* is projected to $\mathcal{X}$
- $\widetilde{\mathcal{X}} \stackrel{\text{def}}{=} \{x \in \mathcal{M} \mid d(x, \mathcal{X}) \leq \|\eta \nabla \tilde{\ell}_t(x_t)\|\}$
  $\widetilde{D} \stackrel{\text{def}}{=} \operatorname{diam}(\widetilde{\mathcal{X}})$



Original set $\mathcal{X}$ vs enlarged set $\widetilde{\mathcal{X}}$

- *Improper* regret: Action in $\widetilde{\mathcal{X}}$, but $u \in \mathcal{X}$
  $\widetilde{R}_T(u) = \mathcal{O}(\frac{\widetilde{D}^2}{\eta} + \eta \zeta_{\widetilde{D}} V_T)$
- Recurrent relationship between η and $\widetilde{D}$

Prior Work

Euclidean Regret Bound:

$R_T(u) = \mathcal{O}(\frac{D^2}{\eta} + \eta V_T)$

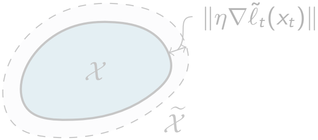
- $D \stackrel{\text{def}}{=} \operatorname{diam}(\mathcal{X})$
- η: step size
- $V_T \stackrel{\text{def}}{=} \sum_{t=1}^{T} \|\nabla \ell_t(x_t) - \nabla \tilde{\ell}_t(x_t)\|^2$

[Wang et al., 2023]:

- General manifolds $\mathcal{M}$, no constraints
- $R_T(u) = \mathcal{O}(\frac{D^2}{\eta} + \eta \zeta_D^2 (V_T + G^2))$
- Recurrent relationship between η and D without assuming boundedness.
- *Assume* the iterates to lie in bounded domain with diameter D

[Hu et al., 2023]:

- Hadamard manifolds $\mathcal{H}$ + constraints $\mathcal{X}$
- Two sequences, only *one* is projected to $\mathcal{X}$
- $\widetilde{\mathcal{X}} \stackrel{\text{def}}{=} \{x \in \mathcal{M} \mid d(x, \mathcal{X}) \leq \|\eta \nabla \tilde{\ell}_t(x_t)\|\}$
  $\widetilde{D} \stackrel{\text{def}}{=} \operatorname{diam}(\widetilde{\mathcal{X}})$



Original set $\mathcal{X}$ vs enlarged set $\widetilde{\mathcal{X}}$

- *Improper* regret: Action in $\widetilde{\mathcal{X}}$, but $u \in \mathcal{X}$
  $\widetilde{R}_T(u) = \mathcal{O}(\frac{\widetilde{D}^2}{\eta} + \eta \zeta_{\widetilde{D}} V_T)$
- Recurrent relationship between η and $\widetilde{D}$

Prior Work

Euclidean Regret Bound:

$R_T(u) = \mathcal{O}(\frac{D^2}{\eta} + \eta V_T)$

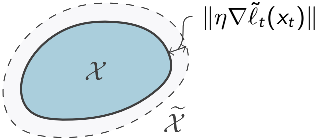
- $D \stackrel{\text{def}}{=} \text{diam}(\mathcal{X})$
- η: step size
- $V_T \stackrel{\text{def}}{=} \sum_{t=1}^T \|\nabla \ell_t(x_t) - \nabla \tilde{\ell}_t(x_t)\|^2$

[Wang et al., 2023]:

- General manifolds $\mathcal{M}$, no constraints
- $R_T(u) = \mathcal{O}(\frac{D^2}{\eta} + \eta \zeta_D^2 (V_T + G^2))$
- Recurrent relationship between η and D without assuming boundedness.
- *Assume* the iterates to lie in bounded domain with diameter D

[Hu et al., 2023]:

- Hadamard manifolds $\mathcal{H}$ + constraints $\mathcal{X}$
- Two sequences, only *one* is projected to $\mathcal{X}$
- $\widetilde{\mathcal{X}} \stackrel{\text{def}}{=} \{x \in \mathcal{M} \mid d(x, \mathcal{X}) \leq \|\eta \nabla \tilde{\ell}_t(x_t)\|\}$
  $\widetilde{D} \stackrel{\text{def}}{=} \text{diam}(\widetilde{\mathcal{X}})$



Original set $\mathcal{X}$ vs enlarged set $\widetilde{\mathcal{X}}$

- *Improper* regret: Action in $\widetilde{\mathcal{X}}$, but $u \in \mathcal{X}$
  $\widetilde{R}_T(u) = \mathcal{O}(\frac{\tilde{D}^2}{\eta} + \eta \zeta_{\tilde{D}} V_T)$
- Recurrent relationship between η and $\widetilde{D}$

What makes this problem hard?

