HOMEWORK SET 0

CSCI5527 Deep Learning (Fall 2023)

Due 11:59 pm, Sep 29 2023

Instruction Your writeup, either typeset or scanned, should be a single PDF file. For problems requiring coding, organize all codes for each problem into a separate Jupyter notebook file (i.e., `.ipynb` file). Your submission to Gradescope should include the single PDF and all notebook files—**please DO NOT zip them**! No late submission will be accepted. For each problem, you should acknowledge your collaborators—**including AI tools**, if any.

About the use of AI tools You are strongly encouraged to use AI tools—they are becoming our workspace friends, such as ChatGPT (https://chat.openai.com/, which does not yet accept PDFs, but we provide the LaTeX source codes for our problems that you can copy and enter), Claude (https://claude.ai/chats, which accepts numerous forms of inputs, including PDFs), and Github Copilot (https://github.com/features/copilot), to help you when trying to solve problems. It takes a bit of practice to ask the right and effective questions/prompts to these tools; we highly recommend that you go through this popular free short course **ChatGPT Prompt Engineering for Developers** offered by https://learn.deeplearning.ai/ to get started.

If you use any AI tools for any of the problems, you should include screenshots of your prompting questions and their answers in your writeup. The answers provided by such AI tools often contain factual errors and reasoning gaps. **So, if you only submit an AI answer with such bugs for any problem, you will obtain a zero score for that problem**. You obtain the scores only when you explain the bugs and also correct them in your own writing. You can also choose not to use any of these AI tools, in which case we will grade based on the efforts you have made.

Notation We will use small letters (e.g., u) for scalars, small boldface letters (e.g., $\boldsymbol{a}$) for vectors, and capital boldface letters (e.g., $\boldsymbol{A}$) for matrices. $\mathbb{R}$ is the set of real numbers. $\mathbb{R}^n$ is the space of n-dimensional real vectors, and similarly $\mathbb{R}^{m\times n}$ is the space of $m \times n$ real matrices. The dotted equal sign $\doteq$ means defining.

Problem 1 (Matrix norms, inner products, traces; 7/15) Recall that for any vector $\boldsymbol{v} \in \mathbb{R}^n$, the ℓ_p norm of $\boldsymbol{v}$ is defined as $\|\boldsymbol{v}\|_p \doteq \left(\sum_i |v_i|^p\right)^{1/p}$. The cases when $p = 1, 2, \infty$ are often used. When $p = 2$, it is also called the Euclidean norm. Similar norms can be defined for matrices. Particularly, the direct generalization of the vector Euclidean norm is the *Frobenius norm* defined as

$$\|\boldsymbol{M}\|_F \doteq \sqrt{\sum_{ij} m_{ij}^2}$$

for a matrix $\boldsymbol{M}$. On the other hand, the inner (i.e., dot) product of matrices is defined similarly to that of vectors. For $\boldsymbol{A}, \boldsymbol{B}$ of the same size, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle \doteq \sum_{ij} a_{ij} b_{ij}$. Obviously, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \langle \boldsymbol{B}, \boldsymbol{A}\rangle$ and $\|\boldsymbol{M}\|_F = \sqrt{\langle \boldsymbol{M}, \boldsymbol{M}\rangle}$. A third notion of interest is the matrix trace, $\operatorname{tr}(\boldsymbol{M}) = \sum_i m_{ii}$, i.e., sum of the diagonal entries, which is only defined for square matrices. **Please provide detailed steps with justification for all problems; jumping into the final results leads to a zero score. Also, if we ask you to use certain facts to prove a thing, you have to use these facts (perhaps plus others); otherwise you get a zero score, e.g., if you prove everything from scratch**.

(a) Show that $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B})$ and so $\|\boldsymbol{M}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{M}^\intercal \boldsymbol{M})}$. (1/15)

(b) Use the fact in (a) to show that $\operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}^\intercal \boldsymbol{A})$. (1/15)

(c) Assume $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same size. In general, $\boldsymbol{A}\boldsymbol{B}^\intercal$ and $\boldsymbol{B}^\intercal\boldsymbol{A}$ have different sizes, but $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}^\intercal) = \operatorname{tr}(\boldsymbol{B}^\intercal\boldsymbol{A})$. Use the fact in (a) to show it! (1/15)

(d) Use the fact in (c) to show that $\operatorname{tr}(\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{M}_3) = \operatorname{tr}(\boldsymbol{M}_3\boldsymbol{M}_1\boldsymbol{M}_2) = \operatorname{tr}(\boldsymbol{M}_2\boldsymbol{M}_3\boldsymbol{M}_1)$, assuming that the sizes of $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$ are compatible with all the matrix multiplications. This is known as the *cyclic property* of the matrix trace. (1/15)

