CS 760: Machine Learning Spring 2022

Topic 2: Review of Linear Algebra

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN

<u>DO NOT POLLUTE!</u> AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

2.1 Fundamental Concepts

Linear algebra lies at the heart of machine learning, studying linear equations, mappings, and their representations in feature spaces. One of the most elemental vector manipulations are linear combinations, which essentially comprise scaling and addition.

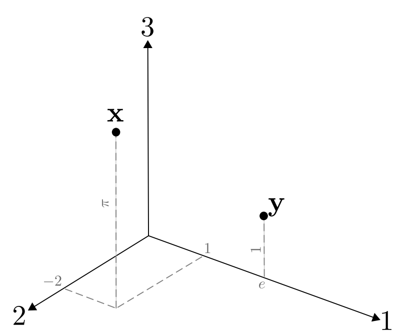
Definition 2.1 (Linear combination, coefficients)**.** A vector $\mathbf{z}$ is a *linear combination* of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_R\}$ if it can be written as

$$\mathbf{z} \;=\; \sum_{r=1}^{R} c_r \mathbf{x}_r \tag{2.1}$$

for some $c_1, \ldots, c_R \in \mathbb{R}$. The scalars $\{c_1, \ldots, c_R\}$ are called the *coefficients* of $\mathbf{z}$ with respect to (w.r.t.) $\{\mathbf{x}_1, \ldots, \mathbf{x}_R\}$.

Example 2.1. Consider the following vectors in $\mathbb{R}^3$:

$$\mathbf{x} \;=\; \begin{bmatrix} 1 \\ -2 \\ \pi \end{bmatrix}, \qquad\qquad \mathbf{y} \;=\; \begin{bmatrix} e \\ 0 \\ 1 \end{bmatrix}.$$



Then

$$\mathbf{z} \;=\; -3\mathbf{x} + 2\mathbf{y} \;=\; -3\begin{bmatrix}1\\-2\\\pi\end{bmatrix} + 2\begin{bmatrix}e\\0\\1\end{bmatrix} \;=\; \begin{bmatrix}-3\\6\\-3\pi\end{bmatrix} + \begin{bmatrix}2e\\0\\2\end{bmatrix} \;=\; \begin{bmatrix}-3+2e\\6\\-3\pi+2\end{bmatrix}$$

is a linear combination of $\mathbf{x}$ and $\mathbf{y}$, with coefficients -3 and 2.

Another fundamental concept is linear independence.

Definition 2.2 (Linear independence)**.** A set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_\mathrm{R}\}$ is *linearly independent* if

$$\sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{c}_\mathrm{r}\mathbf{x}_\mathrm{r} \;=\; \mathbf{0}$$

implies $\mathrm{c}_\mathrm{r} = 0$ for every $\mathrm{r} = 1, \ldots, \mathrm{R}$. Otherwise we say it is *linearly dependent*.

Intuitively, a set of vectors is linearly independent if none of them can be written as a linear combination of the others.

Example 2.2. The following vectors are linearly independent:

$$\mathbf{x}_1 \;=\; \begin{bmatrix}1\\0\\0\end{bmatrix}, \qquad \mathbf{x}_2 \;=\; \begin{bmatrix}0\\1\\0\end{bmatrix}, \qquad \mathbf{x}_3 \;=\; \begin{bmatrix}0\\0\\1\end{bmatrix},$$

because we cannot write either of them as a linear combination of the others. In contrast, the following vectors are linearly dependent:

$$\mathbf{y}_1 \;=\; \begin{bmatrix}5\\5\\2\end{bmatrix}, \qquad \mathbf{y}_2 \;=\; \begin{bmatrix}1\\1\\0\end{bmatrix}, \qquad \mathbf{y}_3 \;=\; \begin{bmatrix}0\\0\\1\end{bmatrix},$$

because we can write $\mathbf{y}_1$ as a linear combination of $\mathbf{y}_2$ and $\mathbf{y}_3$, namely, $\mathbf{y}_1 = 5\mathbf{y}_2 + 2\mathbf{y}_3$.

