Beyond Worst Case Analysis **Fall 2018**

Lecture 1: Linear Algebra and Probability Review

Scribe: Antares Chen, Haaris Khan, Hermish Mehta, Jeff Xu 9/6/2018

We will regularly fall back to a solid understanding of linear algebra and discrete probability throughout our exploration of algorithm analysis regimes that extend beyond worst-case analysis. For this reason, we present a comprehensive review of the basic tools that we will use. This document is split into two sections covering (1) discrete probability and (2) linear algebra.

1.1 Discrete Probability

Moving beyond traditional algorithm analysis, we are often interested in understanding algorithms on typical instances, rather than worse-case pathological examples. Describing naturally occurring inputs precisely, however, requires formalizing ideas about likelihood with respect to large classes of possible instances. *Probability theory* is a branch of mathematics which provides a language to do just this, allowing us to rigorously reason about uncertainty.

1.1.1 Probability Spaces

A *probability space* $(\Omega, \mathcal{F}, \Pr)$ models an underlying uncertain process that produces exactly one *outcome*; this outcome may belong to several *events*, sets of outcomes. The model is characterized by a *sample space* (Ω), the set of all possible outcomes, and a *probability law* $(\Pr)$ which assigns the events in $\mathcal{F}$ *probabilities* according to the following axioms.

1. For all events $A \in \mathcal{F}$, $\Pr(A) \geq 0$.
2. The sample space has unit probability, $\Pr(\Omega) = 1$.
3. Given any sequence of disjoint events (A_i),

$$\Pr(A_1 \cup A_2 \cup \dots) = \Pr(A_1) + \Pr(A_2) + \dots. \tag{1.1}$$

From these basic axioms, we can derive the following properties about probability laws.

Claim 1.1. *Consider any events $A, B \subseteq \Omega$ over the same probability space $(\Omega, \mathcal{F}, \Pr)$. The following properties must hold.*

1. The empty set has 0 probability, $P(\phi) = 0$.
2. If $A \subseteq B$, then $\Pr(A) \leq \Pr(B)$.
3. Probabilities are between $0 \leq \Pr(A) \leq 1$.
4. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Proof.

1. Since ϕ is disjoint from the sample space, Ω, we immediately have $P(\phi) = 1 - P(\Omega) = 0$.

2. Since $A \subseteq B$, we can express B as the union of A and the difference between the two.

$$\begin{aligned}\Pr(B) &= \Pr(A) + \Pr(B - A) \\ &\geq \Pr(A)\end{aligned}$$

3. Notice any event A is a subset of the sample space, so from the above result, we have $\Pr(A) \leq \Pr(\Omega) = 1$. Since probabilities must be non-negative, we get the lower bound from the axioms.

4. We express $A \cup B$ as a union of disjoint sets to apply the probability axioms.

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B - A) \\ &= \Pr(A) + \Pr(B - A) + \Pr(A \cap B) - \Pr(A \cap B) \\ &= \Pr(A) + \Pr(B) - \Pr(A \cap B)\end{aligned}$$

□

When the sample space is *discrete*, finite or countably infinite, the probability law is defined by the probabilities of individual outcomes since for any subset of $A \subseteq \Omega$,

$$\Pr(A) = \sum_{\omega \in A} \Pr(\omega).$$

Here, we use $P(\omega)$ as shorthand to denote the probability of the event $\{w\}$. This motivates referring to the probabilities of singleton events as the probabilities of the outcomes themselves.

1.1.2 Counting

Consider some event $A \subseteq \Omega$ belonging to some finite sample space with n equally likely outcomes. By definition, each outcome in Ω must have probability $1/n$, therefore the probability of A is

$$\begin{aligned}\Pr(A) &= \sum_{\omega \in A} \frac{1}{n} \\ &= \frac{|A|}{|\Omega|}.\end{aligned}$$

Computing the probability of A over a *uniform* distribution, then, is intrinsically linked to counting the elements of A and Ω. Since we are primarily interested in combinatorial objects, we present a few important counting principles through a series of examples.

