Performance estimation of the (stochastic) gradient method for L-smooth functions

Etienne de Klerk

July, 2026

Tilburg University

The gradient method of Cauchy

The gradient descent method of Cauchy

Input: $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}_0 \in \mathbb{R}^n$, number of steps N and $\{t_k\}_{k=0}^{N-1}$ (step lengths).

for $k = 0, \ldots, N-1$

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

We will assume

- f is L-smooth;
- f has a global minimizer $\mathbf{x}_\star$, and write $f(\mathbf{x}_\star) = f_\star$.

Gradient descent for L-smooth functions

L-smooth functions

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

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

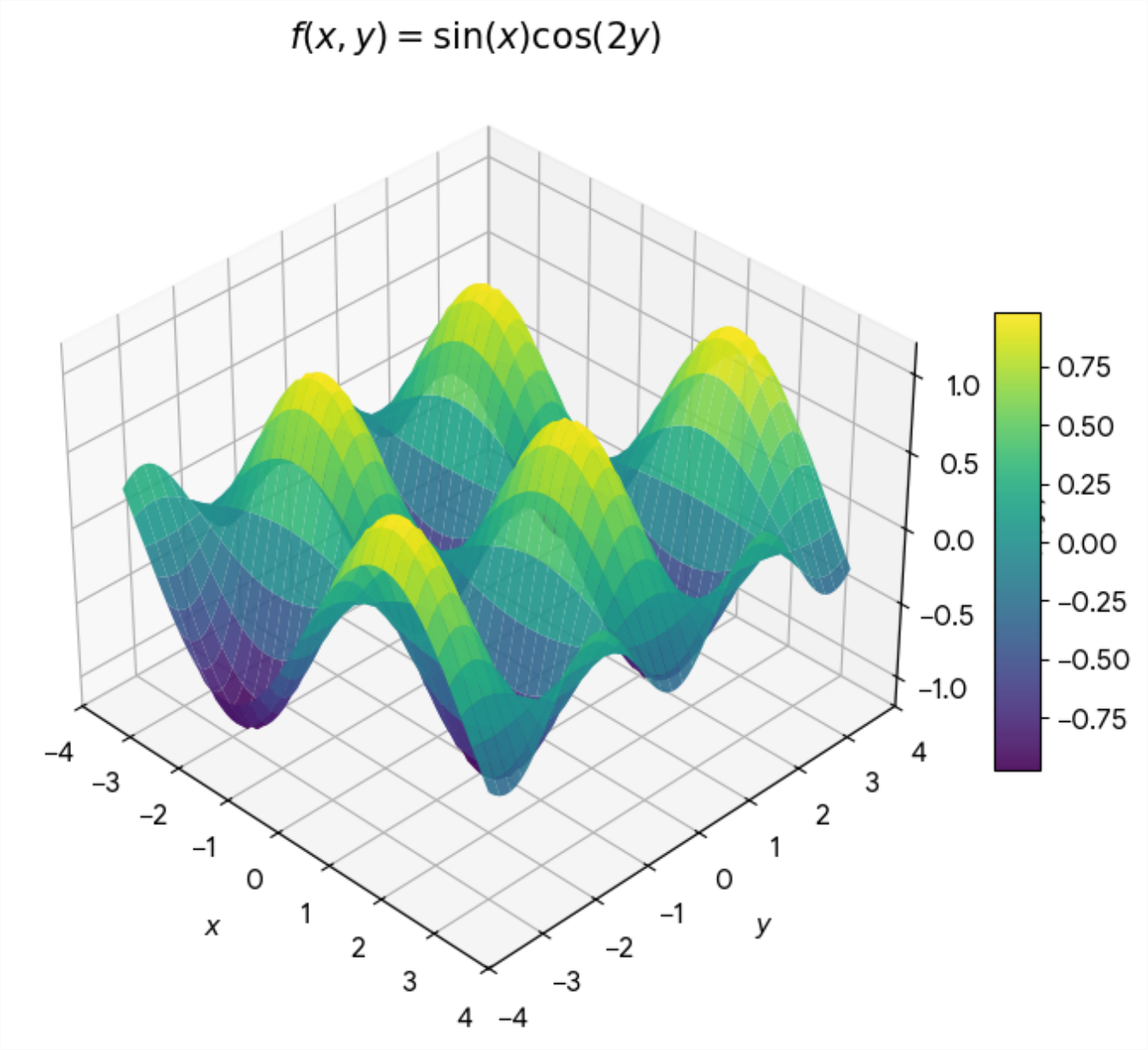
  for some (known) Lipschitz constant $L > 0$.
- **Notation:** $f \in \mathcal{F}_L(\mathbb{R}^n)$.

