Today's Lecture

- Orthonormal vectors
- Gram-Schmidt procedure
- QR factorization
- Orthogonal decomposition induced by a matrix

Orthonormal Set of Vectors

- A set of vectors $u_1, u_2, \ldots, u_k \in R^n$ is
  - normalized if $\|u_i\| = 1$, $i = 1, \ldots, k$.
    (u_i are called unit vectors of direction vectors)
  - orthogonal if $u_i \perp u_j$ for $i \neq j$.
  - orthnormal if both normalized and orthogonal.
- In terms of $U = [u_1, u_2 \ \ldots \ u_k]$, orthonormal means

$$U^T U = I_k$$

Orthonormal Set of Vectors

- Orthonormal vectors are independent.
  Show this from $\alpha_1 u_1 + \alpha_2 u_2 + \ \ldots \ + \alpha_k u_k = 0$.
- Hence $u_1, u_2, \ \ldots \ u_k$ is an orthornormal basis for

$$\text{span}(u_1, u_2, \ \ldots \ u_k) = \mathcal{R}(U)$$

Geometric Properties

- Suppose the columns of $U = [u_1, u_2 \ \ldots \ u_k]$ are orthonormal.
- If $w = Uz$, then $\|w\| = \|z\|$.
  - multiplication by U does not change the norm.
  - the mapping $w = Uz$ is isometric, i.e., it preserves distances.
- To show this using matrices.

$$\begin{aligned}\|w\|^2 = \|Uz\|^2 &= (Uz)^T(Uz) = z^T U^T U z \\ &= z^T z = \|z\|^2\end{aligned}$$

Geometric Properties

- Inner products are also preserved. $\langle Uz, U\tilde{z}\rangle = \langle z, \tilde{z}\rangle$.

  If $w = Uz$ and $\tilde{w} = U\tilde{z}$, then

  $$\langle w, \tilde{w}\rangle = \langle Uz, U\tilde{z}\rangle = (Uz)^T(U\tilde{z}) = z^T U^T U\tilde{z} = \langle z, \tilde{z}\rangle$$

- If norms and inner products are preserved
  $\Rightarrow$ angles are also preserved.

  $$\angle(Uz, U\tilde{z}) = \angle(z, \tilde{z})$$

  i.e., multiplication by U preserves inner products, angles and distances.

Orthonormal Basis for R^n

- Suppose $u_1, u_2 \ldots u_k$ is an orthonormal basis for R^n.

  Then, the matrix $U = [u_1, u_2 \ldots u_k]$ is called orthogonal.
  It is a square matrix and satisfies $U^T U = I$.

  A matrix is never referred to as orthonormal.

- It follows that $U^{-1} = U^T$, and hence also $UU^T = I$, i.e.,

  $$\sum_{i=1}^{n} u_i u_i^T = I$$

Expansion in Orthonormal Basis

- Suppose U is orthogonal, so $x = UU^T x$, i.e.,

  $$x = \sum_{i=1}^{n} (u_i^T x) u_i$$

  - $u_i^T x$ is called the component of x in the direction u_i.
  - $a = U^T x$ resolves x into the vector of its u_i components.
  - $x = Ua$ reconstitutes x from its u_i components.
  - $x = Ua = \sum_{i=1}^{n} a_i u_i$ is called the (u_i-) expansion of x.

Expansion in Orthonormal Basis

- Geometric interpretation.
  If U is orthogonal, then the transformation $w = Uz$
  - preserves the norm of vectors, i.e., $\|Uz\| = \|z\|$.
  - preserves the angles between vectors, i.e., $\angle(Uz, U\tilde{z}) = \angle(z, \tilde{z})$.

- Examples.
  - rotation about some axis.
  - reflection through some plane.

- In fact, the converse is true.
  If U is orthogonal, then the mapping $w = Uz$ is either a rotation or a reflection.

