Directional Smoothness and Gradient Methods: Convergence and Adaptivity

Aaron Mishkin* [1] Ahmed Khaled* [2] Yuanhao Wang [2] Aaron Defazio [3] Robert M. Gower [4]

[1]Stanford University [2]Princeton University [3]FAIR, META [4]CCM, Flatiron Institute

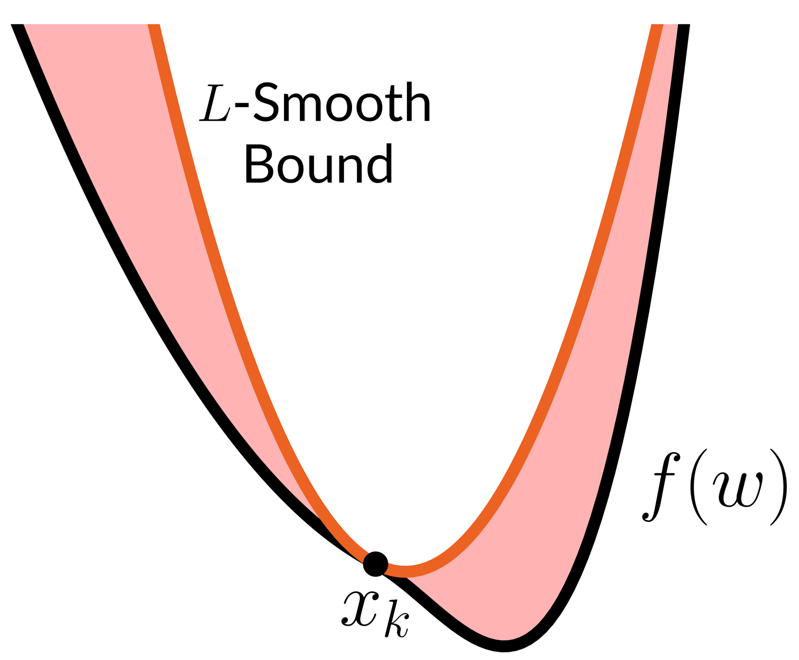
Introduction

Goal: Minimize convex, differentiable function f using GD,

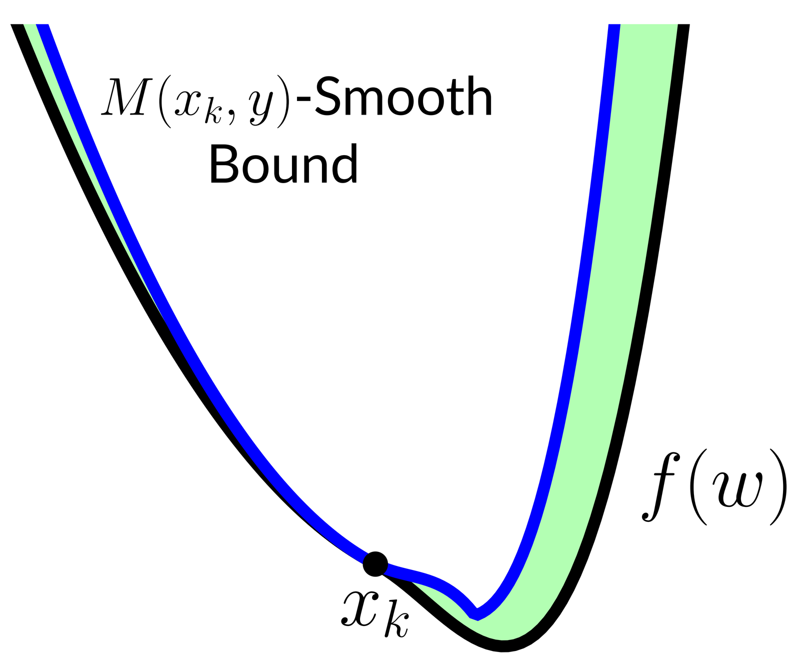
$$x_{k+1} = x_k - \eta_k \nabla f(x_k).$$

Problem: Gradient descent (GD) is an inherently local algorithm, but standard analyses rely on global, worst-case assumptions.

