CS 181 math review section

1 Linear Algebra

1.1 Scalars and Vectors

A **scalar** is a single, real-valued number. We write scalars in lowercase: x. A **vector** is an ordered list of scalars. We write vector names in bold lowercase, and the vector itself as a column of scalars:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{bmatrix}$$

where m is the dimension of the vector ($\mathbf{x} \in \mathbb{R}^m$) and $x_1, \dots, x_m$ are its scalar elements. We follow Bishop and use column vectors by default. Assume that $\mathbf{x}^\top$ is a row vector.

1.2 Matrices

An $n \times m$ matrix is a grid of scalars with n rows and m columns. If we call the matrix $\mathbf{A}$, the element at the i^{th} row and j^{th} column of $\mathbf{A}$ is A_{ij}. Technically, a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$.

The **transpose** of a matrix $\mathbf{A}$ switches the rows and columns, i.e. if $\mathbf{B} = \mathbf{A}^\top$, then $B_{ji} = A_{ij}$ for all values of i and j. $(\mathbf{A}^\top)^\top = \mathbf{A}$. $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$.

The matrix $\mathbf{A}$ is **symmetric** if $A_{ij} = A_{ji}$. Only square matrices can be symmetric. $\mathbf{A}^\top \mathbf{A}$ is symmetric.

The matrix $\mathbf{A}$ is **orthogonal** if $\mathbf{A}^\top = \mathbf{A}^{-1}$.

A **diagonal** matrix is a matrix in which the off-diagonal elements are zero.

1.3 Matrix Multiplication Properties

We expect that you are familiar with matrix multiplication. Recall these properties:

Commutativity does not hold: $\mathbf{AB} \neq \mathbf{BA}$ for some $\mathbf{A}$ and $\mathbf{B}$

Associativity does hold: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$

Identity: $\mathbf{IA} = \mathbf{A}$ and $\mathbf{AI} = \mathbf{A}$

1.4 Matrix Inverse

The **inverse** of a matrix $\mathbf{A}$ is the unique matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is identity matrix. Only square matrices have an inverse, and they are generally invertible. In a rare case a square matrix may not be invertible, in which case it is said to be singular. If $\mathbf{A}$ is a rectangular matrix, the **Moore-Penrose pseudoinverse** is defined as $\mathbf{B} = \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top$ such that $\mathbf{B}\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the $m \times m$ identity matrix if $\mathbf{A}$ is an $n \times m$ matrix.

1.5 Matrix Rank

The rank of a matrix is the dimension of the vector space spanned by its rows or columns. A matrix is full-rank if its rank equals its smaller dimension, and rank-deficient otherwise. A square matrix has an inverse if and only if it is full-rank.

1.6 Matrix Determinant

The general formula is quite complicated, but you should know that $\det(\mathbf{A}) = 0$ if and only if $\mathbf{A}$ is singular (i.e. has no inverse). This is due to the property that $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$, which is undefined if $\det(\mathbf{A}) = 0$. The determinant of a diagonal matrix is the product of its non-zero entries.

1.7 Eigenvalues and Eigenvectors

A matrix $\mathbf{A}$ is said to have eigenvector $\mathbf{v}$ with eigenvalue λ if we can write

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

i.e. the matrix only transforms the vector by its magnitude (i.e. stretching without rotating/shearing) with the amount of stretching given by λ. You will not need to find eigenvalues/eigenvectors, but will need to understand their significance in several parts of the course.

1.8 Positive Definite Matrices

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ is **positive definite** if it satisfies the property

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$$

and **positive semi-definite** if it satisfies

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$$

for every non-zero vector $\mathbf{x} \in \mathbb{R}^m$. Positive definite matrices have all eigenvalues > 0 and positive semi-definite matrices have all eigenvalues ≥ 0.

2 Vector Calculus

2.1 Differentiation

You should be familiar with single-variable differentiation, including properties like

$$\text{Chain rule: } \frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$$
$$\text{Product rule: } \frac{d}{dx}f(x)g(x) = f'(x)g(x) + f(x)g'(x)$$
$$\text{Linearity: } \frac{d}{dx}(af(x) + bg(x)) = af'(x) + bg'(x)$$

The multivariate case is similar, but now we consider the partial derivative of each pair of input and output dimensions and end up with a matrix:

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

(A partial derivative is just like the single-variable derivative, where you treat each variable except the one you're differentiating with respect to as a constant.)

2.2 Gradient Vector

If f is scalar-valued, its derivative is a column vector we call the gradient vector:

$$\frac{df(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \dots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

The gradient vector points in the direction of steepest ascent in $f(\mathbf{x})$. This is useful for optimization.

Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. The same is true here, using the condition $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can search for local minima numerically using **gradient descent**: We start with an initial guess $\mathbf{w}_0$, and then at each step i we update our guess by going in the direction of greatest descent (opposite the direction of the gradient vector)

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \eta \frac{df(\mathbf{w})}{d\mathbf{w}}$$

where η is a learning rate. We stop when the value of the gradient is close to 0.

2.3 Convexity

A scalar-valued function $f(\mathbf{x})$ is convex if the line segment between any two points on the graph of the function lies above or on the graph.
If the line always lies above, the function is strictly convex and the graph of f is cup-shaped. This guarantees that f has a unique global minimum, and that gradient descent will be able to find it. (We will use gradient descent on non-convex functions anyway).

