This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

3.a

In the notation of the problem, for one i, the value of $k_i^\top q$ must be large relative to all other dot products $k_j^\top q$ for $j \neq i$ so that

$$\alpha_i = \frac{\exp(k_i^\top q)}{\sum_{j=1}^{n} \exp(k_j^\top q)}$$

is approximately 1 and all other α_j are close to 0.

3.b

We can set $q = \log(t)(k_a + k_b)$, where $t \gg n$. Since the k_i are pairwise orthogonal and unit length, we have $k_i^\top q = 0$ for $i \notin \{a, b\}$ and $k_i^\top q = \log(t)$ otherwise. Then we obtain an explicit expression for the α_i:

$$\alpha_i = \begin{cases} \frac{1}{2t+n-2} & i \notin \{a, b\} \\ \\ \frac{t}{2t+n-2} & i \in \{a, b\} \end{cases}.$$

With our condition that $t \gg n$, we get $\alpha_a = \alpha_b \approx 1/2$ while the remaining $\alpha_i \approx 0$.

3.c

(i) A similar expression as in (b), namely $q = \log(t)(u_a + u_b)$ with $t \gg n$, will work in nearly the same way. Since the covariance matrices for the k_i are almost 0, the sampled vectors will essentially form an orthonormal set of vectors up to some negligible noise, hence form the desired query vector by the same reasoning as in 2 (b).

(ii) If $\Sigma_a = \alpha\mathbf{I} + \frac{1}{2}(u_a u_a^\top)$ with $\alpha \approx 0$, then k_a will roughly vary in magnitude along the direction of u_a, again up to some negligible noise. The only dot product that will be affected is $k_a^\top q$, which will weight the output vector c more towards u_a if k_a is scaled positively and greater than 1, or more towards k_b otherwise.

3.d

(i) We can set $q_1 = \log(s)u_a$ and $q_2 = \log(t)u_b$ with $s, t \gg n$. If the α_i are the attention weights corresponding to q_1 and the β_i are the weights corresponding to q_2, then we see

$$\alpha_i \approx \begin{cases} 0 & i \neq a \\ 1 & i = a \end{cases}$$

and

$$\beta_i \approx \begin{cases} 0 & i \neq b \\ 1 & i = b \end{cases}$$

by the orthonormality properties (up to small noise) of the $k_i \approx u_i$. This means $c_1 \approx v_a$ and $c_2 \approx v_b$, so their average is $c \approx \frac{1}{2}(v_a + v_b)$.

(ii) Even with $\Sigma_a = \alpha\mathbf{I} + \frac{1}{2}(u_a u_a^\top)$, we still obtain $c \approx \frac{1}{2}(v_a + v_b)$ with the same query vectors as in (i). The issue in part (c) (ii) no longer applies since we are able to separate the the single query vector into two query vectors focusing attention on v_a and v_b separately. Now the averages are applied to the outputs, rather than trying to control a single query to give equal attention to v_a and v_b. The added attention head provides a safeguard against a more noisy key.

3.e

(i) The vector c_2 approximates $v_2 = x_2 = u_2$. Since the u_i are pairwise orthogonal, we get

$$\alpha_{2j} = \begin{cases} \frac{\exp(\beta^2)}{\exp(\beta^2)+2} & j = 2 \\ \\ \frac{1}{\exp(\beta^2)+2} & j \neq 2 \end{cases}.$$

Since $\beta \gg 0$, we see $\alpha_{22} \approx 1$ so $c_2 \approx v_2 = x_2 = u_2$. If we were to set $x_2 = u_a + u_d$, we would instead get attention scores

$$\alpha_{2j} = \begin{cases} \frac{\exp(\beta^2)}{\exp(2\beta^2)+\exp(\beta^2)+1} & j = 1 \\ \\ \frac{\exp(2\beta^2)}{\exp(2\beta^2)+\exp(\beta^2)+1} & j = 2 \\ \\ \frac{1}{\exp(2\beta^2)+\exp(\beta^2)+1} & j = 3 \end{cases}.$$

Since the denominators are square in $\exp(\beta^2)$, the attention weight α_{21} will approach 0 as β gets large, so c_2 will still approximate $v_2 = x_2$, which is now $u_a + u_d$ but still not u_b. A similar calculation shows that setting $x_2 = u_a + u_c$ will make $c_2 \approx x_2 = u_a + u_c$.

