LEAST-SQUARES ESTIMATION

4.1: Deterministic least squares

- Least-squares estimation core of all future work.
- Make multiple measurements of a constant vector X.

$$Y = HX + v, \qquad \text{where}$$

Y	$\in \mathbb{R}^m$,	Vector of measurements; $y_i = H_i^T X + v_i$.
H	$\in \mathbb{R}^{m\times n}$,	Measurement matrix assumed constant and known.
X	$\in \mathbb{R}^n$,	Constant state vector.
v	$\in \mathbb{R}^m$,	Error vector.

- Assume that $m \ge n$ ➠ Too many measurements.
  - Often there is no (exact) solution for X.
  - Therefore, need to estimate X.

GOAL: Find an estimate of X (called $\hat{X}$) given these erroneous measurements.

IDEAL SITUATION: Pick $\hat{X}$ to minimize $|e_X| = |X - \hat{X}|$.

- Not possible since X not available for comparison.
- Instead, define $\hat{Y} = H\hat{X}$, $e_Y = Y - \hat{Y}$, and pick $\hat{X}$ to minimize

$$J = \frac{1}{2}e_Y^T e_Y = \frac{1}{2}[Y - H\hat{X}]^T[Y - H\hat{X}].$$

- Interpretation: Pick $\hat{X}$ so that the square of the outputs agree as much as possible ➠ "least squares".

NOTE: (Vector calculus)

1. $\dfrac{\mathrm{d}}{\mathrm{d}X} Y^T X = Y$.
2. $\dfrac{\mathrm{d}}{\mathrm{d}X} X^T Y = Y$.
3. $\dfrac{\mathrm{d}}{\mathrm{d}X} X^T A X = (A + A^T)X \ldots$ for symmetric $A \ldots\ 2AX$.

- Expanding the cost function:

$$J = \frac{1}{2}[Y - H\hat{X}]^T[Y - H\hat{X}]$$

$$2J = Y^TY - \hat{X}^TH^TY - Y^TH\hat{X} + \hat{X}^TH^TH\hat{X}.$$

- Stationary point at $\mathrm{d}J/\mathrm{d}\hat{X} = 0$.

$$\frac{\mathrm{d}(2J)}{\mathrm{d}\hat{X}} = -2H^TY + 2H^TH\hat{X} = 0.$$

- Least-squares estimator:

$$\hat{X}_{LSE} = (H^TH)^{-1}H^TY = \underbrace{H^{-L}}_{\text{left pseudo-inverse}} Y.$$

- Question: Is this stationary point a minimum?

$$\frac{\mathrm{d}^2J}{\mathrm{d}\hat{X}^2} = H^TH,$$

  and $H^TH > 0$ (generally) if H has rank n or higher.

  - So, stationary point is a minimum if $\operatorname{rank}(H) = n$.

- Question: Does $(H^TH)^{-1}$ exist? (Is H^TH invertible?)

  - If $\operatorname{rank}(H) = n$, yes.

- Geometric interpretation: $\hat{X} = (H^TH)^{-1}H^TY$ is the *projection* of Y onto the subspace spanned by the columns of H. The error is orthogonal to the columns of H.

- Note: We have said nothing (or at least very little) about the form of the measurement errors v.
- Note: In MATLAB, `Xhat=H\Y;`

Deterministic weighted least squares

- Often find that some measurements are better than others, so we want to emphasize them more in our estimate.
- Use a weighting function

$$J_W = \frac{1}{2}e_Y^T W e_Y = \frac{1}{2}[Y - H\hat{X}]^T W [Y - H\hat{X}].$$

- A useful choice of W is $W = \mathrm{diag}(w_i)$, $i = 1 \ldots m$.

1. $w_i > 0$.
2. $\sum_{i=1}^{m} w_i = 1$ (normalized)
3. If y_j is a good measurement (*i.e.,* clean with very small errors), then make w_j relatively large.

- Large w_j puts much more emphasis on that measurement.

$$\frac{\mathrm{d}J_W}{\mathrm{d}\hat{X}} = 0 \qquad \Rightarrow \qquad \hat{X}_{WLSE} = (H^T W H)^{-1} H^T W Y.$$

- Note: $W = \frac{1}{m} I$ recovers least-square estimate.
- If $H \in \mathbb{R}^{m\times n}$ and $m = n$, $\mathrm{rank}(H) = n$ then a unique solution will exist for this $\hat{X}_{LSE}$.
- What if $m > n$? ⇒ We would like to see some averaging (seems like a good thing to try).

- Does $\hat{X} = (H^T H)^{-1} H^T Y$ average?

EXAMPLE: Consider a simple case: x a scalar, m measurements Y so

$$y_i = x + v_i.$$

(*i.e.,* $H_i = 1$ for each).

- So, $H = [\,1 \;\; 1 \;\; \cdots \;\; 1\,]^T \in \mathbb{R}^{m\times 1}$ and $H^T H = m$.

$$\begin{aligned}\hat{X} = (H^T H)^{-1} H^T Y &= \frac{1}{m}[\,1 \;\; 1 \;\; \cdots \;\; 1\,]Y \\ &= \frac{1}{m}\sum_{j=1}^{m} y_j,\end{aligned}$$

i.e., averaging!

