lions@epfl

EE-556: MATHEMATICS OF DATA: FROM THEORY TO COMPUTATION
LABORATORY FOR INFORMATION AND INFERENCE SYSTEMS
FALL 2025

EPFL

INSTRUCTOR: PROF. VOLKAN CEVHER

HEAD TAS: IOANNIS MAVROTHALASSITIS, ZHENYU ZHU

HANDOUT 2

In this handout, we will go through pen and paper exercises to get familiar with fundamental concepts that will be used throughout the course. We will take a look at convergence rates and smooth functions.

1 Interpreting convergence rates

Throughout the course, we will often compare the performance of different methods by looking at convergence plots of different methods. It means that it is very important to be familiar with reading and drawing convergence rates. In the course, you will be mostly confronted with sublinear, linear and quadratic rates of convergence. Assume that you are given a sequence of iterates $(\mathbf{x}_k) \in \mathbb{R}^p$, converging towards a vector $\mathbf{x}^*$.

- Then, the sequence $(\mathbf{x}_k)$ is said to converge **sublinearly** to $\mathbf{x}^*$ if
$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 1$$
- The sequence $(\mathbf{x}_k)$ is said to converge **linearly** to $\mathbf{x}^*$ if, for some $c \in (0, 1)$,
$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = c$$
- The sequence $(\mathbf{x}_k)$ is said to converge **superlinearly** to $\mathbf{x}^*$ if,
$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 0$$
  - The definition above can be refined by defining the *order of convergence*. The sequence $(\mathbf{x}_k)$ is said to converge **with order** $q > 1$ to $\mathbf{x}^*$ if
  $$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|^q} < c$$
  for some $c > 0$, not necessarily smaller than 1. In particular, with $q = 2$, we have **quadratic** convergence.

Problem 1: Convergence rate of different sequences.

For each of the following sequences, find the limit x^* and the the convergence rate.

$$\text{i) } x_k = \frac{1}{k+1} \qquad \text{ii) } x_k = \frac{5k + \log(k)}{3k + 6} \qquad \text{iii) } x_k = \frac{3}{2}\exp(-k/4) \qquad \text{iv) } x_k = \frac{1}{(3k)^2} \qquad \text{v) } x_k = \frac{1}{3^{2^k}}$$

Solution

i) We see immediately that $x^* = \lim_{k\to\infty} \frac{1}{k+1} = 0$. The sequence converges **sublinearly** because
$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \lim_{k\to\infty} \frac{\frac{1}{k+2}}{\frac{1}{k+1}} = \lim_{k\to\infty} \frac{k+1}{k+2} = 1$$

ii) We first compute $x^* = \lim_{k\to\infty} \frac{5k+\log(k)}{3k+6} = \frac{5}{3}$. This is because $\lim_{k\to\infty} \frac{\log(k)}{3k+6} = 0$ through L'Hôpital's rule. The sequences converges **sublinearly** because
$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \lim_{k\to\infty} \frac{\frac{5(k+1)+\log(k)}{3(k+1)+6} - \frac{5}{3}}{\frac{5k+\log(k)}{3k+6} - \frac{5}{3}} = \lim_{k\to\infty} \frac{5(k+1) + \log(k) - 5(k+3)}{3k+9} \frac{3k+6}{5k + \log(k) - 5(k+2)} = \lim_{k\to\infty} \frac{3k+6}{3k+9} = 1$$

iii) We immediately see that $x^* = \lim_{k\to\infty} \frac{3}{2}\exp(-k/4) = 0$. The sequences converges **linearly** because

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \lim_{k\to\infty} \frac{\frac{3}{2}\exp(-(k+1)/4)}{\frac{3}{2}\exp(-k/4)} = \lim_{k\to\infty} \exp\left(-\frac{1}{4}(k+1-k)\right) = \exp\left(-\frac{1}{4}\right) < 1$$

iv) We immediately see that $x^* = \lim_{k\to\infty} \frac{1}{(3k)^2} = 0$. The sequence converges **sublinearly** because

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \lim_{k\to\infty} \frac{\frac{1}{(3k+1)^2}}{\frac{1}{(3k)^2}} = \lim_{k\to\infty} \frac{9k^2}{9k^2+6k+1} = 1$$

iv) We immediately see that $x^* = \lim_{k\to\infty} \frac{1}{(3k)^2} = 0$. The sequence converges **sublinearly** because

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \lim_{k\to\infty} \frac{\frac{1}{(3k+1)^2}}{\frac{1}{(3k)^2}} = \lim_{k\to\infty} \frac{9k^2}{9k^2+6k+1} = 1$$

v) We immediately see that $x^* = \lim_{k\to\infty} \frac{1}{3^{2^k}} = 0$ (as $2^k \to +\infty$ as $k \to \infty$). The sequence converges **superlinearly** because

