POSITIVE DEFINITE MATRICES AND KERNELS

JOEL FRIEDMAN

Contents

Date: Friday 10th October, 2025, at 15:33.
Research supported in part by an NSERC grant.

Disclaimer: The material may sketchy and/or contain errors, which I will elaborate upon and/or correct in class. For those not in CPSC 563F: use this material at your own risk...
This version is A WORK IN PROGRESS: I'll post to the course website when I revise this article.

1. The Main Goals of 536F

CPSC 536F is a new course, focusing on positive (semi)definite matrices and kernels, and some 8-12 applications that arise in computer science, especially machine learning.

The main goal is to build up some intuition about these matrices and kernels, their eigenvectors and eigenfuctions, and how they are used.

A secondary goal is to explain the origin and theory of kernels, to make them easier to design (and seem less "mysterious").

Along the way we may also describe some theoretical aspects of neural networks.

This course is theoretical, meant to compliment more applied courses.

2. Basic Notation

Let us fix some basic terminology and notation that we use throughout this article.

2.1. **Sets.** We use $\mathbb{R}, \mathbb{C}, \mathbb{Z}$ to denote, respectively, the real numbers, the complex numbers, and the integers. We use $\mathbb{N} = \{1, 2, \ldots\}$ to denote the positive integers, and $\mathbb{Z}_{\geq 0} = \{0, 1, \ldots\}$ to denote the set of non-negative integers. For $n \in \mathbb{Z}_{\geq 0}$ we use $[n]$ to denote $\{1, 2, \ldots, n\}$, with $[0] = \emptyset$, the empty set.

If A, B are sets, then $|A|$ denotes the cardinality (size) of A, and we define the set difference "A minus B" as

$$A \setminus B = \{a \in A \mid a \notin B\}.$$

2.2. **Vectors and Matrices.** If A is a set, we use $\mathbb{R}^A$ to denote the set of maps $A \to \mathbb{R}$, and $\mathbb{R}^n$ to denote $\mathbb{R}^{[n]} = \mathbb{R}^{\{1,\ldots,n\}}$ (which is therefore the set of maps $[n] \to \mathbb{R}$). We use the notation $\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$, where bold letters (e.g., $\mathbf{u}$) reserved for *vectors*, i.e., elements of $\mathbb{R}^n$, and non-bold letters (e.g., $u_1, \ldots, u_n$) reserved for the components of $\mathbf{u}$. Similarly, $\mathbf{1} \in \mathbb{R}^n$ to refers to $(1, \ldots, 1) \in \mathbb{R}^n$, and similarly for $\mathbf{0}$; similarly for $\mathbf{1}, \mathbf{0} \in \mathbb{R}^A$ for a set, A. Similarly for $\mathbb{R}$ replaced with $\mathbb{N}, \mathbb{Z}, \mathbb{C}$ or any set.

We will often be interested in maps $A \to \mathbb{N}$, i.e., elements of $\mathbb{N}^A$, where A has no implied order; we still use a bold letter for the vector, e.g., $\mathbf{m} \in \mathbb{N}^A$ or $\mathbf{m}\colon A \to \mathbb{N}$, and use non-bold letters for the components of $\mathbf{m}$, e.g., writing $\mathbf{m} = \{m(a)\}_{a \in A}$. Similarly for $\mathbb{N}$ replaced with $\mathbb{R}, \mathbb{Z}, \mathbb{C}$

We will make use of the inner product or "dot product" of $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$,

$$\mathbf{u} \cdot \mathbf{w} = u_1 w_1 + \cdots + u_n w_n,$$

and similarly for $\mathbf{u}, \mathbf{w} \in \mathbb{C}^n$,

$$\mathbf{u} \cdot \mathbf{w} = \overline{u_1} w_1 + \cdots + \overline{u_n} w_n, \tag{1}$$

where $\overline{u}_i$ denotes the complex conjugate of u_i. [Elsewhere in the literature one uses the complex conjugate, i.e., the sum of $u_i \overline{w_i}$, and one could equally well use this version of the complex dot product throughout this article.]

We use $\mathbb{C}^{m\times n}$ to denote the set of $m\times n$ matrices, M, with entries in $\mathbb{C}$ (similarly for $\mathbb{N}$, $\mathbb{R}$, etc.); for $i\in[n]$ and $j\in[m]$, we use $M_{i,j}$ or M_{ij} or $M(i,j)$ to denote the (i,j)-th entry of M (i.e., the entry in the i-th row and j-th column). Similarly, for $\mathbb{C}^{A\times B}$ when A,B are sets, which do not necessarily come with an implied order.

We use M^{T} to denote the transpose of M. If $\mathbf{u}\in\mathbb{R}^n$, we view $\mathbf{u}$ as an $n\times 1$ matrix; hence it makes sense to write $M\mathbf{u}$ for an $m\times n$ matrix M, and similarly for $\mathbf{w}^{\mathrm{T}}M$ when $\mathbf{w}\in\mathbb{R}^m$. We use I_n to denote the $n\times n$ identity matrix, or simply I if n is clear in context.

In this article most of the linear algebra uses $\mathbb{R}$ as its field of scalars, rather than $\mathbb{C}$. When working over $\mathbb{C}$, we can't use i to both denote $\sqrt{-1}$ and an integer.

It follows that for $\mathbf{u},\mathbf{v}\in\mathbb{R}^n$ one can equally well write

$$\mathbf{u}\cdot\mathbf{v}=\sum_{j=1}^{n}u_jv_j=\mathbf{u}^{\mathrm{T}}\mathbf{v},$$

although technically $\mathbf{u}^{\mathrm{T}}\mathbf{v}$ is a 1×1 matrix. Similarly, if $\mathbf{u},\mathbf{v}\in\mathbb{R}^A$ for a set, A, we will write

$$\mathbf{u}\cdot\mathbf{v}=\sum_{a\in A}^{n}u_av_a=\mathbf{u}^{\mathrm{T}}\mathbf{v},$$

although even if A is finite, i.e., $|A|=n$ for some $n\in\mathbb{N}$, then $\mathbf{u}$ is not really an element of $\mathbb{R}^n$ unless we fix an identification (bijection) of A with $[n]=\{1,\dots,n\}$.

Remark 2.1. If A is infinte, then sums over A, e.g.,

$$\sum_{a\in A}^{n}u_av_a$$

do not necessarily make sense.

For more details on linear algebra, including all the theorems we need, see Appendix **??**.

3. Introduction to the "Kernel Trick"

In this section we introduce the "kernel trick" for a toy model of clustering.

We will spend a fair amount of time discussing "kernel functions" in this course.

3.1. **A Toy Example.** Consider the following setup:

(1) We have a small, finite set, S, such as

$$S=\{\texttt{cow},\texttt{goldfish},\texttt{rabbit}\}.$$

(2) For each $s\in S$, we have a circle, C_s, in $\mathbb{R}^2$. So for each $s\in S$, if (x_s,y_s) is the centre of C_s, and $r_s\geq 0$ is its radius, then

$$C_s=\big\{(x,y)\ \big|\ x=r_s\cos\theta+x_s,\ y=r_s\sin\theta+y_s,\ \text{for some }\theta\in\mathbb{R}\big\}$$

These circles are arbitrary: for example, $C_{\texttt{cow}}$ can intersect $C_{\texttt{goldfish}}$, one can lie inside another, or two circles may even be the same circle.

(3) For each $s\in S$ we have a large set of "training data for s," which we view as a subset $T_s\subset\mathbb{R}^2$.[1] We assume that for each s, T_s looks roughly like a set

[1] If you know what is meant by a *multi-set*, then T_s can be a multi-set, i.e., some points of T_s can "occur multiple times." Multi-sets are very convenient here, but all the ideas can be illustrated using sets.

of random points sampled from C_s (they can have some "noise," meaning that don't have to lie exactly on C_s).

Remark 3.1. You could imagine that the training data $T_s \subset \mathbb{R}^2$ really comes from a bunch of "pictures" of animal s. The pictures themselves may be elements of $\mathbb{R}^N$ for some fixed, large N, but you have designed an algorithm computing a function $f\colon \mathbb{R}^N \to \mathbb{R}^2$ that attempts to "cluster" the pictures of each animal into its own cluster. Maybe you've already built an ANN — artificial neural network — that does an excellent clustering some animals, or the MNIST dataset (or another standard dataset), and now you are hoping it will work well for the animals in S. Of course, we are working with a "toy example:" there is no reason to think f might not cluster well for elements in S, but give clusters that are near perfect circles...

Remark 3.2. We can replace $\mathbb{R}^2$ above with $\mathbb{R}^n$ for some $n \geq 3$; likely n is much smaller than N. One can then map $\mathbb{R}^n$ to a vector that represents all monomials of degree at most 2. More generally, one could map $\mathbb{R}^n$ to all monomials of degree at most d.

Now consider the map $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^6$ given by[2]

$$\Phi(x,y) = (1, \sqrt{2}x, \sqrt{2}y, x^2, \sqrt{2}xy, y^2). \tag{2}$$

It's not hard to see that the average value (or centre of mass) of the set $\Phi(T_s)$ should be different for each circle. For example, say that for some $s \in S$, C_s is the circle of radius r about the origin, $(0,0)$, i.e., the set

$$C_s = \{(x,y) \mid x^2 + y^2 = r^2\}.$$

Over C_s (with its usual arc length), the average value of $\Phi(C_s)$ is:

$$\text{Average}_{x \in C_s}\big(\Phi(x)\big) = \big(1, 0, 0, r^2/2, 0, r^2/2\big); \tag{3}$$

in class we explained roughly why this is true. So one might expect that the average value of T_s would be roughly this.

Exercise 3.1. The general equation of a circle of radius r centred at (x_0, y_0) in the plane is:

$$x(\theta) = x_0 + r\cos\theta, \;\; y(\theta) = y_0 + r\sin\theta.$$

Evaluate:

$$\frac{1}{2\pi}\int \big(1, \sqrt{2}x, \sqrt{2}y, x^2, \sqrt{2}xy, y^2\big)\, d\theta.$$

Are these averages different for different values of θ? What about for an ellipse with axes parallel to the x- and y-axes? What about for general *conic sections* (i.e., ellipses, parabolas, or hyperbolas)?

Now let's say that someone give you a "test point" $(x^{\text{test}}, y^{\text{test}}) \in \mathbb{R}^2$; you want to decide if this is a cow or a goldfish.

Exercise 3.2. (This exercise was done in class in the 2025 version of this course; it was inspired by remarks of YJ.) Fix an $(x^{\text{test}}, y^{\text{test}}) \in \mathbb{R}^2$; say that $(x^{\text{test}}, y^{\text{test}})$ lies on the circle of radius ρ about $(0,0)$, i.e.,

$$\rho^2 = \big(x^{\text{test}}\big)^2 + \big(y^{\text{test}}\big)^2.$$

Show that:

[2]We are grateful to ML who corrected our original omission of the $\sqrt{2}$ factors.

(1)

$$\left\|\Phi\big(x^{\text{test}}, y^{\text{test}}\big) - (1, 0, 0, r^2/2, 0, r^2/2)\right\|^2 = f\big(x^{\text{test}}, y^{\text{test}}\big) + r^2\left(\frac{r^2}{2} - \rho^2\right),$$

where f is a function independent of ρ (and show your work, don't simply cut and paste from your favourite "AI").

(2) Show that for fixed ρ, the function $g\colon \mathbb{R} \to \mathbb{R}$

$$g(r) = r^2\left(\frac{r^2}{2} - \rho^2\right)$$

is minimized at $r = \pm\rho$. [Remark: it may be simpler to work with $h(x) = x(x/2 - a)$ for fixed $a \in \mathbb{R}$.]

One problem with $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^6$ is that it increases the dimension, which is not good news. (The calculation in the Exercise 3.2 should convince you of this.) So we use the "kernel trick," which we now describe.

3.2. **New Notation, and the Fundamental Lemma of the Kernel Trick.**

WARNING 3.3. At this point we are switching notation: we use $\mathbf{x}^{\text{test}} \in \mathbb{R}^2$. Hence $\mathbf{x}^{\text{test}} = (x_1^{\text{test}}, x_2^{\text{test}})$ rather than the notation in the previous subsection, namely $(x^{\text{test}}, y^{\text{test}})$.

We now point out that although $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^6$ in (2) is a (non-linear) map, give a sample of cows and goldfish plus a "test" element $\mathbf{x}^{\text{test}} \in \mathbb{R}^2$, we never really need to work in $\mathbb{R}^6$.

