HOMEWORK SET 0

CSCI 5980/8980 Think Deep Learning (Fall 2020)

Due 11:59 pm, Sep 30 2020

Instruction Please typeset your homework in LaTeX and submit it as a single PDF file in Canvas. No late submission will be accepted. For each problem, your should acknowledge your collaborators if any. For problems containing multiple subproblems, there are often close logic connections between the subproblems. So always remember to build on previous ones, rather than work from scratch.

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 (Chain rules, gradient and Hessian) Recall from calculus that for a multivariate function $f(\boldsymbol{x})$ mapping from $\mathbb{R}^n$ to $\mathbb{R}$, i.e., $f : \mathbb{R}^n \mapsto \mathbb{R}$, the i-th partial derivative of f is defined as $\frac{\partial f}{\partial x_i}$, i.e., the univariate derivative with respect to the i-th variable while holding the other variables constant. This generalizes naturally to the matrix case, where we consider $f(\boldsymbol{X})$ with $\boldsymbol{X} \in \mathbb{R}^{m\times n}$. The (i, j)-th partial derivative of $f(\boldsymbol{X})$ is then $\frac{\partial f}{\partial X_{ij}}$. For either case, the *gradient* is the ordered collection of the partial derivatives. A common convention is to order the partial derivatives into the same shape as the input variable. So for $f(\boldsymbol{x})$, the gradient $\nabla f(\boldsymbol{x}) \in \mathbb{R}^n$ is

$$\nabla f(\boldsymbol{x}) \doteq \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_i} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}.$$

Similarly for $f(\boldsymbol{X})$, the gradient $\nabla f(\boldsymbol{X}) \in \mathbb{R}^{m\times n}$ is

$$\nabla f(\boldsymbol{X}) \doteq \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \cdots & \frac{\partial f}{\partial X_{1j}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f}{\partial X_{i1}} & \cdots & \frac{\partial f}{\partial X_{ij}} & \cdots & \frac{\partial f}{\partial X_{in}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f}{\partial X_{m1}} & \cdots & \frac{\partial f}{\partial X_{mj}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{bmatrix}.$$

Each of the $\frac{\partial f}{\partial x_i}$ (respectively $\frac{\partial f}{\partial X_{ij}}$) is a real-valued function of $\boldsymbol{x}$ (respectively $\boldsymbol{X}$) again. So we can take the partial derivatives with respect to entries in $\boldsymbol{x}$ (respectively $\boldsymbol{X}$) again to obtain the 2nd order partial derivatives like

$$\frac{\partial \left(\frac{\partial f}{\partial x_i}\right)}{\partial x_\ell},$$

which we often write neatly as

$$\frac{\partial^2 f}{\partial x_\ell \partial x_i}.$$

For $f(\boldsymbol{x})$, there are two indices (ℓ, j) to index all the 2nd order partial derivatives, and so it is convenient to form an $n \times n$ matrix whose (ℓ, j)-th entry is $\frac{\partial^2 f}{\partial x_\ell \partial x_i}$ to collect them. We shall write this matrix compactly as $\left[\frac{\partial^2 f}{\partial x_\ell \partial x_i}\right]_{\ell,i} \in \mathbb{R}^{n\times n}$, and this is the *Hessian* for $f(\boldsymbol{x})$:

$$\nabla^2 f(\boldsymbol{x}) \doteq \left[\frac{\partial^2 f}{\partial x_\ell \partial x_i}\right]_{\ell,i}.$$

Under very mild conditions, Hessian for $f(\boldsymbol{x})$ is symmetric, that is, $\frac{\partial^2 f}{\partial x_\ell \partial x_i} = \frac{\partial^2 f}{\partial x_i \partial x_\ell}$ for all i and ℓ. The Hessian for $f(\boldsymbol{X})$ is more complicated. There are four indices to index all the 2nd order partial derivatives, e.g., $\frac{\partial^2 f}{\partial X_{\ell,k} \partial X_{i,j}}$. So it is natural to use a 4-dimensional array (matrices are 2-dimensional arrays) to collect them here, i.e., we need 4-D tensors. We will cover more about tensors when we come to deep learning on images. In practice, one would never explicitly form the Hessian of $f(\boldsymbol{X})$ for computation. We will talk more about this in future lectures and homework sets on numerical optimization.