Problem 1: Where to linearize?

Naive idea for optimistic update:

$$x_{t+1} = \arg\min_{z \in \mathcal{X}} \langle \nabla \ell_t(x_t), \mathrm{Exp}_{x_t}^{-1}(z) \rangle_{x_t} + \frac{1}{2\eta} d(z, x_t)^2$$

$$\tilde{x}_{t+1} = \arg\min_{z \in \mathcal{X}} \langle \nabla \tilde{\ell}_t(x_{t+1}), \mathrm{Exp}_{x_{t+1}}^{-1}(z) \rangle_{x_{t+1}} + \frac{1}{2\eta} d(z, x_t)^2$$

- Linearization breaks g-convexity!
- The first update is *star g-convex* in x_t.
- But the second is *not* g-convex.
- Prior works are based on smart use of parallel transport.
- These parallel transports don't play nicely with projections.

Problem 2: Metric projections are hard

- Metric projections $P_{\mathcal{X}}(x) = \arg\min_{z \in \mathcal{X}} d(x, z)$ are hard in Riemannian geometry!
- **Positive curvature:** In general, *not* non-expansive [Wang et al., 2023a].
- **Hadamard manifolds:** First general proof of linear convergence of Projected RGD is very recent [Martínez-Rubio et al., 2023].

Riemannian Implicit Optimistic Online Gradient Descent (RIOD)

Approach: Minimize regularized *full* loss/hint function, not linearization.

Riemannian **I**mplicit **O**ptimistic Online Gradient **D**escent (**RIOD**)

$$x_0 = \tilde{x}_0 \in \mathcal{X}, \quad \forall t \in \mathbb{N}, \quad \begin{cases} \textit{Choose} & \tilde{\ell}_t \\ \textbf{Play} & \tilde{x}_t \leftarrow \operatorname{arg\,min}_{x \in \mathcal{X}} \left\{ \tilde{\ell}_t(x) + \frac{1}{2\eta} d(x, x_t)^2 \right\} \\ \textit{Observe} & \ell_t \\ \textit{Update} & x_{t+1} \leftarrow \operatorname{arg\,min}_{x \in \mathcal{X}} \left\{ \ell_t(x) + \frac{1}{2\eta} d(x, x_t)^2 \right\} \end{cases}$$

Riemannian Implicit Optimistic Online Gradient Descent (RIOD)

Approach: Minimize regularized *full* loss/hint function, not linearization.

Riemannian **I**mplicit **O**ptimistic Online Gradient **D**escent (**RIOD**)

$$x_0 = \tilde{x}_0 \in \mathcal{X}, \quad \forall t \in \mathbb{N}, \quad \begin{cases} \textit{Choose} & \tilde{\ell}_t \\ \textbf{Play} & \tilde{x}_t \leftarrow \operatorname{arg\,min}_{x \in \mathcal{X}} \left\{ \tilde{\ell}_t(x) + \frac{1}{2\eta} d(x, x_t)^2 \right\} \\ \textit{Observe} & \ell_t \\ \textit{Update} & x_{t+1} \leftarrow \operatorname{arg\,min}_{x \in \mathcal{X}} \left\{ \ell_t(x) + \frac{1}{2\eta} d(x, x_t)^2 \right\} \end{cases}$$

RIOD - Main Result

Theorem:

- $\mathcal{M}$: Hadamard manifold (curvature $\kappa_{\text{min}} \leq 0$).
- $\mathcal{X}$: Compact, g-convex subset $\mathcal{X}$, $\text{diam}(\mathcal{X}) = D$.
- $\ell_t, \tilde{\ell}_t$: g-convex, differentiable in $\mathcal{X}$.

