# COS 514: Fundamentals of Deep Learning

Fall 2025

**Instructor:** Prof. Sanjeev Arora
**TA:** Gon Buzaglo

## Assignment 1

**Instructions:**

- Submission deadline is September 22.
- We recommend start reading all questions as soon as possible, as some are harder than others.
- You may collaborate in groups of up to **3** students.
- If you collaborate on a problem, you must clearly state the names of your collaborators at the beginning of the solution to that problem.
- All group members must declare that they contributed equally to the solutions.
- You must write up your own solutions independently in LaTeX. **Handwritten or scanned solutions will not be accepted.**
- Cite any resources (papers, textbooks, websites) that you use.
- Submit your assignment as a single PDF on gradescope.

# Problems

## Problem 1: Smoothness Inequality

**Definition (L-smooth).** A function $f$ is L-smooth in a domain if for every $w$ in the domain all eigenvalues of $\nabla^2 f(w)$ lie in the interval $[-L, L]$.

**Problem** Prove that

(a) if $f$ is L-smooth then

$$f(w) \leq f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + \frac{L}{2}\|w - w_t\|_2^2.$$

(b) if $f$ is $\beta_1$-smooth and $g$ is $\beta_2$-smooth, then $f + g$ is $(\beta_1 + \beta_2)$-smooth.

## Problem 2: Variance of Stochastic Gradients

Read Section 2.5.3 in the course book before attempting this problem.
**Problem** Suppose the gradient is estimated using a random sample of $B$ datapoints.

(a) Let $\widetilde{\nabla}_t^{(B)}$ be the stochastic gradient at time $t$ when the batchsize is $B$. Suppose the variance of $\widetilde{\nabla}_t^{(1)}$ (defined as $\mathbb{E}\big[\|\widetilde{\nabla}_t^{(1)} - \nabla_t\|^2\big]$) is bounded by $\gamma_1^2$. Show that there exists an upper bound $\gamma_B^2$ on the variance of $\widetilde{\nabla}_t^{(B)}$ that scales with $1/B$.

(b) Compute the asymptotic size of $T$ to find a point with $\|\nabla f(w)\| \leq \epsilon$ depending on $B$ and $\epsilon$. For simplicity, you only need to consider the case when $\eta \leq \frac{1}{\beta}$.

## Problem 3: Lipschitzness (in $x$ and in Parameters) for ReLU Networks

**Definition (Lipschitz).** A function $f : (\mathbb{R}^d, \|\cdot\|_2) \to (\mathbb{R}^k, \|\cdot\|_2)$ is $L$-Lipschitz if

$$\|f(x) - f(y)\|_2 \leq L\,\|x - y\|_2 \qquad \text{for all } x, y \in \mathbb{R}^d.$$

(For scalar $f$, the left-hand side is $|f(x) - f(y)|$.)

Throughout, use the Euclidean norm for vectors. For matrices, unless stated otherwise, use the operator (spectral) norm $\|\cdot\|_2$. Recall $\mathrm{ReLU}(z) = \max\{0, z\}$ is 1-Lipschitz.

(a) **Composition rule.** Let $g : \mathbb{R}^m \to \mathbb{R}^k$ be $L_g$-Lipschitz and $h : \mathbb{R}^d \to \mathbb{R}^m$ be $L_h$-Lipschitz. Prove that $g \circ h$ is $(L_g L_h)$-Lipschitz.

(b) **One-layer ReLU network (Lipschitz in $x$).** Let $f_1(x) = w_0^\top \mathrm{ReLU}(Wx)$. Show that $f_1$ is Lipschitz w.r.t. $x$ and give a bound in terms of $\|w_0\|_2$ and $\|W\|_2$.

(c) **Two-layer ReLU network (Lipschitz in $x$).** Let $f_2(x) = w_0^\top \mathrm{ReLU}\big(W_1\,\mathrm{ReLU}(W_2 x)\big)$. Show that $f_2$ is Lipschitz w.r.t. $x$ and bound its Lipschitz constant in terms of $\|w_0\|_2$, $\|W_1\|_2$, and $\|W_2\|_2$.

(d) **(Lipschitz in parameters.)** Fix an input $x$. For each of the following, prove Lipschitzness w.r.t. the listed parameter block and give a boundsan:

(i) For $f_1$: w.r.t. $w_0$; w.r.t. $W$.

(ii) For $f_2$: w.r.t. $w_0$; w.r.t. $W_1$; w.r.t. $W_2$.

## Problem 4: Constructing an Epsilon-Cover with Random Sampling

In this exercise, we will show that a set of $n$ points sampled uniformly at random from the unit square $[0,1]^2$ forms an $\epsilon$-cover with high probability, if $n$ is large enough. Let $S = \{s_1, \ldots, s_n\}$ be the set of $n$ randomly sampled points.

The strategy will be to first discretize the square into a fine grid of test points, show that our random sample covers all test points, and then use the triangle inequality to argue that the entire square is covered. You may use the inequality $(1-p)^n \le e^{-np}$ for $p \in [0,1]$.

1. Consider a grid of test points $T = \{(i \cdot \frac{\epsilon}{2}, j \cdot \frac{\epsilon}{2}) \mid i, j \in \{0, 1, \ldots, \lceil 2/\epsilon \rceil\}\}$. What is an upper bound on the number of points in this grid, $|T|$? Explain why any point $x \in [0,1]^2$ is within a distance of $\epsilon/2$ of at least one point in $T$. (A rough upper bound is fine for $|T|$).

2. Let $t$ be any single point in the grid $T$. Consider a ball (a disk) $B(t, \epsilon/2)$ of radius $\epsilon/2$ centered at $t$. What is the minimum possible area of the intersection of this ball with the unit square, i.e., $\mathrm{Area}(B(t, \epsilon/2) \cap [0,1]^2)$? (Hint: The minimum area occurs when $t$ is at a corner of the square).

3. For a single random point $s_i \in S$, let $p$ be the probability that it lands inside $B(t, \epsilon/2)$. Using your result from (b), what is a lower bound on $p$? Use this to show that the probability that *none* of the $n$ points in $S$ land in $B(t, \epsilon/2)$ is at most $e^{-n\pi\epsilon^2/16}$.

4. Using the union bound, find a value for $n$ (in terms of $\epsilon$ and $\delta$) that guarantees that with probability at least $1 - \delta$, *every* test point $t \in T$ has at least one sample $s_i \in S$ within a distance of $\epsilon/2$.

5. Finally, argue why the condition you proved in (d) implies that $S$ is an $\epsilon$-cover for the entire unit square $[0, 1]^2$. (Hint: For any point $x \in [0, 1]^2$, use your results from (a) and (d) along with the triangle inequality.)