Example 1.2 (Permutations)**.** Consider some set of n elements. The number of ways to arrange its elements in a sequence is equal to the number of choices of the first element times number of the choices of the second and so on. This will simply be the product of all numbers from 1 to n inclusive.

$$\prod_{k=1}^{n} k = n!$$

This product is denoted $n!$, and is called the *factorial* of n; by convention, we say we define 0! as 1. Each of these sequences, which correspond to arrangements of elements in the set, is called a *permutation.*

Example 1.3 (k-Permutations)**.** Given the same set of n elements, if we are only interested in permutations of exactly $k \leq n$ elements, the computation is slightly different. There are still $(n - i + 1)$ choices for the ith element, but only k choices are made in total. Therefore, the total number of *k-permutations* is

$$\prod_{i=1}^{k}(n - i + 1) = \frac{n!}{(n-k)!}.$$

Example 1.4 (Combinations)**.** Now consider counting the number of different k-element subsets possible from this set of n elements. Notice we can easily count the total number of k-permutations, which is

$$\frac{n!}{(n-k)!}$$

However, since each subset of k elements has $k!$ permutations, each subset will correspond to exactly $k!$ unique k-permutations. Therefore, the total number of these subsets is

$$\frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

The above expression is referred to as a *binomial coefficient*, denoted with parentheses; with integral $n \geq k \geq 0$, it refers to the number of ways to choose k elements from set of n distinct items.

1.1.3 Conditional Probability

While reasoning about uncertainty is powerful, the uncertainty we are concerned about is usually relative to the partial information we may already have. For example, when working with random graphs, we could be interested in path lengths given the graph generated is connected. In probability theory this idea is expressed through through *conditional probability.* Formally, we define the probability of A *conditioned on* B as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \text{ given } \Pr(B) > 0. \tag{1.2}$$

Notice that any fixed event B with non-zero probability induces a valid probability law $\Pr(\cdot|B)$; we can check this by verifying the axioms. In particular, the sample space still has probability 1 and the probabilities of any two disjoint events A_1 and A_2 still sum.

$$\begin{aligned}
\Pr(\Omega|B) &= \frac{\Pr(\Omega \cap B)}{\Pr(B)} \\
&= \frac{\Pr(B)}{\Pr(B)} = 1 \\
\Pr(A_1 \cup A_2|B) &= \frac{\Pr((A_1 \cup A_2) \cap B))}{\Pr(B)} \\
&= \frac{\Pr((A_1 \cap B) \cup (A_2 \cap B))}{\Pr(B)} \\
&= \frac{\Pr(A_1 \cap B)}{\Pr(B)} + \frac{\Pr(A_2 \cap B)}{\Pr(B)} \\
&= \Pr(A_1|B) + \Pr(A_2|B)
\end{aligned}$$

Theorem 1.5 (Bayes' Rule)**.** *Given events A and B over the same probability space $(\Omega, \mathcal{F}, \Pr)$, the following relationship holds, given $\Pr(A) > 0$ and $\Pr(B) > 0$.*

$$\Pr(A|B) = \Pr(A)\frac{\Pr(B|A)}{\Pr(B)} \tag{1.3}$$

Proof.

$$\begin{aligned}
\Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\
&= \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}
\end{aligned}$$

□

Bayes' rule allows us to update the probability of A, given an observation of B. Here, $\Pr(A)$ is called the prior, $\Pr(A|B)$ the posterior and $\Pr(B|A)$ the likelihood. The essence of Bayes' rule is simply that the posterior probability is simply proportional to the prior and likelihood, or

$$\Pr(A|B) \propto \Pr(A)\Pr(B|A).$$

Conditional probability represents an attempt to reason given partial information. However, with two consecutive coin tosses, knowledge about the result of former carries no information about the latter. To that end, we generalize this idea by saying some event A is *independent* from B if

$$\Pr(A \cap B) = \Pr(A)\Pr(B). \tag{1.4}$$

Notice of $\Pr(B)$ is strictly positive, then this equation tells us $\Pr(A|B) = \Pr(A)$. Symmetrically, when $\Pr(A) > 0$, this means $\Pr(B|A) = \Pr(B)$. Hence conditioning on one event does not change the probability of the other

1.1.4 Random Variables

We are often less interested in specific outcomes of an experiment than some value associated with each outcome. For example, when analyzing random networks we may just be interested in the number of connected components rather than individual graphs.