Then for any comparator $u \in \mathcal{X}$,

$$R_T(u) = \sum_{t=1}^{T} \ell_t(x_t) - \ell_t(u) \leq \frac{3D^2}{2\eta} + \eta V_T.$$

- Works for **constrained** setting $\mathcal{X}$.
- Matches **Euclidean** regret ($\mathcal{O}(D^2/\eta + \eta V_T)$) $\rightarrow$ No dependence on geometric constant ζ.
- For $\ell_t, \tilde{\ell}_t$ L-smooth in $\mathcal{X}$: Can handle **inexact** updates.
- Regret is not (oracle) complexity!

RIOD - Main Result

Theorem:

- $\mathcal{M}$: Hadamard manifold (curvature $\kappa_{\text{min}} \leq 0$).
- $\mathcal{X}$: Compact, g-convex subset $\mathcal{X}$, $\text{diam}(\mathcal{X}) = D$.
- $\ell_t, \tilde{\ell}_t$: g-convex, differentiable in $\mathcal{X}$.

Then for any comparator $u \in \mathcal{X}$,

$$R_T(u) = \sum_{t=1}^{T} \ell_t(x_t) - \ell_t(u) \leq \frac{3D^2}{2\eta} + \eta V_T.$$

- Works for **constrained** setting $\mathcal{X}$.
- Matches **Euclidean** regret $(\mathcal{O}(D^2/\eta + \eta V_T)) \rightarrow$ No dependence on geometric constant ζ.
- For $\ell_t, \tilde{\ell}_t$ L-smooth in $\mathcal{X}$: Can handle **inexact** updates.
- Regret is not (oracle) complexity!

RIOD - Main Result

Theorem:

- $\mathcal{M}$: Hadamard manifold (curvature $\kappa_{\min} \leq 0$).
- $\mathcal{X}$: Compact, g-convex subset $\mathcal{X}$, $\operatorname{diam}(\mathcal{X}) = D$.
- $\ell_t, \tilde{\ell}_t$: g-convex, differentiable in $\mathcal{X}$.

Then for any comparator $u \in \mathcal{X}$,

$$R_T(u) = \sum_{t=1}^{T} \ell_t(x_t) - \ell_t(u) \leq \frac{3D^2}{2\eta} + \eta V_T.$$

- Works for **constrained** setting $\mathcal{X}$.
- Matches **Euclidean** regret ($\mathcal{O}(D^2/\eta + \eta V_T)$) $\rightarrow$ No dependence on geometric constant ζ.
- For $\ell_t, \tilde{\ell}_t$ L-smooth in $\mathcal{X}$: Can handle **inexact** updates.
- Regret is not (oracle) complexity!

RIOD Implementation - Smooth losses

Subproblem: $$\underbrace{\ell_t(x)}_{\substack{L\text{-smooth}\\ \text{g-convex}}} + \underbrace{\frac{1}{2\eta}d(x,x_t)^2}_{\substack{(\zeta/\eta)\text{-smooth}\\ (1/\eta)\text{-strongly g-convex}}}$$

Projected RGD:
[Martínez-Rubio et al., 2023]

$$x_{t+1} \leftarrow P_{\mathcal{X}}(\mathrm{Exp}_{x_t}(-\frac{1}{L}\nabla f(x_t)))$$

- RIOD step: $\tilde{\mathcal{O}}(\zeta(L\eta+\zeta))$
- With $\eta = \frac{1}{L}$, we get $\tilde{\mathcal{O}}(\zeta^2)$.

Composite RGD: (functions $f+g$)
[Martínez-Rubio et al., 2024]

$$x_{t+1} \leftarrow \arg\min_{y\in\mathcal{X}} \langle \nabla f(x_t), \mathrm{Exp}_{x_t}^{-1}(y)\rangle + g(y) + \frac{1}{2\gamma}d(y,x_t)^2$$

- RIOD step: $\tilde{\mathcal{O}}(L\eta)$, with $\eta = \frac{1}{L}$ we get $\tilde{\mathcal{O}}(1)$.
- Not necessarily g-convex.
- But: One oracle call per CRGD step.

