

Midterm Review

Leah Dickstein

July 17, 2014

Contents

1	2014-06-03 Channel Estimation	2
1.1	Problem Statement:	2
1.2	Solution:	2
1.3	Multiple Copies of Y	2
2	Another $L[X—Y]$	2
3	Notes from Kalman Filter Wikipedia	2
4	2014-06-05	2
4.1	Q1	2
4.2	Q2	3
4.3	Q3	3
4.4	Q4	3
4.5	Q5	3
5	Notes from EE126 Appendix A	3
6	Proofs about $L[X—Y]$	3
6.1	$L[X—Y,Z] = L[X—Y] + L[X—Z]$	3
6.2	$L[X—Y,Z] = L[X—Y] + L[X—Z-L[Z—Y]]$	3
7	2014-06-09 2014-06-15 Kalman Filter	3
7.1	Problem Setup:	3
7.2	Goal:	3
7.3	Equations:	3
7.4	$\mathbb{E}[X[n+1] Y^{n-1}]$	4
7.5	$\mathbb{E}[Y[n] Y^{n-1}]$	4
7.6	$\mathbb{E}[X[n+1] Y[n] - C\hat{X}[n]]$	4
8	2014-06-16 Underlying X1 and X2	5
8.1	Problem Setup	5
8.2	Solution	5

9	2014-06-18 Conditions of Observability	5
10	2014-06-19 Various Proofs	5
10.1	Best Control	5
10.2	Error of Control Problem	5
10.3	Without System Error, Estimation Error = 0	6
11	2014-06-29 Kalman Filter with Multiplicative Noise	6
11.1	Problem Setup:	6
11.2	Goal:	6
11.3	Equations:	6
11.4	$\mathbb{E}[X[n] Y^{n-1}]$	6
11.5	$\mathbb{E}[Y[n] Y^{n-1}]$	6
11.6	$\mathbb{E}[X[n] Y[n] - CM(n)A\hat{X}[n-1]]$	6
12	Notes from A Mathematical Theory of Communication	7
12.0.1	Introduction	7
12.1	Discrete Noiseless Systems	8
12.1.1	The Discrete Noiseless Channel	8
12.1.2	The Discrete Source of Information	8
12.1.3	The Series of Approximations to English	8
12.1.4	Graphical Representation of a Markoff (Markov?) Process	8
12.1.5	Ergodic and Mixed Sources	8
12.1.6	Choice, Uncertainty and Entropy	9
12.1.7	The Entropy of an Information Source	9
12.1.8	Representations of the Encoding and Decoding Operations	10
12.1.9	The Fundamental Theorem for a Noiseless Channel	10
12.1.10	Discussion and Examples	10
12.2	Discrete Noisy Systems	10
12.2.1	Representation of a Noisy Discrete Channel	10
12.2.2	Equivocation and Channel Capacity	10
12.2.3	The Fundamental Theorem for a Discrete Channel with Noise	10
12.2.4	Discussion	10
12.2.5	Example of a Discrete Channel and its Capacity	10
12.2.6	The Channel Capacity in Certain Special Cases	10
12.2.7	An Example of Efficient Coding	10
13	Results of Simulations	11
13.1	Kalman Filter Additive Noise	11
13.2	Kalman Filter Ignoring Multiplicative Noise	12
13.3	Kalman Filter for Multiplicative Noise	15
13.4	Quantization Noise: Schenato	17
13.5	NonCoherence Estimation: Gireeja	20
13.6	NonCoherence Control: Gireeja	22
13.7	Varying A, NonCoherence Control, Gireeja	23
13.8	Rajasekaran vs Schenato: Prediction vs Estimation	24

1 2014-06-03 Channel Estimation

1.1 Problem Statement:

$$X + Z = Y$$

$$X \sim N(0, \sigma^2)$$

$$Z \sim N(0, 1)$$

$$\mathbb{E}[X] = 0$$

$$\text{Var}[X] = 0$$

1.2 Solution:

$$f_{X|Y=y}(X) = \frac{f_X(X) * f_{Y=y|X}(Y)}{f_{Y=y}(Y=y)}$$

$$\text{Pdf} = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\text{Var}[Y|X] = 1$$

$$\text{Var}[Y] = \sigma^2 + 1$$

$$\text{Var}[X] = \sigma^2$$

$$\text{Var}[Z] = 1$$

$$f_{X|Y=y}(X) = \frac{\sqrt{\sigma^2 + 1}}{\sigma\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2} - \frac{x^2}{2\sigma^2} + \frac{y^2}{2(\sigma^2+1)}}$$

$$\frac{d}{dx} f_{X|Y=y}(X) = \frac{\sigma^2}{\sigma^2 + 1} Y$$

1.3 Multiple Copies of Y

for the Same Realization of X:

It's more effective because there is an averaging effect in the variance of the noise:

$$\text{Var}[Y|X] = \frac{1}{n}$$

$$\text{Var}[Y] = \sigma^2 + \frac{1}{n}$$

$$\text{Var}[X] = \sigma^2$$

$$\text{Var}[Z] = \frac{1}{n}$$

This assumes we are working with Gaussians.

2 Another L[X—Y]

$$\min \mathbb{E}[(\alpha Y + \beta - X)^2]$$

$$\frac{d}{d\alpha} \mathbb{E}[(\alpha Y + \beta - X)^2] = \mathbb{E}[2(\alpha Y + \beta - X)(Y)] = 0$$

$$= \mathbb{E}[\alpha Y^2 + \beta Y - XY] = 0$$

$$= \mathbb{E}[\alpha Y^2] = \mathbb{E}[XY - \beta Y]$$

$$= \alpha = \frac{\mathbb{E}[XY - \beta Y]}{\mathbb{E}[Y^2]}$$

$$= \alpha = \frac{\text{Cov}[XY]}{\text{Var}[Y]}$$

This assumes either $\beta = 0$ or $\mathbb{E}[Y] = 0$. The former we will soon show, and the latter is a result of Y being 0-mean, which is a result of X and Z being 0-mean.

$$\frac{d}{d\beta} \mathbb{E}[(\alpha Y + \beta - X)^2] = \mathbb{E}[2(\alpha Y + \beta - X)] = 0$$

$$= \mathbb{E}[\beta] = \mathbb{E}[X - \alpha Y]$$

$$= \beta = \mathbb{E}[X] - \alpha \mathbb{E}[Y]$$

$$= \beta = 0$$

$$L[X|Y] = \alpha Y + \beta = \frac{\text{Cov}[XY]}{\text{Var}[Y]} Y$$

This assumes we are using a linear estimator for random variables that **aren't necessarily** Gaussian, thus showing that for Gaussians the optimal estimator **is** the linear the estimator.

3 Notes from Kalman Filter Wikipedia

4 2014-06-05

4.1 Q1

$$x[n] = a * x[n-1]$$

$$x[0] \sim N(0, a^{2n})$$

The variance increases so that the probability of being larger remains the same.

4.2 Q2

$$\begin{aligned}x[n] &= a * x[n-1] + w[n-1] \\w[n-1] &\sim N(0, \sigma_w^2) \\x[n] &\sim N(0, \sigma_w^2 \sum_{i=0}^{n-1} a^{2i} + a^{2n})\end{aligned}$$

4.3 Q3

$$\begin{aligned}x[n] &= a * x[n-1] \\y[n] &= c * x[n] \\\hat{x}[n] &= 1/c * y[n]\end{aligned}$$

In this case, the observation is noiseless.

4.4 Q4

$$\begin{aligned}x[n] &= a * x[n-1] + w[n-1] \\y[n] &= c * x[n] \\\hat{x}[n] &= 1^* y[n]\end{aligned}$$

$W[n]$ doesn't matter because it's incorporated into the state, and we're trying to guess the state.

4.5 Q5

Estimate $x[n]$ using memory.

If the observations are noiseless, then memory doesn't matter since we have perfect observation anyway. If observations are noisy, over time the noise $\rightarrow 0$.

If you add noise to the observation, use the $L[X|Y]$ shown above.

