Lecture 13 Outline

Tuesday, October 8, 2019 2:47 PM

Midterm Exam

Wed, October 16

In-class (~1 hr, 20 minutes)

All material through Lecture 10 on linear transformations and HW 1-4

HW Solutions available to view during office hours, beginning Th, Oct 10

Projections

Least Squares Approximations

Orthogonal & Orthonormal Basis

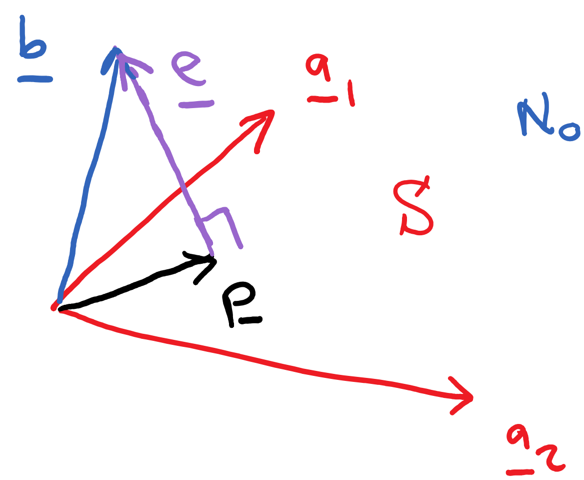
Projections (continued)

Sunday, October 6, 2019 11:34 AM

Projection onto Subspaces

Next consider the projection of a vector $\underline{b}$ in $\mathbb{R}^m$ onto a subspace S in $\mathbb{R}^n$

Let S be spanned by vectors $\underline{a}_1$ and $\underline{a}_2$



Note: $\underline{e} \perp S$

More generally, let the subspace S be spanned by $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n$

We want to find $\underline{A}$, such that

$$\underline{p} = \underline{A}\,\hat{\underline{x}}$$

where the columns of $\underline{A}$ span the subspace and $\hat{\underline{x}}$ is the coordinates ("weights") of the column space of $\underline{A}$

As before, the error vector $\underline{e}$ is perpendicular to subspace S

$$\Rightarrow \quad \underline{e} \cdot (\text{any vector in } S) = 0$$

Thus, since $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n$ are in S, then

$$\underline{a}_1 \cdot \underline{e} = \underline{a}_1^T \underline{e} = 0$$

$$\underline{a}_2 \cdot \underline{e} = \underline{a}_2^T \underline{e} = 0$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\underline{a}_n \cdot \underline{e} = \underline{a}_n^T \underline{e} = 0$$

which can be written as

which can be written as

$$\begin{bmatrix} — & \underline{a}_1^T & — \\ — & \underline{a}_2^T & — \\ & \vdots & \\ — & \underline{a}_n^T & — \end{bmatrix} \underline{e} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\Rightarrow \underline{A}^T \underline{e} = \underline{0}$$

But $\underline{e} = \underline{b} - \underline{p} = \underline{b} - \underline{A}\hat{\underline{x}}$

$$\underline{A}^T \underline{e} = \underline{A}^T (\underline{b} - \underline{A}\hat{\underline{x}}) = 0$$

$$\Rightarrow \boxed{\underline{A}^T \underline{A} \hat{\underline{x}} = \underline{A}^T \underline{b}}$$

or $\hat{\underline{x}} = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{b}$

We wanted a matrix such that the matrix times $\underline{b}$ gives $\underline{p}$

$$\underline{p} = \underline{A}\,\hat{\underline{x}} = \underbrace{\underline{A}\,(\underline{A}^T\underline{A})^{-1}\underline{A}^T}\,\underline{b}$$

This is the matrix that will project any vector $\underline{b}$ in $\mathbb{R}^m$ onto the space spanned by the columns of $\underline{A}$ (in $\mathbb{R}^n$)

What if $\underline{S}$ is spanned by one vector, say $\underline{a}$?

$$\underline{A}\,(\underline{A}^T\underline{A})^{-1}\underline{A}^T = \underline{a}\,(\underline{a}^T\underline{a})^{-1}\underline{a}^T$$

$$= \underbrace{(\underline{a}^T\underline{a})^{-1}}_{\text{scalar}}\,\underline{a}\,\underline{a}^T = \frac{\underline{a}\,\underline{a}^T}{\underline{a}^T\underline{a}}$$

Yes, exactly the previous result!