(e) For any matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ of compatible sizes, we always have $\langle \boldsymbol{A}\boldsymbol{C}\boldsymbol{B}, \boldsymbol{D}\rangle = \langle \boldsymbol{C}\boldsymbol{B}, \boldsymbol{A}^\intercal\boldsymbol{D}\rangle = \langle \boldsymbol{A}\boldsymbol{C}, \boldsymbol{D}\boldsymbol{B}^\intercal\rangle$, i.e., we can always move the **leading** matrix of one side of the inner product to the other side as **leading** matrix **once transposed** (if these matrices are complex valued, should be conjugate transposed), and similarly the **trailing** matrix to the other side as **trailing** matrix once **transposed**. Use the cyclic property of the matrix trace to show this. (1/15)

(f) For $\boldsymbol{M}$, let us perform a *compact SVD* (if not sure, check up Wikipedia! https://en.wikipedia.org/wiki/Singular_value_decomposition#Compact_SVD) to obtain $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal$, so that $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal (not necessarily square) matrices, i.e., $\boldsymbol{U}^\intercal\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^\intercal\boldsymbol{V} = \boldsymbol{I}$. Use the cyclic property of the matrix trace and that $\|\boldsymbol{M}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{M}^\intercal\boldsymbol{M})}$ to show that

$$\|\boldsymbol{M}\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2},$$

assuming the rank of $\boldsymbol{M}$ is r. Here, σ_i's are the singular values of $\boldsymbol{M}$. (1/15)

(g) Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ be three matrices of the same size, and $\odot$ denote the Hadamard (i.e., elementwise https://en.wikipedia.org/wiki/Hadamard_product_(matrices)) product between two matrices. Show that $\langle \boldsymbol{A} \odot \boldsymbol{B}, \boldsymbol{C}\rangle = \langle \boldsymbol{A}, \boldsymbol{B} \odot \boldsymbol{C}\rangle = \langle \boldsymbol{B}, \boldsymbol{A} \odot \boldsymbol{C}\rangle$. (1/15)

Problem 2 (Gradient and Hessian, chain rules, perturbation-expansion method; 6.5/15) Please provide detailed steps with justification for all problems; jumping into the final results leads to a zero score. Also, if we ask you to use certain facts or methods to derive something, you have to use these facts or methods (perhaps plus others); otherwise, you get a zero score.

(a) Let $p(\boldsymbol{X}; \boldsymbol{B}) = \frac{e^{\langle \boldsymbol{B}, \boldsymbol{X}\rangle}}{1+e^{\langle \boldsymbol{B}, \boldsymbol{X}\rangle}}$. Consider the function

$$\begin{aligned} f(\boldsymbol{B}) &= \sum_{i=1}^{N} \left[y_i \log p(\boldsymbol{X}_i; \boldsymbol{B}) + (1 - y_i) \log(1 - p(\boldsymbol{X}_i; \boldsymbol{B})) \right] \\ &= \sum_{i=1}^{N} \left[y_i \langle \boldsymbol{B}, \boldsymbol{X}_i\rangle - \log\left(1 + e^{\langle \boldsymbol{B}, \boldsymbol{X}_i\rangle}\right) \right], \end{aligned}$$

which is the log-likelihood loss for logistic regression with matrix inputs $\{\boldsymbol{X}_i\}_{i=1}^N$ and matrix coefficient $\boldsymbol{B}$, assuming N training points of the form $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^N$. Derive $\nabla f(\boldsymbol{B})$ **using the perturbation-expansion method** (see our review notes https://sunju.org/teach/DL-Fall-2023/calculus_review.pdf) (Hint: by the single-variable Taylor's theorem, $\log(1+e^{z+\delta}) = \log(1+e^z) + \frac{e^z}{1+e^z}\delta + O(\delta^2)$ as $\delta \to 0$). Validate your results either by analytic derivation using the chain rule or by any AI tools or by https://www.matrixcalculus.org/; show your derivation or screenshots. (1.5/15)

(b) In (a), $f(\boldsymbol{B})$ takes a matrix as input and so its Hessian $\nabla^2 f(\boldsymbol{B})$ is most naturally represented as a 4-th order tensor (i.e., 4-dimensional array): this is because each element in $\boldsymbol{B}$ has two indices, and so any of the second-order partial derivatives takes the form of

$$\frac{\partial^2 f}{\partial b_{ij} \partial b_{k\ell}},$$

which may be daunting to work with in practice. Fortunately, for the purpose of numerical computation, it is often sufficient to work with $\nabla^2 f(\boldsymbol{B})[\boldsymbol{V}]$: here, $\boldsymbol{V}$ is a given matrix (direction) and the resulting $\nabla^2 f(\boldsymbol{B})[\boldsymbol{V}]$ is another matrix—both matrices have the same size as that of $\boldsymbol{B}$, and we should think of $\nabla^2 f(\boldsymbol{B})[\boldsymbol{V}]$ as the "Hessian tensor" $\nabla^2 f(\boldsymbol{B})$ takes in $\boldsymbol{B}$ and returns a matrix. To derive $\nabla^2 f(\boldsymbol{B})[\boldsymbol{V}]$, we can apply our usual perturbation-expansion method to $f(\boldsymbol{B} + \delta \boldsymbol{V})$ and then compare it to

