CS 4644/7643: Deep Learning
Summer 2025
Problem Set 1

Instructor: Zsolt Kira
TAs: Mili Das, Yipu Chen, Kausar Patherya
Discussions: https://piazza.com/class/mafsg3dobtu42c

Deadline: 11:59pm June 5th, 2025

Instructions

1. We will be using Gradescope to collect your assignments. Please read the following instructions for submitting to Gradescope carefully!

   - For the **HW1 Theory** component on Gradescope, upload one single PDF containing the answers to all the theory questions. **However, the solution to each problem/subproblem must be on a separate page. When submitting to Gradescope, please make sure to mark the page(s) corresponding to each problem/subproblem.**
   - For the **HW1 Coding** component on Gradescope, please use the `collect_submission.ipynb` notebook provided and upload the resulting **hw1_code_submission.zip** on Gradescope. Please make sure you have saved the most recent version of your code.
   - For the **HW1 Writeup** component on Gradescope, please use the `collect_submission.ipynb` script provided and upload the resulting **hw1_notebook_submission.pdf** on Gradescope.
   - Note: This is a large class and Gradescope's assignment segmentation features are essential. Failure to follow these instructions may result in parts of your assignment not being graded. We will not entertain regrading requests for failure to follow instructions.

2. LaTeX'd solutions are strongly encouraged (solution template available in the zip file in HW1-Theory under the Assignments tab on Canvas), but scanned handwritten copies are acceptable. Hard copies are **not** accepted.

3. We generally encourage you to collaborate with other students.

   You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and *not* as a group activity. Please list the students you collaborated with.

1 Collaborators [0.5 points]

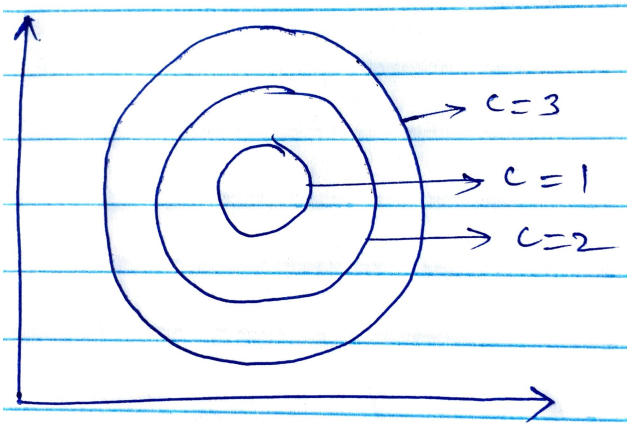
Please list your collaborators and assign this list to the corresponding section of the outline on Gradescope. If you don't have any collaborators, please write 'None' and assign it to the corresponding section of the Gradescope submission regardless.

2 Optimization

1. **[2 points]** Let $f(\mathbf{x})$ be a function, where $\mathbf{x} \in \mathbb{R}^d$. The *level surface* of a function is the set of points with the same value of function, *i.e.*

$$L_c = \{\mathbf{x} \mid f(\mathbf{x}) = c\} \tag{1}$$

for some c $\in \mathbb{R}$.



The above figure shows level surfaces for different values of c for a 2-dimensional case. The points on the perimeter of each curve illustrate the set of points sharing the same value, i.e. $f(x_1, x_2) = c$.

Recall that a gradient vector is defined as $\nabla f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_d}\right]$.

Consider a specific point $\mathbf{x}_0 \in \mathbb{R}^d$. Let us denote the level surface passing through this point by $L_{f(\mathbf{x_0})}$. Let us denote the gradient vector at $\mathbf{x_0}$ by ∇f_0.

Now, consider an arbitrary curve $\mathbf{r}(t) = [x_1(t); x_2(t); \ldots, x_d(t)]$ (parameterized by a scalar t) that passes through $\mathbf{x}_0$, *i.e.* $\mathbf{r}(t_0) = \mathbf{x_0}$ for some $t_0 \in \mathbb{R}$. Furthermore, the curve $\mathbf{r}(t)$ lies within the level surface passing through $\mathbf{x}_0$, *i.e.* $\mathbf{r}(t) \in L_{f(\mathbf{x_0})}$, $\forall t$.

Recall that the *tangent* to a curve is defined as $\frac{\partial \mathbf{r}}{\partial t}$.

With all that context and notation behind us, now comes the 'fun' part – show that ∇f_0 is orthogonal to the tangent of $\mathbf{r}(t)$ at t_0.

Now, please describe in non-technical language what you have just proven. Why might we care about this in the context of deep learning?

2. **[2 points]**

Consider a (not necessarily convex) differentiable function $g : \mathbb{R}^n \to \mathbb{R}$. g has a local minimum at some $\mathbf{w}^t$ if there exists some $\gamma > 0$ such that for all $\mathbf{w} \in \mathbb{R}^n$, $\left\|\mathbf{w}^t - \mathbf{w}\right\|_2 < \gamma \Rightarrow g(\mathbf{w}^t) \leq g(\mathbf{w})$.