Question: Why is the following not true in general?

$$\underline{A}\,(\underline{A}^T\underline{A})^{-1}\underline{A}^T \stackrel{?}{=} \underline{A}\,\underline{A}^{-1}\,\underline{A}^{-T}\underline{A}^T = \underline{I}\,\underline{I} = \underline{I}$$

Because we do not know if $\underset{\sim}{A}^{-1}$ and $\underset{\sim}{A}^{-T}$ exist! For example, $\underset{\sim}{A}$ may not even be square, as in case above with $\underset{\sim}{A} = \underset{\sim}{a}$

But why is $(\underset{\sim}{A}^T \underset{\sim}{A})^{-1}$ ok?

$\underset{\sim}{A} \in M_{mn}$ $\qquad$ $\underset{\sim}{A}^T \in M_{nm}$

$\underset{\sim}{A}^T \underset{\sim}{A} \Rightarrow (n \times m)(m \times n) \Rightarrow n \times n$

where columns of $\underset{\sim}{A}$ span subspace S

What if $\underset{\sim}{A}^{-1}$ does exist?

① $\underset{\sim}{A} \in M_{nn}$

② The columns of $\underset{\sim}{A}$ span $\mathbb{R}^n$

$\Rightarrow$ A vector $\underset{\sim}{b}$ in $\mathbb{R}^n$ projected onto $\mathbb{R}^n$ is nothing but $\underset{\sim}{b}$ itself

$\Rightarrow$ If $\underline{A}^{-1}$ exists, then

$$\underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T \text{ must equal } \underline{I},$$

where

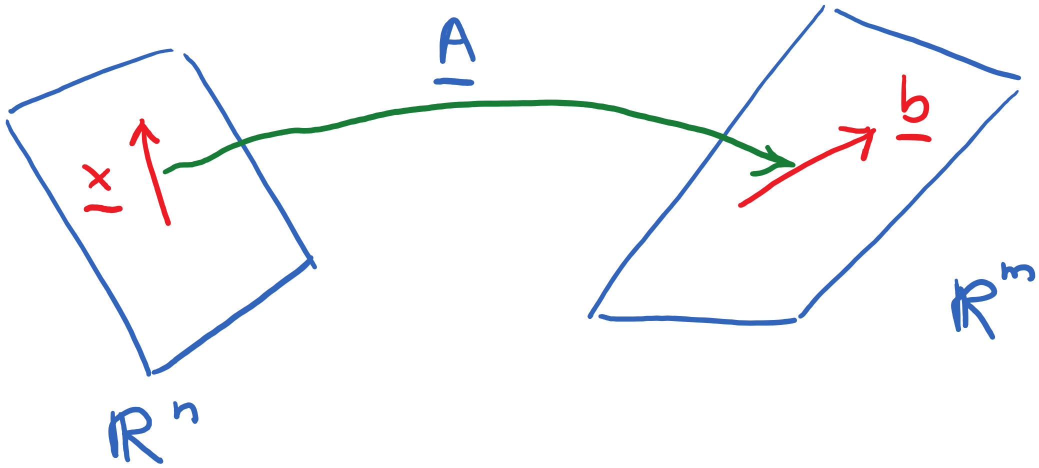
$$\underline{p} = \underline{I}\,\underline{b} = \underline{b}$$

Least Squares Approximations

Sunday, October 6, 2019 7:11 PM

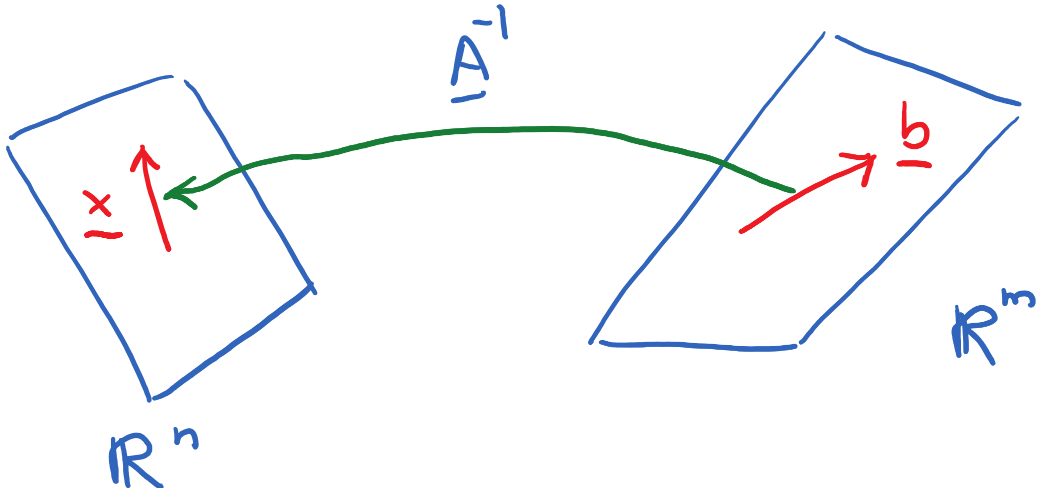
Consider what a linear operator $\underline{A}$ does to $\underline{x}$