Exercise:

Prove the following 'descent lemma' for $f \in \mathcal{F}_L(\mathbb{R}^n)$:

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

Example: a nonconvex, L-smooth function



Trigonometric polynomials are L-smooth

Trigonometric polynomial

A function $f : \mathbb{R} \to \mathbb{R}$ of the form

$$f(x) = \sum_{k=1}^{d} \left(a_k \sin(kx) + b_k \cos(kx) \right),$$

with $a_k, b_k \in \mathbb{R}$ is called a (real) trigonometric polynomial of degree d.

Exercise:

Let $f_i : \mathbb{R} \to \mathbb{R}$ $(i \in [m])$ be trigonometric polynomials with degree $(f_i) = d_i$. Show that $f(\mathbf{x}) := \prod_{i=1}^{m} f_i(x_i)$ is L-smooth. Give an upper bound the value of L in terms of the degrees and coefficients of the f_i.

Known results on convergence rate to stationary point

- (Classical, e.g. Nesterov textbook) When $t_k = \frac{1}{L}$, $k \in \{0, ..., N-1\}$:

$$\min_{0 \le k \le N} \|\nabla f(\mathbf{x}_k)\|^2 \le \left(\frac{2L(f(\mathbf{x}_0) - f_\star)}{N} \right).$$

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

$$\min_{0 \le k \le N} \|\nabla f(\mathbf{x}_k)\|^2 \le \left(\frac{4(f(\mathbf{x}_0) - f_\star)}{\sum_{k=0}^{N-1} t_k(4 - Lt_k)} \right).$$

Reference [Drori-Shamir (2020)]

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

$$\max \left(\min_{0 \le k \le N} \|\nabla f(\mathbf{x}_k)\| \right)$$

s.t. $f(\mathbf{x}_0) - f_\star \le \Delta$

$\mathbf{x}_N, ..., \mathbf{x}_1$ are generated by gradient descent w.r.t. $f, \mathbf{x}_0$

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

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

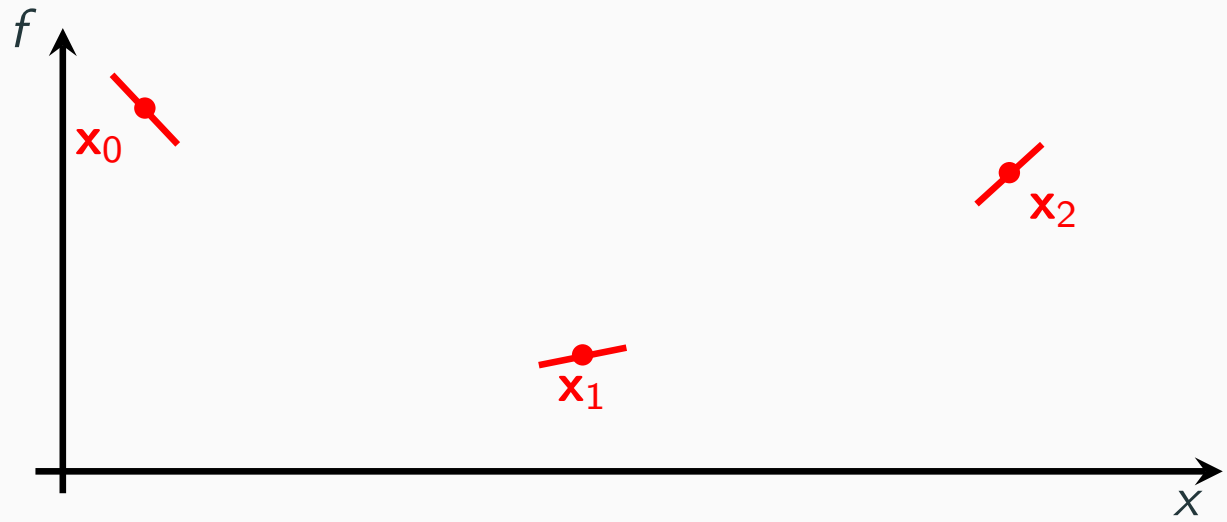
$\mathbf{x}_0 \in \mathbb{R}^n,$

with variables f and $\mathbf{x}_0$, and given parameters $\Delta > 0, L > 0, N > 0$ and the step lengths $t_k > 0$.

