CS 189 Fall 2018 Introduction to Machine Learning DIS1

1 Trace Derivatives

(a) Let $\mathbf{P}$ be a $p \times q$ matrix and $\mathbf{Q}$ be a $q \times p$ matrix. Compute $\frac{\partial \text{trace}(\mathbf{PQ})}{\partial \mathbf{P}}$.

Solution:

Review of abstract definition of derivatives

Let us review the last discussion where we defined derivatives. The main goal in this section is to clarify

- why derivatives of scalar functions with respect to matrices give rise to traces, as seen in Discussion 0.
- what the distinction between a derivative $\frac{\partial f}{\partial x}$ and a gradient $\nabla f(x)$ is.

For rigor we need to introduce a lot of notation here, so please read very carefully.

An abstract concept of a derivatives which is useful here is that of a Frechet derivative: It is a linear map $D_x f : \mathbf{X} \to \Re$ for a function $f : \mathbf{X} \to \Re$, (where $\mathbf{X}$ is the domain of f, and in our use cases $\mathbf{X} = \Re^{k \times d}$ with any $k \geq 1, d \geq 1$, i.e. including space of matrices and vectors) which satisfies the following inequality

$$f(x_0 + \Delta) - \underbrace{(f(x_0) + D_{x_0} f(x - x_0))}_{L_{x_0}(x)} = o(\|\Delta\|)$$

where $f(x) = o(g(x))$ if $\lim_{x \to 0} \frac{f(x)}{g(x)} = 0$ and $\Delta \in \mathbf{X}$. Note that $L_{x_0}(x)$ can be understood as the first order (or affine) approximation of f at point x_0.

Thus saying that a function $f : \mathbf{X} \to \Re$ is differentiable at some x_0 is equivalent to saying that there exists a linear map $D_{x_0} f$. The next important point to note is that **any linear map** on $\mathbf{X} = \Re^{k \times d}$ mapping to the real line, corresponds uniquely to an element $u \in \Re^{d \times k}$. In the case $d = 1$, i.e. in vector spaces, every linear map m applied on some $x \in \Re^{k \times d}$ can be written as $m(x) = ux$, for general $d \geq 1$, i.e. matrix space it is $m(X) = \text{tr}(UX)$.

What we call the derivative at x_0, denoted by $\frac{\partial f}{\partial x}(x_0) := \left.\frac{\partial f}{\partial x}\right|_{x = x_0}$, is now exactly the element in $\Re^{d \times k}$ which corresponds to the Frechet derivative, the linear map $D_{x_0} f$. People sometimes also refer to the transpose of this element in the space $\Re^k$ as a gradient.

To summarize our notation, for vector spaces we write

$$D_{\mathbf{x_0}} f(\mathbf{x}) = \begin{cases} \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x_0})\mathbf{x} \\ \nabla_{\mathbf{x}} f(\mathbf{x_0})^\top \mathbf{x} \end{cases}$$

which is why $\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x_0}) = \nabla_\mathbf{x} f(\mathbf{x_0})^\top$ and for matrix spaces we write

$$D_{\mathbf{X_0}} f(\mathbf{X}) = \mathrm{tr}\left(\frac{\partial f}{\partial \mathbf{X}}(\mathbf{X_0})\mathbf{X}\right)$$

Optional: Let's denote the ij-th element of matrix $\mathbf{X}$ by $X_{ij} = (\mathbf{X})_{ij}$. Now you can also derive that the matrix $\frac{\partial f}{\partial \mathbf{X}}(\mathbf{X_0})$ has to have the elements

$$\left(\frac{\partial f}{\partial \mathbf{X}}(\mathbf{X_0})\right)_{ij} = \frac{\partial f}{\partial X_{ji}}(\mathbf{X}_0)$$

as already claimed in Discussion 0. Do this as an exercise. For this purpose, you need to understand that $\frac{\partial f}{\partial X_{ji}}(\mathbf{X}_0)$ corresponds to the derivative of the function $g : \Re \to \Re$ at $z = (\mathbf{X}_0)_{ji}$ where g is defined by $g(z) = f(\mathbf{X}_0^{ji,z})$ where $\mathbf{X}_0^{ji,z}$ is element-wise defined by $(\mathbf{X}_0^{ji,z})_{k\ell} = (\mathbf{X}_0)_{k\ell}$ for all $(k,\ell) \neq (j,i)$ and $(\mathbf{X}_0^{ji,z})_{ji} = z$.