Prove that if g has a local minimum at some $\mathbf{w}^t$ then the gradient at $\mathbf{w}^t = 0$, and that the converse is not necessarily true.

3. **[3 points]**

   Prove that if a differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ is convex and the gradient at some $\mathbf{w}^*$ is 0, then $\mathbf{w}^*$ is the global minimum of g (**Grading note: Students should NOT be allowed to use first order condition without proof**).

4. **[2 points]**

   During lecture 3, we discussed the softmax function $\mathbf{s}(\mathbf{z})$, which takes a vector input $\mathbf{z}$ and outputs a vector whose ith entry s_i is

$$s_i = \frac{e^{z_i}}{\sum_k e^{z_k}} \tag{2}$$

   The input vector $\mathbf{z}$ to $\mathbf{s}(\cdot)$ is sometimes called the "logits", which just means the unscaled output of previous layers. Derive the gradient of $\mathbf{s}$ with respect to the logits, *i.e.* derive $\frac{\partial \mathbf{s}}{\partial \mathbf{z}}$. Consider re-using your work from HW0.

5. **[3 points: Extra credit for both 4644 and 7643]**

   Recall that a $(d-1)$-dimensional simplex is defined as:

$$\Delta_{d-1} = \{\mathbf{p} \in \mathbb{R}^d \mid \mathbf{1}^T\mathbf{p} = 1 \text{ and } p_i \geq 0 \ \forall i \in \{1, \ldots, d\}\} \tag{3}$$

   In this question, you will develop an interpretation of softmax as a projection operator – that it projects an arbitrary point $\mathbf{x} \in \mathbb{R}^d$ onto the interior of the $d-1$ simplex. Specifically, let $\mathbf{s}(\mathbf{x})$ denote the softmax function (as defined above). Now prove that,

$$\mathbf{s}(\mathbf{x}) = \underset{\mathbf{y} \in \mathbb{R}^d}{\operatorname{argmin}} \quad -\mathbf{x}^T\mathbf{y} - H(\mathbf{y}) \tag{4}$$

$$s.t. \quad \mathbf{1}^T\mathbf{y} = 1, \quad 0 \leq y_i \leq 1 \quad \forall i \tag{5}$$

   where H is the entropy function:

$$H(\mathbf{y}) = -\sum_i y_i \log(y_i) \tag{6}$$

   Now, what does this formal interpretation tell you about the softmax layer in a neural network? Hint: Look at the KKT conditions.

3 Directed Acyclic Graphs (DAG)

One important property for a feed-forward network, as discussed in the lectures, is that it must be a directed acyclic graph (DAG). Recall that a *DAG is a directed graph that contains no directed cycles.* We will study some of its properties in this question.
Let's define a graph $G = (V, E)$ in which V is the set of all nodes as $\{v_1, v_2, ..., v_i, ...v_n\}$ and E is the set of edges $E = \{e_{i,j} = (v_i, v_j) \mid v_i, v_j \in V\}$.

A *topological order of a directed graph* $G = (V, E)$ is an ordering of its nodes as $\{v_1, v_2, ..., v_i, ...v_n\}$ so that for every edge (v_i, v_j) we have $i < j$.
There are several lemmas that can be inferred from the definition of a DAG. One lemma is: if G is a DAG, then G has a node with no incoming edges.

6. **[3 points]**

   Prove that if the graph G is a DAG, then G has a topological ordering.

7. **[3 points]**

   Prove that if the graph G has a topological order, then G is a DAG.

4 Paper Review

The first of our paper reviews for this class comes from a NeurIPS 2019 paper on the topic **'Weight Agnostic Neural Networks'** by **Adam Gaier** and **David Ha** from Google Brain.
The paper presents an interesting proposition that, through a series of experiments, re-examines some fundamental notions about neural networks - in particular, the comparative importance of architectures and weights in a network's predictive performance.
The paper can be viewed here. There's also a helpful interactive webpage with intuitive visualizations to help you understand its key concepts better.
The evaluation rubric for this section is as follows:

8. **[2 points]** Briefly summarize the key contributions, strengths and weaknesses of this paper.

9. **[2 points]** What is your personal takeaway from this paper? This could be expressed either in terms of relating the approaches adopted in this paper to your traditional understanding of learning parameterized models, or potential future directions of research in the area which the authors haven't addressed, or anything else that struck you as being noteworthy.

Guidelines: Please restrict your reviews to no more than 350 words (total length for answers to both the above questions).

5 Implement and train a network on MNIST

10. **[22 points]**

    Please refer to the instructions in the starter code zip file available on Canvas, in HW1-Coding under the Assignments tab, to get started. Please remember to label your pages after you submit your writeup.