Recent advances in semidefinite programming performance estimation

Etienne de Klerk (joint work with Hadi Abbaszadehpeivasti and Moslem Zamani

August 29th, 2022

Tilburg University

The gradient method of Cauchy

The gradient descent method of Cauchy

Input: $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}^1 \in \mathbb{R}^n$, number of steps N and $\{t_k\}_{k=1}^N$ (step lengths).

for $k = 1, \ldots, N$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)$$

The gradient descent method of Cauchy

Input: $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}^1 \in \mathbb{R}^n$, number of steps N and $\{t_k\}_{k=1}^N$ (step lengths).

for $k = 1, \ldots, N$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)$$

Augustin-Louis Cauchy (1789–1857)

(Studied the gradient descent method in 1847.)

History

Cauchy described his method in the short note:

A. Cauchy. Méthode générale pour la résolution des systèmes déquations simultanées. C. R. Acad. Sci. Paris, 25:536–538, 1847.

History

Cauchy described his method in the short note:

A. Cauchy. Méthode générale pour la résolution des systèmes déquations simultanées. C. R. Acad. Sci. Paris, 25:536–538, 1847.

It was motivated by a problem in astronomy: to compute an orbit of a heavenly body as a nonlinear least squares problem:

> *'[to solve] the [algebraic] equations representing the motion of this body, taking as unknowns the elements of the orbit themselves. Then there are six such unknowns.'*

History: convergence analysis

Cauchy did not do any careful convergence analysis:

> *'If the new value of [the residual] is not a minimum, one can deduce, again proceeding in the same way, a [new] value still smaller; and, so continuing, smaller and smaller values will be found, which will converge to a minimal value. [The minimum residual] will always be obtained by the above method, provided that the [starting point is] suitably chosen.'*

History: convergence analysis

Cauchy did not do any careful convergence analysis:

> *'If the new value of [the residual] is not a minimum, one can deduce, again proceeding in the same way, a [new] value still smaller; and, so continuing, smaller and smaller values will be found, which will converge to a minimal value. [The minimum residual] will always be obtained by the above method, provided that the [starting point is] suitably chosen.'*

To obtain more precise convergence results, one needs to restrict the class of objective functions ...

Outline of this talk

We consider the worst-case performance of the gradient method for different classes of non-convex functions:

- First we consider all L-smooth functions ...
- ... and then restrict to L-smooth functions satisfying a growth condition at stationary points, ...
- ... and finally restrict further to functions of bounded minimum curvature (hypoconvex functions).

Gradient descent for L-smooth functions

Gradient descent for L-smooth functions

- We first analyse the worst-case convergence for L-smooth functions,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

  for some (known) Lipschitz constant $L > 0$.

Gradient descent for L-smooth functions

- We first analyse the worst-case convergence for L-smooth functions,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

  for some (known) Lipschitz constant $L > 0$. **Notation:** $f \in \mathcal{F}_L(\mathbb{R}^n)$.

Gradient descent for L-smooth functions

- We first analyse the worst-case convergence for L-smooth functions,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

  for some (known) Lipschitz constant $L > 0$. **Notation:** $f \in \mathcal{F}_L(\mathbb{R}^n)$.
- We also assume f has a global minimizer $\mathbf{x}^\star$, and write $f(\mathbf{x}^\star) = f^\star$.

Gradient descent for L-smooth functions

- We first analyse the worst-case convergence for L-smooth functions,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

  for some (known) Lipschitz constant $L > 0$. **Notation:** $f \in \mathcal{F}_L(\mathbb{R}^n)$.
- We also assume f has a global minimizer $\mathbf{x}^\star$, and write $f(\mathbf{x}^\star) = f^\star$.
- Basic property ('descent lemma') for $f \in \mathcal{F}_L(\mathbb{R}^n)$:

$$f(\mathbf{x}) - f^\star \geq \tfrac{1}{2L}\|\nabla f(\mathbf{x})\|^2, \ \forall \mathbf{x} \in \mathbb{R}^n.$$

Worst-case performance estimation

Main tool: semidefinite programming (SDP) performance estimation, introduced in:

Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

Worst-case performance estimation

Main tool: semidefinite programming (SDP) performance estimation, introduced in:

Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

Basic idea: find the worst-case $f \in \mathcal{F}_L(\mathbb{R}^n)$ for gradient descent as the analytic solution of a tractable optimization problem (an SDP problem).