Proofs of identities

From the definition of the trace, we have

$$\begin{aligned} \frac{\partial \mathrm{trace}(\mathbf{PQ})}{\partial \mathbf{P}_{k\ell}} &= \frac{\partial \sum_{i=1}^p \sum_{j=1}^q P_{ij} Q_{ji}}{\partial P_{k\ell}} \\ &= Q_{\ell k}, \end{aligned}$$

where the last part follows since the term $P_{k\ell}$ appears in the sum once when $i = k$ and $j = \ell$, with a multiplicative scaling $Q_{\ell k}$. Collecting all of these partial derivatives into a matrix using $\left(\frac{\partial\,\mathrm{tr}(\mathbf{PQ})}{\partial \mathbf{P}}\right)_{\ell k} = \frac{\partial \mathrm{trace}(\mathbf{PQ})}{\partial \mathbf{P}_{k\ell}}$ (can be shown from first principle, see solution of Discussion 0) finally gives us the matrix derivative

$$\frac{\partial}{\partial \mathbf{P}} \mathrm{trace}(\mathbf{PQ}) = \mathbf{Q}$$

(b) Let $\mathbf{P}$ be a $p \times q$ matrix and $\mathbf{Q}$ be a $q \times q$ matrix. Compute $\frac{\partial \mathrm{trace}(\mathbf{PQP}^\top)}{\partial \mathbf{P}}$ at $\mathbf{P} = \mathbf{U}$.

Solution:

To prove the second identity we can use the first identity and the product rule which reads:

$$\left.\frac{\partial \mathrm{trace}(\mathbf{PQP}^\top)}{\partial \mathbf{P}}\right|_{\mathbf{P}=\mathbf{U}} = \left.\frac{\partial \mathrm{trace}(\mathbf{UQP}^\top)}{\partial \mathbf{P}}\right|_{\mathbf{P}=\mathbf{U}} + \left.\frac{\partial \mathrm{trace}(\mathbf{PQU}^\top)}{\partial \mathbf{P}}\right|_{\mathbf{P}=\mathbf{U}}.$$

It can also be obtained by first principle (refer to Discussion 0 if you are uncomfortable with this). If we set $\tilde{\mathbf{Q}} = \mathbf{QU}^\top$, then using the first identity we see that the second term above is

$$\frac{\partial \mathrm{trace}(\mathbf{PQU}^\top)}{\partial \mathbf{P}} = \frac{\partial \mathrm{trace}(\mathbf{P}\tilde{\mathbf{Q}})}{\partial \mathbf{P}} = \tilde{\mathbf{Q}} = \mathbf{QU}^\top.$$

We can set $\tilde{\mathbf{Q}} = \mathbf{UQ}$ and use the first identity to compute $\frac{\partial \text{trace}(\tilde{\mathbf{Q}}\mathbf{P}^\top)}{\partial \mathbf{P}}$. Recalling that $\text{trace}(\mathbf{A}) = \text{trace}(\mathbf{A}^\top)$, we then obtain using (a):

$$\frac{\partial \text{trace}(\mathbf{UQP}^\top)}{\partial \mathbf{P}} = \frac{\partial \text{trace}(\tilde{\mathbf{Q}}\mathbf{P}^\top)}{\partial \mathbf{P}} = \frac{\partial \text{trace}(\mathbf{P}\tilde{\mathbf{Q}}^\top)}{\partial \mathbf{P}_{k\ell}} = \tilde{\mathbf{Q}}^\top = \mathbf{Q}^\top\mathbf{U}^\top.$$

Putting everything together, we get

$$\frac{\partial \text{trace}(\mathbf{PQP}^\top)}{\partial \mathbf{P}} = \left.\frac{\partial \text{trace}(\mathbf{UQP}^\top)}{\partial \mathbf{P}}\right|_{\mathbf{P}=\mathbf{U}} + \left.\frac{\partial \text{trace}(\mathbf{PQU}^\top)}{\partial \mathbf{P}}\right|_{\mathbf{P}=\mathbf{U}} = \mathbf{Q}^\top\mathbf{U}^\top + \mathbf{QU}^\top = (\mathbf{Q}^\top + \mathbf{Q})\mathbf{U}^\top$$

2 Unitary invariance

(a) Prove that the regular Euclidean norm (also called the ℓ^2-norm) is unitary invariant; in other words, the ℓ^2-norm of a vector is the same, regardless of how you apply a rigid linear transformation to the vector (i.e., rotate or reflect). Note that rigid linear transformation of a vector $\mathbf{v} \in \mathbb{R}^d$ means multiplying by an orthogonal $\mathbf{U} \in \mathbb{R}^{d\times d}$.

Solution:

Recall that an orthogonal matrix $\mathbf{U}$ is one whose columns are orthonormal — i.e. each has norm 1 and their Euclidean inner products with each other are zero. If $\mathbf{U}$ is orthogonal then this implies that $\mathbf{U}^\top = \mathbf{U}^{-1}$ or $\mathbf{U}^\top\mathbf{U} = \mathbf{U}^{-1}\mathbf{U} = \mathbf{I}$.