(ii) Set $V = (1/\beta^2)(u_b u_b^\top - u_c u_C^\top)$. Then (to save clutter we do not expand x_i, but the penultimate equalities follow from the orthogonality properties of the u_i)

$$v_1 = Vx_1 = (1/\beta^2)(u_b u_b^\top - u_c u_c^\top)x_1 = (1/\beta^2)(u_b u_b^\top x_1 - u_c u_c^\top x_1) = (1/\beta^2)(u_b \beta^2) = u_b,$$

and

$$v_3 = Vx_3 = (1/\beta^2)(u_b u_b^\top - u_c u_c^\top)x_3 = (1/\beta^2)(u_b u_b^\top x_3 - u_c u_c^\top x_3) = (1/\beta^2)(u_b \beta^2 - u_c \beta^2) = u_b - u_c.$$

(Not needed but note $v_2 = 0$ since $x_2 = u_a$ is orthogonal to u_b and u_c.)

We want to construct K and Q such that the following hold: $k_1^\top q_1 = 0$, $k_2^\top q_1 = 0$, $k_3^\top q_1 = \beta^2$, $k_1^\top q_2 = \beta^2$, $k_2^\top q_2 = 0$, and $k_3^\top q_2 = 0$. If these equalities hold, we will get our desired attention scores for c_1 and c_2. We can accomplish this by setting $K = I$, the identity matrix, and $Q = (1/\beta^2)(u_c u_d^\top + u_d u_a^\top)$. Then

$$q_1 = Qx_1 = (1/\beta^2)(u_c u_d^\top + u_d u_a^\top)x_1 = (1/\beta^2)(u_c u_d^\top x_1 + u_d u_a^\top x_1) = (1\beta^2)u_c \beta^2 = u_c$$

and

$$q_2 = Qx_2 = (1/\beta^2)(u_c u_d^\top + u_d u_a^\top)x_2 = (1/\beta^2)(u_c u_d^\top x_2 + u_d u_a^\top x_2) = (1\beta^2)u_d \beta^2 = u_d.$$

We verify that we obtain the desired dot products:

$$\begin{aligned} k_1^\top q_1 &= x_1^\top u_c = (u_d + u_b)^\top u_c = 0 \\ k_2^\top q_1 &= x_2^\top u_c = u_a^\top u_c = 0 \\ k_3^\top q_1 &= x_3^\top u_c = (u_c + u_b)^\top u_c = \beta^2 \\ k_1^\top q_2 &= x_1^\top u_d = (u_d + u_b)^\top u_d = \beta^2 \\ k_2^\top q_2 &= x_2^\top u_d = u_a^\top u_d = 0 \\ k_3^\top q_2 &= x_3^\top u_d = (u_c + u_b)^\top u_d = 0. \end{aligned}$$

From these, we can compute our attention scores:

$$\alpha_{1j} = \begin{cases} \frac{\exp(\beta^2)}{\exp(\beta^2)+2} & j = 3 \\ \\ \frac{1}{\exp(\beta^2)+2} & j \neq 3 \end{cases}$$

$$\alpha_{2j} = \begin{cases} \frac{\exp(\beta^2)}{\exp(\beta^2)+2} & j = 1 \\ \\ \frac{1}{\exp(\beta^2)+2} & j \neq 1 \end{cases}.$$

Since $\beta \gg 0$, we get the desired approximations $c_1 \approx v_3 = u_b - u_c$ and $c_2 \approx v_1 = u_b$.