Known results on convergence rate to stationary point

- (Classical, e.g. Nesterov textbook) When $t_k = \frac{1}{L}$, $k \in \{1, ..., N\}$:

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(\mathbf{x}^k) \right\| \leq \left(\frac{2L(f(\mathbf{x}^1) - f^\star)}{N} \right)^{\frac{1}{2}}.$$

Known results on convergence rate to stationary point

- (Classical, e.g. Nesterov textbook) When $t_k = \frac{1}{L}$, $k \in \{1, ..., N\}$:

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(\mathbf{x}^k) \right\| \leq \left(\frac{2L(f(\mathbf{x}^1) - f^\star)}{N} \right)^{\frac{1}{2}}.$$

- If all $t_k < \frac{1}{L}$:

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(\mathbf{x}^k) \right\| \leq \left(\frac{4(f(\mathbf{x}^1) - f^\star)}{\sum_{k=1}^{N} t_k(4 - Lt_k)} \right)^{\frac{1}{2}}.$$

Drori, Y., Shamir, O.: The complexity of finding stationary points with stochastic gradient descent. In: Proceedings of the 37th International Conference on Machine Learning, pp. 2658–2667 (2020)

SDP performance analysis

Worst-case analysis

$$\max \left(\min_{1 \le k \le N+1} \left\| \nabla f(\mathbf{x}^k) \right\| \right)$$

s.t. $f(\mathbf{x}^1) - f^\star \le \Delta$

$\mathbf{x}^{N+1}, \mathbf{x}^N, ..., \mathbf{x}^2$ are generated by gradient descent w.r.t. $f, \mathbf{x}^1$

$f(\mathbf{x}) - f^\star \ge \frac{1}{2L} \|\nabla f(\mathbf{x})\|^2, \ \forall \mathbf{x} \in \mathbb{R}^n$

$f \in \mathcal{F}_L(\mathbb{R}^n)$

$\mathbf{x}^1 \in \mathbb{R}^n,$

with variables f and $\mathbf{x}^1$, and given parameters $\Delta > 0, L > 0, N > 0$ and the step lengths $t_k > 0$.

Worst-case analysis

$$\max \left(\min_{1 \le k \le N+1} \left\| \nabla f(\mathbf{x}^k) \right\| \right)$$

s.t. $f(\mathbf{x}^1) - f^\star \le \Delta$

$\mathbf{x}^{N+1}, \mathbf{x}^N, \ldots, \mathbf{x}^2$ are generated by gradient descent w.r.t. $f, \mathbf{x}^1$

$f(\mathbf{x}) - f^\star \ge \frac{1}{2L} \|\nabla f(\mathbf{x})\|^2, \ \forall \mathbf{x} \in \mathbb{R}^n$

$f \in \mathcal{F}_L(\mathbb{R}^n)$

$\mathbf{x}^1 \in \mathbb{R}^n,$

with variables f and $\mathbf{x}^1$, and given parameters $\Delta > 0, L > 0, N > 0$ and the step lengths $t_k > 0$.

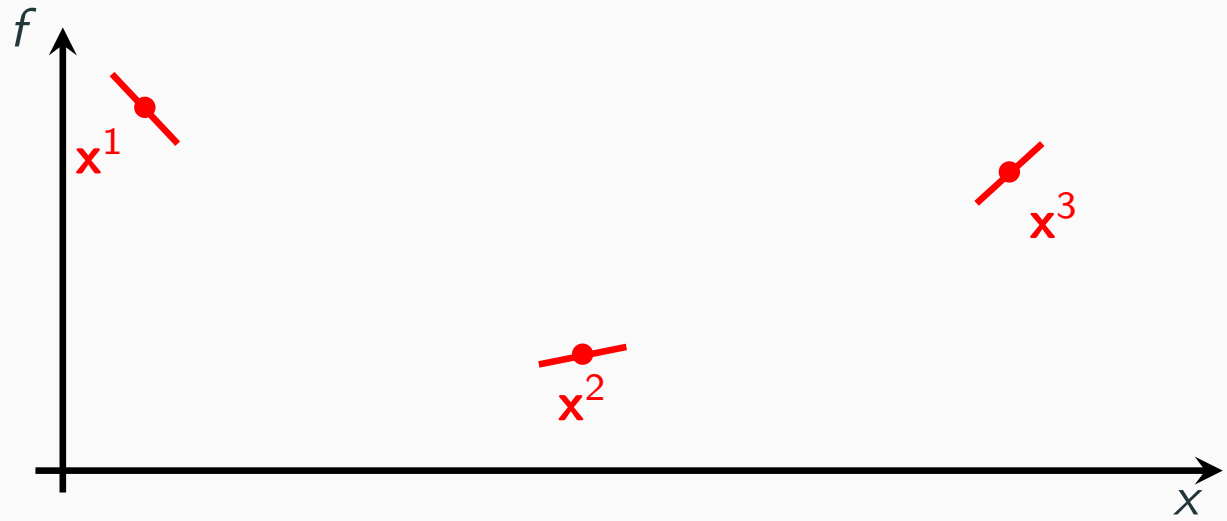
Key idea: This can be solved using semidefinite programming (SDP) by representing $\mathcal{F}_L(\mathbb{R}^n)$ via interpolation.