RIOD Implementation - Smooth losses

Subproblem: $\underbrace{\ell_t(x)}_{\substack{L\text{-smooth} \\ \text{g-convex}}} + \underbrace{\frac{1}{2\eta} d(x, x_t)^2}_{\substack{(\zeta/\eta)\text{-smooth} \\ (1/\eta)\text{-strongly g-convex}}}$

Projected RGD:
[Martínez-Rubio et al., 2023]

$$x_{t+1} \leftarrow P_{\mathcal{X}}(\mathrm{Exp}_{x_t}(-\frac{1}{L}\nabla f(x_t)))$$

- RIOD step: $\tilde{\mathcal{O}}(\zeta(L\eta + \zeta))$
- With $\eta = \frac{1}{L}$, we get $\tilde{\mathcal{O}}(\zeta^2)$.

Composite RGD: (functions $f + g$)
[Martínez-Rubio et al., 2024]

$$x_{t+1} \leftarrow \arg\min_{y \in \mathcal{X}} \langle \nabla f(x_t), \mathrm{Exp}_{x_t}^{-1}(y) \rangle + g(y) + \frac{1}{2\gamma} d(y, x_t)^2$$

- RIOD step: $\tilde{\mathcal{O}}(L\eta)$, with $\eta = \frac{1}{L}$ we get $\tilde{\mathcal{O}}(1)$.
- Not necessarily g-convex.
- But: One oracle call per CRGD step.

RIOD Implementation - Smooth losses

Subproblem: $\underbrace{\ell_t(x)}_{\substack{L\text{-smooth} \\ \text{g-convex}}} + \underbrace{\frac{1}{2\eta} d(x, x_t)^2}_{\substack{(\zeta/\eta)\text{-smooth} \\ (1/\eta)\text{-strongly g-convex}}}$

Projected RGD:
[Martínez-Rubio et al., 2023]

$$x_{t+1} \leftarrow P_{\mathcal{X}}(\mathrm{Exp}_{x_t}(-\frac{1}{L}\nabla f(x_t)))$$

- RIOD step: $\tilde{\mathcal{O}}(\zeta(L\eta + \zeta))$
- With $\eta = \frac{1}{L}$, we get $\tilde{\mathcal{O}}(\zeta^2)$.

Composite RGD: (functions $f + g$)
[Martínez-Rubio et al., 2024]

$$x_{t+1} \leftarrow \arg\min_{y \in \mathcal{X}} \langle \nabla f(x_t), \mathrm{Exp}_{x_t}^{-1}(y) \rangle + g(y) + \frac{1}{2\gamma} d(y, x_t)^2$$

- RIOD step: $\tilde{\mathcal{O}}(L\eta)$, with $\eta = \frac{1}{L}$ we get $\tilde{\mathcal{O}}(1)$.
- Not necessarily g-convex.
- But: One oracle call per CRGD step.

RIOD Implementation - Smooth losses

Subproblem: $\underbrace{\ell_t(x)}_{\substack{L\text{-smooth}\\ \text{g-convex}}} + \underbrace{\frac{1}{2\eta} d(x, x_t)^2}_{\substack{(\zeta/\eta)\text{-smooth}\\ (1/\eta)\text{-strongly g-convex}}}$

Projected RGD:
[Martínez-Rubio et al., 2023]

$$x_{t+1} \leftarrow P_{\mathcal{X}}(\mathrm{Exp}_{x_t}(-\frac{1}{L}\nabla f(x_t)))$$

- RIOD step: $\tilde{\mathcal{O}}(\zeta(L\eta + \zeta))$
- With $\eta = \frac{1}{L}$, we get $\tilde{\mathcal{O}}(\zeta^2)$.

Composite RGD: (functions $f + g$)
[Martínez-Rubio et al., 2024]

$$x_{t+1} \leftarrow \arg\min_{y \in \mathcal{X}} \langle \nabla f(x_t), \mathrm{Exp}_{x_t}^{-1}(y) \rangle + g(y) + \frac{1}{2\gamma} d(y, x_t)^2$$

- RIOD step: $\tilde{\mathcal{O}}(L\eta)$, with $\eta = \frac{1}{L}$ we get $\tilde{\mathcal{O}}(1)$.
- Not necessarily g-convex.
- But: One oracle call per CRGD step.