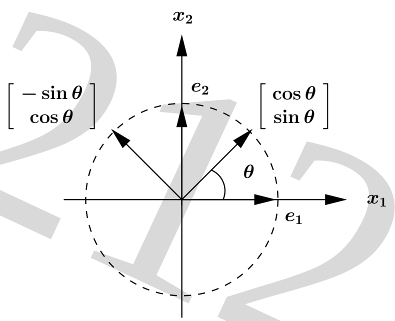
Expansion in Orthonormal Basis

- Example. Rotation by θ in R^2 is given by

$$y = U_\theta x, \qquad U_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$e_1 \rightarrow \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$

$$e_2 \rightarrow \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}$$



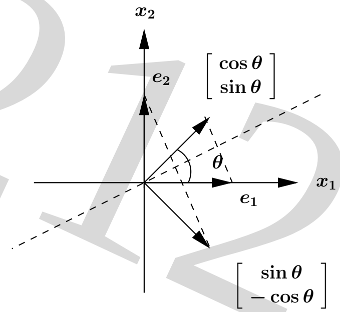
Expansion in Orthonormal Basis

- Example. Reflection across line $x_2 = x_1 \tan(\theta/2)$ is given by

$$y = R_\theta x, \qquad R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

$$e_1 \rightarrow \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$

$$e_2 \rightarrow \begin{bmatrix} \sin\theta \\ -\cos\theta \end{bmatrix}$$



Gram-Schmidt Procedure

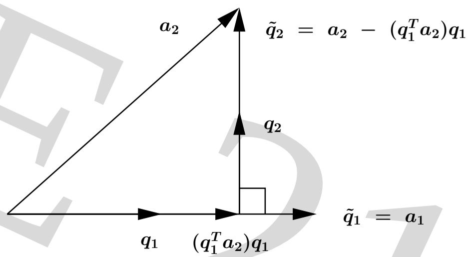
- Given independent vectors $a_1, a_2, \ldots, a_k \in R^n$, find orthonormal vectors $q_1, q_2, \ldots, q_k$ such that

$$\mathrm{span}(a_1, a_2, \ldots, a_r) = \mathrm{span}(q_1, q_2, \ldots, q_r) \text{ for } r \leq k$$

- One can write q_i recursively as follows.
  - let $\tilde{q}_1 = a_1$ and normalize. $q_1 = \tilde{q}_1/\|\tilde{q}_1\|$
  - remove q_1 component from a_2.
  - let $\tilde{q}_2 = a_2 - (q_1^T a_2)q_1$ and normalize. $q_2 = \tilde{q}_2/\|\tilde{q}_2\|$
  - remove q_1, q_2 components from a_3.
  - let $\tilde{q}_3 = a_3 - (q_1^T a_3)q_1 - (q_2^T a_3)q_2$ and normalize. $q_3 = \tilde{q}_3/\|\tilde{q}_3\|$
  - ...

Gram-Schmidt Procedure

- How does this operation work?



- For $i = 1, 2, \ldots, k$, we have

$$a_1 = r_{1i}q_1 + r_{2i}q_2 + \ldots + r_{ii}q_i$$

r_{ij} are easily obtained from the previous procedure.

Gram-Schmidt Procedure

- Written in matrix form.

$$\underbrace{[a_1\ a_2\ \dots\ a_k]}_{A} = \underbrace{[q_1\ q_2\ \dots\ q_k]}_{Q} \underbrace{\begin{bmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ 0 & r_{22} & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{kk} \end{bmatrix}}_{R}$$

Concisely, $A = QR$ where $Q^T Q = I_k$ and R is upper triangular, invertible.

R is full rank and nonsingular since

$$\det R = \prod_{i=1}^{k} r_{ii} \neq 0$$

Gram-Schmidt Procedure

- Note that $r_{ii} \neq 0$.

  If this is not the case, a_i can be written as a linear combination of $a_1, a_2, \dots, a_{i-1}$.

- Comments on $A = QR$.
  - called the QR decomposition of A.
  - never computed using the Gram-Schmidt procedure due to propagation of numerical errors.
  - columns of Q are orthonormal basis for $\mathcal{R}(A)$.

