Linear Algebra Refresher

For clarity we use bold faced letters to denote (column) vectors and upper case letters to denote matrices.

Linear Spaces

A (typically infinite) set of vectors S forms a vector space if $\forall \mathbf{u}, \mathbf{w} \in S$ and $s \in \mathbb{R}$ the vector $a\mathbf{u} + \mathbf{v}$ is in S. Note that by setting $a = -1, \mathbf{u} = \mathbf{v}$ we get that $\mathbf{0} \in S$ and thus $a\mathbf{v} \in S$.

A set of vectors $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^k$ is linearly dependent if there exists a vector $\mathbf{s} \in \mathbb{R}^k$ such that $\sum_i a_i \mathbf{v}^i = \mathbf{0}$. Otherwise, the set is called linearly independent. We can pack the set of vectors $\mathbf{v}^i$ into a matrix V whose i^{th} row is the transpose of $\mathbf{v}^i$. Linear dependence thus means $\mathsf{V}\mathbf{a} = \mathbf{0}$.

Prove to yourself the following two properties:

- If S is linearly dependent and $S \subset T$ then T is linearly dependent.
- If S is linearly independent and $T \subset S$ then T is linearly independent.

The span of a set S is a vector space $\Omega = \text{span}(S)$. It consists of all vectors that can be expressed as linear combinations of the vectors in S. A linearly independent set S is a *basis* for a vector space Ω if $S \subset \Omega$ and $\text{span}(S) = \Omega$. Prove to yourself the following claim:

If S, T are linearly independent sets of a vector space Ω and S is a basis for Ω then $|T| \leq |S|$. Therefore, all possible bases of Ω have the same cardinality. We can thus equate the dimension of Ω with the cardinality of its basis regardless of the (ambient) dimension of $\mathbf{0} \in \Omega$.

Matrices and Vector Spaces

For a matrix $\mathsf{A} \in \mathbb{R}^{m \times n}$ let us denote by $\text{col}_i(\mathsf{A})$ its i^{th} column. The column space of A is $\text{span}(S)$ where $S = \{\text{col}_i(\mathsf{A})\}_{i=1}^n$. Analogously, the row space of A is the column space of $\mathsf{A}^\top$ where $A^\top_{ij} = A_{ji}$. The column (row) rank of a matrix is k if the basis for is column (row) space has cardinality k. For a matrix $\mathsf{A} \in \mathbb{R}^{m \times n}$ its rank $k = \text{rank}(A)$ satisfies $k \leq \min(m, n)$.

The inner (dot) product of two vectors $\mathbf{u}, \mathbf{v}$ is $\mathbf{u} \cdot \mathbf{v} = \sum_i u_i v_i$. Recall that for two matrices $\mathsf{A} \in \mathbb{R}^{m \times r}$ and $\mathsf{B} \in \mathbb{R}^{r \times n}$, their product is a matrix $\mathsf{C} = \mathsf{AB}$ where entries of C are $\mathsf{C}_{ij} = \text{col}_i(\mathsf{A}^\top) \cdot \text{col}_j(\mathsf{B})$. Matrix multiplication is associative, $\mathsf{A}(\mathsf{BC}) = (\mathsf{AB})\mathsf{C}$, distributive $\mathsf{A}(\mathsf{B} + \mathsf{C}) = \mathsf{AB} + \mathsf{AC}$, but *not* commutative $\mathsf{AB} \neq \mathsf{BA}$ for general square matrices $\mathsf{A}, \mathsf{B} \in \mathbb{R}^{n \times n}$.

A square matrix $\mathsf{A} \in \mathbb{R}^{n \times n}$ is *invertible* if $\exists \mathsf{B} \in \mathbb{R}^{n \times n}$ such that $AB = BA = I_n$ where

$$\mathsf{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & & & 1 \end{pmatrix}$$

A matrix is called singular if it does *not* have an inverse. The inverse of a matrix A, denoted A^{-1} is *unique* and $\mathsf{A}^{-1}\mathsf{A} = \mathsf{A}\mathsf{A}^{-1} = \mathsf{I}_n$. Let A, B be non-singular matrices of the same size, then using associativity we get

$$(\mathsf{B}^{-1}\mathsf{A}^{-1})(\mathsf{A}\mathsf{B}) = \mathsf{B}^{-1}(\mathsf{A}^{-1}\mathsf{A})\mathsf{B} = \mathsf{B}^{-1}\mathsf{I}_n\mathsf{B} = \mathsf{I} \quad \Rightarrow \quad (\mathsf{AB})^{-1} = \mathsf{B}^{-1}\mathsf{A}^{-1} .$$

It it less immediate (try to reprove) that a $k \times k$ matrix A is non-singular *iff* $\text{rank}(\mathsf{A}) = k$.

