CS 189:289A Introduction to Machine Learning
Spring 2021 Jonathan Shewchuk **DIS2**

1 Back to Basics: Linear Algebra

Let $X \in \mathbb{R}^{n\times m}$. We introduce some important terms and notation

$w \leftarrow Xv$ $\quad v \in \mathbb{R}^m$
$w = Xv$ $\quad w \in \mathbb{R}^n,$

The **Columnspace**, also called the range, or span, of X is $Range(X) := \{y \mid y = Xv\}$.
The **Rowspace** is $Row(X) := \{y \mid y = X^\top v\}$.
The **Nullspace**, or Kernel, of X is defined is $\mathcal{N}(x) := \{v \mid Xv = 0\}$.
The **Orthongal Complement** of a subspace, U, is a subspace, $U^\perp$ such that $u \in U, u' \in U^\perp \implies \langle u, u'\rangle = 0$

$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

For this problem We do not assume that X has full rank.

$Col(A) = \mathbb{R}^2$

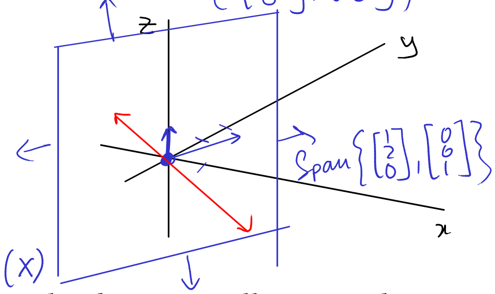
(a) Check the following facts:

$Row(A) = Col(A^T) = Span\left\{\begin{bmatrix}1\\2\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}\right\}$

(i) The $Row(X) = Range(X^\top)$
Rows of X are cols of X^T.

$\begin{bmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{bmatrix}$

$Null(A) = Span\left\{\begin{bmatrix}0\\0\\1\end{bmatrix}\right\}$



(ii) The $\mathcal{N}(X)^\perp = Row(X)$ (Null)
$v \in Null(X) \Rightarrow Xv = \underline{0} \Rightarrow x_i^T v = 0 \;\forall 1 \le i \le n$
$\therefore v \perp$ all vectors in $Row(X) \Rightarrow Row(X) = Null(X)^\perp$

$v \in Null(X)$
$\Updownarrow$
$v \perp Row(X)$

$(Xv)_i = x_i^T v = 0 \;\; \forall i$
$\therefore v \in Null(X)$

(iii) $\underline{\mathcal{N}(X^\top X)} = \underline{\mathcal{N}(X)}$ *Hint: if $v \in \mathcal{N}(X^\top X)$, then $v^\top X^\top X v = 0$.*
$v^T X^T X v = 0 \Rightarrow (Xv)^T (Xv) = 0 \Rightarrow \|Xv\|_2^2 = 0 \Rightarrow Xv = 0$
$\Rightarrow v \in Null(X)$
$v \in Null(X) \Rightarrow v \in Null(X^T X) \because X^T X v = X^T 0 = 0$

(iv) $Row(X^\top X) = Range(X^\top X) = Row(X)$ *Hint: Use the relationship between nullspace and rowspace.* $X^T X$ is symm $\because (X^T X)^T = X^T (X^T)^T = X^T X$
$\therefore Col(X^T X) = Row((X^T X)^T) = Row(X^T X)$
From (iii) $\quad Null(X)^\perp = Row(X) \quad Null(X^T X)^\perp = Null(X)^\perp \Rightarrow Row(X^T X) = Row(X)$ ✓

$(AB)^T = B^T A^T$
$(AB)^{-1} = B^{-1} A^{-1}$

(b) We now prove an important result of linear algebra, the Rank-Nullity theorem. Let $Rank(X) = dim(Range(X)) = dim(Row(X))$ and $Nullity(X) = dim(\mathcal{N}(X))$. The Rank nullity theorem says that for $X \in \mathbb{R}^{nxm}$ we have

$$Rank(X) + Nullity(X) = m$$

Use the above results to prove this theorem. *Hint: The complementary subspace theorem says that for a vector space V and subspace U, we can always find a complementary subspace $U^\perp$ such that $U + U^\perp = V$*

2 Probability Review

There are n archers all shooting at the same target (bulls-eye) of radius 1. Let the score for a particular archer be defined to be the distance away from the center (the lower the score, the better, and 0 is the optimal score). Each archer's score is independent of the others, and is distributed uniformly between 0 and 1. What is the expected value of the worst (highest) score?