This mapping between outcomes and values is codified through *random variables*, functions from the sample space to the real numbers. Its *support* is the set of outcomes with strictly positive probability. Introducing a random variable X induces events denoted as $\{X = x\}$, referring the set of outcomes where X takes on the value x.

$$\begin{aligned} \{X = x\} &= \{\omega \in \Omega : X(\omega) = x\} \\ &\qquad\qquad \text{where } X : \Omega \to \mathbb{R} \end{aligned}$$

The *probability mass function* (PMF) of a discrete random variable is a function which gives the probability of such events for each possible value x. This defines the *distribution* of X, the set of all values and respective probabilities. More precisely,

$$\begin{aligned} p_X(x) &= \Pr(X = x) \\ &= \Pr(\{\omega \in \Omega : X(\omega) = x\}). \end{aligned}$$

Functions of random variables, $Y = g(X)$, define new random variables so that for each outcome where $X(\omega) = x$, $Y(\omega) = g(x)$. This definition allows us compute the probability mass function of the image.

$$\begin{aligned} p_Y(y) &= \sum_{\substack{\omega\in\Omega: \\ Y=y}} \Pr(\omega) \\ &= \sum_{\substack{x: \\ g(x)=y}} \sum_{\substack{\omega\in\Omega: \\ X=x}} \Pr(\omega) \\ &= \sum_{\substack{x: \\ g(x)=y}} p_X(x) \end{aligned}$$

The *expectation* of a random variable represents a weighted mean of its possible values. We will immediately prove that expectation is linear, a powerful result which radically simplifies any expression involving linear combinations of random variables.

$$\mathbb{E}[X] = \sum_x x p_X(x) \tag{1.5}$$

Claim 1.6 (Linearity of Expectation)**.** *For any random variables $X_1, \ldots X_n$ defined over the same sample space with corresponding scalars $c_1, \ldots c_n$.*

$$\mathbb{E}\left[\sum_{i=1}^n c_i X_i\right] = \sum_{i=1}^n c_i \mathbb{E}[X_i] \tag{1.6}$$

Proof.

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^n c_i X_i\right] &= \sum_{\omega\in\Omega} \left(\sum_{i=1}^n c_i X_i(\omega)\right) P(\omega) \\ &= \sum_{i=1}^n c_i \left(\sum_{\omega\in\Omega} X_i(\omega) P(\omega)\right) \\ &= \sum_{i=1}^n c_i \mathbb{E}[X_i] \end{aligned}$$

□

The *variance* measures the expected squared error with respect to the mean, establishing some metric about the spread of the distribution. More precisely,

$$\begin{aligned}\mathbf{Var}(X) &= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] && (1.7)\\ &= \sum_x (x-\mathbb{E}[X])^2 p_X(x). && (1.8)\end{aligned}$$

The variance of a random variable can readily be expressed in terms of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$, the first and second *moments* of X respectively. More generally, the *nth moment* of a random variable is defined as the expectation of X^n.

Claim 1.7 (Moments Expression of Variance)**.** *The variance of a random variable X is the difference between its second moment and the square of its first moment.*

$$\mathbf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{1.9}$$

Proof.

$$\begin{aligned}\mathbf{Var}(X) &= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]\\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]\\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X\mathbb{E}[X]] + \mathbb{E}[X]^2\\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2\\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

□

1.1.5 Markov's Inequality

Often times, having knowledge of the expectation and variance of some non-negative random variable X allows us to derive a bound for the probability of X exceeding some value. To elaborate on this point, we introduce a few inequalities, starting with **Markov's Inequality**

Theorem 1.8. ***Markov's Inequality*** *For some non-negative Random Variable X, whose expectation* $\mathbb{E}[X]$ *that is known, and any $\alpha > 0$, we have*

$$\Pr(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_a a \cdot \Pr(X = a) && (1.10)\\
&\geq \sum_{a \geq \alpha} a \cdot \Pr(X = a) && (1.11)\\
&\geq \alpha \cdot \sum_{a \geq \alpha} \Pr(X = a) && (1.12)\\
&= \alpha \cdot \Pr(X \geq \alpha) && (1.13)\\
\Rightarrow \frac{\mathbb{E}[X]}{\alpha} &\geq \Pr(X \geq \alpha) && (1.14)
\end{aligned}$$

□

Begin by taking the definition of expectation, and lower bounding that value by restricting the values X takes on to be $\geq \alpha$. Then, further lower bound this by considering each $\Pr(X = a)$ to be multiplied by α, and rewrite the line on (1.3) to yield $\Pr(X \geq \alpha)$

(FYI: An analog to this for continuous distributions also works, where we replace sums with integrals. Same underlying logic applies)

While useful as an introduction, Markov's Inequality generally provides a very loose bound. Another layer to consider that provides a more useful result arises when considering the variance of the random variable.

