EE 227A: Convex Optimization and Applications **January 19**

Lecture 2: **Linear Algebra Review**

Lecturer: Mert Pilanci

Reading assignment: Appendix C of BV. Sections 2-6 of the web textbook[1]

2.1 Vectors

2.1.1 Basics

Independence. A set of vectors $x_i \in \mathbf{R}^n$, $i = 1, \ldots, m$ is said to be *independent* if and only if the following condition on a vector $\lambda \in \mathbf{R}^m$:

$$\sum_{i=1}^{m} \lambda_i x_i = 0$$

implies $\lambda = 0$. This means that no vector in the set can be expressed as a linear combination of the others.

Subspace, span. A subspace of $\mathbf{R}^n$ is a subset that is closed under addition and scalar multiplication. As an example, the *span* of a set of vectors $x_i \in \mathbf{R}^n$, $i = 1, \ldots, m$ is defined as

$$\text{span}(x_1, \ldots, x_m) := \left\{ \sum_{i=1}^{m} \lambda_i x_i \ : \ \lambda \in \mathbf{R}^m \right\}.$$

Basis. A basis of $\mathbf{R}^n$ is a set of n independent vectors. The basis of a given subspace $\mathcal{L} \subseteq \mathbf{R}^n$ is any independent set of vectors whose span is $\mathcal{L}$. The number of vectors in the basis is actually independent of the choice of the basis (for example, in $\mathbf{R}^3$ you need two independent vectors to describe a plane containing the origin). This number is called the *dimension* of $\mathcal{L}$.

2.1.2 Scalar product and norms

Scalar product. The scalar product (or, dot product) between two vectors $x, y \in \mathbf{R}^n$ is defined as the scalar $x^T y = \sum_{i=1}^{n} x_i y_i$. More generally, an inner product on $\mathbf{R}^n$ is a bilinear function $\langle \cdot, \cdot \rangle$ that satisfies the properties of symmetry (with respect to a swap in the two

[1] At `https://inst.eecs.berkeley.edu/~ee127a/book/login/index.html`, with login and password both set to `ee127a-web`.

arguments), and positive-definiteness (that is, $\langle x, x\rangle$ is always non-negative, and zero only when $x = 0$). An example of an inner product is the weighted dot product

$$\langle x, y\rangle_\sigma := \sum_{i=1}^{m} \sigma_i^2 x_i y_i, \tag{2.1}$$

where $\sigma \in \mathbf{R}^n$, $\sigma \neq 0$ is given.

Vector norms. A function $f : \mathbf{R}^n \to \mathbf{R}$ is a norm on $\mathbf{R}^n$ if the following three conditions are met:

1. f is convex.
2. f is positively homogeneous, meaning that $f(\alpha x) = \alpha f(x)$ for every $x \in \mathbf{R}^n$ and $\alpha \in \mathbf{R}_+$.
3. f is positive-definite: for every $x \in \mathbf{R}^n$, $f(x) = 0$ implies $x = 0$.

Together, the first two conditions are equivalent to the *triangle inequality*:

$$\forall\, x, y \in \mathbf{R}^n \;:\; f(x + y) \leq f(x) + f(y).$$

Often, norms are denoted $\|\cdot\|$.

Popular norms. There are three very popular norms for a vector $x \in \mathbf{R}^n$:

- The Euclidean norm is $\|x\|_2 := \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{x^T x}$, which corresponds to the usual notion of distance in two or three dimensions.
- The l_1-norm, or Manhattan distance, is $\|x\|_1 = \sum_{i=1}^n |x_i|$. The norm corresponds to the distance travelled on a rectangular grid (such as Manhattan) to go from one point to another.
- The l_∞-norm is given by $\|x\|_\infty := \max_{1\leq i\leq n} |x_i|$.

