BP 的暴力推导

袁振

日期：January 28, 2021

目录

1 BP 算法的暴力推导

1.1 前言

每个跑深度学习的人都知道 BP 算法（后向传播），却很少有人推导过数学公式，虽然推导数学公式需要时间、更需要一些技巧。不知道在其他地方，付出和收获是不是成比例，比如你买了彩票激动的等了一天而毫无所获，但是推导 BP 算法的过程会对以后看很多文章都有帮助。

1.2 Newton's method

先看一个简单的例子，在微积分里我们学过牛顿法找函数的零点，例如给定函数：

$$y = f(w) = aw^2 + bw + c, \quad w \in R$$

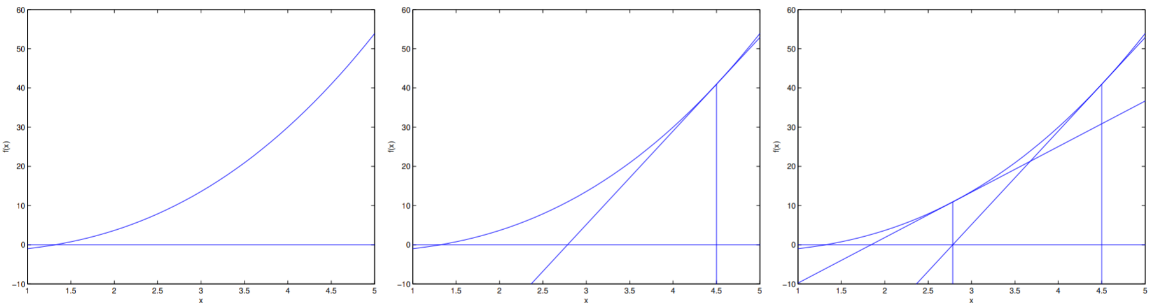

图 1.1: 牛顿法（图片来源）

在给定点 (w_0, y_0) 和其导数 $f'(w_0)$ 后，我们用线性拟合（如果函数是线性函数，便直接得到其零点）：

$$z = f'(w_0)(w - w_0) + y_0$$

来求得改拟合函数的零点：

$$w_1 = w_0 - \frac{f(w_0)}{f'(w_0)}$$

如果 $f(w_1) = 0$，我们的任务完成，否则我们在此基础上继续拟合。

这个过程和 BP 算法非常接近了，如果我们人为给定一个步长 α：

$$w_1 = w_0 - \alpha\frac{f(w_0)}{f'(w_0)}$$

这就是全世界口中的 BP 算法。但现实世界的情况远远没有这么简单，如果你只想简单的了解 BP 算法，下面的内容便不用看了。

1.3 MLP 的 BP 过程

假设我们的 MLP 模型为：

$$z = Ax + a, \quad A \in R^{m\times p}, a \in R^m \tag{1}$$

$$h = \sigma(z) \tag{2}$$

$$y = Bh + b, \quad B \in R^{n\times m}, b \in R^n \tag{3}$$

$$\text{Loss} = L = \frac{1}{2}\|y - t\|^2 \tag{4}$$

其中 A, a, B, b 是模型的参数。

1.4 最简单的方式

最简单的方式是将上面的公式不要写成矩阵和向量的形式，而是写成类似单变量的形式：

$$z_i = \sum_j A_{ij} x_j + a_i \tag{5}$$

$$h_i = \sigma(z_i) \tag{6}$$

$$y_k = \sum_i B_{ki} h_i + b_k \tag{7}$$

$$L = \frac{1}{2}\sum_k (y_k - t_k)^2 \tag{8}$$

于是我们有（非常快）：

$$
\begin{aligned}
\frac{\partial L}{\partial y_k} &= y_k - t_k && (9)\\
\Rightarrow \frac{\partial L}{\partial y} &= (y-t) && (10)\\
\frac{\partial L}{\partial B_{ki}} &= \frac{\partial L}{\partial y_k}\frac{\partial y_k}{\partial B_{ki}} && (11)\\
&= (y_k - t_k)h_i && (12)\\
\Rightarrow \frac{\partial L}{\partial B} &= (y-t)h^T && (13)\\
\frac{\partial L}{\partial b_k} &= \frac{\partial L}{\partial y_k}\frac{\partial y_k}{\partial b_k} && (14)\\
&= (y_k - t_k) && (15)\\
\Rightarrow \frac{\partial L}{\partial b} &= (y-t) && (16)\\
\frac{\partial L}{\partial h_i} &= \sum_k \frac{\partial L}{\partial B_k}\frac{\partial B_k}{\partial h_i} && (17)\\
&= \sum_k (y_k - t_k)B_{ki} && (18)\\
\Rightarrow \frac{\partial L}{\partial h} &= (y-t)^T B && (19)\\
\frac{\partial L}{\partial z_i} &= \frac{\partial L}{\partial h_i}\frac{\partial h_i}{\partial z_i} && (20)\\
