2.b

The cross-entropy is

$$\begin{aligned} CE(\mathbf{y}, \hat{\mathbf{y}}) &= -\sum_i y_i \log(\hat{y}_i) \\ &= -y_i \log(\hat{y}_i), \end{aligned}$$

where in the second line, the summation symbol has been dropped because in Einstein notation the sum is *implied* by the repeated indices *i*. Now we can find the gradient of CE by taking the partial with respect to θ_k.

$$
\begin{aligned}
\frac{\partial CE}{\partial \theta_k} &= -\frac{\partial}{\partial \theta_k} y_i \log(\hat{y}_i) \\
&= -y_i \frac{\partial}{\partial \theta_k} \log(\hat{y}_i) \\
&= -y_i \frac{1}{\hat{y}_i} \frac{\partial}{\partial \theta_k} \hat{y}_i \\
&= -y_i \frac{1}{\hat{y}_i} \frac{\partial}{\partial \theta_k} \frac{e^{\theta_i}}{e^{\theta_m}} \\
&= -y_i \frac{1}{\hat{y}_i} \left[\frac{\frac{\partial}{\partial \theta_k} e^{\theta_i}}{e^{\theta_m}} - \frac{e^{\theta_i}}{(e^{\theta_m})^2} \frac{\partial}{\partial \theta_k} e^{\theta_p} \right] \\
&= -y_i \frac{1}{\hat{y}_i} \left[\frac{e^{\theta_i} \delta_{ik}}{e^{\theta_m}} - \frac{e^{\theta_i}}{(e^{\theta_m})^2} e^{\theta_p} \delta_{pk} \right] \\
&= -y_i \frac{1}{\hat{y}_i} \left[\hat{y}_i \delta_{ik} - \hat{y}_i \hat{y}_p \delta_{pk} \right] \\
&= -y_i \left[\delta_{ik} - \hat{y}_p \delta_{pk} \right]
\end{aligned}
$$

Notice that in the very last line above, the $\hat{y}_i$ terms factored out of the brackets and were cancelled by the $\frac{1}{\hat{y}_i}$ term out front. At this point, you've done all the derivatives so it's just a matter of being careful with your indices as you pop off the delta functions. Since *k* is the subscript wrt which we're taking the derivative, that's the only subscript that should remain at the end of our simplification; all other indices will be summed over somehow. First let's pop off the *p* index, distribute the y_i through, and then pop the *i* index:

$$
\begin{aligned}
&= -y_i \Big[\delta ik - \hat{y}_k \Big] \\
&= -y_i \delta ik + y_i \hat{y}_k \\
&= -y_k + y_i \hat{y}_k
\end{aligned}
$$

Now remember that in Einstein notation all summations are implied by either repeated indices within each term **of a summation** or by the dimensionality of the final answer. We know that the final answer must be a vector indexed only by *k*, so we leave the k's alone, but what's that *i* doing in there? It's an implied sum! To abandon Einstein for a moment and make the sum explicit, this is equivalent to

$$
\begin{aligned}
&= -y_k + \sum_i y_i \hat{y}_k \\
&= -y_k + \hat{y}_k \sum_i y_i \\
&= -y_k + \hat{y}_k .
\end{aligned}
$$

In the first line above we write out the implied summations explicitly. The second line factors out the $\hat{y}_k$ because it has no dependence on i. The final line evaluates the sum over all the elements of the one-hot-vector, which is equal to one. Now we've just derived the gradient with respect to an arbitrary index k. Thus this relationship holds for *all* indices and can be written in vector form:

$$
\frac{\partial CE}{\partial \theta} = \hat{\mathbf{y}} - \mathbf{y}.
$$

2.c

Here's the equations I'm starting with. It's too much for me to write out all the vectors as bold so just know which symbols are scalars / vectors / matrices by context.

$$
\begin{aligned}
t &= xW^{(1)} + b^{(1)} \\
h &= \sigma(t) \\
v &= hW^{(2)} + b^{(2)} \\
\hat{y} &= \mathbf{softmax}(v) = \psi(v) \\
J &= -y_i log(\hat{y}_i)
\end{aligned}
$$