L-smooth Interpolation Problem

Consider an index set I, and given values $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i \in I}$ where $\mathbf{x}^i \in \mathbb{R}^n$, $\mathbf{g}^i \in \mathbb{R}^n$ and $f^i \in \mathbb{R}$.

L-smooth Interpolation Problem

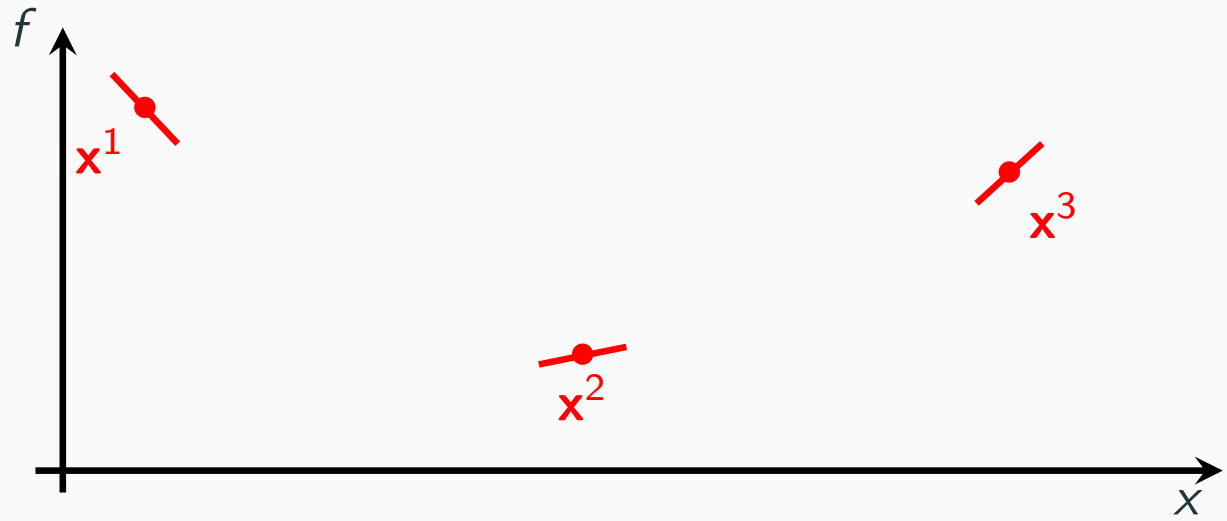
Consider an index set I, and given values $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i \in I}$ where $\mathbf{x}^i \in \mathbb{R}^n$, $\mathbf{g}^i \in \mathbb{R}^n$ and $f^i \in \mathbb{R}$.



$$?\exists \;\; f \in \mathcal{F}_L(\mathbb{R}^n) \colon \; f(\mathbf{x}^i) = f^i, \quad \text{and} \quad \mathbf{g}^i = \nabla f(\mathbf{x}^i), \qquad \forall i \in S.$$

L-smooth Interpolation Problem

Consider an index set I, and given values $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i\in I}$ where $\mathbf{x}^i \in \mathbb{R}^n$, $\mathbf{g}^i \in \mathbb{R}^n$ and $f^i \in \mathbb{R}$.



$$?\exists\ f \in \mathcal{F}_L(\mathbb{R}^n):\ f(\mathbf{x}^i) = f^i, \quad \text{and} \quad \mathbf{g}^i = \nabla f(\mathbf{x}^i), \qquad \forall i \in S.$$

If yes, we say $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i\in I}$ is $\mathcal{F}_L(\mathbb{R}^n)$-interpolable.

L-smooth Interpolation (ctd.)

