Due Date: March 30th (11pm), 2022

Instructions

- *For all questions, show your work!*
- *Starred questions are* ***hard*** *questions, not* ***bonus*** *questions.*
- *Use LaTeX and the template we provide when writing your answers. You may reuse most of the notation shorthands, equations and/or tables. See the assignment policy on the course website for more details.*
- *Unless noted that questions are related, assume that notation and defintions for each question are self-contained and independent.*
- *Submit your answers electronically via Gradescope.*
- *TAs for this assignment are* ***Ankit Vani*** *and* ***Sai Aravind Sreeramadas****.*

Question 1 (6-9-6)**.** This question is about activation functions and vanishing/exploding gradients in recurrent neural networks (RNNs). Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function. When the argument is a vector, we apply σ element-wise. Consider the following recurrent unit:

$$\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$$

1.1 Show that applying the activation function in this way results in an equivalent recurrence as the conventional way of applying the activation function: $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$ (i.e. express $\boldsymbol{g}_t$ in terms of $\boldsymbol{h}_t$). More formally, you need to prove it using mathematical induction. You only need to prove the induction step in this question, assuming your expression holds for time step $t-1$.

*1.2 Let $||\boldsymbol{A}||$ denote the L_2 operator norm [1] of matrix $\boldsymbol{A}$ ($||\boldsymbol{A}|| := \max_{\boldsymbol{x}:||\boldsymbol{x}||=1} ||\boldsymbol{A}\boldsymbol{x}||$). Assume $\sigma(x)$ has bounded derivative, i.e. $|\sigma'(x)| \leq \gamma$ for some $\gamma > 0$ and for all x. We denote as $\lambda_1(\cdot)$ the largest eigenvalue of a symmetric matrix. Show that if the largest eigenvalue of the weights is bounded by $\frac{\delta^2}{\gamma^2}$ for some $0 \leq \delta < 1$, gradients of the hidden state will vanish over time, i.e.

$$\lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2} \quad \implies \quad \left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| \to 0 \text{ as } T \to \infty$$

Use the following properties of the L_2 operator norm

$$||\boldsymbol{A}\boldsymbol{B}|| \leq ||\boldsymbol{A}||\,||\boldsymbol{B}|| \qquad \text{and} \qquad ||\boldsymbol{A}|| = \sqrt{\lambda_1(\boldsymbol{A}^\top \boldsymbol{A})}$$

1.3 What do you think will happen to the gradients of the hidden state if the condition in the previous question is reversed, i.e. if the largest eigenvalue of the weights is larger than $\frac{\delta^2}{\gamma^2}$? Is this condition *necessary* and/or *sufficient* for the gradient to explode ? (Answer in 1-2 sentences).

Answer 1.

1.1 The hypothesis to prove is $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$. Assume that $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$. Then, given the definititions of $\boldsymbol{g}_t$ and $\boldsymbol{h}_t$,

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}) = \sigma(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}) = \sigma(\boldsymbol{h}_t)$$

[1] The L_2 operator norm of a matrix $\boldsymbol{A}$ is is an *induced norm* corresponding to the L_2 norm of vectors. You can try to prove the given properties as an exercise.

1.2 For consecutive units, the Jacobian is

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} = \boldsymbol{W} \frac{\partial \sigma(\boldsymbol{h}_{t-1})}{\partial \boldsymbol{h}_{t-1}}$$

Recall the following properties of 2-norm:

$$||\boldsymbol{AB}|| \leq ||\boldsymbol{A}|| \, ||\boldsymbol{B}|| \qquad \text{and} \qquad ||\boldsymbol{A}|| = \sqrt{\lambda_1(\boldsymbol{A}^\top \boldsymbol{A})}$$

from which we have

$$\left|\left| \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \right|\right| \leq ||\boldsymbol{W}|| \left|\left| \frac{\partial \sigma(\boldsymbol{h}_{t-1})}{\partial \boldsymbol{h}_{t-1}} \right|\right| \leq \frac{\delta}{\gamma}\gamma = \delta$$

which means the 2-norm is bounded by some $\delta \in [0, 1)$. Applying the sub-multiplicativity T times gives

$$\left|\left| \frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0} \right|\right| \leq \prod_{t=1}^{T} \left|\left| \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \right|\right| \leq \delta^T \to 0 \text{ as } T \to \infty$$

1.3 By contraposition, gradients of the hidden state would not become arbitrarily large if the largest eigenvalue of weights is not larger than $\frac{\delta^2}{\gamma^2}$. Thus it is a necessary condition for gradient explosion. It is not sufficient: the product of the norms can be greater than the norm of the product. This can happen if the hidden state is orthogonal to the largest eigenvector of $\boldsymbol{W}$.

