Solution outline Project 2 2022

César Ramírez Ibáñez

December 1, 2022

1 Derivatives of cost functionals

For the loss functionals

$$J_1(\overline{W}) = \sum_{i=1}^{m}(1 - y_i\overline{W}\cdot\overline{X}_i^T)^2 \quad \big(\text{linear regression}\big) \tag{1}$$

$$J_2(\overline{W}) = \sum_{i=1}^{m}\log(1 + e^{-y_i\overline{W}\cdot\overline{X}_i^T}) \quad \big(\text{logistic regression}\big) \tag{2}$$

$$J_3(\overline{W}) = \sum_{i=1}^{m}\log(1 + e^{-y_i\overline{W}\cdot\overline{X}_i^T}) + \frac{\lambda}{2}\|\overline{W}\|_2^2 \quad \big(\text{regularized logistic regression}\big) \tag{3}$$

Let us at the moment take $J := J_2$. I found it easier to directly compute the partial derivatives of J_2. For $\overline{W} = (w_1, ..., w_d)$ we compute w.r.t w_k for $1 \le k \le d$ to get (noticing I write $\overline{X}_i = (x_1^{(i)}, ..., x_d^{(i)})$)

$$\begin{aligned}
&\frac{\partial}{\partial w_k}\log\left(1 + e^{-y_i\overline{W}\cdot\overline{X}_i^T}\right) \\
&\quad = \frac{\partial}{\partial w_k}\log\left(1 + e^{-y_i(w_1x_1^{(i)} + \cdots + w_dx_d^{(i)})}\right) = \frac{-y_ix_k^{(i)}}{1 + e^{-y_i\overline{W}\cdot\overline{X}_i^T}}e^{-y_i\overline{W}\cdot\overline{X}_i^T}
\end{aligned}$$

which multiplying and dividing by $e^{+y_i\overline{W}\cdot\overline{X}_i^T}$ becomes, after introducing the notation

$$p_i := p_i(\overline{W}) = \frac{1}{1 + e^{+y_i\overline{W}\cdot\overline{X}_i^T}} \tag{4}$$

from where we see that the ith gradient contribution in (2) is given by

$$\nabla_W \log\left(1 + e^{-y_i\overline{W}\cdot\overline{X}_i^T}\right) = -y_ip_i\overline{X}_i^T \quad \text{(a column vector!)} \tag{5}$$

and using the notation $P := \mathrm{diag}[p_1, ..., p_d]$, we get after adding all the contributions of type (5) we obtain

$$\nabla J(\overline{W}) = -X^T P\overline{y} = -\sum_{i=1}^{m} y_i p_i \overline{X}_i^T \tag{6}$$

and following the [Aggarwal] notation[1] and emphasizing the $\overline{W}$-dependence of the p_i we obtain

$$H = \left[\frac{\partial \nabla J(\overline{W})}{\partial \overline{W}}\right]^T = -\sum_{i=1}^{m} y_i \left[\frac{\partial}{\partial \overline{W}}\big(p_i(\overline{W})\overline{X}_i^T\big)\right]^T \tag{7}$$

In the denominator layout [Aggarwal] follows, the derivative of the column vector $p_i(\overline{W})\overline{X}_i^T$ w.r.t. the column vector $\overline{W}$ is based on the identity (iii) of his Table 4.2 (b) which is namely, for $\mathbf{x} : \mathbb{R}^d \to \mathbb{R}^d$ (COLUMN vector!) and $g : \mathbb{R}^d \to \mathbb{R}$ scalar-valued:

$$\frac{\partial}{\partial \overline{W}}\Big[g(\overline{W})\mathbf{x}(\overline{W})\Big] = \frac{\partial g}{\partial \overline{W}}\mathbf{x}^T + g(\overline{W})\frac{\partial \mathbf{x}}{\partial \overline{W}} \quad \leftarrow \text{(p. 173 [Aggarwal])} \tag{8}$$

where remember in [Aggarwal] convention the derivative of a scalar function w.r.t to column vector is a column vector. The derivative of a col vector w.r.t. another col vector is a matrix. But [Aggarwal] arranges gradients of scalar functions as columns, so unlike many standard calculus courses, the gradient of each component of a vector field is arranged as a column, whereas many follow the convention that the gradients of each scalar component are arranged in rows.

In our case, each vector $\overline{X}_i$ is independent of $\overline{W}$ and so (8) applied to our case gives (noting that for us, $\mathbf{x} = \overline{X}_i^T$ is a column vector)

$$\frac{\partial}{\partial \overline{W}}\big(p_i(\overline{W})\overline{X}_i^T\big) = \frac{\partial p_i}{\partial \overline{W}}\overline{X}_i = \tag{9}$$

Note

$$\frac{\partial[e^{y_i\overline{W}\cdot\overline{X}_i^T}]}{\partial \overline{W}} = \exp\left(y_i\overline{W}\cdot\overline{X}_i^T\right) y_i\overline{X}_i^T = \frac{(1-p_i)y_i}{p_i}\overline{X}_i^T \tag{10}$$

so that

$$\frac{\partial p_i}{\partial \overline{W}} = -p_i^2 \cdot \frac{(1-p_i)y_i}{p_i}\overline{X}_i^T = -y_i p_i(1-p_i)\overline{X}_i^T \tag{11}$$

so (12) into the expression for H in (7) becomes

$$H = \sum_{i=1}^{m} y_i^2 p_i(1-p_i)\overline{X}_i^T\overline{X}_i = \sum_{i=1}^{m} p_i(1-p_i)\overline{X}_i^T\overline{X}_i \tag{12}$$

where remember $\overline{X}_i$ is a row vector, so $\overline{X}_i^T\overline{X}_i$ is a rank-1 $d \times d$ matrix.[2]

[1] This author follows some convention where a Hessian is **the transpose** of some calculus convention of the differential $\frac{\partial}{\partial \overline{W}}$ that he adopts.

[2] In the convention that $\overline{X}_i$ is a column vector, it would be $\overline{X}_i\overline{X}_i^T$ instead in (12)

For A_1 the characteristic polynomial is $P_1(\lambda) = (\lambda - 3)(\lambda - 1)$ gives eigenvalues arranged into the change of basis matrix

$$C_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \tag{13}$$

we get

$$\begin{aligned} A_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} &= \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \\ &= \frac{3}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} [1 \quad 1] + \frac{1}{2} \begin{bmatrix} -1 \\ 1 \end{bmatrix} [-1 \quad 1] \end{aligned}$$

and since $C_1 = V$ and $C_1^{-1} = U^T$ are orthogonal, we note we have already provided an SVD decomposition and don't have to do anything else. For A_2, $P_2(\lambda) = (\lambda - 3)(\lambda - 1)$ and we get

$$C_2 = \begin{bmatrix} 1/2 & -1/2 \\ 1 & 1 \end{bmatrix} \tag{14}$$

is a matrix that diagonalizes $A_2 = C_2 D_2 C_2^{-1}$. However this is not an SVD decomposition because C_2 (unlike the previous case C_1) is not an orthogonal matrix. To get an SVD decomposition we have to diagonalize the matrix $M = A^T A$ to get $\Sigma^T \Sigma = \text{diag}[9 \quad 1]$ and a change of basis matrix U