Let $y = f(u)$ be differentiable with respect to u and $u = g(x)$ be differentiable with respect to x. The univariate chain rule says

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}.$$

This can be extended to the multivariate case. Let $y = f(u_1, \ldots, u_j, \ldots, u_m)$ be differentiable and $u_1, \ldots, u_j, \ldots, u_m$ be differentiable functions of $(x_1, \ldots, x_i, \ldots, x_n)$. Then, the multivariate chain rule says that

$$\frac{dy}{dx_i} = \sum_j \frac{\partial y}{\partial u_j}\frac{\partial u_j}{\partial x_i} \quad \forall\, i.$$

We can also write the multivariate version in matrix notations. Write $y = f(\boldsymbol{u})$ where $\boldsymbol{u} \in \mathbb{R}^m$, and $\boldsymbol{u}$ as a differentiable function of $\boldsymbol{x} \in \mathbb{R}^n$. Then,

$$\nabla_{\boldsymbol{x}} f = [\boldsymbol{J}_{\boldsymbol{u}}(\boldsymbol{x})]^\intercal \nabla_{\boldsymbol{u}} f,$$

where $\boldsymbol{J}_{\boldsymbol{u}}(\boldsymbol{x}) \in \mathbb{R}^{m\times n}$ is the Jacobian matrix of $\boldsymbol{u}$ with respect to $\boldsymbol{x}$:

$$\boldsymbol{J}_{\boldsymbol{u}}(\boldsymbol{x}) = \left[\frac{\partial u_j}{\partial x_i}\right]_{j,i}.$$

(a) Let $p(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}^\intercal \boldsymbol{x}}}{1+e^{\boldsymbol{\beta}^\intercal \boldsymbol{x}}}$. The log-likelihood for logistic regression with two classes is (assuming N samples of the form $(\boldsymbol{x}_i, y_i)$)

$$\begin{aligned} f(\boldsymbol{\beta}) &= \sum_{i=1}^N \left[y_i \log p(\boldsymbol{x}_i; \boldsymbol{\beta}) + (1-y_i)\log\left(1 - p(\boldsymbol{x}_i; \boldsymbol{\beta})\right)\right] \\ &= \sum_{i=1}^N \left[y_i \boldsymbol{\beta}^\intercal \boldsymbol{x}_i - \log\left(1 + e^{\boldsymbol{\beta}^\intercal \boldsymbol{x}_i}\right)\right]. \end{aligned}$$

Derive the gradient and Hessian of $f(\boldsymbol{\beta})$. Please include your calculation details. (1/12)

(b) Let $\sigma = \tanh(x)$, i.e., the *hyperbolic tangent function*. Derive the gradient of the matrix-variable function $g(\boldsymbol{W}) = \|\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})\|_2^2$, where σ is applied to the vector $\boldsymbol{W}\boldsymbol{x}$ elementwise. This is regression based on a one-layer network. Please include your calculation details. (1/12)

Problem 2 (Matrix norms, inner products, traces) Recall that for any vector $\boldsymbol{v} \in \mathbb{R}^n$, the ℓ_p norm of $\boldsymbol{v}$ is defined as $\|\boldsymbol{v}\|_p \doteq (\sum_i |v_i|^p)^{1/p}$. 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 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.