Global Smoothness



Directional Smoothness



Main Contributions:

- Directional Smoothness: a new, point-wise relaxation of L-smoothness.
- Path-Dependent Rates: guarantees for GD using only local properties of f.
- Adaptive Methods: optimizers that adapt to the directional smoothness.

Directional Smoothness

Global Smoothness: f is L-smooth if for every $x, y \in$ dom(f),

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2,$$

Directional Smoothness: M is a directional smoothness function if,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M(x, y)}{2} \|y - x\|_2^2.$$

We give explicit smoothness functions — no oracles required!

Point-wise Smoothness:

$$D(x, y) = \frac{2 \|\nabla f(x) - \nabla f(y)\|_2}{\|x - y\|_2} \qquad (\leq 2L)$$

Path-wise Smoothness:

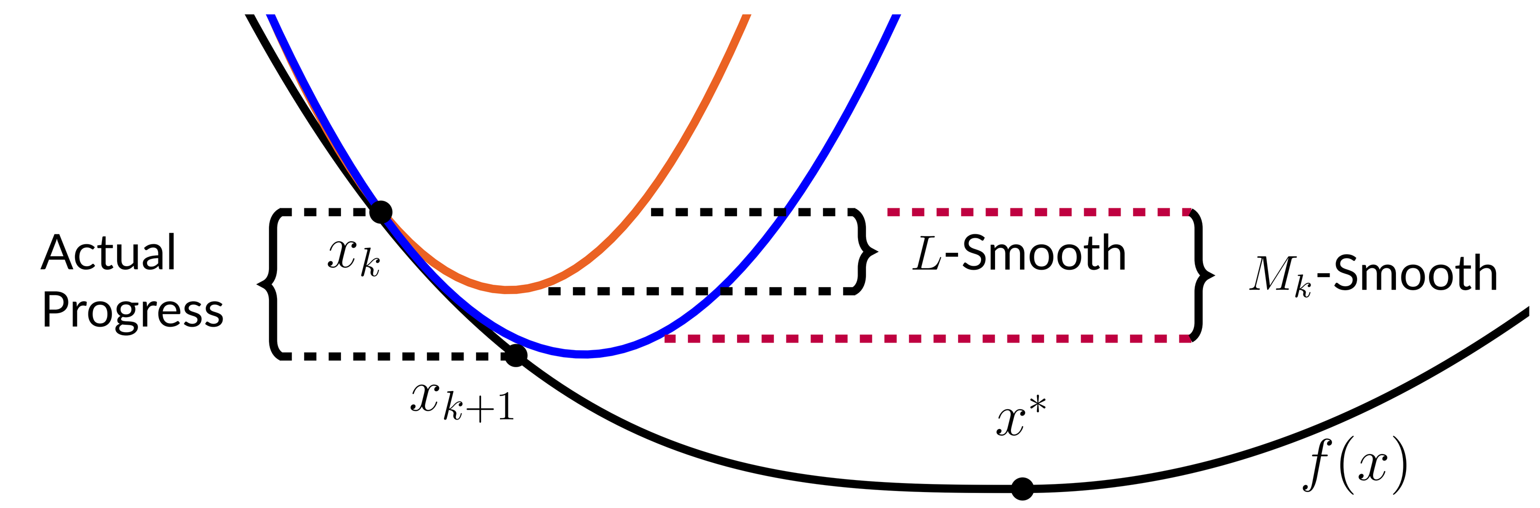
$$A(x, y) = \sup_{t \in [0,1]} \frac{\langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle}{t \|x - y\|_2^2} \qquad (\leq L)$$

Exact (Point-wise) Smoothness:

$$H(x, y) = \frac{2 |f(y) - f(x) - \langle \nabla f(x), y - x \rangle|}{\|y - x\|_2^2} \qquad (\leq L)$$

Easy to compute in hindsight unlike other approaches (Park et al., 2021; Mei et al., 2021).