Take a rotated or reflected version of v to then be $v_2 = \mathbf{U}v$ for an orthogonal matrix $\mathbf{U}$.

$$\|v_2\|_2^2 = \|\mathbf{U}v\|_2^2 = (\mathbf{U}v)^T(\mathbf{U}v) = v^T\mathbf{U}^T\mathbf{U}v = v^Tv = \|v\|_2^2$$

Take the square root of both sides; this is valid since the ℓ^2-norm is always non-negative.

$$\|v_2\|_2 = \|v\|_2$$

Because the lengths are preserved from this rigid linear transformation, geometrically you can see that orthogonal transformations are generalizations of rotations and reflections.

(b) Now show that the Frobenius norm of matrix $\mathbf{A}$ is unitary invariant. The Frobenius norm is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j=1}^n |a_{ij}|^2} = \sqrt{tr(\mathbf{A}^\top\mathbf{A})}$.

Solution:

Suppose that we apply the unitary transformation $\mathbf{U}$ again.

$$\|\mathbf{UA}\|_F^2 = tr((\mathbf{UA})^\top\mathbf{UA})$$

$$
\begin{aligned}
&= tr(\mathbf{A}^\top \mathbf{U}^\top \mathbf{U} \mathbf{A}) \\
&= tr(\mathbf{A}^\top \mathbf{A}) \\
&= \|\mathbf{A}\|_F^2
\end{aligned}
$$

3 Least Squares (using vector calculus)

(a) In ordinary least-squares linear regression, we typically have $n > d$ so that there is no $\mathbf{w}$ such that $\mathbf{X}\mathbf{w} = \mathbf{y}$ (these are typically overdetermined systems — too many equations given the number of unknowns). Hence, we need to find an approximate solution to this problem. The residual vector will be $\mathbf{r} = \mathbf{X}\mathbf{w} - \mathbf{y}$ and we want to make it as small as possible. The most common case is to measure the residual error with the standard Euclidean ℓ^2-norm. So the problem becomes:

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$

Where $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^n$. Derive using vector calculus an expression for an optimal estimate for $\mathbf{w}$ for this problem assuming $\mathbf{X}$ is full rank.

Solution: The work flow is as follows: We first find a critical point by setting the gradient to 0, then show that it is unique under the conditions in the question and finally that it is in fact a minimizer.

Let us first find critical points $\mathbf{w}_{OLS}$ such that the gradient is zero, i.e $\nabla_w \|\mathbf{X}\mathbf{w}_{OLS} - \mathbf{y}\|_2^2\big|_{\mathbf{w}=\mathbf{w}_{OLS}} = 0$. In order to take the gradient, we expand the ℓ^2-norm. First, note the following:

$$
\begin{aligned}
\nabla_w(\mathbf{w}^\top \mathbf{B}\mathbf{w}) &= (\mathbf{B} + \mathbf{B}^\top)\mathbf{w} \\
\nabla_w(\mathbf{w}^\top \mathbf{b}) &= \mathbf{b}
\end{aligned}
$$

We start by expanding the ℓ^2-norm:

$$
\begin{aligned}
&\nabla_w(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) \\
&= \nabla_w((\mathbf{X}\mathbf{w})^T(\mathbf{X}\mathbf{w}) - (\mathbf{X}\mathbf{w})^T(\mathbf{y}) - \mathbf{y}^T(\mathbf{X}\mathbf{w}) + \mathbf{y}^T\mathbf{y}) \quad \text{Combine middle terms, identical scalars.} \\
&= \nabla_w(\mathbf{w}^T\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}) \quad \text{Apply two derivative rules above} \\
&= (\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y} \\
&= 2\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y})
\end{aligned}
$$

Having computed the gradient, we now require it to vanish at the critical point $\mathbf{w} = \mathbf{w}_{OLS}$

$$
\begin{aligned}
\nabla_w \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2\big|_{\mathbf{w}=\mathbf{w}_{OLS}} &= 2\mathbf{X}^T(\mathbf{X}\mathbf{w}_{OLS} - \mathbf{y}) \\
&= 2\mathbf{X}^\top\mathbf{X}\mathbf{w}_{OLS} - 2\mathbf{X}^\top\mathbf{y} = 0
\end{aligned}
$$

$$\implies \mathbf{X}^\top \mathbf{X} \mathbf{w}_{OLS} = \mathbf{X}^\top \mathbf{y}$$

Because X is full rank, $\mathbf{X}^T\mathbf{X}$ is invertible (see question (b)) and thus there is only one vector which satisfies the last equation which reads: $\mathbf{w}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Therefore, there is only one unique critical point.

