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.

1.a

First we recall some basic facts about vector derivatives. Let $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ be an n-dimensional vector of variables, let $f(x_1, x_2, \ldots, x_n, \ldots)$ be a differentiable function of the x_i for all i and possibly some other inputs, and let $\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$ be an arbitrary vector in $\mathbb{R}^n$. Then the definition of $\frac{\partial f}{\partial \mathbf{x}}$ is

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}.$$

Using the definition, we see that

$$\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \left(\sum_{i=1}^n a_i x_i\right)}{\partial \mathbf{x}} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}.$$

Now lets start from the given calculation:

$$\boldsymbol{J}_{\textsf{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\hat{y}_o) = -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log\left(\sum_{w \in \textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\right).$$

Using the chain-rule and the fact above, we see that

$$\begin{aligned} \frac{\partial J}{\partial \boldsymbol{v}_c} &= -\boldsymbol{u}_o + \left(\sum_{w \in \textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\right)^{-1} \left(\sum_{w \in \textsf{Vocab}} \boldsymbol{u}_w \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\right) \\ &= -\boldsymbol{u}_o + \sum_{s \in \textsf{Vocab}} \frac{\boldsymbol{u}_s \exp(\boldsymbol{u}_s^\top \boldsymbol{v}_c)}{\sum_{w \in \textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}. \end{aligned}$$

By definition we have

$$\hat{y}_s = P(O = s | C = c) = \frac{\exp(\boldsymbol{u}_s^\top \boldsymbol{v}_c)}{\sum_{w \in \textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)},$$

so it follows

$$\frac{\partial J}{\partial \boldsymbol{v}_c} = -\boldsymbol{u}_o + \sum_{s \in \textsf{Vocab}} \frac{\boldsymbol{u}_s \exp(\boldsymbol{u}_s^\top \boldsymbol{v}_c)}{\sum_{w \in \textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} = -\boldsymbol{u}_o + \sum_{s \in \textsf{Vocab}} \hat{y}_s u_s,$$

establishing equation (4).

Let u_{ij} be the ij-th entry of $\boldsymbol{U}$, and let V be the size of the vocabulary. Note the k-th entry of $\boldsymbol{u}_w$ is u_{kw}. Using this, the i-th entry of $\boldsymbol{U}\hat{\boldsymbol{y}}$ is

$$\boldsymbol{U}\hat{\boldsymbol{y}}_i = \sum_{w=1}^V \hat{y}_w u_{iw}.$$

This is precisely the i-th entry of $\sum_{w=1}^{V} \hat{y}_w \boldsymbol{u}_w$. Finally, since $\boldsymbol{y}$ is a one-hot vector with 1 in the o-th position, we have

$$-\boldsymbol{U}\boldsymbol{y} = -\begin{bmatrix} u_{1o} \\ \vdots \\ u_{do} \end{bmatrix} = -\boldsymbol{u}_o,$$

where d is the dimension of the word embeddings, which establishes equation (3) from equation (4).

1.b

First, let's recall the definition of a function with respect to a matrix of variables. If $\mathbf{X}$ is an $m \times n$ matrix of variables where x_{ij} is the ij-th entry of $\mathbf{X}$, and if f is a differentiable function of the x_{ij}, then

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}.$$

Note that the derivative identity computed in part (a)

$$\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

also holds when taking the transpose of the dot product

$$\frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$$

since the expression for the dot product is identical. Then, using this identity and the chain-rule (like in part (a)), we can compute the derivatives $\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_w}$ from the expression

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log\left(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\right).$$

- Case 1: $w \neq o$
  Then the first term drops, and by the chain-rule we are left with

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_w} = \boldsymbol{v}_c \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} = \boldsymbol{v}_c \hat{y}_w.$$

- Case 2: $w = o$
  The computation is the same as in the first case, but now the derivative of the first term gives $-\boldsymbol{v}_c$, so we get

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_w} = -\boldsymbol{v}_c + \boldsymbol{v}_c \hat{y}_w = (\hat{y}_w - 1)\boldsymbol{v}_c$$

This establishes equation (7).

Let V be the size of the vocabulary. It follows from the definition of the matrix derivative that

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{U}} = \begin{bmatrix} \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_1} & \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_2} & \cdots & \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_V} \end{bmatrix}.$$

By equation (7), we have

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{U}} = \begin{bmatrix} \hat{y}_1 \boldsymbol{v}_c & \hat{y}_2 \boldsymbol{v}_c & \cdots & \hat{y}_{o-1} \boldsymbol{v}_c & (\hat{y}_o - 1)\boldsymbol{v}_c & \hat{y}_{o+1} \boldsymbol{v}_c & \cdots & \hat{y}_V \boldsymbol{v}_c \end{bmatrix},$$

which is the outer product $\boldsymbol{v}_c(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top$, establishing equation (6).