2.2 Matrices

Matrices are very handy structures to arrange and manipulate vectors.

Example 2.3. We can arrange the vectors in the Example 2.2 in the following matrices:

$$\mathbf{X} \;=\; \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}, \qquad \mathbf{Y} \;=\; \begin{bmatrix}5&5&2\\1&1&0\\0&0&1\end{bmatrix}.$$

2.2.1 Review of Basic Matrix Operations

- **Matrix Multiplication.** Given matrices $\mathbf{A} \in \mathbb{R}^{m \times k}$ and $\mathbf{B} \in \mathbb{R}^{k \times n}$, their product, denoted by $\mathbf{AB}$, is another matrix of size $m \times n$ whose $(i, j)^{th}$ entry is given by:

$$[\mathbf{AB}]_{ij} \ = \ \sum_{\ell=1}^{k} \mathbf{A}_{i\ell}\mathbf{B}_{\ell j}.$$

  Intuitively, the $(i, j)^{th}$ entry of $\mathbf{AB}$ is given by the multiplication of the i^{th} row of $\mathbf{A}$ and the j^{th} column of $\mathbf{B}$. For example, if

$$\mathbf{A} \ = \ \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \qquad \mathbf{B} \ = \ \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}, \tag{2.2}$$

  Then:

$$\mathbf{AB} \ = \ \begin{bmatrix} (1 \cdot 1 + 2 \cdot 3 + 3 \cdot 5) & (1 \cdot 2 + 2 \cdot 4 + 3 \cdot 6) \\ (4 \cdot 1 + 5 \cdot 3 + 6 \cdot 5) & (4 \cdot 2 + 5 \cdot 4 + 6 \cdot 6) \end{bmatrix} \ = \ \begin{bmatrix} 22 & 28 \\ 49 & 64 \end{bmatrix}.$$

- **Scalar Multiplication** Given a scalar c and a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, their product, denoted by $c\mathbf{A}$, is an $m \times n$ matrix whose $(i, j)^{th}$ entry is given by c times the $(i, j)^{th}$ entry of $\mathbf{A}$. For example, with $\mathbf{A}$ as in (2.2), and $c = 7$,

$$c\mathbf{A} \ = \ 7\mathbf{A} \ = \ \begin{bmatrix} 7 & 14 & 21 \\ 28 & 35 & 42 \end{bmatrix},$$

- **Transposition.** The transpose of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, denoted by $\mathbf{A}^{\mathsf{T}}$, is an $n \times m$ matrix whose $(i, j)^{th}$ entry is given by the $(j, i)^{th}$ entry of $\mathbf{A}$. Intuitively, transposing a matrix is like flipping its rows and columns along the diagonal. For example, with $\mathbf{A}$ as in (2.2),

$$\mathbf{A}^{\mathsf{T}} \ = \ \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

- **Trace.** Given a squared matrix $\mathbf{X} \in \mathbb{R}^{m \times m}$, its trace, denoted by $\mathsf{tr}(\mathbf{X})$ is the sum of its diagonal entries:

$$\mathsf{tr}(\mathbf{X}) \ = \ \sum_{i=1}^{m} \mathbf{X}_{ii}.$$

  For example, with $\mathbf{X}$ as in Example 2.3, $\mathsf{tr}(\mathbf{X}) = 3$.

- **Identity Matrix.** The identity matrix of size $m \times m$, denoted by $\mathbf{I}_m$, is a squared matrix whose diagonal entries are all ones, and off-diagonal entries are all zeros. For example, the 3×3 identity matrix is

$$\mathbf{I}_3 \ = \ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

  Whenever there is no possible confusion about the size of the identity matrix, people often drop the m subindex, and simply denote it as $\mathbf{I}$.

