Function Curvature and Algorithm Complexity in Convex Optimization

Juan PEYPOUQUET

Dutch Optimization Seminar, April 29, 2021

First order methods

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable* function with minimizers

First order method: rule $x_k \mapsto x_{k+1} = T_k(x_k)$, where T_k depends on gradient evaluations and possibly some parameters.

First order methods

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable* function with minimizers

First order method: rule $x_k \mapsto x_{k+1} = T_k(x_k)$, where T_k depends on gradient evaluations and possibly some parameters.

- Sufficient decrease:

$$f(x_{k+1}) + \|w_{k+1}\|^2 \leq f(x_k).$$

First order methods

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable* function with minimizers

First order method: rule $x_k \mapsto x_{k+1} = T_k(x_k)$, where T_k depends on gradient evaluations and possibly some parameters.

- Sufficient decrease:

$$f(x_{k+1}) + \|w_{k+1}\|^2 \leq f(x_k).$$

- Identification and stability:

$$x_{k+1} - x_k \;\sim\; w_{k+1} \;\sim\; \nabla f(x_{k(+1)}).$$

Convergence rate and complexity

For first order methods, we have

$$f(x_k) - \min(f) \leq \frac{C}{k},$$

where C depends on x_0 and the parameters of the algorithm.

Convergence rate and complexity

For first order methods, we have

$$f(x_k) - \min(f) \leq \frac{C}{k},$$

where C depends on x_0 and the parameters of the algorithm.

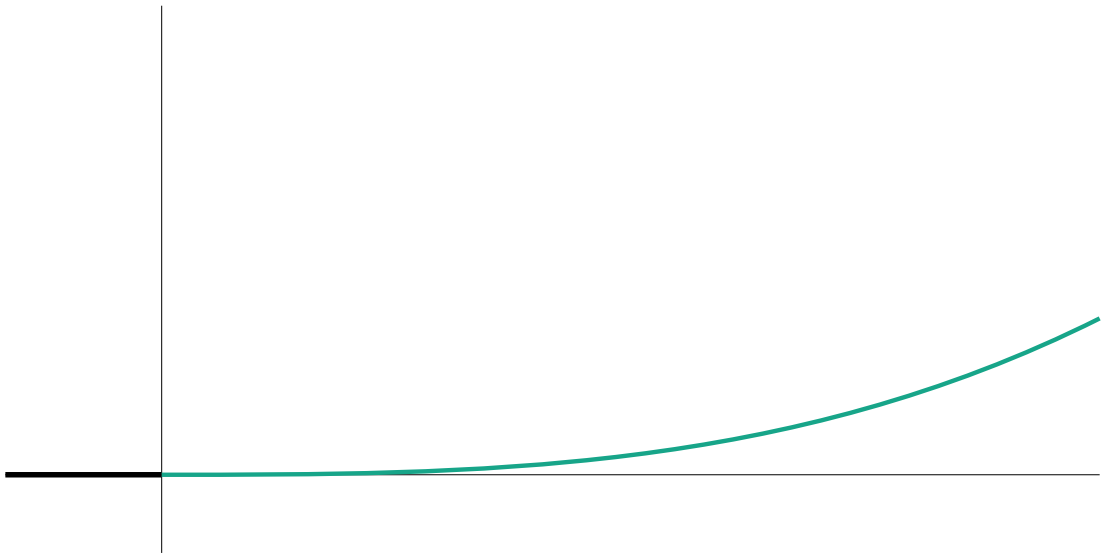
As a consequence, in order to ensure

$$f(x_k) - \min(f) \leq \varepsilon,$$

it is enough to perform $k = \mathcal{O}(\varepsilon^{-1})$ iterations.