Modified Gram-Schmidt Procedure

- What if $a_1, a_2, \dots, a_k$ are dependent?
  We find $\tilde{q}_j = 0$ for some j, which means a_j is linearly dependent on $a_1, a_2, \dots, a_{j-1}$.

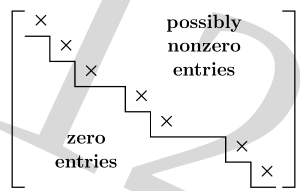
- Modify the algorithm. Skip a_j and move on to a_{j+1} once $\tilde{q}_j = 0$ is encountered.

$$\begin{aligned}
&r = 0; \\
&\text{for } i = 1, \dots, k\ \{ \\
&\qquad \tilde{q} = a_i - \sum_{j=1}^{r} q_j q_j^T a_i; \\
&\qquad \text{if } \tilde{q} \neq 0\ \{\ r = r + 1;\ q_r = \tilde{q}/\|\tilde{q}\|;\ \} \\
&\}
\end{aligned}$$

Modified Gram-Schmidt Procedure

- This results in
  - $r = \text{rank}(A)$.
  - set $q_1, q_2, \dots, q_r$ is an orthonormal basis for $\mathcal{R}(A)$.
  - each a_i is a linear combination of the previously generated q_j's.

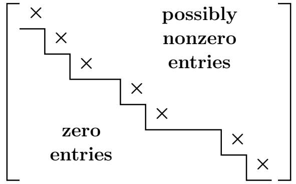
- In matrix form, $A = QR$ where
  - $Q^T Q = I_r$ and
  - $R \in R^{r \times k}$ in upper staircase form, full rank with rank r.



Modified Gram-Schmidt Procedure

- Entries below the staircase are zero.

  Corner entries (marked with $\times$) are nonzero.



- We can move the columns so that all $\times$ entries are towards to the left.

$$A = Q[\tilde{R}\ S]P$$

where

- $Q^TQ = I_r$
- $\tilde{R} \in R^{r\times r}$ is a nonsingular upper triangular matrix.
- $P \in R^{k\times k}$ is a permutaion matrix

Modified Gram-Schmidt Procedure

- Applications.
  - yields an orthonormal basis for $\mathcal{R}(A)$.
  - yields the factorization $A = BC$ with $B \in R^{n\times r}$ and $C \in R^{r\times k}$, $r = \text{rank}(A)$.
  - gives a method for determining if $b \in \text{span}(a_1, a_2, \ldots, a_k)$.

    apply the procedure to $[a_1\ a_2\ \ldots\ a_k\ b]$.
  - staircase pattern in R shows which columns of A are dependent on the previous ones.

Modified Gram-Schmidt Procedure

- Works incrementally.

  One pass of the procedure yields the QR decomposition of $[a_1\ a_2\ \ldots\ a_p]$ for $p = 1, \ldots, k$.

$$[a_1\ a_2\ \ldots\ a_p] = [q_1\ q_2\ \ldots\ q_s]R_p$$

where

- $s = \text{rank}([a_1\ a_2\ \ldots\ a_p])$ and
- R_p is the leading $s \times p$ submatrix of R.

Full QR Factorization

- Using QR factorization on A, write $A = Q_1R_1$. Then, we can also write

$$A = [Q_1\ Q_2]\begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

where $[Q_1\ Q_2]$ is orthogonal, i.e., columns of $Q_2 \in R^{n\times(n-r)}$ are orthonormal and orthogonal to Q_1.

- To find Q_2.
  - find any matrix $\tilde{A}$ such that $[A\ \tilde{A}]$ is full rank (e.g., $\tilde{A} = I$).
  - apply the modified Gram-Schmidt to $[A\ \tilde{A}]$.

Full QR Factorization

- Consequently,
  - Q_1 are orthonormal vectors from columns of A.
  - Q_2 are orthonormal vectors from columns of $\tilde{A}$.

- This means that any set of orthonormal vectors can be extended to an orthonormal basis for R^n.

