Foundations of Deep Learning **Fall 2024**

Exercise 4: Optimization

Lecturer: Aurelien Lucchi

Problem 1 (Characterizations of convex functions):
In class, you have seen that a function $\ell : \mathbb{R}^d \to \mathbb{R}$ is convex if

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d,\ \forall \lambda \in [0,1] : \ell(\lambda \mathbf{x} + (1-\lambda)\mathbf{x}') \leq \lambda \ell(\mathbf{x}) + (1-\lambda)\ell(\mathbf{x}'). \tag{1}$$

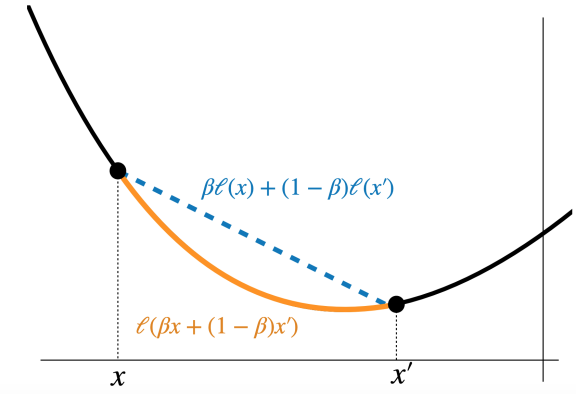
a) Show that, if ℓ is differentiable, this condition implies that

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d : \ell(\mathbf{x}) \geq \ell(\mathbf{x}') + \nabla \ell(\mathbf{x}')^\top (\mathbf{x} - \mathbf{x}'). \tag{2}$$

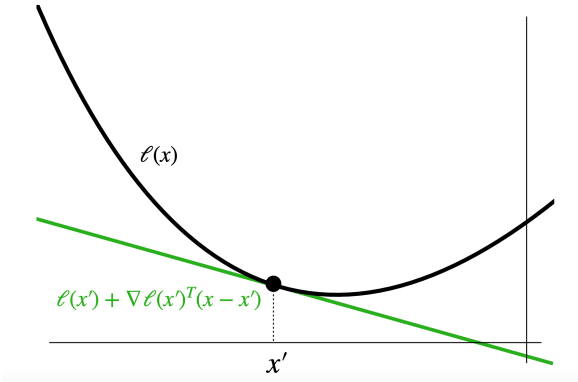
b) Show that, if ℓ is differentiable, conditions (1) and (2) are actually equivalent.

c) Now we assume that ℓ is also Lipschitz-smooth, i.e. there exists $L > 0$ such that $\|\nabla \ell(x) - \nabla \ell(\mathbf{x}')\| \leq L\|\mathbf{x} - \mathbf{x}'\|$, $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. Using the fundamental theorem of calculus as well as Cauchy-Schwarz inequality, show that

$$\ell(\mathbf{x}) \leq \ell(\mathbf{x}') + \nabla \ell(\mathbf{x}')^\top (\mathbf{x} - \mathbf{x}') + \frac{L}{2}\|\mathbf{x} - \mathbf{x}'\|^2 \tag{3}$$



Definition in the lecture



This exercise

Problem 2 (Gradient Descent for Least-square Problem):
Consider the problem of solving the linear system

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

where $\mathbf{X} = (\mathbf{x}_i^\top)_{i=1}^n \in \mathbb{R}^{n\times d}$ is the data matrix, $\mathbf{w} \in \mathbb{R}^d$ the parameter vector of the model, and $\mathbf{y} \in \mathbb{R}^n$ the target. We assume that there exists a unique solution provided by the mean-squared loss

$$\mathbf{w}_* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left(\frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 =: \ell(\mathbf{w}) \right).$$

We consider the gradient descent method

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \ell(\mathbf{w}_t) \tag{4}$$

with a fixed step size η.

a) Calculate the gradient $\nabla\ell(\mathbf{w})$ and use it to re-write the gradient descent iterates of Eq. (4).

b) Calculate the Hessian $\nabla^2\ell(\mathbf{w})$ and argue whether or not $\ell(\mathbf{w})$ is convex.

c) Use the following facts.

- Remember that $\mathbf{y} = \mathbf{X}\mathbf{w}_*$;
- Recall the definition of the operator norm of a matrix:

$$\|\mathbf{A}\|_{\text{op}} = \sup\left\{ \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} : \mathbf{x} \in \mathbb{R}^d \text{ with } \mathbf{x} \neq 0 \right\} = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})},$$

  where $\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ denotes the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$.
- for any matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and vector $\mathbf{v} \in \mathbb{R}^d$, we have $\|\mathbf{A}\mathbf{v}\|_2 \leq \|\mathbf{A}\|_{\text{op}} \cdot \|\mathbf{v}\|_2$, where $\|\mathbf{A}\|_{\text{op}}$ denotes the operator norm of the matrix $\mathbf{A}$;

Show that the following bound on the distance to the optimizer holds

$$\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2 \leq \|\mathbf{I} - \eta\mathbf{X}^\top\mathbf{X}\|_{\text{op}}^t \cdot \|\mathbf{w}_1 - \mathbf{w}_*\|_2 \tag{5}$$

d) Show that, if $\eta < \frac{1}{\|\mathbf{X}^\top\mathbf{X}\|_{\text{op}}}$, i.e. if $\|\eta\mathbf{X}^\top\mathbf{X}\|_{\text{op}} < 1$, then the Gradient Descent converge, i.e. $\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2 \to 0$ as $t \to \infty$.

e) One, obviously cheaper, alternative is to only compute the update (indexed by k) based on the gradient of one specific datapoint $\mathbf{x}_i$. This is the updated of *stochastic* gradient descent (SGD), which is arguably the most widely used optimizer in ML:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta\nabla\ell_i(\mathbf{w}_k), \quad i \in \{1,\dots,n\};\ \eta > 0.$$

where we denote $\ell_i(\mathbf{w}) := \frac{n}{2}(\mathbf{x}_i^\top\mathbf{w} - y_i)^2$ for each $i = 1, ..., n$. In each iteration k, a datapoint $\mathbf{x}_i$ is chosen uniformly at random such that $\mathbb{E}[\nabla\ell_i(\mathbf{w}_k)] = \nabla\ell(\mathbf{w}_k)$.

Show that, given a constant positive step size $\eta \leq \frac{1}{2\|\mathbf{X}^\top\mathbf{X}\|_{\text{op}}}$, we have

$$\mathbb{E}\left[\|\mathbf{w}_{k+1} - \mathbf{w}_*\|_2^2\right] \geq \eta^2\mathbb{E}\left[\|\nabla\ell_i(\mathbf{w}_k)\|_2^2\right] > 0$$

In other words, SGD does not converge to the critical point $\mathbf{w}_*$ on average.