The idea is as follows:

(1) for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^2$, we have

$$\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}') = (1 + \mathbf{x} \cdot \mathbf{x}')^2 = (1 + x_1x_1' + x_2x_2')^2; \tag{4}$$

and

(2) to see whether or not $\Phi(\mathbf{x}^{\text{test}})$ is closer to the average value of Φ applied to a subset, $C \subset \mathbb{R}^2$ (of "`cows`") or of $G \subset \mathbb{R}^2$ (representing "`goldfish`"), it suffices to consider

$$\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')$$

where $\mathbf{x}, \mathbf{x}'$ vary over all values of C, G, and x^{test}.

Remark 3.4. Notice[3] that algorithmically it is not particularly difficult to compute the quantity in (4) as

$$1 + 2x_1x_1' + 2x_2x_2' + x_1^2(x_1')^2 + 2x_1x_2x_1'x_2' + x_2^2(x_2')^2.$$

Hence this doesn't seem like a big savings to write this as

$$= (1 + x_1x_1' + x_2x_2')^2,$$

perhaps a factor of three or four, and even less if you are working with software that computes dot products quickly. However, there are two ways that this part of the kernel trick could ultimately be useful:

[3]We thank TL for this remark.

(1) for analogous kernels, such as
$$(1+\mathbf{x}\cdot\mathbf{x}')^d$$
where d is larger and/or $\mathbf{x},\mathbf{x}'\in\mathbb{R}^n$ with n large, this could represent a bigger savings; and
(2) the study of such kernels may help to explain why certain algorithms aren't working well.

Lemma 3.5. *Let $n\in\mathbb{N}$, $\boldsymbol{\phi}^{\text{test}}\in\mathbb{R}$, and $C,G\subset\mathbb{R}^n$ be two subsets*[4]*. Let*
$$C_{\text{avg}}=\frac{1}{|C|}\sum_{\mathbf{c}\in C}\mathbf{c},\quad G_{\text{avg}}=\frac{1}{|G|}\sum_{\mathbf{g}\in G}\mathbf{g},$$
where are therefore elements of $\mathbb{R}^n$. Then $\boldsymbol{\phi}^{\text{test}}$ is closer to C_{avg} than to G_{avg} iff
$$\|\boldsymbol{\phi}^{\text{test}}-C_{\text{avg}}\|^2<\|\boldsymbol{\phi}^{\text{test}}-G_{\text{avg}}\|^2,$$
and hence iff
$$\frac{1}{|C|}\sum_{\mathbf{c}\in C}\big(\boldsymbol{\phi}^{\text{test}}-\mathbf{c}\big)^2>\frac{1}{|G|}\sum_{\mathbf{g}\in G}\big(\boldsymbol{\phi}^{\text{test}}-\mathbf{g}\big)^2,$$
hence iff
$$\frac{1}{|C|}\sum_{\mathbf{c}\in C}\big(\boldsymbol{\phi}^{\text{test}}\cdot\boldsymbol{\phi}^{\text{test}}-2\boldsymbol{\phi}^{\text{test}}\cdot\mathbf{c}+\mathbf{c}\cdot\mathbf{c}\big)>\frac{1}{|G|}\sum_{\mathbf{g}\in G}\big(\boldsymbol{\phi}^{\text{test}}\cdot x^{\text{test}}-2\boldsymbol{\phi}^{\text{test}}\cdot\mathbf{g}+\mathbf{g}\cdot\mathbf{g}\big)\tag{5}$$
Hence this condition can be expressed completely in terms of the sizes of C,G, and the dot products between elements of C and G and $\boldsymbol{\phi}^{\text{test}}$.

The proof of this lemma should be straightfoward (perhaps assign as an EXERCISE).

3.3. **Symmetric and Positive (Semi)Definite Matrices: Basic Examples.** Before discussing kernel functions abstractly, it may be better to give some examples of positive semidefinite and definite matrices. Most of these examples we will cover in class. In the next section we will prove that if $K\in\mathbb{R}^{n\times n}$ (an $n\times n$ matrix with real entries), then the following are equivalent:

(1) $K^{\mathrm{T}}=K$ (i.e., K is symmetric);
(2) $KU=U\Lambda$ where U is an ($n\times n$ real) orthogonal matrix[5] (i.e., $UU^{\mathrm{T}}=I$, or equivalently U's columns form an orthonormal basis of $\mathbb{R}^n$, i.e., equivavlently U's rows), and Λ is the diagonal matrix
$$\operatorname{diag}(\lambda_1,\dots,\lambda_n)=\begin{bmatrix}\lambda_1&&\\&\ddots&\\&&\lambda_n\end{bmatrix}$$
with $\lambda_j\in\mathbb{R}$ for all j;
(3) K has an orthonormal eigenbasis, with all eigenpairs (eigenvectors and eigenvalues) purely real;
(4) etc.

[4]More generally, C,G can be multi-sets, if you know what this means.

[5]Historically an **orthogonal** matrix is one whose columns (or whose rows) forn an **orthonormal** basis. Yuck...

If $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ is any orthonormal eigenbasis of K, with corresonding eigenvalues $\lambda_1, \ldots, \lambda_n$ (i.e. $K\mathbf{v}_j = \lambda_j \mathbf{v}_j$), then

$$K = \sum_{j=1}^{n} \lambda_j \mathbf{v}_j \mathbf{v}_j^{\mathrm{T}}. \tag{6}$$

This formula can be viewed as a "baby version"[6] of the usual spectral theorem for bounded, self-adjoint operators (on a Hilbert space).

Example 3.6. The matrix

$$K = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} = \mathbb{1}\mathbb{1}^{\mathrm{T}}, \quad \text{where} \quad \mathbb{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

is a rank one matrix. It has one eigenvalue, $\lambda_1 = n$, and the rest, λ_j for $2 \leq j \leq n$ are equal to 0. One normalized eigenvector $\mathbf{v}_1$ corresponding to the eigenvalue λ_1 is the vector $\mathbb{1}/\sqrt{n}$. This illustrates (6)

$$K = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} = \mathbf{v}_1 \mathbf{v}_1^{\mathrm{T}} \lambda_1 = \big(\mathbb{1} \,/\, \sqrt{n}\big)\big(\mathbb{1} \,/\, \sqrt{n}\big)^{\mathrm{T}} n$$

In case we may remark that if $\mathbf{v}_1 = \mathbb{1}/\sqrt{n}$, and $|\lambda_2|, \ldots, |\lambda_n|$ are all "much smaller" than λ_1, then K is "close" to the matrix $\mathbb{1}\mathbb{1}^{\mathrm{T}}(\lambda_1/n)$; this is very important in the theory of *expanding graphs*; we will return to this later in course, in order to compare regular graphs that "expand well" versus those that don't.

Example 3.7. A lot of situations can be intuitively understood by the 2×2 symmetric matrix:

$$\begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

Whose eigenpairs are evident from

$$\begin{bmatrix} a & b \\ b & a \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (a+b) \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} a & b \\ b & a \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = (a-b) \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

For example, if we have a Markov chain consisting of two groups, say Democrats and Republicans, where communication per year from one group to another is rare, but where communication per year within each group is common, then one can model an associate Markov chain (visiting friends, webpages that point to other pages for a PageRank computation, etc.) as a Markov chain

$$\begin{bmatrix} 0.999999 & 0.000001 \\ 0.000001 & 0.999999 \end{bmatrix},$$

which has eigenvalues 1 and 0.999998. For this reason, the *mixing time* of this Markov chain is on the order of 10^6 (years). By contrast, the Markov chain

$$\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

[6]The term "baby version" is not necessarily pejorative; it is often a very simple class of examples of a more general theorem, but one giving essential insight into the theorem.

completely mixes after a single iteration, and has $\lambda_1 = 1$, $\lambda_2 = 0$; hence this is a special case of Example 3.6 multiplied by $1/2$. By contrast, the (irreducible) Markov chain

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

modelled by various phenomena[7] has $\lambda_1 = 1$ and $\lambda_2 = -1$, so $|\lambda_2| = \lambda_1$ and the mixing time of this Markov chain is infinite.

Remark: it is easy to see that if K is a symmetric matrix and $\mathbf{v}, \mathbf{u}$ are eigenvectors of K with distinct eigenvalues, then $\mathbf{v}, \mathbf{u}$ are necessarily orthogonal.

Example 3.8. *Circulant matrices* generalize both Examples 3.6 and 3.7: most generally these are matrices of the form:

$$K = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & \cdots & a_n \\ a_n & a_1 & a_2 & a_3 & \cdots & a_{n-1} \\ a_{n-1} & a_n & a_1 & a_2 & \cdots & a_{n-2} \\ a_{n-2} & a_{n-1} & a_n & a_1 & \cdots & a_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_2 & a_3 & a_4 & & \cdots & a_1 \end{bmatrix},$$

i.e., these are $n \times n$ matrices K whose (i,j)-th entry, K_{ij} is just a function of $i - j$ modulo n. Hence, for $n \geq 3$, these matrices are not necessarily symmetric; these matrices are symmetric iff $a_2 = a_n$, $a_3 = a_{n-1}$, etc. If ζ is any n-th root of unity (i.e., $\zeta \in \mathbb{C}$ satisfies $\zeta^n = 1$), we easily check that (whether or not K is symmetric)

$$K \begin{bmatrix} 1 \\ \zeta \\ \zeta^2 \\ \vdots \\ \zeta^{n-1} \end{bmatrix} = \left(a_1 + \zeta a_2 + \cdots + \zeta^{n-1} a_n\right) \begin{bmatrix} 1 \\ \zeta \\ \zeta^2 \\ \vdots \\ \zeta^{n-1} \end{bmatrix}.$$

This gives an orthogonal set of eigenvectors — provided that we work with "the standard inner product"[8] on $\mathbb{C}^n$. These become an orthonormal set upon dividing each by $\sqrt{n}$.

Example 3.9. The *standard* 2×2 *Hadamard matrix*

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Then H_2 is a symmetric matrix, and we easily check that $H_2 H_2 = 2I$, and hence

$$U_2 = \frac{1}{\sqrt{2}} H_2$$

is a unitary matrix, and we easily check that its eigenvalues are ± 1; similarly the eigenvalues of H_2 are $\pm\sqrt{2}$. (exercise). We easily see that a set of corresponding

[7] Perhaps sneetches?

[8] As mentioned earlier, this is not entirely standard, and you have to specify whether the inner product will be linear on the first or second vector.

eigenvectors are given by $(1, \pm\sqrt{2}-1) \in \mathbb{R}^2$. Let

$$H_4 = H_2 \otimes H_2 = \begin{bmatrix} H_2 & H_2 \\ H_2 & -H_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

where we have used *block matrix form* and the *tensor product* of matrices; more generally if $A \in \mathbb{R}^{m_1 \times n_1}$ and $B \in \mathbb{R}^{m_2 \times n_2}$ are matrices of any dimensions, we define $A \otimes B$ as the $\mathbb{R}^{m_1 m_2 \times n_1 n_2}$ to be the matrix given by

$$(A \otimes B)(\mathbf{u} \otimes \mathbf{v}) = (A\mathbf{u}) \otimes (B\mathbf{v}),$$

where $\mathbf{u} \otimes \mathbf{v}$ is defined as the vector whose components are $u_i v_j$ ranging over all the entries u_i of $\mathbf{u}$ and v_j of $\mathbf{v}$: there is some freedom in how we arrange the entries $u_i v_j$, although it is convenient to choose one of the two standard ways of doing this; so if $\mathbf{u} \in \mathbb{R}^{n_1}$ and $\mathbf{v} \in \mathbb{R}^{n_2}$, then we can set

$$\mathbf{u} \otimes \mathbf{v} = \big(u_1 v_1, u_1 v_2, \ldots, u_1 v_{n_2}, u_2 v_1, \ldots, u_{n_1} v_{n_2}\big), \tag{7}$$

which corresponds to listing the elements of $\mathbf{u}\mathbf{v}^{\mathrm{T}}$ row by row (starting with the first column); the other way is to list these elements column by column, or, equivalently, those of $\mathbf{u}\mathbf{u}^{\mathrm{T}}$ row by row. More generally, we define the *standard* $2^m \times 2^m$ *Hadamard matrix* to be the matrix

$$H_{2^m} = H_2 \otimes H_{2^{m-1}} = H_2 \otimes H_2 \otimes \cdots \otimes H_2 \quad (m \text{ times}).$$

(we easily see that the resulting $2^m \times 2^m$ doesn't depend on which of the standard orderings of components we use to define $\mathbf{u} \otimes \mathbf{v}$). More generally, an $n \times n$ *Hadamard matrix* is an element $H \in \mathbb{R}^{n \times n}$ whose entries are ± 1 and such that $H^2 = I$; it is a currently (2025) open question to know for which values of $n \in \mathbb{N}$ there exists an $n \times n$ Hadamard matrix. The exercises below show that if A is an arbitrary square matrix with (algebraic) eigenvalues $\lambda_1, \ldots, \lambda_n$, and B one with eigenvalues $\mu_1, \ldots, \mu_m$, then the eigenvalues of $A \otimes B$ are $\lambda_i \mu_j$ ranging over all $i \in [n]$ and $j \in [m]$. It easily follows that H_{2^m} has half of its eigenvalues equal to $2^{m/2}$, the other half $-2^{m/2}$.

