Lecture 3

Vector Norms and Linear Least-Squares Problems

3.1 Lecture Objectives

- Review vectors and matrices, as well as their basic properties and operations.
- Review vector norms, which will become critical for a variety of optimization problems (where we will often seek to minimize the norm of some 'error' vector).
- Understand linear least-squares problems, including the algebraic solution and the corresponding geometric interpretation.

3.2 Vectors and Matrices

3.2.1 Vectors

Many signals and images considered in this course will be expressed as vectors. A length-N vector can be defined as:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \tag{3.1}$$

where each of the entries is generally a complex scalar (although some of the applications reviewed in this course will be real-valued, we will keep this general for now).

3.2.2 Matrices

Matrices are central to many of the formulations and algorithms described in this course. Indeed, linear operations on vectors (for instance, many data acquisition systems in medical imaging) can be expressed as matrices. In general, we consider an $M \times N$ matrix $\mathbf{A}$, defined as:

$$\mathbf{A} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & \ddots & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} \tag{3.2}$$

where each of the entries A_{mn} is a complex number. Sometimes we will consider real-valued vectors and matrices too, but we will try to define concepts in terms of complex data to accommodate operations such as the Discrete Fourier Transform (DFT).

3.2.3 Matrix and vector operations

Please make sure that you are familiar with standard matrix and vector operations and concepts, including:

- Transpose and Hermitian transpose
- Matrix-vector and matrix-matrix multiplication
- Inner product between two vectors (and orthogonality when the inner product is zero)
- Linear independence between vectors
- Matrix rank
- Matrix determinant
- Matrix inverse (for square, invertible matrices)
- Matrix pseudoinverse (for any matrix)
- Range-space of a matrix (linear vector space spanned by this matrix, ie: the space of vectors $\mathbf{y}$ such that $\mathbf{Ax} = \mathbf{y}$ for some $\mathbf{x}$)
- Null-space of a matrix (the space of vectors $\mathbf{x}$ such that $\mathbf{Ax} = \mathbf{0}$)

Do not worry if you are not an expert with each of these concepts. However, if you have never heard of some of them, I suggest you skim through a brief matrix algebra introduction, eg: as provided in the beginning chapters/appendix of some of the references for this course (Bertsekas, Nonlinear Programming, 2016; Liang and Lauterbur, Principles of Magnetic Resonance Imaging, 1999; or just google them!). Also, feel free to email the instructor or attend office hours *over the first few weeks of the course* if you are unsure about your level of expertise.

3.2.4 Linear problems (aka linear systems of equations)

In this course, we will encounter multiple instances of linear problems (ie: linear systems of equations):

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{3.3}$$

Note that these linear systems need not have an exact solution. For instance, does the following linear system (where $\mathbf{x}$ is a length-1 vector) have a solution?

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 7 \end{bmatrix} \tag{3.4}$$

(Why or why not?)

How about the following linear system (where $\mathbf{x}$ is a length-2 vector): does it have a solution?

$$\begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{3.5}$$

(Why or why not?)

How about the following linear system (where $\mathbf{x}$ is a length-2 vector): does it have a solution?

$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{3.6}$$

(Why or why not?)

In addition, systems of equations oftentimes have multiple solutions (where in this case $\mathbf{x}$ is a length-3 vector). For instance, consider the system of equations:

$$\begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{3.7}$$

Can you give two or more possible solutions $\mathbf{x}$ to the linear system above? Can you provide a closed form description of *all* possible solutions to this linear system?

3.2.5 General challenges with linear systems

As can be observed in the examples above, linear systems of equations present two general challenges that prevent the solution from being perfect and uniquely defined:

- *Lack of an exact solution.* In many systems of equations, there is often no exact solution $\hat{\mathbf{x}}$ such that $\mathbf{A}\hat{\mathbf{x}} = \mathbf{y}$. In technical terms, this happens when the data vector $\mathbf{y}$ is not in the range space of the system matrix $\mathbf{A}$. In this case, there is always some fitting error $\mathbf{z} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}} \neq \mathbf{0}$, and we typically solve the problem by seeking a solution that minimizes the "size" (eg: a norm) of this fitting error $\mathbf{z}$. See below in this lecture for details.