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \lim_{k\to\infty} \frac{\frac{1}{3^{2^{k+1}}}}{\frac{1}{3^{2^k}}} = \lim_{k\to\infty} 3^{-2^{k+1}+2^k} = \lim_{k\to\infty} 3^{2^k(-2+1)} = \lim_{k\to\infty} 3^{-2^k} = 0$$

However, the convergence can be characterized more precisely. It is in fact **quadratic** as

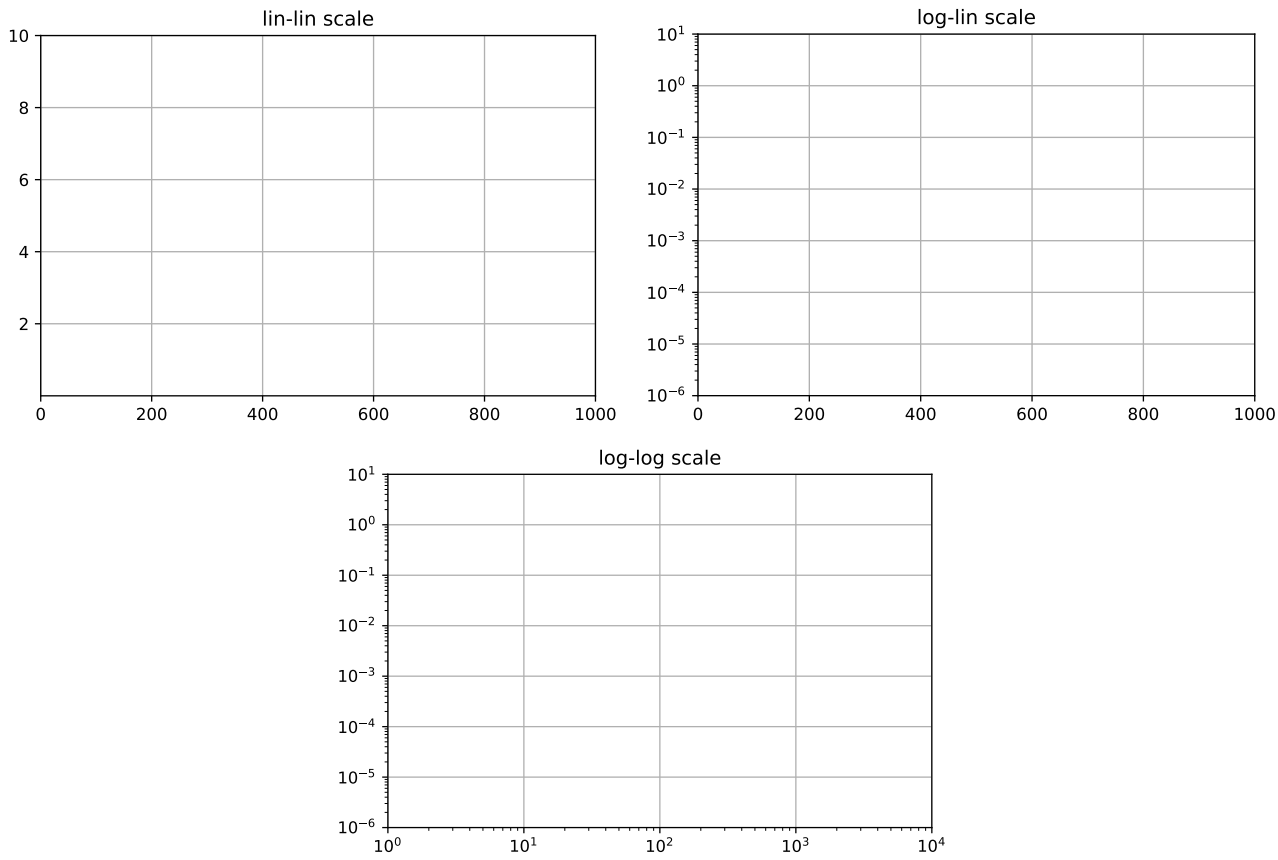
$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|^2} = \lim_{k\to\infty} \frac{\frac{1}{3^{2^{k+1}}}}{\frac{1}{3^{2\cdot 2^k}}} = \lim_{k\to\infty} 3^{-2^{k+1}+2^{k+1}} = 1$$

Problem 2: Drawing convergence rates

Draw the asymptotic rate of convergence $\|\mathbf{x}_k - \mathbf{x}^*\|$ for the following sequences on either of the graphs that feature different scales (lin-lin, log-lin, log-log).

$$\text{i)}\ \frac{1}{k+1} \qquad \text{ii)}\ \frac{1}{k^3+4} \qquad \text{iii)}\ \frac{3}{2}\exp(-k/4) \qquad \text{iv)}\ \frac{1}{(3k)^2}$$

Note. Each of the sequences can be naturally drawn on one of the scales. By natural, we mean that on some scale, the asymptotic behaviour of the sequence will be displayed as a line. For instance, a line in a log-log plot means that $\log(\|\mathbf{x}_k - \mathbf{x}^*\|)$ is a *linear* function of $\log(k)$.