$$f(\boldsymbol{B} + \delta \boldsymbol{V}) = f(\boldsymbol{B}) + \delta \left\langle \nabla f(\boldsymbol{B}), \boldsymbol{V} \right\rangle + \frac{1}{2}\delta^2 \left\langle \boldsymbol{V}, \nabla^2 f(\boldsymbol{B})[\boldsymbol{V}] \right\rangle + o(\delta^2)$$

to extract out the corresponding $\nabla^2 f(\boldsymbol{B})[\boldsymbol{V}]$ term. **Follow this idea** to derive $\nabla^2 f(\boldsymbol{B})[\boldsymbol{V}]$ for $f(\boldsymbol{B})$ in (a). (Hint: by the single-variable Taylor's theorem, $\log(1 + e^{z+\delta}) = \log(1 + e^z) + \frac{e^z}{1+e^z}\delta + \frac{1}{2}\frac{e^z}{(1+e^z)^2}\delta^2 + O(\delta^3)$ as $\delta \to 0$). Validate your results either by analytic derivation using the chain rule (warning: this is going to be a prolonged path) or by any AI tools or by https://www.matrixcalculus.org/; show your derivation or screenshots. (1.5/15)

(c) Consider the regression problem based on a one-layer network: $g(\boldsymbol{W}) = \left\|\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})\right\|_2^2$, where the activation $\sigma = \sin$ is applied elementwise to the vector $\boldsymbol{W}\boldsymbol{x}$. **Use the perturbation-expansion method** to derive $\nabla g(\boldsymbol{W})$. Hint: You may want to first show that $\sigma((\boldsymbol{W} + \boldsymbol{\Delta})\boldsymbol{x}) = \sigma(\boldsymbol{W}\boldsymbol{x}) + \sigma'(\boldsymbol{W}\boldsymbol{x}) \odot (\boldsymbol{\Delta}\boldsymbol{x}) + o(\|\boldsymbol{\Delta}\|_F)$ when $\boldsymbol{\Delta} \to \boldsymbol{0}$, and then plug it back into the square and expand the square using the identity $\|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|_2^2 = \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + \|\boldsymbol{c}\|_2^2 + 2\langle \boldsymbol{a}, \boldsymbol{b}\rangle + 2\langle \boldsymbol{a}, \boldsymbol{c}\rangle + 2\langle \boldsymbol{b}, \boldsymbol{c}\rangle$. Keep in mind that to derive the gradient, any terms of order lower than $\|\boldsymbol{\Delta}\|_F$ are not interesting. Validate your results either by analytic derivation using the chain rule (warning: this is going to be a prolonged path) or by any AI tools or by https://www.matrixcalculus.org/; show your derivation or screenshots. (1.5/15)

(d) Consider $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_2^2$ with $\lambda > 0$. Any $\boldsymbol{x}_0$ is a candidate minimizer of $f(\boldsymbol{x})$ if $\nabla f(\boldsymbol{x}_0) = \boldsymbol{0}$ (i.e., first-order optimality condition). Please find all candidate minimizers of $f(\boldsymbol{x})$. How many can you find in total? Please include your detailed calculation and reasoning; you can use whatever way you want to derive $\nabla f(\boldsymbol{x})$ if necessary. (Hint: recall that a symmetric matrix $\boldsymbol{M}$ is positive definite if and only if $\boldsymbol{v}^\intercal \boldsymbol{M} \boldsymbol{v} > 0$ for all $\boldsymbol{v} \neq 0$; all positive definite matrices are invertible) (1.5/15)

(e) Continuing with (d). What happens when $\lambda < 0$? (0.5/15)

Problem 3 (Conditional probability and Bayes' Rule; 1.5/15) Let $P(A \mid B)$ be the probability of event A given event B. Bayes' theorem states that

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}.$$

In general, for a partition $\{A_i\}$ of the sample space,

$$P(A_i \mid B) = \frac{P(B \mid A_i) P(A_i)}{\sum_i P(B \mid A_i) P(A_i)}.$$

(a) Suppose that women who often drink alcohol are 10 times more likely to develop stomach cancer than women who do not, while the corresponding risk for men is 20 (drinker v.s. non-drinker). Statistical data show that 40% of women drink alcohol often. Now, a woman has been diagnosed with stomach cancer. But due to some privacy issues, it is impossible for you to get other information about this woman. Based on the limited information, is it possible to calculate the probability that she is a drinker? If yes, what is the probability? If not, explain why and what additional information you will need.[1] $(1/15)$

(b) Doctors apply a standardized test for a certain disease. If the patient has the disease, the test shows a positive result with a probability of 99%. However, with 2% probability, a healthy patient can have a positive test result. Statistical data show that 1 of 1000 in the population have the disease. What is the probability that a patient with a positive test is affected by the disease? $(0.5/15)$

[1] All data and information in this question are made up. No medical studies/institutions have proven it.