- **Inverse.** Given a squared matrix $\mathbf{X} \in \mathbb{R}^{m \times m}$, its inverse, denoted by $\mathbf{X}^{-1}$ is a matrix such that $\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I}$. For example, with $\mathbf{X}$ as in Example 2.3, $\mathbf{X}^{-1} = \mathbf{X}$. As another example, consider

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}.$$

  Then

$$\mathbf{X}^{-1} = \begin{bmatrix} 3 & -3 & 1 \\ -2.5 & 4 & -1.5 \\ 0.5 & -1 & 0.5 \end{bmatrix}.$$

  You can verify that $\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I}$. Here is a special trick to invert 2×2 matrices:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

  Of course, this requires that $ad - bc \neq 0$.

- **Hadamard Product.** Given two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, their Hadamard product (aka point-wise product) is a matrix of size $m \times n$, denoted as $\mathbf{A} \odot \mathbf{B}$, whose $(i,j)^{th}$ entry is given by the product of the (i,j) entries of $\mathbf{A}$ and $\mathbf{B}$, i.e.,

$$[\mathbf{A} \odot \mathbf{B}]_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}.$$

  For example, with $\mathbf{A}, \mathbf{B}$ as in (2.2),

$$\mathbf{A} \odot \mathbf{B}^\mathsf{T} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \odot \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 6 & 15 \\ 8 & 20 & 36 \end{bmatrix}.$$

- **Vector Operations.** Notice that vectors are 1-column matrices, so all matrix operators that apply to non-squared matrices also apply to vectors. For instance, with the same setup as in Example 2.2,

$$\mathbf{x}_1^\mathsf{T}\mathbf{x}_2 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = 0.$$

  Notice that $\mathbf{x}_1^\mathsf{T}\mathbf{x}_2 \neq \mathbf{x}_1\mathbf{x}_2^\mathsf{T}$:

$$\mathbf{x}_1\mathbf{x}_2^\mathsf{T} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

- **Norms.** Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, its Frobenius norm is defined as:

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{A}_{ij}^2}.$$

  In the particular case of vectors, this is often called the ℓ_2-norm or Euclidean norm, and is denoted by $\|\mathbf{x}\|_2$, or simply $\|\mathbf{x}\|$. Norms are important because they essentially quantify the *size* of a matrix or vector. Just as there are several ways to quantify the size of a person (e.g., age, height, weight), there are also several ways to quantify the *size* of a matrix or vector, for which we can use different norms. Another example is the ℓ_1-norm, which for matrices is defined as:

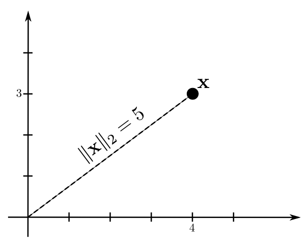
$$\|\mathbf{A}\|_1 \;:=\; \max_{\mathrm{j}} \sum_{\mathrm{i}=1} |\mathbf{A}_{\mathrm{ij}}|,$$

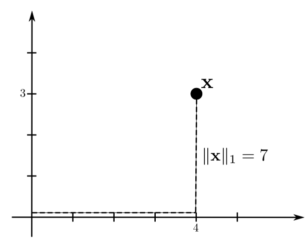
and for vectors is defined as

$$\|\mathbf{x}\|_1 \;:=\; \sum_{\mathrm{j}} |\mathrm{x}_{\mathrm{j}}|,$$

also known as the taxi-cab or Manhattan norm. Intuitively, the ℓ_2-norm measures the *point-to-point* distance, while the ℓ_1-norm measures the *taxi-cab* distance. For example, for the same vector $\mathbf{x} = [4 \;\; 3]^{\mathsf{T}}$, here are two different notions to quantify its size:

$$\|\mathbf{x}\|_2 \;=\; \sqrt{4^2 + 3^2} = 5 \qquad\qquad \|\mathbf{x}\|_1 \;=\; |4| + |3| = 7.$$





The norm $\|\mathbf{x}\|$ of a vector $\mathbf{x}$ is essentially its size. Norms are also useful because they allow us to measure distance between vectors (through their difference). For example, consider the following images:

and vectorize them to produce vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$. We want to do face clustering, i.e., we want to know which images correspond to the same person. If $\|\mathbf{x}-\mathbf{y}\|$ is small (i.e., $\mathbf{x}$ is similar to $\mathbf{y}$), it is reasonable to conclude that the first two images correspond to the same person. If $\|\mathbf{x} - \mathbf{z}\|$ is large (i.e., $\mathbf{x}$ is very different from $\mathbf{z}$), it is reasonable to conclude that the first and second images corresponds to different persons.