- How does weighting change this? Let y_1 be the really good measurement and the rest are all tied for last.

$$W = \begin{bmatrix} w_1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix}.$$

- Let's see how w_1 changes the solution.

$$\begin{aligned}H^T W H &= [\,1 \;\; 1 \;\; \cdots \;\; 1\,] \begin{bmatrix} w_1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \\ &= \underbrace{[\,w_1 \;\; 1 \;\; \cdots \;\; 1\,]}_{H^T W} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = w_1 + (m-1).\end{aligned}$$

- So,

$$\hat{X}_{WLSE} = (H^T W H)^{-1} H^T W Y = \frac{1}{w_1 + (m-1)}(w_1 y_1 + y_2 + y_3 + \cdots + y_m).$$

- If y_i are approx. the same size and $w_1 \to \infty$ then $\hat{X}_{WLSE} = \dfrac{w_1 y_1}{w_1} = y_1$.
- Weighting emphasized our good, clean measurement and eliminated the averaging process to use the good piece of data available. ➠ We see this all the time, very important.

EXAMPLE: Suppose that a number of measurements $y(t_k)$ are made at times t_k with the intent of fitting a parabola to the data.

$$y(t) = x_1 + x_2 t + x_3 t^2$$

with three measurements: $y(0) = 6$; $y(1) = 0$; $y(2) = 0$.

- We have

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}; \qquad Y = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}; \qquad H = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}.$$

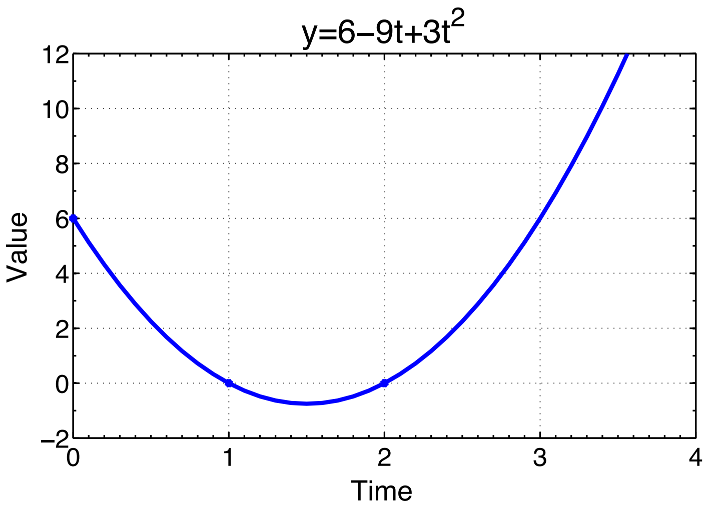
- For $Y = HX + v$ we can solve for the least-squares estimate $\hat{X} = (H^T H)^{-1} H^T Y$. The parabola through the three points is

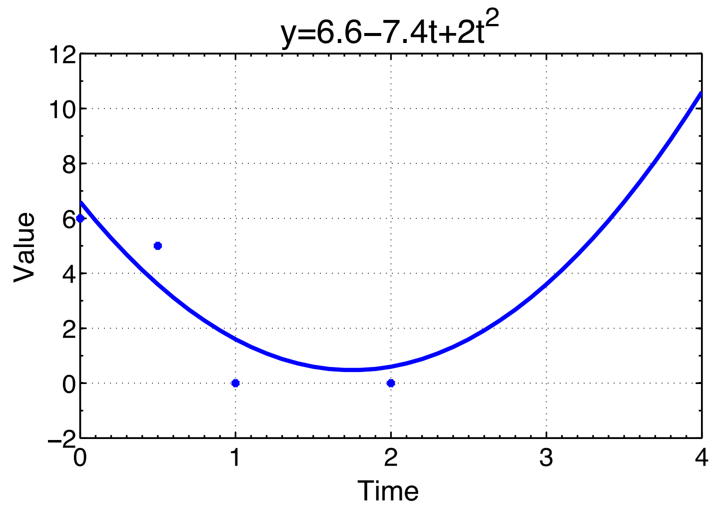
$$y = 6 - 9t + 3t^2.$$

- Now suppose we used more measurements: $y(0.5) = 5$. Error is no longer zero.

$$\text{New } H = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0.5 & 0.25 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}, \qquad e = Y - H\hat{X} = \begin{bmatrix} -0.6 \\ 1.6 \\ -1.2 \\ 0.2 \end{bmatrix}$$

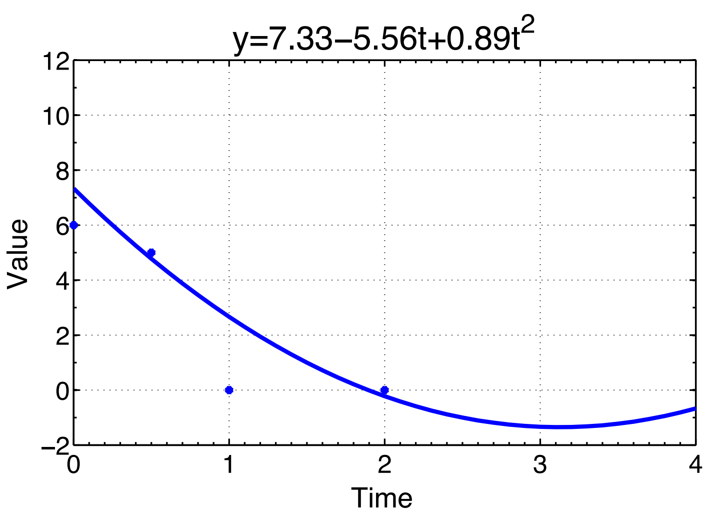
Error is perpendicular to all columns of H.





EXAMPLE: Weighted least squares.