To show that this is the global minimizer, it suffices to show $\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2 \to \infty$ for $\|\mathbf{w}\|_2 \to \infty$. Because $\mathbf{X}$ is full rank, the matrix $\mathbf{X}^\top\mathbf{X}$ is positive definite and therefore we have the eigendecomposition

$$\mathbf{X}^\top\mathbf{X} = \sum_i \lambda_i \mathbf{v}_i^\top \mathbf{v}_i \tag{1}$$

with eigenvalues $\lambda_i > 0$ and orthonormal eigenvectors $\mathbf{v}_i$ and therefore by writing

$$\mathbf{w} = \sum_i \mu_i \mathbf{v}_i \tag{2}$$

we get

$$\begin{aligned}
\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 &= \mathbf{w}^T\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y} \\
&\geq \sum_i \mu_i^2\lambda_i - 2\|\mathbf{w}\|_2\left\|\mathbf{X}^\top\mathbf{y}\right\|_2 + \mathbf{y}^\top y = \sum_i \mu_i^2\lambda_i - 2\|\boldsymbol{\mu}\|_2\left\|\mathbf{X}^\top\mathbf{y}\right\|_2 + \mathbf{y}^\top\mathbf{y} \\
&\geq \min(\lambda_1, \ldots, \lambda_d) \cdot \|\mu\|_2^2 - 2\|\boldsymbol{\mu}\|_2\left\|\mathbf{X}^\top\mathbf{y}\right\|_2 + \mathbf{y}^\top\mathbf{y}
\end{aligned}$$

(in the last step we used the Cauchy Schwarz inequality) where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$, and $\|\mu\|_2 = \|\mathbf{w}\|_2$ because the $\mathbf{v}_i$ are orthonormal. Therefore $\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2$ goes to ∞ as $\|\mu\|_2 = \|\mathbf{w}\|_2 \to \infty$, which shows that$\mathbf{w}_{OLS}$ is the global minimizer of the loss.

(b) How do we know that $\mathbf{X}^\top\mathbf{X}$ is invertible?

Solution: Matrix $\mathbf{X}$ is said to be full rank if n $\geq$ d and its columns are not linear combinations of each other. In this case, $\mathbf{X}^\top\mathbf{X}$ will be positive definite and therefore invertible. If $\mathbf{X}$ is not full rank, at least one of the columns will be a linear combination of the other columns. In this case, the rank of $\mathbf{X}$ will be less than n and $\mathbf{X}^\top\mathbf{X}$ will not be invertible.

In this question, we know that $\mathbf{X}$ has full rank, so if we can show that the rank of $\mathbf{X}$ is equivalent to the rank of $\mathbf{X}^\top\mathbf{X}$, then $\mathbf{X}^\top\mathbf{X}$ has full rank and is therefore invertible. Let us show the ranks are equivalent using nullspaces. Suppose $\mathbf{v}$ is in the nullspace of $\mathbf{X}^\top\mathbf{X}$ meaning $\mathbf{X}^\top\mathbf{X}\mathbf{v} = \mathbf{0}$:

$$\begin{aligned}
\mathbf{X}^\top\mathbf{X}\mathbf{v} &= \mathbf{0} \\
\mathbf{v}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v} &= 0 \\
(\mathbf{X}\mathbf{v})^\top(\mathbf{X}\mathbf{v}) &= 0 \\
\|\mathbf{X}\mathbf{v}\|_2^2 &= 0 \\
\mathbf{X}\mathbf{v} &= 0 \qquad \text{Because the only vector whose length is 0 is the } \mathbf{0} \text{ vector.}
\end{aligned}$$

From this we can see that any $\mathbf{v}$ which is in nullspace of $\mathbf{X}^\top\mathbf{X}$ also needs to be in the nullspace of $\mathbf{X}$. Since $\mathbf{X}$ and $\mathbf{X}^\top\mathbf{X}$ have the same null space, then $\mathbf{X}^\top\mathbf{X}$ should also be full rank and therefore invertible.

(c) What should we do if $\mathbf{X}$ is not full rank?

Solution: (Basic idea) If $\mathbf{X} \in \mathbf{R}^{n\times d}$ is not full rank, there is no unique answer. As we will see later, this is not an issue in ridge regression where we add a penalization to the loss function (thus change the loss function) which forces a unique solution. Another possibility is to use the solution that minimizes the norm of $\mathbf{w}$ (in later lectures we will see why that might be a good thing to do).

The minimum norm solution can be found by using the pseudo-inverse of $\mathbf{X}^\top\mathbf{X}$. The pseudo-inverse of an arbitrary matrix $\mathbf{X}$ is denoted as $\mathbf{X}^\dagger$. More intuitively, $\mathbf{X}^\dagger$ behaves most similarly to the inverse: it is the matrix that, when multiplied by $\mathbf{X}$, minimizes distance to the identity. $\mathbf{X}^\dagger = \text{argmin}_{W\in\mathbf{R}^{n\times d}}\|\mathbf{X}\mathbf{W} - \mathbf{I}_m\|_F$.