&= (y-t)^T B\sigma'(z_i) && (21)\\
\Rightarrow \frac{\partial L}{\partial z} &= (y-t)^T B \odot \sigma'(z) && (22)\\
\frac{\partial L}{\partial A_{ij}} &= \frac{\partial L}{\partial z_i}\frac{\partial z_i}{\partial A_{ij}} && (23)\\
&= (y-t)^T B\sigma'(z_i)x_j && (24)\\
\Rightarrow \frac{\partial L}{\partial A} &= \left((y-t)^T B \odot \sigma'(z)\right) x^T && (25)\\
\frac{\partial L}{\partial a_i} &= \frac{\partial L}{\partial z_i}\frac{\partial z_i}{\partial a_i} && (26)\\
&= (y-t)^T B\sigma'(z_i) && (27)\\
\Rightarrow \frac{\partial L}{\partial a} &= \left((y-t)^T B \odot \sigma'(z)\right) && (28)\\
& && (29)
\end{aligned}
$$

当然还有更加难的一种方式。

1.5 一些简单的矩阵微积分

在得到成就感之前，枯燥的过程是避免不了的，尤其是直接硬邦邦的给一些看起来毫无意义的定义、定理。

定义 1.1.

$$y = f(\vec{x}), \quad \vec{x} \in R^n \tag{30}$$

$$\Rightarrow \frac{\partial y}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} \tag{31}$$

由此我们有：$y = \vec{a}^T \vec{x} \quad \Rightarrow \frac{\partial y}{\partial \vec{x}} = \vec{a}$

定义 1.2.

$$\vec{y} = f(x), \quad y \in R^m \tag{32}$$

$$\Rightarrow \frac{\partial \vec{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \dots & \frac{\partial y_m}{\partial x} \end{bmatrix} \tag{33}$$

定义 1.3.

$$\vec{y} = f(X), \quad X \in R^{m \times n} \tag{34}$$

$$\Rightarrow \frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ & \ddots & \\ \frac{\partial y}{\partial x_{m1}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix} \tag{35}$$

够了够了，我们有了以上这些定义，我们可以做很多的事情了，来让我们推导两个非常有用的公式。在以下的公式中，我们不会给向量用箭头表示，因为根据公式前后我们可以很明显看出。

$$y = x^T A x \quad \Rightarrow \frac{\partial \vec{y}}{\partial x} = Ax + A^T \tag{36}$$

$$\tag{37}$$

1.6 MLP 的 BP 过程

假设我们的 MLP 模型为：

$$z = Ax + a, \quad A \in R^{m \times p}, a \in R^m \tag{38}$$

$$h = \sigma(z) \tag{39}$$

$$y = Bh + b, \quad B \in R^{n \times m}, b \in R^n \tag{40}$$

$$\text{Loss} = L = \frac{1}{2}\|y - t\|^2 \tag{41}$$

其中 A, a, B, b 是模型的参数。

我们从 Loss 将依次求偏导到 A, a, B, b，首先我们有：

$$L = \frac{1}{2}(y - t)^T (y - t) \quad \Rightarrow \frac{\partial L}{\partial y} = (y - t)$$