5 Notes from EE126 Appendix A

6 Proofs about $L[X|Y]$

$$6.1 \quad L[X|Y, Z] = L[X|Y] + L[X|Z]$$

$$6.2 \quad L[X|Y, Z] = L[X|Y] + L[X|Z|L[Z|Y]]$$

7 2014-06-09 2014-06-15 Kalman Filter

7.1 Problem Setup:

$$\begin{aligned}X[n] &= AX[n-1] + W[n-1] \\Y[n] &= CX[n] + V[n]\end{aligned}$$

$$\begin{aligned}\mathbf{X} &\sim N(0, A^{2n} + \sigma_W^2 \sum_{i=0}^{n-1} A^{2i}) \\ \mathbf{Y} &\sim N(0, C^2(A^{2n} + \sigma_W^2 \sum_{i=0}^{n-1} A^{2i}) + \sigma_V^2)\end{aligned}$$

$$\begin{aligned}X(0) &\sim N(0, 1) \\W[n] &\sim N(0, \Sigma_W) \\V[n] &\sim N(0, \Sigma_V)\end{aligned}$$

7.2 Goal:

$$\begin{aligned}\mathbb{E}[X[n+1]|Y^n] &= \hat{X}[n+1] \\Y^n &= (Y[0] \dots Y[n]) \\\mathbb{E}[X[n+1]|Y^n] &= \sqcup \mathbb{E}[X[n]|Y^{n-1}] + \sqcup (Y[n] - \mathbb{E}[Y[n]|Y^{n-1}])\end{aligned}$$

7.3 Equations:

$$\begin{aligned}(1) \quad L[X|Y] &= \mathbb{E}[X] + \frac{\text{cov}(X, Y)}{\text{cov}(Y)}(Y - \mathbb{E}[Y]) \\(2) \quad L[X|Y, Z] &= L[X|Y] + L[X|Z - L[Z|Y]] \\(3) \quad \text{cov}(AX, CY) &= A \text{cov}(X, Y) C' \\(4) \quad \text{if } V, W \perp &\text{cov}(V + W) = \text{cov}(V) + \text{cov}(W)\end{aligned}$$

$$\mathbb{E}[X[n+1]|Y^n] = \mathbb{E}[X[n+1]|Y^{n-1}] + \mathbb{E}[X[n+1]|Y[n] - \mathbb{E}[Y[n]|Y^{n-1}]]$$

7.4 $\mathbb{E}[X[n+1]|Y^{n-1}]$

$$\begin{aligned}
(1) \quad & \mathbb{E}[AX[n] + W[n]|Y^{n-1}] = \mathbb{E}[AX[n]|Y^{n-1}] + \mathbb{E}[W[n]|Y^{n-1}] \\
(2) \quad & = A\hat{X}[n] + \mathbb{E}[W[n]] \\
(3) \quad & = A\hat{X}[n]
\end{aligned}$$

7.5 $\mathbb{E}[Y[n]|Y^{n-1}]$

$$\begin{aligned}
(4) \quad & \mathbb{E}[CX[n] + V[n]|Y^{n-1}] = C\mathbb{E}[X[n]|Y^{n-1}] + \mathbb{E}[V[n]|Y^{n-1}] \\
(5) \quad & = C\hat{X}[n]
\end{aligned}$$

7.6 $\mathbb{E}[X[n+1]|Y[n] - C\hat{X}[n]]$

$$\begin{aligned}
(6) \quad & \mathbb{E}[X[n+1]|Y[n] - C\hat{X}[n]] = \mathbb{E}[AX[n] + W[n]|Y[n] - C\hat{X}[n]] \\
(7) \quad & = \mathbb{E}[AX[n]|Y[n] - C\hat{X}[n]] \\
(8) \quad & = \mathbb{E}[AX[n] - A\hat{X}[n]|Y[n] - C\hat{X}[n]]
\end{aligned}$$

Lemma: $Y^{n-1} \perp Y[n] - \mathbb{E}[Y[n]|Y^{n-1}]$

Strong Induction!

Base Case: $\text{cov}(Y[0], Y[1] - \mathbb{E}[Y[1]|Y[0]]) = 0$

$$\begin{aligned}
(1) \quad & \mathbb{E}[y[0](cax[0] + cw[0] + v[1] - \frac{ac^2\Sigma_{x[0]}}{\Sigma_{y[0]}}y[0])] \\
(2) \quad & = \mathbb{E}[y[0](cax[0] - \frac{ac^2\Sigma_{x[0]}}{\Sigma_{y[0]}}y[0])] \\
(3) \quad & = \mathbb{E}[(cx[0] + v[0])cax[0] - \frac{ac^2\Sigma_{x[0]}}{\Sigma_{y[0]}}y^2[0]] \\
(4) \quad & = \mathbb{E}[c^2ax^2[0] - \frac{ac^2\Sigma_{x[0]}}{\Sigma_{y[0]}}y^2[0]] \\
(5) \quad & = c^2a\Sigma_{x[0]} - \frac{ac^2\Sigma_{x[0]}}{\Sigma_{y[0]}}\Sigma_{y[0]} \\
(6) \quad & = c^2a\Sigma_{x[0]} - ac^2\Sigma_{x[0]} = 0
\end{aligned}$$

Inductive Hypothesis: $\text{cov}(Y[n-1], Y[n] - \mathbb{E}[Y[n]|Y^{n-1}]) = 0 \wedge \dots \wedge \text{cov}(Y[0], Y[n] - \mathbb{E}[Y[n]|Y[0]]) = 0$

Inductive Step: $\text{cov}(Y[n], Y[n+1] - \mathbb{E}[Y[n+1]|Y[n]]) = 0$

$$\mathbb{E}[y[n](cax[n] - \frac{c^2a\Sigma_{x[n]}}{\Sigma_{y[n]}}y[n])] = c^2a\Sigma_{x[n]} - \frac{c^2a\Sigma_{x[n]}}{\Sigma_{y[n]}}\Sigma_{y[n]} = 0$$

In addition,

$$\begin{aligned}
(7) \quad & \forall t \leq n, \quad \text{cov}(y[t], y[n+1] - \mathbb{E}[y[n+1]|y[t]]) = 0 \\
(8) \quad & = \mathbb{E}[y[t](cay^{n+1-t}x[t] - \frac{c^2a^{n+1-t}\Sigma_{x[t]}}{\Sigma_{y[t]}}y[t])] \\
(9) \quad & = \mathbb{E}[c^2a^{n+1-t}x^2[t] - \frac{c^2a^{n+1-t}\Sigma_{x[t]}}{\Sigma_{y[t]}}y^2[t]] \\
(10) \quad & = c^2a^{n+1-t}\Sigma_{x[t]} - \frac{c^2a^{n+1-t}\Sigma_{x[t]}}{\Sigma_{y[t]}}\Sigma_{y[t]} = 0
\end{aligned}$$

If $t = n+1$, $\text{cov}(y[n+1], y[n+1] - \mathbb{E}[y[n+1]|y[n+1]]) = \text{cov}(y[n+1], y[n+1] - y[n+1]) = \text{cov}(y[n+1], 0) = 0$.

The answer is trivial.

$$\text{cov}(Y^{n-1}, Y[n] - \mathbb{E}[Y[n]|Y^{n-1}]) = \mathbb{E}\left[\begin{bmatrix} Y[0] \\ \vdots \\ Y[n-1] \end{bmatrix} \left(Y[n] - \frac{\text{cov}(Y[n], Y^{n-1})}{\text{cov}(Y^{n-1})} \begin{bmatrix} Y[0] \\ \vdots \\ Y[n-1] \end{bmatrix} \right)\right]$$

We have proved for $\forall t < n$ this = 0, therefore the answer is the 0 vector and we prove the Lemma. Since $\hat{X}[n] = \mathbb{E}[X[n]|Y^{n-1}]$, it is the projection of X onto Y^{n-1} . If $Y^{n-1} \perp \hat{Y}$, $\hat{X}[n] \perp \hat{Y}$. We can add if inside the cov() since it's equivalent to adding 0.

$$\text{cov}(AX[n] - A\hat{X}[n], CX[n] - C\hat{X}[n]) = A\text{cov}(X[n] - \hat{X}[n])C'$$

$$S_n = \text{cov}(X[n] - \hat{X}[n])$$

$$\begin{aligned}
& \text{cov}(Y[n] - C\hat{X}[n]) = \text{cov}(CX[n] + V[n] - C\hat{X}[n]) \\
& = \text{cov}(C(X[n] - \hat{X}[n])) + \text{cov}(V[n]) = CS_nC' + \sigma_v^2
\end{aligned}$$

$$K_n = \frac{AS_nC'}{CS_nC' + \sigma_v^2}$$

$$\hat{X}[n+1] = \mathbb{E}[X[n+1]|Y^n]$$

$$= A\hat{X}[n] + \frac{A\text{cov}(X[n] - \hat{X}[n])C'}{C\text{cov}(X[n] - \hat{X}[n])C' + \sigma_v^2} (Y[n] - C\hat{X}[n])$$

8 2014-06-16 Underlying X1 and X2 10 2014-06-19 Various Proofs

8.1 Problem Setup

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} * \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}$$

$$Y(n) = [1 \quad 1] * \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}$$

Calculate $x_1(n), x_2(n)$ from $y(n)$

Calculate $x_1(n), x_2(n)$ from $y(n), y(n-1)$

$$\text{Variations: } Y(n) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}$$

$$Y(n) = [0 \quad 1] * \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}$$

8.2 Solution

$$X[n] = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} * \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix}^{-1} * \begin{bmatrix} y(n-1) \\ y(n) \end{bmatrix}$$

$$X[n] = \begin{bmatrix} 6 & -2 \\ -6 & 3 \end{bmatrix} \begin{bmatrix} y(n-1) \\ y(n) \end{bmatrix}$$

In this case we showed that when dealing with X_1 and X_2 , system error matters. This is because instead of trying to estimate the fluctuating line, we are now trying to estimate the "straight line" where the system "should" go.

