Lecture 4 Part 1: Gradients and Inner Products in Other Vector Spaces

MIT 18.S096 Matrix Calculus For Machine Learning And Beyond

March 5, 2024

Table of contents

Riesz representation theorem

A *Hilbert space* is a continuous vector space with an inner (dot/scalar) product defined. For $\mathbb{R}^n$ (i.e., column vectors), we usually define the inner product as $\boldsymbol{x}\cdot\boldsymbol{y}=\boldsymbol{x}^T\boldsymbol{y}$. For $\mathbb{R}^{n\times m}$ (i.e., matrices), we usually define the inner product as $\text{sum}(\boldsymbol{A}\odot\boldsymbol{B})=(\text{vec}(A))^T(\text{vec}\,B)=\text{tr}(A^TB)$. Three properties of a valid inner product:

1. Symmetric: $x\cdot y=y\cdot x$
2. Linear: $x\cdot(\alpha y+\beta z)=\alpha(x\cdot y)+\beta(x\cdot z)$
3. Non-negative: $x\cdot x=\|x\|^2\geq 0$, $=0$ iff $x=0$

Setup. Let $f(x)$ be a function that maps from a Hilbert space to $\mathbb{R}$. We know that the derivative is the linear operator ("linear form") that takes a dx (a infinitesimal change in the input) to df (a infinitesimal change in the output):

$$df=f'(x)[dx].$$

Riesz representation theorem tells us that if we have a linear function that's "vector in number out", then it can be represented as a dot product with its input. So $f'(x)[dx]$ can be represented as the dot product between some vector and dx, and we call this vector the *gradient*:

$$df=f'(x)[dx]=(\nabla f)\cdot(dx).$$

An observation is that the gradient would always have the same "shape" as x.

General strategy from deriving the gradient. Start with df. Gradually massage it into a dot product between something and dx. That "something" would be the gradient.

Below we show some examples of this procedure from matrix-scalar functions. Note that the stuff on the LHS is equivalent to the stuff on the RHS, which is written in 18.06 style:

$$\begin{aligned} df &= (\nabla f)\cdot \text{dA} \\ &= \text{tr}((\nabla f)^T dA) \end{aligned} \quad \underset{\text{equivalent}}{\Leftrightarrow} \quad \nabla f=\begin{pmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} & \cdots \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Example: gradient of $\|A\|_F$ with respect to A

Function:

$$f(A_{m\times n}) = \|A\|_F = \sqrt{\mathrm{tr}(A^TA)}$$

Deriving the gradient:

$$\begin{aligned}
df &= \frac{1}{2\sqrt{\mathrm{tr}(A^TA)}}\, d(\mathrm{tr}(A^TA)) \quad \text{(scalar chain rule)} \\
&= \frac{1}{2\sqrt{\mathrm{tr}(A^TA)}}\, \mathrm{tr}[d(A^TA)] \quad \text{(trace is linear)} \\
&= \frac{1}{2\sqrt{\mathrm{tr}(A^TA)}} \mathrm{tr}[(dA)^TA + A^TdA] \quad \text{(matrix chain rule)} \\
&= \frac{1}{2\sqrt{\mathrm{tr}(A^TA)}}\Big(\mathrm{tr}[(dA)^TA] + \mathrm{tr}(A^TdA)\Big) \\
&= \frac{1}{2\sqrt{\mathrm{tr}(A^TA)}}\Big(\mathrm{tr}[A^T(dA)] + \mathrm{tr}(A^TdA)\Big) \quad (\mathrm{tr}(A) = \mathrm{tr}(A^T)) \\
&= \frac{1}{\sqrt{\mathrm{tr}(A^TA)}}\, \mathrm{tr}[A^T(dA)] \\
&= \frac{1}{\sqrt{\mathrm{tr}(A^TA)}}\, A\cdot dA \quad \text{(definition of the matrix dot product)} \\
&= \underbrace{\frac{A}{\sqrt{\mathrm{tr}(A^TA)}}}_{\text{gradient}} \cdot dA
\end{aligned}$$

Note that the gradient is simply A divided by $\|A\|_F$.

Example: gradient of x^TAy with respect to A

Function:

$$f(A_{m\times n}) = x^TAy$$

for some constant $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$.

Deriving the gradient:

$$\begin{aligned}
df &= x^TdAy \quad \text{(matrix product rule)} \\
&= \mathrm{tr}(x^TdAy) \\
&= \mathrm{tr}(yx^TdA) \\
&= \underbrace{(xy^T)}_{\text{gradient}} \cdot dA
\end{aligned}$$

Example: gradient of sum(A) with respect to A

Function:

$$f(A_{m\times n}) = \mathrm{sum}(A) = \mathbf{1}^TA\mathbf{1} = \mathrm{sum}(\mathrm{matrix}(1) \odot A)$$

Deriving the gradient (two ways are pretty much equivalent):

First way (using $\mathbf{1}^T A \mathbf{1}$):

$$\begin{aligned} df &= d(\mathbf{1}^T A \mathbf{1}) \\ &= \mathbf{1}^T dA \mathbf{1} \\ &= \mathrm{tr}(\mathbf{1}^T dA \mathbf{1}) \\ &= \mathrm{tr}(\mathbf{1}\mathbf{1}^T dA) \\ &= \underbrace{(\mathbf{1}\mathbf{1}^T)}_{\text{gradient}} \cdot dA \end{aligned}$$

Second way (using $\mathrm{sum}(\mathrm{matrix}(1) \odot A)$):

$$\begin{aligned} df &= \mathrm{sum}(\mathrm{matrix}(1) \odot dA) \quad \text{(both sum and hadamard product are linear)} \\ &= \underbrace{\mathrm{matrix}(1)}_{\text{gradient}} \cdot dA \end{aligned}$$

Lingering questions

- What would be the interpretation of the gradient if I define a weird but valid inner product?