Question 2 (8-8-8)**.** In this question you will demonstrate that an estimate of the first moment of the gradient using an (exponential) running average is equivalent to using momentum, and is biased by a scaling factor. The goal of this question is for you to consider the relationship between different optimization schemes, and to practice noting and quantifying the effect (particularly in terms of bias/variance) of *estimating* a quantity.

Let $\boldsymbol{g}_t$ be an unbiased sample of gradient at time step t and $\Delta\boldsymbol{\theta}_t$ be the update to be made. Initialize $\boldsymbol{v}_0$ to be a vector of zeros.

2.1 For $t \geq 1$, consider the following update rules:

- SGD with momentum:

$$\boldsymbol{v}_t = \alpha \boldsymbol{v}_{t-1} + \epsilon \boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t$$

  where $\epsilon > 0$ and $\alpha \in (0, 1)$.

- SGD with running average of $\boldsymbol{g}_t$:

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\delta \boldsymbol{v}_t$$

  where $\beta \in (0, 1)$ and $\delta > 0$.

Express the two update rules recursively ($\Delta\boldsymbol{\theta}_t$ as a function of $\Delta\boldsymbol{\theta}_{t-1}$). Show that these two update rules are equivalent ; i.e. express (α, ϵ) as a function of (β, δ).

2.2 Unroll the running average update rule, i.e. express $\boldsymbol{v}_t$ as a linear combination of $\boldsymbol{g}_i$'s ($1 \leq i \leq t$).

2.3 Assume $\boldsymbol{g}_t$ has a stationary distribution independent of t. Show that the running average is biased, i.e. $\mathbb{E}[\boldsymbol{v}_t] \neq \mathbb{E}[\boldsymbol{g}_t]$. Propose a way to eliminate such a bias by rescaling $\boldsymbol{v}_t$.

Answer 2.

2.1 For SGD with momentum,

$$\Delta\boldsymbol{\theta}_t = -\alpha \boldsymbol{v}_{t-1} - \epsilon \boldsymbol{g}_t = \alpha \Delta\boldsymbol{\theta}_{t-1} - \epsilon \boldsymbol{g}_t$$

For SGD with running average,

$$\Delta\boldsymbol{\theta}_t = -\delta\beta \boldsymbol{v}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t = \beta \Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t$$

To equalize the two, set $(\alpha, \epsilon) = (\beta, \delta(1-\beta))$, and reversely $(\beta, \delta) = (\alpha, \frac{\epsilon}{1-\alpha})$.

2.2 For SGD with running average we have:

$$\begin{aligned}\boldsymbol{v}_t &= \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \\ &= \beta(\beta \boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t \\ &= (1-\beta)\sum_{i=1}^{t} \beta^{t-i}\boldsymbol{g}_i\end{aligned}$$

2.3 Taking the expectation of the running average yields

$$\begin{aligned}\mathbb{E}[\boldsymbol{v}_t] &= \mathbb{E}\left[(1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i\right] \\ &= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathbb{E}[\boldsymbol{g}_i] && \text{(Linearity)} \\ &= \mathbb{E}[\boldsymbol{g}_t](1-\beta)\sum_{i=1}^{t}\beta^{t-i} && \text{(Stationarity)} \\ &= \mathbb{E}[\boldsymbol{g}_t](1-\beta^t) && \text{(Telescoping sum)}\end{aligned}$$

One can use $\frac{\boldsymbol{v}_t}{1-\beta^t}$ instead as an unbiased estimate.