Solution

i) Starting with $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{1}{k+1}$, we notice that $\log\left(\frac{1}{k+1}\right) = -\log(k+1) \approx -\log(k)$ as $k \to \infty$. So this means that taking the $\log(\|\mathbf{x}_k - \mathbf{x}^*\|) \lessapprox -\log(k)$. This means that taking the log of the precision, we are upper bounded by a function of the log of the iterates. This makes for a natural representation on a log-log plot, where our bound will be a line of slope -1.

ii) Starting with $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{1}{k^3+4}$, we notice that $\log\left(\frac{1}{k^3+4}\right) = -\log(k^3+4) \approx -\log(k^3)$ as $k \to \infty$. So this means that taking the $\log(\|\mathbf{x}_k - \mathbf{x}^*\|) \lessapprox -3\log(k)$. This means that taking the log of the precision, we are upper bounded by a function of the log of the iterates. This makes for a natural representation on a log-log plot, where our bound will be a line of slope -3.

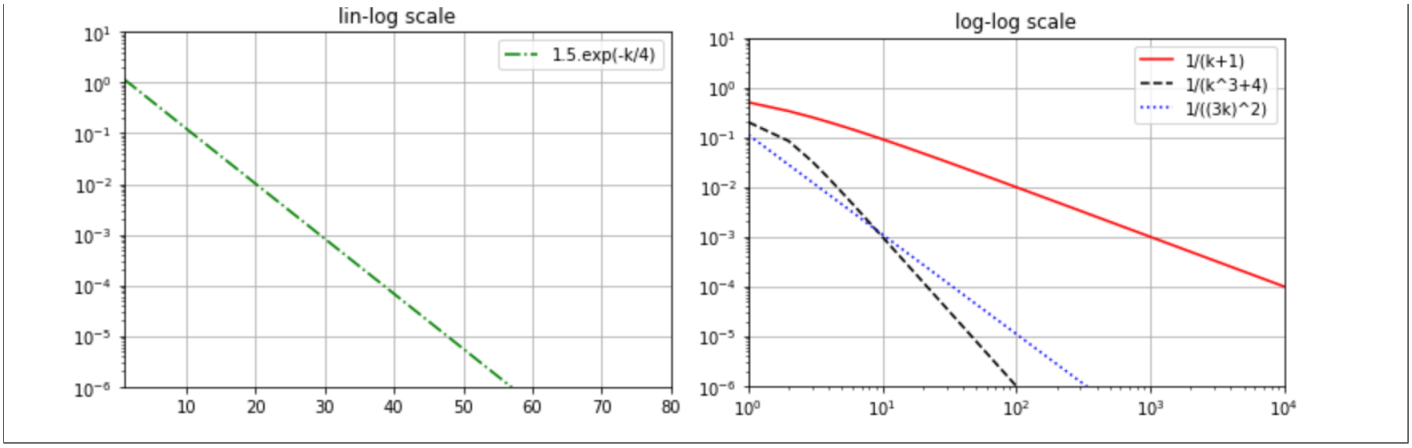
iii) Starting with $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{3}{2}\exp(-k/4)$, we notice that $\log\left(\frac{3}{2}\exp(-k/4)\right)$. Here, we come across our first problem as the basis usually used for a log plot is the basis 10, so we will need to change our basis. We have

$$\log\left(\frac{3}{2}\exp(-k/4)\right) = \log\left(\frac{3}{2}\right) - \log_{10}\left(\exp(-k/4)\right) = \log\left(\frac{3}{2}\right) + \log_{10}\left(10^{-\log_{10}(e)k/4}\right) = \log\left(\frac{3}{2}\right) - \frac{\log_{10}(e)}{4}k.$$

This means that taking the $\log(\|\mathbf{x}_k - \mathbf{x}^*\|) \leq \log\left(\frac{3}{2}\right) - \frac{\log_{10}(e)}{4}k$. This means that taking the log of the precision, we are upper bounded by a linear function of the of the iterates. This makes for a natural representation on a log-lin plot, where our bound will be a line of slope $-\frac{\log_{10}(e)}{4} \approx -0.15$ with the ordinate at the origin being $\log\left(\frac{3}{2}\right)$.

iv) Starting with $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{1}{(3k)^2}$, we notice that $\log\left(\frac{1}{(3k)^2}\right) = -\log((3k)^2) = \log(9) - 2\log(k)$ as $k \to \infty$. So this means that taking the $\log(\|\mathbf{x}_k - \mathbf{x}^*\|) \leq \log(9) - 2\log(k)$. This means that taking the log of the precision, we are upper bounded by a function of the log of the iterates. This makes for a natural representation on a log-log plot, where our bound will be a line of slope -2 with the ordinate at 1 being $\log(9)$.