Norms satisfy the so-called *triangle inequality*: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. This allows you to draw useful conclusions. For example, knowing that $\|\mathbf{x}-\mathbf{y}\|$ is small and that $\|\mathbf{x}-\mathbf{z}\|$ is large allows us to conclude that $\|\mathbf{y} - \mathbf{z}\|$ is also large. Intuitively, this allows us to conclude that if $\mathbf{x}, \mathbf{y}$ correspond to the same person, and $\mathbf{x}$ and $\mathbf{z}$ corresponds to different persons, then $\mathbf{y}$ and $\mathbf{z}$ also correspond to different persons. In other words, nothing weird will happen.

Inner Product. Given two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}}$, their inner product is defined as:

$$\langle \mathbf{A}, \mathbf{B} \rangle \;:=\; \mathsf{tr}(\mathbf{A}^{\mathsf{T}}\mathbf{B}).$$

For example, with $\mathbf{A}$ as in (2.2),

$$\langle \mathbf{A}, \mathbf{A} \rangle \;=\; \mathsf{tr}(\mathbf{A}^\mathsf{T}\mathbf{A}) \;=\; \mathsf{tr}\left(\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \right) \;=\; \mathsf{tr}\left(\begin{bmatrix} 17 & 22 & 27 \\ 22 & 29 & 36 \\ 27 & 36 & 45 \end{bmatrix} \right) \;=\; 91.$$

Notice that for vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^\mathrm{D}$, $\mathbf{x}^\mathsf{T}\mathbf{y}$ will always be a scalar, so we can drop the trace, and simply write $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\mathsf{T}\mathbf{y}$. For example, with the same setup as in Example 2.2:

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle \;=\; \mathbf{x}_1^\mathsf{T}\mathbf{x}_2 \;=\; \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \;=\; 0,$$

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle \;=\; \mathbf{y}_1^\mathsf{T}\mathbf{y}_2 \;=\; \begin{bmatrix} 5 & 5 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \;=\; 10.$$

Inner products are of particular importance because they measure the similarity between matrices and vectors. In particular, recall from your kinder garden classes that the angle θ between two vectors is given by:

$$\cos\theta \;=\; \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}.$$

More generally, the larger the inner product between two objects (in absolute value), the more similar they are.

2.2.2 Why do I care about Matrices?

Matrix operators are useful because they allow us to write otherwise complex burdensome operations in simple and concise matrix form.

Example 2.4. With the same settings as in Examples 2.2 and 2.3, we can write linear combinations using a simple matrix multiplication. For instance, instead of writing

$$\mathbf{y}_1 \;=\; \begin{bmatrix} 5 \\ 5 \\ 2 \end{bmatrix} \;=\; 5 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 5 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \;=\; 5\mathbf{x}_1 + 5\mathbf{x}_2 + 2\mathbf{x}_3,$$

we can let $\mathbf{c} = [5 \;\; 5 \;\; 2]^\mathsf{T}$ be $\mathbf{y}_1$'s coefficient vector, and equivalently write in simpler form:

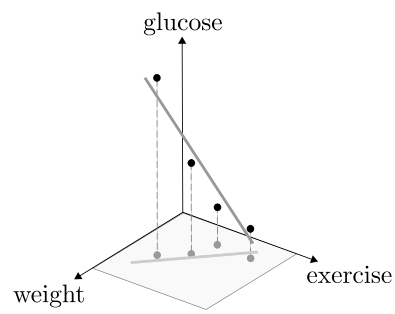
$$\mathbf{y}_1 \;=\; \begin{bmatrix} 5 \\ 5 \\ 2 \end{bmatrix} \;=\; \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 5 \\ 2 \end{bmatrix} \;=\; \mathbf{X}^\mathsf{T}\mathbf{c}.$$

Similar matrix expressions will be ubiquitous throughout this course, so you should start getting familiar and feeling comfortable using matrix operations.