3.4. **Examples from Graph Theory.** Graph theory gives a set of examples of symmetric and positive semidefinite matrices that go a long way to develop intuition regarding eigenvectors and eigenvalues.

3.4.1. *Simple Graphs.*

Definition 3.10. If A is a set, then $\binom{A}{2}$ refers to the set of subsets of A of size 2. A *simple graph* refers to pair $G = (V, E)$ where V, E are both countable (i.e., finite or countably infinite) such that $E \subset \binom{V}{2}$. We also write $G = (V_G, E_G)$ to emphasize G or in discussions more than one graph.

Remark 3.11. For some applications of graph theory, it is enough to work with simple graphs. For many applications of graph theory, many important ideas and methods become needlessly awkward or impossible unless one works with a more general notion of graph (examples of these are ideas are covering maps and regular graphs, especially *relative expanders*).

There are five important matrices we associate with any simple graph, G: A_G — its *adjacency matrix*, D_G — its *degree matrix*, Λ_G — its *Laplacian*, and ∂_G — its *incidence matrix*, B_G — its *non-backtracking matrix*, most of which we will soon define. So the symbol $A_G, A_{G'}, A_{\tilde{G}}, A_H$ etc. usually refers to the adjacency matrix of a graph (not necessarily a simple graph); similarly for $V_G, E_G, D_G, \Lambda_G, \partial_G$, etc.

Definition 3.12. Let $G = (V_G, E_G)$ be a simple graph. Then A_G refers to the $V_G \times V_G$ matrix whose entries are 0's and 1's according to the following rule: $A_G(v, v') = 1$ if $\{v, v'\} \in E_G$ (i.e., if v, v' are *adjacent*, i.e., v, v' are *joined by an edge*); if not, then $A_G(v, v') = 0$. Hence A_G can be viewed as an element of $\mathbb{R}^{V_G \times V_G}$. If V_G is finite, and $|V_G| = n$, then we can order the vertices of G as $V_G = \{v_1, \dots, v_n\}$, whereby A_G is an $n \times n$ matrix.

Therefore A_G is a symmetric matrix, and its eigenvalues are real and can be ordered:

$$\lambda_n(A_G) \le \cdots \le \lambda_2(A_G) \le \lambda_1(A_G). \tag{8}$$

Remark 3.13. Let be V a set, and $A \in \mathbb{R}^{V \times V}$ be symmetric (i.e., $A^{\mathrm{T}} = A$), and whose entries are 0's and 1's, such that all the diagonal entries of A vanish (i.e., equal 0). Then there is a unique simple graph $G = (V_G, E_G)$ such that $V_G = V$ and $A_G = A$.

Definition 3.14. Let G be a simple graph, and $v, v' \in V_G$. Then a *walk in G (of length k) (beginning at v and ending at v')* refers to a sequence of vertices of G, i.e., of elements of V_G,

$$w = (v_0, v_1, \dots, v_k)$$

such that $v_0 = v$ and $v_k = v'$.

If G is a simple graph, then it is not hard to see that for each $k \in \mathbb{N} = \{1, 2, \dots, \}$, the $V_G \times V_G$ matrix A_G^k (meaning $(A_G)^k$) has an important meaning, namely that

$$\forall v, v' \in V_G \quad (A_G^k)(v, v') = \big|\{ \text{ walks in } G \text{ of length } k \text{ beginning at } v \text{ and ending at } v' \ \}\big|.$$

Example 3.15. If $\mathbb{B}^n$ is the n-dimension Boolean hypercube, which is a simple graph with vertex set $\{0,1\}^n$ (also commonly $\{1,-1\}^n$ or $\{F,T\}^n$), and has edges joining two vertices of Hamming distance 1, then $\mathbb{B}^n$ satisfies

$$\mathbb{B}^n = \mathbb{B}^1 \times \mathbb{B}^{n-1},$$

where $\times$ is the usual *Cartesian product* of graphs. One way of definiting Cartesian product (we will draw pictures in class when we cover it) is that if A_G denotes the adjacency matrix of a graph, $G = (V_G, E_G)$, then

$$A_{G \times H} = A_G \otimes I_{V_H} + I_{V_G} \otimes A_H.$$

Exercises will include computing the adjacency matrix eigenvalues of a *cycle of length n* and its products.

3.5. **Spectral Gaps in Regular Graphs.** In 2025 we discussed some basic aspects of spectral theory (the eigenvalues and eigenvectors) of d-regular matrices. The point is that adjacency matrix eigenvalues give intuition regarding the "expanding" properties of graphs. Here are some main points.

In 2025, we computed the eigenvalues of grid graphs, G, (which we took to mean the product of two cycle graphs, which is therefore a 4-regular graphs, looking like

a grid but with "wrap around" edges). We showed that such a graph, G, being the product of two cycles, say of lengths N_1 and N_2, has adjacency eigenvalues

$$\lambda_{j_1,j_2} = 4 - 2\cos(2\pi j_1/N_1) - 2\cos(2\pi j_2/N_2)$$

for $j_1 \in \mathbb{Z}/N_1\mathbb{Z}$ and $j_2 \in \mathbb{Z}/N_2\mathbb{Z}$. Hence the largest A_G eigenvalue is 4, the next largest is

$$4 - 2\cos\big(2\pi/\max(N_1, N_2)\big),$$

which for $n = N_1N_2$ (the number of vertices of this graph) is roughly

$$4 - \frac{C_1}{(\max(N_1, N_2))^2}$$

for a constant C_1, and $\max(N_1, N_2)$ is anywhere between

$$\sqrt{n} \le \max(N_1, N_2) \le n/3$$

(since simple graphs don't have cycles of length 1 or 2). Also the smallest of this graph is -4 iff N_1, N_2 are both even.

We also made the following definitions and claims.

Definition 3.16. Let $G = (V_G, E_G)$ be a simple graph. For each $v \in V_G$, we define the *degree of v (in G)*, denoted $\deg_G(v)$, to be the number of edges upon which v is incident, i.e., the number of vertices adjacent to v, i.e., $A_G\mathbb{1}_v$, where $\mathbb{1}_v$ is the vector that is 1 in its v-component, and 0 elsewhere. We say that G is *d-regular* if $V \neq \emptyset$, and for all $v \in V$, $\deg_G(v) = d$; equivalently $A_G\mathbf{1} = d\mathbf{1}$.

Recall from (8), we typically order the eigenvalues of A_G (for any graph, G) from largest to smallest, i.e.,

$$\lambda_n(A_G) \le \cdots \le \lambda_2(A_G) \le \lambda_1(A_G). \tag{9}$$

The following facts are not difficult to prove (perhaps after a few pointers).

Proposition 3.17. *Let G be a d-regular graph, and order its adjacency eigenvalues as usual, i.e., Then the following hold:*

1. *$\lambda_1(A_G) = d$;*
2. *the multiplicity of d as an eigenvalue of A_G is the number of* connected components *of the graph G;*
3. *$\lambda_n(A_G) \ge -d$;*
4. *the multiplicity of $-d$ as an eigenvalue of A_G is the number of connected components of the graph G that are* bipartite graphs.

For definitions and a proof, see Exercise C.1.

Definition 3.18. For any graph, we define its *adjacency spectral gap* to be $\lambda_1(G) - \lambda_2(G)$, and its *subdominant adjacency spectral radius* to be

$$\lambda(G) = \max_{2\le i\le n} |\lambda_i(G)|. \tag{10}$$

Hence a graph has 0 spectral gap iff it is disconnected (not connected). The definition of $\lambda(G)$ is standard in the literature on *expanders*. To understand the interest in $\lambda(G)$, we remark that the spectral decomposition

$$A_G = \sum_{i=1}^{n} \lambda_i(A_G)\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i$$

for an orthonormal eigenbasis $\mathbf{v}_1,\dots,\mathbf{v}_n$ ($A\mathbf{v}_i=\lambda_i\mathbf{v}_i$) implies that

$$A_G=\frac{d}{n}\begin{bmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&\ddots&\vdots\\1&1&\cdots&1\end{bmatrix}+\mathcal{E}_G,$$

therefore $\mathcal{E}_G$ is symmetric, with eigenvalues 0 and $\lambda_2(G),\dots,\lambda_n(G)$. The spectral theorem implies that

$$\|\mathcal{E}_G\|_{L^2(V_G)}=\lambda(G)$$

From this observation, we easily get the following remarkably useful result in applications.

[In early October, 2025, we didn't yet cover the stuff below.]

Proposition 3.19. *Let G be a d-regular graph. Then for any $U,W\subset V_G$, and any $k\in\mathbb{N}$, we have*

$$\Big|\mathbb{1}_U^{\mathrm{T}}A_G^k\mathbb{1}_W-(d^k/n)|U|\,|W|\Big|\le\big(\lambda(G)\big)^k\sqrt{\frac{|U|\,(n-|U|)}{n}}\sqrt{\frac{|W|\,(n-|W|)}{n}}\le\big(\lambda(G)\big)^k\sqrt{|U|\,|W|}.$$

(Here $\mathbb{1}_U$ denotes the indicator function of U, i.e., $\sum_{u\in U}e_u$, where e_u is the standard basis vector, i.e., the function that is 1 on U and 0 elsewhere, i.e., on $V_G\setminus U$.)

The special case $k=1$ of the above immediately implies the weaker (but simpler) estimate

$$\Big|e(U,W)-(d/n)|U|\,|W|\Big|\le\lambda(G)\sqrt{|U|\,|W|}$$

which is often called the *expander mixing lemma*.

ADD MORE STUFF HERE OR IN THE EXERCISES?

3.6. **Laplacians in Graph Theory.** If $G=(V,E)$ is a simple graph, let $\partial\in\mathbb{R}^{E\times V}$ be any matrix such that $\phi(e,v)=0$ if e is not incident upon v, and if $e=\{v_1,v_2\}$ then one of $\phi(v_1),\phi(v_2)$ equals 1, the other -1.[9] Then we define the *graph Laplacian of G*, denoted Δ_G, to be the matrix

$$\Delta_G=\partial\partial^{\mathrm{T}},$$

which doesn't depend on the ±1 choice above, and is easily seen to equal

$$\Delta_G=\partial\partial^{\mathrm{T}}=D_G-A_G,$$

where A_G is the adjacency matrix of G, and D_G is the *degree counting matrix* of G, namely a diagonal matrix whose (v,v)-entry equals $\deg_G(v)$ (i.e., the number of edges incident upon v, i.e., the number of vertices adjacent to v).

[9]The choice may seem "ad hoc," but this is really a way of simplifying the fact that ∂ is canonically defined, but the choice of ±1 above is really a choice of a basis for $\mathbb{R}^2/\mathrm{diag}$, where diag is the diagonal subspace of $\mathbb{R}^2$. This is explained in [Fri15].

3.7. **Kernels: The Low Road.** One way to describe kernel functions is by first describing what is meant by a positive (semi)definite $n \times n$ matrix, and then use this to define kernel functions. In the 2025 version of this course, we took this route.

Definition 3.20. Let $K \in \mathbb{R}^{n\times n}$ be a real, $n \times n$ matrix. We say that K is:

(1) *symmetric* if $K^{\mathrm{T}} = K$, i.e., $K_{ij} = K_{ji}$ for all $i, j \in [n]$;
(2) *positive semidefinite* if

$$\forall \mathbf{v} \in \mathbb{R}^n, \mathbf{v}^{\mathrm{T}} K \mathbf{v} \geq 0; \tag{11}$$

and
(3) (11) holds with equality iff $\mathbf{v} = \mathbf{0}$.

