Matrix derivatives and matrix norm facts

Stephen Becker

Laboratoire Jacques-Louis Lions, University Paris-6
IBM Research; University of Colorado

October 14, 2014

Abstract

Originally from 4/15/2010. See appendix A in Boyd and Vandenberghe for some good stuff too Typed up 2/10/2013; then, old notes (3/18/10) about matrix norms added in August 2013. Adding info on Hessian in October 2014.

1 Matrix derivatives

Note: I write $\mathcal{A}^*$ to denote adjoint of a general linear operator $\mathcal{A}$, and I write X^T to denote the adjoint of a matrix (not necessarily the transpose), in order to distinguish matrices from more general linear operators. For example, we might define $\mathcal{A}(X) = A\text{vec}(X)$.

1.1 Method 1

Let $f(U,V) = \frac{1}{2}\|\mathcal{A}(UV^T) - b\|_2^2$ for a linear operator $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^r$, and U is $m \times k$ and V is $n \times k$. Question: what is ∇f_U (and same for V)? i.e. how to correctly apply chain rule with matrix variables.

This method due to Mike McCoy?

Part 1 Rewrite $f = \frac{1}{2}\langle \mathcal{A}(UV^T), \mathcal{A}(UV^T)\rangle - \langle b, \mathcal{A}(UV^T)\rangle + \frac{1}{2}\|b\|^2$. Define a new linear operator $\mathcal{A}_V(U) = \mathcal{A}(UV^T)$, so $f = \frac{1}{2}\langle \mathcal{A}_V(U), \mathcal{A}_V(U)\rangle - \langle b, \mathcal{A}_V(U)\rangle + \frac{1}{2}\|b\|^2$

Part 2 Derivatives. We will use product rule, thinking of U as U_1 and U_2 as separate variables.

$$\begin{aligned}
\nabla_U \frac{1}{2}\left\langle \mathcal{A}_V(U), \mathcal{A}_V(U)\right\rangle &= \nabla_U \frac{1}{2}\left\langle U, \mathcal{A}_V^*\mathcal{A}_V(U)\right\rangle && (1.1)\\
&= \nabla_{U_1,U_2} \frac{1}{2}\left\langle U_1, \mathcal{A}_V^*\mathcal{A}_V(U_2)\right\rangle && (1.2)\\
&= \frac{1}{2}(\nabla_{U_1}\left\langle U_1, \mathcal{A}_V^*\mathcal{A}_V(U_2)\right\rangle + \nabla_{U_2}\left\langle \mathcal{A}_V^*\mathcal{A}_V(U_1), U_2\right\rangle && (1.3)\\
&= \frac{1}{2}(\mathcal{A}_V^*\mathcal{A}_V(U_2) + \mathcal{A}_V^*\mathcal{A}_V(U_1)) && (1.4)\\
&= \mathcal{A}_V^*(\mathcal{A}_V(U)) && (1.5)
\end{aligned}$$

Part 3 If $\mathcal{A}_V(U) = \mathcal{A}(UV^T)$, what is $\mathcal{A}_V^*$? More generally, does $\left\langle \mathcal{A}XB, Y\right\rangle = \langle X, \mathcal{A}^*YB^*\rangle$? Yes, by using the circular properties of trace.

$$\left\langle \mathcal{A}_V(U), h\right\rangle = \left\langle \mathcal{A}(UV^T), h\right\rangle = \left\langle UV^T, \mathcal{A}^*(h)\right\rangle = \left\langle U, \mathcal{A}^*(h)V\right\rangle \equiv \left\langle U, \mathcal{A}_V^*(h)\right\rangle$$

so $\mathcal{A}_V^*(h) = \mathcal{A}^*(h)V$.