- Q_1 and Q_2 are called complementary since
  - their ranges are orthogonal and
  - together they span R^n.

Orthogonal Decomposition Induced by A

- What is the use of the full QR decomposition?

$$A^T = [R_1^T \; 0] \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}$$

Thus,

$$z \in \mathcal{N}(A^T) \Leftrightarrow Q_1^T z = 0 \Leftrightarrow z \in \mathcal{R}(Q_2)$$

i.e., columns of Q_2 are an orthonormal basis for $\mathcal{N}(A^T)$.

- Recall that the columns of Q_1 are an orthonormal basis for $\mathcal{R}(A)$.

Orthogonal Decomposition Induced by A

- Thus, we can say that $\mathcal{R}(A)$ and $\mathcal{N}(A^T)$ are complementary subspaces.
  - they are orthogonal. $\mathcal{R}(A) \perp \mathcal{N}(A^T)$.
  - together, they span R^n.

- Fancy notation for complementary subspaces.

$$\mathcal{R}(A) \overset{\perp}{+} \mathcal{N}(A^T) = R^n$$

Every $y \in R^n$ can be written uniqiely as $y = z + w$ with $z \in \mathcal{R}(A)$ and $w \in \mathcal{N}(A^T)$.

Least-squares Method

- Least-squares solution of overdetermined equations.
- Least-squares estimation
- Least-squares data fitting
- Least-squares solution via QR factorization
- Application to system identification

Overdetermined Linear Equations

- Consider $y = Ax$ where $A \in R^{m \times n}$ is tall, i.e., $m > n$.
  - this is termed as overdetermined set of linear equations. (more equations than unknowns).
  - for most y, we cannot solve for x.

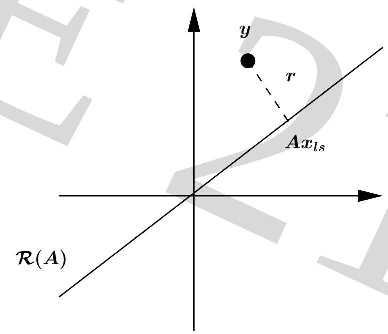
- Find an approximate solution to $y = Ax$.
  - define residual or error $r = Ax - y$.
  - find $x = x_{ls}$ that minimizes $\|r\|$.

  x_{ls} is called the least-squares solution to $y = Ax$.

Overdetermined Linear Equations

- Geometric interpretation.

  Ax_{ls} is a point in $\mathcal{R}(A)$ closest to y.
  Ax_{ls} is the projection of y on $\mathcal{R}(A)$.



Least-squares Solution

- Assume A is full rank and tall.

  To find x_{ls}, take the square of the residual

$$\|r\|^2 = x^T A^T A x - 2y^T A x + y^T y$$

- Minimize with respect to x.

$$2x^T A^T A - 2y^T A = 0$$

  This gives the normal equations,

$$A^T A x = A^T y$$

Least-squares Solution

- Since A is full rank, $A^T A$ is nonsingular. This leads to the well-known formula,

$$x_{ls} = (A^T A)^{-1} A^T y$$

  - x_{ls} is a linear function of y.
  - $x_{ls} = A^{-1} y$ if A is square.
  - x_{ls} solves $y = Ax_{ls}$ if $y \in \mathcal{R}(A)$.

- $A^\dagger = (A^T A)^{-1} A^T$ is called the pseudoinverse of A. $A^\dagger$ is a left inverse of A (full rank and tall)

$$A^\dagger A = (A^T A)^{-1} A^T A = I$$

Least-squares Solution

- Ax_{ls} (projection of y onto $\mathcal{R}(A)$) is linear.

$$Ax_{ls} \; = \; A(A^TA)^{-1}A^Ty$$

The expression $A(A^TA)^{-1}A^T$ is called the projection matrix.