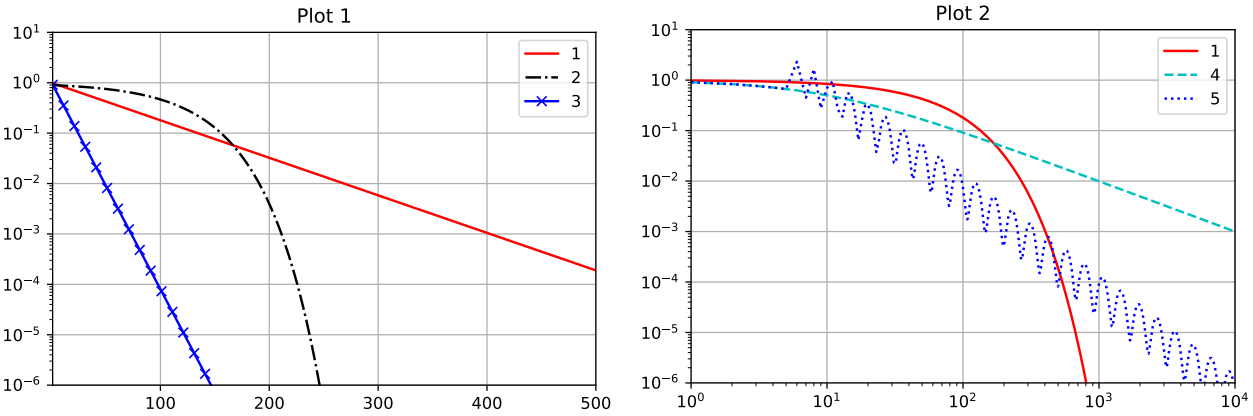
All the sequences are displayed below.



Problem 3: Reading convergence plots

On plots 1 and 2 below, the convergence rates of 5 methods are displayed (method 1 is displayed on both plots).

1. Characterize the rate of convergence (sublinear, linear, or quadratic) for each of the methods. Justify your answer.
2. Establish more precisely the order of convergence of methods 1, 3, 4 and 5 by reading the plots.
   Hint. Find the slopes of the different lines and map, and use the scale of the plot to write the rate of convergence of the method.
3. Rank methods 1 to 5 from the slowest to the fastest **asymptotic** rate of convergence, using the fact that method 1 is displayed on both plots.



Solution

From the previous problem, we've established that on a log-lin plot, linearly-converging sequences will be displayed as lines, and that on a log-log plot, sublinearly-converging sequences will be represented as a line. We've also seen how to relate the slope of the line to a typical function: a log-lin plot with a line of slope $-q$ will correspond to a convergence rate of the order $c10^{-qk}$. Similarly, a log-log plot with a line of slope $-q$ will correspond to a convergence rate of the order $c_1 \frac{1}{(k+c_2)^q}$.

With this information, we see on plot 1 that the method 1 has a slope $-q \approx -\frac{3}{400}$, as it descends 3 order of magnitudes in 400 iterations, its convergence rate is then $c10^{-\frac{3k}{400}}$.. On plot 1 as well, we see that method 2 converges **superlinearly**, as it is not displayed as a line on the log-lin plot. Similarly, we see that method 3 has a slope $-q \approx -\frac{4}{100}$, its convergence rate is then $c10^{-\frac{4k}{100}}$.

Then, reading from plot 2, we can see that method 4 has a slope $-q = -1$, as it descends 1 order of magnitude and advances 1 order of magnitude at once. Its rate is then of the form $c_1 \frac{1}{(k+c_2)}$. Finally, even though method 5 oscillates, one can see that there is an average slope of $-q = -2$ and so its convergence rate is of the form $c_1' \frac{1}{(k+c_2')^2}$

Based on the convergence rates and on method 1 represented on both plots, we can sort the asymptotic speeds of convergence of the methods as follows:

$$4 < 5 < 1 < 3 < 2$$

Problem 4: Convergence in accuracy against convergence in iterations

Up to now, we have considered convergence in **iteration**, as a function of k. However, it is common to view the convergence as a function of the time require to reach a given accuracy ϵ. If we know that the $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{1}{k+1}$, the convergence in ϵ tries to characterize the order of convergence as a function of the desired accuracy instead of the number of iterations. In practice, this amounts to find $K(\epsilon)$ such that $\forall k \in \mathbb{N}$, $k \geq K(\epsilon) \Rightarrow \|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \epsilon$.

