13 Derivatives of Eigenproblems

13.1 Differentiating on the Unit Sphere

Geometrically, we know that velocity vectors (equivalently, tangents) on the sphere are orthogonal to the radii. Out differentials say this algebraically, since given $x \in \mathbb{S}^n$ we have $x^T x = 1$, this implies that

$$2x^T dx = d(x^T x) = d(1) = 0.$$

In other words, at the point x on the sphere (a radius, if you will), dx, the linearization of the constraint of moving along the sphere satisfies $dx \perp x$. This is our first example where we have seen the infinitesimal perturbation dx being constrained. See Figure 16.

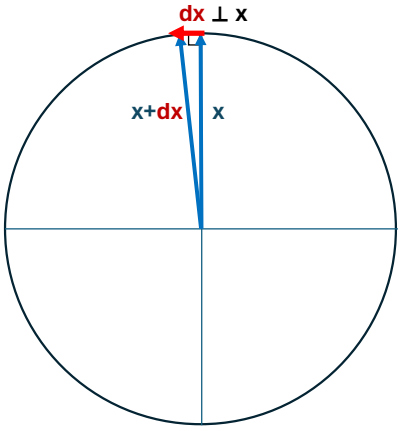

Figure 16: Differentials on a sphere ($x^T x = 1$): the differential dx is constrained to be perpendicular to x.

13.1.1 Special Case: A Circle

Let us simply consider the unit circle in the plane where $x = (\cos\theta, \sin\theta)$ for some $\theta \in [0, 2\pi)$. Then,

$$x^T dx = (\cos\theta, \sin\theta) \cdot (-\sin\theta, \cos\theta) d\theta = 0.$$

Here, we can think of x as "extrinsic" coordinates, in that it is a vector in $\mathbb{R}^2$. On the other hand, θ is an "intrinsic" coordinate, as every point on the circle is specified by one θ.

13.1.2 On the Sphere

You may remember that the rank-1 matrix xx^T, for any unit vector $x^T x = 1$, is a **projection matrix** (meaning that it is equal to its square and it is symmetric) which projects vectors onto their components in the direction of x. Correspondingly, $I - xx^T$ is also a projection matrix, but onto the directions *perpendicular* to x: geometrically, the matrix removes components in the x direction. In particular, if $x^T dx = 0$, then $(I - xx^T)dx = dx$. It follows

that if $x^T dx = 0$ and A is a symmetric matrix, we have

$$\begin{aligned} d\left(\frac{1}{2}x^T Ax\right) &= (Ax)^T dx \\ &= x^T A(dx) \\ &= x^T A(I - xx^T)dx \\ &= ((I - xx^T)Ax)^T dx. \end{aligned}$$

In other words, $(I - xx^T)Ax$ is the gradient of $\frac{1}{2}x^T Ax$ *on the sphere.*

So what did we just do? To obtain the gradient on the sphere, we needed (i) a linearization of the function that is correct on tangents, and (ii) a direction that *is* tangent (i.e. satisfies the linearized constraint). Using this, we obtain the gradient of a general scalar function on the sphere:

Theorem 60

Given $f : \mathbb{S}^n \to \mathbb{R}$, we have

$$df = g(x)^T dx = ((I - xx^T)g(x))^T dx.$$

The proof of this is precisely the same as we did before for $f(x) = \frac{1}{2}x^T Ax$.

13.2 Differentiating on Orthogonal Matrices

Let Q be an orthogonal matrix. Then, computationally (as is done in the Julia notebook), one can see that $Q^T dQ$ is an anti-symmetric matrix (sometimes called skew-symmetric).

Definition 61

A matrix M is anti-symmetric if $M = -M^T$. Note that all anti-symmetric matrices thus have zeroes on their diagonals.

In fact, we can prove that $Q^T dQ$ is anti-symmetric.

Theorem 62

Given Q is an orthogonal matrix, we have that $Q^T dQ$ is anti-symmetric.

Proof. The constraint of being orthogonal implies that $Q^T Q = I$. Differentiating this equation, we obtain

$$Q^T dQ + dQ^T Q = 0 \implies Q^T dQ = -(Q^T dQ)^T.$$

This is precisely the definition of being anti-symmetric. □

Before we move on, we may ask what the dimension of the "surface" of orthogonal matrices is in $\mathbb{R}^{n^2}$.

When $n = 2$, all orthogonal matrices are rotations and reflections, and rotations have the form

$$Q = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

Hence, when $n = 2$ we have one parameter.