2.4 Logarithms

Recall the properties of logs:

$$\text{Product rule: } \log(ab) = \log(a) + \log(b)$$
$$\text{Power rule: } \log(a^b) = b\log(a)$$

Thanks to linearity of the derivative, we can make our lives easier by taking the logarithm of the function we are minimizing before differentiating. (Minimizing the log-objective is equivalent to minimizing the objective because the logarithm function is monotonically increasing.)

2.5 Matrix cookbook

`http://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf` contains all of the facts about matrices and their derivatives. We recommend keeping it open while you attempt derivations.

3 Probability

3.1 Random Variables

A random variable can either be discrete or continuous. A discrete random variable X takes one of m values from sample space $\mathcal{X}$, each with a corresponding probability $p(x)$ for $x \in \mathcal{X}$.

We say that $x \sim X$ (x is sampled from X) when the value of x is picked in accordance with the distribution of X. Note that it is also common to notate the probability mass function of $x \sim X$ as $p(x)$.

A continuous random variable can take on a continuous range of values. One example of a continuous random variable is the amount of rain in a given day (it can be 1 inch, 1.25 inches, 6.245 inches, etc.). For continuous random variables, we will use $p(x)$ for the probability density function. Note that this is the same notation used for the probability mass function in discrete random variables, but in this case $p(x)$ should be thought of as not the probability of x but the probability *density* at x. Among other things, this means $p(x)$ can be greater than 1.

3.2 Expected Value

The **expected value** (or *expectation, mean*) of a random variable can be thought of as the "weighted average" of the possible outcomes of the random variable.
For discrete random variables:

$$\mathbb{E}_{x\sim p(x)}[X] = \sum_{x\in\mathcal{X}} x \cdot p(x)$$
$$\mathbb{E}[f(X)] = \sum_{x\in\mathcal{X}} f(x)p(x)$$

For continuous random variables:

$$\mathbb{E}[X] = \int_{\mathcal{X}} x \cdot p(x)dx$$
$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)p(x)dx$$

The most important property of expected values is the **linearity of expectation**. For **any** two random variables X and Y, scaling coefficients a and b, and a constant c, the following property holds:

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

The above is true regardless of whether X and Y are dependent or independent.

3.3 Variance

The variance of a random variable is its expected squared deviation from its mean

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

In other words, variance is a measure of the spread of a random variable. High variance variables are more spread out.

3.4 Conditional Probability

Receiving information about the value of a random variable Y can change the distribution of another variable X. We write the new conditional random variable as $X|Y$, and the new conditional distribution as $p(x|y)$.

Note that it is common to determine the expectation and variance of a variable. If X and Y are random variables, then $\mathbb{E}[X|Y]$ is a random variable too, because it can take on several values depending on Y.

Adam's law (law of iterated expectations) gives

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

There is an analogous property for variances (Eve's Law, or law of total variance)

$$\text{var}[X] = \mathbb{E}[\text{var}[X|Y]] + \text{var}[\mathbb{E}[X|Y]]$$

3.5 Joint Distributions

A joint distribution is a distribution over 2 or more random variables. For example, say you have the variables X and Y. The joint distribution is the function $p(x, y) = p(X = x, Y = y)$. When given a joint distribution, it is common to want to "marginalize" one or more of the variables. To do this, we use the "sum" rule, which is that

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$
$$p(x) = \int_{\mathcal{Y}} p(x, y) dy$$

where the first equation is for discrete random variables, and the second for continuous variables.

3.6 Product Rule

The product rule gives the joint probability $p(x, y)$ as the product of a conditional probability and a marginal probability:

$$\begin{aligned} p(x, y) &= p(x|y)p(y) \\ &= p(y|x)p(x) \end{aligned}$$

which can be extended to as many variables as you want

$$\begin{aligned} p(x_1, ..., x_n) &= p(x_1|x_2, ..., x_n)p(x_2, ..., x_n) \\ &= ... = p(x_1|x_2, ..., x_n)...p(x_{n-1}|x_n)p(x_n) \end{aligned}$$

3.7 Bayes' Theorem

This is a central theorem that we will use repeatedly in this course, and is an extension of the product rule.

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Since we are conditioning on y, y is held constant, and that means $p(y)$ is just a normalization constant. As a result, we often write the above property as

$$p(x|y) \propto p(y|x)p(x)$$

To see this concretely in terms of machine learning: say we observe data D, and we are interested in parameters $\mathbf{w}$. We can write the *posterior distribution* of the parameters given data by by using Bayes' theorem.

$$\underbrace{p(\mathbf{w}|D)}_{\text{posterior}} = \frac{\overbrace{p(D|\mathbf{w})}^{\text{likelihood}}\overbrace{p(\mathbf{w})}^{\text{prior}}}{\underbrace{p(D)}_{\text{evidence}}}$$

Related to this, a maximum a posteriori (MAP) estimate for the parameter $\mathbf{w}$ is the value

$$\arg\max_{\mathbf{w}} p(\mathbf{w}|D)$$

A maximum likelihood estimate (MLE) is the value

$$\arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

We will cover these in more depth in the course.