Theorem (Taylor, Hendrickx, and Glineur (2017))

The following statements are equivalent:

1. $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i \in I}$ *is* $\mathcal{F}_L(\mathbb{R}^n)$*-interpolable;*
2. $\forall i, j \in I$*:*

$$\tfrac{1}{2L}\left\|\mathbf{g}^i - \mathbf{g}^j\right\|^2 - \tfrac{L}{4}\left\|\mathbf{x}^i - \mathbf{x}^j - \tfrac{1}{L}(\mathbf{g}^i - \mathbf{g}^j)\right\|^2 \leq f^i - f^j - \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j \right\rangle.$$

L-smooth Interpolation (ctd.)

Theorem (Taylor, Hendrickx, and Glineur (2017))

The following statements are equivalent:

1. $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i \in I}$ *is* $\mathcal{F}_L(\mathbb{R}^n)$*-interpolable;*
2. $\forall i, j \in I$*:*

$$\tfrac{1}{2L}\left\|\mathbf{g}^i - \mathbf{g}^j\right\|^2 - \tfrac{L}{4}\left\|\mathbf{x}^i - \mathbf{x}^j - \tfrac{1}{L}(\mathbf{g}^i - \mathbf{g}^j)\right\|^2 \leq f^i - f^j - \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j \right\rangle.$$

A.B. Taylor, J.M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization* 27(3), 1283–1313 (2017)

L-smooth Interpolation: refininement

Theorem (Drori, Shamir (2020))

By the previous theorem, the following statements are equivalent:

1. $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i \in I}$ *is* $\mathcal{F}_L(\mathbb{R}^n)$*-interpolable;*
2. $\forall i, j \in I$*:*

$$\tfrac{1}{2L}\left\|\mathbf{g}^i - \mathbf{g}^j\right\|^2 - \tfrac{L}{4}\left\|\mathbf{x}^i - \mathbf{x}^j - \tfrac{1}{L}(\mathbf{g}^i - \mathbf{g}^j)\right\|^2 \leq f^i - f^j - \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j \right\rangle.$$

L-smooth Interpolation: refininement

Theorem (Drori, Shamir (2020))

By the previous theorem, the following statements are equivalent:

1. $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i\in I}$ *is* $\mathcal{F}_L(\mathbb{R}^n)$*-interpolable;*
2. $\forall i, j \in I$*:*

$$\tfrac{1}{2L}\left\|\mathbf{g}^i - \mathbf{g}^j\right\|^2 - \tfrac{L}{4}\left\|\mathbf{x}^i - \mathbf{x}^j - \tfrac{1}{L}(\mathbf{g}^i - \mathbf{g}^j)\right\|^2 \leq f^i - f^j - \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j \right\rangle.$$

Moreover, w.l.o.g., the interpolating $f \in \mathcal{F}_L(\mathbb{R}^n)$ *satisfies*

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(x) = \min_{i\in I} f_i - \tfrac{1}{2L}\|\mathbf{g}^i\|^2.$$

SDP formulation (tight!)

$$
\begin{aligned}
\max \ & \left(\min_{1 \le k \le N+1} \left\| \mathbf{g}^k \right\| \right) \\
\text{s.t. } & \tfrac{1}{2L} \left\| \mathbf{g}^i - \mathbf{g}^j \right\|^2 - \tfrac{L}{4} \left\| \mathbf{x}^i - \mathbf{x}^j - \tfrac{1}{L}(\mathbf{g}^i - \mathbf{g}^j) \right\|^2 \le f^i - f^j - \\
& \qquad \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j \right\rangle \quad i,j \in \{1, ..., N+1, \star\} \\
& \mathbf{x}^{k+1} = \mathbf{x}^k - t_k \mathbf{g}^k \quad k \in \{1, ..., N+1\} \\
& f^k \ge f^\star + \tfrac{1}{2L} \|\mathbf{g}^k\|^2 \quad k \in \{1, ..., N+1\} \\
& f^1 - f^\star \le \Delta \\
& \mathbf{g}^\star = 0,
\end{aligned}
$$

with variables $\mathbf{x}^1$, f^k, $\mathbf{g}^k$ $(1 \le k \le N+1)$, and given parameters L, Δ, N and t_k, $(1 \le k \le N+1)$.

Worst-case convergence results

Main result

Theorem

Consider N steps of the gradient method with step lengths $t_k \in (0, \frac{\sqrt{3}}{L})$ for $k \in \{1, ..., N\}$, applied to some $f \in \mathcal{F}_L(\mathbb{R}^n)$ with starting point $\mathbf{x}^1$.