- We now have an optimal residual (in the least-squares sense).

$$r \; = \; Ax_{ls} \; - \; y \; = \; [A(A^TA)^{-1}A^T - I]y$$

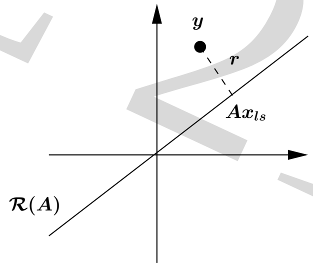
Least-squares Solution

- Orthogonality principle.

Residual r is orthogonal to $\mathcal{R}(A)$.

$$\langle r, Az \rangle \; = \; y^T[A(A^TA)^{-1}A^T - I]Az \; = \; 0$$

for all $z \; \in \; R^n$.



Least-squares Estimation

- Many applications can be categorized as either inversion, estimation and resconstruction.

Many have the form

$$y \; = \; Ax \; + \; v$$

- x is what we want to estimate or reconstruct.
- y is the sensor measurement(s).
- v is an unknown noise or measurement error (assumed small).
- ith row of A characterizes the ith sensor.

Least-squares Estimation

- Least-squares approach. Choose $\hat{x}$ (the estimate) such that it minimizes

$$A\hat{x} \; - \; y$$

i.e., the deviation between

- our actual observation y, and
- what we would observe if $x \; = \; \hat{x}$ and there were no noise, i.e., $v \; = \; 0$.

- The least-squares estimate is simply $\hat{x} \; = \; (A^TA)^{-1}A^Ty$.

Least-squares Estimation

- Linear measurement with noise : $y = Ax + v$ with A full rank and tall.

- Consider a linear estimator of the form $\hat{x} = By$.

  The estimator is called unbiased if $\hat{x} = x$ whenever $v = 0$, i.e., no estimation error when there is no noise.

  Same as $BA = I$, i.e., B is the left inverse of A.

- Estimation error of an unbiased linear estimator is

$$x - \hat{x} = x - B(Ax + v) = -Bv$$

  Then, we would like B to be 'small' and also $BA = I$.

Least-squares Estimation

- Recall our pseudoinverse definition.

  $A^\dagger = (A^T A)^{-1} A^T$ is the smallest left inverse of A.

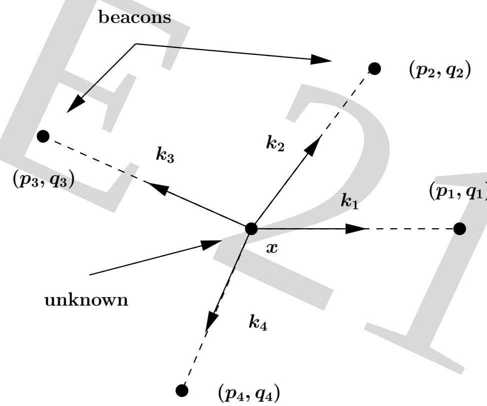
- Thus, for any B with $BA = I$, we have

$$\sum_{i,j} B_{ij}^2 \geq \sum_{i,j} \left(A_{ij}^\dagger\right)^2$$

- Least-squares provides the best linear unbiased estimator (BLUE).

Least-squares Estimation

- Example. Navigation using range measurements from distant beacons.



Least-squares Estimation

- Beacons far from the unknown position $x \in R^2$.

- Distance vector $\rho \in R^4$ is a function of $(x_1, x_2) \in R^2$.

$$\rho_i(x_1, x_2) = \sqrt{(x_1 - p_i)^2 + (x_2 - q_i)^2}$$

- Linearize around $x_0 = 0$.

$$f(x) - f(x_0) \approx Df(x_0)(x - x_0)$$

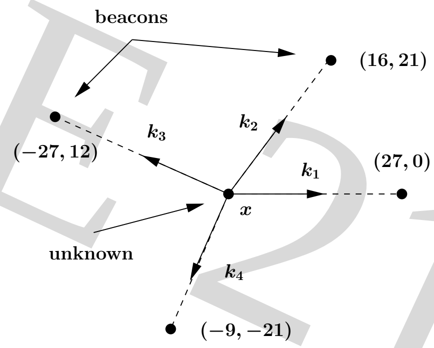
- Changes in range measurements $y \in R^4$ (with noise v)

$$y = \begin{bmatrix} -k_1^T \\ -k_2^T \\ -k_3^T \\ -k_4^T \end{bmatrix} x + v$$

where k_i is a unit vector from the 0 to a beacon i.

