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.

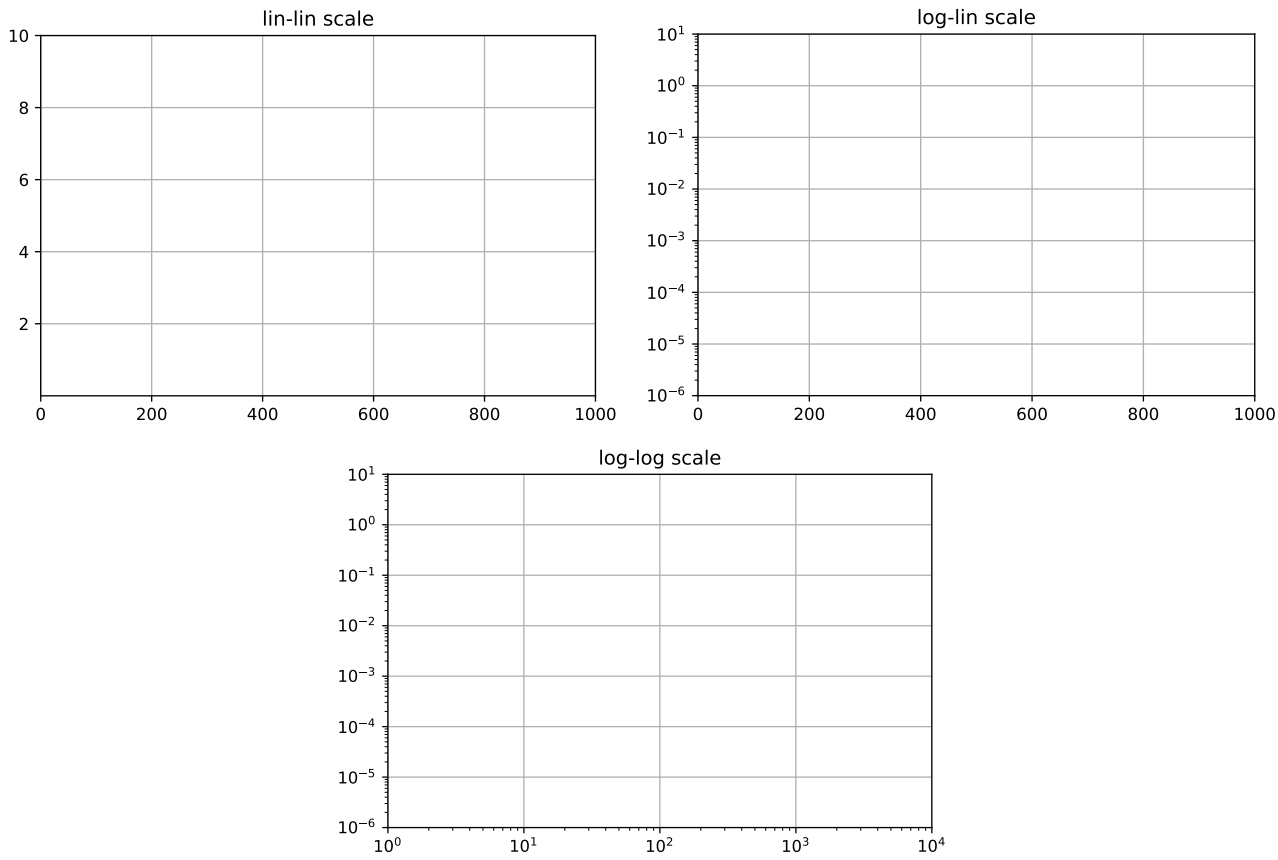
$$\text{i) } x_k = \frac{1}{k+1} \qquad \text{ii) } x_k = \frac{5k + \log(k)}{3k + 6} \qquad \text{iii) } x_k = \frac{3}{2}\exp(-k/4) \qquad \text{iv) } x_k = \frac{1}{(3k)^2} \qquad \text{v) } x_k = \frac{1}{3^{2^k}}$$

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).

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

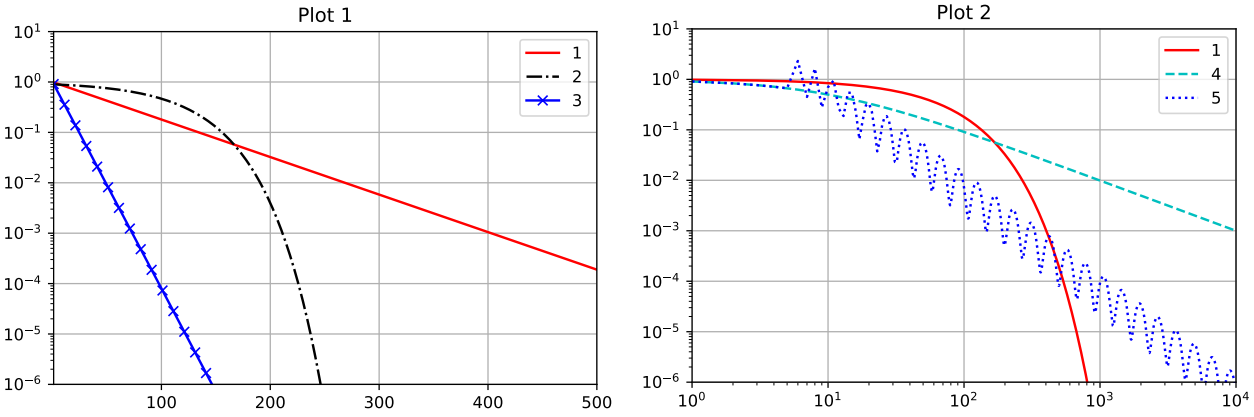
Note. Each of the sequences can be naturally drawn on one of the scales. By natural, we mean that on some scale, the asymptotic behaviour of the sequence will be displayed as a line. For instance, a line in a log-log plot means that $\log(\|\mathbf{x}_k - \mathbf{x}^*\|)$ is a *linear* function of $\log(k)$.



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.



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

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

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.

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.

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

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?

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.

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.

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

   (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.

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.