1. In the Transformer, positional embeddings enable a query and a key to match one another based on their relative position, rather than based on their content. The positional embedding is a $d_{\text{model}}$-dimensional vector whose $(2i)\text{th}$ and $(2i+1)\text{st}$ elements, at time $t$, are:

$$e_{2i}^t = \sin \left( \frac{t}{10000^{2i/d_{\text{model}}}} \right)$$

$$e_{2i+1}^t = \cos \left( \frac{t}{10000^{2i/d_{\text{model}}}} \right)$$

These are then added to the query and key prior to computing attention, thus

$$e^t = [e_0^t, e_1^t, \cdots, e_{d_{\text{model}}-1}^t]$$

$$E = \begin{bmatrix} e_0 \\ \vdots \\ e_{n-1} \end{bmatrix}$$

$$\text{head}_i = \text{Attention} \left( (Q + E)W_i^Q, (K + E)W_i^K, (V + E)W_i^V \right),$$

where the matrices $W_i^Q$ and $W_i^K$ each are of dimension $d_{\text{model}} \times d_k$.

Consider the matrix $A = W_i^K W_i^Q^T$. All parts of this problem will ask you to calculate input-output relations of the Attention operation by thinking about the values of the following submatrix:

$$A_{(2i;2i+1),(2j;2j+1)} = \begin{bmatrix} A_{2i,2j} & A_{2i,2j+1} \\ A_{2i+1,2j} & A_{2i+1,2j+1} \end{bmatrix},$$

(a) (1 point) What should be the values of $A_{(2i;2i+1),(2j;2j+1)}$ so that the attention, $\alpha_{\tau,t}$, is maximized when the time alignment of the key (the time index $\tau$ of vector $k^\tau$) precedes the time index of the query (time index $t$ of vector $q^t$) by exactly $T$ time steps ($\tau = t - T$)?

**Solution:** We want to maximize the dot product

$$k^\tau W_i^K u_i^Q, q^t = k^\tau A q^t$$

This is maximized when the vectors $k^\tau$ and $A q^t$ point in the same direction, i.e., we want

$$\begin{bmatrix} \sin \left( \frac{\tau}{10000^{2i/d_{\text{model}}}} \right) \\ \cos \left( \frac{\tau}{10000^{2i/d_{\text{model}}}} \right) \end{bmatrix} = \begin{bmatrix} A_{2i,2j} & A_{2i,2j+1} \\ A_{2i+1,2j} & A_{2i+1,2j+1} \end{bmatrix} \begin{bmatrix} \sin \left( \frac{t}{10000^{2i/d_{\text{model}}}} \right) \\ \cos \left( \frac{t}{10000^{2i/d_{\text{model}}}} \right) \end{bmatrix}$$
This can be accomplished by using trig identities such as
\[
\sin(\alpha - \beta) = \sin \alpha \cos \beta - \cos \alpha \sin \beta \\
\cos(\alpha - \beta) = \cos \alpha \cos \beta + \sin \alpha \sin \beta,
\]
where we set
\[
\alpha = \frac{t}{10000^{2i/d_{\text{model}}}} \\
\beta = \frac{T}{10000^{2i/d_{\text{model}}}}
\]
This gives us the solution:
\[
\begin{bmatrix}
A_{2i,2j} & A_{2i,2j+1} \\
A_{2i+1,2j} & A_{2i+1,2j+1}
\end{bmatrix}
= 
\begin{bmatrix}
\cos \beta - \sin \beta \\
\sin \beta & \cos \beta
\end{bmatrix}
\]

(b) (1 point) Suppose that you require a less-precise time alignment. Suppose that you want the attention to be maximized when \(t - \tau \in [T - \frac{B}{2}, T + \frac{B}{2}]\), i.e., the separation between \(\tau\) and \(t\) should be \(T \pm \frac{B}{2}\). This can be accomplished by starting with the \(A\) matrix you computed in part (a), and then zeroing some of its elements. Which elements of \(A\) should be zeroed so that \(\alpha_{\tau,t}\) is maximum for an \(t - \tau \in [T - \frac{B}{2}, T + \frac{B}{2}]\)?

