Directional Smoothness and Gradient Methods

Convergence and Adaptivity

Aaron Mishkin* Ahmed Khaled* Yuanhao Wang
Aaron Defazio Robert M. Gower





Overview

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

Overview

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

Contributions:

- We propose directional smoothness, a point-wise relaxation of Lipschitz continuous gradients (a.k.a. L-smoothness).

Overview

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

Contributions:

- We propose directional smoothness, a point-wise relaxation of Lipschitz continuous gradients (a.k.a. L-smoothness).
- We use directional smoothness to derive path-dependent sub-optimality bounds for GD.

Overview

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

Contributions:

- We propose directional smoothness, a point-wise relaxation of Lipschitz continuous gradients (a.k.a. L-smoothness).
- We use directional smoothness to derive path-dependent sub-optimality bounds for GD.
- We prove that the Polyak step-size and Normalized GD match the fast rates of GD with strongly adapted step-sizes.

Background on L-Smoothness

Setting: minimize a convex, differentiable function f using GD:

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

Background on L-Smoothness

Setting: minimize a convex, differentiable function f using GD:

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

Standard Analyses assume that ∇f is L-Lipschitz.

- L is the smallest constant such that for every x, y,

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

Background on L-Smoothness

Setting: minimize a convex, differentiable function f using GD:

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

Standard Analyses assume that ∇f is L-Lipschitz.

- L is the smallest constant such that for every x, y,

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

- To hold globally, the Lipschitz constant must reflect the worst-case growth of f,

$$L = \sup_{x,y} \frac{\|\nabla f(x) - \nabla f(y)\|_2}{\|x - y\|_2}.$$

Directional Smoothness

Definition: M is a directional smoothness function if $\forall x, y$,

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

Directional Smoothness

Definition: M is a directional smoothness function if $\forall x, y$,

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

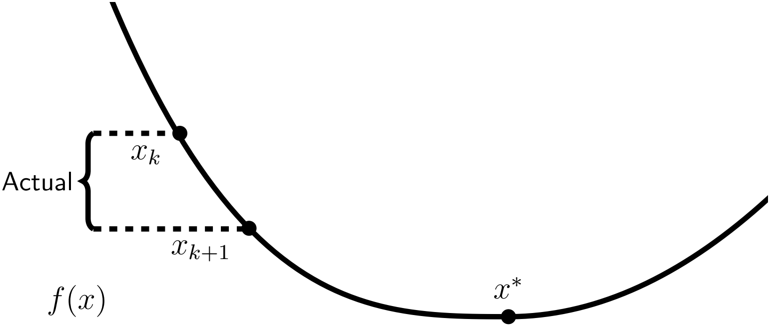
Tighter quadratic bounds imply a tighter descent lemma!

Directional Smoothness

Definition: M is a directional smoothness function if $\forall x, y$,

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

Tighter quadratic bounds imply a tighter descent lemma!

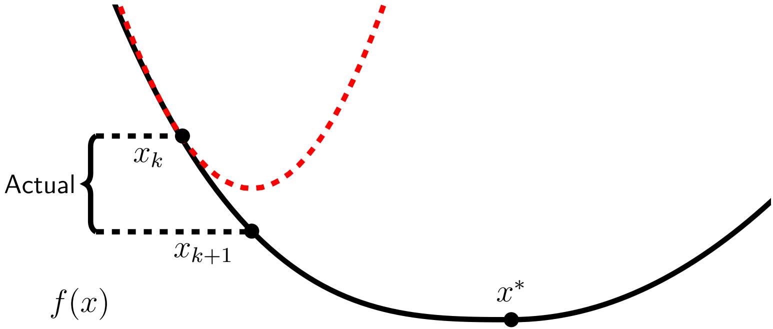


Directional Smoothness

Definition: M is a directional smoothness function if $\forall x, y$,

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

Tighter quadratic bounds imply a tighter descent lemma!