Theorem 1.9. ***Chebychev's Inequality*** *For any $t > 0$, we have*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{Var(X)}{t^2}$$

Proof. Recall $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, it's now a one line proof from Markov. □

To illustrate the effects, we'll consider the following: Let X be a non-negative random variable where $\mathbb{E}[X] = 2$ and $Var(X) = \frac{1}{4}$. We can find the probability that X takes on a value over 4 using both inequalities:

Markov's Inequality

$$\begin{aligned}
\Pr(X \geq 4) &\leq \frac{\mathbb{E}[X]}{4}\\
&= \frac{1}{2}
\end{aligned}$$

Chebyshev's Inequality

$$\begin{aligned}
\Pr(|X - 2| \geq 2) &\leq \frac{\frac{1}{4}}{2^2}\\
&= \frac{1}{16}
\end{aligned}$$

1.2 Linear Algebra

In this section, we review linear algebra facts that will be fundamental to the analysis we perform in this reading group. We begin with basic definitions discussing hermitian and symmetric matrices as well as inner products and norms. We then recall definitions and properties of eigenvectors and eigenvalues. The next section contains a critical fact about real symmetric matrices: the spectral theorem, which we then use to discuss positive semi-definite matrices. Finally, this linear algebra review closes with a derivation of the variational characterization of eigenvalues.

1.2.1 Basic Definitions

We begin by reviewing some basic definitions. Let A be a real $m \times n$ matrix (often denoted as $A \in \mathbb{R}^{m\times n}$), then the *transpose* $A^\top$ is given by $(A^\top)_{ij} = A_{ji}$. When $A^\top = A$, then we call A a *symmetric* matrix. We can note an algebraic identity that will come handy throughout our studies:

Claim 1.10. *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$. Then $(AB)^\top = B^\top A^\top$.*

Proof. For any $i, j = 1, \dots, n$, we have the following

$$\left((AB)^\top\right)_{ij} = (AB)_{ji} = \sum_{k=1}^{n} A_{jk}B_{ki} = \sum_{k=1}^{n} B_{ki}A_{jk} = \sum_{k=1}^{n} (B^\top)_{ik}(A^\top)_{kj} = (B^\top A^\top)_{ij} \qquad \square$$

Next, let us consider vectors in n-dimensional space. Recall that we may define the *standard inner product* between two vectors $x, y \in \mathbb{R}^n$ as the following.

$$\langle x, y\rangle = x^\top y = \sum_{i=1}^{n} x_i y_i$$

If $\langle x, y\rangle = 0$, then we say that x and y are *orthogonal.* Using the previous claim, we can show another identity that will be useful for us throughout our readings.

Claim 1.11. *For any square matrix $A \in \mathbb{R}^{n\times n}$ and vectors $x, y \in \mathbb{R}^n$, it holds that $\langle Ax, y\rangle = \langle x, A^\top y\rangle$.*

Proof. For any two vectors $x, y \in \mathbb{R}^n$

$$\langle Ax, y\rangle = (Ax)^\top y$$

If we consider A, x as matrices, then claim 1.10 gives $(Ax)^\top y = x^\top A^\top y = \langle x, A^\top y\rangle$ as required. $\square$

The standard inner product is also related to Euclidean distance in that $\langle x, x\rangle = x^\top x = \sum_{i=1}^{n} x_i^2$. Specifically, $\langle x, x\rangle$ is the Euclidean distance from the origin to x squared. We can generalize the notion of distance on $\mathbb{R}^n$ through a *norm* function. A norm is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ admitting the following properties:

(1) $\|x\| \geq 0$ for any $x \in \mathbb{R}^n$ and $\|x\| = 0$ if and only if $x = 0$.

(2) $\|x+y\| \leq \|x\| + \|y\|$ for any $x, y \in \mathbb{R}^n$ (this is called the triangle inequality).

(3) $\|\alpha x\| = |\alpha|\|x\|$ for any $\alpha \in \mathbb{R}, x \in \mathbb{R}^n$.

Indeed, one can verify that the Euclidean distance of a vector x satisfies all the above properties and is thus a norm. We often call this the ℓ_2-norm and denote the ℓ_2-norm of vector x by $\|x\|_2$. Finally, we say that if $\|x\|_2 = 1$, then x is a *unit vector* and if $\langle x, y\rangle = 0$ where x, y are unit vectors, then we say x, y are *orhthonormal* to each other.