1.6.1 对 B 的偏导

$$
\begin{aligned}
\frac{\partial L}{\partial B} &= \frac{1}{2}\frac{\partial (Bh+b-t)^T(Bh+b-t)}{\partial B} \\
&= \frac{1}{2}\frac{\partial h^TB^TBh + 2c^TBh + \text{常数}}{\partial B} && (\text{令 } c = b-t)
\end{aligned}
$$

我们令 $B = \begin{bmatrix} - & b_1^T & - \\ & \vdots & \\ - & b_n^T & - \end{bmatrix}$ 有：

$$
\begin{aligned}
\frac{\partial h^TB^TBh}{\partial B} &= \frac{\partial \begin{bmatrix} h^Tb_1 & \dots & h^Tb_n \end{bmatrix} \begin{bmatrix} h^Tb_1 \\ \vdots \\ h^Tb_n \end{bmatrix}}{\partial B} \\
&= \frac{\partial (h^Tb_1)^2 + \dots (h^Tb_n)^2}{\partial B} \\
\Rightarrow \frac{\partial h^TB^TBh}{\partial B_{ij}} &= \frac{\partial (h^Tb_i)^2}{\partial B_{ij}} \\
&= 2h_j(h^Tb_i) \\
&= 2(Bh)_ih_j \\
\Rightarrow \frac{\partial h^TB^TBh}{\partial B} &= 2Bhh^T
\end{aligned}
$$

我们还有：

$$
\begin{aligned}
2\frac{\partial c^TBh}{\partial B} &= 2\frac{\partial c^T \begin{bmatrix} h^Tb_1 \\ \vdots \\ h^Tb_n \end{bmatrix}}{\partial B} \\
&= 2\frac{\partial c_1h^Tb_1 + \dots c_nh^Tb_n}{\partial B} \\
\Rightarrow 2\frac{\partial c^TBh}{\partial B_{ij}} &= 2\frac{\partial c_ih^Tb_i}{\partial B_{ij}} \\
&= 2c_ih_j \\
\Rightarrow 2\frac{\partial c^TBh}{\partial B} &= 2ch^T
\end{aligned}
$$

综合起来，我们有：

$$
\begin{aligned}
\frac{\partial L}{\partial B} &= Bhh^T + ch^T \\
&= (y-t)h^T && (Bh + c = y - t)
\end{aligned}
$$

1.6.2 对 b 的偏导

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{1}{2}\frac{\partial (b+c)^T(b+c)}{\partial b} \\
&= \frac{1}{2}\frac{\partial b^T b + 2c^T b + \text{常数}}{\partial b} && (\text{令 } c = Bh - t) \\
&= y - t && (b + c = y - t)
\end{aligned}
$$

1.6.3 对 h 的偏导

如果想要得到 A, a 的偏导，我们必须要先解决 h，这和 b 过程一致，在这里我用同上来省略推导过程：

$$
\frac{\partial L}{\partial h} = B^T(y - t) \tag{同上}
$$

1.6.4 对 z 的偏导

至此，我们遇到了第一个让人困惑的问题，这个 σ 函数让人束手无策，如果我们写出来，问题便解决了：

$$
\sigma\left(\begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}\right) = \begin{bmatrix} \sigma(z_1) \\ \vdots \\ \sigma(z_m) \end{bmatrix}
$$

$$
\frac{\partial L}{\partial z} = \begin{bmatrix} \frac{\partial L}{\partial \sigma(z_1)} \\ \vdots \\ \frac{\partial L}{\partial \sigma(z_m)} \end{bmatrix} \odot \begin{bmatrix} \sigma'(z_1) \\ \vdots \\ \sigma'(z_m) \end{bmatrix}
$$