- Incorporate into the estimator design that some measurements may be better than others.
- Let $W = \text{diag}\{0.05, 0.8, 0.05, 0.1\}$. Emphasize $y(0.5)$.



- New error vector:

$$e = [-1.33\ 0.22\ \ -2.67\ 0.22]^T.$$

- No longer perpendicular to the columns of H.

4.2: Stochastic least squares

- Slightly different formulation: Same results but different interpretation.

$$Y = HX + v, \qquad \text{where}$$

Y	$\in \mathbb{R}^m$,	Vector of measurements; $y_i = H_i^T X + v_i$.
H	$\in \mathbb{R}^{m\times n}$,	Measurement matrix assumed constant and known.
X	$\in \mathbb{R}^n$,	Completely unknown (no statistical model).
v	$\in \mathbb{R}^m$,	Rand vector $\sim \mathcal{N}(0, R)$, R diagonal (X, v) independent.

- Noise v must be Gaussian for our linear method to be optimal. . . otherwise, nonlinear methods must be used.
- Use maximum likelihood approach ➠ Select estimate $\hat{X}$ of X to be the value of X that maximizes the probability of our observations Y.

TWO STEPS:

- Find the pdf of Y given unknown constant parameter X: $f_{Y;X}(y; X)$.
  - Note: $f_{Y;X}(y; X)$ works pretty much like the conditional pdf, $f_{Y|X}(y|x)$ except that it recognizes that X is not a random variable per se since it does not have a pdf.
  - Read $f_{Y;X}(y; X)$ as "the pdf of Y parameterized by X".
- Select $\hat{X} = X$ value that yields a maximum value of $f_{Y;X}(y; X)$.

1. What is the distribution of $f_{Y;X}(y; X)$?

   - If v is Gaussian, and X an unknown (but constant) parameter, then $Y = HX + v$ must be Gaussian.

- Therefore, the distribution of Y parameterized by X is Gaussian. To determine the full pdf, must find mean and covariance:

$$
\begin{aligned}
\mathbb{E}[Y;X] &= \mathbb{E}[HX+v;X] \\
&= \mathbb{E}[HX;X] + \mathbb{E}[v;X] \\
&= HX. \\
\Xi_{Y;X} &= \mathbb{E}[(Y-\bar{y})(Y-\bar{y})^T;X] \\
&= \mathbb{E}[YY^T - \bar{y}Y^T - Y\bar{y}^T + \bar{y}\bar{y}^T;X] \\
&= \mathbb{E}[YY^T;X] - \bar{y}\bar{y}^T \\
&= \mathbb{E}[(HX+v)(HX+v)^T;X] - (HX)(HX)^T \\
&= \mathbb{E}[vv^T] = R.
\end{aligned}
$$

- So, $f_{Y;X}(y;X) \sim \mathcal{N}(HX,R)$

$$
= \frac{1}{(2\pi)^{n/2}|R|^{1/2}} \exp\left\{ -\underbrace{\frac{1}{2}(Y-HX)^T R^{-1}(Y-HX)}_{J} \right\}.
$$

2. Now, pick $\hat{X} = X$ that maximizes $f_{Y;X}(y;X)$.

   - Achieved by minimizing exponent of $\exp\{-J\}$.
   - $\hat{X} = \arg\min_X \left\{ \frac{1}{2}(Y-HX)^T R^{-1}(Y-HX) \right\}$.

- This is a weighted least-squares problem where $W = R^{-1}$. Then

$$
\hat{X} = (H^T R^{-1} H)^{-1} H^T R^{-1} Y.
$$

Consistent with previous interpretation?

4.3: Metrics for our estimates

1. "Bias": Is $\mathbb{E}[X-\hat{X}]=0$ for all m large enough to obtain a solution?
2. "Consistency": Does $\lim_{m\to\infty}\mathbb{E}[(X-\hat{X})^T(X-\hat{X})]=0$? That is, does $\hat{X}$ converge to X in mean-square as we collect more data?
3. "Minimum-Variance": Is it the best estimate?

Metrics of WLSE

- Use $W=R^{-1}$ where $\mathbb{E}[vv^T]=R$.

BIAS: Note

$$X-\hat{X}=X-\underbrace{(H^TR^{-1}H)^{-1}H^TR^{-1}}_{H_R^{-L}}\underbrace{(HX+v)}_{Y}$$

$$=X-(H_R^{-L}HX+H_R^{-L}v).$$

- Now, $H_R^{-L}H=(H^TR^{-1}H)^{-1}(H^TR^{-1}H)=I$, so

$$X-\hat{X}=-H_R^{-L}v$$

$$\mathbb{E}[X-\hat{X}]=-\mathbb{E}[H_R^{-L}v]=0$$

since $\mathbb{E}[v]=0$ and we assumed that H, W are known (deterministic). Therefore, WLSE unbiased by zero-mean noise.

CONSISTENCY: $\lim_{m\to\infty}Q_1=\mathbb{E}[(X-\hat{X})^T(X-\hat{X})]=0$?

- Know that $X-\hat{X}=-H_R^{-L}v$.
- Define $Q_2=\mathbb{E}[(X-\hat{X})(X-\hat{X})^T]$. Q_1 is an inner product; Q_2 is an outer product.
- Since $z^Tz=\mathrm{trace}(zz^T)$ then $Q_1=\mathrm{trace}(Q_2)$.