Given the convergence rate $\|\mathbf{x}_k - \mathbf{x}^*\|$ of a sequence, express the number of iterations required to reach a accuracy ϵ for

$$\text{i) } \frac{1}{k+1} \qquad \text{ii) } \frac{1}{k^3+4} \qquad \text{iii) } \frac{3}{2}\exp(-k/4) \qquad \text{iv) } \frac{1}{(3k)^2} \qquad \text{v) } \frac{1}{3^{2^k}} \qquad \text{vi) } \frac{4}{\sqrt{k+3}}$$

Solution

Convergence in iterations and in accuracy are simply inversely related to each other. We can generally proceed as follows: Start from the fact that for some f, $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{1}{k+1} \leq f(k) \leq \epsilon \; \forall k \in \mathbb{N}$ and for some ϵ. Then, invert the condition $f(k) \leq \epsilon$ to find $k \geq K(\epsilon)$ with $K(\cdot) = f^{-1}(\cdot)$ (the sign reverts because k and ϵ are inversely related to each other). We then have

i) $K(\epsilon) = \frac{1}{\epsilon} - 1 = O\left(\epsilon^{-1}\right)$

ii) $K(\epsilon) = \sqrt[3]{\frac{1}{\epsilon} - 4} = O\left(\epsilon^{-1/3}\right)$

iii) $K(\epsilon) = \frac{1}{4}\log\left(\frac{3}{2\epsilon}\right) = O\left(\log \epsilon^{-1}\right)$

iv) $K(\epsilon) = \frac{1}{3}\sqrt{\frac{1}{\epsilon}} = O\left(\epsilon^{-1/2}\right)$

v) $K(\epsilon) = \log_2(\log_3(\epsilon)) = O\left(\log(\log(\epsilon))\right)$

vi) $K(\epsilon) = \left(\frac{4}{\epsilon}\right)^2 - 3 = O\left(\epsilon^{-2}\right)$

2 Smooth functions

Throughout the course, we will frequently encounter L-smooth functions.

Definition. A function $f : Q \to \mathbb{R}$ is said to be L-smooth with respect to a pair of dual norms $(\|\cdot\|, \|\cdot\|_*)$ if there exists some $L > 0$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in Q. \tag{1}$$

The Lipschitz constant of the the gradient L, also called the smoothness constant, can be computed in several ways, and we will explore different ways to obtain it in the following exercises.

Problem 5: Lipschitz gradient in the one-dimensional case

In a single dimensional case, we have a function $f : Q \subseteq \mathbb{R} \to \mathbb{R}$. The equation (1) can be restated as

$$|f'(x) - f'(y)| \leq L|x - y| \quad \forall \mathbf{x}, \mathbf{y} \in Q.$$

Prove that the smoothness constant L can be computed as the maximum of the absolute value of the second derivative, i.e. $L = \max_{z \in Q} |f''(z)|$.

HINT. Use the mean value theorem.

REMARK. This statement can be extended to higher dimensional cases, but one needs to be careful to appropriately define the norms that will be used.

Solution

Recall the mean-value theorem, expressed directly for f'.

Theorem 2.1. *if f' is a continuous function on the closed interval $[x, y]$ and differentiable on the open interval (x, y), then there exists a point*

z in (x, y) such that

$$\frac{f'(x) - f'(y)}{x - y} = f''(z)$$

We can note immediately that taking the absolute value on both sides, we have

$$\left|\frac{f'(x) - f'(y)}{x - y}\right| = \frac{|f'(x) - f'(y)|}{|x - y|} = |f''(z)|.$$

Now starting from the definition of the Lipschitz-gradient, we notice that the case $|x - y| = 0$ is trivial, as any L will satisfy the requirement. We can then restrict ourselves to the case $|x - y| \neq 0$. Dividing on both sides by $|x - y|$, we have

$$\frac{|f'(x) - f'(y)|}{|x - y|} \leq L \quad \forall \mathbf{x}, \mathbf{y} \in Q.$$

From the mean-value theorem, there exists a $z \in [x, y]$ such that

$$\frac{|f'(x) - f'(y)|}{|x - y|} = |f''(z)|$$

Looking now at the right hand side, for $z \in [x, y]$, we have $|f''(z)| \leq \max_{z \in [x,y]} |f''(z)| \leq \max_{z \in Q} |f''(z)|$. Setting $L = \max_{z \in Q} |f''(z)|$, we obtain the desired statement.