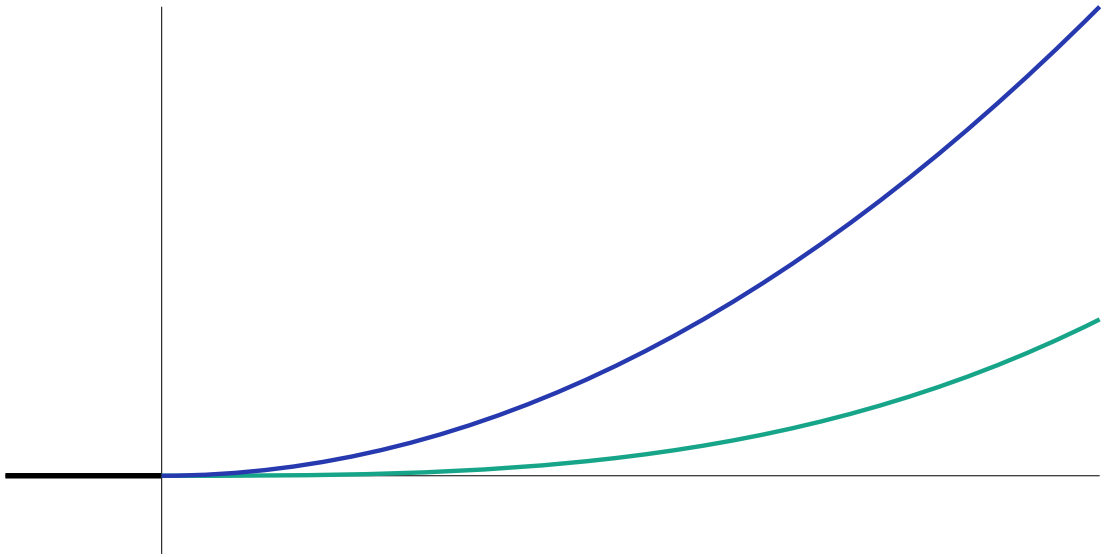
Complexity

But that is the worst-case scenario. In real life, complexity depends on the steepness of the function:



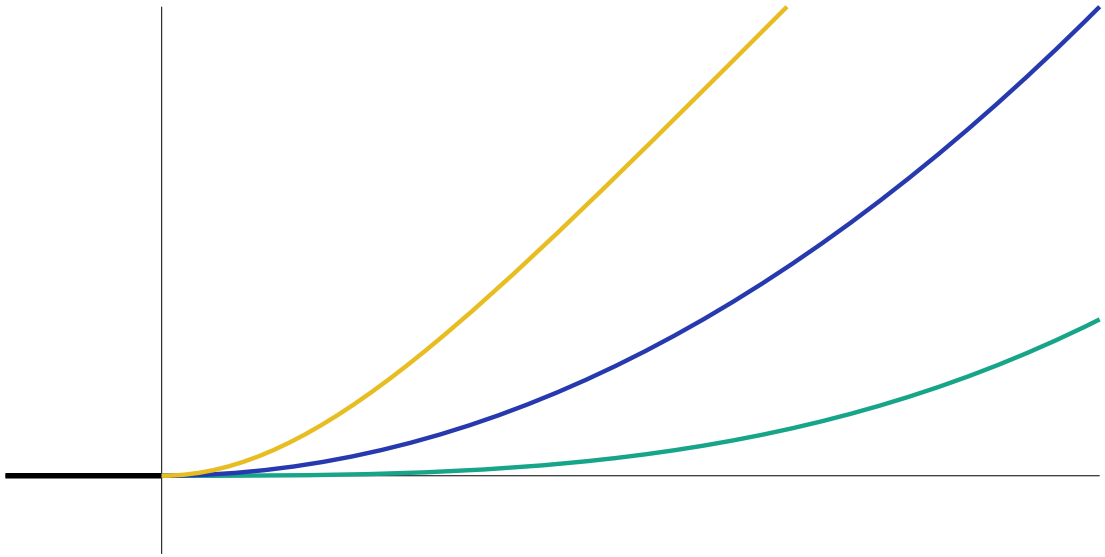
Complexity

But that is the worst-case scenario. In real life, complexity depends on the steepness of the function:



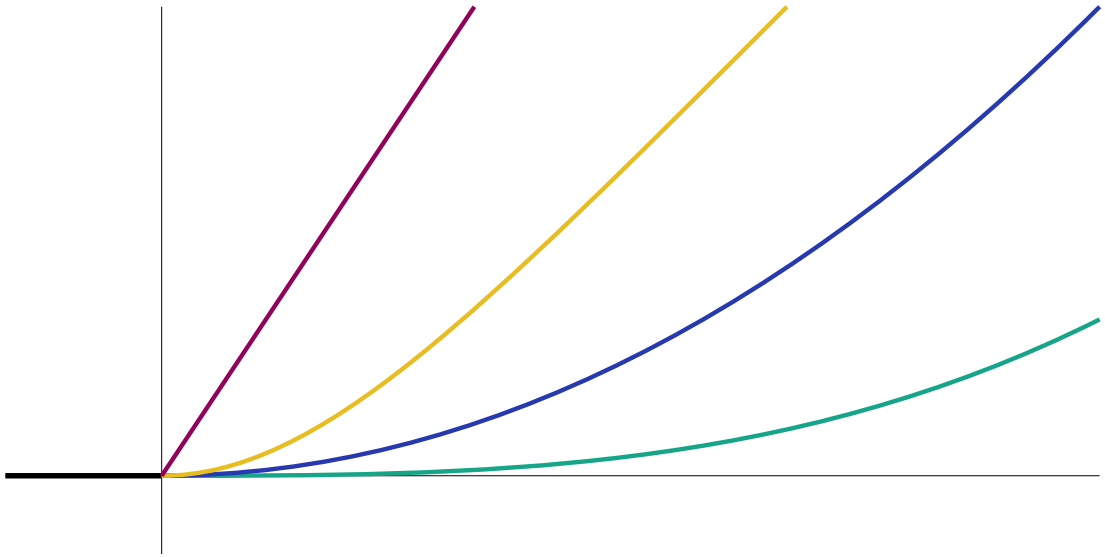
Complexity

But that is the worst-case scenario. In real life, complexity depends on the steepness of the function:



Complexity

But that is the worst-case scenario. In real life, complexity depends on the steepness of the function:



Quantifying steepness

S denotes the set of minimizers of f. For simplicity, we set $\min(f) = 0$

Error bound:

$$f(x) \geq \varphi\big(\text{dist}(x, S)\big).$$

Here, φ represents a height-to-distance relation.

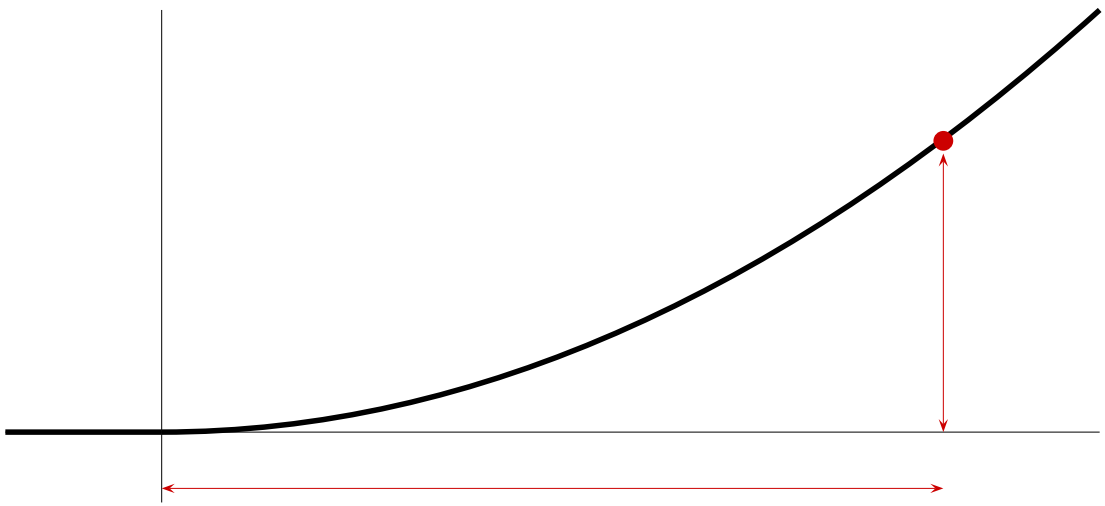
Quantifying steepness

S denotes the set of minimizers of f. For simplicity, we set $\min(f) = 0$

Error bound:

$$f(x) \geq \varphi\big(\text{dist}(x, S)\big).$$

Here, φ represents a height-to-distance relation.



Quantifying steepness

S denotes the set of minimizers of f. For simplicity, we set $\min(f) = 0$

Error bound:

$$f(x) \geq \varphi\big(\text{dist}(x, S)\big).$$

Common situation: $f(x) \geq \mu\big(\text{dist}(x, S)\big)^{\theta}$, with $\theta \geq 1$.

Quantifying steepness

S denotes the set of minimizers of f. For simplicity, we set $\min(f) = 0$

Error bound:

$$f(x) \geq \varphi\big(\text{dist}(x, S)\big).$$

Common situation: $f(x) \geq \mu\big(\text{dist}(x, S)\big)^{\theta}$, with $\theta \geq 1$.