**Solution:** The period of the sinusoids in dimensions \(2i\) and \(2i + 1\) is \(2\pi10000^{2i/d_{\text{model}}}\). Let’s say that the peak of a sinusoid has a width of about 1 radian, i.e., about \(10000^{2i/d_{\text{model}}}\) time steps. Thus we want to zero out any elements of \(A\) such that

\[
B > 10000^{2i/d_{\text{model}}}
\]

Solving, we find that we should zero out the elements of \(A\) such that

\[
i < \frac{d_{\text{model}} \log B}{2 \log(10000)}
\]

(c) (1 point) The previous parts have considered a head that cares about relative position. In this part, instead, consider a head that only cares whether or not a query and key have the same content. Suppose, for example, that for this head, \(A\) is the identity matrix. Then, assuming that \(|k^\tau| = 1\) and \(|q^t| = 1\), the inner product \(k^\tau A q^t : T = k^\tau q^t : T\) is maximized if \(k^\tau = q^t\).

Suppose that \(k^\tau = q^t\) and \(A\) is the identity matrix, but we also have to consider the contribution of the positional embeddings. Let’s define \(\tilde{q}\) and \(\tilde{k}\) to be the position-enhanced embeddings, thus

\[
\tilde{q}_{2i} = q_{2i} + e_{2i}^t \\
\tilde{k}_{2i} = k_{2i} + e_{2i}^\tau
\]

What are the mean and variance of the inner product \(\tilde{k}^\tau \tilde{q}^t : T\), assuming that \(k^\tau = q^t\), \(|k^\tau| = 1\) and \(|q^t| = 1\)?

Hint: what are the mean and variance of \(\sin \alpha \sin \beta\) if \(\alpha\) and \(\beta\) are each independent random variables uniformly distributed between 0 and \(2\pi\)? What are the mean and variance of \(k_{2i}^\tau e_{2i}^t\) if \(k_{2i}^\tau\) is a zero-mean, unit-variance random variable independent of \(e_{2i}^t\)? How many such terms are added together in order to compute the inner product \(\tilde{k}^\tau \tilde{q}^t : T\)?
Solution: The expected value and variance of \( \sin \alpha \sin \beta \), if \( \alpha \) and \( \beta \) are each independent random variables uniformly distributed between 0 and \( 2\pi \), are

\[
E[\sin \alpha \sin \beta] = E[\sin \alpha] E[\sin \beta] = 0
\]

\[
E[\sin^2 \alpha \sin^2 \beta] = E[\sin^2 \alpha] E[\sin^2 \beta] = \left( \frac{1}{2} \right) \left( \frac{1}{2} \right) = \frac{1}{4}
\]

The mean and variance of \( k^\tau_2 e^t_2 \) are

\[
E[k^\tau_2 \sin \beta] = E[k^\tau_2] E[\sin \beta] = 0
\]

\[
E[(k^\tau_2)^2 \sin^2 \beta] = E[(k^\tau_2)^2] E[\sin^2 \beta] = 1 \times \frac{1}{2} = \frac{1}{2}
\]

Thus

\[
E[k^\tau \tilde{q}^{t,T}] = k^\tau \tilde{q}^{t,T} + E[k^\tau e^t] + E[e^t \tilde{q}^{t,T}] + E[e^t e^t]
\]

\[
= k^\tau \tilde{q}^{t,T} = 1
\]

The variance of each element in the sum is

\[
\text{Var}
\left( \tilde{q}^{t_2}_2 \tilde{k}^{\tau_2}
\right) = \text{Var}
\left( k^\tau_2 \tilde{q}^{t_2}_2
\right) + \text{Var}
\left( k^\tau_2 e^{t_2}_2
\right) + \text{Var}
\left( e^{t_2}_2 \tilde{q}^{t_2}_2
\right) + \text{Var}
\left( e^{t_2}_2 e^{t_2}_2
\right)
\]

\[
= 0 + \frac{1}{2} + \frac{1}{2} + \frac{1}{4}
\]

\[
= \frac{5}{4}
\]

and so the total variance is \( \frac{5}{4} d_{\text{model}} \).