The l_p-norm is a class that includes the previous ones (in an asymptotic sense in the case of the l_∞ norm), and is defined as

$$\|x\|_p := \left(\sum_{i=1}^{p} |x_i|^p\right)^{1/p},$$

where $p \geq 1$.

There are many other norms that are important or interesting in some applications. For example, for $k \in \{1, \ldots, n\}$ we can define

$$\|x\|_{1,k} := \sum_{i=1}^{k} |x|_{[i]}$$

where for every i, $|x|_{[i]}$ is the i-th largest absolute value of elements of x. The norm is a kind of mixture between the l_1- and l_∞-norms, respectively obtained upon setting $k = n$ and $k = 1$.

Finally, any scalar product $\langle \cdot, \cdot \rangle$ generates a norm, defined as $\|x\| := \sqrt{\langle x, x \rangle}$. For example, the Euclidean norm is generated by the ordinary scalar product. Another example is the norm induced by the inner product defined in (2.1), which is the weighted Euclidean norm

$$\|x\| = \sqrt{\sum_{i=1}^{n} \sigma_i^2 x_i^2}.$$

Cauchy-Schwartz inequalities, angles, dual norm. The Cauchy-Schwartz inequality states that

$$\forall\, x, y \in \mathbf{R}^n \ : \ x^T y \leq \|x\|_2 \cdot \|y\|_2.$$

When none of the vectors involved is zero, we can define the corresponding angle as θ such that

$$\cos\theta = \frac{x^T y}{\|x\|_2 \|y\|_2}.$$

(The notion generalizes the usual notion of angle between two directions in two dimensions.)

Cauchy-Schwartz inequalities can be obtained for norms other than the Euclidean. For example,

$$\forall\, x, y \in \mathbf{R}^n \ : \ x^T y \leq \|x\|_\infty \cdot \|y\|_1.$$

More generally, to any norm $\|\cdot\|$ we can associate a *dual norm*, usually denoted $\|\cdot\|_*$, and defined as

$$\|y\|_* := \max_x \ x^T y \ : \ \|x\| \leq 1.$$

(Check this is indeed a norm.) By construction, the norm $\|\cdot\|$ and its dual satisfy the (generalized) Cauchy-Schwartz inequality

$$\forall\, x, y \in \mathbf{R}^n \ : \ x^T y \leq \|x\| \cdot \|y\|_*.$$

In this setting, the Euclidean norm is its own dual; and the l_1- and l_∞-norms are dual of each other.

Orthogonal basis. A basis $(u_i)_{i=1}^n$ is said to be *orthogonal* if $u_i^T u_j = 0$ if $i \neq j$. If in addition, $\|u_i\|_2 = 1$, we say that the basis is *orthonormal.*

2.2 Matrices

2.2.1 Basics

Matrices (in say, $\mathbf{R}^{m \times n}$) can be viewed simply as a collection of vectors of same size. Alternatively, a matrix can be see as a (linear) operator from the "input" space $\mathbf{R}^n$ to the

"output" space $\mathbf{R}^m$. Both points of view are useful.

Transpose, trace and scalar product. The transpose of a matrix A is denoted by A^T, and is the matrix with (i,j) element A_{ji}, $i = 1, \ldots, m$, $j = 1, \ldots, n$.

The *trace* of a square $n \times n$ matrix A, denoted by $\mathbf{Tr}\, A$, is the sum of its diagonal elements: $\mathbf{Tr}\, A = \sum_{i=1}^n A_{ii}$.

We can define the scalar product between two $m \times n$ matrices A, B via

$$\langle A, B \rangle = \mathbf{Tr}\, A^T B = \sum_{i=1}^m \sum_{j=1}^m A_{ij} B_{ij}.$$

In this definition, both A, B are viewed as long vectors with all the columns stacked on top of each other, and the scalar product is the ordinary scalar product between the two vectors.

