ECE C147/C247, Winter 2023 **Discussion 1**
Neural Networks & Deep Learning Prof. J. Kao
UCLA; Department of ECE TAs: T. Monsoor, R. Gore, K. Kamalakannan

1 Probability basics

In statistical signal processing, manipulation of probability expressions is very important. The two tools from probability theory that we will be using frequently to manipulate the expressions are:

- Law of total probability
- Probability chain rule

1.1 Law of total probability

If $A_1, A_2, \ldots, A_n$ forms a partition of the sample space S, then the probability of an event B is given by

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$

Using the definition of conditional probability, we can rewrite the above expression as

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Thus, the probability of the event B is the sum of the conditional probabilities $P(B|A_i)$ weighted by the probability $P(A_i)$.

1.2 Probability chain rule

For any events A and B, we have

$$P(A \cap B) = P(A)P(B|A)$$

More generally, for any events $A_1, A_2, \ldots, A_n$ we get

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = \prod_{k=1}^{n} P(A_k| \cap_{j=1}^{k-1} A_j)$$

You can use induction to show the general version of the probability chain rule.

1.3 Practice problems on probability basics

1. A computer manufacturer uses chips from three sources. Chips from sources A, B, and C are defective with probabilities 0.005, 0.001, and 0.010, respectively. If a randomly selected chip is found to be defective, find:

   (a) The probability that the manufacturer was A

   (b) The probability that the manufacturer was C

   Assume that the proportions of chips from A, B, and C are 0.5, 0.1, and 0.4, respectively.

2. A box contains three coins. One of the coins is a two-headed coin, the second is a fair coin, and the third is a biased coin with $P(head) = p$. One of the coins is picked at random and flipped:

   (a) Find the probability that the coin shows head. What is it's maximum value?

   (b) If the coin shows head, find the probability that it is the fair coin. What are it's maximum and minimum values?

3. Suppose that the reliability of a chest X-ray test for the detection of tuberculosis is specified as follows: of people with tuberculosis, 90% of the X-ray examinations detect the disease but 10% go undetected. Of people free of tuberculosis, 99% of the X-rays are judged free of the disease, but 1% are diagnosed as showing the tuberculosis. From a large population of which only 0.1% have tuberculosis, one person is selected at random, given a chest X-ray, and the radiologist reports the presence of tuberculosis. What is the probability that the person has tuberculosis?

2 Linear Algebra basics

Proficiency in linear algebra is required to have a solid foundation in machine learning. The two linear algebra concepts that we will be using repeatedly in this course are:

- Decomposition of matrices
- Derivatives of vectors and matrices

In this discussion we will focus on the decomposition of matrices, namely:

- Eigendecomposition
- Singular value decomposition

2.1 Eigendecomposition

Every real symmetric matrix, $A \in \mathbf{R}^{n\times n}$, can be factored as

$$A = Q\Delta Q^T$$

where $\Delta = diag(\lambda_1, \lambda_2, \cdots, \lambda_n) \in \mathbf{R}^{n\times n}$. More specifically we have,

- The set $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ are called the eigenvalues of A and can be found by solving the equation

$$det(A - \lambda I) = 0$$

- The columns of Q are an orthonormal set of n eigenvectors and can be found by solving the equation

$$Av_i = \lambda_i v_i$$

2.2 Singular value decomposition

Let $A \in \mathbf{R}^{m\times n}$ be a matrix with rank r, then there exists orthogonal matrices $U \in \mathbf{R}^{m\times m}$ and $V \in \mathbf{R}^{n\times n}$ such that

$$A = U\Sigma V^T$$

where $\Sigma = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}$, $S = diag(\sigma_1, \sigma_2, \cdots, \sigma_r) \in \mathbf{R}^{r\times r}$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. More specifically, we have

- The set $\{\sigma_1, \sigma_2, \cdots, \sigma_r\}$ are called the non-zero singular values of A and can be calculated as

$$\sigma_i(A) = \lambda_i^{\frac{1}{2}}(A^T A) = \lambda_i^{\frac{1}{2}}(AA^T)$$

- The columns of U are called the left singular vectors of A (and are the orthonormal eigenvectors of AA^T).
- The columns of V are called the right singular vectors of A (and are the orthonormal eigenvectors of $A^T A$).

2.3 Practice problems on linear algebra basics

1. Show the following properties for matrices
   (a) If $b^T Ab > 0$ for all $b \in \mathbf{R}^n$, then all eigenvalues of A are positive.
   (b) If $A \in \mathbf{R}^{n\times n}$ is an orthogonal matrix, then all eigenvalues of A have norm 1.
   (c) If $A \in \mathbf{R}^{m\times n}$ is a matrix with rank r, then

$$\sigma_i(A) = \lambda_i^{\frac{1}{2}}(AA^T)$$

2. Determine if the following statements are true or false. If they are true, give a justification. If they are false, give a counterexample.
   (a) If $\mathbf{A}$ is an orthogonal matrix, then $\mathbf{A}^T$ is also an orthogonal matrix.
   (b) If $\mathbf{A}$ and $\mathbf{B}$ are orthogonal $n \times n$ matrices, then $\mathbf{AB}$ is orthogonal.
   (c) If $\mathbf{A}$ and $\mathbf{B}$ are orthogonal $n \times n$ matrices, then $\mathbf{A} + \mathbf{B}$ is orthogonal.
   (d) If the column vectors of $\mathbf{A}$ are orthonormal, then the row vectors of $\mathbf{A}$ must also be orthonormal.