9 2014-06-18 Conditions of Observability

$$\begin{bmatrix} C \\ \vdots \\ CA^{\lfloor \frac{m}{n} \rfloor} \end{bmatrix} * \begin{bmatrix} x_1(n - \lfloor \frac{m}{n} \rfloor) \\ \vdots \\ x_n(n - \lfloor \frac{m}{n} \rfloor) \end{bmatrix} = \begin{bmatrix} y(n - \lfloor \frac{m}{n} \rfloor) \\ \vdots \\ y(n) \end{bmatrix}$$

Check that $\begin{bmatrix} C \\ \vdots \\ CA^{\lfloor \frac{m}{n} \rfloor} \end{bmatrix}$ is full rank, then delete

$\lfloor \frac{m}{n} \rfloor * n - m = -m \bmod n$ lines and solve for x .

$$\begin{bmatrix} C \\ \vdots \\ CA^{\lfloor \frac{m}{n} \rfloor} \end{bmatrix} \text{ must be full rank, or span } X.$$

10.1 Best Control

$$\begin{aligned} & \min \mathbb{E}[||x(n+1)||^2] \\ &= \min \mathbb{E}[||ax(n) + w(n) + u(n)||^2] \\ &= \min \mathbb{E}[||ax(n) + w(n) + \alpha y(n) + \beta||^2] \\ &= \min \mathbb{E}[||ax(n) + w(n) + \alpha cx(n) + \alpha v(n) + \beta||^2] \end{aligned}$$

$$\begin{aligned} \frac{d}{d\alpha} \mathbb{E}[] &= \mathbb{E}[2(ax(n) + w(n) + \alpha cx(n) + \alpha v(n) + \beta)(cx(n) + v(n))] = 0 \\ &= \mathbb{E}[acx^2(n) + \alpha(c^2x^2(n) + v^2(n))] = 0 \\ &= \alpha \mathbb{E}[c^2x^2(n) + v^2(n)] = -\mathbb{E}[acx^2(n)] \\ &= \alpha = \frac{-acVar(x)}{Var(Y)} \end{aligned}$$

$$\begin{aligned} \frac{d}{d\beta} \mathbb{E}[] &= \mathbb{E}[2(ax(n) + w(n) + \alpha cx(n) + \alpha v(n) + \beta)] = 0 \\ &= \mathbb{E}[\beta] = 0 \end{aligned}$$

$$u(n) = \frac{-acVar(X)}{Var(Y)} Y(n) = -L[X|Y]$$

10.2 Error of Control Problem

$$\begin{aligned} & \mathbb{E}[||x(n) + u(n)||^2] \\ &= \mathbb{E}[||x(n) - \frac{ac\Sigma_X}{\Sigma_Y}(cx(n) + v(n))||^2] \\ &= \mathbb{E}[||(1 - \frac{ac\Sigma_X}{\Sigma_Y})x(n) - \frac{ac\Sigma_X}{\Sigma_Y}v(n)||^2] \\ &\rightarrow \boxed{(1 - \frac{ac\Sigma_X}{\Sigma_Y})^2 \Sigma_X + (\frac{ac\Sigma_X}{\Sigma_Y})^2 \Sigma_V} \end{aligned}$$

10.3 Without System Error, Estimation Error = 0

In this system, we assume $a \leq 1$.

$$\begin{aligned}\Sigma_n &= \text{cov}(x(n) - \hat{x}(n)) \\ S_n &= A^2 \Sigma_{n-1} + \Sigma_W = A^2 \Sigma_{n-1} \\ \Sigma_n &= (1 - KnC)S_n = A^2(1 - KnC)\Sigma_{n-1} \\ &\rightarrow 0 < \left(1 - \frac{c^2 S_n}{c^2 S_n + \Sigma_V}\right) < 1 \\ \left(1 - \frac{c^2 S_n}{c^2 S_n + \Sigma_V}\right) < 1 &\rightarrow \frac{c^2 S_n}{c^2 S_n + \Sigma_V} > 0\end{aligned}$$

It is easy to see the error coefficient is greater than 0, because 1 - fraction > 0 . By definition $c \neq 0$, since the signal must have *some* power. By definition Σ_n starts out > 0 , thus $S_n > 0$, thus $\left(1 - \frac{c^2 S_n}{c^2 S_n + \Sigma_V}\right) < 1$ and $n \rightarrow \infty \implies \Sigma_n \rightarrow 0$.

11 2014-06-29 Kalman Filter with Multiplicative Noise

11.1 Problem Setup:

$$\begin{aligned}X[n] &= AX[n-1] + BW[n-1] \\ Y[n] &= U[n]CX[n] + V[n]\end{aligned}$$

$$\begin{aligned}X(0) &\sim N(0, S(0)) \\ W[n] &\sim N(0, \Sigma_W) = N(0, Q(n)) \\ V[n] &\sim N(0, \Sigma_V) = N(0, R(n)) \\ U(n) &\sim N(\mu, \sigma^2) = N(M(n), N(n))\end{aligned}$$

11.2 Goal:

$$\begin{aligned}\hat{X}[n] &= \mathbb{E}[X[n]|Y^n] \\ Y^n &= (Y[0] \dots Y[n]) \\ \mathbb{E}[X[n]|Y^n] &= \sqcup \mathbb{E}[X[n]|Y^{n-1}] + \sqcup (Y[n] - \mathbb{E}[Y[n]|Y^{n-1}])\end{aligned}$$

It is necessary and sufficient: $\tilde{X}[n] = X[n] - \hat{X}[n] \perp Y^n$ or $Y[n] - Y^{n-1}$.

11.3 Equations:

$$\begin{aligned}(11) \quad L[X|Y] &= \mathbb{E}[X] + \frac{\text{cov}(X, Y)}{\text{cov}(Y)}(Y - \mathbb{E}[Y]) \\ (12) \quad L[X|Y, Z] &= L[X|Y] + L[X|Z] - L[Z|Y] \\ (13) \quad \text{cov}(AX, CY) &= A\text{cov}(X, Y)C' \\ (14) \quad \text{if } V, W \perp \text{cov}(V + W) &= \text{cov}(V) + \text{cov}(W)\end{aligned}$$

$$\mathbb{E}[X[n]|Y^n] = \mathbb{E}[X[n]|Y^{n-1}] + \mathbb{E}[X[n]|Y[n] - \mathbb{E}[Y[n]|Y^{n-1}]]$$

11.4 $\mathbb{E}[X[n]|Y^{n-1}]$

$$\begin{aligned}(1) \quad \mathbb{E}[AX[n-1] + BW[n-1]|Y^{n-1}] &= \mathbb{E}[AX[n-1]|Y^{n-1}] + \mathbb{E}[BW[n-1]|Y^{n-1}] \\ (2) \quad &= A\hat{X}[n-1] + B\mathbb{E}[W[n-1]] \\ (3) \quad &= A\hat{X}[n-1]\end{aligned}$$

11.5 $\mathbb{E}[Y[n]|Y^{n-1}]$

$$\begin{aligned}(4) \quad \mathbb{E}[U[n]CX[n] + V[n]|Y^{n-1}] &= \mathbb{E}[U[n]CX[n]|Y^{n-1}] \\ (5) \quad &= C\mathbb{E}[U[n]|Y^{n-1}]\mathbb{E}[X[n]|Y^{n-1}] \\ (6) \quad &= C\mathbb{E}[U[n]] * A\hat{X}[n-1] = CM(n)\hat{X}[n-1]\end{aligned}$$

11.6 $\mathbb{E}[X[n]|Y[n] - CM(n)A\hat{X}[n-1]]$

$$\text{Note: } P(n) = \text{cov}(X[n] - A\hat{X}[n-1]) = \mathbb{E}[X^2[n] - 2X[n]A\hat{X}[n-1] + A^2\hat{X}^2[n-1]]$$

Numerator:

$$(7) \quad cov(X[n], U[n]CX[n] + V[n] - CM[n]A\hat{X}[n-1])$$

$$(8) \quad = cov(X[n], C(U[n]X[n] - M[n]A\hat{X}[n-1]))$$

$$(9) \quad = cov(X[n] - A\hat{X}[n-1], C(U[n]X[n] - M[n]A\hat{X}[n-1]))$$

$$(10) \quad = C\mathbb{E}[U[n]X^2[n] - U[n]AX[n]\hat{X}[n-1] - M[n]X[n]A\hat{X}[n-1] + M[n]A^2\hat{X}^2[n-1]]$$

$$(11) \quad = C(\mathbb{E}[U[n]]\mathbb{E}[X^2[n]] - \mathbb{E}[U[n]]\mathbb{E}[AX[n]\hat{X}[n-1]] - M[n]\mathbb{E}[X[n]A\hat{X}[n-1]] + M[n]\mathbb{E}[A^2\hat{X}^2[n-1]])$$

$$(12) \quad = C(M[n]\mathbb{E}[X^2[n]] - M[n]\mathbb{E}[AX[n]\hat{X}[n-1]] - M[n]\mathbb{E}[X[n]A\hat{X}[n-1]] + M[n]\mathbb{E}[A^2\hat{X}^2[n-1]])$$

$$(13) \quad = CM[n](\mathbb{E}[X^2[n]] - \mathbb{E}[AX[n]\hat{X}[n-1]] - \mathbb{E}[X[n]A\hat{X}[n-1]] + \mathbb{E}[A^2\hat{X}^2[n-1]])$$

$$(14) \quad = CM[n]P[n]$$

Denominator:

$$(15) \quad cov(U[n]CX[n] + V[n] - M[n]CA\hat{X}[n-1])$$

$$(16) \quad = cov(C(U[n]X[n] - M[n]A\hat{X}[n-1])) + cov(V[n])$$

$$(17) \quad = C^2 cov(U[n]X[n] - M[n]X[n] + M[n]X[n] - M[n]A\hat{X}[n-1]) + R[n]$$

$$(18) \quad = C^2 cov((U[n] - M[n])X[n] + M[n](X[n] - A\hat{X}[n-1])) + R[n]$$

$$(19) \quad = C^2 \mathbb{E}[|(U[n] - M[n])X[n] + M[n](X[n] - A\hat{X}[n-1])|^2] + R[n]$$

$$(20) \quad = C^2 \mathbb{E}[(U[n] - M[n])^2 X^2[n] + 2(U[n] - M[n])X[n]M[n](X[n] - A\hat{X}[n-1]) + M^2[n](X[n] - A\hat{X}[n-1])^2] + R[n]$$

$$(21) \quad = C^2 \mathbb{E}[(U[n] - M[n])^2 X^2[n] + M^2[n](X[n] - A\hat{X}[n-1])^2] + R[n]$$

$$(22) \quad = C^2 (\mathbb{E}[U^2[n] - 2U[n]M[n] + M^2[n]]X^2[n] + M^2[n]\mathbb{E}[(X[n] - A\hat{X}[n-1])^2]) + R[n]$$

$$(23) \quad = C^2 (\mathbb{E}[U^2[n] - M^2[n]]X^2[n] + M^2[n]P[n]) + R[n]$$

$$(24) \quad = C^2 (\mathbb{E}[N(n) * X^2[n]] + M^2[n]P[n]) + R[n]$$

$$(25) \quad = C^2 (N[n]S[n] + M^2[n]P[n]) + R[n]$$

$$Kf = \frac{M[n]P[n]C}{C^2 N[n]S[n] + C^2 M^2[n]P[n] + R[n]}$$

$$\hat{X}[n] = \mathbb{E}[X[n]|Y^n] = A\hat{X}[n-1] + Kf(Y[n] - CM(n)A\hat{X}[n-1])$$

12 Notes from A Mathematical Theory of Communication

12.0.1 Introduction

- Similar to the coding learned in CS70 (Hamming, RSA, Error Correcting Codes), an important aspect of communication is being able to *distinguish* messages. Meaning of messages doesn't matter, but the code for the messages must be a distance apart

- "The significant aspect is that the actual message is one *selected from a set* of possible messages"
- Log function used to measure information produced when a message is chosen, 3 reasons
 1. Useful: Time, bandwidth, etc. vary linearly with $\log(\text{number of possibilities})$
 2. Intuitive: Doubling the possibilities also doubles the amount of information when using the log function
 3. Mathematically suitable: it helps the math work out
- Converting to bits is easy: $\log_2 M = \log_{10} M / \log_{10} 2 = 3.32 \log_{10} M$
- Five parts to the communication system:
 1. Information Source
 2. Transmitter
 3. Channel
 4. Receiver
 5. Destination
- Three general types of communication systems: Discrete, continuous and mixed

12.1 Discrete Noiseless Systems

12.1.1 The Discrete Noiseless Channel

- The capacity C of a discrete channel $C = \lim_{T \rightarrow \infty} \frac{\log N(T)}{T}$
- **Theorem 1:** Let $b_{ij}^{(s)}$ be the duration of the s^{th} symbol which is allowable in state i and leads to state j .
 $C = \log W$ where W is the largest real root of:
 $|\sum_s W^{-b_{ij}^{(s)}} - \delta_{ij}| = 0$

12.1.2 The Discrete Source of Information

- We want to give the shortest codes to the letters/words/phrases with the highest probability, to minimize bits that need to be sent across
- This system that produces a sequence of symbols based on probability = stochastic process; a stochastic process that does this = discrete source

- Simple to Complicated: Choosing letters independently, choosing letters based on probability, choosing letters based on transition probability (based on what the previous letter was), choosing letters based on previous two letters (trigram), choosing letters n -gram, choosing words independently, etc.

12.1.3 The Series of Approximations to English

- Zero-order approximation = Symbols independent and equiprobable
- First-order = Symbols independent but with frequencies of English text
- Second-order = Digram structure
- Third-order = Trigram
- First-order word approximation
- Second-order word approximation
- "Note that these samples have reasonably good structure out to about twice the range that is taken into account in their construction" e.g. If the process ensures reasonable text for two-letter sequences, what actually results is usually reasonable four-letter sequences

12.1.4 Graphical Representation of a Markoff (Markov?) Process

Markoff process = information source if a letter is produced for each transition between states

?? "The states will correspond to the 'residue of influence' from preceding letters"

12.1.5 Ergodic and Mixed Sources

- Ergodic process = statistical homogeneity = as the lengths of the sequences increase, the probabilities approach limits independent of the sequence
- Two properties necessary and sufficient:
 1. The graph doesn't have two isolated parts A and B that can't reach each other

2. A closed series of lines in the graph = "circuit"; number of lines = "length" of the "circuit". The GCD of the lengths of all circuits = 1 \implies no periodicity
- If a condition is violated, separate the graph into sub-graphs that satisfy the conditions. The source = "mixed" made of pure components $L = p_1 L_1 + p_2 L_2 + p_3 L_3 + \dots$, with p_i = the probability you start in any of the sub-graphs (since isolated, you would never leave the sub-graph)
- "Except when the contrary is stated we shall assume a source to be ergodic. This assumption enables one to identify averages along a sequence with averages over the ensemble of possible sequences."
- ?? P_i = probability of state i , $p_i(j)$ = probability of transition from i to j ; for process to be stationary equilibrium conditions must be satisfied: $P_j = \sum_i P_i p_i(j)$

12.1.6 Choice, Uncertainty and Entropy

- At what rate is information produced? How much "choice" is involved in the selection of the event or of how uncertain we are of the outcome? Measure of property $H(p_1, p_2, \dots, p_n)$ needs to satisfy three properties:
 1. H should be continuous in p_i
 2. If all the p_i are equal ($p_i = \frac{1}{n}$), H is a monotonic increasing function of n (more choices = more uncertainty)
 3. If a choice is split into two successive choices, H = weighted sum of individual values of H (each choice should be factored in)
- **Theorem 2:** $H = -K \sum_{i=1}^n p_i \log p_i$
- $H(X)$ = entropy for random variable X , but X isn't argument to function it acts like a label
- **INSERT FIGURE 7**
- More properties:
 1. $H = 0$ iff all p_i but one = 0, meaning there's only one choice and no uncertainty
 2. H is max and = to $\log(n)$ if uniform distribution (all $p_i = \frac{1}{n}$)

3. $H(x, y) \leq H(x) + H(y)$ The uncertainty of joint is less than/equal to sum of individual uncertainties
4. Any change toward equalization of the probabilities increases H , especially any "averaging" operation
5. $p_i(j) = \frac{p(i,j)}{\sum_j p(i,j) = \frac{p(i,j)}{p(i)}}$ This looks just like stuff done in CS70
Conditional entropy $H_x(y) = -\sum_{i,j} p(i,j) \log p_i(j)$
 $\implies H(x, y) = H(x) + H_x(y)$ Joint entropy = Uncertainty in X + what we don't know about Y given X (like innovation)
6. $H(y) \geq H_x(y)$ They are equal if X and Y are independent; knowing X can only provide information so can only decrease uncertainty/entropy

12.1.7 The Entropy of an Information Source

- Entropy per second $H = \sum_i f_i H_i$ where F_i is average frequency of state i
- **Theorem 3:** When N is large, $H \approx \frac{\log 1/p}{N}$ very close! MEANS entropy = normalization of $-\log p$; the amount of information we gain is related to how low a probability of the sequence occurring
- **Theorem 4:** $\lim_{N \rightarrow \infty} \frac{\log n(q)}{N} = H$ where $n(q)$ = number of messages we must take from the set in order of decreasing probability to have total probability q of those taken; applying Theorem 3 to subsets of sequences
- **Theorem 5:** Let $p(B_i)$ = probability of sequence B_i , let $G_N = -\frac{1}{N} \sum_i p(B_i) \log p(B_i)$ summing over all sequences B_i containing N symbols, then G_N is a monotonic decreasing function of N and $\lim_{N \rightarrow \infty} G_N = H$; This basically restates Theorem 2 but defines K for large N and applies it to longer sequences
- **Theorem 6:** Let $p(B_i, S_j)$ be probability of B_i followed by symbol S_j , let $F_N = -\sum_{i,j} p(B_i, S_j) \log p_{B_i}(S_j)$, summing over total length N , F_N is monotonic decreasing in N and:
 $G_N = \frac{1}{N} \sum_{n=1}^N F_n \mid F_N \leq G_N \mid \lim_{N \rightarrow \infty} F_N = H$
- F_N is the entropy of the N^{th} order approximation of the information source; F_N is the conditional entropy of the next symbol when $(N-1)$ are known, G_N is the entropy per symbol of blocks of N symbols

- Relative entropy = ratio of entropy of source to its max value with same symbols = measure of max compression possible
- redundancy = 1 - relative entropy
- Max entropy = max possible states/symbols/choices = max rate of transmission; redundancy = "safety check" of codes to help prevent corruption

- **IMPORTANT/RELEVANT** "In general, ideal or nearly ideal encoding requires a long delay in the transmitter and receiver. In the noiseless case which we have been considering, the main function of this delay is to allow reasonably good matching of probabilities to corresponding lengths of the sequences."
- Maximum entropy based on statistical conditions determines channel capacity

12.1.8 Representations of the Encoding and Decoding Operations

Probably don't understand this part well enough

- Transducer = encoding/decoding info at transmitter/receiver
- **Theorem 7:** The output of a finite state transducer driven by a finite state statistical source is a finite state statistical source, with entropy (per unit time) \leq to that of the input. If the transducer is non-singular they are equal
- Transducer (aka encoding/decoding) can only decrease entropy/uncertainty
- **Theorem 8:** Let the system of constraints considered a channel have capacity $C = \log W$. If $p_{ij}^{(s)} = \frac{B_i}{B_j} W^{-l_{ij}^{(s)}}$, where $l_{ij}^{(s)}$ is duration of s^{th} symbol from i to j and $B_i = \sum_{s,j} B_j W^{l_{ij}^{(s)}}$ then H is maximized and $= C$.