Same idea as before: 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 S, and given values $\{(\mathbf{x}_i, \mathbf{g}_i, f_i)\}_{i \in I}$ 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 $\{(\mathbf{x}_i, \mathbf{g}_i, f_i)\}_{i \in S}$ is $\mathcal{F}_L(\mathbb{R}^n)$-interpolable.

L-smooth Interpolation (ctd.)

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

The following statements are equivalent:

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

$$\tfrac{1}{2L} \|\mathbf{g}_i - \mathbf{g}_j\|^2 - \tfrac{L}{4} \left\|\mathbf{x}_i - \mathbf{x}_j - \tfrac{1}{L}(\mathbf{g}_i - \mathbf{g}_j)\right\|^2 \leq f_i - f_j - \langle \mathbf{g}_j, \mathbf{x}_i - \mathbf{x}_j \rangle .$$

Reference:

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)

SDP formulation

$$
\begin{aligned}
&\max \left(\min_{0 \le k \le N} \|\mathbf{g}_k\|^2 \right) \\
&\text{s.t. } \tfrac{1}{2L} \|\mathbf{g}_i - \mathbf{g}_j\|^2 - \tfrac{L}{4} \left\|\mathbf{x}_i - \mathbf{x}_j - \tfrac{1}{L}(\mathbf{g}_i - \mathbf{g}_j)\right\|^2 \le f_i - f_j - \\
&\qquad \langle \mathbf{g}_j, \mathbf{x}_i - \mathbf{x}_j \rangle \quad i,j \in S := \{0, ..., N, \star\} \\
&\quad \mathbf{x}_{k+1} = \mathbf{x}_k - t_k \mathbf{g}_k \quad k \in \{0, ..., N\} \\
&\quad f_k \ge f_\star + \tfrac{1}{2L} \|\mathbf{g}_k\|^2 \quad k \in \{0, ..., N\} \\
&\quad f_0 - f_\star \le \Delta \\
&\quad \mathbf{g}_\star = 0,
\end{aligned}
$$

with variables $\mathbf{x}_k$, f_k, $\mathbf{g}_k$ $(0 \le k \le N)$, and given parameters L, Δ, N and t_k, $(0 \le k \le N-1)$.

SDP formulation (ctd.)

- We may eliminate $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$ using $\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \mathbf{g}_k$.
- We may also assume w.l.o.g. that $\mathbf{x}_\star = \mathbf{0}$ and $f_\star = 0$ as before.
- Then the objective and constraints become linear functions of the $(N+2) \times (N+2)$ Gram matrix:

$$G := \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{x}_0 \rangle & \langle \mathbf{x}_0, \mathbf{g}_0 \rangle & \cdots & \langle \mathbf{x}_0, \mathbf{g}_N \rangle \\ \langle \mathbf{g}_0, \mathbf{x}_0 \rangle & \langle \mathbf{g}_0, \mathbf{g}_1 \rangle & \cdots & \langle \mathbf{g}_0, \mathbf{g}_N \rangle \\ \vdots & & \ddots & \vdots \\ \langle \mathbf{g}_N, \mathbf{x}_0 \rangle & \langle \mathbf{g}_N, \mathbf{g}_1 \rangle & \cdots & \langle \mathbf{g}_N, \mathbf{g}_N \rangle \end{pmatrix}$$

  and the scalar variables $f_0, \ldots, f_N$.
- If we view the matrix G as an (unknown) postive semidefinite matrix variable, then we obtain an SDP problem.

SDP formulation exercise

Exercise:

Write the SDP problem explicitly for the case $N = 1$ and step length $1/L$.

1. What is the size of the SDP matrix variable here?
2. For which values of n will the bound be tight? (rank condition)
3. How many linear constraints are there?
4. What does the dual SDP problem look like?

Performance estimation results

Tight convergence rate

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 $\mathbf{x}_0$. Then

$$\min_{0 \leq k \leq N} \|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{4(f(\mathbf{x}_0) - f_\star)}{\sum_{k=0}^{N-1} \min\{-L^2 t_k^3 + 4t_k, -Lt_k^2 + 4t_k\} + \frac{2}{L}},$$

and this bound is tight in some cases.

