# Alice's Adventures in a Differentiable Wonderland

Errata list

"I knew who I was this morning, but I've changed a few times since then."

*Chapter 5, Advice from a Caterpillar*

This is a list of corrections to the version currently available on arXiv.[1] I thank everyone who suggested these edits. Minor typos are not shown. I will periodically update the arXiv version to incorporate these changes.

## Corrections to v2

- **Chapter 4** (Box C.4.2): there was an unnecessary parameter (`w`).
- **Section 9.3.1** (*Dropout*): the role of $p$ and $1-p$ was inverted, so that *low drop* means $p = 0.8$ or $p = 0.9$, and:

$$\text{Dropout}(\mathbf{X}) = \begin{cases} \dfrac{\mathbf{M} \odot \mathbf{X}}{p} & \texttt{[training]} \\ \mathbf{X} & \texttt{[inference]} \end{cases}$$

- **Chapter 10** (Definition D.10.1): the number of parameters for self-attention is $2ke + ve$, not $k(2e + v)$.

[1] https://arxiv.org/abs/2404.17625

- **Chapter 12** (Laplacian): an index was missing on the definition of diffusion by the Laplacian:

$$[\mathbf{L}\mathbf{x}]_i = \sum_{(i,j)\in\mathcal{E}} A_{ij}(x_i - x_j) \tag{E.12.1}$$

- **Section 12.2.2** (*Graph convolutional layer*): the matrix $\mathbf{W}$ has shape $(c, c')$, not $(c', c)$.

## Corrections to v1

- **Chapter 3**, page 44: training set and test set should have an empty intersection, not union ($\mathcal{S}_n \cap \mathcal{T}_m = \emptyset$).

- **Chapter 6**: the indices in (E.6.6) are inverted and there is an extra term, the correct equation is:

$$\nabla^\top_{\mathbf{w}_i} y = \mathbf{1}^\top \left[\partial_{\mathbf{h}_{l-2}} \mathbf{h}_{l-1}\right] \cdots \left[\partial_{\mathbf{h}_i} \mathbf{h}_{i+1}\right] \left[\partial_{\mathbf{w}_i} \mathbf{h}_i\right] \tag{E.6.6}$$

- **Chapter 7**, Figure F.7.2: the rightmost pooled value (in red) should be 3.0, not 2.7.

- **Chapter 8**, Section 8.4.2: we parameterize each element $p(x_i \mid x_{:i}, c)$ of the product, not the entire product, so the correct equation is:

$$p(x_i \mid x_{:i}, c) \approx \text{Categorical}(x_i \mid f(x_{:i}, c))$$

- **Chapter 7**, Eq. (E.7.5), we can make the offset a function of the index in order to use it separately for $i$ and $j$:

$$t(i) = i - k - 1 \tag{E.7.5}$$

  With this notation, the equation for the convolution becomes:

$$H_{ijz} = \sum_{i'=1}^{2k+1} \sum_{j'=1}^{2k+1} \sum_{d=1}^{c} [W]_{i',j',z,d} [X]_{i'+t(i),j'+t(j),d}$$

- **Chapter 8**, page 138: the last generated value corresponds to the last input to the model:

$$\begin{bmatrix} - \\ - \\ - \\ \widehat{\mathbf{x}}_6 \end{bmatrix} = f\left(\begin{bmatrix} \mathbf{x}_2 \\ \widehat{\mathbf{x}}_3 \\ \widehat{\mathbf{x}}_4 \\ \widehat{\mathbf{x}}_5 \end{bmatrix}\right)$$

- **Chapter 11**, Eq. (E.11.2), the formula of cross-attention has a typo:

$$\text{CA}(\mathbf{X}, \mathbf{Z}) = \text{SA}(\mathbf{X}, \mathbf{Z}, \mathbf{Z}) = \text{softmax}\left(\frac{\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{Z}^\top}{\sqrt{k}}\right)\mathbf{Z}\mathbf{W}_v \qquad \text{(E.11.2)}$$

- **Chapter 12**, page 212: to make the polynomial layer clearer we remove self-loops from the adjacency matrix and write:

$$\mathbf{H} = \phi\left(\mathbf{X}\mathbf{W}_0 + \mathbf{A}\mathbf{X}\mathbf{W}_1 + \mathbf{A}^2\mathbf{X}\mathbf{W}_2\right)$$

  with three trainable parameters $\mathbf{W}_0$, $\mathbf{W}_1$, and $\mathbf{W}_2$ to handle self-loops, neighbors, and neighbors of neighbors respectively.

- **Appendix A**, Section A.1: most values in Table A.1 were inconsistent. In addition, Eq. (E.A.2) had a typo:

$$p(w) = \sum_r p(w, r) = \sum_r p(w \mid r)p(r) \qquad \text{(E.A.2)}$$