12.1.9 The Fundamental Theorem for a Noiseless Channel

- **Theorem 9:** Let a source have entropy H (bits per symbol) and channel have capacity C (bits per second), then it's possible to go up to an average rate of information $\frac{C}{H}$ but not possible to exceed that
- **QUESTION MARK REREAD THIS**

12.1.10 Discussion and Examples

- "The source as seen from the channel through the transducer should have the same statistical structure as the source which maximizes the entropy of the channel"

12.2 Discrete Noisy Systems

12.2.1 Representation of a Noisy Discrete Channel

12.2.2 Equivocation and Channel Capacity

12.2.3 The Fundamental Theorem for a Discrete Channel with Noise

12.2.4 Discussion

12.2.5 Example of a Discrete Channel and its Capacity

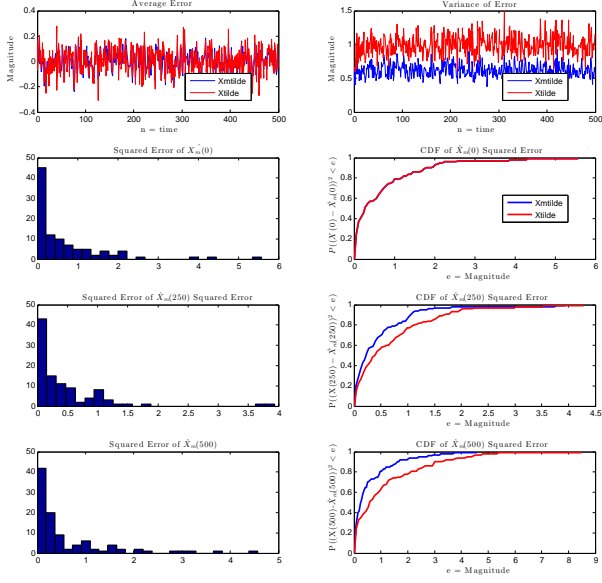
12.2.6 The Channel Capacity in Certain Special Cases

12.2.7 An Example of Efficient Coding

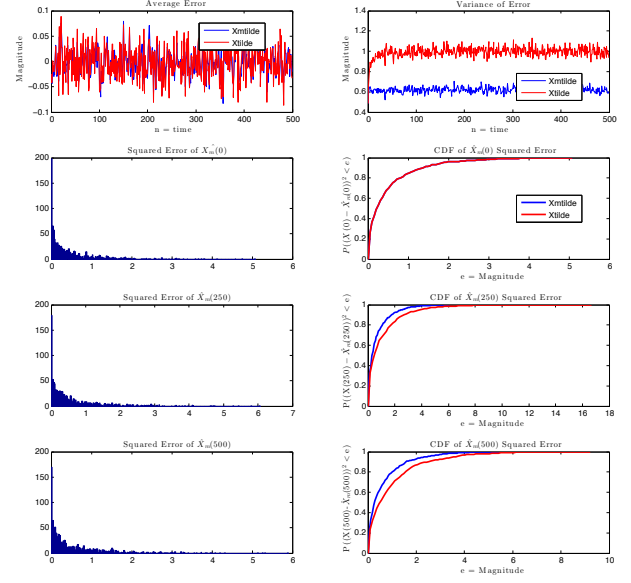
13 Results of Simulations

13.1 Kalman Filter Additive Noise

A = 1; V = 1; W = 1; M = 100



A = 1; V = 1; W = 1; M = 1000



These two figures represent a Kalman Filter with Additive Noise estimating a system with only additive noise. X_{tilde} represents a memoryless estimate, while X_{mtilde} represents the Kalman Filter's estimate.

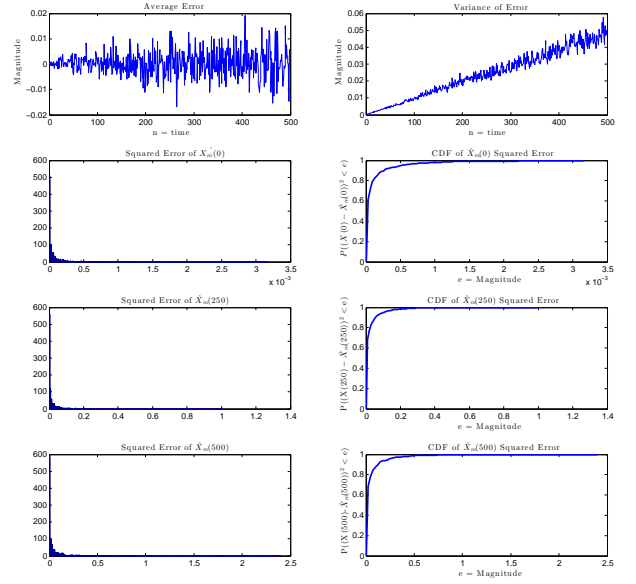
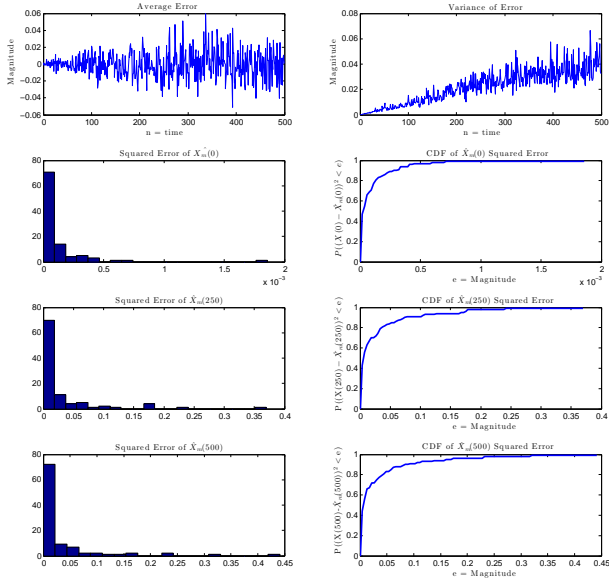
- The estimation error variance is bounded
- The Kalman Filter performs better than the optimal memoryless estimator; over many trials it's clearer the error variance is lower
- If $0 < A < 1$, $X_{mtilde} \rightarrow X_{tilde}$. We showed earlier that the Kalman Filter is only intended for an estimation system, not a control system, and as the value of the state converges to 0 the estimation system behaves like the control problem. Since the value of the state is so close to 0 at each timestep, memory provides no additional benefit/utility to estimation.
- The CDF of estimation error is affected by C, V, and W

Note: CDF of $\hat{X}_m(n)$ Squared Error means the plots are of the estimation error.

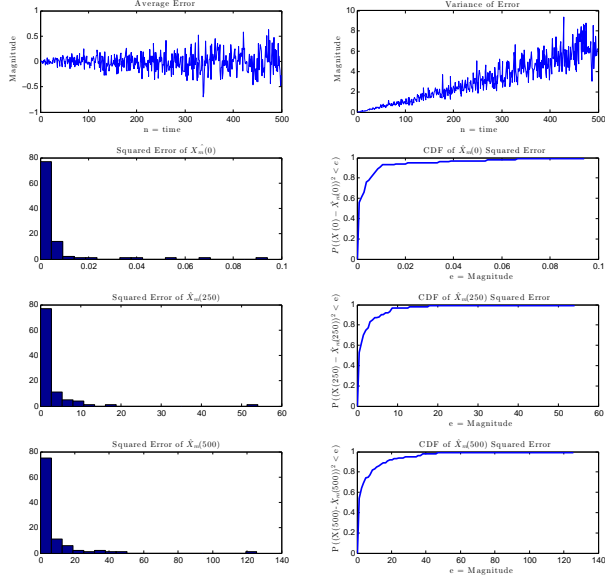
13.2 Kalman Filter Ignoring Multiplicative Noise