Problem 6: Lipschitz gradient in the quadratic case

We now move to a multidimensional case, where we have $f : \mathbb{R}^p \to \mathbb{R}$ defined as $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x}$, where A is a symmetric matrix. We will explore a different way to compute the Lipschitz constant of the gradient in this setting.
Given a pair of dual norms $(\|\cdot\|_p, \|\cdot\|_q)$ with $\frac{1}{p} + \frac{1}{q} = 1$, prove that when

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_q \leq L\|\mathbf{x} - \mathbf{y}\|_p \quad \forall \mathbf{x}, \mathbf{y} \in Q,$$

then $L = \|A\|_{p \to q}$.

HINT. Recall the definition of the operator norm from the lecture.

$$\|A\|_{p \to q} := \sup_{\mathbf{x}: \|\mathbf{x}\|_p \leq 1} \|A\mathbf{x}\|_q$$

Solution

First of all, we need to compute $\nabla f(\mathbf{x}) = A\mathbf{x}$. We start from the definition of the Lipschitz gradient (1), provided the given pair of dual norms, we have

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_q \leq L\|\mathbf{x} - \mathbf{y}\|_p \quad \forall \mathbf{x}, \mathbf{y} \in Q.$$

Substituting the gradient of f, we have $\|A(\mathbf{x}-\mathbf{y})\|_q \leq L\|\mathbf{x}-\mathbf{y}\|_p$. Setting $\mathbf{z} = \mathbf{x}-\mathbf{y}$, and considering the case $\mathbf{z} \neq 0$ (the case $\mathbf{z} = 0$ leads to a trivial setting where any L can be chosen), we can divide by $\|\mathbf{z}\|_p$ on both sides. Then, we can upper bound the left-hand side by maximizing over $\mathbf{z}$. Noting that the formulation is equivalent to the definition of the operator norm, we can identify the RHS expression with L for the desired result:

$$\frac{\|A\mathbf{z}\|_q}{\|\mathbf{z}\|_p} \leq \sup_{\mathbf{z}':\mathbf{z}' \neq \mathbf{0}} \frac{\|A\mathbf{z}'\|_q}{\|\mathbf{z}'\|_p} = \sup_{\substack{\mathbf{z}':\mathbf{z}' \neq \mathbf{0} \\ \|\mathbf{z}'\|_p \leq 1}} \|A\mathbf{z}'\|_q =: \|A\|_{p \to q} =: L$$

Problem 7: Operator norms in action

1. Given $A \in \mathbb{R}^{m \times n}$ and $a_i^\top$ the i-th row of A, prove that the operator norm $\|\mathbf{A}\|_{1 \to \infty} = \max_{i \in \{1,\dots,m\}} \|a_i\|_\infty$.

2. Consider the matrix

$$A = \begin{bmatrix} 2 & -\frac{1}{\sqrt{2}} & -1 \\ -\frac{1}{\sqrt{2}} & 3 & -\frac{1}{\sqrt{2}} \\ -1 & -\frac{1}{\sqrt{2}} & 2 \end{bmatrix}.$$

Compute the Lipschitz constant of the gradient of $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$ in the following settings.

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_\infty \leq L\|\mathbf{x} - \mathbf{y}\|_1 \qquad \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$$

Are the values of L equal? How do you interpret the result?

Solution

1. Let $A \in \mathbb{R}^{n\times n}$ and $a_i^\top$ denote i-th row of A. Let us first prove that the norm $\|A\|_{1\to\infty} = \max_{i\in\{1,\dots,n\}} \|a_i\|_1$

$$\|Ax\|_\infty = \max_{i\in\{1,\dots,m\}} |\langle a_i, x\rangle| \leq \max_{i\in\{1,\dots,m\}} \|a_i\|_p \|x\|_q \leq \max_{i\in\{1,\dots,m\}} \|a_i\|_p,$$

where p is such that $\|\cdot\|_p$ is the dual norm of $\|\cdot\|_q$. When $q = 1$, we get $p = \infty$, which implies that we take the maximal ℓ_∞ norm of a row. This amounts to taking the maximum entry of the matrix.