- Then,

$$\begin{aligned} Q_2 &= \mathbb{E}[H_R^{-L} v v^T H_R^{-LT}] = H_R^{-L} \mathbb{E}[v v^T] H_R^{-LT} \\ &= (H^T R^{-1} H)^{-1} H^T R^{-1} R R^{-1} H (H^T R^{-1} H)^{-1} \\ &= (H^T R^{-1} H)^{-1}. \end{aligned}$$

- Therefore, for consistency, need to check

$$\lim_{m\to\infty} Q_1 = \lim_{m\to\infty} \text{trace}\{(H^T R^{-1} H)^{-1}\} \stackrel{?}{=} 0.$$

EXAMPLE: $y_i = x + v_i$, m measurements.

- $v_i \sim \mathcal{N}(0, \sigma^2)$ and i.i.d. ➠ $V \sim \mathcal{N}(0, \sigma^2 I)$ and $R = \sigma^2 I$.
- $H = [\, 1 \;\; 1 \;\; \cdots \;\; 1 \,]^T$ and $H^T H = m$.
- Test:

$$\begin{aligned} \lim_{m\to\infty} \text{trace}\{(H^T R^{-1} H)^{-1}\} &= \lim_{m\to\infty} \text{trace}\{(H^T (\sigma^2 I)^{-1} H)^{-1}\} \\ &= \lim_{m\to\infty} \text{trace}\left\{ \left(\frac{H^T H}{\sigma^2} \right)^{-1} \right\} \\ &= \lim_{m\to\infty} \frac{\sigma^2}{m} \to 0. \end{aligned}$$

Therefore, consistent.

MINIMUM-VARIANCE: An estimator $\hat{X}$ is called a minimum-variance estimator if

$$\mathbb{E}[(\hat{X} - X)^T (\hat{X} - X)] \le \mathbb{E}[(\hat{X}' - X)^T (\hat{X}' - X)]$$

where $\hat{X}'$ is any other estimator. Here, we assume unbiased: $\mathbb{E}[\hat{X}] = \mathbb{E}[\hat{X}'] = X$.

- Special case: Linear unbiased estimators. Consider *any* linear unbiased estimator.

$$\hat{X}' = BY,$$

where $Y = HX + v$. ($\mathbb{E}[v] = 0,\ \Xi_v = \sigma^2 I$).

- We will show that among all estimators of this form, the one with the minimum variance property is the least-squares estimate

$$\hat{X}_{LS} = (H^T H)^{-1} H^T Y.$$

- $\mathbb{E}[\hat{X}'] = \mathbb{E}[BY] = \mathbb{E}[BHX + Bv] = BHX$.
- But, $\mathbb{E}[\hat{X}'] = X$ since assumed unbiased. Therefore $BHX = X$ or $BH = I$.

$$\begin{aligned}\Xi_{\hat{X}'} &= \mathbb{E}[(\hat{X}' - X)(\hat{X}' - X)^T]\\ &= \mathbb{E}[(BHX + Bv - X)(BHX + Bv - X)^T]\\ &= \mathbb{E}[Bvv^T B^T]\\ &= \sigma^2 BB^T.\end{aligned}$$

- To find the estimator with the minimum variance, find B subject to $BH = I$ to make $\text{trace}(\sigma^2 BB^T)$ as small as possible.
- Without loss of generality, write

$$\hat{X}' = BY = (B_o + \bar{B})Y$$

where $B_o = (H^T H)^{-1} H^T$, the least-squares coefficients.

$$\begin{aligned}\text{trace}(\sigma^2 BB^T) &= \text{trace}(\sigma^2 (B_o + \bar{B})(B_o + \bar{B})^T)\\ &= \text{trace}(\sigma^2 (B_o B_o^T + B_o \bar{B}^T + \bar{B} B_o^T + \bar{B}\bar{B}^T)).\end{aligned}$$

- Now, $BH = I$, so $(B_o + \bar{B})H = I$. By definition of B_o we have $I + \bar{B}H = I$ or $\bar{B}H = 0$ and $H^T\bar{B}^T = 0$.
- Therefore $B_o\bar{B}^T = (H^TH)^{-1}H^T\bar{B}^T = 0$.
- Therefore $\bar{B}B_o^T = \bar{B}H(H^TH)^{-1} = 0$. So,

$$\text{trace}(\sigma^2BB^T) = \text{trace}(\sigma^2(B_oB_o^T + \bar{B}\bar{B}^T)),$$

  but for any matrix B the diagonal terms of BB^T are always sums of squares and hence non-negative. Therefore, the above equation is minimized when $\bar{B} = 0$.
- Conclusion:

$$\hat{X}_{LS} = (H^TH)^{-1}H^TY$$

  is the minimum-variance, unbiased linear estimate of X.
  (BLUE="Best Linear Unbiased Estimator")

4.4: Recursive estimation

- All of the processing so far has been "batch" ➠ Collect ALL the data and reduce it at once.
- Problem: If a new piece of data comes along, we have to repeat the entire calculation over again!
- Would like to develop a *RECURSIVE* form of the estimator so that we can easily include new data as it is obtained ➠ *REAL TIME.*

1. Set up data collection.
2. Discuss batch process and analyze it to develop recursive form.
3. Look at properties of new estimator.

Basic example

- Data collection in two lumps. Collect two vectors y_1 and y_2.

1. $y_1 = H_1 X_1 + v_1$ and $v_1 \sim \mathcal{N}(0, R_1)$. Assume X constant but no statistical properties known. Use maximum likelihood.
2. More data *from same X.* $(X_1 = X_2)$. $y_2 = H_2 X_2 + v_2$ and $v_2 \sim \mathcal{N}(0, R_2)$.