A = 1; V = 0.01; W = 1; M = 100

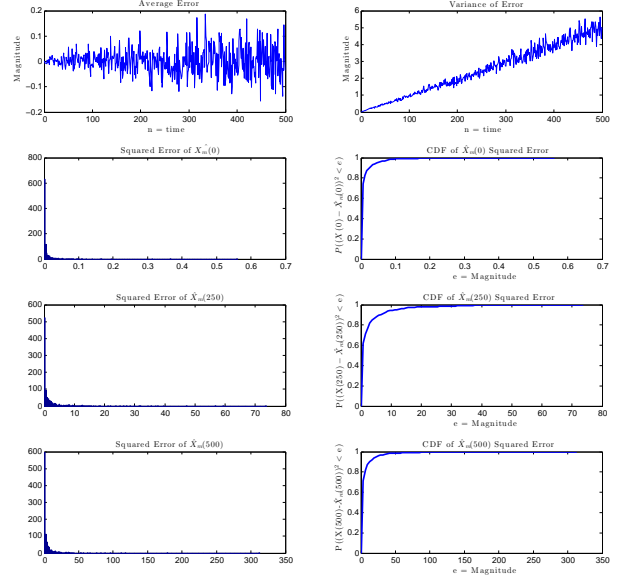
A = 1; V = 0.01; W = 1; M = 1000



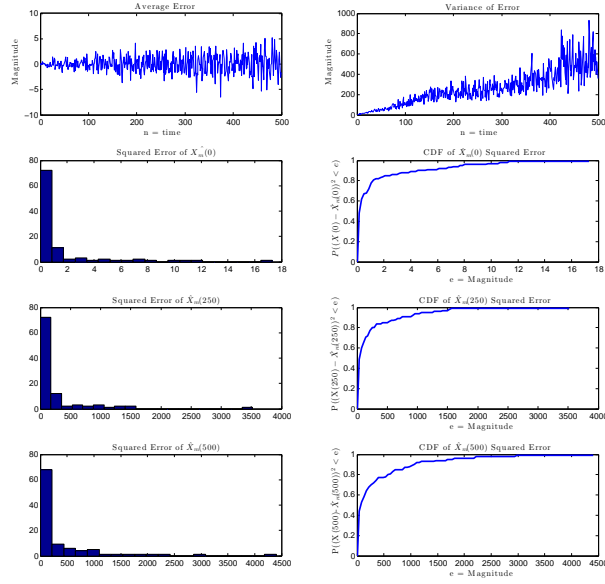
A = 1; V = 0.1; W = 1; M = 100



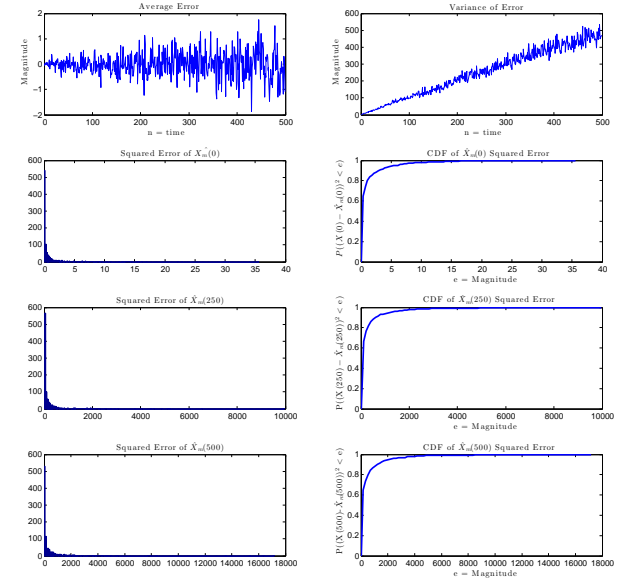
A = 1; V = 0.1; W = 1; M = 1000



A = 1; V = 1; W = 1; M = 100

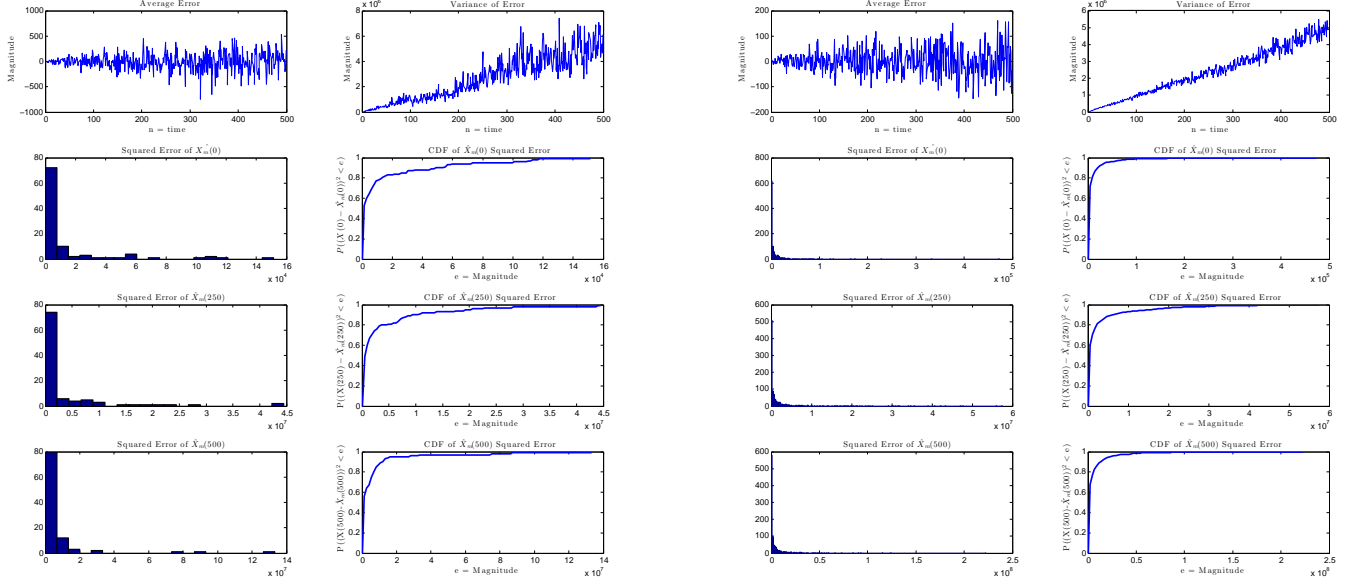


A = 1; V = 1; W = 1; M = 1000



A = 1; V = 100; W = 1; M = 100

A = 1; V = 100; W = 1; M = 1000

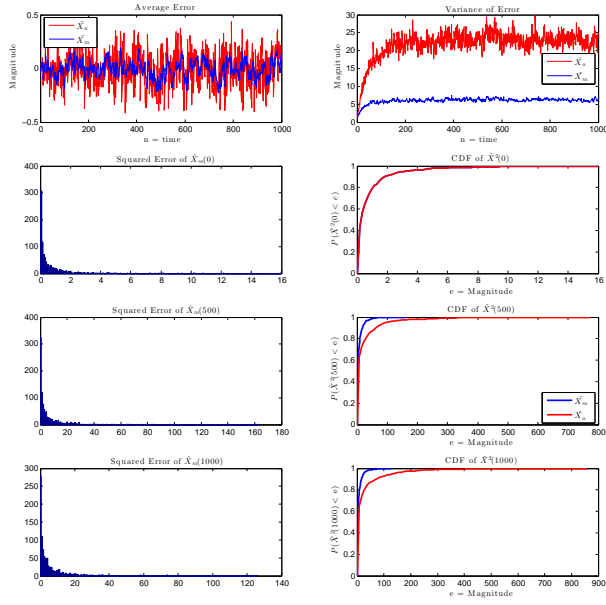


These figures represent a Kalman Filter for Additive Noise with a system that has state multiplicative noise and additive noise but no observation noise. The estimation error variance steadily increases because the multiplicative noise causes the signal to increase, thus increasing the error from the multiplicative noise. In this setup, the multiplicative error was $(1 + r(n)) = (1 + \text{normrnd}(0, \text{varv}))$. The four figures represent ascending levels of V (or $\frac{1}{p}$), with the estimation error variance increasing as V increases.

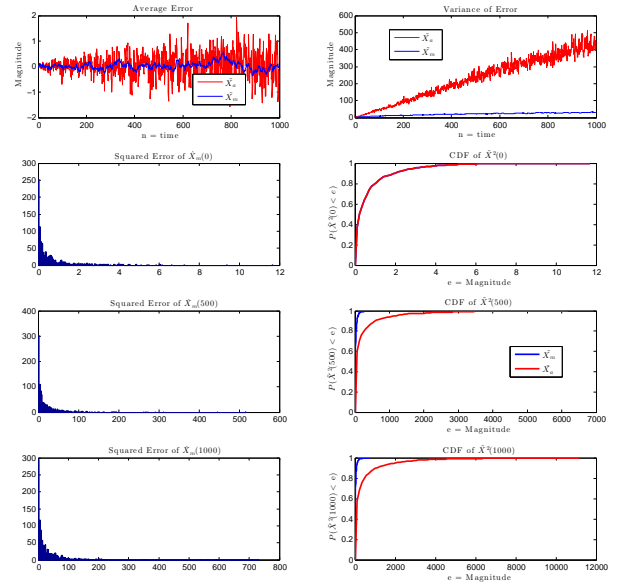
Note: CDF of $\hat{X}_m(n)$ Squared Error means the plots are of the estimation error.