Note. The proof for $\|A\|_{2\to 2}$ is a more involved, but we also provide it for completeness..

$$\begin{aligned}
\|\mathbf{A}\|_{2\to 2} = \sup_{\|\mathbf{x}\|_2\leq 1} \|\mathbf{A}\mathbf{x}\|_2 &= \sup_{\|\mathbf{x}\|_2\leq 1} \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\|_2 \quad \text{(using SVD of } \mathbf{A}\text{)} \\
&= \sup_{\|\mathbf{x}\|_2\leq 1} \|\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\|_2 \quad \text{(rotational invariance of } \|\cdot\|_2\text{)} \\
&= \sup_{\|\mathbf{z}\|_2\leq 1} \|\mathbf{\Sigma}\mathbf{z}\|_2 \quad \text{(letting } \mathbf{V}^T\mathbf{x} = \mathbf{z}\text{)} \\
&= \sup_{\|\mathbf{z}\|_2\leq 1} \sqrt{\sum_{i=1}^{\min(n,p)} \sigma_i^2 z_i^2} = \sigma_{\max} = \|\mathbf{A}\|
\end{aligned}$$

2. For our particular application, we directly see that $\|A\|_{1\to\infty} = 3$, and by taking a singular value decomposition of the matrix, we obtain $\|A\|_{2\to 2} = 4$. We see then that different norm can give us more or less flat landscapes. This can impact the speed of convergence of gradient methods, where the step size is given by $\frac{1}{L}$. A small L will mean that the method will be able to take larger steps.

Problem 8: The importance of choosing the smoothness norm

1. During the lectures we saw that the L-smoothness of a function f gives rise to local quadratic upper-bounds. The iterative minimization of these upper bounds recovers the well-known Gradient Descent (GD) method. As a warm-up, let us remind ourselves of the computation.

   Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L_2-smooth and recall from the lecture that this implies

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_2}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathbb{R}^d. \tag{2}$$

   Show that the minimizer in y of the right-hand side of (2) is

$$y^* = x - \frac{1}{L_2}\nabla f(x). \tag{3}$$

   Observe that setting $x = x_k$ and letting $x_{k+1} := y^*$ in (3) results precisely in the update rule of GD.

   Solution

   Note that the RHS is strongly-convex in y and thus has a unique minimum. Thee first order optimality condition gives us:

$$\nabla f(x) - L_2(x - y^*) = 0 \iff y^* = x - \frac{1}{L_2}\nabla f(x)$$

2. In point 1. we arrive at the GD update rule by considering the smoothness of f with respect to the Euclidean norm. However, smoothness may be considered with respect to arbitrary norms $\|\cdot\|_p$, and its general expression is given by

$$\|\nabla f(x) - \nabla f(y)\|_q \leq L_p\|x - y\|_p, \tag{4}$$

where $\|z\|_q := \max_{\|t\|_p \leq 1} \langle z, t \rangle$ is the dual norm of $\|\cdot\|_p$. As in the case of smoothness with respect to $\|\cdot\|_2$, smoothness with respect to $\|\cdot\|_p$ induces a local quadratic upper bound as follows:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_p}{2}\|x - y\|_p^2, \quad \forall x, y \in \mathbb{R}^d. \tag{5}$$

By iteratively minimizing the right-hand side of (5) and depending on the chosen p, one arrives at various non-Euclidean gradient methods. The choice of norm is important as it can result in asymptotically faster gradient methods than the traditional GD. An example can be found in the work of [3], who leveraged smoothness with respect to $\|\cdot\|_\infty$ to obtain superior convergence for the maximum s-t flow and maximum concurrent multicommodity flow problems.

In the following, we will guide you in discovering the update rule that emerges from considering smoothness in the ℓ_∞-norm. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex with L_∞-Lipschitz gradient $\|\nabla f(x) - \nabla f(y)\|_1 \leq L_\infty \|x - y\|_\infty$.

(a) Define

$$[x]^\# := \arg\max_{s \in \mathbb{R}^d} \left\{ \langle x, s \rangle - \frac{1}{2}\|s\|_\infty^2 \right\}. \tag{6}$$

Show that $\|x\|_1 \operatorname{sgn}(x) \in [x]^\#$, i.e. that it is a maximizer of the expression in (6).

Hint: You can use Hölder's inequality below to find an upper bound, then show that it is correspondingly attained.

$$|\langle x, y \rangle| \leq \|x\|_p \|y\|_q \qquad \forall p, q \in [1, \infty] \text{ s.t. } \frac{1}{p} + \frac{1}{q} = 1 \text{ (with the convention that } \frac{1}{\infty} = 0).$$

Solution

Note that if $x = 0$, then $\arg\max_{s \in \mathbb{R}^d} \left\{ \langle x, s \rangle - \frac{1}{2}\|s\|_\infty^2 \right\} = 0 = \|x\|_1 \operatorname{sgn}(x)$.
Otherwise, we proceed by first deriving an upper-bound using Hölder's inequality:

$$\begin{aligned} \langle x, s \rangle - \frac{1}{2}\|s\|_\infty^2 &\leq \|x\|_1 \|s\|_\infty - \frac{1}{2}\|s\|_\infty^2 \\ &= -\frac{1}{2}\left(\|s\|_\infty - \|x\|_1\right)^2 + \frac{1}{2}\|x\|_1^2 \\ &\leq \frac{1}{2}\|x\|_1^2 \qquad \forall s \in \mathbb{R}^d \end{aligned}$$