Linear Systems of Equations

Suppose we are given a matrix $\mathsf{X} \in \mathbb{R}^{n \times d}$ and a vector $\mathbf{y} \in \mathbb{R}^n$. The matrix X is called the data or measurements matrix and the vector $\mathbf{y}$ consists of targets. Each row of X is called an example where the i^{th} example is denoted $\mathbf{x}^i$. The number of rows n of X is thus the number of examples and each example dimension is d. Many learning tasks are concerned with finding a vector $\mathbf{w} \in \mathbb{R}^d$ such that $\mathsf{X}\mathbf{w} \approx \mathbf{y}$. We leave the semantic of the approximation to the machine learning course and recap properties of the problem when we seek an exact solution to the system $\mathsf{X}\mathbf{w} = \mathbf{y}$.

The system $\mathsf{X}\mathbf{w} = \mathbf{y}$ has no solution iff $\mathbf{y}$ is not in the column space of X. Moreover, if the system has at least one solution, and $\text{rank}(\mathsf{X}) < d$, then it has infinitely many solutions. A rectangular matrix $\mathsf{X} \in \mathbb{R}^{n \times d}$ with $n \leq d$ is of full rank if $\text{rank}(\mathsf{X}) = n$. If X is of full rank, then the system $\mathsf{X}\mathbf{w} = \mathbf{y}$.

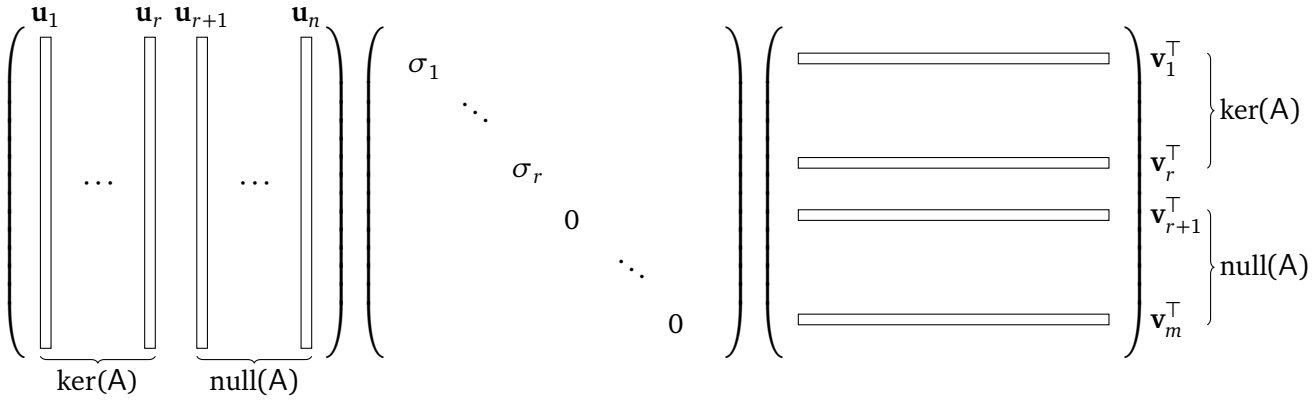
To find a solution for the system of equations we can use a sequence of elementary row operations such as the Gauss-Jordan Elimination. This method applies elementary row operations to the matrix X and the free vector $\mathbf{y}$ and reduces the problem to a simple form from which the solution is immediately obtained. Regression problems in statistical machine learning are concerned with more complex settings where there is no solution to the system or we need to find a solution vector of certain properties when from infinitely many solutions.

Singular Value Decomposition

Singular Value Decomposition (SVD) is a major analysis tool in statistical machine learning and data science. We provide here only a brief illustrative description and revisit SVD in the appropriate context during the workshop. Any real-valued matrix $\mathsf{A} \in \mathbb{R}^{\times n}$ can be represented as the product of three special, albeit not unique matrices, $\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{V}^\top$. The matrices U, V are orthonormal and D is a $m \times n$ matrix with only main diagonal non-zero elements, $D_{i,i} \geq D_{i+1,i+1}$ for $r \leq \min(m, n)$,

$$\mathsf{U}^\top\mathsf{U} = \mathsf{I}_\mathsf{m} \qquad \mathsf{V}^\top\mathsf{V} = \mathsf{I}_\mathsf{n} \qquad \forall i \neq j : \mathsf{D}_{\mathsf{i},\mathsf{j}} = 0 \; .$$

