CS 189 Introduction to Machine Learning
Summer 2019 Marc Khoury & Brijen Thananjeyan **DIS2**

1 Back to Basics: Linear Algebra

Let $X \in \mathbb{R}^{n \times d}$. We do not assume that X is full rank.

(a) Give the definition of the rowspace, columnspace, and nullspace of X.

(b) Check the following facts:

  (a) The rowspace of X is the columnspace of $X^\top$, and vice versa.
  (b) The nullspace of X and the rowspace of X are orthogonal complements.
  (c) The nullspace of $X^\top X$ is the same as the nullspace of X. *Hint: if v is in the nullspace of $X^\top X$, then $v^\top X^\top X v = 0$.*
  (d) The columspace and rowspace of $X^\top X$ are the same, and are equal to the rowspace of X. *Hint: Use the relationship between nullspace and rowspace.*

2 Concentration Inequalities

For a given random variable, we are often interested in computing bounds on its tail, or on the probability that it deviates from its mean. In this problem we will prove a concentration inequality for *sub-exponential* random variables. A random variable X with mean $E[X] = \mu$ is *sub-exponential* if there are non-negative parameters (v, b) such that

$$E[e^{\lambda(X-\mu)}] \leq e^{\frac{1}{2}v^2\lambda^2} \text{ for all } |\lambda| < \frac{1}{b} \tag{1}$$

We will prove that if X is sub-exponential,

$$P(X \geq \mu + t) \leq \begin{cases} e^{-\frac{t^2}{2v^2}} & \text{if } 0 \leq t \leq \frac{v^2}{b} \\ e^{-\frac{t}{2b}} & \text{if } t > \frac{v^2}{b} \end{cases} \tag{2}$$

(a) We will prove the case for $\mu = 0$. First, use a Chernoff bound to show that

$$P(X \geq t) \leq \exp\{-\lambda t + \frac{\lambda^2 v^2}{2}\} \tag{3}$$

for any $\lambda \in [0, \frac{1}{b})$.

(b) Define $g(\lambda) = -\lambda t + \frac{\lambda^2 v^2}{2}$. Compute the optimal values of λ which minimize g under the constraint that $\lambda \in (0, \frac{1}{b}]$.

(c) Use these values to deduce the tightest possible bound for the inequality from the first part, completing the proof.

3 Vector Calculus

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, $\frac{\partial}{\partial \mathbf{x}}$ denotes the derivative with respect to $\mathbf{x}$, while $\nabla_{\mathbf{x}}$ denote the gradient with respect to $\mathbf{x}$. Compute the following:

(a) $\frac{\partial \mathbf{w}^T \mathbf{x}}{\partial \mathbf{x}}$ and $\nabla_{\mathbf{x}}(\mathbf{w}^T \mathbf{x})$

(b) $\frac{\partial (\mathbf{w}^T \mathbf{A} \mathbf{x})}{\partial \mathbf{x}}$ and $\nabla_{\mathbf{x}}(\mathbf{w}^T \mathbf{A} \mathbf{x})$

(c) $\frac{\partial (\mathbf{w}^T \mathbf{A} \mathbf{x})}{\partial \mathbf{w}}$ and $\nabla_{\mathbf{w}}(\mathbf{w}^T \mathbf{A} \mathbf{x})$

(d) $\frac{\partial (\mathbf{w}^T \mathbf{A} \mathbf{x})}{\partial \mathbf{A}}$ and $\nabla_{\mathbf{A}}(\mathbf{w}^T \mathbf{A} \mathbf{x})$

(e) $\frac{\partial (\mathbf{x}^T \mathbf{A} \mathbf{x})}{\partial \mathbf{x}}$ and $\nabla_{\mathbf{x}}(\mathbf{x}^T \mathbf{A} \mathbf{x})$

(f) $\nabla_{\mathbf{x}}^2(\mathbf{x}^T \mathbf{A} \mathbf{x})$