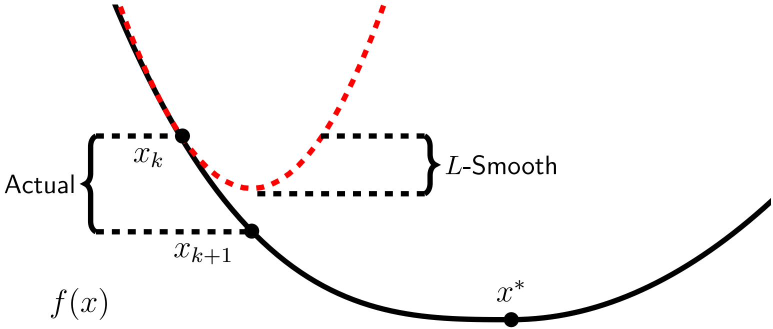


Directional Smoothness

Definition: M is a directional smoothness function if $\forall x, y$,

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

Tighter quadratic bounds imply a tighter descent lemma!

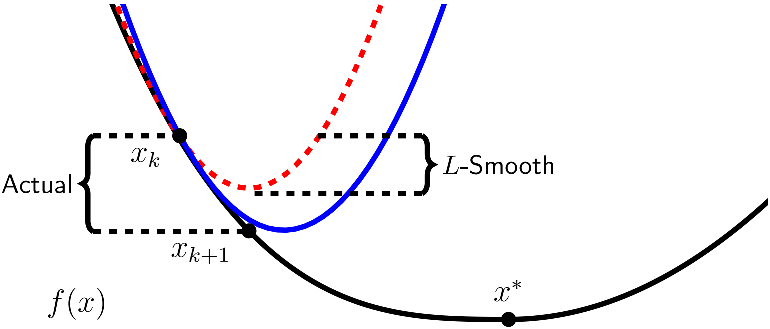


Directional Smoothness

Definition: M is a directional smoothness function if $\forall x, y$,

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

Tighter quadratic bounds imply a tighter descent lemma!

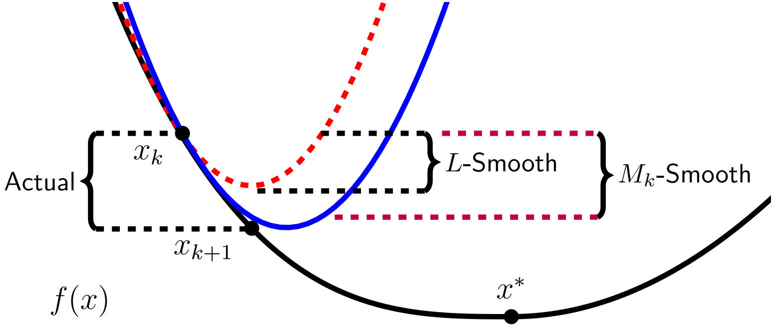


Directional Smoothness

Definition: M is a directional smoothness function if $\forall x, y$,

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

Tighter quadratic bounds imply a tighter descent lemma!



Concrete Directional Smoothness Functions

We give explicit directional smoothness functions — no oracles!

Concrete Directional Smoothness Functions

We give explicit directional smoothness functions — no oracles!

- **Point-wise Smoothness**:

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

Concrete Directional Smoothness Functions

We give explicit directional smoothness functions — no oracles!

- **Point-wise Smoothness**:

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

- **Path-wise Smoothness**:

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

Concrete Directional Smoothness Functions

We give explicit directional smoothness functions — no oracles!