(a) Show that $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B})$ and so $\|\boldsymbol{M}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{M}^\intercal \boldsymbol{M})}$. (1/12)

(b) Show that $\operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}^\intercal \boldsymbol{A})$. (1/12)

(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 $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}^\intercal) = \operatorname{tr}(\boldsymbol{B}^\intercal \boldsymbol{A})$. Show it! (1/12)

(d) 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 matrix traces. (Hint: think of (c)) (1/12)

(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**. Why? (Hint: think of the above results and also try to remember this important property that will be useful for calculation later) (1/12)

(f) For $\boldsymbol{M}$, let's 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 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 $\boldsymbol{M}$. (1/12)

Problem 3 (Taylor expansion of multivariate functions) We learned Taylor's theorem for single variable functions in our first calculus course. For a function f that is twice-differentiable around x,

$$f(x+\delta) = f(x) + f'(x)\,\delta + \frac{1}{2} f''(x)\,\delta^2 + o(|\delta|^2) \quad \text{as } \delta \to 0,$$

i.e., the approximation error of the 2nd order Taylor expansion $f(x) + f'(x)\,\delta + \frac{1}{2} f''(x)\,\delta^2$ is lower order than $|\delta|^2$ when δ is sufficiently small. The result generalizes naturally to the multivariate case. For $f : \mathbb{R}^n \to \mathbb{R}$ that is twice-differentiable around $\boldsymbol{x}$,

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = f(\boldsymbol{x}) + [\nabla f(\boldsymbol{x})]^\intercal \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^\intercal \nabla^2 f(\boldsymbol{x})\,\boldsymbol{\delta} + o\left(\|\boldsymbol{\delta}\|_2^2\right).$$

Similarly, for $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ that is twice-differentiable around $\boldsymbol{X}$,

$$f(\boldsymbol{X}+\boldsymbol{\Delta}) = f(\boldsymbol{X}) + \operatorname{tr}\left([\nabla f(\boldsymbol{X})]^\intercal \boldsymbol{\Delta}\right) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Delta}^\intercal \nabla^2 f(\boldsymbol{X})\,\boldsymbol{\Delta}\right) + o\left(\|\boldsymbol{\Delta}\|_F^2\right).$$

Consider the one-layer network regression again: $g(\boldsymbol{W}) = \|\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})\|_2^2$ with $\sigma(x) = \tanh(x)$. Let's try to work out its 1st order Taylor expansion by direct expansion as follows.

- 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 $\boldsymbol{\Delta} \to \boldsymbol{0}$. Here, both σ and σ' are applied elementwise, and $\odot$ denotes the elementwise (Hadamard) product. (1/12)
- So $\boldsymbol{y} - \sigma((\boldsymbol{W}+\boldsymbol{\Delta})\boldsymbol{x}) = (\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})) - \sigma'(\boldsymbol{W}\boldsymbol{x}) \odot (\boldsymbol{\Delta}\boldsymbol{x}) - o(\|\boldsymbol{\Delta}\|_F)$ when $\boldsymbol{\Delta} \to \boldsymbol{0}$. Substitute this back into the square and use the identity $\|\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}\|_2^2 = \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + \|\boldsymbol{c}\|_2^2 + 2\boldsymbol{a}^\intercal\boldsymbol{b} + 2\boldsymbol{a}^\intercal\boldsymbol{c} + 2\boldsymbol{b}^\intercal\boldsymbol{c}$ to obtain the first-order approximation to $g(\boldsymbol{W}+\boldsymbol{\Delta})$. Remember that any terms lower order than $\|\boldsymbol{\Delta}\|_F$ are not interesting and we can always assume $\boldsymbol{\Delta}$ as small as needed. (1/12)
- Substitute the result from **Problem 1(b)** into the 1st order Taylor expansion formula above and compare it to the result obtained here. Are they equal or not? (0.5/12)

Problem 4 (Conditional probability and Bayes' Rule) 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 drink alcohol often are 10 times more likely to develops a 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 the 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 no, explain why and what extra information you will need.[1] (1/12)

(b) Doctors apply a standardized test for a certain disease. If the patient has the disease, the test shows a positive result with a 99% chance. However, with 2% probability a healthy patient can have a positive test. Statistical data show that 1 out of 1000 in the population have the disease. What is the probability for a patient with a positive test to be affected by the disease? (0.5/12)

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