在此我们不继续推下去了，因为我怕转移我的注意力，再提示一遍，我们需要的是 A, a 的偏导。

算了，我还是推导一下，为了简便，我们设 $\sigma(z_i) = z_i^2$ 而再深度学习中，这个函数一般是：ReLU、Thanh 等等。为了更加简明，我们令 $\Sigma = \begin{bmatrix} \sigma(z_1) \\ \vdots \\ \sigma(z_m) \end{bmatrix}$

$$
\begin{aligned}
\frac{\partial L}{\partial z} &= \frac{1}{2}\frac{\partial \Sigma^T B^T B\Sigma + 2b^T B\Sigma}{\partial \Sigma} \odot \begin{bmatrix} \sigma'(z_1) \\ \vdots \\ \sigma'(z_m) \end{bmatrix} \\
&= (B^T B\Sigma + B^T b) \odot \begin{bmatrix} \sigma'(z_1) \\ \vdots \\ \sigma'(z_m) \end{bmatrix}
\end{aligned}
$$

我们不用管 $\begin{bmatrix} \sigma'(z_1) \\ \vdots \\ \sigma'(z_m) \end{bmatrix}$ 因为，这就是单变量的微分（逐一求导即可）。

1.6.5 对 A 的偏导

很重要的一点是，我们现在将 $\sigma(Ax+a)$ 看作 $(Ax+a)$ 即看作没有非线性变换，同样我们将 A, B 写作：$A = \begin{bmatrix} - & a_1^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix}$，$B = \begin{bmatrix} - & b_1^T & - \\ & \vdots & \\ - & b_n^T & - \end{bmatrix}$

$$
\begin{aligned}
\frac{\partial L}{\partial A} &= \frac{\partial \sigma^T(Ax+a)B^TB\sigma(Ax+a) + 2c^TB\sigma(Ax+a)}{\partial A} && (\text{令 } c = b - t) \\
&= \frac{\partial (Ax+a)^TB^TB(Ax+a) + 2c^TB(Ax+a)}{\partial A} \\
&= \frac{\partial x^TA^TB^TBAx + 2a^TB^TBAx + a^TB^TBa + 2c^TBAx + 2c^TBa}{\partial A} \\
&= \frac{\partial \left(x^TA^TB^TBA + 2a^TB^TBA + 2c^TBA\right)x}{\partial A} \\
&= \left[(b_1Axb_1 + \cdots + b_nAxb_n)\,x^T + B^TBax^T + B^Tcx^T\right] \\
&= \left[\left(b_1^TAxb_1 + \cdots + b_n^TAxb_n\right) + (a^Tb_1b_1 + \cdots + a^Tb_nb_n) + (c_1b_1 + \cdots + c_nb_n)\right]x^T \\
&= \left[\left(b_1^T(Ax+a) + c_1\right)b_1 + \cdots + \left(b_n^T(Ax+a) + c_n\right)b_n\right]x^T \\
&= \begin{bmatrix} | & & | \\ b_1 & \dots & b_n \\ | & & | \end{bmatrix} \begin{bmatrix} (y_1 - t_1) \\ \vdots \\ (y_n - t_n) \end{bmatrix} x^T && (b_i^T(Ax+a) + c_i = (y_i - t_i)) \\
&= B^T(y-t)x^T
\end{aligned}
$$

Now，我们将 $\sigma(Ax+b)$ 恢复成非线性变换，直接的：

$$
\begin{aligned}
\frac{\partial L}{\partial A} &= \frac{\partial \sigma^T(Ax+a)B^TB\sigma(Ax+a) + 2c^TB\sigma(Ax+a)}{\partial A} && (\text{令 } c = b - t) \\
&= \left(B^T(y-t) \odot \sigma'(z)\right)x^T
\end{aligned}
$$

1.6.6 对 a 的偏导

最后的成就感，留给大家做练习题吧。