Question 3 (8-8-6-9-3)**.** In this question, you will analyze the performance of dot-product attention and derive an efficient approximation of it. Consider that *multi-head* dot-product attention for a sequence of length n is defined as follows:

$$\begin{aligned}\text{MultiHead}(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{K}}, \bar{\boldsymbol{V}}) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)\boldsymbol{W}^O \\ \text{where} \quad \text{head}_i &= \text{Attention}_{\text{std}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \quad (\text{here, } \boldsymbol{Q} := \bar{\boldsymbol{Q}}\boldsymbol{W}_i^Q, \boldsymbol{K} := \bar{\boldsymbol{K}}\boldsymbol{W}_i^K, \boldsymbol{V} := \bar{\boldsymbol{V}}\boldsymbol{W}_i^V) \\ &= \text{softmax}_{\text{row}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}\end{aligned}$$

where $\bar{\boldsymbol{Q}}, \bar{\boldsymbol{K}}, \bar{\boldsymbol{V}} \in \mathbb{R}^{n \times d_{\text{model}}}$ are the queries, keys, and values, and $\boldsymbol{W}_i^Q, \boldsymbol{W}_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ $\forall i$, and $\boldsymbol{W}_O \in \mathbb{R}^{hd_v \times d_{model}}$ are the weights. The softmax subscript "row" indicates that the softmax is computed along the rows, and the Attention subscript "std" indicates that this is the standard variant (we will see other variants later in the question). For this question, you can assume that $d_k = d_v = d_{\text{model}}$ and call the value d.

For calculating the time and space complexities, you can also assume that matrix multiplications are performed naively. As an example, for $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$ where $\boldsymbol{A} \in \mathbb{R}^{p \times q}$, $\boldsymbol{B} \in \mathbb{R}^{q \times r}$, and $\boldsymbol{C} \in \mathbb{R}^{p \times r}$, the time complexity is $\Theta(pqr)$ due to the three nested loops, and the space complexity is $\Theta(pq + qr + pr)$ from storing the inputs and the result.

3.1 What is the time and space complexity of the attention operation carried out by a single head in Θ-notation in terms of n and d ? Use your answer to calculate the time and space complexity of multi-head dot-product attention in terms of n, d, and h, assuming that the heads are computed sequentially. For very long sequences, where does the bottleneck lie ?

For the remaining parts, let us focus on the attention operation carried out by a single head. Furthermore, you can omit the scaling factor $\sqrt{d}$ without loss of generality by considering that $\boldsymbol{Q}$ and $\boldsymbol{K}$ can be scaled as desired.

3.2 Let us consider an alternative form of attention, one that performs row-wise softmax on $\boldsymbol{Q}$ and column-wise softmax on $\boldsymbol{K}$ separately as follows:

$$\text{Attention}_{\text{separable}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}_{\text{row}}(\boldsymbol{Q})\text{softmax}_{\text{col}}(\boldsymbol{K})^\top \boldsymbol{V}.$$

Prove that $\text{softmax}_{\text{row}}(\boldsymbol{Q})\text{softmax}_{\text{col}}(\boldsymbol{K})^\top$ produces valid categorical distributions in every row, like $\text{softmax}_{\text{row}}(\boldsymbol{Q}\boldsymbol{K}^\top)$. If $n \gg d$, show that $\text{Attention}_{\text{separable}}$ can be faster and requires less space than $\text{Attention}_{\text{std}}$. Is $\text{Attention}_{\text{separable}}$ as expressive as $\text{Attention}_{\text{std}}$?

(Hint: For a valid categorical distribution $\boldsymbol{p} \in \mathbb{R}^d$ over d categories, $p_i \geq 0 \, \forall i \in \{1, \ldots, d\}$ and $\sum_{i=1}^{d} p_i = 1$.)

3.3 Verify that the standard attention can be written as

$$\text{Attention}_{\text{std}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{V}$$

where $\boldsymbol{A} = \exp\left(\boldsymbol{Q}\boldsymbol{K}^\top\right)$ and $\boldsymbol{D} = \text{diag}(\boldsymbol{A}\mathbf{1})$, where exp is an element-wise operation, diag creates a diagonal matrix from a vector, and $\mathbf{1}$ is a vector of ones. Note that you can store diagonal matrices in linear space and compute matrix multiplications with them in linear time.

