NOTES ON DERIVATIVES OF SVD

Daniel Povey

Johns Hopkins University,
3400 N Charles St, Baltimore, MD 21218
dpovey@gmail.com
(Written in 2018)

ABSTRACT

These are some notes on how to backprop through a special type of matrix function that's based on applying a scalar function to the singular values of a matrix and reconstructing. This type of problem arises when estimating fMLLR transforms with spherical or unit covariance matrices. This document is not about finding derivatives w.r.t. the SVD operation itself.

Index Terms— SVD, derivatives

1. INTRODUCTION

We are interested in the matrix-valued function $F(\mathbf{A})$, operating on real-valued square matrices, that applies a scalar function $f(\lambda)$ to the singular values of that matrix. We'll assume that the scalar $f(\lambda)$ is defined, and differentiable, for $\lambda \geq 0$. To establish notation, if $\mathbf{A}$ may be singular-value decomposed as:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \tag{1}$$

(with, of course, $\mathbf{U}$ and $\mathbf{V}$ orthogonal and $\mathbf{\Lambda}$ diagonal and nonnegative), then

$$F(\mathbf{A}) = \mathbf{U}F(\mathbf{\Lambda})\mathbf{V}^T \tag{2}$$

where $F(\mathbf{\Lambda})$ is computed by applying the scalar function $f(\cdot)$ to the diagonal elements of $\mathbf{\Lambda}$. TODO: we'd like to establish under exactly what conditions this function is well-defined... I believe it's necessary that either $f(0) = 0$, or for all the singular values of $\mathbf{A}$ be positive.

2. DERIVING THE DERIVATIVE COMPUTATION

This section contains the derivation of our method to compute the derivatives. If you just want a how-to, please skip to the Summary (next section).

We want to backprop through $F(\mathbf{A})$. Suppose we are trying to find the derivative of a scalar function g that depends on $F(\mathbf{A})$. Let us use the notation $\bar{\mathbf{X}}$ for the derivative of g w.r.t. a matrix $\mathbf{X}$, and we'll use a notation where the i, j'th element of $\bar{X}$ is the derivative of g w.r.t. the i, j'th element of $\mathbf{X}$ (i.e., there is no transpose).

Suppose we are interested in the derivatives around a particular value of $\mathbf{A}$, then we'll first do the SVD

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \tag{3}$$

and, defining

$$\mathbf{D} = F(\mathbf{\Lambda}), \tag{4}$$

where $F(\cdot)$ above reduces to applying $f(\cdot)$ on the diagaonal elements, we compute the output $\mathbf{B} = F(\mathbf{A})$ as:

$$\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{5}$$

Our method of computing the derivatives here is to treat $\mathbf{U}$ and $\mathbf{V}$ as constants but treat $\mathbf{\Lambda}$ as a matrix that is not necessarily diagonal. Getting the derivative $\bar{\mathbf{D}}$ fairly straightforward. The nontrivial part is finding the derivatives for Equation 4, including the case where $\mathbf{\Lambda}$ has small nonzero elements off the diagonal.

When we change a diagonal element of $\mathbf{\Lambda}$ it will only affect the corresponding diagonal element of $\mathbf{D}$. When we change an off-diagonal element $\lambda_{i,j}$ of $\mathbf{\Lambda}$ by a small amount, it only has the potential to affect the four elements $\lambda_{i,j}$, $\lambda_{j,i}$, $\lambda_{i,i}$ and $\lambda_{j,j}$; and, as we will see below, it actually only affects the off-diagonal ones.

First, to state the obvious: for the diagonal elements,

$$\frac{\partial d_{i,i}}{\partial \lambda_{i,i}} = f'(\lambda_{i,i}). \tag{6}$$

2.1. Derivatives of off-diagonal elements (different singular values)

To analyze the derivatives for the off-diagonal elements, we can consider the case of a 2 by 2 matrix, since if we are changing one off-diagonal element we could always reorder the dimensions to make those two dimensions adjacent, and the other dimensions are irrelevant.