2.3 Subspaces

Subspaces are essentially high-dimensional lines. A 1-dimensional subspace is a line, a 2-dimensional subspaces is a plane, and so on. Subspaces are useful because data often lies near subspaces.

Example 2.5. The following data lie near a 1-dimensional subspace (line):



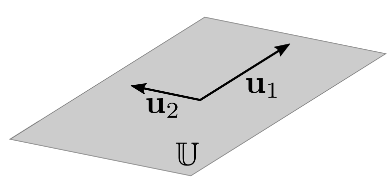
In higher dimensions subspaces may be harder to visualize, so you will have to use imagination to decide how a higher-dimensional subspace looks. Luckily, we have a precise and formal mathematical way to define them:

Definition 2.3 (Subspace)**.** A subset $\mathbb{U} \subseteq \mathbb{R}^{\mathrm{D}}$ is a *subspace* if for every $\mathrm{a}, \mathrm{b} \in \mathbb{R}$ and every $\mathbf{u}, \mathbf{v} \in \mathbb{U}$, $\mathrm{a}\mathbf{u} + \mathrm{b}\mathbf{v} \in \mathbb{U}$.

Definition 2.4 (Span)**.** $\mathrm{span}[\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}}]$ is the set of all linear combinations of $\{\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}}\}$. More formally,

$$\mathrm{span}[\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}}] \;:=\; \left\{ \mathbf{x} \in \mathbb{R}^{\mathrm{D}} \;:\; \mathbf{x} = \sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{c}_{\mathrm{r}} \mathbf{u}_{\mathrm{r}} \;\text{ for some } \mathrm{c}_1, \ldots, \mathrm{c}_{\mathrm{R}} \in \mathbb{R} \right\}.$$

For any vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}} \in \mathbb{R}^{\mathrm{D}}$, $\mathrm{span}[\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}}]$ is a subspace. For example, here is a subspace $\mathbb{U}$ (plane) spanned by two vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$:



Definition 2.5 (Basis)**.** A set of linearly independent vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}}\}$ is a *basis* of a subspace $\mathbb{U}$ if each $\mathbf{v} \in \mathbb{U}$ can be written as

$$\mathbf{v} \;=\; \sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{c}_{\mathrm{r}} \mathbf{u}_{\mathrm{r}}$$

for a *unique* set of coefficients $\{\mathrm{c}_1, \ldots, \mathrm{c}_{\mathrm{R}}\}$.

Definition 2.6 (Orthogonal)**.** A collection of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{R}\}$ is *orthogonal* if $\langle \mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j} \rangle = 0$ for every $\mathrm{i} \neq \mathrm{j}$.

If $\|\mathbf{x}\| = 1$, we say $\mathbf{x}$ is a *unit* vector, or that it is *normalized.* Similarly, a collection of normalized, orthogonal vectors is called *orthonormal.* There is a tight relation between inner products and norms. The following is one of the most important and useful inequalities that describe this relationship.

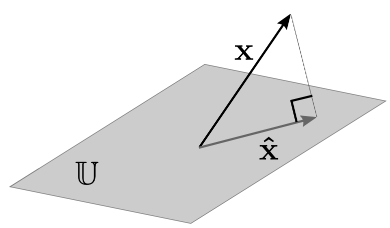
Proposition 2.1 (Cauchy-Schwartz inequality)**.** For every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^\mathrm{D}$,

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \;\leq\; \|\mathbf{x}\|\|\mathbf{y}\|.$$

Furthermore, if $\mathbf{y} \neq \mathbf{0}$, then equality holds if and only if $\mathbf{x} = \mathrm{a}\mathbf{y}$ for some $\mathrm{a} \in \mathbb{R}$.