(a) Define a random variable Z that corresponds with the worst (highest) score.

(b) Derive the Cumulative Distribution Function (CDF) of Z.

(c) Let X be a non-negative random variable. The Tail-Sum formula states that

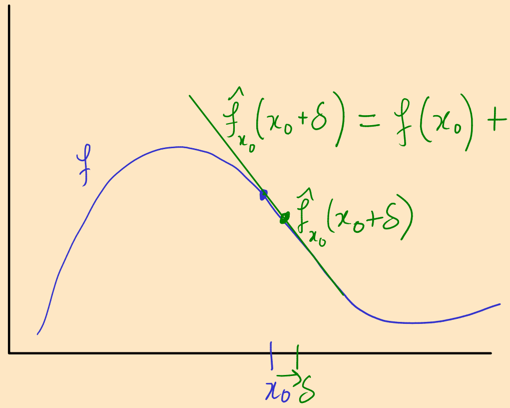
$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)dt$$

Using both the Tail-Sum formula and the CDF of Z derived above, calculate the expected value of Z *Hint: Write $\mathbb{P}(X \geq t)$ in terms of the CDF of X*

(d) Consider what happens to $\mathbb{E}[Z]$ as $n \to \infty$. Does this match your intuition?

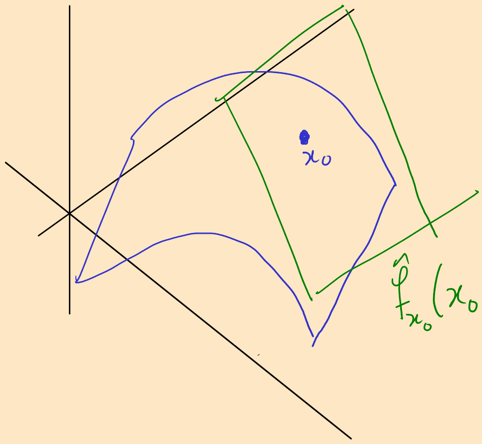
$f: \mathbb{R} \to \mathbb{R}$

$f: \mathbb{R}^n \to \mathbb{R}$



$$\hat{f}_{x_0}(x_0+\delta) = f(x_0) + \left(\frac{df}{dx}\Big|_{x_0}\right)\delta$$

$\hat{f}_{x_0}(x_0+\delta)$



$$\hat{f}_{x_0}(\underbrace{x_0+\Delta}_{\in \mathbb{R}^n}) = f(x_0) + \underbrace{\boxed{\frac{\partial f}{\partial x}}}_{\text{deriv}}\Delta$$

$[\quad\quad]\,[\;]$

$f = x^2 + y^2$

$\frac{\partial f}{\partial x} = 2x$

$\nabla_x f \in \mathbb{R}^n$ $\quad \underbrace{\frac{\partial f}{\partial x}}_{\in \mathbb{R}^{1\times n}} : \mathbb{R}^n \to \mathbb{R}$, linear

$\underbrace{\left[- \frac{\partial f}{\partial x_i} -\right]}_{n}$ "derivative"

$f: \mathbb{R}^n \to \mathbb{R}^m$

$$\hat{f}_{x_0}(\underbrace{x_0+\Delta}_{\in\mathbb{R}^n}) = f(x_0) + \boxed{\frac{\partial f}{\partial x}}\Delta$$

$\frac{\partial f}{\partial x}$: $\mathbb{R}^n \to \mathbb{R}^m$ linear, $\in \mathbb{R}^{m\times n}$