Definition 3.21. Let $\mathcal{X}$ be a set. A *kernel (function) on* $\mathcal{X}$ is a function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For each finite subset $\mathcal{X}' \subset \mathcal{X}$ of size $m \in \mathbb{N}$, writing $\mathcal{X}'$ as $\{x_1, \ldots, x_m\}$, we identify the restriction, $k|_{\mathcal{X}'\times\mathcal{X}'}$, of k to $\mathcal{X}' \times \mathcal{X}'$, with the $m \times m$ real matrix

$$k|_{(x_1,\ldots,x_m)} \stackrel{\text{def}}{=} \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_m) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_m) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_m,x_1) & k(x_m,x_2) & \cdots & k(x_m,x_m) \end{bmatrix}.$$

[We use the notation $k|_{(x_1,\ldots,x_m)}$ to stress that the resulting matrix depends on the order of $x_1, \ldots, x_m$.] Sometimes, when the order of the elements of $\mathcal{X}'$ is understood (or unimportant), we simply write $k|_{\mathcal{X}'\times\mathcal{X}'}$ or just $k|_{\mathcal{X}'}$; alternatively, we also call $\mathcal{X}'$ a *finite ordered set* if $\mathcal{X}'$ is finite, and if it comes with an ordering $x_1, \ldots, x_m$ of its elements. We say that: k is *symmetric* (respectively, *positive semidefinite*, *positive definite*, etc.) if for all finite ordered sets, $\mathcal{X}'$, $k|_{\mathcal{X}'}$ is symmetric (respectively, positive semidefinite, etc.).

Remark 3.22. Some authors will use the term *kernel (function)* to mean a kernel (function) that is necessarily symmetric and positive semidefinite. These occur commonly in what people in AI/ML/etc. call *kernel methods*.

Example 3.23. Let $\mathcal{X} = [n] = \{1, \ldots, n\}$, and let $[n]$ be given its usual ordering. Hence $\mathcal{X}$ is a set with an implied order, so a kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be identified with an $n \times n$ matrix.

Example 3.24. Let $\mathcal{X} = [3] = \{1, 2, 3\}$, and let k be the kernel function identified with the element $K \in \mathbb{R}^{3\times 3}$ given as follows:

(1) Let $\Phi\colon [3] \to \mathbb{R}^2$ be the function

$$\Phi(1) = \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \quad \Phi(2) = \begin{bmatrix} -17 \\ 5 \end{bmatrix}, \quad \Phi(3) = \begin{bmatrix} \pi \\ e \end{bmatrix}$$

(making $\Phi(3)$ both irrational and transcendental).
(2) Define $k(x, x') = \Phi(x) \cdot \Phi(x')$.

K is therefore the 3×3 real matrix:

$$K = \begin{bmatrix} \begin{bmatrix}5\\1\end{bmatrix}\cdot\begin{bmatrix}5\\1\end{bmatrix} & \begin{bmatrix}5\\1\end{bmatrix}\cdot\begin{bmatrix}-17\\5\end{bmatrix} & \begin{bmatrix}5\\1\end{bmatrix}\cdot\begin{bmatrix}\pi\\e\end{bmatrix} \\ \begin{bmatrix}-17\\5\end{bmatrix}\cdot\begin{bmatrix}5\\1\end{bmatrix} & \begin{bmatrix}-17\\5\end{bmatrix}\cdot\begin{bmatrix}-17\\5\end{bmatrix} & \begin{bmatrix}-17\\5\end{bmatrix}\cdot\begin{bmatrix}\pi\\e\end{bmatrix} \\ \begin{bmatrix}\pi\\e\end{bmatrix}\cdot\begin{bmatrix}5\\1\end{bmatrix} & \begin{bmatrix}\pi\\e\end{bmatrix}\cdot\begin{bmatrix}-17\\5\end{bmatrix} & \begin{bmatrix}\pi\\e\end{bmatrix}\cdot\begin{bmatrix}\pi\\e\end{bmatrix} \end{bmatrix} = \begin{bmatrix} 26 & -80 & 5\pi+e \\ -80 & 314 & -17\pi+5e \\ 5\pi+e & -17\pi+5e & \pi^2+e^2 \end{bmatrix}$$

Note that if $\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3)\in\mathbb{R}^3$, then

$$\boldsymbol{\alpha}^{\mathrm{T}} K \boldsymbol{\alpha} = \sum_{i,j=1}^{3} \alpha_i\alpha_j k(i,j) = \left\| \alpha_1 \begin{bmatrix}5\\1\end{bmatrix} + \alpha_2 \begin{bmatrix}-17\\5\end{bmatrix} + \alpha_3 \begin{bmatrix}\pi\\e\end{bmatrix} \right\|^2 \tag{12}$$

$$= \Big\| \alpha_1\Phi(1) + \alpha_2\Phi(2) + \alpha_3\Phi(3) \Big\|^2. \tag{13}$$

Since no two of $\Phi(1),\Phi(2),\Phi(3)$ are colinear, we have that each 2×2 principal minor of K is positive definite. However, for any solution of

$$\alpha_1\Phi(1) + \alpha_2\Phi(2) + \alpha_3\Phi(3) = \begin{bmatrix}0\\0\end{bmatrix}$$

we have that

$$\boldsymbol{\alpha}^{\mathrm{T}} K \boldsymbol{\alpha} = \sum_{i,j} \alpha_i\alpha_j k(i,j) = 0.$$

One can, of course, see that K is of rank at most 2 by writing K more simply as:

$$K = \begin{bmatrix}5\\-17\\\pi\end{bmatrix}\begin{bmatrix}5 & -17 & \pi\end{bmatrix} + \begin{bmatrix}1\\5\\e\end{bmatrix}\begin{bmatrix}1 & 5 & e\end{bmatrix} = \begin{bmatrix}5 & 1\\-17 & 5\\\pi & e\end{bmatrix}\begin{bmatrix}5 & -17 & \pi\\1 & 5 & e\end{bmatrix}$$

Hence

$$K = M^{\mathrm{T}} M \quad \text{where} \quad M = \begin{bmatrix}5 & -17 & \pi\\1 & 5 & e\end{bmatrix}. \tag{14}$$

The above example generalizes in an evident fashion. Here is the ML/AI terminology.

Definition 3.25. Let $\mathcal{X}$ be a set. If $\Phi\colon \mathcal{X}\to\mathbb{R}^f$ for some $f\in\mathbb{N}$, we define the *resulting (bilinear) kernel associated to* Φ to be the kernel function $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ given by

$$k(x,x') = \Phi(x)\cdot\Phi(x').$$

Φ is often called the *feature map*, and f the *number of features*. k might be called a *finite dimensional kernel*.

The following theorem is more or less evident from (12) and (13) in Example 3.24.

Theorem 3.26. *Let $\Phi\colon \mathcal{X}\to\mathbb{R}^f$ be a set theoretic map, and k the associated kernel function. Then k is positive semidefinite (we understand this to mean that k is also symmetric). Moreover, k is positive definite iff for all $n\in\mathbb{N}$ and distinct $x_1,\dots,x_n\in\mathcal{X}$, we have that that the only solution in $\alpha\in\mathbb{R}^n$ to*

$$\alpha_1\Phi(x_1) + \cdots + \alpha_n\Phi(x_n) = \mathbf{0}$$

is the trivial solution $\boldsymbol{\alpha} = \mathbf{0}$; equivalently, Φ is injective, and $\Phi(\mathcal{X})$ is a set of linearly independent vectors in $\mathbb{R}^f$.

Corollary 3.27. *Let k be the kernel function associated to a "feature map" $\Phi\colon \mathcal{X} \to \mathbb{R}^f$. If $f < |\mathcal{X}|$, then k is not positive definite.*

Remark 3.28. The "kernel trick" in Subsections 3.1 and 3.2 used the fact that for the kernel function there, namely k associated to

$$\Phi(\mathbf{x}) = \big(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2\big),$$

can be done by a computation in $\mathbb{R}^2$, namely

$$k(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}') = \big(1 + \mathbf{x} \cdot \mathbf{x}'\big)^2.$$

Hence, if we don't mind making computations in $\mathbb{R}^f$ for a function $\Phi\colon \mathcal{X} \to \mathbb{R}^f$, then you can still speak of a "feature map," number of features, etc., without having a "kernel trick" at hand.

3.8. **Kernels: The High Road.** One can define kernel functions in general, and then to discuss the special case of kernel functions on a finite set, which are $n \times n$ matrices. This was not done in the 2025 version of this course.

Definition 3.29. Let $\mathcal{X}$ be a set. A *kernel on* $\mathcal{X}$ is a function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We say that:

(1) k is *symmetric* if for all $x, x' \in \mathcal{X}$, $k(x, x') = k(x', x)$;
(2) k is *positive semidefinite* if k is *symmetric* for any $n \in \mathbb{N} = \{1, 2, \ldots\}$, any $x_1, \ldots, x_n \in \mathcal{X}$, and all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, we have

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0, \tag{15}$$

(3) k is *positive definite* if it is positive semidefinite, and if equality in (15) (i.e., the left-hand-side equals 0) iff all the $\alpha_1, \ldots, \alpha_n$ are zero.

WARNING 3.30. In the definition below we will eventually see that for $\mathcal{X} = \mathbb{R}^2$, the function $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$k(x, x') = \Phi(x) \cdot \Phi(x') \tag{16}$$

with Φ as in (2) is a positive definite kernel. This is **not trivial** to prove. It can also be a **serious misconception** to regard $\mathcal{X}$ as a vector space in this context, for the following reason: $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^6$ is **not a linear function**. This makes it **non-trivial (but still quite easy, once you see the trick)** to prove that k above is a positive **semidefinite** kernel function. **Proving that a kernel is positive definite can be a much more subtle matter** (although not hard, once you see a few tricks).

WARNING 3.31. If U, V are finite dimensional real vector spaces of positive dimension, and $\Phi\colon U \to V$ a linear[10] map, then $k\colon U \times U \to \mathbb{R}$ given by (16) will **never be positive definite**. Indeed, for any $u \in U$ we have

$$k(u, u) + k(u, -u) + k(-u, u) + k(-u, -u) = k(u, u)(1 - 1 - 1 + 1) = 0.$$

Hence for $x_1 = u$ and $x_2 = -u$ and $\alpha_1 = \alpha_2 = 1$ we have

$$\sum_{i,j=1}^{2} \alpha_i \alpha_j k(x_i, x_j) = 0.$$

[10] Added "linear" thanks to ML.

* * *

You likely have seen such "kernels" when $\mathcal{X}$ is finite, as the following example shows.

Example 3.32. Let $n \in \mathbb{N}$ and $\mathcal{X} = [n] = \{1, \dots, n\}$. To any kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, or equivalently $k\colon [n] \times [n] \to \mathbb{R}$, we associate the $n \times n$ matrix K given by

$$K = \begin{bmatrix} k(1,1) & k(1,2) & \cdots & k(1,n) \\ k(2,1) & k(2,2) & \cdots & k(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ k(n,1) & k(n,2) & \cdots & k(n,n) \end{bmatrix}$$

Then k is symmetric, positive semidefinite, and positive definite, respectively, iff K, as a matrix, is, respectively, symmetric, positive semidefinite, and positive definite in the usual sense of linear algebra. So k is symmetric iff $K^{\mathrm{T}} = K$, where K^{T} is the transpose of K. In the next section we will review the notion of positive semidefinite and definite matrices, and discuss their spectral properties (i.e., eigenvalues and eigenvectors).

WARNING 3.33. We now see that an $n \times n$ matrix is the same thing as a kernel $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is **the finite set** $\mathcal{X} = [n] = \{1, \dots, n\}$. So if $\mathcal{X}$ is an **infinite set**, morally a kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a "**square matrix of infinite dimension**." If $\mathcal{X}$ is a real vector space (of positive dimension), then $\mathcal{X}$ is uncountably infinite; ideally this should seem rather weird at first; of course, as we get used to kernels on $\mathcal{X} = \mathbb{R}^n$, we will see that kernels naturally arise when, for example, we solve **linear** ODEs (ordinary differential equations) or **linear** PDEs (partial differential equation); **however, these kernel functions use the linear structure of functions** $\mathcal{X} \to \mathbb{R}$ where $\mathcal{X} = \mathbb{R}^n$. Ultimately these functions often satisfy $k(x, x') = f(x - x')$, which reflects the **group theoretic structure of** $\mathbb{R}^n$**, not its linear structure**.

