Differentiating the Singular Value Decomposition

James Townsend

August 10, 2016

1 The low rank case

Let $\mathbf{A}$ be an $m \times n$ matrix of rank $k \leq \min(m, n)$. Then we may decompose $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, where $\mathbf{U}$ is $m \times k$, $\mathbf{S}$ is $k \times k$ diagonal, $\mathbf{V}$ is $n \times k$ and the matrices $\mathbf{U}$ and $\mathbf{V}$ satisfy the relation

$$\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k. \tag{1}$$

In this case the differential of $\mathbf{A}$ may be expressed as

$$\mathrm{d}\mathbf{A} = \mathrm{d}\mathbf{U}\mathbf{S}\mathbf{V}^\top + \mathbf{U}\mathrm{d}\mathbf{S}\mathbf{V}^\top + \mathbf{U}\mathbf{S}\mathrm{d}\mathbf{V}^\top. \tag{2}$$

The constraint (1) implies that the diffentials $\mathrm{d}\mathbf{U}$ and $\mathrm{d}\mathbf{V}$ are also constrained: focussing on $\mathbf{U}$ for a moment, taking the differential of (1) gives

$$\mathrm{d}\mathbf{U}^\top\mathbf{U} + \mathbf{U}^\top\mathrm{d}\mathbf{U} = \mathbf{0}. \tag{3}$$

So the matrix $\mathrm{d}\Omega_\mathbf{U} = \mathbf{U}^\top\mathrm{d}\mathbf{U}$ is *skew-symmetric*. In fact, if we fix an $m \times (m - k)$ matrix $\mathbf{U}_\perp$ such that $\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}$ is an orthogonal matrix (this could be computed using the Gram-Schmidt process) then we may expand $\mathrm{d}\mathbf{U}$ as

$$\mathrm{d}\mathbf{U} = \mathbf{U}\mathrm{d}\Omega_\mathbf{U} + \mathbf{U}_\perp\mathrm{d}\mathbf{K}_\mathbf{U} \tag{4}$$

where $\mathrm{d}\mathbf{K}_\mathbf{U}$ is an unconstrained $(m - k) \times k$ matrix. Similarly we may expand $\mathrm{d}\mathbf{V}$ as

$$\mathrm{d}\mathbf{V} = \mathbf{V}\mathrm{d}\Omega_\mathbf{V} + \mathbf{V}_\perp\mathrm{d}\mathbf{K}_\mathbf{V} \tag{5}$$

where $\mathrm{d}\Omega_\mathbf{V} = \mathbf{V}^\top\mathrm{d}\mathbf{V}$ is $k \times k$ skew-symmetric and $\mathrm{d}\mathbf{K}_\mathbf{V}$ is an $(n - k) \times k$ matrix. See [1] for more detail. Left-multiplying (2) by $\mathbf{U}^\top$ and right-multiplying by $\mathbf{V}$ gives

$$\mathbf{U}^\top\mathrm{d}\mathbf{A}\mathbf{V} = \mathrm{d}\Omega_\mathbf{U}\mathbf{S} + \mathrm{d}\mathbf{S} + \mathbf{S}\mathrm{d}\Omega_\mathbf{V}^\top. \tag{6}$$

Since $\mathrm{d}\Omega_\mathbf{U}$ and $\mathrm{d}\Omega_\mathbf{V}$ are skew-symmetric, they have zero diagonal and thus the products $\mathrm{d}\Omega_\mathbf{U}\mathbf{S}$ and and $\mathbf{S}\mathrm{d}\Omega_\mathbf{V}^\top$ must also have zero diagonal. This means that we can split (6) into two components as follows. Letting $\mathrm{d}\mathbf{P} := \mathbf{U}^\top\mathrm{d}\mathbf{A}\mathbf{V}$ and using $\circ$ to denote the Hadamard product, the diagonal component of (6) is

$$\mathrm{d}\mathbf{S} = \mathbf{I}_k \circ \mathrm{d}\mathbf{P} \tag{7}$$

and the off diagonal

$$\bar{\mathbf{I}}_k \circ \mathrm{d}\mathbf{P} = \mathrm{d}\Omega_{\mathbf{U}}\mathbf{S} - \mathbf{S}\mathrm{d}\Omega_{\mathbf{V}} \tag{8}$$

where $\bar{\mathbf{I}}_k$ denotes the $k \times k$ matrix with zero diagonal and ones everywhere else.