Let's consider the matrix

$$\mathbf{M} = \begin{pmatrix} a & \delta \\ 0 & b \end{pmatrix}, \tag{7}$$

with δ assumed to be small, $a \geq 0$, $b \geq 0$ and $a \neq b$.

To apply the function $F(\mathbf{M})$, we need to diagonalize $\mathbf{M}$ by multiplying by orthogonal matrices on the left and right; we can treat these orthogonal matrices as elements of the SVD of $\mathbf{M}$. Since $\mathbf{M}$ is already close to being diagonal, we will multiply it by expressions of the form $\begin{pmatrix} 1 & \epsilon \\ -\epsilon & 1 \end{pmatrix}$, you can verify that (ignoring ϵ^2), matrices of this form are orthogonal.

We'll try to solve the equation

$$\mathbf{U}^T\mathbf{M}\mathbf{V} = \text{diagonal} \tag{8}$$

which will immediately give us the SVD of $\mathbf{M}$. We'll let $\mathbf{U}^T = \begin{pmatrix} 1 & \epsilon_1 \\ -\epsilon_1 & 1 \end{pmatrix}$ and $\mathbf{V} = \begin{pmatrix} 1 & \epsilon_2 \\ -\epsilon_2 & 1 \end{pmatrix}$. Multiplying out $\mathbf{U}^T\mathbf{M}\mathbf{V}$ and ignoring products of "small" terms δ with ϵ, we get:

$$\begin{pmatrix} 1 & \epsilon_1 \\ -\epsilon_1 & 1 \end{pmatrix}\begin{pmatrix} a & \delta \\ 0 & b \end{pmatrix}\begin{pmatrix} 1 & \epsilon_2 \\ -\epsilon_2 & 1 \end{pmatrix} \simeq \begin{pmatrix} a & \delta + \epsilon_2 a + \epsilon_1 b \\ -\epsilon_1 a - \epsilon_2 b & b \end{pmatrix} \quad (9)$$

(TODO: obviously need to make this more rigorous, with $o(\delta^2)$ and so on). Because we want to diagonalize $\mathbf{M}$, we want to equate (9) with $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$. That implies that we need to solve the following two equations:

$$-\epsilon_1 a - \epsilon_2 b = 0 \quad (10)$$

$$\delta + \epsilon_2 a + \epsilon_1 b = 0 \quad (11)$$

From (10), we have $\epsilon_1 = -\frac{b}{a}\epsilon_2$; substituting that into (11), we get

$$\epsilon_2 = \frac{-\delta a}{a^2 - b^2} \quad (12)$$

and then:

$$\epsilon_1 = \frac{\delta b}{a^2 - b^2}. \quad (13)$$

We can write, again being a bit sloppy,

$$\mathbf{M} \simeq \mathbf{U}\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}\mathbf{V}^T = \begin{pmatrix} 1 & -\epsilon_1 \\ \epsilon_1 & 1 \end{pmatrix}\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}\begin{pmatrix} 1 & -\epsilon_2 \\ \epsilon_2 & 1 \end{pmatrix} \quad (14)$$

After applying $f(\cdot)$ to the singular values, we can get $\mathbf{N} = F(\mathbf{M})$ as:

$$\begin{aligned} \mathbf{N} &= F(\mathbf{M}) \\ &\simeq \mathbf{U}\begin{pmatrix} f(a) & 0 \\ 0 & f(b) \end{pmatrix}\mathbf{V}^T \\ &= \begin{pmatrix} 1 & -\epsilon_1 \\ \epsilon_1 & 1 \end{pmatrix}\begin{pmatrix} f(a) & 0 \\ 0 & f(b) \end{pmatrix}\begin{pmatrix} 1 & -\epsilon_2 \\ \epsilon_2 & 1 \end{pmatrix} \\ &\simeq \begin{pmatrix} f(a) & -\epsilon_2 f(a) - \epsilon_1 f(b) \\ \epsilon_1 f(a) + \epsilon_2 f(b) & f(b) \end{pmatrix}, \end{aligned} \quad (15)$$

where in (15) we are again ignoring terms that are products of epsilons. Notice that, in the case where f is the identity function, (15) reduces to $\begin{pmatrix} a & \delta \\ 0 & b \end{pmatrix}$, which is expected because then $\mathbf{N} = \mathbf{M}$.