Now the full derivative of J wrt x, using the chain rule and temporarily ignoring the subscripts / indices:

$$
\frac{\partial J}{\partial x} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v} \frac{\partial v}{\partial h} \frac{\partial h}{\partial t} \frac{\partial t}{\partial x}
$$

Okay now that we've written it out, put in the subscripts:

$$
\frac{\partial J}{\partial x_k} = \frac{\partial J}{\partial \hat{y}_a} \frac{\partial \hat{y}_a}{\partial v_b} \frac{\partial v_b}{\partial h_c} \frac{\partial h_c}{\partial t_d} \frac{\partial t_d}{\partial x_k}
$$

Now the problem reduces to finding each vector / matrix term in the chain rule equation. Let's start from the left:

$$\begin{aligned}\frac{\partial J}{\partial \hat{y}_a} &= -y_i \frac{1}{\hat{y}_i} \frac{\partial \hat{y}_i}{\partial \hat{y}_a} \\ &= -\frac{y_i}{\hat{y}_i} \delta_{ia} \\ &= -\frac{y_a}{\hat{y}_a}\end{aligned}$$

Note that for this term, in the last line the δ_{ia} is evaluated because the index that it is removing (i) is isolated within that term; it will not appear anywhere else in the equation and so δ_{ia} will only affect this term. This will NOT be the case for subsequent deltas. Next...

$$\begin{aligned}\frac{\partial \hat{y}_a}{\partial v_b} &= \frac{\partial}{\partial v_b} \psi(v_a) \\ &= \hat{y}_a \delta_{ab} - \hat{y}_a \hat{y}_b\end{aligned}$$

I skipped all the steps here because this is identical work from problem 2.b. This delta (δ_{ab}) must be kept for now because it connects the two indices from the LHS of the equation; these indices will appear in other multiplicative terms. Next...

$$\begin{aligned}\frac{\partial v_b}{\partial h_c} &= \frac{\partial}{\partial h_c} \left[h_i W^{(2)}_{ib} + b_b \right] \\ &= W^{(2)}_{ib} \frac{\partial h_i}{\partial h_c} \\ &= W^{(2)}_{ib} \delta_{ic} \\ &= W^{(2)}_{cb}\end{aligned}$$

Don't get confused by my use of the index i. Since we eliminated i earlier when calculating $\frac{\partial J}{\partial \hat{y}_a}$, the index i is therefore free to use again. δ_{ic} was evaluated here because the index that it was eliminating (i) was internal to this term (doesn't appear on the LHS and doesn't appear in any other multiplicative terms) and thus would not affect any other terms we're computing. Next...

$$\begin{aligned}\frac{\partial h_c}{\partial t_d} &= \frac{\partial}{\partial t_d}\sigma(t_c) \\ &= \sigma(t_c)(1-\sigma(t_c))\frac{\partial t_c}{\partial t_d} \\ &= \sigma(t_c)(1-\sigma(t_c))\delta_{cd} \\ &= \sigma'(t_c)\delta_{cd}\end{aligned}$$

In the above calculation, I'm making use of the result from 2.a. Don't forget to do the chain rule on the argument of σ, because that gives you a delta function you need. Last term...

$$\frac{\partial t_d}{\partial x_k} = W^{(1)}_{kd}$$

Is basically just a repeat of the work from $\frac{\partial v_b}{\partial h_c}$. Now we put all the terms together.

$$\begin{aligned}\frac{\partial J}{\partial x_k} &= \frac{\partial J}{\partial \hat{y}_a}\frac{\partial \hat{y}_a}{\partial v_b}\frac{\partial v_b}{\partial h_c}\frac{\partial h_c}{\partial t_d}\frac{\partial t_d}{\partial x_k} \\ &= -\frac{y_a}{\hat{y}_a}\left[\hat{y}_a\delta_{ab} - \hat{y}_a\hat{y}_b\right]W^{(2)}_{cb}\sigma'(t_c)\delta_{cd}W^{(1)}_{kd} \\ &= -y_a\left[\delta_{ab} - \hat{y}_b\right]W^{(2)}_{cb}\sigma'(t_c)\delta_{cd}W^{(1)}_{kd} \\ &= -y_a\delta_{ab}W^{(2)}_{cb}\sigma'(t_c)\delta_{cd}W^{(1)}_{kd} + y_a\hat{y}_bW^{(2)}_{cb}\sigma'(t_c)\delta_{cd}W^{(1)}_{kd}\end{aligned}$$