In applications we will sometime want kernels to be positive definite, not merely positive semidefinite.

Remark 3.34. If $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite, then we must have $k(\mathbf{x}, \mathbf{x}) > 0$; if k is positive semidefinite, then we must have $k(\mathbf{x}, \mathbf{x}) \geq 0$, but equality can occur. (For example if k is identically 0, then k is positive semidefinite.)

Example 3.35. Let $\mathcal{X} = \mathbb{R}^2$, and let $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^6$ be given as in (2). For $x, x' \in \mathbb{R}^2$, let

$$k(x, x') = \Phi(x) \cdot \Phi(x').$$

For any $x \in \mathcal{X} = \mathbb{R}^2$ we have

$$k(x, x) = \big(x \cdot x + 1\big)^2 > 0.$$

However, we will soon prove that k is not positive definite.

WARNING 3.36. In the next subsection we will prove that if $\mathcal{X}$ is any set, and $\Phi\colon \mathcal{X} \to \mathbb{R}^n$ is any (set theoretic) map, then

$$k(x, x') = \Phi(x) \cdot \Phi(x')$$

is positive semidefinite, but cannot be positive definite unless $|\mathcal{X}| \leq n$, where $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$ (i.e., we aren't using anything about the linear structure of $\mathcal{X}$,

even if $\mathcal{X}$ is a vector space). Hence this is **never** the case if $\mathcal{X}$ is an infinite set; for this reason we will want to work in infinite dimensional inner product spaces; it is most convenient to work in the special case where the space is closed, i.e., a *Hilbert space*, although this is not strictly necessary, depending on what assumptions you want to make on the kernel function, k.

WARNING 3.37. In mathematics, one usually writes $k\colon X \times X \to \mathbb{R}$ and $k(x, y)$, instead of $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k(x, x')$ that is more common in ML (machine learning).

3.9. **Kernels on Finite Sets.** There is a standard theorem in linear algebra that if K is an $n \times n$ matrix, then:

(1) K is positive semidefinite matrix iff it can be written as $K = M^{\mathrm{T}}M$ for some matrix M; and
(2) K is positive definite matrix iff it can be written as $K = M^{\mathrm{T}}M$ for some matrix M whose kernel is trivial, i.e.,

$$\ker(M) \stackrel{\text{def}}{=} \{\mathbf{v} \mid M\mathbf{v} = \mathbf{0}\}$$

equals $\{0\}$ (it follows that M must be $m \times n$ where $m \geq n$, but one can, moreover, find an M with $K = M^{\mathrm{T}}M$ that is $n \times n$).

The "if" part is easy to prove; the "only if" will be proven later on in these notes. This therefore is a theorem about kernel functions $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} = [n]$ (or any finite set).

Proposition 3.38. *Let $\mathcal{X} = [n]$ be a finite set, and let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function that is positive semidefinite (respectively, definite). The there is a function $\Phi\colon \mathcal{X} \to \mathbb{R}^f$ such that k is the kernel function associated to Φ. More precisely, if $K \in \mathbb{R}^{n\times n}$ is the matrix associated to k, and if $K = M^{\mathrm{T}}M$ for some M with $\operatorname{rank}(K) = \operatorname{rank}(M) = r$, then we may take $f = r$.*

Proof. We have $K = M^{\mathrm{T}}M$ with $\operatorname{rank}(M) = r = \operatorname{rank}(K)$. Then it suffices to take $\Phi(j)$ be the j-th column of M (see (14) for an example that illustrates the idea). The details are an exercise left to the reader, namely Exercise B.5. □

3.10. **More Examples of Kernel Functions.** If $\mathcal{X}$ is a finite set of size n, then kernel functions $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are essentially the same as $n \times n$ matrices. Hence we have the following simple observation.

Proposition 3.39. *Let $p(x)$ be a polynomial or a convergent power series that takes $[0, \infty)$ to itself (respectively, to $(0, \infty)$). Then if K is a positive semidefinite matrix, then is $p(K)$ positive semidefinite (respectively, positive definite).*

Proof. By the spectral theorem (which we prove in the next subsection), K has an orthonormal basis, and hence $KU = U\Lambda$ where U is an orthogonal matrix, and Λ a diagonal matrix. Hence $K = U\Lambda U^{-1}$, and therefore $p(K) = Up(\Lambda)U^{-1}$. □

Example 3.40. Consider $p(x) = e^{xt}$ for some $t \in \mathbb{R}$. Then if K is postive semi-definite, then $p(K) = e^{Kt}$ is positive definite.

Theorem 3.41. *Let $k, \tilde{k}$ be two positive semidefinite (respectively definite) kernel function $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then so is their pointwise product, $k\tilde{k}$, i.e.,*

$$(k\tilde{k})(x, x') = k(x, x')\tilde{k}(x, x').$$

Proof. It suffices to check this in the case where $\mathcal{X}$ is finite, and then to use Definition 3.21.

If $\mathcal{X} = [n]$, then we know that $k, \tilde{k}$ are the kernels associated to, respectively, $\Phi\colon \mathcal{X} \to \mathbb{R}^f$ and $\tilde{\Phi}\colon \mathcal{X} \to \mathbb{R}^{\tilde{f}}$ (by Proposition 3.38).

[A subtlety in this proof is that $f, \tilde{f}$ above can be as large as $|\mathcal{X}| = n$, so $f, \tilde{f}$ depend on the size of the finite set $\mathcal{X}$; hence for an infinite $\mathcal{X}$, there is no one value of $f, \tilde{f}$ that works on all finite subsets of $\mathcal{X}$. We will return to this subtlety later.]

Set $\Phi \otimes \tilde{\Phi}\colon \mathcal{X} \to \mathbb{R}^f \otimes \mathbb{R}^{f'} \simeq \mathbb{R}^{f\tilde{f}}$ to be the map given by

$$(\Phi \otimes \tilde{\Phi})(x) = \Phi(x) \otimes \tilde{\Phi}(x).$$

Using Exercise B.6, part (a), we see that

$$k(x,x')\tilde{k}(x,x') = \big(\Phi(x)\cdot\Phi(x')\big)\big(\tilde{\Phi}(x)\cdot\tilde{\Phi}(x')\big) = (\Phi\otimes\tilde{\Phi})(x)\cdot(\Phi\otimes\tilde{\Phi})(x'),$$

and so kk' is the kernel function associated to $\Phi \otimes \tilde{\Phi}$. □

More information on kernels can be found in the Exercises sections (WHICH SPECIFICALLY?).

3.11. **A Positive Definite Kernel Function on $\mathbb{R}$: PDE's in $\mathbb{R}^n$ for $n = 1$, i.e., ODE's, for the Reluctant Reader.** In this subsection we give a kernel function on $\mathbb{R}$ that is positive definite. This comes from the Laplacian there, i.e., the map $w(x) \mapsto w''(x)$.

3.11.1. *Ancient Aspects.* Many aspects of the spectral decomposition of the one-dimensional Laplacian were well known to the ancient Greeks, and are easy to demonstrate on any string instrument, such a guitar, violin, or a lyre (if you happen to have one lying around). For example, if you pluck a guitar string near its bridge, you will hear a "twangy" sound that represents hearing the higher harmonics of the *Laplacian*: this means that you are modeling the guitar string by the real interval $[0, L] \subset \mathbb{R}$ for some $L > 0$, and you are interested in the ODE (with various physical constants suppressed, unimportant for mathematical intuition):

$$w''(x) + \lambda w(x) = 0 \quad \forall x \in (0, L)$$

subject to the boundary condition

$$w(0) = w(L) = 0$$

which is called the *Dirichlet boundary conditions*. Of course, the solution to this ODE gives the sine wave eigenfunctions:

$$w_m(x) = \sin(2\pi m x), \quad \lambda_m = (2\pi m)^2, \quad m \in \mathbb{N},$$

where the w_m are the "harmonics" and $\sqrt{\lambda_m}$ represents the "frequencies" when solving the wave equation $u_{xx} = u_{tt}$.

3.11.2. *The Poisson Equation in $\mathbb{R}^n$, $n = 1$.* Our positive definite kernel function $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that we will construct will be an example of a *Green's function* in the (very simple case) of solving the Poisson equation on the interval $[0, L]$ subject to Dirichlet conditions. In other words, we are given a function $f\colon [0, L] \to \mathbb{R}$, and we want to solve the equation

$$\forall x \in (0, L), \quad w''(x) = f(x), \qquad \text{s.t.} \quad w(0) = w(L) = 0.$$

We claim that it suffices to solve this in the case where $f(x) = \delta_y(x)$, where $\delta_y(x)$ is the *Dirac delta function* at $x = y$.

The only problem is that the Dirac delta function is not a function in the classical sense, but is a *generalized function*; however, the Dirac delta function provides crucial information to mathematicians since — according to ChatGPT and Google Gemini (although they slightly differ on precise dates, at least in 2025), the 1820's (e.g., Fourier, Poisson, Cauchy, and Gauss), some 100 years before Dirac's influential work. The notion of a *generalized function* is foundational to many parts of analysis, and goes back to ...

TO BE CONTINUED

3.11.3. *Some Examples.* Let $f(x) = 6x$. Then integrating twice we get

$$w''(x) = x^3 + C_0 + C_1 x.$$

Hence the solution to the Dirichlet

TO BE CONTINUED

3.11.4. *What is the Dirac delta function?* In class we gave the usual intuitive description of the Dirac delta function: it is not a function in the classical sense, but a "generalized function" such that for any function $f(x)$ we have

$$\int f(t)\delta_y(t)\,dt = f(y).$$

So $\delta_y(x)$ is a function that is zero "away from $x = y$," but has

$$\int \delta_y(t)\,dt = 1.$$

This notion of a "generalized function" can be made rigorous, and then $\delta_y(x)$ is a very important function: any other real valued function f can be written as the integral

3.11.5. *The ReLU, Heaviside, and Dirac Delta Functions.* Consider the function:

$$\mathrm{ReLU}(x) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } x \leq 0, \text{ and} \\ x & \text{if } x \geq 0. \end{cases}$$

This is called the *ReLU function* (rectified linear unit) in ML (machine learning); it is also encounted in options trading. This function is the heart of solution of the Poisson equation for the following reason: one can say that the derivative of $\mathrm{ReLU}(x)$ is the classical Heaviside function:

$$\mathrm{Heaviside}(x) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } x \leq 0, \text{ and} \\ 1 & \text{if } x \geq 0. \end{cases}$$

The derivative of the Heaviside function does not exist in the usual sense, but we are going to see that one can work with *generalized functions*, and in this setting the derivative of the Heaviside function does exist and is just $\delta_0(x)$, called a *Dirac delta function*.

TO BE CONTINUED

3.11.6. *The Differential Equations $w' = f(x)$ and $w'' = f(x)$.* The usual ordinary differential equation $w' = f(x, w)$, where w' means dw/dx, has received tons of attention, for many reasons. However, the differential equation

$$w' = \frac{dw}{dx} = f(x)$$

does not recieve much attention: the reasons is that it has a simple solution, namely this means that w is the integral of $f(x)$, typically written as

$$w(x) = \int f(x)\,dx + C$$

understanding the integral here is indefinite. Similarly the equation

$$w'' = f(x)$$

has a general solution

$$w(x) = \int \left(\int f(x)\,dx \right) + C_1 x + C_0,$$

i.e., w is $f(x)$ integrated twice, and the solution is only ambiguous up to a linear term $C_1 x + C_0$; in other words, if

$$w'' = f(x) \quad \text{and} \quad \tilde{w}'' = f(x),$$

then $(w - \tilde{w})'' = 0$, and therefore $w - \tilde{w} = C_1 x + C_0$.

3.11.7. *Abstractions Aside...* Ultimately, we want to convince you that the solution to

$$\forall x \in (0, L), \quad w''(x) = f(x), \qquad \text{s.t.} \quad w(0) = w(L) = 0,$$

is given by the unique function

$$w(x) = \int f(y) G(x, y)\,dy,$$

and that for each fixed y,

$$G(x, y) = \mathrm{ReLU}(x - y) + C_1(y)x + C_0(y),$$

where $C_1(y), C_0(y)$ are the unique constants (with y fixed) making $G(x, y)$ satisfy

$$G(0, y) = G(L, y) = 0.$$

TO BE CONTINUED

3.12. **History of Kernels.** Toward the end of this course we will explore the version for an infinite set, $\mathcal{X}$; when Aronsjan organized the study of kernel functions [**?**, **?**], he pointed out that the abstract theorem was proven by Moore [**?**] and named a *reproducing kernel in Hilbert space* (RKHS these days), although is tied to (and perhaps implicit in) Mercer's work [**?**].