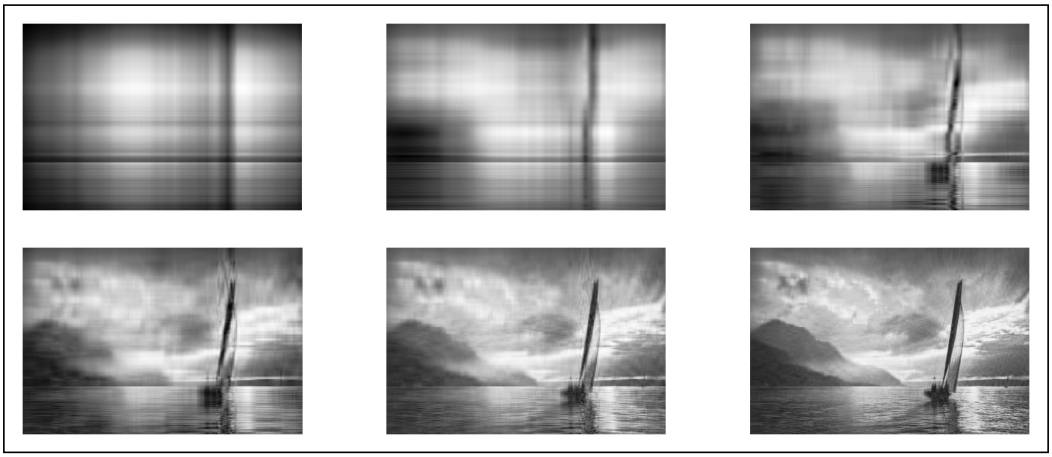
The singular values, the non-zero diagonal elements of D, are typically denoted as $\sigma_i = D_{i,i}$ and satisfy $\sigma_1 \geq \sigma_2 \ldots \sigma_r > 0$. Here is an illustration of SVD:



The first r columns of U form a basis for the column vectors of A. The span of these r vectors is called the kernel of A. The rest of the vectors $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are orthogonal to any vector in the span of the columns of A, meaning $A^\top \mathbf{u}_j = 0$ for $j > r$. The span of these vectors is called the null-space of A. An analogous representation holds for the rows of A using the matrix V.

An alternative view is to say that $\mathsf{A} = \sum_{i=1}^{r} \sigma_r \mathbf{u}_i \mathbf{v}_i^\top$ where $\forall i \neq j : \mathbf{u}_i \cdot \mathbf{u}_j = 0$ and $\mathbf{v}_i \cdot \mathbf{v}_j = 0$.

SVD is a very useful tool for low-rank approximation of matrices, replacing $\sigma_{r'+1}, \ldots, \sigma_{r'}$ where with zeros and keeping only $r' \ll r$ columns of U, V we can obtain a compact representation of A. This representation can be used for compression, denoising, and learning, see notebook `svdemo` which illustrates using SVD for compression.



Square Symmetric Matrices

Square symmetric matrices play an important role in numerous scientific domains. In machine learning we encounter such matrices in second order analysis of learning functions. Of a particular interest is the spectrum, namely singular values, of square symmetric matrices. While for general matrices singular value decomposition requires two different matrices, for square symmetric matrices a single orthonormal matrix suffices,

$$\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{U}^\top \quad \text{where} \quad \mathsf{U}^\top\mathsf{U} = \mathsf{I}_n \ .$$

However, the singular values when using $\mathsf{V} = \mathsf{U}$ are general real numbers. The singular value decomposition of these matrices is closely related to eigen-value decomposition, colloquially often called principle component analysis (PCA). Concretely, the columns of U are the eigenvectors of A thus for a column vector $\mathbf{v} = \mathsf{U}_{\star,i}$ of U we have that $\mathsf{A}\mathbf{v} = \sigma_i \mathbf{v}$.

When $D_{i,i} = \sigma_i \geq 0$ for all i we say that the matrix is positive semi-definite (PSD) and designated it by writing $\mathsf{A} \succcurlyeq 0$. PSD matrices play an important role in distance learning for example. Now, let $\mathsf{M} \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Then, the following properties are all equivalent:

- M is PSD: $M \succcurlyeq 0$.
- For any $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}^\top \mathsf{M}\, \mathbf{v} \geq 0$.
- There exists a matrix A such that $\mathsf{M} = \mathsf{A}^\top \mathsf{A}$.
- All eigenvalues of M are non-negative.

These equivalences are almost immediate to prove given the singular value decomposition theorem.