Properties of the Trace and Matrix Derivatives

John Duchi

Contents

1 Notation

A few things on notation (which may not be very consistent, actually): The columns of a matrix $A \in \mathbb{R}^{m\times n}$ are a_1 through a_n, while the rows are given (as vectors) by $\tilde{a}_1^T$ throught $\tilde{a}_m^T$.

2 Matrix multiplication

First, consider a matrix $A \in \mathbb{R}^{n\times n}$. We have that

$$AA^T = \sum_{i=1}^{n} a_i a_i^T,$$

that is, that the product of AA^T is the sum of the outer products of the columns of A. To see this, consider that

$$(AA^T)_{ij} = \sum_{p=1}^{n} a_{pi} a_{pj}$$

because the i, j element is the i^{th} row of A, which is the vector $\langle a_{1i}, a_{2i}, \cdots, a_{ni}\rangle$, dotted with the j^{th} column of A^T, which is $\langle a_{1j}, \cdots, a_{nj}$.

If we look at the matrix AA^T, we see that

$$AA^T = \begin{bmatrix} \sum_{p=1}^n a_{p1}a_{p1} & \cdots & \sum_{p=1}^n a_{p1}a_{pn} \\ \vdots & \ddots & \vdots \\ \sum_{p=1}^n a_{pn}a_{p1} & \cdots & \sum_{p=1}^n a_{pn}a_{pn} \end{bmatrix} = \sum_{i=1}^n \begin{bmatrix} a_{i1}a_{i1} & \cdots & a_{i1}a_{in} \\ \vdots & \ddots & \vdots \\ a_{in}a_{i1} & \cdots & a_{in}a_{in} \end{bmatrix} = \sum_{i=1}^n a_i a_i^T$$

3 Gradient of linear function

Consider Ax, where $A \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$. We have

$$\nabla_x Ax = \begin{bmatrix} \nabla_x \tilde{a}_1^T x \\ \nabla_x \tilde{a}_2^T x \\ \vdots \\ \nabla_x \tilde{a}_m^T x \end{bmatrix} = \begin{bmatrix} \tilde{a}_1 & \tilde{a}_2 & \cdots & \tilde{a}_m \end{bmatrix} = A^T$$

Now let us consider $x^T Ax$ for $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$. We have that

$$x^T Ax = x^T[\tilde{a}_1^T x \ \tilde{a}_2^T x \ \cdots \ \tilde{a}_n^T x]^T = x_1\tilde{a}_1^T x + \cdots + x_n\tilde{a}_n^T x$$

If we take the derivative with respect to one of the x_ls, we have the l component for each $\tilde{a}_i$, which is to say a_{il}, and the term for $x_l\tilde{a}_l^T x$, which gives us that

$$\frac{\partial}{\partial x_l} x^T Ax = \sum_{i=1}^n x_i a_{il} + \tilde{a}_l^T x = a_l^T x + \tilde{a}_l^T x.$$

In the end, we see that

$$\nabla_x x^T Ax = Ax + A^T x.$$

4 Derivative in a trace

Recall (as in *Old and New Matrix Algebra Useful for Statistics*) that we can define the differential of a function $f(x)$ to be the part of $f(x+dx) - f(x)$ that is linear in dx, i.e. is a constant times dx. Then, for example, for a vector valued function $\boldsymbol{f}$, we can have

$$\boldsymbol{f}(\boldsymbol{x} + d\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{f}'(\boldsymbol{x})d\boldsymbol{x} + \text{(higher order terms)}.$$

In the above, $\boldsymbol{f}'$ is the derivative (or Jacobian). Note that the gradient is the transpose of the Jacobian.

Consider an arbitrary matrix A. We see that

$$\frac{\operatorname{tr}(AdX)}{dX} = \frac{\operatorname{tr}\begin{bmatrix} \tilde{a}_1^T dx_1 & & \\ & \ddots & \\ & & \tilde{a}_n^T dx_n \end{bmatrix}}{dX} = \frac{\sum_{i=1}^n \tilde{a}_i^T dx_i}{dX}.$$

Thus, we have

$$\left[\frac{\operatorname{tr}(AdX)}{dX}\right]_{ij} = \left[\frac{\sum_{i=1}^n \tilde{a}_i^T dx_i}{\partial x_{ji}}\right] = a_{ij}$$

so that

$$\frac{\operatorname{tr}(AdX)}{dX} = A.$$

Note that this is the Jacobian formulation.

5 Derivative of product in trace

In this section, we prove that

$$\nabla_A \mathrm{tr} AB = B^T$$

$$\begin{aligned}
\mathrm{tr}AB &= \mathrm{tr}\begin{bmatrix} \longleftarrow \vec{a_1} \longrightarrow \\ \longleftarrow \vec{a_2} \longrightarrow \\ \vdots \\ \longleftarrow \vec{a_n} \longrightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{b_1} & \vec{b_2} & \cdots & \vec{b_n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \\
&= \mathrm{tr}\begin{bmatrix} \vec{a_1}^T\vec{b_1} & \vec{a_1}^T\vec{b_2} & \cdots & \vec{a_1}^T\vec{b_n} \\ \vec{a_2}^T\vec{b_1} & \vec{a_2}^T\vec{b_2} & \cdots & \vec{a_2}^T\vec{b_n} \\ \vdots & & \ddots & \vdots \\ \vec{a_n}^T\vec{b_1} & \vec{a_n}^T\vec{b_2} & \cdots & \vec{a_n}^T\vec{b_n} \end{bmatrix} \\
&= \sum_{i=1}^m a_{1i}b_{i1} + \sum_{i=1}^m a_{2i}b_{i2} + \ldots + \sum_{i=1}^m a_{ni}b_{in} \\
\Rightarrow \quad \frac{\partial \mathrm{tr}AB}{\partial a_{ij}} &= b_{ji} \\
\Rightarrow \quad \nabla_A \mathrm{tr}AB &= B^T
\end{aligned}$$