[To be fair, there was a flurry of work on kernel functions around 1900 and into the early-mid 1900's; and some of this work had never been widely known — or perhaps mostly forgotten — by the time of Aronszajn's work [**?**], likely due to the importance of Hilbert-Schmidt kernels, Bergman kernels, etc., that likely eclipsed other work. Aronszajn [**?**] organized many aspects of work on kernels, and popularized many aspects of their study (including giving a historical account of a lot of work to date at the time).]

To be continued.

4. Symmetric Matrices and Rayleigh Quotients

In this section we give the basic theory of symmetric matrices.

4.1. **Symmetric Matrices and Rayleigh Quotients: Basic Theorems.** Let $A \in \mathbb{R}^n$, i.e., A is an $n \times n$ matrix with real entries. We say that A is *symmetric* if $A = A^{\mathrm{T}}$, where A^{T} is the transpose of A. We use the notation

$$\mathcal{R}_A(\mathbf{v}) = \frac{(A\mathbf{v}) \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}} = \frac{\mathbf{v}^{\mathrm{T}} A \mathbf{v}}{\mathbf{v}^{\mathrm{T}} \mathbf{v}}$$

which we call the *Raleigh quotient of* A, which we view as a real-valued function on $\mathbb{R}^n \setminus \{0\}$; clearly $\mathcal{R}_A$ is invariant under scaling.

Theorem 4.1. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric. Then A has real eigenvalues*

$$\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A),$$

with a real, orthonormal eigenbasis, i.e., $A\mathbf{v}_j = \lambda_j \mathbf{v}_j$ for $j \in [n]$, such that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are orthonormal (i.e., $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ if $i \neq j$, and $\mathbf{v}_i \cdot \mathbf{v}_i = 1$.

Proof. Let $\mathbb{S}^{n-1}$ be the set of unit vectors in $\mathbb{R}^n$. We easily see that $\|A\mathbf{v}\|$ is bounded over all $\mathbf{v} \in \mathbb{S}^{n-1}$, and hence $\mathcal{R}_A$ is bounded on $\mathbb{S}^{n-1}$ (and therefore over all of $\mathbb{R}^n \setminus \{0\}$). Let $\mathbf{v} \in \mathbb{S}^{n-1}$ be a vector at which $\mathcal{R}_A$ attains its maximum.

Let us prove that $A\mathbf{v} = \lambda \mathbf{v}$ for some $\lambda \in \mathbb{R}$. Let $\mathbf{u}$ be any vector orthogonal to $\mathbf{v}$; then we easily see that for small ϵ,

$$\big(A(\mathbf{v} + \epsilon \mathbf{u})\big) \cdot (\mathbf{v} + \epsilon \mathbf{u}) = (A\mathbf{v}) \cdot (\mathbf{v}) + 2\epsilon A\mathbf{v} \cdot \mathbf{u} + O(\epsilon^2)$$

(using $A^{\mathrm{T}} = A$) and

$$(\mathbf{v} + \epsilon \mathbf{u}) \cdot (\mathbf{v} + \epsilon \mathbf{u}) = \mathbf{v} \cdot \mathbf{v} + O(\epsilon^2) = 1 + O(\epsilon^2)$$

(using $\mathbf{v} \cdot \mathbf{v} = 1$ and $\mathbf{v} \cdot \mathbf{u} = 0$). It follows that

$$\mathcal{R}_A(\mathbf{v} + \epsilon \mathbf{u}) = \mathcal{R}_A(\mathbf{v}) + 2\epsilon (A\mathbf{v}) \cdot \mathbf{u} + O(\epsilon^2).$$

Since $\mathcal{R}_A$ is maximized at $\mathbf{v}$, it follows that

$$(A\mathbf{v}) \cdot \mathbf{u} = 0.$$

Since $\mathbf{u}$ was an arbitrary vector orthogonal to $\mathbf{v}$, it follows that $A\mathbf{v}$ is orthogonal to each vector orthogonal to $\mathbf{v}$, and hence $A\mathbf{v} = \lambda \mathbf{v}$ for some $\lambda \in \mathbb{R}$.

Let $\lambda_1 = \lambda$ and $\mathbf{v}_1 = \mathbf{v}$.

Next consider $\mathcal{R}_A$ restricted to vertors orthogonal to $\mathbf{v}_1$, and say that this maximum, restricted to $\mathbb{S}^{n-1}$, is attained at $\mathbf{v}_2$. By considering

$$\mathcal{R}_A(\mathbf{v}_2 + \epsilon \mathbf{u})$$

over all $\mathbf{u}$ that are orthogonal to both $\mathbf{v}_2, \mathbf{v}_1$ we similarly show that $(A\mathbf{v}_2)\mathbf{u} = 0$. Hence $A\mathbf{v}_2$ is a multiple of $\mathbf{v}_1$ and $\mathbf{v}_2$; since

$$(A\mathbf{v}_2) \cdot \mathbf{v}_1 = \mathbf{v}_2 \cdot (A\mathbf{v}_1) = \lambda \mathbf{v}_2 \cdot \mathbf{v}_1 = 0,$$

we have that $A\mathbf{v}_2$ is orthogonal to $\mathbf{v}_1$; since $\mathbf{v}_1, \mathbf{v}_2$ are orthogonal, we have $A\mathbf{v}_2 = \lambda_2 \mathbf{v}_2$ for some λ_2.

Next note that for $j = 1, 2$ we have

$$\mathcal{R}_A(\mathbf{v}_i) = \frac{(A\mathbf{v}_i) \cdot \mathbf{v}_i}{\mathbf{v}_i \cdot \mathbf{v}_i} = \frac{\lambda_i \mathbf{v}_i \cdot \mathbf{v}_i}{\mathbf{v}_i \cdot \mathbf{v}_i} = \lambda_i.$$

It follows that $\lambda_1 \ge \lambda_2$, since λ_1 is the maximum of $\mathcal{R}_A$ over all $\mathbb{R}^n \setminus \{0\}$.

Similarly, for any $j = 3, 4, \dots, n$, we inductively find $\mathbf{v}_j$ as the maximum of $\mathcal{R}_A$ restricted to vectors orthogonal to $\mathbf{v}_1, \dots, \mathbf{v}_{j-1}$, and show that $A\mathbf{v}_j = \lambda_j \mathbf{v}_j$ for some $\lambda_j \in \mathbb{R}$, and that $\lambda_j \le \lambda_{j-1}$. □

Note that the last paragraph of this proof can be used to prove the following more general (and sometimes useful) fact.

Lemma 4.2. *Let A be a symmetric $n \times n$ matrix, and $\mathbf{v}_1, \dots, \mathbf{v}_{j-1} \in \mathbb{R}^n$ be mutually orthogonal vectors with $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for $i \in [j-1]$. Then if $\mathbf{v}_j$ is a vector on which $\mathcal{R}_A$ takes its maximum over all vectors orthogonal to $\mathbf{v}_1, \dots, \mathbf{v}_{j-1}$, then $A\mathbf{v}_j = \lambda_j \mathbf{v}_j$ for some $\lambda_j \in \mathbb{R}$ (and $\mathcal{R}_A(\mathbf{v}_j) = \lambda_j$).*

Appendix A. Exercises on Kernel Methods

A.1. Cows and Goldfish/Ghosts/etc.

Exercise A.1. Let $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^6$ be given by (2), i.e.,

$$\Phi(x, y) = (1, \sqrt{2}x, \sqrt{2}y, x^2, \sqrt{2}xy, y^2). \tag{17}$$

Recall that we have argued by symmetry that

$$\int_{\theta=0}^{\theta=2\pi} \Phi(r\cos\theta, r\sin\theta)\frac{d\theta}{2\pi} = (0, 0, 0, r^2/2, 0, r^2/2). \tag{18}$$

Finish the calculation in class that shows that for any $x, y, \rho \in \mathbb{R}$ with $\rho \ge 0$ such that $x^2 + y^2 = \rho^2$, we have

$$f(r) = \|(0, 0, 0, r^2/2, 0, r^2/2) - \Phi(x, y)\|^2$$

attains its minimum value at $r^2 = \rho^2$.

Exercise A.2. By a *PDF (probability density function)* on $[0, 2\pi]$ we mean any function $p\colon [0, 2\pi] \to \mathbb{R}_{\ge 0}$ such that p is piecewise continuous (this condition can be weakened) and

$$\int_{\theta=0}^{\theta=2\pi} p(\theta)\, d\theta = 1.$$

Let $\Psi\colon [0, 2\pi] \to \mathbb{R}^m$ be any continuous map for some m, i.e.,

$$\Psi(\theta) = \big(\psi_1(\theta), \dots, \psi_m(\theta)\big)$$

where each ψ_i is a continuous map $[0, 2\pi] \to \mathbb{R}$. For any such Ψ, and any PDF, p, the *p-expected value of* Ψ refers to

$$\mathbb{E}_p[\Psi] \stackrel{\text{def}}{=} \int_0^{2\pi} \Psi(\theta)\, p(\theta) d\theta,$$

i.e., the vector

$$(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^m, \quad \text{where} \quad \alpha_i = \int_0^{2\pi} \psi_i(\theta) p(\theta)\, d\theta.$$

A.2(a) The *uniform density function* is the PDF on $[0, 2\pi]$ given by $p_{\text{unif}}(\theta) = 1/(2\pi)$. By a direct integration of trigonometric functions, show that for Φ as in (17) we have

$$\mathbb{E}_{p_{\text{unif}}}\big[\Phi(r\cos\theta, r\sin\theta)\big] = (1, 0, 0, r^2/2, 0, r^2/2)$$

(we proved this in class using various symmetry arguments).

A.2(b) Let

$$p(\theta) = \begin{cases} 1/(3\pi) & \text{if } 0 \le \theta \le \pi, \text{ and} \\ 2/(3\pi) & \text{if } \pi \le \theta \le 2\pi, \end{cases}$$

which we easily see is a PDF. Compute

$$\mathbb{E}_p\big[\Phi(r\cos\theta, r\sin\theta)\big].$$

[Hence $\mathbb{E}_p$ of anything generally depends on p.]

A.2(c) Say that p is a PDF where $p(\pi - \theta) = p(\theta)$ for all $\theta \in \mathbb{R}$.[11] [12] Which components of

$$\mathbb{E}_p\big[\Phi(r\cos\theta, r\sin\theta)\big]$$

necessarily vanish? Explain. [Hint: you can reduce this expected value to an integral from $-\pi/2 \le \theta \le \pi/2$ (or $[0,\pi/2] \cup [3\pi/2, 2\pi]$ or something similar), or you can think of the symmetry argument we used in class to establish (18).]

A.2(d) Say that p is a PDF where for all θ,

$$p(\pi - \theta) = p(\theta) = p(-\theta) = p(\pi + \theta).$$

Show that

$$\mathbb{E}_p\big[\Phi(r\cos\theta, r\sin\theta)\big] = (1, 0, 0, r_1^2, 0, r_2^2),$$

where r_1, r_2 are reals satisfying $r_1^2 + r_2^2 = r^2$. [Hint: you can reduce this expected value to an integral from $0 \le \theta \le \pi/2$, or you can think of the symmetry argument we used in class to establish (18).]