Main result

Theorem

Consider N steps of the gradient method with step lengths $t_k \in (0, \frac{\sqrt{3}}{L})$ for $k \in \{1, ..., N\}$, applied to some $f \in \mathcal{F}_L(\mathbb{R}^n)$ with starting point $\mathbf{x}^1$. Then

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(\mathbf{x}^k) \right\|^2 \leq \frac{4(f(\mathbf{x}^1) - f^\star)}{\sum_{k=1}^{N} \min\{-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k\} + \frac{2}{L}},$$

and this bound is tight in some cases.

Main result

Theorem

Consider N steps of the gradient method with step lengths $t_k \in (0, \frac{\sqrt{3}}{L})$ for $k \in \{1, ..., N\}$, applied to some $f \in \mathcal{F}_L(\mathbb{R}^n)$ with starting point $\mathbf{x}^1$. Then

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(\mathbf{x}^k) \right\|^2 \leq \frac{4(f(\mathbf{x}^1) - f^\star)}{\sum_{k=1}^{N} \min\{-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k\} + \frac{2}{L}},$$

and this bound is tight in some cases.

Corollary 1

Let $t_k = \frac{1}{L}$ for $k \in \{1, ..., N\}$. Then

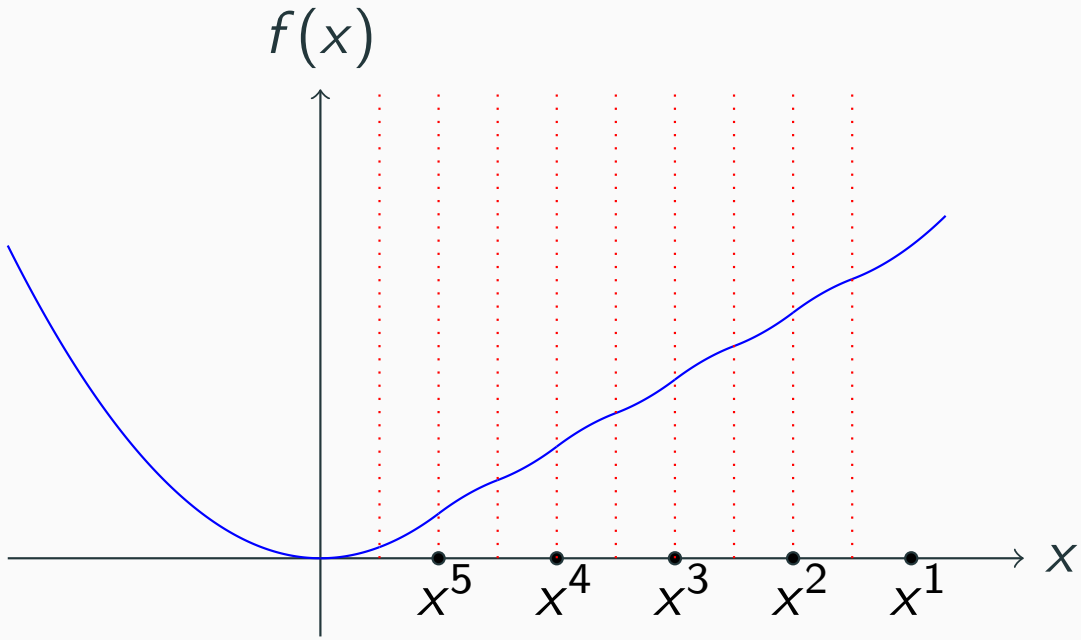
$$\min_{1 \leq k \leq N+1} \left\| \nabla f(\mathbf{x}^k) \right\| \leq \left(\frac{4L(f(\mathbf{x}^1) - f^\star)}{3N+2} \right)^{\frac{1}{2}}.$$

Proof sketch

- The proof follows from weak duality for the SDP, given the right Lagrange multipliers of the SDP constraints.
- The multipliers were first guessed by solving the SDP numerically for various choices of L, Δ, N, ...
- ... and noting the optimal multipliers.
- Finally, the correctness of the inequality was verified through (symbolic) calculation. (A dual feasible solution was constructed.)
- The fact that the bound can be tight is demonstrated by a family of examples ... [next slide]

Example (picture only)

We can construct piecewise quadratic univariate f that are L-smooth with $f^\star = 0$ and $x^\star = 0$, that attain the bound in the corollary for suitable x^1, and all $t_k = \frac{1}{L}$.



(Graph of constructed f for $L = 1$, $\Delta = 2$, $N = 4$.)

Second corollary

Corollary 2

Let f be an L-smooth function. Then the optimal step size for the gradient method is given by

$$t_k = \frac{\sqrt{\frac{4}{3}}}{L} \quad \forall k \in \{1, \ldots, N\},$$

provided that $t_k \in (0, \frac{\sqrt{3}}{L})$ for all $k \in \{1, \ldots, N\}$.

