Gradient Notes

Christopher Yeh

April 8, 2024

1 Jacobian

Consider a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$. Then the Jacobian is the matrix

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}$$

where element-wise $\mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j}$.

If $f : \mathbb{R}^n \to \mathbb{R}$ is a scalar-valued function with vector inputs, then its gradient is just a special case of the Jacobian with shape $1 \times n$.

2 Softmax Cross-Entropy Loss w.r.t. Logits

We want to compute the gradient for the cross-entropy loss $J \in \mathbb{R}$ between our predicted softmax probabilities $\hat{y}$ and the true one-hot probabilities y. Both y and $\hat{y}$ are vectors of the same length. They can be either row or column vectors; the result is the same.

We are given the following:

1. $\hat{y} = \text{softmax}(\theta)$
2. y is a one-hot vector, where $y_k = 1$ and $y_{c \neq k} = 0$
3. y, $\hat{y}, \theta \in \mathbb{R}^n$

The cross-entropy loss J is computed as follows. The second line expands out the softmax function.

$$J(\theta) = CE(y, \hat{y}) = -\sum_c y_c \log \hat{y}_c = -\log \hat{y}_k$$
$$= -\log \frac{e^{\theta_k}}{\sum_c e^{\theta_c}} = \log\left(\sum_c e^{\theta_c}\right) - \theta_k$$

The gradient of the loss w.r.t. the logits θ is

$$\frac{\partial J}{\partial \theta_i} = \frac{e^{\theta_i}}{\sum_c e^{\theta_c}} - \mathbf{1}[i = k] \qquad \longrightarrow \qquad \boxed{\nabla_\theta J = \hat{y} - y}$$

3 Matrix times column vector w.r.t. matrix

Given $z = Wx$ and $r = \frac{\partial J}{\partial z}$, what is $\frac{\partial J}{\partial W}$?

1. $z \in \mathbb{R}^n$ and $x \in \mathbb{R}^m$ are column vectors
2. $W \in \mathbb{R}^{n \times m}$ is a matrix
3. $J \in \mathbb{R}$ is some scalar function of z
4. $r \in \mathbb{R}^{1 \times n}$ is the Jacobian of J w.r.t. z

Note on notation: Technically, J is a scalar-valued function taking nm inputs (the entries of W). This means the Jacobian $\frac{\partial J}{\partial W}$ should be a $1 \times nm$ vector, which isn't very useful. Instead, we will let $\frac{\partial J}{\partial W}$ be a $n \times m$ matrix where $\left(\frac{\partial J}{\partial W}\right)_{ij} = \frac{\partial J}{\partial W_{ij}}$.

$$\frac{\partial J}{\partial W} = \begin{bmatrix} \frac{\partial J}{\partial W_{1,1}} & \cdots & \frac{\partial J}{\partial W_{1,m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial J}{\partial W_{n,1}} & \cdots & \frac{\partial J}{\partial W_{n,m}} \end{bmatrix}$$

Naively, we can write

$$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial z}\frac{\partial z}{\partial W} = r\frac{\partial z}{\partial W}$$

However, it is unclear how to derive $\frac{\partial z}{\partial W}$, since this is the gradient of a vector w.r.t. a matrix. This gradient would have to be 3-dimensional, and multiplying the vector r by this 3-D tensor is not well-defined. Thus, we instead have to take the element-wise derivative $\frac{\partial J}{\partial W_{ij}}$.

Note that z_k is the dot-product between the k-th row of W and x.

$$z_k = \sum_{l=1}^m W_{kl} x_l \qquad\qquad \frac{\partial}{W_{ij}} z_k = \sum_{l=1}^m x_l \frac{\partial}{W_{ij}} W_{kl}$$

Clearly, $\frac{\partial}{W_{ij}} W_{kl} = 1$ only when $i = k$ and $j = l$, and 0 otherwise. Thus, $\frac{\partial}{W_{ij}} z_k = \mathbf{1}[k = i] x_j$. Another way we can write this is

$$\frac{\partial z}{W_{ij}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ x_j \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i\text{th element}$$

Now we can compute

$$\frac{\partial J}{\partial W_{ij}} = \sum_k \frac{\partial J}{\partial z_k} \frac{\partial z_k}{\partial W_{ij}} = \sum_k r_k \mathbf{1}[k = i] x_j = r_i x_j$$

where the summation comes from the Chain Rule. (Every change to W_{ij} influences each z_k which in turn influences J, so the total effect of W_{ij} on J is the sum of the influences of each z_k on J).