Range, nullspace, rank. The range of a $m \times n$ matrix A is defined as the following subset of $\mathbf{R}^m$:

$$\mathcal{R}(A) := \{Ax \ : \ x \in \mathbf{R}^n\}.$$

The nullspace of A is given by

$$\mathcal{N}(A) := \{x \in \mathbf{R}^n \ : \ Ax = 0\}.$$

The rank of a matrix A is the dimension of its range; it is also the rank of A^T. Alternatively, it is equal to n minus the dimension of its nullspace. A basic result of linear algebra states that any vector in $\mathbf{R}^n$ can be decomposed as $x = y + z$, with $y \in \mathcal{N}(A)$, $z \in \mathcal{R}(A^T)$, and z, y are orthogonal. (One way to prove this is via the singular value decomposition, seen later.)

The notions of range, nullspace and rank are all based on viewing the matrix as an operator.

Fundamental theorem of linear algebra. Let $A \in \mathbf{R}^{m \times n}$ be a matrix. The fundamental theorem of linear algebra states that any vector in the input space $\mathbf{R}^n$ can be expressed as the sum of two orthogonal vectors: one that lies in the span of A and the other that lies in the nullspace of A^T.

Theorem 1. *For any $m \times n$ matrix A, the input space $\mathbf{R}^n$ can be decomposed as the direct sum of two orthogonal subspaces $\mathcal{N}(A)$ and the range of its transpose, $\mathcal{R}(A^T)$. That is, we have*

$$\mathbf{R}^n = \mathcal{N}(A) \oplus \mathcal{R}(A^T).$$

As a corollary, since $\dim \mathcal{R}(A^T) = \dim \mathcal{R}(A) = \mathbf{Rank}(A)$, we obtain that

$$\dim \mathcal{N}(A) + \mathbf{Rank}(A) = n.$$

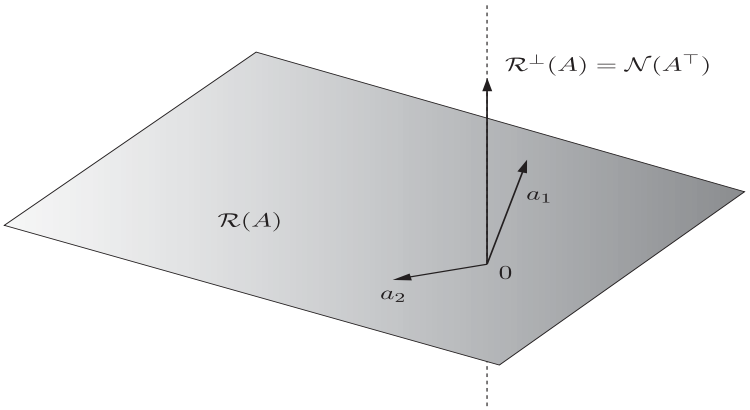

Figure 2.1. Illustration of the fundamental theorem of linear algebra with a 3×2 matrix $A = [a_1, a_2]$.

Orthogonal matrices. A square, $n \times n$ matrix $U = [u_1, \ldots, u_n]$ is orthogonal if its columns form an orthonormal basis (note the unfortunate wording). The condition $u_i^T u_j = 0$ if $i \neq j$, and 1 otherwise, translates in matrix terms as $U^T U = I_n$ with I_n the $n \times n$ identity matrix.

Orthogonal matrices are sometimes called rotations. Indeed, they do not change the Euclidean norm of the input: for every $x \in \mathbf{R}^n$, we have $\|Ux\|_2 = \|x\|_2$ (why?).

2.2.2 Matrix norms

There are many ways to define the norm of a matrix $A \in \mathbf{R}^{m \times n}$.