RIOD Implementation - Smooth losses

$$\text{Subproblem:} \quad \underbrace{\ell_t(x)}_{\substack{L\text{-smooth} \\ \text{g-convex}}} \quad + \quad \underbrace{\frac{1}{2\eta} d(x, x_t)^2}_{\substack{(\zeta/\eta)\text{-smooth} \\ (1/\eta)\text{-strongly g-convex}}}$$

Projected RGD:
[Martínez-Rubio et al., 2023]

$$x_{t+1} \leftarrow P_{\mathcal{X}}(\mathrm{Exp}_{x_t}(-\frac{1}{L}\nabla f(x_t)))$$

- RIOD step: $\tilde{\mathcal{O}}(\zeta(L\eta + \zeta))$
- With $\eta = \frac{1}{L}$, we get $\tilde{\mathcal{O}}(\zeta^2)$.

Composite RGD: (functions $f + g$)
[Martínez-Rubio et al., 2024]

$$x_{t+1} \leftarrow \arg\min_{y \in \mathcal{X}} \langle \nabla f(x_t), \mathrm{Exp}_{x_t}^{-1}(y) \rangle + g(y) + \frac{1}{2\gamma} d(y, x_t)^2$$

- RIOD step: $\tilde{\mathcal{O}}(L\eta)$, with $\eta = \frac{1}{L}$ we get $\tilde{\mathcal{O}}(1)$.
- Not necessarily g-convex.
- But: One oracle call per CRGD step.

Riemannian Min-Max Optimization

Riemannian Min-Max Optimization

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

- $\mathcal{X} \subseteq \mathcal{M}$, $\mathcal{Y} \subseteq \mathcal{N}$ g-convex (compact) subsets.
- f is g-convex in x, g-concave in y.
- f is L-smooth in x and y.

Applications:

- Distributionally robust linear quadratic control [Taskesen et al., 2023].
- Robust version of finite-sum, g-convex optimization problem [Zhang et al., 2023, Jordan et al., 2022].

RIODA: Main Results

Theorem: RIODA converges in $\tilde{\mathcal{O}}(LR^2/\epsilon)$ iterations in the L-smooth, g-convex/g-concave case.

Total Gradient Complexity: Iterations $\times$ Subproblem cost

RIODA-PRGD: *(Constrained)*

- Improvement: $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon}\zeta^{5.5})$ to $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon}\zeta^{2})$.
- No knowledge of initial distance R, Lipschitz constant G required.
- Simpler algorithm (two loops vs five loops).

RIODA-CRGD:

- Full gradient complexity: $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon})$.
- Matches Euclidean lower bounds (up to logs) - No ζ poly factors.

RIODA: Main Results

Theorem: RIODA converges in $\tilde{\mathcal{O}}(LR^2/\epsilon)$ iterations in the L-smooth, g-convex/g-concave case.

Total Gradient Complexity: Iterations $\times$ Subproblem cost

RIODA-PRGD: *(Constrained)*

- Improvement: $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon}\zeta^{5.5})$ to $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon}\zeta^{2})$.
- No knowledge of initial distance R, Lipschitz constant G required.
- Simpler algorithm (two loops vs five loops).

RIODA-CRGD:

- Full gradient complexity: $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon})$.
- Matches Euclidean lower bounds (up to logs) - No ζ poly factors.

RIODA: Main Results

Theorem: RIODA converges in $\tilde{\mathcal{O}}(LR^2/\epsilon)$ iterations in the L-smooth, g-convex/g-concave case.

Total Gradient Complexity: Iterations $\times$ Subproblem cost

RIODA-PRGD: *(Constrained)*

- Improvement: $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon}\zeta^{5.5})$ to $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon}\zeta^{2})$.
- No knowledge of initial distance R, Lipschitz constant G required.
- Simpler algorithm (two loops vs five loops).

RIODA-CRGD:

- Full gradient complexity: $\tilde{\mathcal{O}}(\frac{LR^2}{\epsilon})$.
- Matches Euclidean lower bounds (up to logs) - No ζ poly factors.