Path-Dependent Rates



- Directional smoothness $\implies$ more progress than L-smoothness!

Approach: study local behavior of GD along $\{x_k\}$ using $M(x_k, x_{k+1})$.

Proposition (Strongly Convex): Let $\Delta_i = \|x_i - x_0\|_2^2$ and $M_i = M(x_i, x_{i+1})$. If f is μ-strongly convex, then GD with step-size sequence $\{\eta_k\}$ satisfies,

$$\Delta_k \leq \left[\prod_{i=0}^{k} \frac{|1 - \mu\eta_i|}{1 + \mu\eta_i}\right] \Delta_0 + \sum_{i=0}^{k} \left[\prod_{j>i} \frac{|1 - \mu\eta_j|}{1 + \mu\eta_j}\right] \eta_i^2 (M_i \eta_i - 1) \|\nabla f(x_k)\|_2^2.$$

- Fast rates when η_k are adapted, meaning $\eta_k \leq 1/M(x_k, x_{k+1})$.
- Describes worst-case "blow-up" when η_k are not adapted.

Proposition (Convex): Let $\Delta_i = \|x_i - x_0\|_2^2$ and $M_i = M(x_i, x_{i+1})$. If f is convex, then GD with step-size sequence $\{\eta_k\}$ satisfies,

$$\min_{i \in [k]} f(x_i) - f(x^*) \leq \frac{\Delta_0 + \sum_{i=0}^{k} \eta_i^2 (\eta_i M_i - 1) \|\nabla f(x_i)\|_2^2}{2 \sum_{i=0}^{k} \eta_i},$$

Definition: η_k is strongly adapted to smoothness function M if,

$$\eta_k = \frac{1}{M(x_k, x_k - \eta_k \nabla f(x_k))}.$$

Strongly adapted step-sizes get path-dependent rates.

Global Smoothness

$$\min_{i \in [k]} f(x_i) - f(x^*) \leq \frac{L \Delta_0}{k + 1}$$

Directional Smoothness

$$\min_{i \in [k]} f(x_i) - f(x^*) \leq \left[\frac{\sum_{i=0}^{k} M_i}{k + 1}\right] \frac{\Delta_0}{k + 1}$$

The Quadratic Case

Problem: strongly adapted η_k require solving an implicit equation.

Lemma: If $f(x) = \frac{1}{2} x^\top B x - c^\top x$, then the point-wise smoothness is given by,

$$D(x_k, x_{k+1}(\eta_k)) = \frac{\|B \nabla f(x_k)\|_2}{\|\nabla f(x_k)\|_2}.$$

- This recovers a classic step-size for quadratic optimization proposed by Dai & Yang (2006)!

Adaptive Methods

Question: Can we obtain path-dependent rates for convex functions without computing strongly adapted step-sizes?

First Attempt: Modify exponential search (Carmon & Hinder, 2022).

Theorem (informal): If f is convex and L-smooth, then exponential search requires at most $2K \log \log(2\eta_0 / L)$ iterations of GD to find η^* yielding the path-dependent convergence rate:

$$f(\bar{x}_K) - f(x^*) \leq \frac{\|x_0 - x_*\|^2}{2K} \left[\frac{\sum_{i=0}^{K} M(x'_{i+1}, x'_i) \|\nabla f(x'_i)\|^2}{\sum_{i=0}^{K} \|\nabla f(x'_i)\|^2}\right],$$

Problem: Only adapts to smoothness along virtual sequence $\{x'_k\}$.