The top-right element in (15), divided by δ, is the derivative of $n_{1,2}$ w.r.t. $m_{1,2}$:

$$\frac{\partial n_{1,2}}{\partial m_{1,2}} = \frac{-\epsilon_2 f(a) - \epsilon_1 f(b)}{\delta} \quad (16)$$

$$= \frac{af(a) - bf(b)}{a^2 - b^2}. \quad (17)$$

Doing the same procedure for the lower-right element in (15), we get:

$$\frac{\partial n_{2,1}}{\partial m_{1,2}} = \frac{\epsilon_1 f(a) + \epsilon_2 f(b)}{\delta} \quad (18)$$

$$= \frac{bf(a) - af(b)}{a^2 - b^2}. \quad (19)$$

It is reassuring to notice that if $f(\cdot)$ is the identity function, then (17) reduces to 1 and (19) to 0.

2.2. Derivatives of off-diagonal elements (identical nonzero singular values)

The expressions in (17) and (19) are not defined where $a = b$. Here we figure out the value in this case by finding the limiting value of those expressions as a and b approach each other. This is a rather sloppy way to do it; we'll explain below how we could derive it more correctly using the SVD.

If a and b are very close, then let us write $b = a + \delta$ (note: this is a different δ than before), and $f(b) = f(a) + \delta f'(a)$. Remembering that $a^2 - b^2$ factorizes as $(a-b)(a+b)$, and again ignoring products of deltas, we get that for $b = a + \delta$,

$$\frac{\partial n_{1,2}}{\partial m_{1,2}} = \frac{af(a) - (a+\delta)(f(a) + \delta f'(a))}{(a - (a+\delta))(a + (a+\delta))} \quad (20)$$

$$\simeq \frac{-\delta(f(a) + af'(a))}{-2\delta a} \quad (21)$$

$$= \frac{af'(a) + f(a)}{2a} \quad (22)$$

and:

$$\frac{\partial n_{2,1}}{\partial m_{1,2}} = \frac{(a+\delta)f(a) - a(f(a) + \delta f'(a))}{(a - (a+\delta))(a + (a+\delta))} \quad (23)$$

$$\simeq \frac{\delta(f(a) - af'(a))}{-2\delta a} \quad (24)$$

$$= \frac{af'(a) - f(a)}{2a} \quad (25)$$

2.2.1. *Sketch of more correct derivation*

This is a sketch of how we could derive the expressions above more properly. Basically, matrices of the form $\begin{pmatrix} a & \delta \\ 0 & a \end{pmatrix}$ appear to have singular value decompositions of the form:

$$\begin{pmatrix} a & \delta \\ 0 & a \end{pmatrix} = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -\sqrt{2} \end{pmatrix}\begin{pmatrix} 1 + \frac{\delta}{2} & 0 \\ 0 & 1 - \frac{\delta}{2} \end{pmatrix}\begin{pmatrix} \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -\sqrt{2} \end{pmatrix} \quad (26)$$

and we believe we can probably use this fact to obtain those same expressions.

2.3. Derivatives of off-diagonal elements (pairs of zero singular values)

The approach above doesn't work when both identical singular values are zero. We need to remind the reader that unless $f(0) = 0$, this type of function is not well defined (since the SVD of a zero matrix can use arbitrary orthogonal matrices for $\mathbf{U}$ and $\mathbf{V}$. So our approach will only be valid for $f(0) = 0$. For sufficiently small matrices, the function $F(\mathbf{A})$ equals $f'(0)\mathbf{A} + O(\epsilon^2)$, which (ignoring the smaller term) is just multiplication by a scalar. So, using notation similar to the above, if $a = b = 0$ (i.e. $m_{1,1} = m_{2,2} = 0$), then

$$\frac{\partial n_{1,2}}{\partial m_{1,2}} = f'(0) \quad (27)$$

$$\frac{\partial n_{1,2}}{\partial m_{2,1}} = 0. \quad (28)$$

3. SUMMARY

In this section we present the backprop for this function as a "how-to". The idea is that you should be able to read this section independently of the derivation above.