Part 4 Putting it all together. Including the linear term,

$$\boxed{\nabla_U f = \mathcal{A}_V^*(\mathcal{A}_V(U) - b) = \mathcal{A}^*(\mathcal{A}(UV^T) - b)V.}$$

Similarly, to find ∇_{V^T}, use $\mathcal{A}_U(V^T) = \mathcal{A}(UV^T)$ so $\mathcal{A}_U^*(h) = U^T\mathcal{A}^*(h)$, so

$$\boxed{\nabla_{V^T} f = U^T\mathcal{A}^*(\mathcal{A}(UV^T) - b).}\text{, i.e.,}\boxed{\nabla_V f = (\mathcal{A}^*(\mathcal{A}(UV^T) - b))^T U.}$$

If the problem is symmetric and $U = V$, then we just add these two gradients together, looking at ∇_V not ∇_{V^T}. See subsection 3.

1.2 Another approach

Due to Alex Gittens. Write $\frac{1}{2}\|\mathcal{A}(UV^T) - b\|^2 = f(g(U))$ where f is the quadratic and $g(U) = \mathcal{A}(UV^T) - b$. Then $D_U g$ is complicated, but its directional derivative is simple. $(D_U g)(h) = \mathcal{A}(hV^T)$ by the fact that the derivative of a linear operator is a constant. Also, $(D_X f(X))(h) = \langle X, h\rangle = X^T h$. So

$$(D_U f(g(U)))(h) = (D_{g(U)=\text{“}X\text{”}} f) \circ (D_U g)(h) \tag{1.6}$$

$$= \left\langle \text{“}X\text{”}, \mathcal{A}(hV^T) \right\rangle \tag{1.7}$$

$$= \left\langle \mathcal{A}(UV^T) - b, \mathcal{A}(hV^T) \right\rangle \tag{1.8}$$

Now to get rid of h, write it in the form $\left\langle D_U, h \right\rangle$, so

$$\left\langle \mathcal{A}^*(\mathcal{A}(UV^T) - b), hV^T \right\rangle = \left\langle \underbrace{\mathcal{A}^*(\mathcal{A}(UV^T) - b)V}_{D_U}, h \right\rangle.$$

1.3 Special case: symmetry

Added Jan 2013. Let $R = U = V$, then by product rule [edit: check this!]

$$\nabla_R f = \nabla_U f + \nabla_V f.$$

But we cannot just say this is $2\nabla_U f$ in general. We have $\nabla_{V^T} f = U^T\mathcal{A}^*(\mathcal{A}(UV^T) - b)$, so what is $\nabla_V f$? Let $Q = \mathcal{A}^*(\mathcal{A}(RR^T) - b)$, then $\nabla_{V^T} f = U^T Q$, so $\nabla_V = Q^T U = Q^T R$. So

$$\boxed{f(R) = \frac{1}{2}\|\mathcal{A}(RR^T) - b\|^2, \quad \text{then} \quad \nabla_R f(R) = (Q + Q^T)R \quad \text{where } Q = \mathcal{A}^*(\mathcal{A}(RR^T) - b)}.$$

It looks like Q should be Hermitian/symmetric since it is of the form $\mathcal{A}^*\mathcal{A}$, but since these are operators, I don't know. (They are self-adjoint when viewed as operators, but need not be the same as a Hermitian matrix) and at least the $\mathcal{A}^*b$ term is not guaranteed to be Hermitian/symmetric...

1.4 Similar question: sub-differentials of convex function with complex parts

We can write the total-variation operator as $f(x) = \|x\|_{TV} = \|Wx\|_1$ where $\|y\|_1 = \sum_{i=1}^n |y_i|$ and $|\cdot|$ is the complex modulus. The linear map W takes $\mathbb{R}^n$ to $\mathbb{C}^n$ or $\mathbb{C}^{2n}$. So f is a real-valued function from $\mathbb{R}^n$ to $\mathbb{R}$, but if we want the subdifferential of f, how do we apply the chain rule with W since W is complex? If we apply $W^*(z)$, do we need to take the real part of z?

