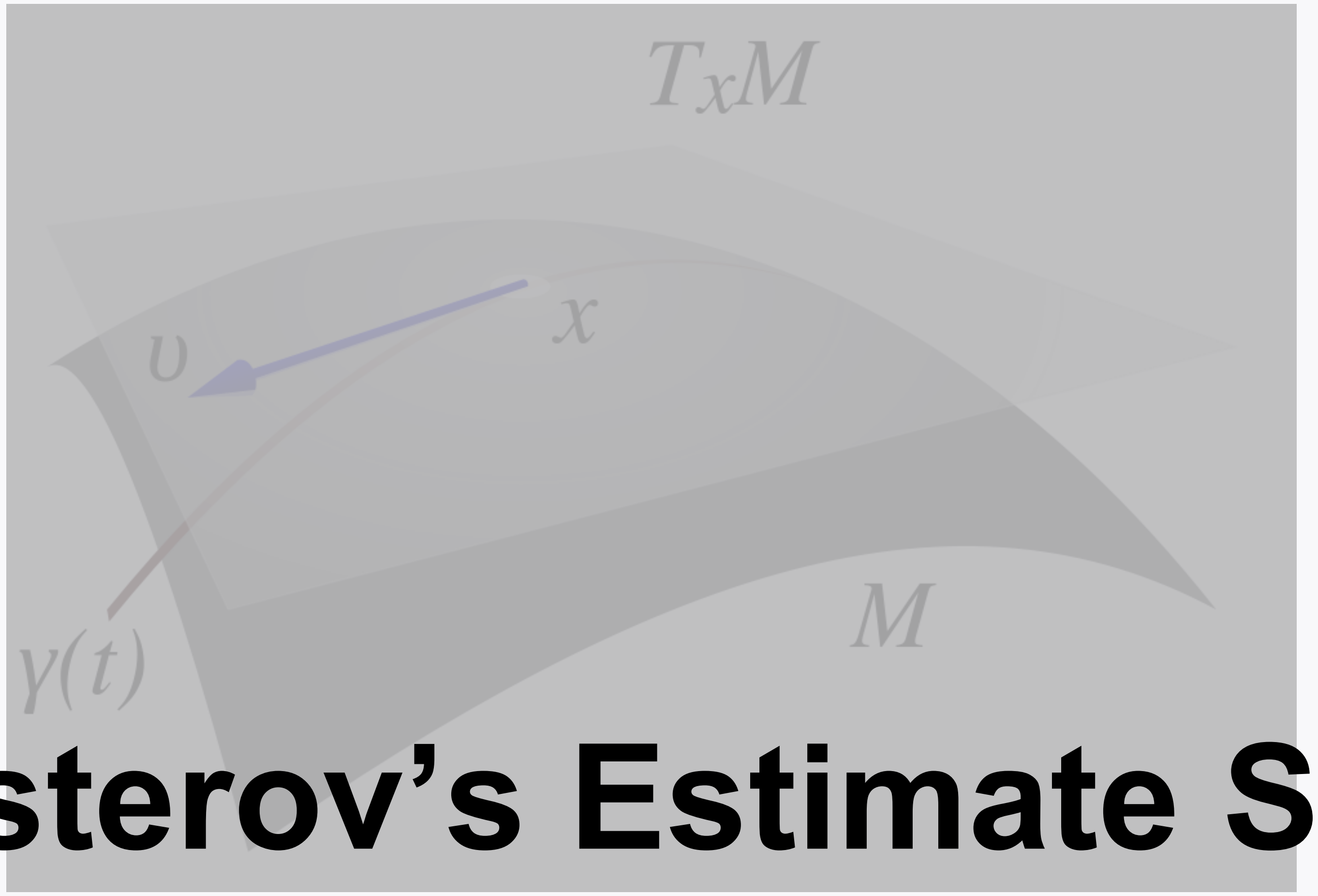


From Nesterov's Estimate Sequence To Riemannian Acceleration

Kwangjun Ahn, Suvrit Sra

COLT 2020

arXiv: https://arxiv.org/abs/2001.08876

Riemannian Optimization?

- (Euclidean) Optimization: $f\colon \mathbb{R}^n \to \mathbb{R}$

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Riemannian Optimization: $f\colon M \to \mathbb{R}$

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

M= a Riemannian manifold

Accel. Gradient Method!