- Strong convexity implies $S = \{x^*\}$ and $\theta = 2$.

Quantifying steepness

S denotes the set of minimizers of f. For simplicity, we set $\min(f) = 0$

Error bound:

$$f(x) \geq \varphi\big(\text{dist}(x, S)\big).$$

Common situation: $f(x) \geq \mu\big(\text{dist}(x, S)\big)^{\theta}$, with $\theta \geq 1$.

- Strong convexity implies $S = \{x^*\}$ and $\theta = 2$.
- Sharpness implies $\theta = 1$.

Quantifying steepness

For simplicity, we still set $\min(f) = 0$

Łojasiewicz inequality:

$$1 \leq \|\nabla(\phi \circ f)(x)\|$$

Quantifying steepness

For simplicity, we still set $\min(f) = 0$

Łojasiewicz inequality:

$$1 \leq \|\nabla\big(\phi \circ f\big)(x)\| = \phi'\big(f(x)\big)\|\nabla f(x)\|$$

Quantifying steepness

For simplicity, we still set $\min(f) = 0$

Łojasiewicz inequality:

$$1 \leq \|\nabla\big(\phi \circ f\big)(x)\| = \phi'\big(f(x)\big)\|\nabla f(x)\| \quad \Longleftrightarrow \quad f(x) \leq \psi\big(\|\nabla f(x)\|\big).$$

This gives a height-to-slope relation.

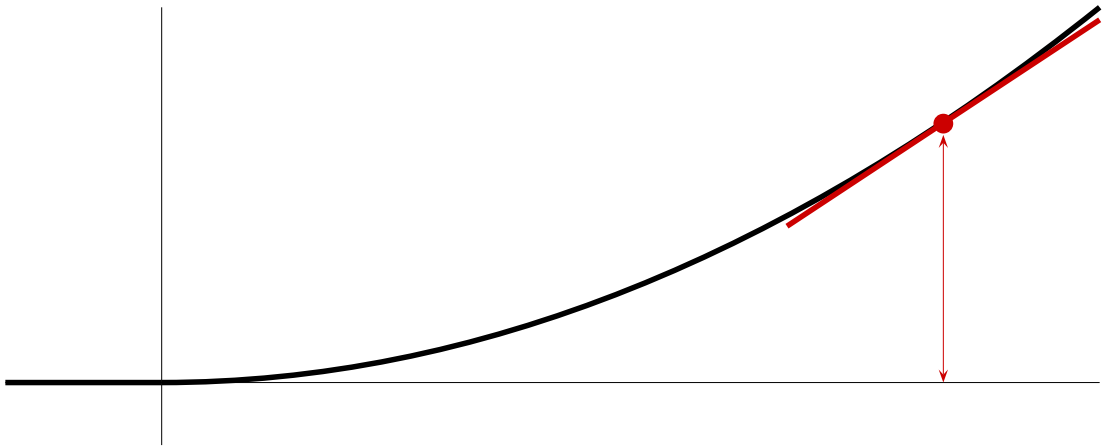
Quantifying steepness

For simplicity, we still set $\min(f) = 0$

Łojasiewicz inequality:

$$1 \leq \|\nabla\big(\phi \circ f\big)(x)\| = \phi'\big(f(x)\big)\|\nabla f(x)\| \quad \Longleftrightarrow \quad f(x) \leq \psi\big(\|\nabla f(x)\|\big).$$

This gives a height-to-slope relation.