Let us now consider a variant $\text{Attention}_{\text{approx}}$ where the elements a_{ij} of $\boldsymbol{A}$ can be represented as $a_{ij} = f(\boldsymbol{q}_i)^\top f(\boldsymbol{k}_j)$ for some $f : \mathbb{R}^d \to \mathbb{R}_+^m$, where $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$ are the ith row of $\boldsymbol{Q}$ and the jth row of $\boldsymbol{K}$ respectively.

If $n \gg m$ and $n \gg d$, how can you use this formulation to make attention efficient ? What is the time and space complexity of $\text{Attention}_{\text{approx}}$? You can assume that f takes $\Theta(md)$ time and space.

(Hint: Decompose the matrix $\boldsymbol{A}$.)

*3.4 Prove that in $\text{Attention}_{\text{std}}$,

$$a_{ij} = \exp\left(\frac{-\|\boldsymbol{q}_i\|^2}{2}\right) \cdot \mathbb{E}_{\boldsymbol{x} \in \mathcal{N}(\mathbf{0},\mathbf{I})}\left[\exp(\boldsymbol{x}^\top \boldsymbol{q}_i)\exp(\boldsymbol{x}^\top \boldsymbol{k}_j)\right] \cdot \exp\left(\frac{-\|\boldsymbol{k}_j\|^2}{2}\right).$$

Use this result to devise the function $f : \mathbb{R}^d \to \mathbb{R}_+^m$ introduced in the previous part, such that $\text{Attention}_{\text{approx}}$ approximates the expectation in $\text{Attention}_{\text{std}}$ by sampling.

(Hint 1: If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, $p(\boldsymbol{x}) = (2\pi)^{-d/2}\exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}\|^2\right)$ and $\int_{\boldsymbol{x}} p(\boldsymbol{x})d\boldsymbol{x} = 1$.)

(Hint 2: $\boldsymbol{x}^\top \boldsymbol{y} = -\frac{1}{2}(\boldsymbol{x}^\top \boldsymbol{x} - (\boldsymbol{x} + \boldsymbol{y})^\top(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{y}^\top \boldsymbol{y})$. This can be useful when starting the proof in the reverse direction.)

3.5 Discuss the implications of the choice of m for $\text{Attention}_{\text{approx}}$. What are the trade-offs to think about ?

Answer 3.

3.1 For $\text{Attention}_{\text{std}}$, the time complexity is $\Theta\left(n^2 d\right)$ and the space complexity is $\Theta\left(n^2 + nd\right)$.

For multi-head attention, we have to additionally consider the weights $\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$, $\boldsymbol{W}_i^V$ $\forall i$, and $\boldsymbol{W}^O$, their multiplication with $\bar{\boldsymbol{Q}}$, $\bar{\boldsymbol{K}}$, $\bar{\boldsymbol{V}}$, and the concatenation of the heads respectively. Since the heads are computed sequentially, we do not need to store the $n \times n$ intermediate matrices for each head simultaneously.

Thus, for MultiHead, the time complexity is $\Theta\left(hn^2 d + hnd^2\right)$ and the space complexity is $\Theta\left(n^2 + hd^2 + hnd\right)$.

(Alternatively, if you include the computation of $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ as a part of single-head, then the single-head time and space complexities are $\Theta(n^2 d + nd^2)$ and $\Theta(n^2 + d^2 + nd)$, and the multi-head complexities remain the same.)

When n is large, the n^2 costs dominate, which occur within the $\text{Attention}_{\text{std}}$ operation for the matrix multiplications and computation of softmax.

3.2 Let $\boldsymbol{Y} = \text{softmax}_{\text{row}}(\boldsymbol{Q})\text{softmax}_{\text{col}}(\boldsymbol{K})^\top$. Let $\boldsymbol{q}_i$ and $\boldsymbol{k}_i$ refer to the column vectors obtained by selecting row i from $\text{softmax}_{\text{row}}(\boldsymbol{Q})$ and $\text{softmax}_{\text{col}}(\boldsymbol{K})$ respectively. Furthermore, let y_{ij}, q_{ij}, and k_{ij} denote the elements of $\boldsymbol{Y}$, $\text{softmax}_{\text{row}}(\boldsymbol{Q})$, and $\text{softmax}_{\text{col}}(\boldsymbol{K})$ at positions i, j respectively.