1.5 Hessians

Again, we have $f(\mathbf{u},\mathbf{v}) = \frac{1}{2}\|\mathcal{A}(\mathbf{u}\mathbf{v}^T) - b\|_2^2$ and we want to compute the Hessian. We can find the Hessian (and derivative) by expanding out and collecting terms of the right order. We have, with $\nabla f_0 \stackrel{\text{def}}{=} \nabla f(u_0, v_0)$ and likewise for $\nabla^2 f_0$,

$$\begin{aligned} f(\mathbf{u}+u_0, \mathbf{v}+v_0) &= f(u_0,v_0) + \langle (\mathbf{u};\mathbf{v}), \nabla f_0\rangle + \langle (\mathbf{u};\mathbf{v}), \nabla^2 f_0 \cdot (\mathbf{u};\mathbf{v})\rangle \\ &= \frac{1}{2}\|\mathcal{A}((\mathbf{u}+u_0)(\mathbf{v}+v_0)^T) - b\|^2 \end{aligned}$$

We now expand this out, which is quite unpleasant. We denote constants in black (normal text), linear terms in *blue*, quadratic terms in *red*, and higher order terms in *green*.

$$\ldots = \frac{1}{2}\underbrace{\|\mathcal{A}((u+u_0)(v+v_0)^T)\|^2}_{I} + \frac{1}{2}\|b\|^2 - \langle \mathcal{A}^* b, \mathbf{u}v_0^T + u_0\mathbf{v}^T\rangle - \langle \mathcal{A}^* b, \mathbf{u}\mathbf{v}^T\rangle$$

and

$$\begin{aligned} I &= \langle (\mathbf{u}+u_0)(\mathbf{v}+v_0)^T, \underbrace{\mathcal{A}^*(\mathcal{A}(\mathbf{u}+u_0)(\mathbf{v}+v_0)^T)}_{z}\rangle \\ &= \langle \mathbf{u}\mathbf{v}^T, z\rangle + \langle \mathbf{u}v_0^T, z\rangle + \langle u_0\mathbf{v}^T, z\rangle + \langle u_0 v_0^T, z\rangle \\ &= \begin{aligned} &\left(\langle \mathbf{u}\mathbf{v}^T, z - \mathcal{A}^*\mathcal{A}(u_0v_0^T)\rangle + \langle \mathbf{u}\mathbf{v}^T, \mathcal{A}^*\mathcal{A}(u_0v_0^T)\rangle\right) \\ &+ \left\langle \mathbf{u}v_0^T, \mathcal{A}^*\mathcal{A}(\mathbf{u}\mathbf{v}^T + u_0\mathbf{v}^T + \mathbf{u}v_0^T + u_0v_0^T)\right\rangle \\ &+ \left\langle u_0\mathbf{v}^T, \mathcal{A}^*\mathcal{A}(\mathbf{u}\mathbf{v}^T + u_0\mathbf{v}^T + \mathbf{u}v_0^T + u_0v_0^T)\right\rangle \\ &+ \left\langle u_0v_0^T, \mathcal{A}^*\mathcal{A}(\mathbf{u}\mathbf{v}^T + u_0\mathbf{v}^T + \mathbf{u}v_0^T + u_0v_0^T)\right\rangle \end{aligned} \end{aligned}$$

Note that terms like $\mathcal{A}^*(\mathcal{A}(u_0v_0^T))$ are NOT symmetric. We can ignore the constant and higher order terms. Denote $x_0 = u_0v_0^T$, and $Q_0 = \mathcal{A}^*(\mathcal{A}(x_0))$ (which is not symmetric unless $\mathcal{A}^*$ happens to have its range be symmetric matrices), and the residual $\boxed{R_0 = Q_0 - \mathcal{A}^*(b)}$,