The proof is by minimizing the right-hand-side in the theorem over the t_k.

Substituting this step length leads to

$$\min_{1 \leq k \leq N+1} \left\| \nabla f(\mathbf{x}^k) \right\| \leq \left(\frac{6\sqrt{3}L(f(\mathbf{x}^1) - f^\star)}{8N + 3\sqrt{3}} \right)^{\frac{1}{2}}.$$

Better constant (≈ 1.299) than for $t_k = \frac{1}{L}$ ($4/3 = 1.333\ldots$).

Adding the Polyak-Łojasiewicz (PŁ) inequality

The Polyak-Łojasiewicz (PŁ) inequality

Definition

A function f is said to satisfy the PŁ inequality on $X \subseteq \mathbb{R}^n$ if there exists $\mu_p > 0$ such that

$$f(\mathbf{x}) - f^\star \leq \frac{1}{2\mu_p} \|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in X. \tag{5.1}$$

The Polyak-Łojasiewicz (PŁ) inequality

Definition

A function f is said to satisfy the PŁ inequality on $X \subseteq \mathbb{R}^n$ if there exists $\mu_p > 0$ such that

$$f(\mathbf{x}) - f^\star \leq \frac{1}{2\mu_p} \|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in X. \tag{5.1}$$

- An f that satisfies the PŁ inequality is also known as *gradient dominated*.

The Polyak-Łojasiewicz (PŁ) inequality

Definition

A function f is said to satisfy the PŁ inequality on $X \subseteq \mathbb{R}^n$ if there exists $\mu_p > 0$ such that

$$f(\mathbf{x}) - f^\star \leq \frac{1}{2\mu_p} \|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in X. \tag{5.1}$$

- An f that satisfies the PŁ inequality is also known as *gradient dominated*.
- Under the PŁ inequality assumption, every stationary point of f is a global minimizer.

The Polyak-Łojasiewicz (PŁ) inequality

Definition

A function f is said to satisfy the PŁ inequality on $X \subseteq \mathbb{R}^n$ if there exists $\mu_p > 0$ such that

$$f(\mathbf{x}) - f^\star \leq \frac{1}{2\mu_p} \|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in X. \tag{5.1}$$

- An f that satisfies the PŁ inequality is also known as *gradient dominated*.
- Under the PŁ inequality assumption, every stationary point of f is a global minimizer.
- PŁ is weaker than convexity.

Gradient descent and the PŁ inequality

Classical linear convergence rate result:

Theorem

Let f be L-smooth and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. If $t_1 = \frac{1}{L}$, we have

$$f(\mathbf{x}^2) - f^\star \leq \frac{L - \mu_p}{L} \left(f(\mathbf{x}^1) - f^\star\right).$$

Reference:

Polyak, B.T.: Gradient methods for the minimisation of functionals. USSR Computational Mathematics and Mathematical Physics 3(4), 864–878 (1963)

New result via SDP performance estimation

Via suitable SDP performance estimation:

Theorem

Let f be L-smooth and let f satisfy PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. If $t_1 = \frac{1}{L}$, we have

$$f(\mathbf{x}^2) - f^\star \leq \left(\frac{L - \mu_p}{L + \frac{1}{2}\mu_p}\right)\left(f(\mathbf{x}^1) - f^\star\right).$$

New result via SDP performance estimation

Via suitable SDP performance estimation:

Theorem

Let f be L-smooth and let f satisfy PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. If $t_1 = \frac{1}{L}$, we have

$$f(\mathbf{x}^2) - f^\star \leq \left(\frac{L - \mu_p}{L + \frac{1}{2}\mu_p}\right)\left(f(\mathbf{x}^1) - f^\star\right).$$

Thus we improve the constant from $\frac{L-\mu_p}{L}$ to $\frac{L-\mu_p}{L+\frac{1}{2}\mu_p}$.

New result via SDP performance estimation

Via suitable SDP performance estimation:

Theorem

Let f be L-smooth and let f satisfy PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. If $t_1 = \frac{1}{L}$, we have

$$f(\mathbf{x}^2) - f^\star \leq \left(\frac{L - \mu_p}{L + \frac{1}{2}\mu_p}\right)\left(f(\mathbf{x}^1) - f^\star\right).$$

Thus we improve the constant from $\frac{L-\mu_p}{L}$ to $\frac{L-\mu_p}{L+\frac{1}{2}\mu_p}$.