First, we see that $y_{ij} = \boldsymbol{q}_i^\top \boldsymbol{k}_j$. Since $q_{ij} \geq 0, k_{ij} \geq 0 \, \forall i, j$, it is clear that $y_{ij} \geq 0$.

Then, we have the sum the elements in the ith row of Y

$$\begin{aligned}\sum_j y_{ij} = \sum_j \boldsymbol{q}_i^\top \boldsymbol{k}_j = \sum_j \sum_l q_{il} k_{jl} &= \sum_l \left(q_{il} \sum_j k_{jl} \right) \\ &= \sum_l q_{il} && \text{(columns in } \text{softmax}_{\text{col}}(\boldsymbol{K}) \text{ sum to 1)} \\ &= 1. && \text{(rows in } \text{softmax}_{\text{row}}(\boldsymbol{Q}) \text{ sum to 1)}\end{aligned}$$

Since all elements of $\boldsymbol{Y}$ are non-negative and every row sums to 1, each row of $\boldsymbol{Y}$ represents a categorical distribution.

If we set the order of multiplications to

$$\text{Attention}_{\text{separable}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}_{\text{row}}(\boldsymbol{Q}) \left(\text{softmax}_{\text{col}}(\boldsymbol{K})^\top \boldsymbol{V} \right),$$

we can change the time complexity of attention from $\Theta(n^2 d)$ to $\Theta(nd^2)$, and the space complexity from $\Theta(n^2 + nd)$ to $\Theta(d^2 + nd)$, which are more efficient when $n \gg d$.

In general, $\text{Attention}_{\text{separable}}$ is not as expressive as $\text{Attention}_{\text{std}}$, since $\text{Attention}_{\text{std}}$ can represent all possible mappings possible through $\text{Attention}_{\text{separable}}$, but the inverse is not true. The direct interactions between $\boldsymbol{Q}$ and $\boldsymbol{K}$ in $\text{Attention}_{\text{std}}$ cannot be expressed in $\text{Attention}_{\text{separable}}$.

3.3 Using the definition of softmax, $\text{softmax}_{\text{row}}(\boldsymbol{M})$ can be written as $\boldsymbol{D}^{-1}\boldsymbol{A}$ where $\boldsymbol{A} = \exp(\boldsymbol{M})$ and $\boldsymbol{D} = \text{diag}(\boldsymbol{A}\boldsymbol{1})$. Thus, we can write $\text{Attention}_{\text{std}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{V}$ where $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{K}^\top$. Let us apply $f : \mathbb{R}^d \to \mathbb{R}_+^m$ to each row of $\boldsymbol{Q} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{K} \in \mathbb{R}^{n \times d}$:

$$\hat{\boldsymbol{Q}} = f_{\text{row}}(\boldsymbol{Q}), \quad \hat{\boldsymbol{K}} = f_{\text{row}}(\boldsymbol{K}),$$

to obtain $\hat{\boldsymbol{Q}} \in \mathbb{R}^{n \times m}$ and $\hat{\boldsymbol{K}} \in \mathbb{R}^{n \times m}$.

Then, we can decompose $\boldsymbol{A} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{K}}^\top$, and rewrite attention as

$$\text{Attention}_{\text{approx}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{D}^{-1}(\hat{\boldsymbol{Q}}(\hat{\boldsymbol{K}}^\top \boldsymbol{V})), \qquad \text{where } \boldsymbol{D} = \text{diag}\left(\hat{\boldsymbol{Q}}(\hat{\boldsymbol{K}}^\top \mathbf{1})\right),$$

where the parentheses represent the order of matrix operations to obtain the time complexity $\Theta(nmd)$ and space complexity $\Theta(nm + nd + md)$.