- y_1, y_2 may be measurements at one time or two distinct times.
- Eventually, would like to use
  - Part 1 of the estimate process $y_1 \to \hat{X}_1$.
  - Part 2 of the estimate process $\hat{X}_1$ and $y_2 \to \hat{X}_2$.
- Start with batch approach to find $\hat{X}_2$.
  - Final result after all data has been reduced and used.

- Can write $\hat{X}_2$ as $\hat{X}_2 = \hat{X}_1 + \delta x$ so that δx is clearly a function of y_2.
  - Then, we have the update/recursion that we really need.

BATCH:

$$\underbrace{\begin{bmatrix} y_1 \\ \hdashline y_2 \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} H_1 \\ \hdashline H_2 \end{bmatrix}}_{H} X + \underbrace{\begin{bmatrix} v_1 \\ \hdashline v_2 \end{bmatrix}}_{v}$$

so that $Y = HX + v$.

- We will assume that $v \sim \mathcal{N}(0, R)$, where
$$R = \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix}.$$
That is, no correlation between v_1 and v_2.

- If R_1 and R_2 are diagonal this is not a bad assumption.
  - Noises not correlated within either data stream, so not correlated between data-collection processes either.

- Solution: $(H^T R^{-1} H)\hat{X}_2 = H^T R^{-1} Y$.

1. $H^T R^{-1} H = \begin{bmatrix} H_1^T \;\vdots\; H_2^T \end{bmatrix} \begin{bmatrix} R_1^{-1} & \vdots & 0 \\ \hdashline 0 & \vdots & R_2^{-1} \end{bmatrix} \begin{bmatrix} H_1 \\ \hdashline H_2 \end{bmatrix} =$
   $[(H_1^T R_1^{-1} H_1) + (H_2^T R_2^{-1} H_2)]$.
2. $H^T R^{-1} = \begin{bmatrix} H_1^T R_1^{-1} \;\vdots\; H_2^T R_2^{-1} \end{bmatrix}$. Therefore,
$$[(H_1^T R_1^{-1} H_1) + (H_2^T R_2^{-1} H_2)]\hat{X}_2 = H_1^T R_1^{-1} y_1 + H_2^T R_2^{-1} y_2.$$

- Further analysis: Define

1. $\hat{X}_2 = \hat{X}_1 + \delta x$.
2. $\hat{X}_1 = (H_1^T R_1^{-1} H_1)^{-1} H_1^T R_1^{-1} y_1$.

GOAL: Find δx as a function of y_2, $\hat{X}_1$ ➠ Things we know and new things we measured.

- Consistent with batch estimate if same data used. Batch can easily handle correlated v.

SOLUTION: Let $Q_2 = [(H_1^T R_1^{-1} H_1) + (H_2^T R_2^{-1} H_2)]^{-1}$. Let $Q_1 = [H_1^T R_1^{-1} H_1]^{-1}$.

- Batch solution becomes (Second line: $Q_1^{-1}\hat{X}_1 = H_1^T R_1^{-1} y_1$)

$$\begin{aligned}
Q_2^{-1}\hat{X}_2 &= H_1^T R_1^{-1} y_1 + H_2^T R_2^{-1} y_2 \\
Q_2^{-1}\hat{X}_1 + Q_2^{-1}\delta x &= Q_1^{-1}\hat{X}_1 + H_2^T R_2^{-1} y_2 \\
Q_2^{-1}\delta x &= (Q_1^{-1} - Q_2^{-1})\hat{X}_1 + H_2^T R_2^{-1} y_2 \\
Q_2^{-1}\delta x &= -H_2^T R_2^{-1} H_2 \hat{X}_1 + H_2^T R_2^{-1} y_2 = H_2^T R_2^{-1}(y_2 - H_2\hat{X}_1) \\
\delta x &= \underbrace{\left[H_1^T R_1^{-1} H_1 + H_2^T R_2^{-1} H_2\right]^{-1}}_{Q_2} H_2^T R_2^{-1} \underbrace{(y_2 - \underbrace{H_2\hat{X}_1}_{\hat{y}_2})}_{\text{prediction error}}.
\end{aligned}$$

- In desired form since $\delta x = \text{fn}(y_2, \hat{X}_1)$.
- Recall Q_1 from our consistency check. $Q_1 = \mathbb{E}\left[(X - \hat{X}_1)(X - \hat{X}_1)^T\right]$. Called the *ERROR COVARIANCE MATRIX.*
- $X - \hat{X}_1 = H_R^{-L} v$. Therefore

$$\begin{aligned}
Q_1 &= \mathbb{E}\left[H_R^{-L} v v^T H_R^{-LT}\right] \\
&= H_R^{-L} R H_R^{-LT} = \left(H_1^T R^{-1} H_1\right)^{-1}.
\end{aligned}$$

Same as defined above!

- Note:

$$\begin{aligned} Q_2 &= \left(H^T R^{-1} H\right)^{-1} \\ &= \left[H_1^T R_1^{-1} H_1 + H_2^T R_2^{-1} H_2\right]^{-1} \\ &= \left[Q_1^{-1} + H_2^T R_2^{-1} H_2\right]^{-1}, \end{aligned}$$

or, the simple update formula

$$Q_2^{-1} = Q_1^{-1} + H_2^T R_2^{-1} H_2.$$

Recursive Estimation

- $\hat{X}_1 = Q_1 H_1^T R_1^{-1} y_1; \qquad Q_1^{-1} = H_1^T R_1^{-1} H_1.$
- $\hat{X}_2 = \hat{X}_1 + Q_2 H_2^T R_2^{-1}\left[y_2 - H_2 \hat{X}_1\right]; \quad Q_2^{-1} = Q_1^{-1} + H_2^T R_2^{-1} H_2.$
- The $y_2 - H_2\hat{X}_1$ term is called the "innovations process" or the "prediction error".
- Innovation compares the new measurement with prediction based on old estimate. ➠ What is new in this data?