- Problem. Determine x given y.

- Let us put in some numbers.



Actual position $x = (3, 5)$.
The measurements are $y = (-2.48, -5.78, 1.26, 5.79)$.

- Based on the coordinates of the beacons, we have

$$y = \begin{bmatrix} -1 & 0 \\ -0.61 & -0.80 \\ 0.91 & -0.41 \\ 0.39 & 0.92 \end{bmatrix} x + v$$

- At the very least, we can figure out the position from two measurements.

  Take y_1 and y_2. The position can be found by considering the first two rows of the linear equation.

$$\hat{x} = By = \left[\begin{bmatrix} -1 & 0 \\ -0.61 & -0.80 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right] y$$

- Thus, the estimated position (relative to the origin) is

$$\hat{x} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0.76 & -1.26 & 0 & 0 \end{bmatrix} y = \begin{bmatrix} 2.48 \\ 5.38 \end{bmatrix}$$

  This gives an error norm of $\|\hat{x} - x\| = 0.64$.

- Using the least-squares method.

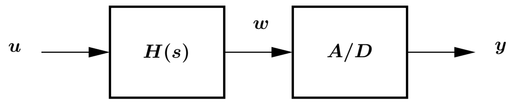
$$\hat{x} = A^{\dagger} y = \begin{bmatrix} -0.45 & -0.17 & 0.46 & 0.06 \\ 0.13 & -0.44 & -0.38 & 0.54 \end{bmatrix} y = \begin{bmatrix} 3.02 \\ 4.86 \end{bmatrix}$$

  The error norm is 0.14.

- Matrices B and $A^{\dagger}$ are both left inverses of A.

  The larger entries in B lead to a larger estimation error.

- Example. Signal estimation.



- Input u is piecewise constant and changes every 1 *sec*.

$$u(t) = x_j, \qquad j - 1 \leq t < j,\ j = 1 \ldots 10$$

- The input is passed through a filter with impulse response $h(t)$.

$$w(t) = \int_0^t h(t - \tau)u(\tau)d\tau$$

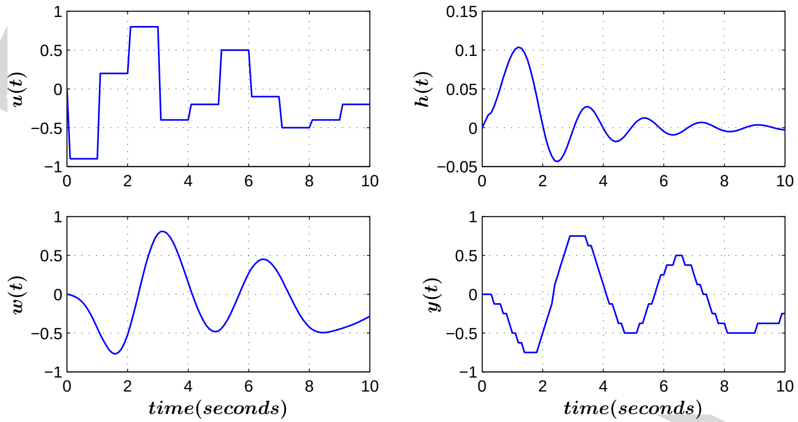
- The signal is then sampled at 10 Hz and 3-bit quantized

$$y_i = Q[w(0.1i)]$$

where $Q(\cdot)$ is a 3-bit quantizing function.

- Problem. Estimate the input ($x \in R^{10}$) from $y \in R^{100}$.

- Illustration. Consider the following signals



- Our linear equation is $y = Ax + v$.