3.4 We show that

$$\begin{aligned}
a_{ij} &= \exp\left(\frac{-\|\boldsymbol{q}_i\|^2}{2}\right) \cdot \mathbb{E}_{\boldsymbol{x}\in\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\exp(\boldsymbol{x}^\top \boldsymbol{q}_i)\exp(\boldsymbol{x}^\top \boldsymbol{k}_j)\right] \cdot \exp\left(\frac{-\|\boldsymbol{k}_j\|^2}{2}\right) \\
&= \exp\left(\frac{-\|\boldsymbol{q}_i\|^2}{2}\right) \cdot \left((2\pi)^{-d/2}\int_{\boldsymbol{x}} \exp\left(\frac{-\|\boldsymbol{x}\|^2}{2}\right)\exp(\boldsymbol{x}^\top \boldsymbol{q}_i)\exp(\boldsymbol{x}^\top \boldsymbol{k}_j)d\boldsymbol{x}\right) \cdot \exp\left(\frac{-\|\boldsymbol{k}_j\|^2}{2}\right) \\
&\qquad \text{(Hint 1)} \\
&= (2\pi)^{-d/2}\int_{\boldsymbol{x}} \exp\left(\frac{-\boldsymbol{q}_i^\top \boldsymbol{q}_i - \boldsymbol{k}_j^\top \boldsymbol{k}_j - \boldsymbol{x}^\top \boldsymbol{x} + 2\boldsymbol{x}^\top(\boldsymbol{q}_i + \boldsymbol{k}_j)}{2}\right) d\boldsymbol{x} \\
&= (2\pi)^{-d/2}\int_{\boldsymbol{x}} \exp\left(\frac{-\boldsymbol{q}_i^\top \boldsymbol{q}_i - \boldsymbol{k}_j^\top \boldsymbol{k}_j - \|\boldsymbol{x} - (\boldsymbol{q}_i + \boldsymbol{k}_j)\|^2 + (\boldsymbol{q}_i + \boldsymbol{k}_j)^\top(\boldsymbol{q}_i + \boldsymbol{k}_j)}{2}\right) d\boldsymbol{x} \\
&= (2\pi)^{-d/2}\exp\left(\frac{-\boldsymbol{q}_i^\top \boldsymbol{q}_i + (\boldsymbol{q}_i + \boldsymbol{k}_j)^\top(\boldsymbol{q}_i + \boldsymbol{k}_j) - \boldsymbol{k}_j^\top \boldsymbol{k}_j}{2}\right)\int_{\boldsymbol{x}} \exp\left(\frac{-\|\boldsymbol{x} - (\boldsymbol{q}_i + \boldsymbol{k}_j)\|^2}{2}\right) d\boldsymbol{x} \\
&= \exp\left(-\frac{1}{2}(\boldsymbol{q}_i^\top \boldsymbol{q}_i - (\boldsymbol{q}_i + \boldsymbol{k}_j)^\top(\boldsymbol{q}_i + \boldsymbol{k}_j) + \boldsymbol{k}_j^\top \boldsymbol{k}_j)\right) && \text{(Hint 1)} \\
&= \exp(\boldsymbol{q}_i^\top \boldsymbol{k}_j), && \text{(Hint 2)}
\end{aligned}$$

which is the value of a_{ij} in $\text{Attention}_{\text{std}}$.
Now, we want f such that

$$\begin{aligned}
a_{ij} &= \\
f(\boldsymbol{q}_i)^\top f(\boldsymbol{k}_j) &= \exp\left(\frac{-\|\boldsymbol{q}_i\|^2}{2}\right) \cdot \mathbb{E}_{\boldsymbol{x}\in\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\exp(\boldsymbol{x}^\top \boldsymbol{q}_i)\exp(\boldsymbol{x}^\top \boldsymbol{k}_j)\right] \cdot \exp\left(\frac{-\|\boldsymbol{k}_j\|^2}{2}\right)
\end{aligned}$$

We can design f such that m samples are used to estimate the expectation above.

$$f(\boldsymbol{z}) = \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{z}\|^2\right)}{\sqrt{m}} \begin{bmatrix} \exp(\boldsymbol{x}_1^\top \boldsymbol{z}) \\ \vdots \\ \exp(\boldsymbol{x}_m^\top \boldsymbol{z}) \end{bmatrix},$$

for any $\boldsymbol{z} \in \mathbb{R}^d$ and fixed vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
We have now worked through some of the core ideas of *Performers*, an efficient variant of Transformers.

3.5 The setting of m allows a trade-off between efficiency and the quality of approximation. Larger m requires more memory and computation but would result in lower variance in the estimation of $\boldsymbol{A}$.