$$\underline{A}\,\underline{x} = \underline{b}$$



Given $\underline{A}$ & $\underline{x}$, there definitely is always a vector $\underline{b}$

Is the reverse true?



This mapping only exists, if $\underline{A}^{-1}$ exists, which is not always true

If $\underline{A}^{-1}$ does not exist, then can we find an approximate solution, called $\hat{\underline{x}}$, that does lie in $\mathbb{R}^n$?

Let's try to minimize $\underline{e} = \underline{b} - \underline{p} = \underline{b} - \underline{A}\,\hat{\underline{x}}$

This is a projection onto a subspace!

Using $\underline{I} - \underline{A}$

The solution that minimizes $\underline{e}$ is

$$\underline{A}^T \underline{A}\,\hat{\underline{x}} = \underline{A}^T \underline{b}$$

Here $\hat{\underline{x}}$ is the solution that minimizes

$$\underline{e} = \underline{A}\,\hat{\underline{x}} - \underline{b} \quad \text{or} \quad \|\underline{e}\|_2 = \|\underline{A}\,\hat{\underline{x}} - \underline{b}\|_2$$

Consider the following:

If $\underline{A}^{-1}$ does not exist, then $\underline{A}\,\underline{x} = \underline{b}$ has a solution iff $\underline{b}$ is in the $C(\underline{A})$

If $\underline{b}$ is not in $C(\underline{A})$, then project $\underline{b}$ onto $C(\underline{A})$

Don't forget, $\hat{\underline{x}}$ does not solve $\underline{A}\,\underline{x} = \underline{b}$

Note: The columns of $\underline{A}$ must still be independent for this to work

The set of equations given by

$$\underline{A}^T \underline{A}\,\underline{x} = \underline{A}^T \underline{b}$$

are called the <u>Normal Equations</u>

$$\underline{A}\,\underline{x} = \underline{b} \implies \underline{A}^T \underline{A}\,\underline{x} = \underline{A}^T \underline{b}$$

$$\implies \underline{x} = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{b}$$

This is very difficult to compute

Why? Consider the condition number of $\underline{A}^T \underline{A}$

$$\kappa(\underline{A}^T \underline{A}) = ?$$

First, look at $\kappa(\underline{A}\,\underline{B})$

$$\kappa(\underline{A}\,\underline{B}) = \|\underline{A}\,\underline{B}\| \, \|(\underline{A}\,\underline{B})^{-1}\| = \|\underline{A}\,\underline{B}\| \, \|\underline{B}^{-1}\underline{A}^{-1}\|$$

$$\leq \|\underline{A}\| \, \|\underline{B}\| \, \|\underline{B}^{-1}\underline{A}^{-1}\|$$

$$\leq \|\underline{A}\| \, \|\underline{B}\| \, \|\underline{B}^{-1}\| \, \|\underline{A}^{-1}\| = \kappa(\underline{A})\kappa(\underline{B})$$

$$\Rightarrow \kappa(\underline{A}\,\underline{B}) \sim \kappa(\underline{A})\,\kappa(\underline{B})$$

Also $\kappa(\underline{A}^T) = \kappa(\underline{A})$ (will show later)

Then,

$$\kappa(\underline{A}^T\underline{A}) = \kappa(\underline{A}^T)\kappa(\underline{A}) = \kappa^2(\underline{A})$$

$\Rightarrow$ if $\underline{A}$ is not well-conditioned, then

$\underline{A}^T\underline{A}$ is even worse!