2.4 Projections

In words, the projection $\hat{\mathbf{x}}$ of a vector $\mathbf{x}$ onto a subspace $\mathbb{U}$ is the vector in $\mathbb{U}$ that is closest to $\mathbf{x}$:



More formally,

Definition 2.7 (Projection)**.** The *projection* of $\mathbf{x} \in \mathbb{R}^\mathrm{D}$ onto subspace $\mathbb{U}$ is the vector $\hat{\mathbf{x}} \in \mathbb{U}$ satisfying

$$\|\mathbf{x} - \hat{\mathbf{x}}\| \;\leq\; \|\mathbf{x} - \mathbf{u}\| \qquad \text{for every } \mathbf{u} \in \mathbb{U}.$$

Notice that if $\mathbf{x} \in \mathbb{U}$, then $\hat{\mathbf{x}} = \mathbf{x}$. The following proposition tells us exactly how to compute projections.

Proposition 2.2. Let $\{\mathbf{u}_1, \dots, \mathbf{u}_\mathrm{R}\}$ be an orthonormal basis of $\mathbb{U}$. The projection of $\mathbf{x} \in \mathbb{R}^\mathrm{D}$ onto $\mathbb{U}$ is given by

$$\hat{\mathbf{x}} \;=\; \sum_{\mathrm{r}=1}^{\mathrm{R}} \langle \mathbf{x}, \mathbf{u}_\mathrm{r} \rangle \mathbf{u}_\mathrm{r}.$$

In other words, the coefficient of $\mathbf{x}$ w.r.t. $\mathbf{u}_\mathrm{r}$ is given by $\langle \mathbf{x}, \mathbf{u}_\mathrm{r} \rangle$.

Furthermore, the following proposition tells us that we can compute projections very efficiently: just using a simple matrix multiplication! This makes projections very attractive in practice. For example, as we saw before, data often lies near subspaces. We can measure how close using the norm of the *residual* $\mathbf{x} - \hat{\mathbf{x}}$.

Proposition 2.3 (Projector operator)**.** Let $\mathbf{U} \in \mathbb{R}^{\mathrm{D}\times\mathrm{R}}$ be a basis of $\mathbb{U}$. The *projection operator* $\mathbf{P}_{\mathbf{U}} : \mathbb{R}^{\mathrm{D}} \to \mathbb{U}$ that maps any vector $\mathbf{x} \in \mathbb{R}^{\mathrm{D}}$ to its projection $\hat{\mathbf{x}} \in \mathbb{U}$ is given by:

$$\mathbf{P}_{\mathbf{U}} \;=\; \mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}.$$

Notice that if $\mathbf{U}$ is orthonormal, then $\mathbf{P}_{\mathbf{U}} = \mathbf{U}\mathbf{U}^\mathsf{T}$.

Proof. Since $\hat{\mathbf{x}} \in \mathbb{U}$, that means we can write $\hat{\mathbf{x}}$ as $\mathbf{U}\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^{\mathrm{R}}$. We thus want to find the $\mathbf{c}$ that minimizes:

$$\|\mathbf{x} - \mathbf{U}\mathbf{c}\|_2^2 \;=\; (\mathbf{x} - \mathbf{U}\mathbf{c})^\mathsf{T}(\mathbf{x} - \mathbf{U}\mathbf{c}) \;=\; \mathbf{x}^\mathsf{T} - 2\mathbf{c}^\mathsf{T}\mathbf{U}^\mathsf{T}\mathbf{x} + \mathbf{c}^\mathsf{T}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{c}.$$