Taking the transpose of (8) yields

$$\bar{\mathbf{I}}_k \circ \mathrm{d}\mathbf{P}^\top = -\mathbf{S}\mathrm{d}\Omega_{\mathbf{U}} + \mathrm{d}\Omega_{\mathbf{V}}\mathbf{S}. \tag{9}$$

Now right multiply (8) by $\mathbf{S}$, left multiply (9) by $\mathbf{S}$ and add. This gives

$$\bar{\mathbf{I}}_k \circ \left[\mathrm{d}\mathbf{P}\mathbf{S} + \mathbf{S}\mathrm{d}\mathbf{P}^\top\right] = \mathrm{d}\Omega_{\mathbf{U}}\mathbf{S}^2 - \mathbf{S}^2\mathrm{d}\Omega_{\mathbf{U}}, \tag{10}$$

which is solved by

$$\mathrm{d}\Omega_{\mathbf{U}} = \mathbf{F} \circ \left[\mathrm{d}\mathbf{P}\mathbf{S} + \mathbf{S}\mathrm{d}\mathbf{P}^\top\right] \tag{11}$$

where $\mathbf{F}_{ij} = \begin{cases} \frac{1}{s_j^2 - s_i^2} & i \neq j \\ 0 & i = j \end{cases}$. By a similar process,

$$\mathrm{d}\Omega_{\mathbf{V}} = \mathbf{F} \circ \left[\mathbf{S}\mathrm{d}\mathbf{P} + \mathrm{d}\mathbf{P}^\top\mathbf{S}\right]. \tag{12}$$

Finally, to find $\mathrm{d}\mathbf{K}_{\mathbf{U}}$, we left multiply (2) by $\mathbf{U}_\perp^\top$, which yields

$$\mathbf{U}_\perp^\top\mathrm{d}\mathbf{A} = \mathrm{d}\mathbf{K}_{\mathbf{U}}\mathbf{S}\mathbf{V}^\top \tag{13}$$

which implies that

$$\mathrm{d}\mathbf{K}_{\mathbf{U}} = \mathbf{U}_\perp^\top\mathrm{d}\mathbf{A}\mathbf{V}\mathbf{S}^{-1}. \tag{14}$$

By a similar line of reasoning,

$$\mathrm{d}\mathbf{K}_{\mathbf{V}} = \mathbf{V}_\perp^\top\mathrm{d}\mathbf{A}^\top\mathbf{U}\mathbf{S}^{-1}. \tag{15}$$

All of this derivation can now be combined into formulae for the differentials $\mathrm{d}\mathbf{U}$, $\mathrm{d}\mathbf{S}$ and $\mathrm{d}\mathbf{V}$ in terms of $\mathrm{d}\mathbf{A}$, $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V}$. We use the identity $\mathbf{U}_\perp\mathbf{U}_\perp^\top = \mathbf{I} - \mathbf{U}\mathbf{U}^\top$ to eliminate $\mathbf{U}_\perp$ and $\mathbf{V}_\perp$.

$$\mathrm{d}\mathbf{U} = \mathbf{U}\left(\mathbf{F} \circ \left[\mathbf{U}^\top\mathrm{d}\mathbf{A}\mathbf{V}\mathbf{S} + \mathbf{S}\mathbf{V}^\top\mathrm{d}\mathbf{A}^\top\mathbf{U}\right]\right) + \left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)\mathrm{d}\mathbf{A}\mathbf{V}\mathbf{S}^{-1} \tag{16}$$

$$\mathrm{d}\mathbf{S} = \mathbf{I}_k \circ \left[\mathbf{U}^\top\mathrm{d}\mathbf{A}\mathbf{V}\right] \tag{17}$$

$$\mathrm{d}\mathbf{V} = \mathbf{V}\left(\mathbf{F} \circ \left[\mathbf{S}\mathbf{U}^\top\mathrm{d}\mathbf{A}\mathbf{V} + \mathbf{V}^\top\mathrm{d}\mathbf{A}^\top\mathbf{U}\mathbf{S}\right]\right) + \left(\mathbf{I}_n - \mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{A}^\top\mathbf{U}\mathbf{S}^{-1} \tag{18}$$