6 Derivative of function of a matrix

Here we prove that

$$\nabla_{A^T} f(A) = (\nabla_A f(A))^T.$$

$$\begin{aligned}
\nabla_{A^T} f(A) &= \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{21}} & \cdots & \frac{\partial f(A)}{\partial A_{n1}} \\ \frac{\partial f(A)}{\partial A_{12}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{n2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{1n}} & \frac{\partial f(A)}{\partial A_{2n}} & \cdots & \frac{\partial f(A)}{\partial A_{nn}} \end{bmatrix} \\
&= (\nabla_A f(A))^T
\end{aligned}$$

7 Derivative of linear transformed input to function

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose we have a matrix $A \in \mathbb{R}^{n \times m}$ and a vector $x \in \mathbb{R}^m$. We wish to compute $\nabla_x f(Ax)$. By the chain rule, we have

$$\begin{aligned}
\frac{\partial f(Ax)}{\partial x_i} &= \sum_{k=1}^n \frac{\partial f(Ax)}{\partial (Ax)_k} \cdot \frac{\partial (Ax)_k}{\partial x_i} = \sum_{k=1}^n \frac{\partial f(Ax)}{\partial (Ax)_k} \cdot \frac{\partial (\tilde{a}_k^T x)}{\partial x_i} \\
&= \sum_{k=1}^n \frac{\partial f(Ax)}{\partial (Ax)_k} \cdot a_{ki} = \sum_{k=1}^n \partial_k f(Ax) a_{ki} \\
&= a_i^T \nabla f(Ax).
\end{aligned}$$

As such, $\nabla_x f(Ax) = A^T \nabla f(Ax)$. Now, if we would like to get the second derivative of this function (third derivatives would be a little nice, but I do not like tensors), we have

$$\begin{aligned}
\frac{\partial^2 f(Ax)}{\partial x_i \partial x_j} &= \frac{\partial}{\partial x_j} a_i^T \nabla f(Ax) = \frac{\partial}{\partial x_j} \sum_{k=1}^{n} a_{ki} \frac{\partial f(Ax)}{\partial (Ax)_k} \\
&= \sum_{l=1}^{n} \sum_{k=1}^{n} a_{ki} \frac{\partial^2 f(Ax)}{\partial (Ax)_k \partial (Ax)_l} a_{li} \\
&= a_i^T \nabla^2 f(Ax) a_j
\end{aligned}$$

From this, it is easy to see that $\nabla_x^2 f(Ax) = A^T \nabla^2 f(Ax) A$.

8 Funky trace derivative

In this section, we prove that

$$\nabla_A \mathrm{tr} ABA^T C = CAB + C^T AB^T.$$

In this bit, let us have $AB = f(A)$, where f is matrix-valued.

$$\begin{aligned}
\nabla_A \mathrm{tr} ABA^T C &= \nabla_A \mathrm{tr} f(A) A^T C \\
&= \nabla_\bullet \mathrm{tr} f(\bullet) A^T C + \nabla_\bullet \mathrm{tr} f(A) \bullet^T C \\
&= (A^T C)^T f'(\bullet) + (\nabla_{\bullet^T} \mathrm{tr} f(A) \bullet^T C)^T \\
&= C^T AB^T + (\nabla_{\bullet^T} \mathrm{tr} \bullet^T C f(A))^T \\
&= C^T AB^T + ((Cf(A))^T)^T \\
&= C^T AB^T + CAB
\end{aligned}$$

9 Symmetric Matrices and Eigenvectors

In this we prove that for a symmetric matrix $A \in \mathbb{R}^{n \times n}$, all the eigenvalues are real, and that the eigenvectors of A form an orthonormal basis of $\mathbb{R}^n$.

First, we prove that the eigenvalues are real. Suppose one is complex: we have

$$\bar{\lambda} x^T x = (Ax)^T x = x^T A^T x = x^T Ax = \lambda x^T x.$$

Thus, all the eigenvalues are real.

Now, we suppose we have at least one eigenvector $v \neq 0$ of A. Consider a space W of vectors orthogonal to v. We then have that, for $w \in W$,

$$(Aw)^T v = w^T A^T v = w^T Av = \lambda w^T v = 0.$$

Thus, we have a set of vectors W that, when transformed by A, are still orthogonal to v, so if we have an original eigenvector v of A, then a simple inductive argument shows that there is an orthonormal set of eigenvectors.

To see that there is at least one eigenvector, consider the characteristic polynomial of A:

$$\mathcal{X}(A) = \det(A - \lambda I).$$

The field is algebraicly closed, so there is at least one complex root r, so we have that $A - rI$ is singular and there is a vector $v \neq 0$ that is an eigenvector of A. Thus r is a real eigenvalue, so we have the base case for our induction, and the proof is complete.