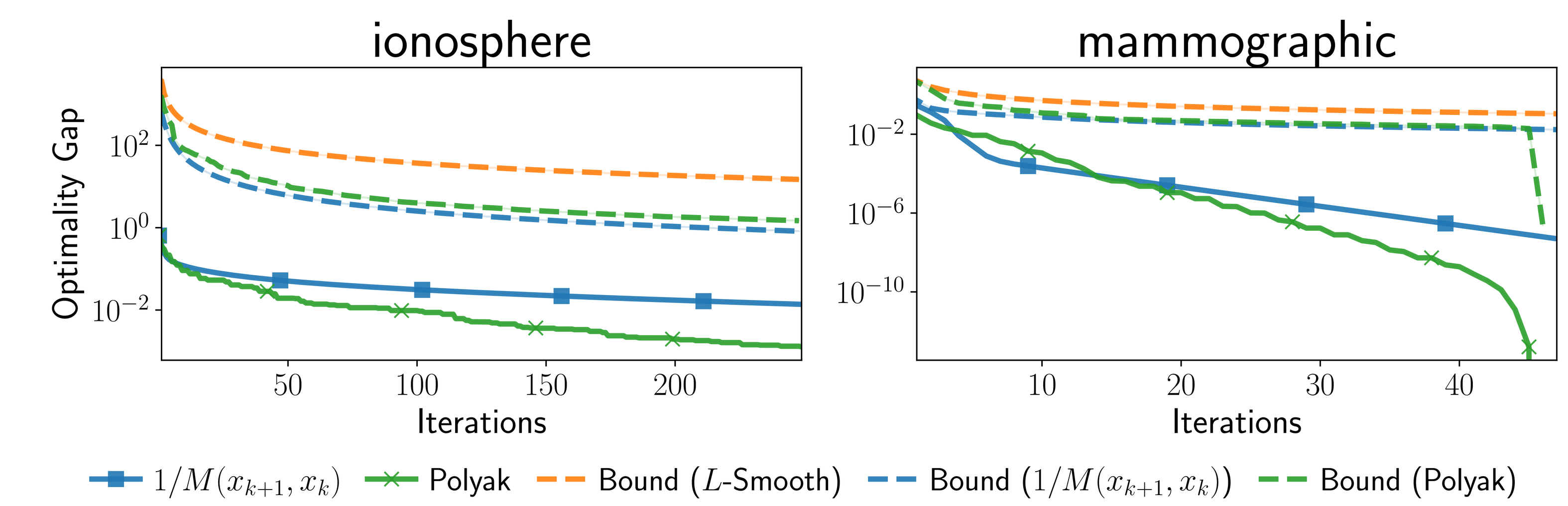
Second Attempt: Polyak step-size: $\eta_k = \gamma(f(x_k) - f(x^*)) / \|\nabla f(x_k)\|_2^2$