A first class of matrix norms, which can be called *vector-based,* can be derived by simply collecting the elements of the matrix into a big vector, and defining the matrix norm to be the norm of that vector. A popular choice in this class is the *Frobenius* norm, which corresponds to the Euclidean norm of the vector formed with its elements:

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

Another class of matrix norm can be obtained as *induced* by a vector norm. Specifically, let $\|\cdot\|_{\text{in}}$, $\|\cdot\|_{\text{out}}$ be two vector norms defined on $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively. Then we define the norm of a $m \times n$ matrix A as

$$\|A\| := \max_x \ \|Ax\|_{\text{out}} \ : \ \|x\|_{\text{in}} \leq 1.$$

It turns out that the above indeed defines a matrix norm. This class of norms views A not as a vector, but as a linear operator, and the norm measures the maximum norm (measured with the output norm $\|\cdot\|_{\text{out}}$) that the operator can achieve with bounded inputs (with bounds measured via the "input" norm $\|\cdot\|_{\text{in}}$).

One popular choice corresponds to the case when both input and output norms are Euclidean. This norm is called the *largest singular value* norm, for reasons visited later.

Some norms are both vector-based and induced. The Frobenius norm is not induced; and the largest singular value norm is not vector-based.

2.2.3 Matrix description of subspaces

Linear and affine subspace. A subspace in $\mathbf{R}^n$ can always be described as the nullspace of a matrix A:

$$\mathcal{L} = \{x \in \mathbf{R}^n \ : \ Ax = 0\}.$$

The dimension of $\mathcal{L}$ is the rank of the matrix A. The subspace above is simply the span of the columns of A.

A subset of the form

$$\mathcal{L} = \{x \in \mathbf{R}^n \ : \ Ax = b\},$$

with $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, is referred to as an *affine subspace.*

Hyperplanes. A hyperplane in $\mathbf{R}^n$ is a set described by one affine constraint. Hence, it is an affine subspace of dimension $n-1$. It can be described by one vector $a \in \mathbf{R}^n$ and one scalar b:

$$\mathcal{H} = \left\{x \in \mathbf{R}^n \ : \ a^T x = b\right\}.$$

2.3 Symmetric Matrices

2.3.1 Definition and examples

Definition. A square matrix $A \in \mathbf{R}^{n\times n}$ is *symmetric* if and only if $A = A^T$. The set of symmetric $n \times n$ matrices is denoted $\mathcal{S}^n$.

Examples. Perhaps the simplest example of symmetric matrices is the class of diagonal matrices, which are non-zero only on their diagonal. If $\lambda \in \mathbf{R}^n$, we denote by $\mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, or $\mathbf{diag}(\lambda)$ for short, the $n \times n$ (symmetric) diagonal matrix with λ on its diagonal.

Another important case of a symmetric matrix is of the form uu^T, where $u \in \mathbf{R}^n$. The matrix has elements $u_i u_j$, and is symmetric. Such matrices are called *dyads.* If $\|u\|_2 = 1$, then the dyad is said to be normalized.

A symmetric matrix is a way to describe a weighted, undirected graph: each edge in the graph is assigned a weight A_{ij}. Since the graph is undirected, the edge weight is independent of the direction (from i to j or vice-versa). Hence, A is symmetric.

Another interesting special case of a symmetric matrix is the Jacobian of a function at a given point, which is the matrix containing the second derivatives of the function. Here, we invoke the fact that the second-derivative is independent of the order in which derivatives are taken.

Finally, any quadratic function $q : \mathbf{R}^n \to \mathbf{R}$ can be written as

$$q(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}^T A \begin{pmatrix} x \\ 1 \end{pmatrix},$$

for an appropriate symmetric matrix $A \in \mathcal{S}^{(n+1)}$. If q is a quadratic form (meaning that there are no linear or constant terms in it), then we can write $q(x) = x^T A x$ where now $A \in \mathcal{S}^n$.

2.3.2 Eigenvalue decomposition

A fundamental result of linear algebra states that any symmetric matrix can be decomposed as a weighted sum of normalized dyads that are orthogonal to each other.