Interesting observation:

For $f \in \mathcal{F}_L(\mathbb{R}^n)$, the gradient descent method with step length $1/L$ is linearly convergent if and only if the PŁ inequality holds on $\{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$.

Restriction to functions of bounded curvature

Functions of bounded curvature

- The function f has a maximum curvature $0 \leq L < \infty$ if

$$\mathbf{x} \mapsto \frac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x}) \text{ is convex } \ldots$$

Functions of bounded curvature

- The function f has a maximum curvature $0 \leq L < \infty$ if

$$\mathbf{x} \mapsto \tfrac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x}) \text{ is convex ...}$$

- ... and minimum curvature $-\infty < \mu \leq L$ if

$$f(\mathbf{x}) - \tfrac{\mu}{2}\|\mathbf{x}\|^2 \text{ is convex.}$$

Functions of bounded curvature

- The function f has a maximum curvature $0 \leq L < \infty$ if

$$\mathbf{x} \mapsto \tfrac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x}) \text{ is convex ...}$$

- ... and minimum curvature $-\infty < \mu \leq L$ if

$$f(\mathbf{x}) - \tfrac{\mu}{2}\|\mathbf{x}\|^2 \text{ is convex.}$$

Notation: $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$.

- $\mathcal{F}_{0,L}(\mathbb{R}^n)$: L-smooth convex functions.
- $\mathcal{F}_{-L,L}(\mathbb{R}^n)$: L-smooth functions.
- If $f \in C^2(\mathbb{R}^n)$, then $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ iff

$$\mu I \preceq \nabla^2 f(\mathbf{x}) \preceq LI, \quad \forall\, \mathbf{x} \in \mathbb{R}^n.$$

Interpolation Theorem

Theorem (Rotaru, Glineur, Panagiotis (2022))

Let $L \in (0, \infty]$ and $\mu \in (-\infty, L]$. The following statements are equivalent:

1. $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i \in I}$ *is $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$-interpolable;*
2. $\forall i, j \in I$*:*

$$\begin{aligned}&\left(\tfrac{1}{L}\left\|\mathbf{g}^i - \mathbf{g}^j\right\|^2 + \mu\left\|\mathbf{x}^i - \mathbf{x}^j\right\|^2 - \tfrac{2\mu}{L}\left\langle \mathbf{g}^j - \mathbf{g}^i, \mathbf{x}^j - \mathbf{x}^i\right\rangle\right)\\ \leq &2(1-\tfrac{\mu}{L})\left(f^i - f^j - \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j\right\rangle\right).\end{aligned}$$

Interpolation Theorem

Theorem (Rotaru, Glineur, Panagiotis (2022))

Let $L \in (0, \infty]$ and $\mu \in (-\infty, L]$. The following statements are equivalent:

1. $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i \in I}$ *is $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$-interpolable;*
2. $\forall i, j \in I$*:*

$$\begin{aligned}&\left(\tfrac{1}{L}\left\|\mathbf{g}^i - \mathbf{g}^j\right\|^2 + \mu\left\|\mathbf{x}^i - \mathbf{x}^j\right\|^2 - \tfrac{2\mu}{L}\left\langle \mathbf{g}^j - \mathbf{g}^i, \mathbf{x}^j - \mathbf{x}^i\right\rangle\right)\\ \leq &2(1-\tfrac{\mu}{L})\left(f^i - f^j - \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j\right\rangle\right).\end{aligned}$$

Rotaru, T., Glineur, F., & Patrinos, P. (2022). Tight convergence rates of the gradient method on hypoconvex functions. arXiv preprint arXiv:2203.00775.

SDP formulation: one step of gradient method

$$
\begin{aligned}
\max\ & \frac{f^2 - f^\star}{f^1 - f^\star} \\
\text{s.t. } & \frac{1}{2(1-\frac{\mu}{L})}\left(\frac{1}{L}\left\|\mathbf{g}^i - \mathbf{g}^j\right\|^2 + \mu\left\|\mathbf{x}^i - \mathbf{x}^j\right\|^2 - \frac{2\mu}{L}\left\langle \mathbf{g}^j - \mathbf{g}^i, \mathbf{x}^j - \mathbf{x}^i\right\rangle\right) \leq \\
& f^i - f^j - \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j\right\rangle \quad i,j \in \{1,2\} \\
& \mathbf{x}^2 = \mathbf{x}^1 - t_1\mathbf{g}^1 \\
& f^k \geq f^\star \quad k \in \{1,2\} \\
& f^k - f^\star \leq \frac{1}{2\mu_p}\|\mathbf{g}^k\|^2, \quad k \in \{1,2\}.
\end{aligned}
$$

with variables $\mathbf{x}^k$, f^k, $\mathbf{g}^k$, $k \in \{1,2\}$, and given parameters μ, L, μ_p and t_1.