1.2.2 Eigenvectors and Eigenvalues

It is interesting to simply consider square matrices as this allows us to study eigenvectors and eigenvalues. Given $A \in \mathbb{R}^{n\times n}$, a non-zero vector $v \in \mathbb{R}^n$ is an *eigenvector* of A if for some $\lambda \in \mathbb{R}$, we have $Av = \lambda v$. We call λ the *eigenvalue* corresponding to the eigenvector v. We also note that the set of eigenvalues for a particular A is also called the *spectrum* of A.

One can intuitively think of v as a vector for which A's action is to scale it along the direction of v. To that end, λ corresponding to v is exactly the factor by which it is scaled. To actually compute the eigenvalues of a matrix A, one generally looks at A's characteristic polynomial. The *characteristic polynomial* of $A \in \mathbb{R}^{n\times n}$ is given by

$$p(t) = \det(A - tI_n)$$

where t is an indeterminate and I_n is the $n \times n$ identity matrix. One might recall then that λ is an eigenvalue of A if and only if $p(\lambda) = 0$. Not only does it suffice to determine the roots of $p(t)$ to find the eigenvalues of A, but it also means that A can have at most n complex eigenvalues. This follows since the fundamental theorem of algebra states that any degree n polynomial, such as $p(t)$, may have at most n complex roots.

1.2.3 Symmetric Real Matrices and the Spectral Theorem

Most of the matrices we deal with in our readings will specifically be symmetric and have real elements. For reasons that we will encounter later on, this is quite spectacular as it allows us to use a very strong property regarding the matrix's spectrum. In particular, if A is a real symmetric matrix, then A is fully characterized by its spectrum in the sense that it can be decomposed into a sum of rank-one matrices scaled by its eigenvalues. This is known as the *spectral theorem* now stated below.

Theorem 1.12. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix, then the following statements hold.*

1. *A has not necessarily distinct real eigenvalues $\lambda_1, \ldots, \lambda_n$ corresponding to real eigenvectors $v_1, \ldots, v_n$*
2. *For any $i, j = 1, \ldots, n$ we have that $\langle v_i, v_j\rangle = 0$ if $i \neq j$ and 1 otherwise. In particular, the set $\{v_1, \ldots, v_n\}$ forms an orthonormal (eigen)basis of $\mathbb{R}^n$.*
3. *A admits a spectral decomposition $A = \sum_{i=1}^n \lambda_i v_i v_i^\top$.*

To prove this theorem, we will need two lemmas. The first allows us to only consider real eigenvalues and eigenvectors.

Lemma 1.13. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric. If λ is an eigenvalue of A then $\lambda \in \mathbb{R}$ and it corresponds to at least one real eigenvector v.*

Proof. Briefly recall some facts on complex numbers. If $x = a + bi \in \mathbb{C}$, then the *complex conjugate* of x is given by $\overline{x} = a - bi$. The complex conjugate of a complex vector or matrix is simply defined element wise. Finally, note that $\langle \overline{v}, v \rangle \geq 0$ with equality holding if and only if $v \neq 0$. This is as

$$\begin{pmatrix} a_1 - b_1 i & \cdots & a_n - b_n i \end{pmatrix} \begin{pmatrix} a_1 + b_1 i \\ \vdots \\ a_n + b_n i \end{pmatrix} = (a_1^2 + b_1^2) + \ldots + (a_n^2 + b_n^2)$$

Let us now demonstrate that A has only real eigenvalues. Consider any λ with eigenvector v and observe that by taking the complex conjugate of $Av = \lambda v$, we derive the following.

$$\overline{Av} = \overline{\lambda v} \qquad \Longleftrightarrow \qquad \overline{A}\overline{v} = \overline{\lambda}\overline{v} \qquad \Longleftrightarrow \qquad A\overline{v} = \overline{\lambda}\overline{v}$$

In particular, we have $\overline{A} = A$ since A is a real matrix. Now, writing $\overline{v}^\top Av$ in two ways using the fact that $A = A^\top$ derives for us:

$$\begin{aligned} \overline{v}^\top Av &= \overline{v}^\top (Av) = \overline{v}(\lambda v) = \lambda \langle \overline{v}, v \rangle \\ \overline{v}^\top Av &= (A\overline{v})^\top v = (\overline{\lambda}\overline{v})^\top v = \overline{\lambda} \langle \overline{v}, v \rangle \end{aligned}$$