Precisely, for every $A \in \mathcal{S}^n$, there exist numbers $\lambda_1, \ldots, \lambda_n$ and an orthonormal basis $(u_1, \ldots, u_n)$, such that

$$A = \sum_{i=1}^n \lambda_i u_i u_i^T.$$

In a more compact matrix notation, we have $A = U \Lambda U^T$, with $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, and $U = [u_1, \ldots, u_n]$.

The numbers $\lambda_1, \ldots, \lambda_n$ are called the eigenvalues of A, and are the roots of the characteristic equation

$$\det(\lambda I - A) = 0,$$

where I_n is the $n \times n$ identity matrix. For arbitrary square matrices, eigenvalues can be complex. In the symmetric case, the eigenvalues are always real. Up to a permutation, eigenvalues are unique, in the sense that there are only n (possibly distinct) solutions to the above equation.

The vectors u_i, $i = 1, \ldots, n$, are called the (normalized) *eigenvectors* of A. In contrast with eigenvalues, there is no unicity property here. For example, the identity matrix has any (unit-norm) vector as eigenvector. However, if all the eigenvalues are distinct, then eigenvectors are unique (up to a change in sign).

It is interesting to see what the eigenvalue decomposition of a given symmetric matrix A tells us about the corresponding quadratic form, $q_A(x) := x^T A x$. With $A = U \Lambda U^T$, we have

$$q_A(x) = (U^T x)^T \Lambda (U^T x) = \sum_{i=1}^n \lambda_i (u_i^T x)^2.$$

The eigenvalue decomposition thus corresponds to the decomposition of the corresponding quadratic form into a sum of squares.

2.3.3 Positive semi-definite matrices

Definition. A matrix $A \in \mathcal{S}^n$ is said to be *positive-definite* (resp. *positive semi-definite*) if and only if all the eigenvalues are positive (resp. non-negative). We use the acronyms PD and PSD for these properties. The set of $n \times n$ PSD matrices is denoted $\mathcal{S}^n_+$, while that of PD matrices is written $\mathcal{S}^n_{++}$. Often, we use the notation $A \succeq 0$ (resp. $A \succ$) for the PSD (resp. PD) property.

In terms of the associated quadratic form $q_A(x) = x^T Ax$, the interpretation is as follows. A matrix A is PD if and only if q_A is a positive-definite function, that is, $q_A(x) = 0$ if and only if $x = 0$. Indeed, when $\lambda_i > 0$ for every i, then the condition

$$q_A(x) = \sum_{i=1}^{n} \lambda_i (u_i^T x)^2 = 0$$

trivially implies $u_i^T x = 0$ for every i, which can be written as $Ux = 0$. Since U is orthogonal, it is invertible, and we conclude that $x = 0$. Thus, to any PD matrix A, we can associate a norm, $\|x\|_A := \sqrt{x^T Ax}$.

Square root and Cholesky decomposition. For PD matrices, we can generalize the notion of ordinary square root of a non-negative number. Indeed, if A is PSD, there exist a unique PD matrix, denoted $A^{1/2}$, such that $A = (A^{1/2})^2$. If A is PD, then so is its square root.

Any PSD matrix can be written as a product $A = LL^T$ for an appropriate matrix L. The decomposition is not unique, and $R = A^{1/2}$ is only a possible choice. If A is positive-definite, then we can choose L to be lower triangular, and invertible. The decomposition is then known as the Cholesky decomposition. The corresponding weighted norm $\|x\|_A$ mentioned above is then simply the Euclidean norm of $L^T x$.

Examples and interpretations. A well-known example of a PSD matrix is the covariance matrix associated with a random variable in $\mathbf{R}^n$. This matrix is defined as

$$\Sigma = \mathbf{E}(x - \hat{x})(x - \hat{x})^T,$$

where $\hat{x} := \mathbf{E}\, x$, and $\mathbf{E}$ denotes the expectation operator associated with the distribution of the random variable x.