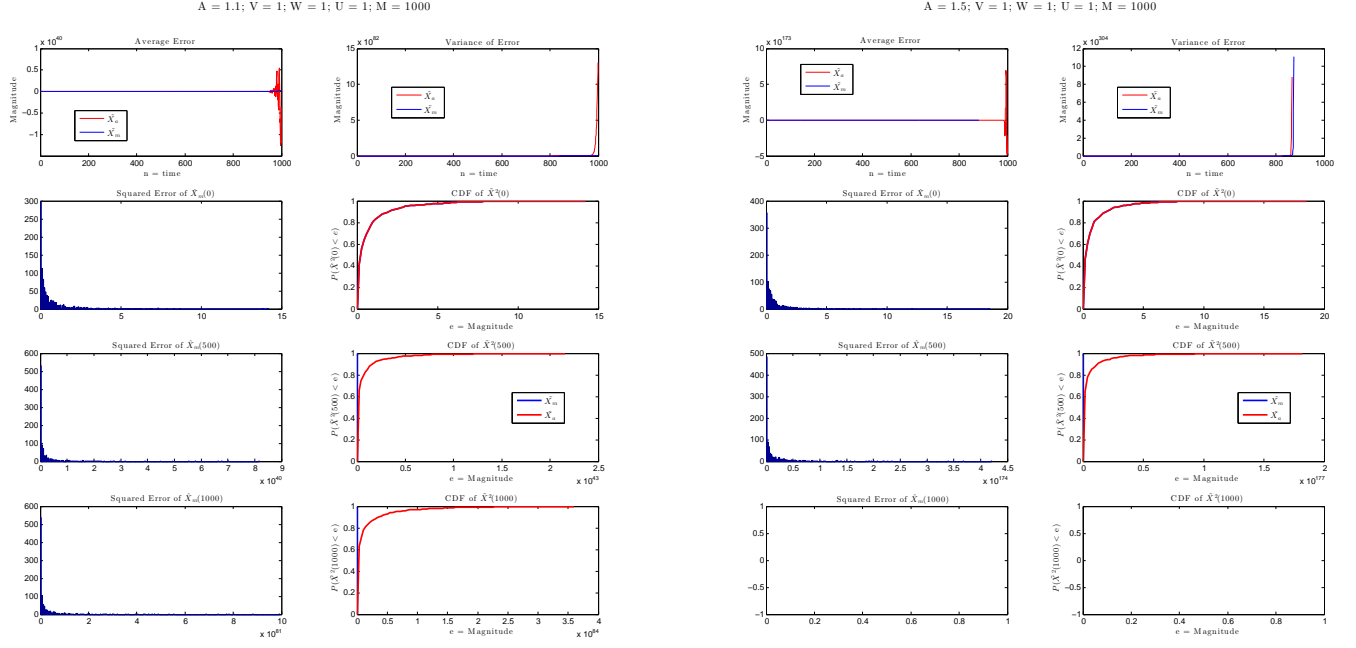
13.3 Kalman Filter for Multiplicative Noise

$A = 0.99; V = 1; W = 1; U = 1; M = 1000$



$A = 1; V = 1; W = 1; U = 1; M = 1000$



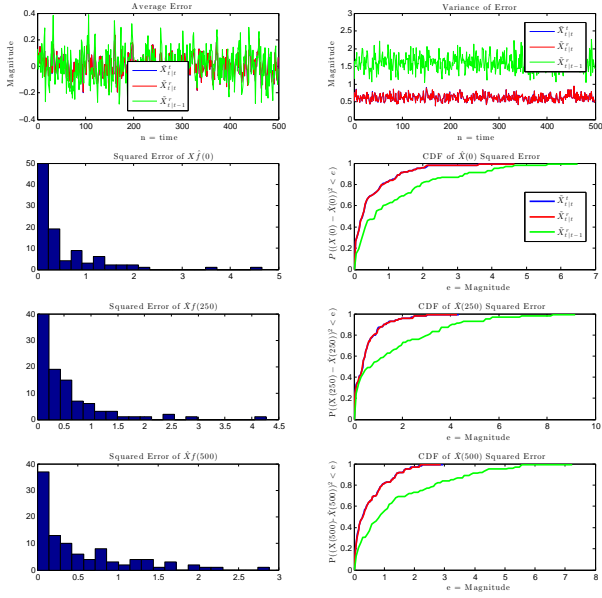


This is the implementation of a Kalman Filter for Multiplicative Noise (Rajasekaran) with a system that has both multiplicative and additive noise. \tilde{X}_a represents the estimation error for a Kalman Filter for only additive noise, while \tilde{X}_m represents the new Kalman Filter. It is apparent the new Kalman Filter is doing much better; both have bounded error variances, but \tilde{X}_m is significantly lower.

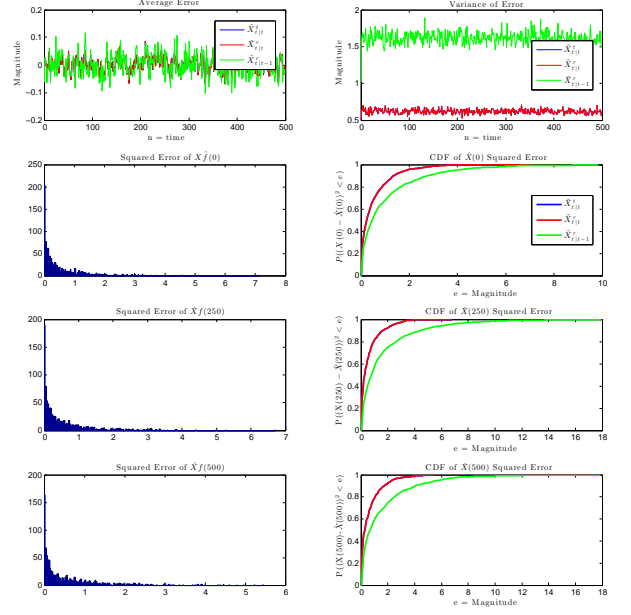
In the upper left corner, $A = 0.99$. Since $A < 1$, the state converges toward 0 and the estimation error variance for X_a is bounded. That being said, $\text{Var}(\tilde{X}_m)$ is still better (bounded at a lower value) than $\text{Var}(\tilde{X}_a)$. In the upper right, $A = 1$. $\text{Var}(\tilde{X}_a)$ matches the behavior we would expect, which is increasing linearly similar to the previous "Ignoring Multiplicative Noise with $A = 1$ " plots above. $\text{Var}(\tilde{X}_m)$ is bounded, again as we would expect since we're using a Kalman Filter intended for multiplicative noise. In the lower left, $A = 1.1$. In this case, you can see the estimation error variance explodes, with $\text{Var}(\tilde{X}_a)$ increasing up to 10^{82} . $\text{Var}(\tilde{X}_m)$, on the other hand, remains at its low bounded value. In the lower right, $A = 1.5$. The system is growing too fast over $n = 1000$ timesteps, and Matlab can't handle the numbers because they're getting too large. Both estimation error variances appear large because of value truncation/roundoff error done in Matlab. This simply means values $A > 1.5$ aren't testable in Matlab, even if Rajasekaran's Kalman Filter applies.