- $A \in R^{100\times 10}$ is given by

$$A_{ij} = \begin{cases} \int_{j-1}^{j} h(0.1i - \tau)d\tau & j \leq \text{ceil}(0.1i) \\ 0 & \text{otherwise} \end{cases}$$

- $v \in R^{100}$ is the 3-bit quantization error.

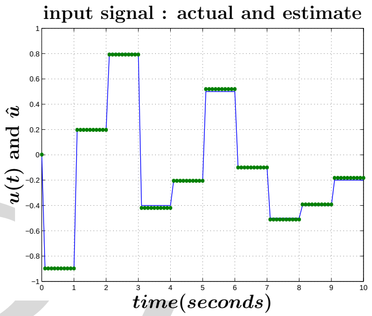
$$v_i = y_i - Q[w(0.1i)] \Rightarrow |v_i| \leq 0.125$$

- The least-squares estimate is

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

RMS error is

$$\frac{\|\hat{x} - x\|}{\sqrt{10}} = 0.011$$



- If we did not filter the input the RMS error would have been 0.035 (due to quantization errors).

We get lower RMS error if we filter the signal first before sampling (and quantizing).

- Given functions $f_1, \ldots, f_n : S \rightarrow R$ (regressors or basis functions).

Also given m data points or measurements, (s_i, g_i), $i = 1, \ldots, m$ where $s_i \in S$, $g_i \in R$ and $m \gg n$.

- Problem. Find the coefficients $x_1, \ldots, x_n \in R$ so that

$$x_1 f_1(s_i) + \ldots + x_n f_n(s_i) \approx g_i, \qquad i = 1, \ldots, m$$

i.e., find a linear combination of the regressors that fits the data.

- Least-squares fit.

Choose x to minimize the mean square error

$$\sum_{i=1}^{m} [x_1 f_1(s_i) + \ldots + x_n f_n(s_i) - g_i]^2$$

- Written in matrix form, the mean square error is

$$\|Ax - g\|^2$$

where $A_{ij} = f_j(s_i)$.

- Assuming A is tall and full rank, the least-squares fit x is given by

$$x = (A^T A)^{-1} A^T g$$

and the corresponding function is

$$f_{ls}(s) = x_1 f_1(s) + \ldots + x_n f_n(s)$$

- Applications include
  - interpolation, extrapolation, smoothing of data
  - developing simple, approximate model of data

Least-squares Polynomial Fitting

- Problem. Fit a polynomial of degree $< n$,

$$p(t) = a_0 + a_1 t + \ldots + a_{n-1} t^{n-1}$$

to the data (t_i, y_i), $i = 1, \ldots, m$.

- Basis functions are $f_j(t) = t^{j-1}$, $j = 1, \ldots, n$.

- Matrix A is a Vandermonde matrix which has the form $A_{ij} = t^{j-1}$, i.e.,

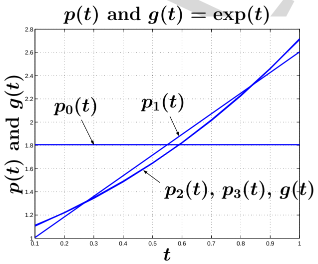
$$\begin{bmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \ldots & t_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_m & t_m^2 & \ldots & t_m^{n-1} \end{bmatrix}$$

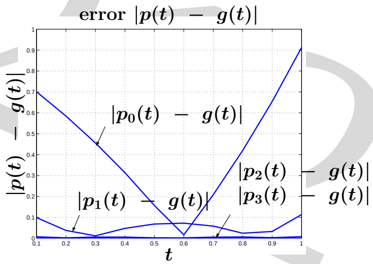
Least-squares Polynomial Fitting

- Assuming $t_k \neq t_l$ for $k \neq l$ and $m \geq n$, then we can say that A full rank.
  - we need to show that $Aa = 0 \Rightarrow a = 0$.
  - if $Aa = 0$, then the polynomial $p(t) = a_0 + a_1 t + \ldots + a_{n-1} t^{n-1}$ vanishes at m points $t_1, \ldots, t_m$.
  - using the fundamental theorem of algebra, $p(t)$ can have no more than $n - 1$ zeros, so $p(t)$ is identically zero, and thus $a = 0$.
  - columns of A are independent $\Rightarrow A$ is full rank.