Since $v \neq 0$, it cannot be that $\langle \overline{v}, v \rangle = 0$ hence it must be that $\lambda = \overline{\lambda}$ which is true if and only if $\lambda \in \mathbb{R}$. Finally, to demonstrate that λ always corresponds to at least one real eigenvector v, suppose that $v \in \mathbb{C}^n$. If so, then we can write $v = x + iy$ for two vectors $x, y \in \mathbb{R}^n$. With $Av = \lambda v$, we have that

$$Av = \lambda v \qquad \Longleftrightarrow \qquad A(x + iy) = \lambda(x + iy) \qquad \Longleftrightarrow \qquad Ax + iAy = \lambda x + i\lambda y$$

Since $v \neq 0$, it cannot be that $x = y = 0$ and so either $Ax = \lambda x$ or $Ay = \lambda y$. Consequently, $\lambda \in \mathbb{R}$ implies either x or y are real eigenvectors of A corresponding to λ as required. □

The second lemma states that eigenvectors corresponding to different eigenvalues are orthogonal.

Lemma 1.14. *Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric with two distinct eigenvalues $\lambda_1 \neq \lambda_2$ corresponding to eigenvectors v_1, v_2 respectively. Then v_1 and v_2 are orthogonal.*

Proof. Recall that because $A = A^\top$, claim 1.11 implies $\langle Av_1, v_2 \rangle = \langle v_1, A^\top v_2 \rangle = \langle v_1, Av_2 \rangle$. This means

$$\langle Av_1, v_2 \rangle - \langle v_1, Av_2 \rangle = 0 \qquad \Longleftrightarrow \qquad \langle \lambda_1 v_1, v_2 \rangle - \langle v_1, \lambda_2 v_2 \rangle = 0 \qquad \Longleftrightarrow \qquad (\lambda_1 - \lambda_2)\langle v_1, v_2 \rangle = 0$$

Because $\lambda_1 \neq \lambda_2$ by assumption, we have that $\langle v_1, v_2 \rangle = 0$ as required. □

We are now ready to prove the spectral theorem.

Proof. (theorem 1.12) We will begin by proving items (1) and (2) via induction on n. The case $n = 1$ holds trivially. For $\big(a\big)$, take the eigenvalue to be a and the eigenvector to be $\big(1\big)$. Let us now assume that statements (1) and (2) hold for all symmetric matrices $A \in \mathbb{R}^{(n-1)\times(n-1)}$ and demonstrate it holds for dimension n. To apply our inductive hypothesis, we will need to restrict A into an $(n-1) \times (n-1)$ matrix.

Fix λ_1 as a real eigenvalue of A corresponding to eigenvector v_1 and consider the dimension $n-1$ vector space V containing all vectors orthogonal to v_1. Now let V have an orthonormal basis $\{b_1, \ldots, b_{n-1}\}$[1] and assign $P = \big(b_1 \cdots b_{n-1}\big)$. Observe that $P^\top P = PP^\top = I_{n-1}$ since $P^\top P = \langle b_i, b_j\rangle$. Because b_i, b_j are orthonormal, we have $\langle b_i, b_j\rangle = 0$ if $i \neq j$ and 1 otherwise.

Now let $A' = P^\top AP$ and observe that $A' \in \mathbb{R}^{(n-1)\times(n-1)}$. Furthermore, A' is symmetric as

$$(P^\top AP)^\top = P^\top (P^\top A)^\top = P^\top AP$$

We can now apply the inductive hypothesis. Suppose A' has eigenvalues $\lambda_2, \ldots, \lambda_n$ corresponding to orthonormal eigenvectors $v'_2, \ldots, v'_n$ such that $\langle v'_2, \ldots, v'_n\rangle$ forms an orthonormal set. Using this, we first show that statement (1) holds for A. For any $i = 2, \ldots, n$, we have that

$$A'v_i = \lambda_i v_i \qquad \Longleftrightarrow \qquad P^\top APv_i = \lambda_i v_i \qquad \Longleftrightarrow \qquad PP^\top APv_i = P(\lambda_i v_i) \qquad \Longleftrightarrow \qquad APv'_i = \lambda_i Pv'_i$$

Define $v_i = Pv'_i$ and note that λ_i must now be an eigenvalue of A with eigenvector v_i. Consequently, $\lambda_1, \ldots, \lambda_n$ corresponding to $v_1, \ldots, v_n$ form n (not necessarily distinct) eigenvalues of A. By lemma 1.13, these must be real thus satisfying (1).