For example, if $\kappa(\underline{A}) \sim 10^3$, then

$$\kappa(\underline{A}^T\underline{A}) \sim 10^6$$

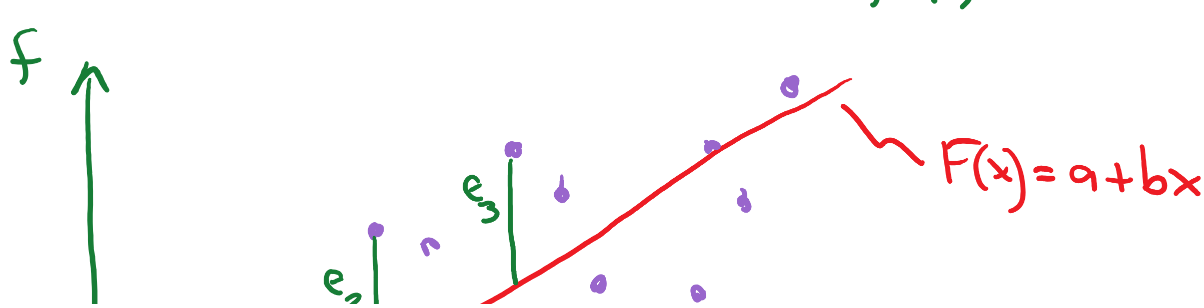
Note: In a few weeks, we will discuss matrix decompositions that allow one to solve

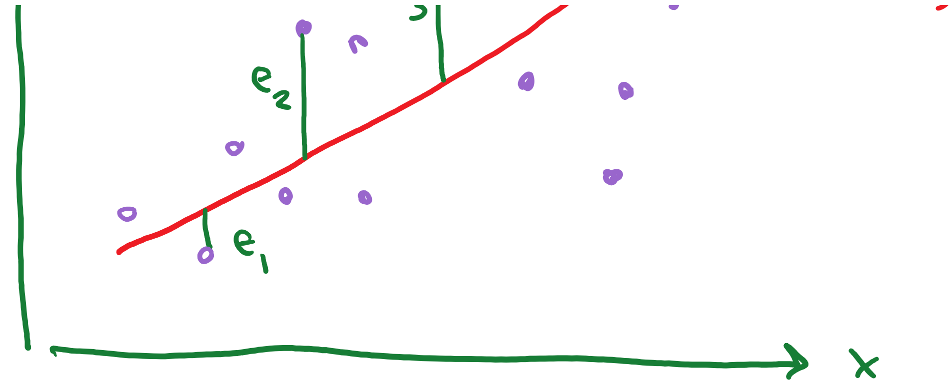
$$\underline{A}^T\underline{A}\,\underline{x} = \underline{A}^T\underline{b} \quad \text{(or similar)}$$

without issue

Example: Curve Fitting

Consider many data points: (x_i, f_i)





Want to find an approximate relation

$F(x) = a + bx$ that describes the data

Want to minimize the error written as

$e_1^2 + e_2^2 + \dots + e_m^2$ for m points

At each point, we wish that

$$a + bx_i = f_i$$

would hold, or

$$\begin{aligned} a + bx_1 &= f_1 \\ a + bx_2 &= f_2 \\ &\vdots \\ a + bx_m &= f_m \end{aligned}$$

Here x_i & f_i are known, while a & b are the unknowns

$$\Rightarrow \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}}_{\underline{A}} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

If at least one of the x_i are distinct, then columns of $\underline{A}$ are independent

If there are more than two data points, then $\underline{A}^{-1}$ does not exist, and this is called an over-constrained system, but $\underbrace{(\underline{A}^T\underline{A})^{-1}}_{\text{size? } 2\times 2 \text{ is full rank } (2)}$ does exist Why?

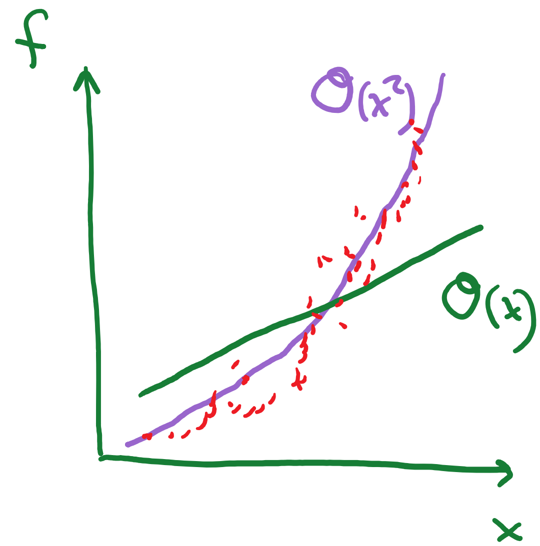
$\Rightarrow$ Solve

$$\underline{A}^T \underline{A} \begin{bmatrix} a \\ b \end{bmatrix} = \underline{A}^T \underline{f}$$

for a & b to minimize $\|\underline{e}\|_2$

Generalize to higher order:

$$F(x) = a + bx + cx^2$$



$$a + bx_1 + cx_1^2 = f_1$$
$$a + bx_2 + cx_2^2 = f_2$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$a + bx_m + cx_m^2 = f_m$$