At this point, everything has been expanded so that we can see all of the individual terms. This is important because the implied sums and the kronecker deltas apply per summation term only. Now let's apply δ_{ab} in the first term and sum over the one-hot y_a in the second term (like we did in prob 2.b), then recombine the y terms, and apply δ_{cd}.

$$
\begin{aligned}
&= -y_b W^{(2)}_{cb} \sigma'(t_c) \delta_{cd} W^{(1)}_{kd} + \hat{y}_b W^{(2)}_{cb} \sigma'(t_c) \delta_{cd} W^{(1)}_{kd} \\
&= \left[\hat{y}_b - y_b\right] W^{(2)}_{cb} \sigma'(t_c) \delta_{cd} W^{(1)}_{kd} \\
&= \left[\hat{y} - y\right]_b W^{(2)}_{cb} \sigma'(t_c) \delta_{cd} W^{(1)}_{kd} \\
&= \left[\hat{y} - y\right]_b W^{(2)}_{cb} \sigma'(t_c) W^{(1)}_{kc}
\end{aligned}
$$

Now the final step is to convert this back to vector notation. We know the outside index is k, so we'll want to transpose $W^{(1)}$ so that k is outside and so that c aligns with the $\sigma'(t_c)$. Will also want to transpose $W^{(2)}$ so that the b and c indices align with the terms on either side of it. Thus

$$
\frac{\partial J}{\partial x_k} = \left[\hat{y} - y\right]_b W^{(2)T}_{bc} \sigma'(t_c) W^{(1)T}_{ck}.
$$

Now all of the repeated indices are sequential / next to each other and the outside index k is the index from the LHS. This is tricky to piece back together on account of the c index appearing thrice. But you can convince yourself that $\left[\hat{y} - y\right]_b W^{(2)T}_{bc} = \left[\left(\hat{y} - y\right) W^{(2)T}\right]_c$ is a vector indexed by c. Further, the c^{th} element of this vector must also be multiplied by the c^{th} element of the vector $\sigma'(t)$, which would be the hadamard

product:

$$\left[\left(\hat{y}-y\right)W^{(2)T}\right]c\sigma'(t_c) = \left[\left(\hat{y}-y\right)W^{(2)T} \circ \sigma'(t)\right]c$$

If you've followed the solution this far, then it's apparent that the final step leads to the answer:

$$\frac{\partial J}{\partial x_k} = \left(\hat{y}-y\right)W^{(2)T} \circ \sigma'(t)W^{(1)T}.$$

2.g

Now we have to compute the gradients with respect to the weights and biases.

With respect to $W^{(2)}$

Note that I'm going to drop the superscript on W and b for this calculation, however it is implied and will be reintroduced at the end. First write out the gradient with the chain rule.

$$\frac{\partial J}{\partial W_{cd}} = \frac{\partial J}{\partial \hat{y}_a}\frac{\partial \hat{y}_a}{\partial v_b}\frac{\partial v_b}{\partial W_{cd}}$$

We've already calculated the first two terms in the chain rule, so simply need the ultimate term.

$$
\begin{aligned}
\frac{\partial v_b}{\partial W_{cd}} &= \frac{\partial}{\partial W_{cd}} \left[hW + b \right]_b \\
&= \frac{\partial}{\partial W_{cd}} \left[h_i W_{ib} + b_b \right] \\
&= h_i \frac{\partial W_{ib}}{\partial W_{cd}} + 0 \\
&= h_i \delta_{ic} \delta_{bd} \\
&= h_c \delta_{bd}
\end{aligned}
$$

Now plug that back in

$$
\begin{aligned}
\frac{\partial J}{\partial W_{cd}} &= \frac{\partial J}{\partial \hat{y}_a} \frac{\partial \hat{y}_a}{\partial v_b} \frac{\partial v_b}{\partial W_{cd}} \\
&= \left[\hat{y}_b - y_b \right] h_c \delta_{bd} \\
&= \left[\hat{y}_d - y_d \right] h_c \\
&= h_c \left[\hat{y}_d - y_d \right]
\end{aligned}
$$