Special Cases

1. First set of data collected was not very good, so we get a poor first estimate. $Q_1^{-1} \approx 0$.
   - Therefore, $Q_2^{-1} \approx H_2^T R_2^{-1} H_2$, and

$$\begin{aligned} \hat{X}_2 &= \hat{X}_1 + \left(H_2^T R_2^{-1} H_2\right)^{-1} H_2^T R_2^{-1}\left[y_2 - H_2\hat{X}_1\right] \\ &= \left(H_2^T R_2^{-1} H_2\right)^{-1} H_2^T R_2^{-1} y_2. \end{aligned}$$

   - Use only second data set to form estimate.

2. Second measurement poor. $R_2 \to \infty$.

   - Therefore $Q_2 \approx Q_1$ and the update gain

$$Q_2 H_2^T R_2^{-1} \approx Q_1 H_2^T R_2^{-1} \to 0.$$

   - If $y - H_1\hat{X}_1$ small, $\hat{X}_2 \approx \hat{X}_1$. Not much updating.

EXAMPLE: First take k measurements. $y_i = x + v_i$. $R_1 = I$, $H_i = I$. Therefore,

$$\hat{X}_1 = \frac{1}{k}\sum_{i=1}^{k} y_i; \qquad Q_1 = (H_1^T H_1)^{-1} = \frac{1}{k} I.$$

- Take one more measurement: $y_{k+1} = x + v_{k+1}$. $R_2 = I$. $H_2 = I$.

$$Q_2^{-1} = Q_1^{-1} + H_2^T H_2 = (k+1)I. \quad \Rightarrow \quad Q_2 = \frac{1}{k+1} I$$

$$\begin{aligned}\hat{X}_2 &= \hat{X}_1 + Q_2 H_2^T\left(y_{k+1} - H_2\hat{X}_1\right)\\ &= \hat{X}_1 + \frac{1}{k+1}\left(y_{k+1} - \hat{X}_1\right)\\ &= \frac{k\hat{X}_1 + y_{k+1}}{k+1}.\end{aligned}$$

- Update is a weighted sum of $\hat{X}_1$ and y_{k+1}.
- For equal noises, note that we get very small updates as $k \to \infty$.
- If the noise on y_{k+1} small, $R_2 = \sigma^2 I$, where $\sigma^2 \ll 1$

$$Q_2^{-1} = Q_1^{-1} + H_2^T R_2^{-1} H_2 = k + 1/\sigma^2 \quad \ldots \quad Q_2 = \frac{\sigma^2}{\sigma^2 k + 1} I.$$

- Now,

$$\hat{X}_2 = \hat{X}_1 + \frac{\sigma^2}{\sigma^2 k + 1}\frac{1}{\sigma^2}\left(y_{k+1} - \hat{X}_1\right)$$

$$= \frac{\sigma^2 k \hat{X}_1 + y_{k+1}}{\sigma^2 k + 1}.$$

- As $\sigma^2 \to 0$, $\hat{X}_2 \approx y_{k+1}$, as expected.

General form of recursion

Initialize algorithm with $\hat{X}_0$ and $Q_0^{-1} \approx 0$.

for $k = 0 \ldots \infty$,

% Update covar matrix.

$Q_{k+1} = \left[Q_k^{-1} + H_{k+1}^T R_{k+1}^{-1} H_{k+1}\right]^{-1}$.

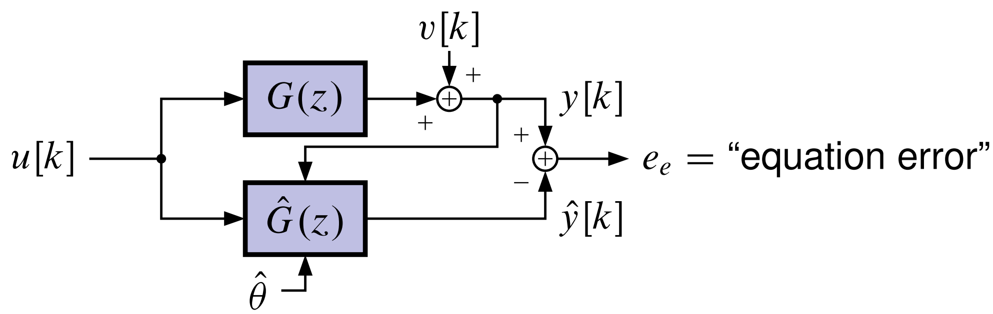
% Update estimate.