Thus the full matrix $\frac{\partial J}{\partial W}$ is the outer product $\frac{\partial J}{\partial W} = r^T x^T$ (recall that r is a row vector).

4 Row vector times matrix w.r.t. matrix

The problem setup is basically the same as the previous case, except with row vectors instead of column vectors.

Given $z = xW$ and $r = \frac{\partial J}{\partial z}$, what is $\frac{\partial J}{\partial W}$?

1. $z \in \mathbb{R}^{1 \times n}$ and $x \in \mathbb{R}^{1 \times m}$ are row vectors
2. $W \in \mathbb{R}^{m \times n}$ is a matrix
3. $J \in \mathbb{R}$ is some scalar function of z
4. $r \in \mathbb{R}^{1 \times n}$ is the Jacobian of J w.r.t. z

Similar to the previous case, we have

$$z_l = \sum_{k=1}^{n} x_k W_{kl}$$

$$\frac{\partial}{W_{ij}} z_l = \sum_{k=1}^{n} x_k \frac{\partial}{W_{ij}} W_{kl} = \mathbf{1}[j = l] x_i$$

Now we can compute

$$\frac{\partial J}{\partial W_{ij}} = \sum_l \frac{\partial J}{\partial z_l} \frac{\partial z_l}{\partial W_{ij}} = \sum_l r_l \mathbf{1}[j = l] x_i = x_i r_j$$

Thus the full matrix $\frac{\partial J}{\partial W}$ is the outer product $\frac{\partial J}{\partial W} = x^T r$ (recall that both x and r are row vectors).

5 Scalar Function of Matrix Multiplication w.r.t. Matrix

Let $B = XY$ be some matrix multiplication. Let A be a scalar that is a function of B, where $\frac{\partial A}{\partial B}$ is known. We want to find $\frac{\partial A}{\partial X}$ and $\frac{\partial A}{\partial Y}$.

1. Let $X \in \mathbb{R}^{n\times m}$ and $Y \in \mathbb{R}^{m\times p}$.
2. This means that $B, \frac{\partial A}{\partial B} \in \mathbb{R}^{n\times p}$.

Note on notation: Technically, A is a scalar-valued function taking np inputs (the entries of B). This means the Jacobian $\frac{\partial A}{\partial B}$ should be a $1 \times np$ vector, which isn't very useful. Instead, we will let $\frac{\partial A}{\partial B}$ be a $n \times p$ matrix where $\left(\frac{\partial A}{\partial B}\right)_{ij} = \frac{\partial A}{\partial B_{ij}}$. We define $\frac{\partial A}{\partial X}$ and $\frac{\partial A}{\partial Y}$ similarly.

Naively, we can write $\frac{\partial A}{\partial X} = \frac{\partial A}{\partial B}\frac{\partial B}{\partial X}$. However, it is unclear how to derive $\frac{\partial B}{\partial X}$, since this is the gradient of a matrix w.r.t. another matrix. This gradient would have to be 4-dimensional, and multiplying the matrix $\frac{\partial A}{\partial B}$ by this 4-D tensor is not well-defined. Thus, we instead take the element-wise derivative $\frac{\partial A}{\partial X_{ij}}$.