A.2(e) Show that in (d), if we additionally assume that $p(\theta) = p(\pi/2 - \theta)$ for all $\theta \in \mathbb{R}$, then $r_1^2 = r_2^2 = r^2$. [Hint: this symmetry means that for all (x, y) on the unit circle we have $\tilde{p}(x, y) = \tilde{p}(y, x)$, where $\tilde{p}$ is given by $\tilde{p}(r\cos\theta, r\sin\theta) = p(\theta)$. Even if you don't use this hint, it should provide some intuition.] [13]

Exercise A.3. Let $\mathcal{D} \subset \mathbb{R}^2$ be an open subset of finite area. If $\phi\colon \mathbb{R}^2 \to \mathbb{R}$ is a function, we define the $\mathcal{D}$-average of ϕ to be

$$\mathrm{Avg}_{\mathcal{D}}(\phi) = \frac{\int_{\mathcal{D}} \phi(x, y)\, dx\, dy}{\mathrm{Area}(\mathcal{D})},$$

assuming the above integral makes sense. Similarly, if $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^m$ for some m, where $\Phi = (\phi_1, \ldots, \phi_m)$ and we define

$$\mathrm{Avg}_{\mathcal{D}}(\Phi) = \big(\mathrm{Avg}_{\mathcal{D}}(\phi_1), \ldots, \mathrm{Avg}_{\mathcal{D}}(\phi_m)\big).$$

For any $\epsilon > 0$, let[14]

$$\mathcal{D}_\epsilon = \big\{(x, y) \;\big|\; 1 \le x^2 + y^2 \le (1 + \epsilon)^2\big\}.$$

---

[11] Here we are viewing θ as making sense for all $\theta \in \mathbb{R}$ via the usual convention θ and $\theta + 2\pi$ refer to the same angle. Hence θ means the same thing as $\theta + 2\pi k$ for any $k \in \mathbb{Z}$. Hence θ can viewed as specifying an element $\mathbb{R}/(2\pi)\mathbb{Z}$.

[12] If the reader wishes to stick to θ between 0 and 2π, then we have $p(\pi - \theta) = p(\theta)$ for $\theta \in [0, \pi]$ and then we have $p(3\pi - \theta) = p(\theta)$ for $\theta \in [\pi, 2\pi]$.

[13] Corrected in 2025 by Z.J. and (later... again) by T.L.

[14] Two corrections here made in 2025 by T.L.

A.3(a) Show that

$$1/2 \le \mathrm{Avg}_{\mathcal{D}_\epsilon}(x^2) \le (1+\epsilon)^2/2$$

using (18) and the fact that $dx\,dy = r\,dr\,d\theta$.

A.3(b) Fix an $r \in \mathbb{R}_{>0}$, and let

$$\mathcal{D}_\epsilon = \big\{(x,y) \;\big|\; r^2 \le x^2+y^2 \le r^2+\epsilon\big\}.$$

Reasoning similarly, show that

$$\lim_{\epsilon\to 0} \mathrm{Avg}_{\mathcal{D}_\epsilon}(x^2) = r^2/2,$$

and similarly compute all of

$$\lim_{\epsilon\to 0} \mathrm{Avg}_{\mathcal{D}_\epsilon}\big(\Phi(x,y)\big),$$

with Φ given by (17), i.e.,

$$\Phi(x,y) = (1, \sqrt{2}x, \sqrt{2}y, x^2, \sqrt{2}xy, y^2).$$

Exercise A.4. Let $F(x,y) = (x/a)^2 + (y/b)^2$ for some fixed $a,b \in \mathbb{R}_{>0}$. Hence $F(x,y) = 1$ describes the ellipse

$$\big\{(x,y) \in \mathbb{R}^2 \;\big|\; x = a\cos\theta,\ y = b\sin\theta,\ \text{for some } \theta \in [0,2\pi]\big\}.$$

For real $\epsilon > 0$, let

$$\mathcal{D}_\epsilon = \big\{(x,y) \;\big|\; 1 \le F(x,y) \le 1+\epsilon\big\}. \tag{19}$$

With Φ as in (17), show that

$$\lim_{\epsilon\to 0} \mathrm{Avg}_{\mathcal{D}_\epsilon}\big(\Phi(x,y)\big) = \big(1,0,0,a^2/2,0,b^2/2\big). \tag{20}$$

[Hint: consider the transformation $\tilde{x} = ax$ and $\tilde{y} = by$, and use the fact that $d\tilde{x}\,d\tilde{y}$ is a fixed constant times $dx\,dy$ (namely $d\tilde{x}\,d\tilde{y} = ab\,dx\,dy$).]

Exercise A.5. Consider Exercise A.5, but say that we chose a different function $F(x,y)$ such that $F(x,y) = 1$ iff (x,y) lies on the ellipse $(x/a)^2 + (y/b)^2 = 1$. Say that we define $\mathcal{D}_\epsilon$ as in (19). Does

$$\lim_{\epsilon\to 0} \mathrm{Avg}_{\mathcal{D}_\epsilon}\big(\Phi(x,y)\big)$$

always equal $\big(1,0,0,a^2/2,0,b^2/2\big)$, or does this limit depend on the choice of F? [Hint: If it helps, you could first consider the special case where $a = b = \rho$; ultimately, however, you want an argument that works for any positive a,b.] [Hint:[15] you may use the co-area formula if you know what this means (however you can do this computation by hand without this formula).]

Exercise A.6. Let $(x^{\mathrm{test}}, y^{\mathrm{test}}) \in \mathbb{R}^2$ be fixed. Consider the function

$$g(a,b) \stackrel{\mathrm{def}}{=} \big\|\big(1,0,0,a^2/2,0,b^2/2\big) - \Phi(x^{\mathrm{test}}, y^{\mathrm{test}})\big\|^2 \tag{21}$$

with Φ (as usual) as in (17). (To understand why we are interested in g, consider (20).)

A.6(a) For which values of $a,b \in \mathbb{R}$ does $g(a,b)$ attain its minimum value? Show that these are when $a^2 = 2(x^{\mathrm{test}})^2$ and $b^2 = 2(y^{\mathrm{test}})^2$. [Note that it follows that $(x^{\mathrm{test}}, y^{\mathrm{test}})$ lies on the ellipse $x^2/a^2 + y^2/b^2 = 1$; of course, $(x^{\mathrm{test}}, y^{\mathrm{test}})$ lies on many ellipses centred at the origin.]

[15] We thank ?? in the 2025 version of this course for making this remark.

A.6(b) Say that we insist that $a = b$. So for which value of a does $g(a, a)$ attain its minimum? [Hint: We've already solved this in Exercise A.1.]

A.6(c) Say that in part (a), $y^{\text{test}} = 0$. Give a very short argument – directly from (21) (and without using the result in part (a)), that for any $a, b \in \mathbb{R}$ with $b > 0$, $g(a, 0) < g(a, b)$.

Appendix B. Exercises in Linear Algebra

Exercise B.1. B.1(a) Show that the following vectors in $\mathbb{R}^n$, for $n \geq 5$, are mutually orthogonal:

$$(1, 1, \ldots, 1),\ (-1, 1, 0, \ldots, 0),\ (-1, -1, 2, 0, \ldots, 0),\ (-1, -1, -1, 3, 0, \ldots, 0),$$
$$(-1, -1, -1, -1, 4, 0, \ldots, 0).$$

B.1(b) Say that you assign a linear algebra class to find an orthonormal eigenbasis for the matrix:

$$K = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

with the instruction that when K has a multiple eigenvalue, λ, then vectors are to be found using the standard method of solving the linear system $(K - \lambda I)\mathbf{v} = \mathbf{0}$ (with "fixed" and "free" variables, or however you call them), and then using Gram-Schmidt to turn your eigenvectors into an orthonormal system. What are the first 4 orthonormal eigenvectors for K with eigenvalue $\lambda = 0$ that your students will produce?

[Rhetorical question: will you then teach them about complex numbers and circulant matrices?]

Exercise B.2. Show that if $\zeta_1 \zeta_2$ are two distinct n-th roots of unity (i.e., $\zeta_1^n = \zeta_2^n = 1$ and $\zeta_1 \neq \zeta_2$), then the vectors:

$$(1, \zeta_1, \zeta_1^2, \ldots, \zeta_1^{n-1}), (1, \zeta_2, \zeta_2^2, \ldots, \zeta_2^{n-1})$$

are orthogonal in (either of the two) usual complex dot products. [Hint: $\overline{\zeta} = \zeta^{-1}$ for any complex number, ζ, such that $|\zeta| = 1$.]

Exercise B.3. * In class we briefly mentioned that the tensor product of two $\mathbb{R}$-vector spaces is an *initial object* in a certain "category" of biliear forms. This is important to understand conceptually: in particular, if U, W are $\mathbb{R}$-linear vector spaces, then it makes no sense to write $U \otimes V$; you have to understand that this is an initial object in a category, and there are many possible conventions to write this as an $\mathbb{R}$-linear vector space.[16] (In class we gave the usual convention, which has many good properties.)

B.3(a) Recall that if U, W are $\mathbb{R}$-linear vector spaces, then a map

$$\phi \colon U \times W \to Z$$

is *bilinear* if for all $u_1, u_2 \in U$, $w_1, w_2 \in W$, and reals $\alpha_1, \alpha_2, \beta_1, \beta_2$ we have

$$\phi(\alpha_1 u_1 + \alpha_2 u_2, \beta_1 w_1 + \beta_2 w_2)$$
$$= \alpha_1\beta_1\phi(u_1, w_1) + \alpha_1\beta_2\phi(u_1, w_2) + \alpha_2\beta_1\phi(u_2, w_1) + \alpha_2\beta_2\phi(u_2, w_2).$$

[16] This question inspired by office hours with anonymousA.

Show that any such ϕ, and any $w \in W$, the map $u \mapsto \phi(u,w)$ is linear map.

B.3(b) Let $\{u_1,\dots,u_m\}$ be a basis for U (hence U is finite dimensional) and similarly $\{w_1,w_2,\dots,w_n\}$ be a basis for W. Let Z be any vector space. Show that if $\{z_{ij}\}_{i\in[m],j\in[n]}$ is any set of vectors in Z, then there is a unique bilinear map $\phi\colon U\times W\to Z$ such that

$$\phi(u_i,w_j)=z_{ij}.$$

B.3(c) Let $\phi_1\colon U\times W\to Z_1$ and $\phi_2\colon U\times W\to Z_2$ be two bilinear forms. A *morphism from ϕ_1 to ϕ_2* is a linear map

$$f\colon Z_1\to Z_2$$

such that $\phi_2 = f\circ\phi_1$.

(i) Show that if

$$\bigl\{\phi_1(u_i,w_j)\bigr\}_{i\in[m],j\in[n]}$$

is any set mn linearly indepedent vectors in Z (hence $\dim(Z)\ge mn$), then for any ϕ_2 there is at least one morphism from $\phi_1\to\phi_2$.

(ii) Show that if, in addition, $\dim(Z)=mn$, then there is a unique morphism from ϕ_1 to ϕ_2. In this case we say that ϕ_1 is an *initial object (in the category of bilinear forms)*.

B.3(d) Show that if ϕ_1,ϕ_2 are any initial objects as above, then there is a unique morphism from ϕ_1 to ϕ_2. Hence the initial objects are "unique up to unique isomorphism."

B.3(e) Find a "terminal object" in this category, meaning a bilinear map ϕ such for each other bilinear form ϕ_1, there is a unique morphism $\phi_1\to\phi$. Is the terminal object unique?

Exercise B.4. Exercise 3.7 in Supplemental Notes and Homework, https://www.cs.ubc.ca/~jf/courses/531F.S2021/homework.pdf from CPSC 531F (2021) https://www.cs.ubc.ca/~jf/courses/531F.S2021/index.html . [This exercise is about the Fibonacci graph and its adjacency matrix (and its powers).]

Exercise B.5. In this exercise we are proving what is in fancy schmancy language can be called the *representability theorem* for kernels over a finite set. This is just Proposition 3.38.

B.5(a) Say that $K\in\mathbb{R}^{n\times n}$ and $M\in\mathbb{R}^{f\times n}$ satisfy $K=M^{\mathrm{T}}M$. Let $k\colon[n]\times[n]\to\mathbb{R}$ be the associated kernel function (i.e., $k(i,j)$ is the (i,j)-th entry of K), and let

$$\Phi\colon[n]\to\mathbb{R}^f$$

by the function where $\Phi(i)$ is the i-th column of K. Then show that

$$k(i,j)=\Phi(i)\cdot\Phi(j).$$

B.5(b) Assume the following result: if $K\in\mathbb{R}^{n\times n}$ is any symmetric, positve semi-definite matrix of rank $r=\mathrm{rank}(K)$, then there is an $M\in\mathbb{R}^{r\times n}$ such that $K=M^{\mathrm{T}}M$. Given this, prove Proposition 3.38.

B.5(c) In part (b), is it possible for there to exist an $N\in\mathbb{R}^{f'\times n}$ where $f'<r$ but $M=N^{\mathrm{T}}N$? Explain briefly.