When $n = 3$, airplane pilots know about "roll, pitch, and yaw", which are the three parameters for the orthogonal matrices when $n = 3$. In general, in $\mathbb{R}^{n^2}$, the orthogonal group has dimension $n(n-1)/2$.

There are a few ways to see this.

- Firstly, orthogonality $Q^TQ = I$ imposes $n(n+1)/2$ constraints, leaving $n(n-1)/2$ free parameters.
- When we do QR decomposition, the R "eats" up $n(n+1)/2$ of the parameters, again leaving $n(n-1)/2$ for Q.
- Lastly, If we think about the symmetric eigenvalue problem where $S = Q\Lambda Q^T$, S has $n(n+1)/2$ parameters and Λ has n, so Q has $n(n-1)/2$.

13.2.1 Differentiating the Symmetric Eigendecomposition

Let S be a symmetric matrix, Λ be diagonal containing eigenvalues of S, and Q be orthogonal with column vectors as eigenvectors of S such that $S = Q\Lambda Q^T$. [For simplicity, let's assume that the eigenvalues are "simple" (multiplicity 1); repeated eigenvalues turn out to greatly complicate the analysis of perturbations because of the ambiguity in their eigenvector basis.] Then, we have

$$dS = dQ\,\Lambda Q^T + Q\,d\Lambda\,Q^T + Q\Lambda dQ^T,$$

which may be written as

$$Q^T dS\,Q = Q^T dQ\Lambda - \Lambda Q^T dQ + d\Lambda.$$

As an exercise, one may check that the left and right hand sides of the above are both symmetric. This may be easier if one looks at the diagonal entries on their own, as there $(Q^T dS\,Q)_{ii} = q_i^T dS\,q_i$. Since q_i is the ith eigenvector, this implies $q_i^T dS\,q_i = d\lambda_i$. (In physics, this is sometimes called the "Hellman–Feynman" theorem, or non-degenerate first-order eigenvalue-perturbation theory.)

Sometimes we think of a curve of matrices $S(t)$ depending on a parameter such as time. If we ask for $\frac{d\lambda_i}{dt}$, this implies it is thus equal to $q_i^T \frac{dS(t)}{dt} q_i$. So how can we get the gradient $\nabla\lambda_i$ for one of the eigenvalues? Well, firstly, note that

$$\operatorname{tr}(q_i q_i^T)^T dS) = d\lambda_i \implies \nabla\lambda_i = q_i q_i^T.$$

What about the eigenvectors? Those come from off diagonal elements, where for $i \neq j$,

$$(Q^T dS\,Q)_{ij} = \left(Q^T \frac{dQ}{dt}\right)_{ij} (\lambda_j - \lambda_i).$$

Therefore, we can form the elements of $Q^T \frac{dQ}{dt}$, and left multiply by Q to obtain $\frac{dQ}{dt}$ (as Q is orthogonal).

It is interesting to get the second derivative of eigenvalues when moving along a line in symmetric matrix space. For simplicity, suppose Λ is diagonal and $S(t) = \Lambda + tE$. Therefore, differentiating

$$\frac{d\Lambda}{dt} = \operatorname{diag}\left(Q^T \frac{dS(t)}{dt} Q\right),$$

we get

$$\frac{d^2\Lambda}{dt^2} = \operatorname{diag}\left(Q^T \frac{d^2S(t)}{dt^2} Q\right) + 2\operatorname{diag}\left(Q^T \frac{dS(t)}{dt} \frac{dQ}{dt}\right).$$

Evaluating this at $Q = I$ and recognizing the first term is zero as we are on a line, we have that

$$\frac{d^2\Lambda}{dt^2} = 2\operatorname{diag}\left(E \cdot \frac{dQ}{dt}\right),$$

or

$$\frac{d^2\Lambda}{dt^2} = 2\sum_{k\neq i} E_{ik}^2/(\lambda_i - \lambda_k).$$

Using this, we can write out the eigenvalues as a Taylor series:

$$\lambda_i(\epsilon) = \lambda_i + \epsilon E_{ii} + \epsilon^2 \sum_{k\neq i} E_{ik}^2/(\lambda_i - \lambda_k) + \dots.$$

(In physics, this is known as second-order eigenvalue perturbation theory.)

18.S096 Matrix Calculus for Machine Learning and Beyond
Independent Activities Period (IAP) 2023