Corollary 1

Let $t_k = \frac{1}{L}$ for $k \in \{0, ..., N-1\}$. Then

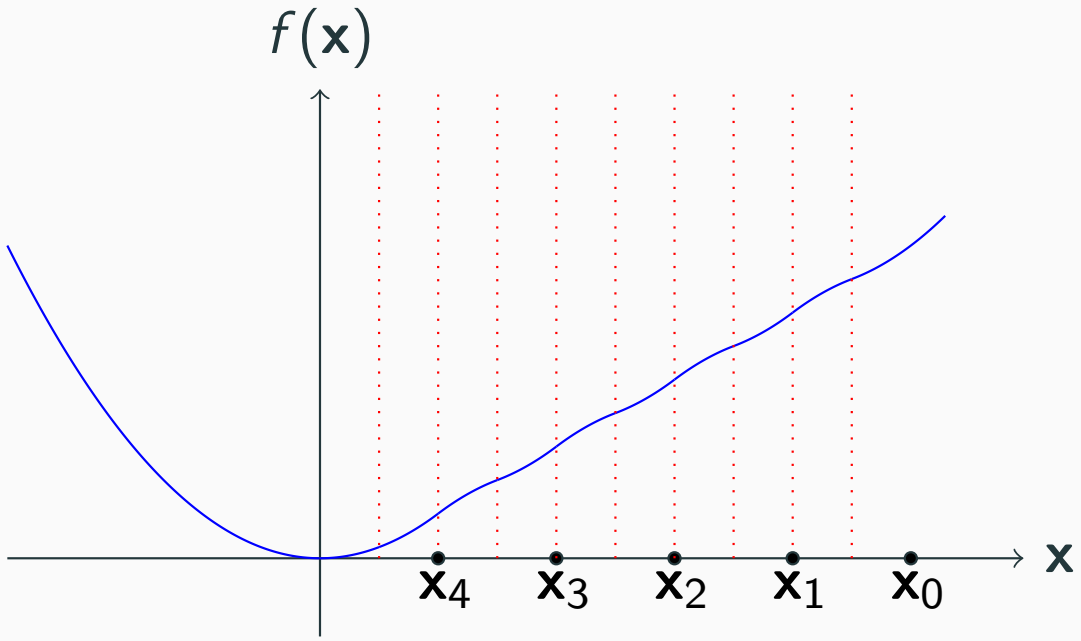
$$\min_{0 \leq k \leq N} \|\nabla f(\mathbf{x}_k)\|^2 \leq \left(\frac{4L(f(\mathbf{x}_0) - f_\star)}{3N+2}\right).$$

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 $\mathbf{x}_\star = 0$, that attain the bound in the corollary for suitable $\mathbf{x}_0$, 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

$$t_k = \frac{\sqrt{\frac{4}{3}}}{L} \quad \forall k \in \{0, \ldots, N-1\},$$

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

The proof is by minimizing the right-hand-side in the theorem over the t_k. Substituting this step length leads to

$$\min_{0 \leq k \leq N} \|\nabla f(\mathbf{x}_k)\| \leq \left(\frac{6\sqrt{3}L(f(\mathbf{x}_0) - f_\star)}{8N + 3\sqrt{3}} \right)^{\frac{1}{2}}$$

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

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

Exercises:

1. Under the PŁ inequality assumption, every stationary point of f is a global minimizer.
2. PŁ is weaker than strong convexity.

The Polyak-Łojasiewicz (PŁ) inequality: exercise

Exercise:

Consider a nonlinear least squares (regression) problem:

$$f_\star = f(\mathbf{x}_\star) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \sum_{i=1}^{m} r_i^2(\mathbf{x}).$$

The $m \times n$ Jacobian matrix ($m \leq n$) associated with the r_i's:

$$J(\mathbf{x}) = \begin{pmatrix} \frac{\partial r_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial r_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_m(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial r_m(\mathbf{x})}{\partial x_n} \end{pmatrix}.$$

Assume $f_\star = 0$, and that $\lambda_{\min}\left(J(\mathbf{x})J(\mathbf{x})^\top\right) \geq \sigma > 0$ on a open set X that contains $\mathbf{x}_\star$. Show that f satisfies the PŁ inequality on X with constant $\mu_p = 2\sigma$.

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 $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$. If $t_0 = \frac{1}{L}$, we have

$$f(\mathbf{x}_1) - f_\star \leq \frac{L - \mu_p}{L} \left(f(\mathbf{x}_0) - f_\star\right).$$

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 $X = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$. If $t_0 = \frac{1}{L}$, we have

$$f(\mathbf{x}_1) - f_\star \leq \left(\frac{L - \mu_p}{L + \frac{1}{2}\mu_p}\right)(f(\mathbf{x}_0) - f_\star).$$

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 $\{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$.