To find the Hessian at x_0, we fit the above to the form $f(x+x_0) = f(x_0) + \langle x, \nabla f(x_0)\rangle + \frac{1}{2}\langle x, \nabla^2 f(x_0)\, x\rangle + \ldots$ (and note that there is already a 1/2 included in the Hessian term). So

$$\begin{aligned} \left\langle (\mathbf{u};\mathbf{v}), \begin{pmatrix} H_{uu} & H_{uv} \\ H_{vu} & H_{vv}\end{pmatrix} (\mathbf{u};\mathbf{v})\right\rangle &= -2\langle \mathcal{A}^* b, \mathbf{u}\mathbf{v}^T\rangle + \langle \mathbf{u}\mathbf{v}^T, \mathcal{A}^*\mathcal{A}(x_0)\rangle + \\ &\quad \langle x_0, \mathcal{A}^*\mathcal{A}(\mathbf{u}\mathbf{v}^T\rangle + \langle \mathbf{u}v_0^T, \mathcal{A}^*\mathcal{A}(u_0\mathbf{v}^T + \mathbf{u}v_0^T)\rangle + \langle \mathbf{v}^T, \mathcal{A}^*\mathcal{A}(u_0\mathbf{v}^T + \mathbf{u}v_0^T)\rangle \end{aligned}$$

We have the $-2\langle \mathcal{A}^* b, \mathbf{u}\mathbf{v}^T\rangle$ term which could go with either H_{uv} or H_{vu}, so we split it into two and give half to each term. Also note that we are not really multiplying $H_{uu}\mathbf{u}$ but rather applying it in an operator sense, so it may not be standard matrix multiplication. That is, the terms like H_{uu} are really $(H_{uu}(u_0,v_0))\,(\mathbf{u})$. By regrouping, we have

$$\begin{aligned} (H_{uu}(u_0,v_0))\,(\mathbf{u}) &= \left(\mathcal{A}^*\mathcal{A}(\mathbf{u}v_0^T)\right) v_0 \\ (H_{uv}(u_0,v_0))\,(\mathbf{v}) &= \left(\mathcal{A}^*\mathcal{A}(u_0\mathbf{v}^T)\right) v_0 + R_0\mathbf{v} \\ (H_{vu}(u_0,v_0))\,(\mathbf{u}) &= \left(\mathcal{A}^*\mathcal{A}(\mathbf{u}v_0)^T)\right)^T u_0 + R_0^T\mathbf{u} \\ (H_{vv}(u_0,v_0))\,(\mathbf{v}) &= \left(\mathcal{A}^*\mathcal{A}(v_0\mathbf{v}^T)\right)^T u_0 \end{aligned}$$

For code, see `/Users/srbecker/Documents/MATLAB/smoothing_BruerTroppCevher/checkHessian.m`.

Sanity check To check that we are doing something reasonable, let's look at the linear terms. We have

$$f(\mathbf{u}+u_0,\mathbf{v}+v_0) = \ldots + \frac{1}{2}\langle \mathbf{u}v_0^T, \mathcal{A}^*\mathcal{A}(u_0v_0^T)\rangle + \frac{1}{2}\langle u_0v_0^T, \mathcal{A}^*\mathcal{A}(\mathbf{u}v_0^T)\rangle - \langle \mathcal{A}^*b, \mathbf{u}v_0^T\rangle +$$
$$\frac{1}{2}\langle u\mathbf{v}^T, \mathcal{A}^*\mathcal{A}(u_0v_0^T)\rangle + \frac{1}{2}\langle u_0v_0^T, \mathcal{A}^*\mathcal{A}(u_0\mathbf{v}^T)\rangle - \langle \mathcal{A}^*b, u_0\mathbf{v}^T\rangle + \ldots$$