Question 4 (4-5-6-6)**.** In this question, you will reconcile the relationship between L2 regularization and weight decay for the Stochastic Gradient Descent (SGD) and Adam optimizers. Imagine you are training a neural network (with learnable weights θ) with a loss function $L(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)})$, under two different schemes. The *weight decay* scheme uses a modified SGD update rule: the weights θ decay exponentially by a factor of λ. That is, the weights at iteration $i+1$ are computed as

$$\theta_{i+1} = \theta_i - \eta \frac{\partial L(f(\mathbf{x}^{(i)}, \theta_i), \mathbf{y}^{(i)})}{\partial \theta_i} - \lambda \theta_i$$

where η is the learning rate of the SGD optimizer. The *L2 regularization* scheme instead modifies the loss function (while maintaining the typical SGD or Adam update rules). The modified loss function is

$$L_{\text{reg}}(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)}) = L(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)}) + \gamma \|\theta\|_2^2$$

4.1 Prove that the *weight decay* scheme that employs the modified SGD update is identical to an *L2 regularization* scheme that employs a standard SGD update rule.

4.2 This question refers to the Adam algorithm as described in the lecture slide (also identical to Algorithm 8.7 of the deep learning book). It turns out that a one-line change to this algorithm gives us Adam with an L2 regularization scheme. Identify the line of the algorithm that needs to change, and provide this one-line modification.

4.3 Consider a "decoupled" weight decay scheme for the original Adam algorithm (see lecture slides, or equivalently, Algorithm 8.7 of the deep learning book) with the following two update rules.

- The **Adam-L2-reg** scheme computes the update by employing an L2 regularization scheme (same as the question above).
- The **Adam-weight-decay** scheme computes the update as $\boldsymbol{\Delta}\theta = -\left(\epsilon \frac{\hat{\mathbf{s}}}{\sqrt{\hat{\mathbf{r}}}+\delta} + \lambda\theta\right)$.

Now, assume that the neural network weights can be partitioned into two disjoint sets based on their gradient magnitude: $\theta = \{\theta_{\text{small}}, \theta_{\text{large}}\}$, where each weight $\theta_s \in \theta_{\text{small}}$ has a much smaller gradient magnitude than each weight $\theta_l \in \theta_{\text{large}}$. Using this information provided, answer the following questions. In each case, provide a brief explanation as to why your answer holds.

(a) Under the **Adam-L2-reg** scheme, which set of weights among θ_{small} and θ_{large} would you expect to be regularized (i.e., driven closer to zero) more strongly than the other ? Why ?

(b) Would your answer change for the **Adam-weight-decay** scheme ? Why/why not ?

(Note: for the two sub-parts above, we are interested in the rate at which the weights are regularized, *relative* to their initial magnitudes.)

4.4 In the context of all of the discussion above, argue that weight decay is a better scheme to employ as opposed to L2 regularization ; particularly in the context of adaptive gradient based optimizers. (Hint: think about how each of these schemes regularize each parameter, and also about what the overarching objective of regularization is).

Answer 4.

4.1 For the *L2 regularization* scheme, the update rule for standard SGD is

$$\theta_{i+1} = \theta_i - \eta \frac{\partial L_{\text{reg}}(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)})}{\partial \theta_i} = \theta_i - \eta \frac{\partial L(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)})}{\partial \theta_i} - 2\eta\gamma\theta_i$$

This is identical to the update rule under the weight decay scheme when $\lambda = 2\eta\gamma$.

4.2 The "*Compute gradient*" step needs to be changed to

$$\mathbf{g} \leftarrow \frac{1}{m} \nabla_\theta \sum_i L_{\text{reg}}(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)})$$

4.3 (a) The larger weights θ_{large} are regularized less strongly than the smaller weights θ_{small}. This is because Adam-L2-reg adapts the gradients of the L2 regularization loss term (which includes the gradients of the usual loss term plus the gradients of the L2 penalty over the weights). This results in the weights with larger gradient magnitudes being adapted less than the weights with smaller gradient magnitudes.

(b) In this case, each weight is adapted by the same amount λ.

4.4 The objective of a regularization term (or of weight decay) is to prevent weights from deviating strongly from their current values. In the context of Adam, L2 regularization results in each weight having it's own update rate based on it's gradient magnitude, which can be undesirable (because weights with low gradient magnitudes are not as strongly regularized, and can deviate from their previous values). However, weight decay ensures that each parameter is regularized uniformly, and achieves better regularization.