3.1. Forward pass

We are given square real matrix-valued input $\mathbf{A}$ and a scalar function $f(\cdot)$ defined and differentiable for positive real input. We compute $\mathbf{B} = F(\mathbf{A})$ by doing the singular value decomposition:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \tag{29}$$

with $\mathbf{U}$ and $\mathbf{V}$ orthogonal and $\mathbf{\Lambda}$ diagonal with positive diagonal elements, then compute:

$$\mathbf{D} = F(\mathbf{\Lambda}) \tag{30}$$

where in this case $F(\cdot)$ just means applying the scalar function $f(\cdot)$ to the diagonal elements; and we'll output the following expression:

$$\mathbf{B} = F(\mathbf{A}) = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{31}$$

3.2. Backward pass

We are doing backprop the derivatives of g, and we are given $\bar{\mathbf{B}} = \frac{\partial g}{\partial \mathbf{B}}$. The notation we'll use is that the i,j'th element of $\bar{\mathbf{B}}$ is the derivative of g w.r.t. the i,j'th element of $\mathbf{B}$ (i.e. there is no transpose). Our aim is to compute $\bar{\mathbf{A}}$, which is the derivative of g w.r.t. $\mathbf{A}$.

We compute the quantity:

$$\bar{\mathbf{D}} = \mathbf{U}^T\bar{\mathbf{B}}\mathbf{V} \tag{32}$$

and the next step is to compute $\bar{\mathbf{\Lambda}}$ from $\bar{\mathbf{D}}$; this is the derivative of g w.r.t. $\mathbf{\Lambda}$, not making the assumption that $\mathbf{\Lambda}$ is constrained to be diagonal.

For notation: let λ_i be the i'th diagonal element of $\mathbf{\Lambda}$, and d_i be the i'th diagonal element of $\mathbf{D} = F(\mathbf{\Lambda})$.

We compute $\bar{\mathbf{\Lambda}}$ as follows. First, for the diagonal elements, set:

$$\bar{\lambda}_{i,i} = f'(\lambda_i)\bar{d}_{i,i} \tag{33}$$

Then, for each $i \neq j$, we do as follows. If λ_i is not extremely close to λ_j (e.g. they differ by more than 10^{-7} relatively), then set:

$$\begin{aligned}\bar{\lambda}_{i,j} &= \bar{d}_{i,j}\frac{\lambda_i d_i - \lambda_j d_j}{\lambda_i^2 - \lambda_j^2} \\ &\quad + \bar{d}_{j,i}\frac{\lambda_j d_i - \lambda_i d_j}{\lambda_i^2 - \lambda_j^2},\end{aligned} \tag{34}$$

where the two terms in (34) correspond to Equations 17 and 19 respectively. On the other hand, if λ_i and λ_j are extremely close but nonzero, then with $\lambda = \frac{1}{2}(\lambda_i + \lambda_j)$, let:

$$\begin{aligned}\bar{\lambda}_{i,j} &= \bar{d}_{i,j}\frac{\lambda f'(\lambda) + d_i}{2\lambda} \\ &\quad + \bar{d}_{j,i}\frac{\lambda f'(\lambda) - d_i}{2\lambda}\end{aligned} \tag{35}$$

The two terms in (35) correspond to (22) and (25) respectively.

If λ_i and λ_j are both zero, then let:

$$\bar{\lambda}_{i,j} = \bar{d}_{i,j}f'(0) \tag{36}$$

Once the derivative $\bar{\mathbf{\Lambda}}$ is obtained as described above, we can compute our answer:

$$\bar{\mathbf{A}} = \mathbf{U}\bar{\mathbf{\Lambda}}\mathbf{V}^T. \tag{37}$$