This simplifies to

$$\langle \mathbf{u}, \nabla_{\mathbf{u}} f(u_0,v_0)\rangle = \frac{1}{2}\langle \mathbf{u}, Q_0v_0\rangle + \langle Q_0v_0, \mathbf{u}\rangle - \langle (\mathcal{A}^*b)v_0, \mathbf{u}\rangle = \langle \mathbf{u}, R_0v_0\rangle, \quad \text{hence} \quad \boxed{\nabla_{\mathbf{u}} f(u_0,v_0) = R_0v_0.}$$

This agrees with our previous definition. For v, we have

$$\langle \mathbf{v}, \nabla_{\mathbf{v}} f(u_0,v_0)\rangle = \frac{1}{2}\langle Q_0^Tu_0, \mathbf{v}\rangle + \frac{1}{2}\langle Q_0, u_0\mathbf{v}^T\rangle - \langle u_0^T(\mathcal{A}^*b), \mathbf{v}^T\rangle, \quad \text{hence} \quad \boxed{\nabla_{\mathbf{v}} f(u_0,v_0) = R_0^Tu_0.}$$

2 Matrix norms

From Alex Gittens, 3/18/10. Some facts

Define

$$\|A\|_{p\to q} = \max_{\|x\|_p\le 1} \|Ax\|_q.$$

For example, $\|A\|_2 = \|A\|_{2\to 2}$ is the spectral norm, and $\|A\|_{1\to\infty}$ is the maximum entry in absolute value, e.g., $\|\operatorname{vec}(A)\|_\infty$.

The dual norm is $\|A\|^* = \max_{\|X\|\le 1} \langle A,\, X\rangle$.

Facts

1. $\|A\|_{p\to q} \le \|U\|_{r\to q}\|V^T\|_{p\to r}$ for all U, V such that $A = UV^T$. This follows from sub-multiplicativity.
2. $\|A^T\|_{p\to q} = \|A\|_{q'\to p'}$ where $1/p + 1/p' = 1$ and similarly for q'. For example, $\|V^T\|_{1\to 2} = \|V\|_{2\to\infty}$.

In general, there is no closed form for the dual of the $p-q$ norm using p and q. Actually, there is, but not easy.

Generally, computing $\|A\|_{p\to q}$ is NP-Hard except for some combinations of $p, q \in \{1, 2, \infty\}$. See Joel's table.

2.1 Grothendieck

Main Grothendieck (Grothendiek?) bound, where κ_G is a constant.

$$\boxed{\gamma_2(A) \le \nu_1(A) \le \kappa_G\, \gamma_2(A)}$$

which also implies

$$\nu_1^*(A) \le \gamma_2^*(A) \le \kappa_G\, \nu_1^*(A)$$

Definition of γ_2 norm. Let $A = UV^T$. Using our elementary facts,

$$\|A\| \le \|U\|_{2\to\infty}\|V^T\|_{1\to 2} = \|U\|_{2\to\infty}\|V\|_{2\to\infty}$$

(Q: which norm was this? Seems like it should be $\|A\|_{1\to\infty}$, but that one we know how to compute).

Thus, define

$$\boxed{\gamma_2(A) = \inf_{U,V:UV^T=A} \|U\|_{2\to\infty}\|V\|_{2\to\infty}}$$

This can be computed exactly as an SDP; so can its dual, γ_2^*.

Now, we may really be interested in the following nuclear norm

$$\nu_1(A) = \|A\|^*_{\infty\to 1} = \inf_d \|d\|_{\ell_1} \,:\, A = \sum_i d_i m_i$$

where m_i is a rank 1 sign matrix. This is very hard to compute, but using Grothendieck's inequality, we can bound it. Not that $\|A\|_{\infty\to 1}$ is also hard.

2.2 misc

$u_i \otimes v_i = vu^T$ "usually". So $(u_i \otimes v_i)x = v\,\langle u,\, x\rangle$.

$$\|A\|^*_{p\to q} = \inf_{A=\sum_i u_i\otimes v_i}\left(\sum_i \|u_i\|_{p'}\|v_1\|_q\right)$$

(Q: should that last q be q'?)