1.1 Reverse mode AD updates

Suppose we have an objective function $f(\mathbf{x})$ whose gradient we wish to calculate. Use the shorthand $\bar{\cdot} = \nabla_{\cdot} f$ to denote the grad of f with respect to $\cdot$, so the gradient we are looking for is $\bar{\mathbf{x}}$. Suppose that at some stage during the computation of f, we take a a matrix $\mathbf{A}(\mathbf{x})$ and compute its svd $\mathbf{U}(\mathbf{x})\mathbf{S}(\mathbf{x})\mathbf{V}(\mathbf{x})^\top$

We may write

$$\mathrm{d}f = \mathrm{tr}(\overline{\mathbf{U}}^\top \mathrm{d}\mathbf{U}) + \mathrm{tr}(\overline{\mathbf{S}}^\top \mathrm{d}\mathbf{S}) + \mathrm{tr}(\overline{\mathbf{V}}^\top \mathrm{d}\mathbf{V}). \tag{19}$$

To get the reverse mode AD update, we need to use the formulae (16), (17) and (18), and massage the right hand side into the form $\mathrm{tr}(\overline{\mathbf{A}}^\top \mathrm{d}\mathbf{A})$, then $\overline{\mathbf{A}}$ will be what we need for the update. Let us look first at the term $\mathrm{tr}(\overline{\mathbf{S}}^\top \mathrm{d}\mathbf{S})$. Using (17), this can be written as

$$\begin{aligned}
\mathrm{tr}(\overline{\mathbf{S}}^\top \mathrm{d}\mathbf{S}) &= \mathrm{tr}\left(\overline{\mathbf{S}}^\top \left(\mathbf{I}_k \circ \left[\mathbf{U}^\top \mathrm{d}\mathbf{A}\mathbf{V}\right]\right)\right) &(20)\\
&= \mathrm{tr}\left(\mathbf{U}^\top \mathrm{d}\mathbf{A}\mathbf{V}\left(\mathbf{I}_k \circ \overline{\mathbf{S}}\right)\right) &(21)\\
&= \mathrm{tr}\left(\mathbf{V}\left(\mathbf{I}_k \circ \overline{\mathbf{S}}\right)\mathbf{U}^\top \mathrm{d}\mathbf{A}\right) &(22)
\end{aligned}$$

using formula 65 of [2]. The expansion of $\mathrm{tr}(\overline{\mathbf{U}}^\top \mathrm{d}\mathbf{U})$ is a little longer...

$$\mathrm{tr}(\overline{\mathbf{U}}^\top \mathrm{d}\mathbf{U}) = \mathrm{tr}\left(\overline{\mathbf{U}}^\top \left[\mathbf{U}\left(\mathbf{F} \circ \left[\mathbf{U}^\top \mathrm{d}\mathbf{A}\mathbf{V}\mathbf{S} + \mathbf{S}\mathbf{V}^\top \mathrm{d}\mathbf{A}^\top \mathbf{U}\right]\right) + \left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)\mathrm{d}\mathbf{A}\mathbf{V}\mathbf{S}^{-1}\right]\right). \tag{23}$$

The right hand side is a sum of two terms. Again using formula 65 of [2] and the fact that $\mathbf{F}^\top = -\mathbf{F}$, the first term is

$$\mathrm{tr}\left(\overline{\mathbf{U}}^\top \mathbf{U}\left(\mathbf{F} \circ \left[\mathbf{U}^\top \mathrm{d}\mathbf{A}\mathbf{V}\mathbf{S} + \mathbf{S}\mathbf{V}^\top \mathrm{d}\mathbf{A}^\top \mathbf{U}\right]\right)\right) = \mathrm{tr}\left(\left[\mathbf{U}^\top \mathrm{d}\mathbf{A}\mathbf{V}\mathbf{S} + \mathbf{S}\mathbf{V}^\top \mathrm{d}\mathbf{A}^\top \mathbf{U}\right]\left(\mathbf{F} \circ \mathbf{U}^\top \overline{\mathbf{U}}\right)\right) \tag{24}$$