Łojasiewicz inequalities and error bounds are equivalent*

Łojasiewicz inequalities and error bounds are equivalent*

Theorem

There exist $\mu > 0$ *and* $\theta > 1$ *such that*

$$f(x) \geq \mu\big(dist(x, S)\big)^{\theta}$$

if, and only if, there exist $\mu^* > 0$ *and* $\theta^* > 1$ *such that*

$$f(x) \leq \mu^* \|\nabla f(x)\|^{\theta^*}.$$

The numbers θ *and* θ^* *are Hölder conjugates.*

Łojasiewicz inequalities and error bounds are equivalent*

Theorem

There exist $\mu > 0$ *and* $\theta > 1$ *such that*

$$f(x) \geq \mu\big(dist(x, S)\big)^{\theta}$$

if, and only if, there exist $\mu^* > 0$ *and* $\theta^* > 1$ *such that*

$$f(x) \leq \mu^* \|\nabla f(x)\|^{\theta^*}.$$

The numbers θ *and* θ^* *are Hölder conjugates.*

The result holds with $\theta = 1$ *and* $\theta^* = \infty$ *as well.*

Back to first order methods

Rates of convergence under Łojasiewicz inequalities

We have

$$f(x_{k+1}) + \|\nabla f(x_{k+1})\|^2 \leq f(x_k).$$

Back to first order methods

Rates of convergence under Łojasiewicz inequalities

We have

$$f(x_{k+1}) + \|\nabla f(x_{k+1})\|^2 \leq f(x_k).$$

Using

$$f(x) \geq \mu\big(\text{dist}(x, S)\big)^\theta \qquad \Longleftrightarrow \qquad f(x) \leq \mu^* \|\nabla f(x)\|^{\theta^*},$$

Back to first order methods

Rates of convergence under Łojasiewicz inequalities

We have

$$f(x_{k+1}) + \|\nabla f(x_{k+1})\|^2 \leq f(x_k).$$

Using

$$f(x) \geq \mu\big(\text{dist}(x, S)\big)^\theta \qquad \Longleftrightarrow \qquad f(x) \leq \mu^* \|\nabla f(x)\|^{\theta^*},$$

we obtain

$$f(x_{k+1}) + b\, f(x_{k+1})^{2q} \leq f(x_k),$$

with $q = 1 - \frac{1}{\theta}$ and $b > 0$.

Back to convergence rates and complexity

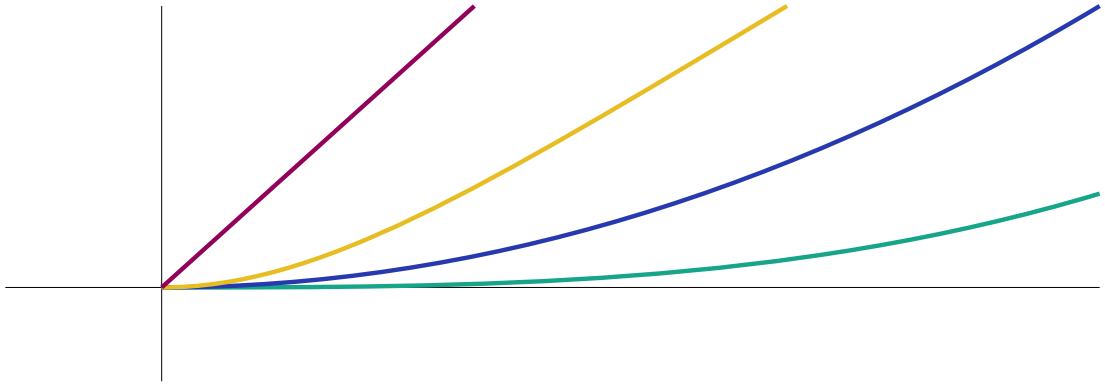
Suppose $f(x) \geq \mu(\text{dist}(x, S))^{\theta}$

$\theta = 1$: convergence in a finite number of steps (explicit number).

$\theta \in (1, 2)$: superlinear convergence.

$\theta = 2$: linear convergence.

$\theta > 2$: sublinear convergence.



Challenges

- Continuous-time dissipative dynamical systems

Challenges

- Continuous-time dissipative dynamical systems
- Global vs local

Challenges

- Continuous-time dissipative dynamical systems
- Global vs local
- Given f, compute θ
  - Unstable for the sum!

Challenges

- Continuous-time dissipative dynamical systems
- Global vs local
- Given f, compute θ
  - Unstable for the sum!
- Beyond first order methods
  - Higher order, accelerated, primal-dual, trust region, derivative-free

Challenges

- Continuous-time dissipative dynamical systems
- Global vs local
- Given f, compute θ
  - Unstable for the sum!
- Beyond first order methods
  - Higher order, accelerated, primal-dual, trust region, derivative-free
- Beyond optimization
  - Saddle points, equilibrium problems, games

Challenges

- Continuous-time dissipative dynamical systems
- Global vs local
- Given f, compute θ
  - Unstable for the sum!
- Beyond first order methods
  - Higher order, accelerated, primal-dual, trust region, derivative-free
- Beyond optimization
  - Saddle points, equilibrium problems, games
- Beyond deterministic algorithms
  - Stochastic gradient descent, variance reduction, data uncertainty