Since this is convex in $\mathbf{c}$, we can use elemental optimization to find the desired minimizer, i.e., we will take derivative w.r.t. $\mathbf{c}$, set to zero and solve for $\mathbf{c}$. To learn more about how to take derivatives w.r.t. vectors and matrices see *Old and new matrix algebra useful for statistics* by Thomas P. Minka. The derivative w.r.t. $\mathbf{c}$ is given by:

$$-2\mathbf{U}^\mathsf{T}\mathbf{x} + 2\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{c}.$$

Setting to zero and solving for $\mathbf{c}$ we obtain:

$$\hat{\mathbf{c}} \;:=\; \underset{\mathbf{c}\in\mathbb{R}^{\mathrm{R}}}{\arg\min}\; \|\mathbf{x} - \mathbf{U}\mathbf{c}\|_2^2 \;=\; (\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}\mathbf{x},$$

where we know $\mathbf{U}^\mathsf{T}\mathbf{U}$ is invertible because $\mathbf{U}$ is a basis by assumption, so its columns are linearly independent. It follows that

$$\hat{\mathbf{x}} \;=\; \mathbf{U}\hat{\mathbf{c}} \;=\; \underbrace{\mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}}_{\mathbf{P}_{\mathbf{U}}}\mathbf{x},$$

as claimed. If $\mathbf{U}$ is orthonormal, then $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}$, and hence $\mathbf{P}_{\mathbf{U}}$ simplifies to $\mathbf{U}\mathbf{U}^\mathsf{T}$. Notice that $\hat{\mathbf{c}}$ are the coefficients of $\hat{\mathbf{x}}$ w.r.t. the basis $\mathbf{U}$. □

2.5 Gram-Schmidt Orthogonalization

Orthonormal bases have very nice and useful properties. For example, in Proposition 2.3, if the basis $\mathbf{U}$ is orthonormal, then the projection operator is simplified into $\mathbf{U}\mathbf{U}^\mathsf{T}$, which requires much less computations than $\mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}$. The following procedure tells us exactly how to transform an arbitrary basis into an orthonormal basis.

Proposition 2.4 (Gram-Schmidt procedure)**.** Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_\mathrm{R}\}$ be a basis of $\mathbb{U}$. Let

$$\begin{aligned} \mathbf{v}'_\mathrm{r} &= \begin{cases} \mathbf{u}_1 & \mathrm{r} = 1, \\ \mathbf{u}_\mathrm{r} - \sum_{\mathrm{k}=1}^{\mathrm{r}-1} \langle \mathbf{u}_\mathrm{r}, \mathbf{v}_\mathrm{k} \rangle \mathbf{v}_\mathrm{k} & \mathrm{r} = 2, \ldots, \mathrm{R}, \end{cases} \\ \mathbf{v}_\mathrm{r} &= \mathbf{v}'_\mathrm{r} / \|\mathbf{v}'_\mathrm{r}\|. \end{aligned}$$

Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_\mathrm{R}\}$ are an orthonormal basis of $\mathbb{U}$.

Proof. We know from Proposition 2.2 that $\sum_{\mathrm{k}=1}^{\mathrm{r}-1} \langle \mathbf{u}_\mathrm{r}, \mathbf{v}_\mathrm{k} \rangle \mathbf{v}_\mathrm{k}$ is the projection of $\mathbf{u}_\mathrm{r}$ onto span$[\mathbf{v}_1, \ldots, \mathbf{v}_{\mathrm{r}-1}]$. This implies $\mathbf{v}'_\mathrm{r}$ is the orthogonal residual of $\mathbf{u}_\mathrm{r}$ onto span$[\mathbf{v}_1, \ldots, \mathbf{v}_{\mathrm{r}-1}]$, and hence it is orthogonal to $\{\mathbf{v}_1, \ldots, \mathbf{v}_{\mathrm{r}-1}\}$, as desired. $\mathbf{v}_\mathrm{r}$ is simply the normalized version of $\mathbf{v}'_\mathrm{r}$. □

2.6 Conclusions

These notes review several fundamental concepts of linear algebra that lie at the heart of machine learning, and will be crucial in *every* topic to be studied in this course.