Another important example is geometric. For a PD matrix P, and vector $\hat{x}$, the set

$$\mathcal{E}(\hat{x}, P) := \left\{ x \ : \ (x - \hat{x})^T P^{-1} (x - \hat{x}) \leq 1 \right\}$$

is an ellipsoid, with center $\hat{x}$. Its principal axes are given by the orthogonal basis that diagonalizes P, and the semi-axis lengths are the eigenvalues. (Check what happens when P is proportional to the identity.) If P is factored as $P = LL^T$ for some (lower-triangular) matrix L, then the ellipsoid can be interpreted as the affine transformation of a unit Euclidean ball:

$$\mathcal{E}(\hat{x}, P) = \left\{ \hat{x} + Lu \ : \ \|u\|_2 \leq 1 \right\}.$$

2.4 Singular Value Decomposition

The singular value decomposition (SVD) of a matrix expresses the matrix as a three-term product involving simpler matrices. This is one of the most powerful tools of linear algebra, allowing to prove most of the results involving rank, range, nullspace, as well as solving simple optimization problems such as least-squares.

The SVD theorem. Recall from that any matrix $A \in \mathbf{R}^{m\times n}$ with rank one can be written as

$$A = \sigma u v^T,$$

where $u \in \mathbf{R}^m$, $v \in \mathbf{R}^n$, and $\sigma > 0$.

It turns out that a similar result holds for matrices of arbitrary rank r. That is, we can express any matrix $A \in \mathbf{R}^{m\times n}$ with rank one as *sum* of rank-one matrices

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T,$$

where $u_1, \ldots, u_r$ are mutually orthogonal, $v_1, \ldots, v_r$ are also mutually orthogonal, and the σ_i's are positive numbers called the *singular values* of A. In the above, r turns out to be the rank of A.

The following important result applies to any matrix A, and allows to understand the structure of the mapping $x \to Ax$.

Theorem 2. *An arbitrary matrix $A \in \mathbf{R}^{m\times n}$ admits a decomposition of the form*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T = U\tilde{S}V^T, \quad \tilde{S} := \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix},$$

where $U \in \mathbf{R}^{m\times m}$, $V \in \mathbf{R}^{n\times n}$ are both orthogonal matrices, and the matrix S is diagonal:

$$S = \mathbf{diag}(\sigma_1, \ldots, \sigma_r),$$

where the positive numbers $\sigma_1 \geq \ldots \geq \sigma_r > 0$ are unique, and are called the /singular values/ of A. The number $r \leq \min(m, n)$ is equal to the rank of A, and the triplet $(U, \tilde{S}, V)$ is called a /singular value decomposition/ (SVD) of A. The first r columns of U: u_i, $i = 1, \ldots, r$ (resp. V: v_i, $i = 1, \ldots, r$) are called left (resp. right) singular vectors of A, and satisfy

$$Av_i = \sigma_i u_i, \quad A^T u_i = \sigma_i v_i, \quad i = 1, \ldots, r.$$

Note that in the theorem, the zeros appearing alongside S are really blocks of zeros. They may be empty, for example if $r = n$, then there are no zeros to the right of S.

One of the consequences of the SVD theorem is that the rank of a matrix equals to that of its transpose.

Computing the SVD. The SVD of a $m \times n$ matrix A can be easily computed via a sequence of linear transformations. The complexity of the algorithm, expressed roughly as the number of floating point operations per seconds it requires, grows as $O(nm \min(n, m))$. This can be substantial for large, dense matrices.

For sparse matrices, we can speed up the computation if we are interested only in the largest few singular values and associated singular vectors. The *power iteration* algorithm allows to do this, under the generic assumption on the input matrix that its largest singular value is distinct from the second. The algorithm can be written as

$$p \to P(Aq), \quad q \to P(A^T p),$$

where P is the projection on the unit circle (assigning to a non-zero vector v its scaled version $v/\|v\|_2$).