- Yurii Nesterov 80's

Accel. Gradient Descent:

For $t = 0,1,2,\ldots$

$$x_{t+1} = y_t + \alpha_{t+1}(z_t - y_t)$$
$$y_{t+1} = x_{t+1} - \gamma_{t+1}\nabla f(x_{t+1})$$
$$z_{t+1} = x_{t+1} + \beta_{t+1}(z_t - x_{t+1}) - \eta_{t+1}\nabla f(x_{t+1})$$

Accel. Gradient Method: Theory

- Yurii Nesterov 80's

C.f. Gradient Descent:

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

$$f(x_t) - f(x_*) \leq O\left(\left(1 - \frac{\boldsymbol{\mu}}{\boldsymbol{L}}\right)^t\right)$$

For ϵ-approx. solution,

We need $O\left(\frac{\boldsymbol{L}}{\boldsymbol{\mu}} \log\left(\frac{1}{\epsilon}\right)\right)$ many iterations.

Nesterov showed:

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

$$f(y_t) - f(x_*) \leq O\left(\left(1 - \sqrt{\frac{\boldsymbol{\mu}}{\boldsymbol{L}}}\right)^t\right)$$

For ϵ-approx. solution,

We only need $t \geq O\left(\sqrt{\frac{\boldsymbol{L}}{\boldsymbol{\mu}}} \log\left(\frac{1}{\epsilon}\right)\right)$.

➔ Acceleration!

➔(and indeed optimal for this class!)

Natural Question..

- Could we develop such **landmark** result for **curved** spaces (Riem. manifolds)?
- Turns out to be challenging question:
  - Liu et al.'17 (*NIPS*) reduces the task to solving nonlinear equations.
    - Not clear whether whether these equations are even feasible or tractably solvable.
  - Alimisis et al.'20 (*AISTATS*): Continuous dynamic approach
    - Not clear whether the discretization yields accel.
  - Most concrete result: Zhang-Sra'18 (*COLT*)
    - proposed an alg. guaranteed to accel. **locally**.

Global accel? ➔ **Open!**

Challenge!

- Nesterov’s analysis is called the ***Estimate Sequence*** *technique*

- Nesterov’s analysis relies on **linear** structure!
  - not clear if it generalizes to **non-linear space** like Riem. manifolds.

- Nesterov’s analysis entails non-trivial algebraic tricks!
  - Hard to understand; its scope has puzzled researchers for years.

Riemannian Accel. GD

(Euclidean) Accel. Gradient Descent:

$$x_{t+1} = y_t + \alpha_{t+1}(z_t - y_t)$$
$$y_{t+1} = x_{t+1} - \gamma_{t+1}\nabla f(x_{t+1})$$
$$z_{t+1} = x_{t+1} + \beta_{t+1}(z_t - x_{t+1}) - \eta_{t+1}\nabla f(x_{t+1})$$

Riemannian Accel. Gradient Descent:

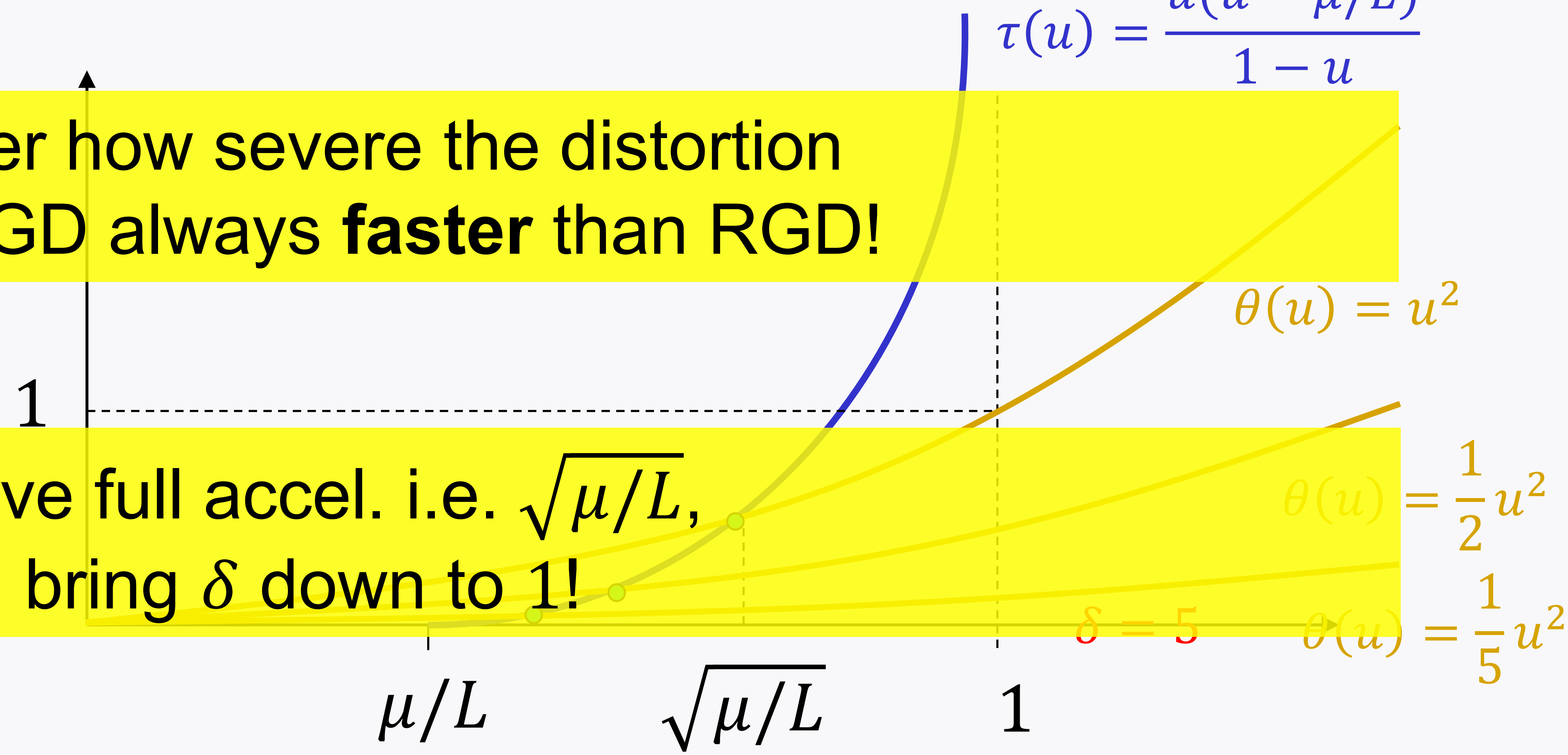
$$x_{t+1} = Exp_{y_t}\left(\alpha_{t+1} \cdot Exp_{y_t}^{-1}(z_t)\right)$$
$$y_{t+1} = Exp_{x_{t+1}}\left(-\gamma_{t+1} \cdot \nabla f(x_{t+1})\right)$$
$$z_{t+1} = Exp_{x_{t+1}}\left(\beta_{t+1} \cdot Exp_{x_{t+1}}^{-1}(z_t) - \eta_{t+1}\nabla f(x_{t+1})\right)$$

Space is curved, causes "distortion"

1. How does this affect the convergence rate?

- Non-linear recursive relation: $\frac{\xi_{t+1}(\xi_{t+1}-\mu/L)}{(1-\xi_{t+1})} = \frac{1}{\delta}\xi_t^2$

Severer the distortion gets,
Slower the convergence rate becomes!



No matter how severe the distortion
Riem. AGD always **faster** than RGD!

To achieve full accel. i.e. $\sqrt{\mu/L}$,
we need bring δ down to 1!

How do we control/estimate the distortion?

Global Accel for Riem. Case!

Thm 2. Given: $\xi_0 > 0$

Find $\xi_{t+1} \in (2\mu\Delta, 1)$ such that

the magnitude of metric distortion at iteration t

$$\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta)}{(1 - \xi_{t+1})} = \frac{1}{\boldsymbol{\delta}_{t+1}} \xi_t^2$$

where $\boldsymbol{\delta}_{t+1} = \boldsymbol{T}(\boldsymbol{d}(\boldsymbol{x}_t, \boldsymbol{z}_t))$ for some computable function T.

$$f(y_{t+1}) - f(x_*) \le O\big((1 - \xi_1)(1 - \xi_2) \cdots (1 - \xi_{t+1})\big)$$

s.t. (1) $\xi_t > \mu/L$ for all t. (2) ξ_t quickly converges to $\sqrt{\mu/L}$.

strictly **faster** than (nonaccel) GD! quickly acheives **full** acceleartion!

Open problem

Obtaining acceleration the non-strongly convex case?

Remarks

- ★ Using strongly convex perturbation can be done
- ★ But, extra $O(\log 1/\epsilon)$ factor
- ★ More crucially, our current proof needs to ensure all iterates remain within a set of specific size to be able to ensure acceleration. Removing this limitation is valuable