- *Multiple choices of* $\mathbf{x}$ *give the same* $\mathbf{Ax}$. Even if we can fit the data $\mathbf{y}$ perfectly, or if we know what fitting error $\mathbf{z}$ we are OK with, there are cases where the system matrix $\mathbf{A}$ is such that multiple (infinitely many) different choices of $\mathbf{x}$ lead to the same vector $\mathbf{Ax}$. In technical terms, this problem arises when the matrix $\mathbf{A}$ has a "null space" (also known as "kernel"), ie: there is a subspace $\Omega \subset \mathbf{R}^N$, such that $\mathbf{Ax_1} = \mathbf{0}$ whenever $\mathbf{x_1} \in \Omega$. In these cases, if $\mathbf{x_0}$ is a solution that matches our data optimally, then any solution $\mathbf{x_0} + \mathbf{x_1}$ (where $\mathbf{x_1} \in \Omega$) will also match our data optimally since $\mathbf{A}(\mathbf{x_0} + \mathbf{x_1}) = \mathbf{Ax_0} + \mathbf{Ax_1} = \mathbf{Ax_0}$.

These challenges lead to the need to develop a systematic way to pick a solution in the face of such ambiguities. Importantly, note that both of these challenges can occur at the same time in the same system of equations. Specific ways to address these challenges (often based on minimization of vector norms) are reviewed in the remainder of this lecture and in subsequent lectures.

3.3 Vector Norms

A norm describes the size or 'length' of a vector $\mathbf{x}$. Although we are used to the traditional Euclidean norm $\sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_N|^2}$, a variety of norms can be defined, with the following conditions:

1. A norm $\|\mathbf{x}\|$ is always ≥ 0, with $\|\mathbf{x}\| = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$
2. $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$ for any scalar c
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)

Note that there is an enormous number of different ways to measure the size of a vector, ie: different vector norms. In this course, we will pay particular attention to the set of norms called ℓ_p norms. For a given value of $p \geq 1$, the ℓ_p norm of a vector is defined as:

$$\|\mathbf{x}\|_p = \left[\sum_n |x_n|^p\right]^{1/p} \tag{3.8}$$

Question: why do we impose the condition that $p \geq 1$ *for* ℓ_p *norms to be actual norms?*

Importantly, the choice of p gives the norm very distinct properties. For instance, consider the three following choices:

- ℓ_1 norm: $\|\mathbf{x}\|_1 = \sum_n |x_n|$ is the sum of the magnitude of the vector components, and is also called the Manhattan norm.
- ℓ_2 norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_n |x_n|^2}$ is the standard Euclidean norm.
- ℓ_∞ norm: $\|\mathbf{x}\|_\infty = \max_n |x_n|$

Throughout this course, we will be using the ℓ_1 and ℓ_2 norms quite frequently, so it is important to be familiar with them. We may also use weighted versions of these norms, defined as (for any given norm ℓ_p):

$$\|\mathbf{x}\|_{p,\mathbf{W}} = \|\mathbf{W}\mathbf{x}\|_p \tag{3.9}$$

for certain choices of matrix $\mathbf{W}$. In this course, we will typically use weighted norms with a diagonal matrix $\mathbf{W}$ such that its diagonal entries are positive and non-zero. These weighted norms allow us to place more weight (ie: "give more importance") on specific entries of our vector, and they are important in a number of image reconstruction and processing applications.

Question: why do we impose the condition that given a diagonal matrix $\mathbf{W}$ none of its diagonal elements be zero in order for the corresponding weighted norms to be actual norms?

3.4 Linear Least-Squares Problems

Let us consider our standard linear problem as defined before:

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{3.10}$$

and focus specifically on the case where an exact solution does not exist, for instance because the measurement vector $\mathbf{y}$ has noise in it. In this case, a very common approach to solving this problem is to seek the best solution $\hat{\mathbf{x}}$ in the sense of matching the measurement vector as closely as possible. How do we define 'as closely as possible'? Here is where the norms come in, as a measure of size (or distance):

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_p \tag{3.11}$$

ie: we seek the solution $\hat{\mathbf{x}}$ such that $\mathbf{A}\hat{\mathbf{x}}$ is as close as possible to $\mathbf{y}$ as measured by a norm (say, an ℓ_p norm since these are the norms we have been focusing on).

Question: will the solution to the optimization problem in Equation 3.11 above depend on the choice of norm? Can you demonstrate it with a simple 2-dimensional example?

For the remainder of this lecture, we will focus on the ℓ_2 norm case for Equation 3.11, which leads to the standard linear least-squares problem.

$$\begin{aligned}
\hat{\mathbf{x}} &= \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 && (3.12)\\
&= \arg\min_{\mathbf{x}} \sqrt{\sum_n |[\mathbf{A}\mathbf{x}]_n - y_n|^2} && (3.13)\\
&= \arg\min_{\mathbf{x}} \sum_n |[\mathbf{A}\mathbf{x}]_n - y_n|^2 && (3.14)\\
&= \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 && (3.15)\\
& && (3.16)
\end{aligned}$$

Note that the removal of the square root in the equation above does not affect the solution of the optimization problem, since the square root is a monotonically increasing operation. We will derive the solution to this problem next.