To demonstrate (2), note that by construction v_1 is orthogonal to all $v_2, \ldots, v_n$ since we chose V to be the space of vectors orthogonal to v_1. It suffices to demonstrate that $\{v_2, \ldots, v_n\}$ forms an orthonormal set. As $v'_2, \ldots, v'_n$ are orthogonal, we have for any i, j:

$$\langle v_i, v_j\rangle = \langle Pv'_i, Pv'_j\rangle = \langle v'_i, P^\top Pv'_j\rangle = \langle v'_i, v'_j\rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

We need not worry if v_1 is a unitary vector since we can re-complete this argument with scaled vector $\frac{v_1}{\|v_1\|_2}$.

To finish the proof, we demonstrate claim (3). Let us consider any vector $x \in \mathbb{R}^n$. Since $\{v_1, \ldots, v_n\}$ forms a basis under statement (2), we can write x under this basis as $x = \sum_{i=1}^n c_i v_i$. By orthonormality

$$\langle x, v_i\rangle = \left\langle \sum_{i'=1}^n c_{i'} v_{i'}, v_i \right\rangle = c_i \langle v_i, v_i\rangle = c_i$$

Therefore, $x = \sum_{i=1}^n \langle x, v_i\rangle v_i$ which can further be rewritten as

$$\sum_{i=1}^n \langle x, v_i\rangle v_i = \sum_{i=1}^n v_i v_i^\top x = \left(\sum_{i=1}^n v_i v_i^\top\right) x$$

[1] Such a set always exists as one can always apply the Gram-Schmidt procedure.

Multiplying both sides by A derives and recalling that $v_1, \dots, v_n$ are eigenvectors

$$Ax = A\Big(\sum_{i=1}^{n} v_i v_i^\top\Big)x = \Big(\sum_{i=1}^{n}(Av_i)v_i^\top\Big)x = \Big(\sum_{i=1}^{n}\lambda_i v_i v_i^\top\Big)x$$

The left and right hand-side agree on every $x \in \mathbb{R}^n$ hence $A = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$ as required. □

We complete this section with a few remarks on theorem 1.12. Our choice of B in the proof is not entirely arbitrary because B is in fact a bijective map from our orthogonal vector space V to $\mathbb{R}^{n-1}$ (i.e. we constructed B to be the change of base matrix into $\mathbb{R}^{n-1}$). Because $B^\top B = I_n$ (also called a *unitary* matrix), $B^\top = B^{-1}$. One may recall that if B is a bijective map, B^{-1} is then the inverse map implying that we can consider B^{-1} as the change of base matrix from $\mathbb{R}^{n-1}$ back into V.

Therefore, the inductive case is actually performing a change of basis in order to create a matrix $A' \in \mathbb{R}^{(n-1)\times(n-1)}$. This proof can actually be made much cleaner if we extend this intuition and consider A as a linear map between vector spaces. However, this requires an even more extensive discussion into topics tangential to our readings. If interested, we refer the reader to Friedberg, Insel, and Spence's *Linear Algebra*.

Finally, the spectral theorem gives us a very nice intuitive way of thinking about symmetric real matrices A. In particular, we can think of A as a matrix whose action simply scales any vector along the directions of the orthonormal eigenbasis since A can always be decomposed into a sum of rank-one matrices scaled by its eigenvalues.

1.2.4 Rayleigh Quotient and Courant-Fischer Theorem

In terms of designing algorithms that rely on the spectrum of a real symmetric matrix, the algebraic definition $Av = \lambda v$ may be too rigid. Certainly when discussing spectral graph theory and spectral methods, we instead often use the variational characterization of eigenvalues. To begin, we define the *Rayleigh quotient* of a real symmetrix matrix given a vector $x \neq 0$ as the following.

$$\frac{x^\top Ax}{x^\top x}$$

The *variational characterization* of eigenvalues then defines the eigenvalues of A as an optimization problem over its Rayleigh quotient. For example, the following lemma is a variational characterization of the largest eigenvalue of matrix A.