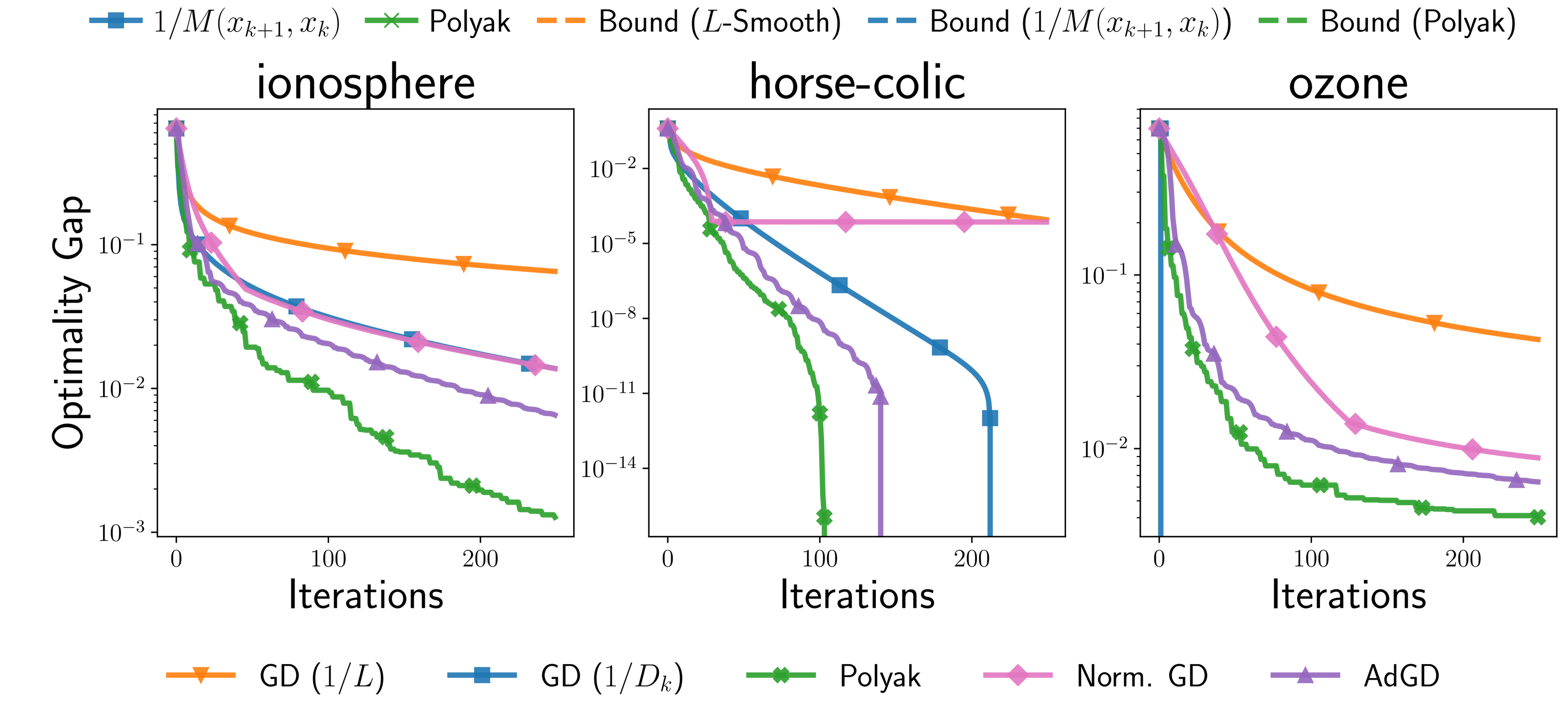
Theorem (Polyak Step-Size): If f is convex and differentiable, then GD with the Polyak step-size using $\gamma = 1.5$ satisfies,

$$\min_{i \in [k]} f(x_i) - f(x^*) \leq 3 \left[\frac{\sum_{i=0}^{k} M_i}{k + 1}\right] \frac{\Delta_0}{k + 1}$$

- Matches rate for strongly adapted step-sizes up to a constant!
- Polyak step-size is "adaptive" to any choice of smoothness M.

Experiments





References

Carmon, Y. and Hinder, O. Making SGD parameter-free. In Loh, P. and Raginsky, M. (eds.), *COLT 2022*. PMLR, 2022.

Dai, Y. and Yang, X. A new gradient method with an optimal stepsize property. *Computational Optimization and Applications*, 33(1), 2006.

Mei, J., Gao, Y., Dai, B., Szepesvári, C., and Schuurmans, D. Leveraging non-uniformity in first-order non-convex optimization. In Meila, M. and Zhang, T. (eds.), *ICML 2021*. PMLR, 2021.

Park, J.-H., Salgado, A. J., and Wise, S. M. Preconditioned accelerated gradient descent methods for locally lipschitz smooth objectives with applications to the solution of nonlinear PDEs. *Journal of Scientific Computing*, 89(1):17, 2021.

* Equal Contribution

Correspondence: amishkin@cs.stanford.edu