$\hat{X}_{k+1} = \hat{X}_k + Q_{k+1} H_{k+1}^T R_{k+1}^{-1} \left[y_{k+1} - H_{k+1}\hat{X}_k\right]$.

endfor

4.5: Example: Equation-error system identification

- Some types of system identification can be solved using least-squares optimization.
- One is known as equation error, and is computed as shown in the diagram:



- In the diagram, given measurements of $\{u[k], y[k]\}$, $\hat{y}[k]$ is computed to be

$$\hat{y}[k] = -\hat{a}_1 y[k-1] - \cdots - \hat{a}_n y[k-n] + \hat{b}_1 u[k-1] + \cdots + \hat{b}_n u[k-n].$$

- Note that $\hat{y}[k] = y[k]$ only when there is no source of error. Not equal if noisy measurements or plant model errors.
- At each k, we denote this equation error as $e_e[k] = y[k] - \hat{y}[k]$.

$$\begin{aligned} e_e[k] &= y[k] + \hat{a}_1 y[k-1] + \cdots + \hat{a}_n y[k-n] - \hat{b}_1 u[k-1] - \cdots - \hat{b}_n u[k-n] \\ &= y[k] - a_e[k]\hat{\theta} \end{aligned}$$

where $a_e[k] = \begin{bmatrix} -y[k-1] & -y[k-2] & \cdots & u[k-1] & u[k-2] & \cdots \end{bmatrix}$.

- Let $E_e = \begin{bmatrix} e_e[1] & \cdots & e_e[n] \end{bmatrix}^T$ then $E_e = Y - A_e\hat{\theta}$.
- Summary:

$$J = \min_{\hat{\theta}} f(E_e), \qquad E_e = Y - A_e\hat{\theta},$$

and E_e is linear in $\hat{\theta}$!

- Some choices for $f(\cdot)$:

1. $\min_{\hat{\theta}} \sum_{k=1}^{n} |e[k]| = \|e[k]\|_1$.
2. $\min_{\hat{\theta}} \sum_{k=1}^{n} e^2[k] = \|e[k]\|_2 = \min_{\hat{\theta}} E_e^T E_e$.
3. $\min_{\hat{\theta}} \max_{k} |e[k]| = \|e[k]\|_\infty$.

- An analytic solution exists for (2). The other two cases may be solved with Linear Programming.

Least-squares equation error

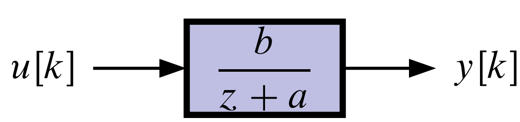
- Given $\{u[k], y[k]\}$, form Y, A_e.
- $\min_{\hat{\theta}} E_e^T E_e = \min_{\hat{\theta}} (Y - A_e\hat{\theta})^T (Y - A_e\hat{\theta}) \Rightarrow A_e^T A_e \hat{\theta} = A_e^T Y$, the MMSE solution.
- If A_e is full rank, $(A_e^T A_e)^{-1}$ exists and

$$\hat{\theta} = (A_e^T A_e)^{-1} A_e^T Y.$$

- When is A_e full rank?

1. $n >$ size$(\hat{\theta})$.
2. $u[k]$ is “sufficiently exciting”.
3. $\hat{\theta}$ is identifiable (one unique $\hat{\theta}$).

EXAMPLE: First-order system.



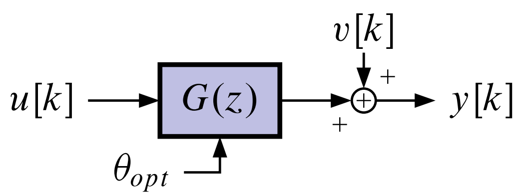
- $e_e[k] = y[k] + \hat{a}y[k-1] - \hat{b}u[k-1]$.

$$E_e = \underbrace{\begin{bmatrix} y[1] \\ y[2] \\ y[3] \end{bmatrix}}_{Y} - \underbrace{\begin{bmatrix} -y[0] & u[0] \\ -y[1] & u[1] \\ -y[2] & u[2] \end{bmatrix}}_{A_e} \underbrace{\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}}_{\hat{\theta}}$$

$$\hat{\theta} = (A_e^T A_e)^{-1} A_e^T Y.$$

Stochastic performance of least squares

- We are interested in the consistency of the least-squares estimate solution $\hat{\theta}$ when our system measurements contain noise.



- Specifically, does $\mathbb{E}[\hat{\theta}] \to \theta_{opt}$ as number of measurements$\to \infty$, and if so, what about the variance of the error $\mathbb{E}\left[(\hat{\theta} - \theta_{opt})^T (\hat{\theta} - \theta_{opt})\right]$?
- In the following, assume θ_{opt} exists and

$$y[k] = a_e[k]\theta_{opt} + e_e[k]$$

or,

$$Y = A_e\theta_{opt} + E_e.$$

- The asymptotic least-square estimate

$$\hat{\theta} = \mathbb{E}[\hat{\theta}(\infty)]$$

can be determined by taking the expected value of the normal equations

$$A_e^T A_e \hat{\theta} = A_e^T Y$$

$$\mathbb{E}\left[A_e^T A_e \hat{\theta}\right] = \mathbb{E}\left[A_e^T Y\right]$$

with A_e full rank and $R_A = \mathbb{E}\left[A_e^T A_e\right]$

$$\mathbb{E}[\hat{\theta}] \to R_A^{-1}\mathbb{E}\left[A_e^T Y\right].$$

- Now, $Y = A_e\theta_{opt} + E_e$

$$\begin{aligned}\mathbb{E}[\hat{\theta}] &\to R_A^{-1}\mathbb{E}\left[A_e^T[A_e\theta_{opt} + E_e]\right] \\ &= \theta_{opt} + R_A^{-1}\mathbb{E}\left[A_e^T E_e\right].\end{aligned}$$