Remarks

Discussion

The sharpest results for gradient descent for L-smooth f are from:

H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani. *Optimization Letters*, **16**, 1649–1661 (2022), ...

... and with PŁ added:

H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani. Conditions for linear convergence of the gradient method for non-convex optimization. *Optimization Letters* **17**, 1105–1125 (2023).

SDP performance estimation may also be used to analyse the stochastic gradient method [Drori and Shamir (2020)]; next ...

The stochastic gradient method

The stochastic gradient descent method

Input: $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}_0 \in \mathbb{R}^n$, number of steps N and $\{t_k\}_{k=1}^N$ (step lengths).

for $k = 1, \ldots, N$

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

where the ξ^k are independent random variables with

- $\mathbb{E}[\xi^k] = 0$ for all k,
- $\sum_{i=1}^n \mathrm{Var}[\xi_i^k] = \mathbb{E}[\langle \xi^k, \xi^k \rangle] \leq \sigma^2$ for all k and some given $\sigma > 0$.

NB: the expectation refers to the joint distribution of the random variables $\mathbf{x}_k$, $\nabla f(\mathbf{x}_k)$, ξ^k after N steps.

The stochastic gradient descent method: performance

Theorem

Let $t_k = \frac{1}{L}$ for $k \in \{0, ..., N-1\}$. Then

$$\min_{0 \leq k \leq N} \mathbb{E}\left[\|\nabla f(\mathbf{x}_k)\|^2\right] \leq \left(\frac{2L(f(\mathbf{x}_0) - f_\star)}{N}\right) + \sigma^2.$$

More general results in:

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23(4) (2013), pp. 2341–2368.

One may again formulate tighter bounds using SDP ... next slides.

SDP formulation ... first try

$$\max \left(\min_{0 \le k \le N} \mathbb{E}\left[\|\mathbf{g}_k\|^2 \right] \right)$$

$$\text{s.t. } \frac{1}{2L} \|\mathbf{g}_i - \mathbf{g}_j\|^2 - \frac{L}{4} \left\| \mathbf{x}_i - \mathbf{x}_j - \frac{1}{L}(\mathbf{g}_i - \mathbf{g}_j) \right\|^2 \le f_i - f_j - \langle \mathbf{g}_j, \mathbf{x}_i - \mathbf{x}_j \rangle \quad i,j \in \{0,...,N,\star\}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k(\mathbf{g}_k + \xi^k) \quad k \in \{0,...,N\}$$

$$f_k \ge f_\star + \frac{1}{2L}\|\mathbf{g}_k\|^2 \quad k \in \{0,...,N\}$$

$$f_0 - f_\star \le \Delta$$

$$\mathbf{g}_\star = 0,$$

with random variables $\mathbf{x}_k$, ξ^k, f_k, $\mathbf{g}_k$ $(0 \le k \le N)$, and given parameters L, Δ, N and t_k, $(0 \le k \le N-1)$.

What is the problem here?

SDP formulation (ctd.)

- We may eliminate $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$ using $\mathbf{x}_{k+1} = \mathbf{x}_k - t_k(\mathbf{g}_k + \xi^k)$.
- Replace each constraint by its expected value.
- Then the objective and constraints become linear functions of $\mathbb{E}[\langle \mathbf{x}_0, \mathbf{g}_k \rangle]$, $\mathbb{E}[\langle \mathbf{g}_i, \mathbf{g}_j \rangle]$, $\mathbb{E}[\langle \xi^i, \xi^j \rangle]$, etc.
- We still need to add constraints $\mathbb{E}[\langle \xi^i, \xi^j \rangle] = 0$ if $i \neq j$ (independence), and $\mathbb{E}[\langle \xi^k, \xi^k \rangle] \leq \sigma^2$, etc.
- Once again, we obtain an SDP problem (why?)
- The SDP gives improved bounds for specific choices of N, the step lengths, and other parameters like σ. [Drori-Shamir (2020)]

SDP formulation exercise

Exercise:

Write the SDP problem explicitly for the case $N = 2$ and $t_k = 1/L$ for $k \in \{1, 2\}$.

1. What is the size of the SDP matrix variable here?
2. How many linear constraints are there?
3. What happens when $\sigma = 0$?

Computer session

Computer session

- Use Chrome to open (or scan the QR code below): `https://github.com/PerformanceEstimation/Tutorial-Europt-2026`
- Click on Part 2 in left-hand-side menu.
- Click on `Europt2026_Tutorial_Part2.ipynb`
- Click on Open in Colab