Conclusion

RIOD: An *inexact, implicit, optimistic* algorithm for online constrained Riemannian optimization (Hadamard manifolds).

- Addresses key limitations of prior Riemannian online optimistic methods (in-manifold constraints, improper regret).
- Matches best Euclidean regret bounds.

RIODA: An inexact, implicit algorithm for smooth, g-convex/g-concave min-max problems (Hadamard manifolds) built on RIOD.

- RIODA-CRGD: Matching upper and lower bounds (up to logs)
- RIODA-PRGD: Improve rates, relax assumptions over SOTA.

Conclusion

RIOD: An *inexact, implicit, optimistic* algorithm for online constrained Riemannian optimization (Hadamard manifolds).

- Addresses key limitations of prior Riemannian online optimistic methods (in-manifold constraints, improper regret).
- Matches best Euclidean regret bounds.

RIODA: An inexact, implicit algorithm for smooth, g-convex/g-concave min-max problems (Hadamard manifolds) built on RIOD.

- RIODA-CRGD: Matching upper and lower bounds (up to logs)
- RIODA-PRGD: Improve rates, relax assumptions over SOTA.

Thank you! Questions?

arXiv:2403.10429

Homepage

References

Allen-Zhu, Z., Garg, A., Li, Y., de Oliveira, R. M., and Wigderson, A. (2018).
Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing.
In Diakonikolas, I., Kempe, D., and Henzinger, M., editors, *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 172–181. ACM.

Criscitiello, C. and Boumal, N. (2023).
Curvature and complexity: Better lower bounds for geodesically convex optimization.
In Neu, G. and Rosasco, L., editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 2969–3013. PMLR.

Hu, Z., Wang, G., and Abernethy, J. D. (2023).
Minimizing Dynamic Regret on Geodesic Metric Spaces.
In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 4336–4383. PMLR.

References

Jordan, M., Lin, T., and Vlatakis-Gkaragkounis, E.-V. (2022).
First-Order Algorithms for Min-Max Optimization in Geodesic Metric Spaces.
Advances in Neural Information Processing Systems, 35:6557–6574.

Martínez-Rubio, D., Roux, C., Criscitiello, C., and Pokutta, S. (2023).
Accelerated methods for riemannian min-max optimization ensuring bounded geometric penalties.
CoRR, abs/2305.16186.

Martínez-Rubio, D., Roux, C., and Pokutta, S. (2024).
Convergence and trade-offs in riemannian gradient descent and riemannian proximal point.
CoRR, abs/2403.10429.

References

Taskesen, B., Iancu, D. A., Koçyigit, Ç., and Kuhn, D. (2023).
Distributionally robust linear quadratic control.
In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wang, X., Yuan, D., Hong, Y., Hu, Z., Wang, L., and Shi, G. (2023).
Riemannian optimistic algorithms.
arXiv preprint arXiv:2308.16004v1.

Zhang, P., Zhang, J., and Sra, S. (2023).
Sion's minimax theorem in geodesic metric spaces and a riemannian extragradient algorithm.
SIAM J. Optim., 33(4):2885–2908.

Connection to Riemannian Lower Bounds

Known lower bound: L-smooth g-convex *minimization* $\Omega(\zeta)$ [Criscitiello and Boumal, 2023].

Our rate: $T = \tilde{\mathcal{O}}(LR^2/\epsilon)$.

Does this contradict lower bounds?

- Lower bound is in Hyperbolic space where: $f(\bar{x}) - f(x^*) \lesssim Ld(\bar{x}, x^*)^2/\zeta$.
- In other words: $\epsilon = \mathcal{O}(\frac{LR^2}{\zeta})$.
- Our rate $T = \tilde{\mathcal{O}}(LR^2/\epsilon)$ is larger than $\Omega(\zeta)$.

Takeaway Messages:

- Implicit hardness from geometry.
- Upper and lower bounds match up to log factors for Riemannian min-max problems.
- Not the case for g-convex minimization: $\tilde{\mathcal{O}}(\zeta + \sqrt{\frac{\zeta LR^2}{\epsilon}})$ vs $\Omega(\zeta)$.