$\left[\begin{matrix} = \\ = \end{matrix}\right][\;]$

"Jacobian"

$\nabla_x f = \left(\frac{\partial f}{\partial x}\right)^T \in \mathbb{R}^{n\times m}$

$$\begin{array}{c} \\ j \to \end{array}\left[\begin{matrix} \\ \frac{\partial f_j}{\partial x_i} \\ \\ \end{matrix}\right]$$
$\uparrow$
i

$f: \mathbb{R}^{m\times n} \to \mathbb{R}$

$$\hat{f}_{x_0}(x_0 + \Delta) = \underline{f(x_0)} + \operatorname{trace}\left(\underbrace{\frac{\partial f}{\partial x}}_{n\times m}\overbrace{\Delta}^{m\times n}\right)$$

$A, B \in \mathbb{R}^{m\times n}$

$\sum_{ij} A_{ij} B_{ij}$

$$\left\langle \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\rangle_F = 1+0+0+1 = 2$$

$$= \operatorname{tr}\left(\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$$\operatorname{tr}\left(\begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 1 & 9 & 3 \end{bmatrix}\right) = 1+2+3 = 6$$

3 Vector Calculus

[1]

Below, $\mathbf{x} \in \mathbb{R}^d$ means that $\mathbf{x}$ is a $d \times 1$ column vector with real-valued entries. Likewise, $\mathbf{A} \in \mathbb{R}^{d \times d}$ means that $\mathbf{A}$ is a $d \times d$ matrix with real-valued entries. In this course, we will by convention consider vectors to be column vectors.

Consider $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{d \times d}$. In the following questions, $\nabla_{\mathbf{x}}$ denotes the gradient with respect to $\mathbf{x}$, which, by convention, is a column vector.

Calculate the following derivatives.

$$\begin{bmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_d \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = w$$

(a) $\underbrace{\nabla_{\mathbf{x}}(\mathbf{w}^\top \mathbf{x})}_{\text{linear}}$ $\qquad f(x) = \sum_{i=1}^{d} w_i x_i \qquad \frac{\partial f}{\partial x_i} = w_i$

$f(x) = \underline{a}x$ $\qquad \frac{\partial f}{\partial x} = w^T$

(b) $\nabla_{\mathbf{x}}(\underbrace{\mathbf{w}^\top \mathbf{A}}_{u^T}\mathbf{x}) = A^T w \qquad \frac{\partial f}{\partial x} = w^T A$

$tr(w^T \Delta x) = tr(x w^T \Delta)$

(c) $\nabla_{\mathbf{A}}(\mathbf{w}^\top \mathbf{A}\mathbf{x})$ $\qquad f(A) = w^T A x \qquad f(A + \Delta) = w^T(A + \Delta)x = w^T A x + w^T \Delta x$

$= f(A) + tr(\underbrace{\boxed{x w^T}}_{\frac{\partial f}{\partial A}}\Delta)$

$\| \; w x^T$

$tr(ABC) = tr(CAB) = tr(BCA)$

$\sum w_j A_{ji} x_i$

(d) $\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x})$

$\frac{\partial f}{\partial A_{ij}} = w_j x_i \qquad w x^T$

(e) $\nabla^2_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x})$

Now let's apply our identities derived above to a practical problem. Given a design matrix $X \in \mathbb{R}^{nxd}$ and label vector $Y \in \mathbb{R}^n$, the Ordinary least squares regression problem becomes

$$w^* = min_w \frac{1}{2}\|Xw - Y\|_2^2$$

(f) Using parts (a) - (e), derive a necessary condition for w^*. *Note: We do not necessarily assume X is full rank!*

---

[1] Good resources for matrix calculus are:

- The Matrix Cookbook: `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
- Wikipedia: `https://en.wikipedia.org/wiki/Matrix_calculus`
- Khan Academy: `https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives`
- YouTube: `https://www.youtube.com/playlist?list=PLSQl0a2vh4HC5feHa6Rc5c0wbRTx56nF7`.