- So, the least-squares estimate is unbiased if

$$\mathbb{E}\left[A_e^T E_e\right] = 0.$$

- Since

$$\begin{aligned}\mathbb{E}\left[A_e^T E_e\right] &= \mathbb{E}\left[\sum_{k=1}^{\infty} a_e^T[k]e_e[k]\right] \\ &= \sum_{k=1}^{\infty}\mathbb{E}\left[a_e^T[k]e_e[k]\right],\end{aligned}$$

we know that the estimate will be unbiased if for every k

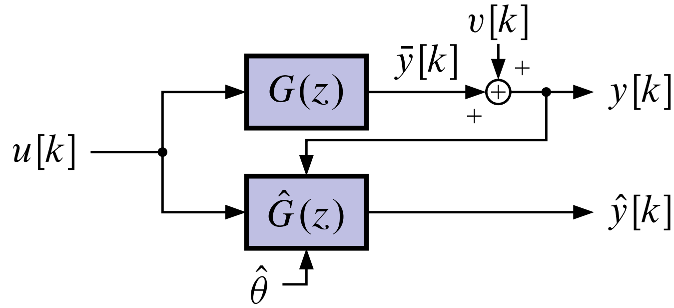
$$\mathbb{E}\left[a_e^T[k]e_e[k]\right] = 0.$$

- Let's check equation-error system ID for bias. Let

$$Y(z) = \frac{B(z)}{z^n + A(z)}U(z) + V(z)$$

or

$$\begin{aligned}y[k] = &-a_1\bar{y}[k-1] - \cdots - a_n\bar{y}[k-n] \\ &+b_1u[k-1] + \cdots + b_nu[k-n] + v[k].\end{aligned}$$



- Now, $y[k] = \bar{y}[k] + v[k]$ or $\bar{y}[k] = y[k] - v[k]$.

$$\begin{aligned} y[k] = -a_1 y[k-1] - \cdots - a_n y[k-n] \\ +b_1 u[k-1] + \cdots + b_n u[k-n] \\ +v[k] + a_1 v[k-1] + \cdots + a_n v[k-n]. \end{aligned}$$

- Check for bias:

$$y[k] = a_e[k]\theta + e_e[k]$$

where

$$a_e[k] = \begin{bmatrix} -y[k-1] & \cdots & -y[k-n], & u[k-1] & \cdots & u[k-n] \end{bmatrix}$$

$$e_e[k] = v[k] + a_1 v[k-1] + \cdots + a_n v[k-n].$$

$$\mathbb{E}\left[a_e^T[k] e_e[k]\right] = \mathbb{E}\left[\begin{bmatrix} -y[k-1] \\ \vdots \\ -y[k-n] \\ u[k-1] \\ \vdots \\ u[k-n] \end{bmatrix} e_e[k]\right]$$

$$= \mathbb{E}\left[\begin{bmatrix} -\bar{y}[k-1]-v[k-1] \\ \vdots \\ -\bar{y}[k-n]-v[k-n] \\ u[k-1] \\ \vdots \\ u[k-n] \end{bmatrix} \begin{matrix} [v[k]-a_1v[k-1]\cdots \\ -a_nv[k-n]] \end{matrix}\right]$$

$\neq 0,$ even for white $v[k]$!

- Therefore, the least-squares estimation error results in a solution that is biased.

$$\mathbb{E}[\hat{\theta}_e] \not\to \theta_{opt}$$

unless

1. $v[k] \equiv 0$ or
2. $a_i = 0$ for $i = 1 \dots n$ (FIR) *and* $v[k]$ is white.

EXAMPLE: $G(z) = \dfrac{b}{z-a}$. $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$.

- Assume $u[k]$ is zero-mean white noise with variance σ_u^2 and $v[k]$ is zero-mean white noise with variance σ_v^2.
- So,

$$\bar{y}[k] = a\bar{y}[k-1] + bu[k-1]$$

$$y[k] - v[k] = a\left(y[k-1] - v[k-1]\right) + bu[k-1]$$

so that

$$\begin{aligned} y[k] &= ay[k-1] + bu[k-1] + v[k] - av[k-1] \\ &= \begin{bmatrix} y[k-1] & u[k-1] \end{bmatrix}\theta + e_e[k] \end{aligned}$$

$$= a_e[k]\theta + e_e[k].$$

- The expected asymptotic estimate $\hat{\theta}$ is

$$\hat{\theta} = \mathbb{E}\left[a_e^T[k]a_e[k]\right]^{-1}\mathbb{E}\left[a_e^T[k]y[k]\right]$$

where

$$\mathbb{E}\left[a_e^T[k]a_e[k]\right] = \mathbb{E}\left[\begin{array}{cc} y^2[k-1] & y[k-1]u[k-1] \\ y[k-1]u[k-1] & u^2[k-1] \end{array}\right] = \left[\begin{array}{cc} \sigma_y^2 & 0 \\ 0 & \sigma_u^2 \end{array}\right]$$

and

$$\mathbb{E}\left[a_e^T[k]y[k]\right] = \mathbb{E}\left[\begin{array}{c} y[k]y[k-1] \\ y[k]u[k-1] \end{array}\right] = \mathbb{E}\left[\begin{array}{c} a\sigma_y^2 - a\sigma_v^2 \\ b\sigma_u^2 \end{array}\right].$$

- Then,

$$\hat{\theta} = \left[\begin{array}{c} a(1-\sigma_v^2/\sigma_y^2) \\ b \end{array}\right] = \theta_{opt} + \text{bias}.$$

- We can express this bias term as a function of SNR.