In the second line I made use of the fact that we've already calculated $\frac{\partial J}{\partial \hat{y}_a} \frac{\partial \hat{y}_a}{\partial v_b}$ twice, two different ways, thus far in this homework so I'm just going to use the result from now on. The final result here is actually a matrix formed by the outer product of h with $\hat{y} - y$. There are two ways you could write it

$$
\begin{aligned}
\frac{\partial J}{\partial W^{(2)}} &= h \otimes \left[\hat{y} - y \right] \\
&= h^T \left[\hat{y} - y \right]
\end{aligned}
$$

I prefer the former tensor product notation, but in this case the second line reminds us that the vectors in this problem statement are actually *row vectors* instead of the more common convention of using column vectors.

With respect to $b^{(2)}$

This is substantially the same, only now we must calculate a different ultimate term in the chain rule

$$\begin{aligned}\frac{\partial v_b}{\partial b_c} &= \frac{\partial}{\partial b_c}\left[hW + b\right]_b \\ &= \frac{\partial}{\partial b_c}\left[h_i W_{ib} + b_b\right] \\ &= 0 + \frac{\partial b_b}{\partial b_c} \\ &= \delta_{bc}\end{aligned}$$

Now plug that in to the full chain rule

$$\begin{aligned}\frac{\partial J}{\partial b_c} &= \frac{\partial J}{\partial \hat{y}_a}\frac{\partial \hat{y}_a}{\partial v_b}\frac{\partial v_b}{\partial b_c} \\ &= \left[\hat{y}_b - y_b\right]\delta_{bc} \\ &= \left[\hat{y}_c - y_c\right]\end{aligned}$$

So it appears that the gradient of the bias is simply the error in the prediction

$$\frac{\partial J}{\partial b^{(2)}} = \hat{y} - y$$

With respect to $W^{(1)}$

Again I'll drop the superscript until the very end. Write out the chain rule, very much like from problem 2.c

$$\frac{\partial J}{\partial W_{fg}} = \frac{\partial J}{\partial \hat{y}_a}\frac{\partial \hat{y}_a}{\partial v_b}\frac{\partial v_b}{\partial h_c}\frac{\partial h_c}{\partial t_d}\frac{\partial t_d}{\partial W_{fg}}$$

Looking closely at this, the only term we haven't already calculated somewhere in this homework is the final one. Even the final term is effectively identical to the work we've already done wrt $W^{(2)}$. Being careful about index placement, we can see by inspection that

$$\frac{\partial t_d}{\partial W_{fg}} = x_f \delta_{dg}$$

Thus

$$\begin{aligned}
\frac{\partial J}{\partial W_{fg}} &= \left[\hat{y} - y\right]_b W_{cb} \sigma'(t_c) \delta_{cd} x_f \delta_{dg} \\
&= \left[\hat{y} - y\right]_b W_{cb} \sigma'(t_c) x_f \delta_{cg} \\
&= \left[\hat{y} - y\right]_b W_{gb} \sigma'(t_g) x_f \\
&= \left[\hat{y} - y\right]_b W^T_{bg} \sigma'(t_g) x_f \\
&= \left[(\hat{y} - y) W^T\right]_g \sigma'(t_g) x_f \\
&= \left[(\hat{y} - y) W^T \circ \sigma'(t)\right]_g x_f \\
&= x_f \left[(\hat{y} - y) W^T \circ \sigma'(t)\right]_g
\end{aligned}$$

In the above we pop the δ_{cd}, then pop δ_{cg}, then transpose W, then begin progressively converting back into matrix notation. The final line swaps the f and g terms so that they align with the order specified by the gradient (i.e. f comes first and g comes second in $\frac{\partial J}{\partial W_{fg}}$).

Converting wholly to matrix notation

$$\frac{\partial J}{\partial W^{(1)}} = x \otimes \left[(\hat{y} - y) W^{(2)T} \circ \sigma'(t) \right]$$

With respect to $b^{(1)}$

At this point, you might have noticed a pattern.

$$\frac{\partial J}{\partial b^{(1)}} = (\hat{y} - y) W^{T} \circ \sigma'(t)$$