Theorem 1.15. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric with (not necessarily distinct) eigenvalues $\lambda_1 \geq \dots \geq \lambda_n$ corresponding to eigenvectors $v_1, \dots, v_n$. Then*

$$\lambda_1 = \max_{x\in\mathbb{R}^n : x\neq 0} \frac{x^\top Ax}{x^\top x} = \max_{x\in\mathbb{R}^n : \|x\|_2^2 = 1} x^\top Ax$$

Proof. By theorem 1.12, A admits the orthonormal basis $\{v_1, \dots, v_n\}$. For any $x \in \mathbb{R}^n$, let us write it under this basis as $x = \sum_{i=1}^{n} c_i v_i$. The expression $x^\top Ax$ is then given by.

$$x^\top Ax = \Big(\sum_{i=1}^{n} c_i v_i\Big)^\top A\Big(\sum_{i=1}^{n} c_i v_i\Big) = \Big(\sum_{i=1}^{n} c_i v_i\Big)^\top\Big(\sum_{i=1}^{n} c_i Av_i\Big) = \sum_{i=1}^{n} c_i^2 \lambda_i$$

With a similar calculation, we can deduce that $x^\top x = \sum_{i=1}^n c_i^2$. Recalling that λ_1 is the largest eigenvalue

$$\frac{x^\top Ax}{x^\top x} = \frac{\sum_{i=1}^n c_i^2 \lambda_i}{\sum_{i=1}^n c_i^2} \leq \frac{\lambda_1 \sum_{i=1}^n c_i^2}{\sum_{i=1}^n c_i^2} = \lambda_1$$

Thus the Rayleigh quotient of A is bounded above by λ_1. Since the right-hand side of the equality is a maximization problem, we can deduce equality by demonstrating a vector that evaluates the Rayleigh quotient to λ_1. Indeed with $x = v_1$, we have

$$\frac{v_1^\top A v_1}{v_1^\top v_1} = \lambda_1 v_1^\top v_1 = \lambda_1$$

as required. □

We can interpret this theorem geometrically as saying finding the greatest eigenvalue of A is equivalent to finding the largest value of $x^\top Ax$ (also called the quadratic form of A) over the unit hypersphere (i.e. x constrainted by $\|x\|_2^2 = 1$). We can extend the above theorem to characterize other eigenvalues in A.

Theorem 1.16. *Suppose $A \in \mathbb{R}^n$ is symmetric with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ corresponding to eigenvectors $v_1, \ldots, v_n$. Denote V_k as the space of all vectors that are orthogonal to $v_1, \ldots, v_{k-1}$, then the following holds.*

$$\lambda_k = \max_{x \in V_k} \frac{x^\top Ax}{x^\top x}$$

We omit the proof as it is very similar to the argument in theorem 1.15. It is also true that you can represent the minimum eigenvalue as a minimization problem. We can also characterize finding eigenvalues as a minimization problem.

Theorem 1.17. *Suppose $A \in \mathbb{R}^n$ is symmetric with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ corresponding to eigenvectors $v_1, \ldots, v_n$. Then*

$$\lambda_n = \min_{x \in \mathbb{R}^n : x \neq 0} \frac{x^\top Ax}{x^\top x} = \min_{x \in \mathbb{R}^n : \|x\|_2^2 = 1} x^\top Ax$$

Additionally, denote V_k as the space of all vectors that are orthogonal to $v_k + 1, \ldots, v_n$. It follows that

$$\lambda_k = \min_{x \in V_k} \frac{x^\top Ax}{x^\top x}$$

The proof is quite similar to theorem 1.15 as well, simply replace the upper-bound on $x^\top Ax$ with a lower-bound using v_n. The above three theorems have very similar forms. Indeed, these statements stem as a corollary from the most generalized version of the variational characterization theorems known as the Courant-Fisher theorem. This states that

Theorem 1.18. (Courant-Fischer) *Let $A \in \mathbb{R}^n$ be symmetric with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. Then*

$$\lambda_k = \max_{V \subseteq R^n : \dim(V) = k} \min_{x \in V} \frac{x^\top Ax}{x^\top x} = \min_{V \subseteq R^n : \dim(V) = n-k+1} \max_{x \in V} \frac{x^\top Ax}{x^\top x}$$

It is likely that we will not see the Courant-Fisher theorem many times. However, we will certainly encounter the first few theorems in this section quite frequently.