一、标量函数

1．N 层神经网络

$$Y_0 = X_0 = input$$

$$output = Y_n$$

$$Y_i = f_i(X_i) = f_i(A_i Y_{i-1} + B_i)$$

$$X_i = A_i Y_{i-1} + B_i$$

$$L = Loss = (Y_n - Y)^2 = (Y_n - Y)^T (Y_n - Y)$$

2．基本微分推导（×表示各元素分别相乘）

$$dL = d[(Y_n - Y)^T (Y_n - Y)]$$

$$dL = dY_n^T (Y_n - Y) + (Y_n - Y)^T d(Y_n - Y)$$

$$dL = 2(Y_n - Y)^T dY_n$$

$$dL = 2(Y_n - Y)^T df_n(A_n Y_{n-1} + B_n)$$

$$dL = 2(Y_n - Y)^T \left[\frac{\partial f_n}{\partial X_n} \times (dA_n Y_{n-1} + A_n dY_{n-1} + dB_n)\right]$$

$$dL = tr\left(2\left[(Y_n - Y) \times \frac{\partial f_n}{\partial X_n}\right]^T (dA_n Y_{n-1} + A_n dY_{n-1} + dB_n)\right)$$

3．第 N 层导数

$$K_n = 2(Y_n - Y) \times \frac{\partial f_n}{\partial X_n}$$

$$dL = tr(K_n^T dA_n Y_{n-1} + K_n^T A_n dY_{n-1} + K_n^T dB_n)$$

$$dL = tr(Y_{n-1} K_n^T dA_n + K_n^T A_n dY_{n-1} + K_n^T dB_n)$$

$$dL = tr((K_n Y_{n-1}^T)^T dA_n + (A_n^T K_n)^T dY_{n-1} + K_n^T dB_n)$$

$$\frac{\partial L}{\partial A_n} = K_n Y_{n-1}^T, \qquad \frac{\partial L}{\partial B_n} = K_n$$

4．第 n-1 层导数：上述dL中间项

$$dL_1 = (A_n^T K_n)^T dY_{n-1}, \qquad K_n = 2(Y_n - Y) \times \frac{\partial f_n}{\partial X_n}$$

$$dL_1 = (A_n^T K_n)^T \left[\frac{\partial f_{n-1}}{\partial X_{n-1}} \times (dA_{n-1} Y_{n-2} + A_{n-1} dY_{n-2} + dB_{n-1})\right]$$

$$dL_1 = \left(A_n^T K_n \times \frac{\partial f_{n-1}}{\partial X_{n-1}}\right)^T (dA_{n-1}Y_{n-2} + A_{n-1}dY_{n-2} + dB_{n-1})$$

$$K_{n-1} = A_n^T K_n \times \frac{\partial f_{n-1}}{\partial X_{n-1}} = (K_n^T A_n)^T \times \frac{\partial f_{n-1}}{\partial X_{n-1}}$$

$$\frac{\partial L}{\partial A_{n-1}} = K_{n-1}Y_{n-2}^T, \qquad \frac{\partial L}{\partial B_{n-1}} = K_{n-1}$$

5．第 i 层导数

$$K_n = 2(Y_n - Y) \times \frac{\partial f_n}{\partial X_n}$$

$$K_i = A_{i+1}^T K_{i+1} \times \frac{\partial f_i}{\partial X_i} = (K_{i+1}^T A_{i+1})^T \times \frac{\partial f_i}{\partial X_i}$$

$$\frac{\partial L}{\partial A_i} = K_i Y_{i-1}^T, \qquad \frac{\partial L}{\partial B_i} = K_{i-1}$$

6．函数导数 1（向量）

$$f_1(X) = \tanh(X), \qquad \frac{\partial f_1}{\partial X} = \frac{1}{\cosh^2(X)}$$

7．函数导数 2（矩阵）

$$f_1(X) = \tanh(X)$$

$$\frac{\partial f_1}{\partial X} = \frac{1}{\cosh^2(X)} = diag\left(\left[\frac{1}{\cosh^2(x_i)}\right]\right)$$

$$f_2(X) = softmax(X) = \left[\frac{e^{x_i}}{\sum e^x}\right]$$

$$\frac{\partial f_2}{\partial x_j} = \frac{e^{x_i}\sum e^x - e^{x_i}e^{x_j}}{(\sum e^x)^2} \; or \; \frac{-e^{x_i}e^{x_j}}{(\sum e^x)^2} = f_2(x_i)_i - f_2(x_i)_i f_2(x_i)_j \; or \; -f_2(x_i)_i f_2(x_i)_j$$

$$\frac{\partial f_2}{\partial X} = diag([f_2(X)]) - [f_2(X)_0 f_2(X), f_2(X)_1 f_2(X), \ldots, f_2(X)_n f_2(X)] =$$

$$\begin{bmatrix} f(X)_0 - f(X)_0 f(X)_0 & -f(X)_0 f(X)_1 & \cdots & -f(X)_0 f(X)_n \\ -f(X)_1 f(X)_0 & f(X)_1 - f(X)_1 f(X)_1 & \cdots & -f(X)_1 f(X)_n \\ \cdots & \cdots & \cdots & \cdots \\ -f(X)_n f(X)_0 & -f(X)_n f(X)_1 & \cdots & f(X)_n - f(X)_n f(X)_n \end{bmatrix}$$

二、矩阵函数

1．基本微分推导

$$dL = 2(Y_n - Y)^T df_n(X_n), \qquad X_n = A_n Y_{n-1} + B_n$$

$$dL = 2(Y_n - Y)^T \left(\frac{\partial f_n}{\partial X_n}\right)^{\mathrm{T}} vec(dX_n)$$

$$dL = 2(Y_n - Y)^T \left(\frac{\partial f_n}{\partial X_n}\right)^{\mathrm{T}} (dA_n Y_{n-1} + A_n dY_{n-1} + dB_n)$$

$$dL = tr\left(2\left(\frac{\partial f_n}{\partial X_n}(Y_n - Y)\right)^T (dA_n Y_{n-1} + A_n dY_{n-1} + dB_n)\right)$$

2．第 i 层导数（与上一章一致）

$$dL_1 = (A_n^T K_n)^T dY_{n-1}, \qquad K_n = 2\frac{\partial f_n}{\partial X_n}(Y_n - Y)$$

$$dL_1 = (A_n^T K_n)^T \left(\frac{\partial f_{n-1}}{\partial X_{n-1}}\right)^T (dA_{n-1} Y_{n-2} + A_{n-1} dY_{n-2} + dB_{n-1})$$

$$dL_1 = \left(\frac{\partial f_{n-1}}{\partial X_{n-1}} A_n^T K_n\right)^T (dA_{n-1} Y_{n-2} + A_{n-1} dY_{n-2} + dB_{n-1})$$

$$K_i = \frac{\partial f_{n-1}}{\partial X_{n-1}} A_n^T K_n = \frac{\partial f_{n-1}}{\partial X_{n-1}} (K_n^{\mathrm{T}} \mathrm{A_n})^{\mathrm{T}}$$

$$\frac{\partial L}{\partial A_i} = K_i Y_{i-1}^T, \qquad \frac{\partial L}{\partial B_i} = K_{i-1}$$