- **Point-wise Smoothness**:

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

- **Path-wise Smoothness**:

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

- **Exact Smoothness**:

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

Path-Dependent Convergence Rates

Theorem (Convex)

If f is convex, then GD with any step-sizes $\{\eta_k\}$ satisfies,

Path-Dependent Convergence Rates

Theorem (Convex)

If f is convex, then GD with any step-sizes $\{\eta_k\}$ satisfies,

$$\min_{k\in[K]} f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^{K} \eta_i} + \frac{\sum_{i=0}^{K} \eta_i^2(\eta_i M_i - 1)\|\nabla f(x_i)\|_2^2}{2\sum_{i=0}^{K} \eta_i},$$

Path-Dependent Convergence Rates

Theorem (Convex)

If f is convex, then GD with any step-sizes $\{\eta_k\}$ satisfies,

$$\min_{k \in [K]} f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^{K} \eta_i} + \frac{\sum_{i=0}^{K} \eta_i^2 (\eta_i M_i - 1) \|\nabla f(x_i)\|_2^2}{2\sum_{i=0}^{K} \eta_i},$$

- If $\eta_k = 1/M(x_{k+1}(\eta_k), x_k)$ is strongly adapted, then we get path-dependent rates:

$$\min_{k \in [K]} f(x_k) - f(x^*) \leq \left[\frac{\sum_{i=0}^{K} M_i}{K+1}\right] \frac{\|x_0 - x^*\|_2^2}{K+1}$$

Strongly Adapted Step-Sizes

Computing strongly adapted step-sizes is an implicit equation,

$$\eta_k = 1/M(x_{k+1}(\eta_k), x_k).$$

Strongly Adapted Step-Sizes

Computing strongly adapted step-sizes is an implicit equation,

$$\eta_k = 1/M(x_{k+1}(\eta_k), x_k).$$

Does any method obtain match the strongly-adapted rate without knowing M?

The Polyak Step-Size

Polyak Step-size: assuming knowledge of $f(x^*)$, set $\gamma \geq 1$ and

$$\eta_k = \frac{f(x_k) - f(x^*)}{\gamma \|\nabla f(x_k)\|_2^2}.$$

The Polyak Step-Size

Polyak Step-size: assuming knowledge of $f(x^*)$, set $\gamma \geq 1$ and

$$\eta_k = \frac{f(x_k) - f(x^*)}{\gamma \|\nabla f(x_k)\|_2^2}.$$

Theorem (Informal)

If f is convex, then GD with Polyak step-size and $\gamma = 1.5$ satisfies

The Polyak Step-Size

Polyak Step-size: assuming knowledge of $f(x^*)$, set $\gamma \geq 1$ and

$$\eta_k = \frac{f(x_k) - f(x^*)}{\gamma \|\nabla f(x_k)\|_2^2}.$$

Theorem (Informal)

If f is convex, then GD with Polyak step-size and $\gamma = 1.5$ satisfies

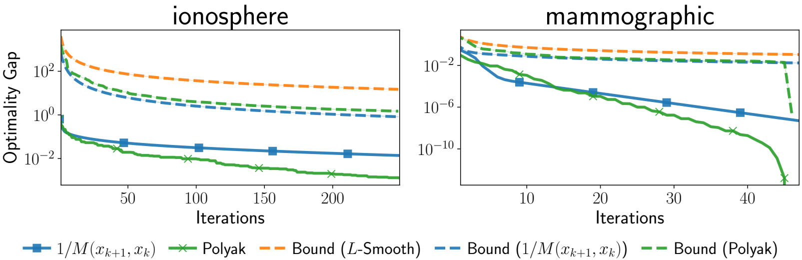
$$\min_{k \in [K]} f(x_k) - f(x^*) \leq \frac{3\|x_0 - x^*\|_2^2}{K} \left[\frac{\sum_{i=0}^{K} M(x_{i_1}, x_i)}{K} \right],$$

Directional Smoothness: Tighter Rates in Practice

How does this path-dependent theory compare to standard L-smooth rates?

Directional Smoothness: Tighter Rates in Practice

How does this path-dependent theory compare to standard L-smooth rates?



Learn more at our poster!

Thursday Dec. 12 at 4:30 p.m

References I