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 \end{bmatrix}}_{\underline{A}} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

Notice that $F(x)$

$$\text{Solve} \quad \underline{A}^T \underline{A} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \underline{A}^T \underline{f}$$

Notice that $F(x)$ can be non-linear, but least squares is still a linear problem!

Multidimensional:

Fit $F(x,y) = a + bx + cy + dxy$

to the data (x_i, y_i, f_i)

$$\begin{bmatrix} 1 & x_1 & y_1 & x_1 y_1 \\ 1 & x_2 & y_2 & x_2 y_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & y_m & x_m y_m \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

Generic functions:

Imagine data (x_i, g_i) for $i = 1, 2, \dots, m$

Let the interpolent be

$$F(x) = a f_0(x) + b f_1(x) + c f_2(x) + \dots$$

where $f_0(x), f_1(x), f_2(x), \dots$ are some functions e.g. $f_0(x) = x$

$f_1(x) = 1 - x$

$f_2(x) = x^2 - x$

$\vdots$

Then,

$$\underbrace{\begin{bmatrix} f_0(x_1) & f_1(x_1) & f_2(x_1) & \cdots \\ f_0(x_2) & f_1(x_2) & f_2(x_2) & \cdots \\ \vdots & \vdots & \vdots & \\ f_0(x_m) & f_1(x_m) & f_2(x_m) & \cdots \end{bmatrix}}_{\underline{A}} \begin{bmatrix} a \\ b \\ c \\ \vdots \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix}$$

Solve $\underline{A}^T \underline{A} \begin{bmatrix} a \\ b \\ c \\ \vdots \end{bmatrix} = \underline{A}^T \underline{g}$

Aside: Many processes exhibit power law behavior $F(x) \cong a x^\beta$

Then, $\underbrace{\ln F}_{} \cong \underbrace{\ln a}_{} + \beta \underbrace{\ln x}_{}$

Then, $\underbrace{\ln F}_{G} = \underbrace{\ln a}_{\alpha} + \beta \underbrace{\ln x}_{y}$

Use least squares to estimate $\alpha + \beta$

Normal Equations: Error Minimization

Consider fitting $f(x) = a + bx$ to a set of data (x_i, f_i)

Resulting system

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

$$\underline{A} \quad \underline{x} \quad = \quad \underline{b}$$

It was stated that the solution to

$$\underline{A}^T \underline{A} \, \underline{x} = \underline{A}^T \underline{b}$$

minimizes the error $\|\underline{e}\|_2$

Let's show the connection. Here,

$$e = \sum \left((a + bx_i) - f_i \right)^2$$

$$= \sum \left(a + bx_i - f_i \right)^2$$

For the minimum, the gradient wrt

a + b must be zero

$$\frac{\partial e}{\partial a} = \sum \left(a + bx_i - f_i \right) = 0$$

$$\frac{\partial e}{\partial b} = \sum x_i \left(a + bx_i - f_i \right) = 0$$

Rewrite as a linear system:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum f_i \\ \sum x_i f_i \end{bmatrix}$$

for n points (x_i, f_i)

Does $\underline{A}^T \underline{A}\, \underline{x} = \underline{A}^T \underline{b}$ give the same 2×2 system?

For simplicity, let $n = 3$

$(x_1, f_1), (x_2, f_2), (x_3, f_3)$

$$\underline{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}, \quad \underline{x} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\underline{A}^T \underline{A} = \begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} = \begin{bmatrix} 3 & x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 & x_1^2 + x_2^2 + x_3^2 \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$\underline{A}^T \underline{b} = \begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} f_1 + f_2 + f_3 \\ x_1 f_1 + x_2 f_2 + x_3 f_3 \end{bmatrix}$$

$$= \begin{bmatrix} \sum f_i \\ \sum x_i f_i \end{bmatrix}$$

$\Rightarrow \underline{A}^T \underline{A}\, \underline{x} = \underline{A}^T \underline{b}$ results in the same system that minimizes $\underline{e}$ ✓