3.4.1 Algebraic solution

Using basic calculus, in order to minimize $\|\mathbf{Ax}-\mathbf{y}\|_2^2$ we can look for the choice of $\mathbf{x}$ (which we call $\hat{\mathbf{x}}$) such that the gradient of $\|\mathbf{Ax}-\mathbf{y}\|_2^2$ is zero, ie:

$$\mathbf{A}^T(\mathbf{A}\hat{\mathbf{x}}-\mathbf{y})=0 \tag{3.17}$$

(let's stick to real-valued vectors and matrices, for simplicity), which can be rewritten as:

$$\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}}=\mathbf{A}^T\mathbf{y} \tag{3.18}$$

Now, if the matrix $\mathbf{A}^T\mathbf{A}$ has an inverse matrix $(\mathbf{A}^T\mathbf{A})^{-1}$, we can apply it to both sides of the equation above, ie:

$$(\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{A})\hat{\mathbf{x}}=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \tag{3.19}$$

or in other words,

$$\hat{\mathbf{x}}=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \tag{3.20}$$

And this is our linear least squares solution.

Question: In general, can we find closed-form solutions like the one in Equation 3.20 if we choose a different norm (with $p \neq 2$) to solve our linear problem in Equation 3.11?

Question: In general, can we find closed-form solutions like the one in Equation 3.20 if we choose a weighted ℓ_2 norm (with $p=2$ but weighted by some matrix $\mathbf{W}$ as described above) to solve our linear problem in Equation 3.11?

Observation: Note that what Equation 3.20 above tells us is that we have a closed form solution for the linear least-squares problem, and further our solution is linear in the data. This is remarkable and very important for many applications.

3.4.2 Geometric interpretation

As indicated in Figure, the vector $\mathbf{A}\hat{\mathbf{x}}-\mathbf{y}$ is orthogonal (perpendicular) to every vector in the range space of $\mathbf{A}$. What does it mean for two vectors $\mathbf{y_1}$ and $\mathbf{y_2}$ to be 'orthogonal' in the context of Euclidean vector spaces? It means that their inner product is zero, ie:

$$\mathbf{y_1}^T\mathbf{y_2}=\sum_n y_{1,n}y_{2,n}=0$$

(which resembles the algebraic approach described above, if we define $\mathbf{y_1}$ as $\mathbf{Ax}$ and $\mathbf{y_2}$ as $\mathbf{A}\hat{\mathbf{x}}-\mathbf{y}$). Note that the orthogonality condition $(\mathbf{A}\hat{\mathbf{x}}-\mathbf{y})^T(\mathbf{Ax})=0$ needs to be satisfied for any possible $\mathbf{x}$, and therefore leads to the same optimality condition $\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}^T\mathbf{y} = 0$ shown above. This interpretation is depicted graphically in figure 3.1. Note that this concept of orthogonality extends to other vector spaces by defining suitable inner products and their corresponding norms - but we will cross that bridge when we get there.

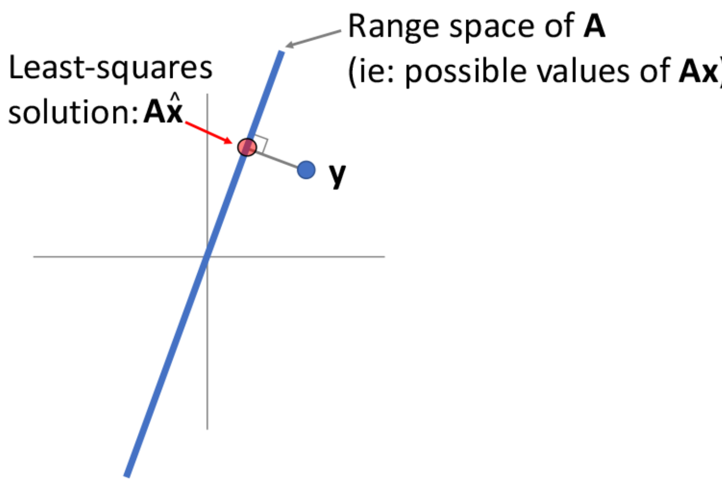

Figure 3.1: Geometric interpretation of the solution to a linear least-squares problem.

3.5 Discussion

In this lecture, we have reviewed vectors and matrices, as well as vector norms and linear problems. We have placed particular emphasis on linear least-squares problems, that have closed-form solutions with elegant algebraic and geometric properties. These properties of linear least squares problems and solutions, which are interesting in low-dimensional problems like the one drawn above, become critical in high dimensional problems (eg: when the space has thousands, or millions of dimensions instead of two dimensions). In these cases, the ability to understand the algebraic and geometric properties of our solutions can give us tools to quickly and intuitively solve problems that would otherwise appear insurmountable.