Exercise B.6. Recall that if S, T are finite sets, then the *tensor product* of two vectors $\mathbf{u} \in \mathbb{R}^S$ and $\mathbf{v} \in \mathbb{R}^T$ is the vector in $\mathbb{R}^S \otimes \mathbb{R}^T$, which, identifying[17] $\mathbb{R}^S \otimes \mathbb{R}^T$ with $\mathbb{R}^{S \times T}$, is the vector $\mathbf{u} \otimes \mathbf{v} \in \mathbb{R}^S \otimes \mathbb{R}^T$, whose (s,t) component is $u(s)v(t)$. [This is essentially the tensor product of (7), but here we work more canonically, i.e., not needlessly ordering the elements of S and T.]

B.6(a) Show that if $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^S$, $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^T$, then

$$(\mathbf{u} \cdot \mathbf{u}')(\mathbf{v} \cdot \mathbf{v}') = (\mathbf{u} \otimes \mathbf{v})(\mathbf{u}' \otimes \mathbf{v}').$$

B.6(b) If $\mathcal{S} \colon \mathbb{R}^{S_1} \to \mathbb{R}^{S_2}$ and $\mathcal{T} \colon \mathbb{R}^{T_1} \to \mathbb{R}^{T_2}$ are linear maps, then there is a unique linear map $\mathcal{W} \colon \mathbb{R}^{S_1 \times T_1} \to \mathbb{R}^{S_2 \times T_2}$ that satisfies

$$\forall \mathbf{u} \in \mathbb{R}^{S_1},\ \mathbf{v} \in \mathbb{R}^{T_1}, \quad \mathcal{W}(\mathbf{u} \otimes \mathbf{v}) = (\mathcal{S}\mathbf{u}) \otimes (\mathcal{T}\mathbf{v}),$$

and this linear map $\mathcal{W}$, and that $A \in \mathbb{R}^{S_1 \times S_2}$ is the matrix representing $\mathcal{S}$ (in the standard way, using the standard bases for $\mathbb{R}^{S_1}$ and $\mathbb{R}^{S_2}$), and similarly for $B \in \mathbb{R}^{T_1 \times T_2}$, then $\mathcal{W}$ coincides with $A \otimes B$ as described in class (in 2025) and touched upon briefly in Subsection 3.3. [One writes $\mathcal{W} = \mathcal{S} \otimes \mathcal{T}$, and $\mathcal{W}$ is called the *tensor product* or *Kronecker product* of $\mathcal{S}$ and $\mathcal{T}$.] [Hint: look at the class 2025 notes, and show that the tensor product of matrices, as we defined it there (and visualized it in terms of block matrices) is an example of such a map $\mathcal{W}$. Now you have to prove the uniqueness of $\mathcal{W}$.]

Appendix C. Exercises in Spectral Graph Theory

Exercise C.1. The point of this exercise is to prove Proposition 3.19. So let $G = (V, E)$ be a simple graph that is d-regular.

C.1(a) Show that $\lambda_1(A_G) \leq d$. We suggest the following approach: assume that $A_G \mathbf{u} = \lambda \mathbf{u}$ with $\mathbf{u} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ and $\lambda > d$. By possibly replacing $\mathbf{u}$ with $-\mathbf{u}$, we may assume $\mathbf{u}$ has at least one positive component; let the maximum component of $\mathbf{u}$ be $u(v)$ for some $v \in V_G$ (v is not necessarily unique). Argue that it is impossible to have

$$\lambda u(v) = \sum_{v' \sim v} u(v').$$

[Hint: if you like, you can assume after scaling that $u(v) = 1$ and hence $u(v') \leq 1$ for all $v' \in V_G$.]

C.1(b) Using the same approach, show that if $A_G \mathbf{u} = d\mathbf{u}$ and $\mathbf{u}$ has a positive component, and if the maximum component of $\mathbf{u}$ is attained at $v \in V_G$, then for all $v' \sim v$ we have $u(v') = u(v)$.

C.1(c) Deduce from (c) that if G is connected, then the only eigenvectors with eigenvalue d are a multiple of $\mathbf{1}$ (the all 1's vector).

C.1(d) Deduce that the multiplicity of d as an eigenvalue is the number of connected components of G.

C.1(e) Similarly deduce that if $\lambda_n \geq -d$.

[17] If $\{e_s\}_{s \in S}$ and $\{e_t\}_{t \in T}$ are the standard bases of $\mathbb{R}^S$ and $\mathbb{R}^T$ respectively, then we may identify $e_s \otimes e_t$ with $e_{(s,t)} \in \mathbb{R}^{S \times T}$, which sets up the bijectiion. One can do similarly with any finite dimensional vector spaces, U, V, and any chosen bases for U and V, and the identification is easily seen to be independent of the choses bases for U and V; see Exercise **??** (perhaps this will be added in 2025, perhaps later).

C.1(f) Similarly deduce that if $A_G\mathbf{u} = -d\mathbf{u}$ and $\mathbf{u}$ has a positive component, and if the maximum component of $\mathbf{u}$ is attained at $v \in V_G$, then for all $v' \sim v$ we have $u(v') = -u(v)$.

C.1(g) Similarly deduce that if G is connected and $-d$ is an eigenvalue, then G is bipartite.

C.1(h) Similarly deduce that the multiplicity of $-d$ as an eigenvalue is the number of connected components of G that are bipartite.

[Note: this type of approach is essentially called a "maximum principle" in ODE's and PDE's: you consider the maximum value of a function, and use it to deduce interesting conclusions and/or contradictions.]

Appendix K. Possible Exercises

These exercises MAY BE assigned in 2025.

Exercise K.1. The point of this exercise is to the many ways one can learn from ChatGPT or a similar generative AI tool. This is more my personal experiment with ChatGPT and ..., but you can probably generate something pretty similar. I'm using the $20.00(USD)/month verion. Type the following questions into ChatGPT or a similar LLM (large language model) AI (artifical intelligence algorithm). Type the following or some reasonable versions thereof:

(1)

Appendix L. Exercises That Will Not Be Assigned in 2025

Exercise L.1. Add something here.

Appendix Z. Glossary of Some ML (Machine Learning) Terminology

This is a glossary of some machine learning terminology, translated into mathematics and, at times, into English. [I am making this glossary as much for me as for the reader.]

activation function: The function of the inputs that a particular node in an ANN (artificial neural network) outputs to the nodes in the next layer; e.g., a function $f\colon \mathbb{R}^n \to \mathbb{R}$ given by $f(x_1, \dots, x_n) = g(w_1x_1 + \cdots + w_nx_n + b)$, where g is a fixed function (e.g., ReLu, "sigmoid" (e.g., logistic, tanh)) for each node (often the same for each layer or for the entire network), and where $w_1, \dots, w_n \in \mathbb{R}$ — the *weights* — and $b \in \mathbb{R}$ — the *bias* — are parameters that vary from node to node. The particular function g is often chosen to make it feasible to compute a good set of parameters for each node (so that the network computes the desired output over various inputs).

AI: artificial intelligence; not a precise term. This is an umbrella term for a set of computer algorithms that supposedly has a sort of "artificial intelligence" that tries to mimic human intelligence.

ANN: artificial neural network; a fairly precise term. A network meant to simulate a real life brain. Usually the network is a set of nodes (or vertices), typically arranged in layers, where the first layer consists of "input" to the ANN, and the last layer is the "output" of the ANN. The nodes of the network are meant to model real life neurons in a brain which, roughly speaking, have many inputs (dendrites) and a single output (axon) that "fires" an electrical impulse (action potential) when its inputs reach a certain

"electrical threshold." In ANNs, the activation function is rarely a threshold function, since these functions would be too hard to tune (i.e., determine good parameter settings for the threshold functions).

AS: artificial stupidity. A ridiculing and/or cynical term for an AI and/or ML algorithm that performs poorly.

binary classifier: The output is $\{0, 1\}$.

linear separator: A hyperplane (codim 1 affine subspace) in $\mathbb{R}^n$ separating two datasets, e.g., one representing cows, the other goldfish.

kernel functions: Particular kernel functions (in the mathematical sense) — most often positive (semi)definite kernels — of interest to ML; i.e., maps $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (or $\mathbb{C}$) whose restriction to each finite subset of $\mathcal{X}$ yields a symmetric, positive (semi)definite matrix. This includes:

- **Exponential kernal:** $k(\mathbf{x}, \mathbf{y}) = e^{\beta \mathbf{x} \cdot \mathbf{y}}$ or a truncation of the power series.
- **Gaussian kernel:** the function $k(\mathbf{x}, \mathbf{y}) = e^{|\mathbf{x}-\mathbf{y}|/\sigma^2}$, based on the classical *fundamental solution* of the heat equation in $\mathbb{R}^n$, $k(\mathbf{x}, t) = (4\pi t)^{-n/2} e^{|\mathbf{x}|^2/(4t)}$.
- **Polynomial Kernel:** the (degree d polynomial kernel) function $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + c)^d$ for $c \in \mathbb{R}_{>0}$ and $d \in \mathbb{N}$.

logistic function (or ODE): (sometimes called *the* sigmoid function). Solutions to $y' = y(1 - y)$ (restricted to $0 < y = y(x) < 1$); aside from $y' = y$, this is one of the simplest ODE's in modeling: for $y > 0$ small, this behaves like $y' = y$; and y' approaches 0 as y approaches 1. Explicitly: $y(x) = 1/(1 + Ce^{-x})$, although one often translates and scales this function and/or the ODE (this ODE is not linear).

ML: machine learning; not a precise term. One typically thinks of ANNs (artificial neural networks), which are built as general purpose algorithms with nodes arranged in layers, each node having a number of parameters that are "learned" or "optimized" through "training data." Then point is the ANN used should be "general purpose," rather than an algorithm specific to the task at hand. Hence a computer chess playing program that is designed by consulting experts in chess and using their knowledge would not be considered an ML algorithm, but rather a broader class of algorithms known as AI (artificial intelligence).

MNIST dataset: an early and influential dataset used to design and test ML algorithms, specifically ANNs. This consists of a some 70,000 number of handwritten images of the digits 0-9; 60,000 of the images are designated as "training images," used to set the parameters of the ANN (or other algorithm) — also described as the *learning phase* of the ANN or algorithm — and 10,000 designated as "testing images" to see if the algorithm with the parameters found do a good job correctly identifying a handwritten digit correctly. These days there are a large number of similarly well-known "benchmark" datasets used by ANN designers to compete with one another.

one-hot: a standard basis vector. Say that $S = \{\text{cat}, \text{dog}, \text{frog}\}$, which we identify with $[3] = \{1, 2, 3\}$, and say that we want to identify each picture as an element of S. Then the *one-hot* values are $(1, 0, 0), (0, 1, 0), (0, 0, 1) \in \mathbb{R}^3$, identified with $\mathbb{R}^S$.

ReLU: $\mathrm{ReLU}(x) = x^+ \stackrel{\text{def}}{=} \max(0, x)$, the *positive part of* x, i.e., the *Rectified Linear Unit*, used also in describing (financial) option payouts. Its derivative is the Heaviside function, and its second derivative is the Dirac delta function.

separator: see "linear separator." Could also refer to separators in graph theory. A "quadratic separator" may refer to a "linear separator" for the kernel map $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} = \mathbb{R}^n$ given by $k(x, y) = (x \cdot y + c)^2$.

sigmoid: this is both an umbrella term and a specific term: the specific term is the "logistic" function, i.e., solution to $y' = y(1 - y)$, often scaled (this ODE is non-linear) and translated; more generally, it refers to any "S-shaped" curve like the "logistic" function (e.g., tanh).

softmax: $\mathrm{softmax}(x_1, \dots, x_n) \stackrel{\text{def}}{=} (e^{x_1}, \dots, e^{x_n}) / \sum_{i=1}^n e^{x_i}$. This is a smoothed version of the max function; of course, this can be adjusted by scaling the vector $(x_1, \dots, x_n)$; similar to the stochastic vector $(e^{-\beta E_1}, \dots, e^{-\beta E_n}) / \sum_{i=1}^n e^{-\beta E_i}$ in statistical mechanics ($\beta = -1$ is the softmax).

TensorFlow (Keras): one of a number of popular ways to play around with NN's. `tensorflow.keras` is a Python library.

References

[Fri15] Joel Friedman, *Sheaves on graphs, their homological invariants, and a proof of the Hanna Neumann conjecture: with an appendix by Warren Dicks*, Mem. Amer. Math. Soc. **233** (2015), no. 1100, xii+106, With an appendix by Warren Dicks. MR 3289057

Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, CANADA.

Email address: jf@cs.ubc.ca

URL: http://www.cs.ubc.ca/~jf