Least-squares Polynomial Fitting

- Example. Find a polynomial approximation for $g(t) = \exp(t)$.
  - take $m = 10$ points between $t = 0$ and $t = 1$.
  - least-squares fit for degrees 0, 1, 2 and 3.
  - RMS errors are 0.5141, 0.0643, 0.0052 and 0.0003, respectively.





Least-squares from QR Decomposition

- Let $A \in R^{m \times n}$ tall and full rank.

  Factor as $A = QR$ with $Q^T Q = I_n$ and $R \in R^{n \times n}$ upper triangular and invertible.

- The pseudoinverse is

$$(A^T A)^{-1} A^T = (R^T Q^T Q R)^{-1} R^T Q^T = R^{-1} Q^T$$

  Thus, the least-squares solution is $x_{ls} = R^{-1} Q^T y$.

- The corresponding residual (of the optimal solution) is

$$\|y - Ax_{ls}\| = \|(I - QQ^T)y\| = \sqrt{\|y\|^2 - \sum_{i=1}^{n} (q_i^T y)^2}$$

Least-squares from QR Decomposition

- In terms of the full QR decomposition $A = [Q\ \tilde{Q}] \begin{bmatrix} R \\ 0 \end{bmatrix}$,

$$\|y - Ax_{ls}\| = \|\tilde{Q}\tilde{Q}^T y\| = \|\tilde{Q}^T y\| = \sqrt{\sum_{i=1}^{m-n} (\tilde{q}_i^T y)^2}$$

- What does this mean?
  - $QQ^T y$ is the part of y we can match by a linear combination of a_i's.
  - $\tilde{Q}\tilde{Q}^T y$ is the part of y orthogonal to $\mathcal{R}(A)$.

How Many Regressors Do You Need?

- Consider the set of least-squares problems

$$\min_{x_i} \left\| \sum_{i=1}^{p} x_i a_i - y \right\|$$

for $p = 1, \ldots, n$ and regressors $a_1, \ldots, a_p$.

- Solve for x_i's.
  - approximate y by a linear combination of $a_1, \ldots, a_p$.
  - project y onto span$\{a_1, \ldots, a_p\}$.
  - find x_i's, i.e., components of y on $a_1, \ldots, a_p$.
  - we get a better fit as p increases; optimal residual decreases.

How Many Regressors Do You Need?

- Solution for each $p \leq n$ is given by

$$x_{ls} = R_p^{-1} Q_p^T y$$

where
  - R_p is the leading $p \times p$ matrix of R.
  - $Q_p = [q_1\ \ldots\ q_p]$ are the first p columns of Q.

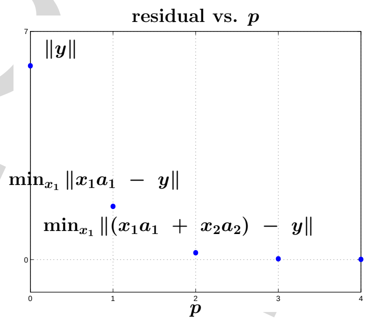
- Thus, we can get the solution to the set of n least-squares problem from one QR decomposition.

How Many Regressors Do You Need?

- The residuals are given by

$$\min_{x_i} \left\| \sum_{i=1}^{p} x_i a_i - y \right\| = \sqrt{\|y\|^2 - \sum_{i=1}^{p} (q_i^T y)^2}$$

- How well does the linear combination of $a_1, \ldots, a_p$ match y for $p = 0, \ldots, n$?



- Orthonormal vectors.
  - Gram-Schmidt procedure
  - QR factorization
  - orthogonal decomposition induced by a matrix

- Least-squares problems.
  - least-squares solution of overdetermined equations.
  - least-squares estimation
  - least-squares data fitting
  - least-squares solution via QR factorization