First, we compute the derivatives for each element of B w.r.t. each element of X and Y.

$$\frac{\partial}{\partial X_{i,j}} B_{k,l} = \frac{\partial}{\partial X_{i,j}} (X_{k,:} \cdot Y_{:,l}) = \mathbf{1}[k = i] Y_{j,l}$$
$$\frac{\partial}{\partial Y_{i,j}} B_{k,l} = \frac{\partial}{\partial Y_{i,j}} (X_{k,:} \cdot Y_{:,l}) = \mathbf{1}[l = j] X_{k,i}$$

Now we can use the (multi-path) chain rule to take the derivative of A w.r.t. each element of X and Y.

$$\frac{\partial A}{\partial X_{i,j}} = \sum_{k,l} \frac{\partial A}{\partial B_{k,l}} \frac{\partial B_{k,l}}{\partial X_{i,j}} = \sum_{k,l} \frac{\partial A}{\partial B_{k,l}} \mathbf{1}[k = i] Y_{j,l} = \sum_{l} \frac{\partial A}{\partial B_{i,l}} Y_{j,l} = \left(\frac{\partial A}{\partial B}\right)_{i,:} \cdot Y_{j,:}$$
$$\frac{\partial A}{\partial Y_{i,j}} = \sum_{k,l} \frac{\partial A}{\partial B_{k,l}} \frac{\partial B_{k,l}}{\partial Y_{i,j}} = \sum_{k,l} \frac{\partial A}{\partial B_{k,l}} \mathbf{1}[l = j] X_{k,i} = \sum_{k} \frac{\partial A}{\partial B_{k,j}} X_{k,i} = \left(\frac{\partial A}{\partial B}\right)_{:,j} \cdot X_{:,i}$$

Combining these element-wise derivatives yields the matrix equations

$$\boxed{\frac{\partial A}{\partial X} = \frac{\partial A}{\partial B} \cdot Y^T \qquad \text{and} \qquad \frac{\partial A}{\partial Y} = X^T \cdot \frac{\partial A}{\partial B}}$$

6 Scalar Function of Matrix-Vector Broadcast Sum w.r.t. Vector

Let A be a scalar that is a function of a matrix $B \in \mathbb{R}^{n \times m}$, and suppose $\frac{\partial A}{\partial B}$ is known. Let $B = X + y$ be some broadcasted sum between a matrix X and a row-vector $y \in \mathbb{R}^{1 \times m}$. We want to find $\frac{\partial A}{\partial y}$.

Intuitively, notice that any change in y directly and linearly affects every row of B. Each row of B in turn affects A. Therefore,

$$\frac{\partial A}{\partial y} = \sum_i \frac{\partial A}{\partial B_i} \frac{\partial B_i}{\partial y} = \sum_i \frac{\partial A}{\partial B_i} \cdot I = \sum_i \frac{\partial A}{\partial B_i}$$

where the B_i is the i-th row of B.

Alternatively, we can write this broadcasted sum properly as

$$B = X + \mathbf{1} \cdot y$$

where $\mathbf{1}$ is a n-dimensional column vector. Then we can use the gradient rules derived earlier to get the equivalent result.

$$\frac{\partial A}{\partial y} = \mathbf{1}^T \cdot \frac{\partial A}{\partial B} = \sum_i \frac{\partial A}{\partial B_i}$$

7 Scalar Function of Matrix-Vector Broadcast Product

Let A be a scalar that is a function of a matrix $B \in \mathbb{R}^{n \times m}$, and suppose $\frac{\partial A}{\partial B}$ is known. Let $B = y \cdot X$ be a broadcasted Hadamard (element-wise) product between a row vector $y \in \mathbb{R}^{1 \times m}$ and matrix X. In other words, the i-th row of B is computed by the Hadamard product $B_i = y \odot X_i$. We want to find $\frac{\partial A}{\partial y}$ and $\frac{\partial A}{\partial X}$.

We first find $\frac{\partial A}{\partial y}$. Intuitively, any change in y directly affects every row of B by a factor of the same row in X. Each row of B in turn affects A. Therefore,

$$\frac{\partial A}{\partial y} = \sum_i \frac{\partial A}{\partial B_i} \frac{\partial B_i}{\partial y} = \sum_i \frac{\partial A}{\partial B_i} \odot X_i$$

Next we find $\frac{\partial A}{\partial X}$. We can find this element-wise, then compose the entire gradient. Note that changing X_{ij} only affects B_{ij} by a scale of y_j. No other indices in B are affected.

$$\frac{\partial A}{\partial X_{ij}} = \frac{\partial A}{\partial B_{ij}} y_j$$
$$\frac{\partial A}{\partial X} = y \cdot \frac{\partial A}{\partial B}$$