Next we show that, replacing s with $\|x\|_1 \operatorname{sgn}(x)$ attains the above upper bound:

$$\begin{aligned} \langle x, \|x\|_1 \operatorname{sgn}(x) \rangle - \frac{1}{2}\| \|x\|_1 \operatorname{sgn}(x) \|_\infty^2 &= \|x\|_1 \underbrace{\langle x, \operatorname{sgn}(x) \rangle}_{=\|x\|_1} - \frac{\|x\|_1^2}{2} \underbrace{\| \operatorname{sgn}(x) \|_\infty^2}_{=1} \\ &= \frac{\|x\|_1^2}{2} \end{aligned}$$

We conclude thus that $\|x\|_1 \operatorname{sgn}(x)$ is a maximizer of the expression in (6).

(b) Using inequality (5) adapted to the $\|\cdot\|_\infty$ norm, show that the minimizer in y of its right-hand side is given by

$$y^* = x - \frac{1}{L_\infty}\|\nabla f(x)\|_1 \operatorname{sgn}\left(\nabla f(x)\right).$$

Similar to point 1., observe how letting $x = x_k$ and $x_{k+1} := y^*$ gives us an update rule. This type of update pertains to the so-called SignGD method.

Hint: Write down the relevant arg min expression and then try to transform it equivalently such that the arg max formulation from (6) appears.

Remark: For those interested in doing further reading on the topic, [2] and [1] are good places to start.

Solution

The induced quadratic upper bound is:

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\infty}{2}\|x - y\|_\infty^2}_{=\text{RHS}}, \quad \forall x, y \in \mathbb{R}^d. \tag{7}$$

We now seek to find the $y^* = \arg\min_{y \in \mathbb{R}^d}\{\text{ RHS }\}$:

$$\begin{aligned}
y^* &= \arg\min_{y \in \mathbb{R}^d} \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\infty}{2}\|x - y\|_\infty^2 \right\} \\
&= \arg\min_{y \in \mathbb{R}^d} \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_\infty}{2}\|x - y\|_\infty^2 \right\} \\
&= \arg\max_{y \in \mathbb{R}^d} \left\{ -\langle \nabla f(x), y - x \rangle - \frac{L_\infty}{2}\|x - y\|_\infty^2 \right\} \\
&= \arg\max_{y \in \mathbb{R}^d} \left\{ \langle \nabla f(x), x - y \rangle - \frac{L_\infty}{2}\|x - y\|_\infty^2 \right\} \\
&= \arg\max_{y \in \mathbb{R}^d} \left\{ \left\langle \frac{1}{L_\infty} \nabla f(x), x - y \right\rangle - \frac{1}{2}\|x - y\|_\infty^2 \right\}
\end{aligned}$$

Note that, since the maximization is done over the entire space $\mathbb{R}^d$ and x is an arbitrary and fixed point, we have that:

$$\max_{y \in \mathbb{R}^d} \left\{ \left\langle \frac{1}{L_\infty} \nabla f(x), x - y \right\rangle - \frac{1}{2}\|x - y\|_\infty^2 \right\} = \max_{s \in \mathbb{R}^d} \left\{ \left\langle \frac{1}{L_\infty} \nabla f(x), s \right\rangle - \frac{1}{2}\|s\|_\infty^2 \right\}, \tag{8}$$

where we defined $s := x - y$.

From point (a) we know that $s^* = \|\frac{1}{L_\infty}\nabla f(x)\|_1 \operatorname{sgn}(\frac{1}{L_\infty}\nabla f(x)) \overset{L_\infty \geq 0}{=} \frac{1}{L_\infty}\|\nabla f(x)\|_1 \operatorname{sgn}(\nabla f(x))$ is a maximizer for the right-hand side expression in (8). Since $s^* = x - y^*$ (as x is fixed), we obtain that $y^* = x - \frac{1}{L_\infty}\|\nabla f(x)\|_1 \operatorname{sgn}(\nabla f(x))$, as required.

References

[1] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signsgd: Compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*, 2018.

[2] D. E. Carlson, E. Collins, Y.-P. Hsieh, L. Carin, and V. Cevher. Preconditioned spectral descent for deep learning. In *Advances in Neural Information Processing Systems*, pages 2971–2979, 2015.

[3] J. A. Kelner, Y. T. Lee, L. Orecchia, and A. Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 217–226. SIAM, 2014.