$$\begin{aligned}
&= \mathrm{tr}\left(\mathbf{V}\mathbf{S}\left(\mathbf{F} \circ \mathbf{U}^\top \overline{\mathbf{U}}\right)\mathbf{U}^\top \mathrm{d}\mathbf{A} - \mathbf{V}\mathbf{S}\left(\mathbf{F} \circ \overline{\mathbf{U}}^\top \mathbf{U}\right)\mathbf{U}^\top \mathrm{d}\mathbf{A}\right) &(25)\\
&= \mathrm{tr}\left(\mathbf{V}\mathbf{S}\left(\mathbf{F} \circ \left[\mathbf{U}^\top \overline{\mathbf{U}} - \overline{\mathbf{U}}^\top \mathbf{U}\right]\right)\mathbf{U}^\top \mathrm{d}\mathbf{A}\right) &(26)
\end{aligned}$$

The second term is more straightforward to deal with

$$\mathrm{tr}\left(\overline{\mathbf{U}}^\top \left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)\mathrm{d}\mathbf{A}\mathbf{V}\mathbf{S}^{-1}\right) = \mathrm{tr}\left(\mathbf{V}\mathbf{S}^{-1}\overline{\mathbf{U}}^\top \left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)\mathrm{d}\mathbf{A}\right) \tag{27}$$

and therefore

$$\mathrm{tr}(\overline{\mathbf{U}}^\top \mathrm{d}\mathbf{U}) = \mathrm{tr}\left(\mathbf{V}\left[\mathbf{S}\left(\mathbf{F} \circ \left[\mathbf{U}^\top \overline{\mathbf{U}} - \overline{\mathbf{U}}^\top \mathbf{U}\right]\right)\mathbf{U}^\top + \mathbf{S}^{-1}\overline{\mathbf{U}}^\top \left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)\right]\mathrm{d}\mathbf{A}\right). \tag{28}$$

A similar derivation leads to

$$\mathrm{tr}(\overline{\mathbf{V}}^\top \mathrm{d}\mathbf{V}) = \mathrm{tr}\left(\left[\mathbf{V}\left(\mathbf{F} \circ \left[\mathbf{V}^\top \overline{\mathbf{V}} - \overline{\mathbf{V}}^\top \mathbf{V}\right]\right)\mathbf{S} + \left(\mathbf{I}_m - \mathbf{V}\mathbf{V}^\top\right)\overline{\mathbf{V}}\mathbf{S}^{-1}\right]\mathbf{U}^\top \mathrm{d}\mathbf{A}\right). \tag{29}$$

Putting all of this together leads to the update equation

$$\overline{\mathbf{A}} = \left[\mathbf{U}\left(\mathbf{F}\circ\left[\mathbf{U}^\top\overline{\mathbf{U}} - \overline{\mathbf{U}}^\top\mathbf{U}\right]\right)\mathbf{S} + \left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)\overline{\mathbf{U}}\mathbf{S}^{-1}\right]\mathbf{V}^\top + \tag{30}$$

$$\mathbf{U}\left(\mathbf{I}_k\circ\overline{\mathbf{S}}\right)\mathbf{V}^\top + \mathbf{U}\left[\mathbf{S}\left(\mathbf{F}\circ\left[\mathbf{V}^\top\overline{\mathbf{V}} - \overline{\mathbf{V}}^\top\mathbf{V}\right]\right)\mathbf{V}^\top + \mathbf{S}^{-1}\overline{\mathbf{V}}^\top\left(\mathbf{I}_n - \mathbf{V}\mathbf{V}^\top\right)\right] \tag{31}$$

by taking the transposes of the expressions above and noting that the matrices $\mathbf{F}\circ\left[\mathbf{U}^\top\overline{\mathbf{U}} - \overline{\mathbf{U}}^\top\mathbf{U}\right]$ and $\mathbf{F}\circ\left[\mathbf{V}^\top\overline{\mathbf{V}} - \overline{\mathbf{V}}^\top\mathbf{V}\right]$ are symmetric.

References

[1] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[2] Thomas P. Minka. Old and new matrix algebra useful for statistics. `http://research.microsoft.com/en-us/um/people/minka/papers/matrix/minka-matrix.pdf`.