New results

Bounds after one step of gradient method

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0, \infty), \mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$.

Bounds after one step of gradient method

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0,\infty), \mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. Then

If $t_1 \in \left(0, \frac{1}{L}\right)$:

$$\sqrt{\frac{f(\mathbf{x}^2) - f^\star}{f(\mathbf{x}^1) - f^\star}}$$

$$\leq \frac{\mu_p (1 - L t_1) + \sqrt{(L-\mu)(\mu - \mu_p)(2 - L t_1)\mu_p t_1 + (L-\mu)^2}}{L - \mu + \mu_p}$$

$$< 1.$$

Main result (ctd.)

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0, \infty), \mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. Suppose that $\mathbf{x}^2$ is generated by the gradient descent

Main result (ctd.)

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0, \infty), \mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. Suppose that $\mathbf{x}^2$ is generated by the gradient descent then

If $t_1 \in \left[\frac{1}{L}, \frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}\right]$**:**

$$\frac{f(\mathbf{x}^2) - f^\star}{f(\mathbf{x}^1) - f^\star} \leq \left(\frac{(Lt_1 - 2)(\mu t_1 - 2)\mu_p t_1}{(L + \mu - \mu_p)\, t_1 - 2} + 1\right)$$

Main result (ctd.)

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0,\infty), \mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. Suppose that $\mathbf{x}^2$ is generated by the gradient descent then

If $t_1 \in \left[\frac{1}{L}, \frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}\right]$**:**

$$\begin{aligned} \frac{f(\mathbf{x}^2) - f^\star}{f(\mathbf{x}^1) - f^\star} &\leq \left(\frac{(Lt_1 - 2)(\mu t_1 - 2)\mu_p t_1}{(L + \mu - \mu_p)\, t_1 - 2} + 1\right) \\ &= \frac{L - \mu_p}{L + \frac{1}{2}\mu_p} \text{ if } t_1 = 1/L \text{ and } \mu = -L. \end{aligned}$$

Main result (ctd.)

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0,\infty), \mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. Suppose that $\mathbf{x}^2$ is generated by the gradient descent then

If $t_1 \in \left[\frac{1}{L}, \frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}\right]$**:**

$$\begin{aligned} \frac{f(\mathbf{x}^2) - f^\star}{f(\mathbf{x}^1) - f^\star} &\leq \left(\frac{(Lt_1 - 2)(\mu t_1 - 2)\mu_p t_1}{(L + \mu - \mu_p)\, t_1 - 2} + 1\right) \\ &= \frac{L - \mu_p}{L + \frac{1}{2}\mu_p} \text{ if } t_1 = 1/L \text{ and } \mu = -L. \end{aligned}$$

Thus we recover the earlier result for L-smooth functions ($\mu = -L$) for step length $1/L$.

Main result (ctd.)

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0, \infty), \mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. Suppose that $\mathbf{x}^2$ is generated by the gradient descent

Main result (ctd.)

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ with $L \in (0, \infty), \mu \in (-\infty, 0]$ and let f satisfy the PŁ inequality on $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^1)\}$. Suppose that $\mathbf{x}^2$ is generated by the gradient descent then

If $t_1 \in \left(\frac{3}{\mu+L+\sqrt{\mu^2-L\mu+L^2}}, \frac{2}{L}\right)$**:**

$$\frac{f(\mathbf{x}^2) - f^\star}{f(\mathbf{x}^1) - f^\star} \leq \frac{(Lt_1 - 1)^2}{(Lt_1 - 1)^2 + \mu_p t_1(2 - Lt_1)}.$$

Final remarks

Discussion

- The new results for gradient descent for L-smooth f are from: *Optimization Letters*, **16**, 1649–1661 (2022), ...

Discussion

- The new results for gradient descent for L-smooth f are from: *Optimization Letters*, **16**, 1649–1661 (2022), ...
- ... and with PŁ added: arXiv:2204.00647

Discussion

- The new results for gradient descent for L-smooth f are from: *Optimization Letters*, **16**, 1649–1661 (2022), ...
- ... and with PŁ added: arXiv:2204.00647
- Good introduction to SDP performance estimation: the PhD thesis by Yoel Drori or the PhD thesis by Adrien Taylor (both online).

The End