Geometry. The theorem allows to decompose the action of A on a given input vector as a three-step process. To get Ax, where $x \in \mathbf{R}^n$, we first form $\tilde{x} := V^T x \in \mathbf{R}^n$. Since V is an orthogonal matrix, V^T is also orthogonal, and $\tilde{x}$ is just a rotated version of x, which still lies in the input space. Then we act on the rotated vector $\tilde{x}$ by scaling its elements. Precisely, the first r elements of $\tilde{x}$ are scaled by the singular values $\sigma_1, \ldots, \sigma_r$; the remaining $n - r$ elements are set to zero. This step results in a new vector $\tilde{y}$ which now belongs to the output space $\mathbf{R}^m$. The final step consists in rotating the vector $\tilde{y}$ by the orthogonal matrix U, which results in $y = U\tilde{y} = Ax$.

Example: Assume A has the simple form

$$A = \begin{pmatrix} 1.3 & 0 \\ 0 & 2.1 \\ 0 & 0 \end{pmatrix},$$

then for an input vector x in $\mathbf{R}^2$, Ax is a vector in $\mathbf{R}^3$ with first component $1.3x_1$, second component $2.1x_2$, and last component zero.

To summarize, the SVD theorem states that any matrix-vector multiplication can be decomposed as a sequence of three elementary transformations: a rotation in the input space, a scaling that goes from the input space to the output space, and a rotation in the output space. In contrast with symmetric matrices, input and output directions are different.

Low-rank approximations. For a given matrix $A \in \mathbf{R}^{n \times m}$, and $k \leq \min(m, n)$, the low-rank approximation problem is

$$\min_X \ \|A - X\|_F \ : \ \mathbf{Rank}(A) = k.$$

The problem can be solved via the SVD of A; precisely, if an SVD of A is known:

$$A = \sum_{=1}^{\min(m,n)} \sigma_i u_i v_i^T$$

then a minimizer is

$$\hat{A} = \sum_{=1}^{k} \sigma_i u_i v_i^T.$$

Exercises

1. Many supervised learning problems can be expressed in the generic form

   $$\min_w \ f(X^T w) + \lambda \|w\|_2^2,$$

   where $X = [x_1, \ldots, x_m] \in reals^{n \times m}$ is the "data" matrix, f is a "loss" function that depends on the model, and λ is a regularization parameter that allows to trade-off model accuracy on the training set and test set error. For example, in linear least-squares regression, $f(\xi) = \|\xi - y\|_2^2$, where y is the vector of responses. Show that we can without loss of generality impose that the solution lies in the span of the data matrix. Show that it implies that the problem's solution depends only on the "kernel matrix" $K := X^T X$. What is the implication of this fact on the complexity of the problem with respect to n, the number of features?

2. We consider a set of m data points $x_i \in \mathbf{R}^n$, $i = 1, \ldots, m$. We seek to find a line in $\mathbf{R}^n$ such that the sum of the squares of the distances from the points to the line is minimized. To simplify, we assume that the line goes through the origin.

   (a) Consider a line that goes through the origin $\mathcal{L} := \{tu \ : \ t \in \mathbf{R}\}$, where $u \in \mathbf{R}^n$ is given. (You can assume without loss of generality that $\|u\|_2 = 1$.) Find an expression for the projection of a given point x on $\mathcal{L}$.

   (b) Now consider the m points and find an expression for the sum of the squares of the distances from the points to the line $\mathcal{L}$.

   (c) Explain how you would find the answer to the problem via SVD.

   (d) Sovle the problem without the restriction that the line has to pass through the origin.

   (e) Consider a variant of the problem in which the line is replaced with a hyperplane.

3. What is the link between the SVD of a matrix A and the EVD of the related symmetric matrices AA^T and $A^T A$?

4. *Power iteration as block-coordinate descent.* Consider the rank-one approximation problem

   $$\min_{p,q} \ \|A - pq^T\|_F$$

   Solving for p, q separately, derive the power iteration algorithm.