13.4 Quantization Noise: Schenato

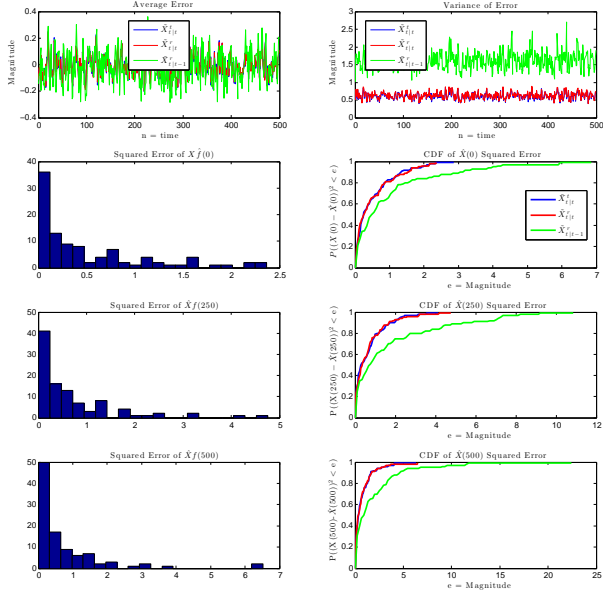
A = 1; V = 1; W = 1; P = 100; M = 100



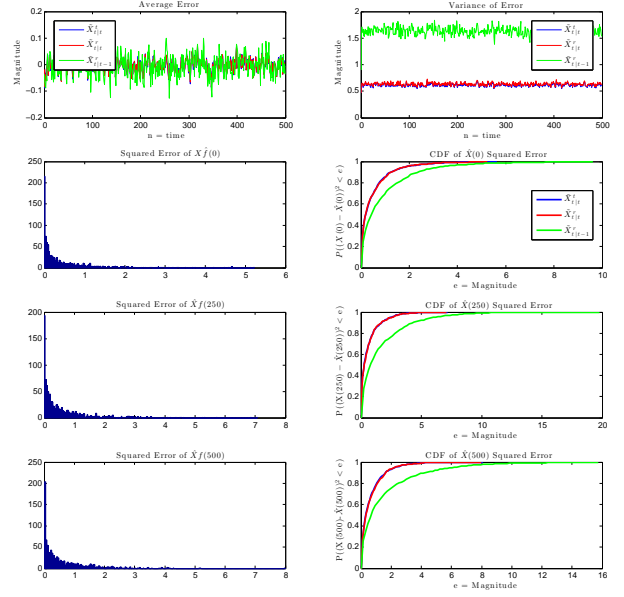
A = 1; V = 1; W = 1; P = 100; M = 1000



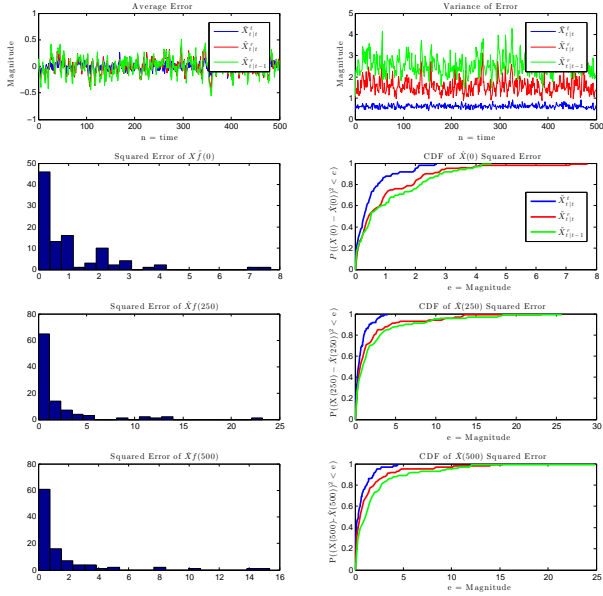
A = 1; V = 1; W = 1; P = 10; M = 100



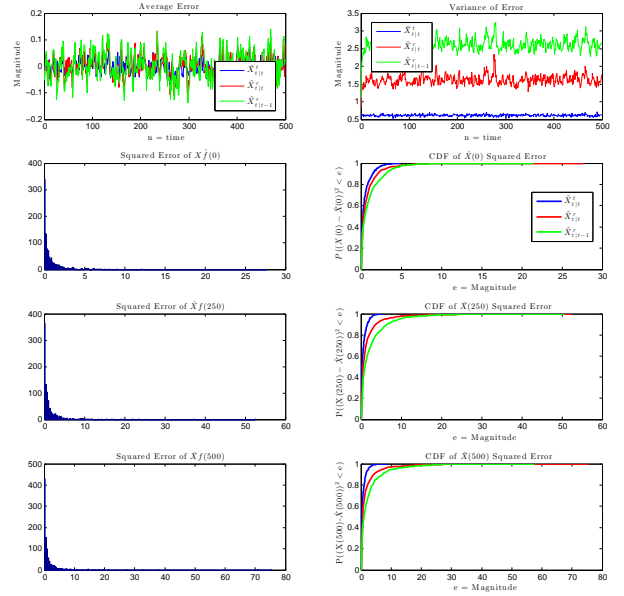
A = 1; V = 1; W = 1; P = 10; M = 1000

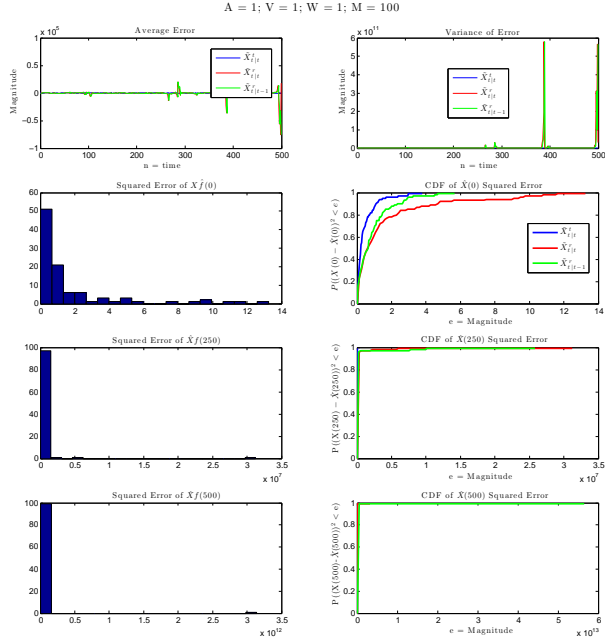


A = 1; V = 1; W = 1; P = 1; M = 100



A = 1; V = 1; W = 1; P = 1; M = 1000

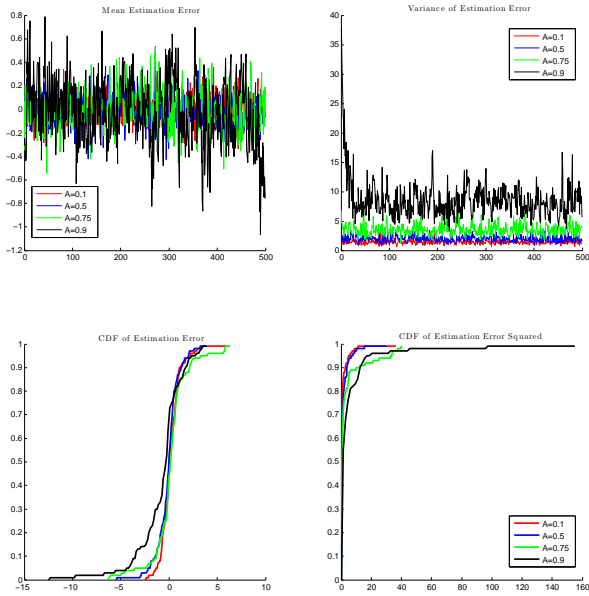




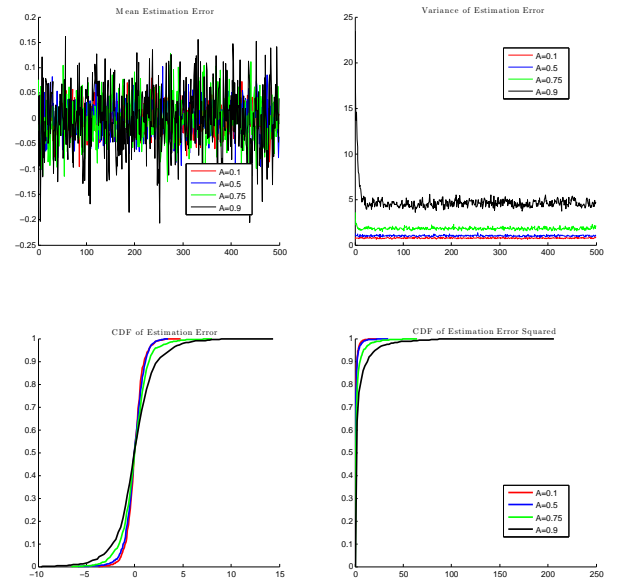
These figures are the implementation of Schenato's paper, with multiplicative noise but no packet drops. Multiplicative (quantization) noise happens over the channel between the transmitter and the receiver. $\tilde{X}_{t|t}^t$ represents the transmitter's estimate of the state, using a Kalman Filter for additive noise (because so far there has only been additive noise.) $\tilde{X}_{t|t-1}^r$ represents a prediction of the state post-quantization noise, and $\tilde{X}_{t|t}^r$ represents the estimation of the state post-quantization noise. When the power is extremely small (thus the multiplicative noise has large variance), $\tilde{X}_{t|t-1}^r$ does better than $\tilde{X}_{t|t}^r$. This is because the multiplicative noise has such a huge effect, the prediction with the moderate growth A is closer to the real state than the totally inaccurate noisy signal.

13.5 NonCoherence Estimation: Gireeja

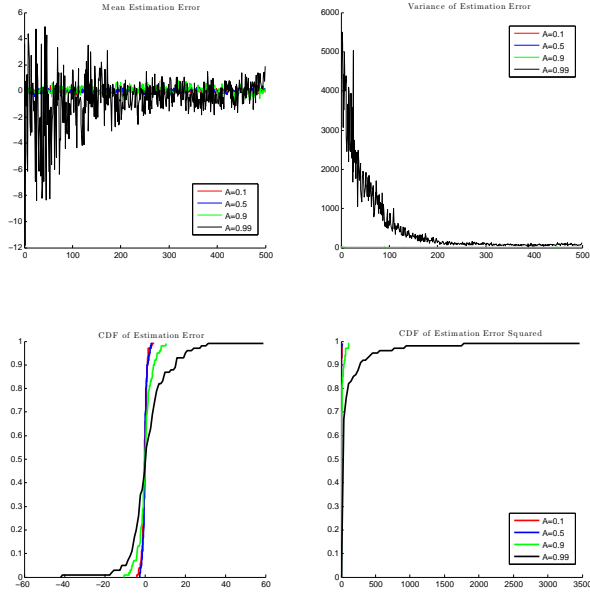
Gireeja NonCoherence Paper Varying A M = 100



Gireeja NonCoherence Paper Varying A M = 1000

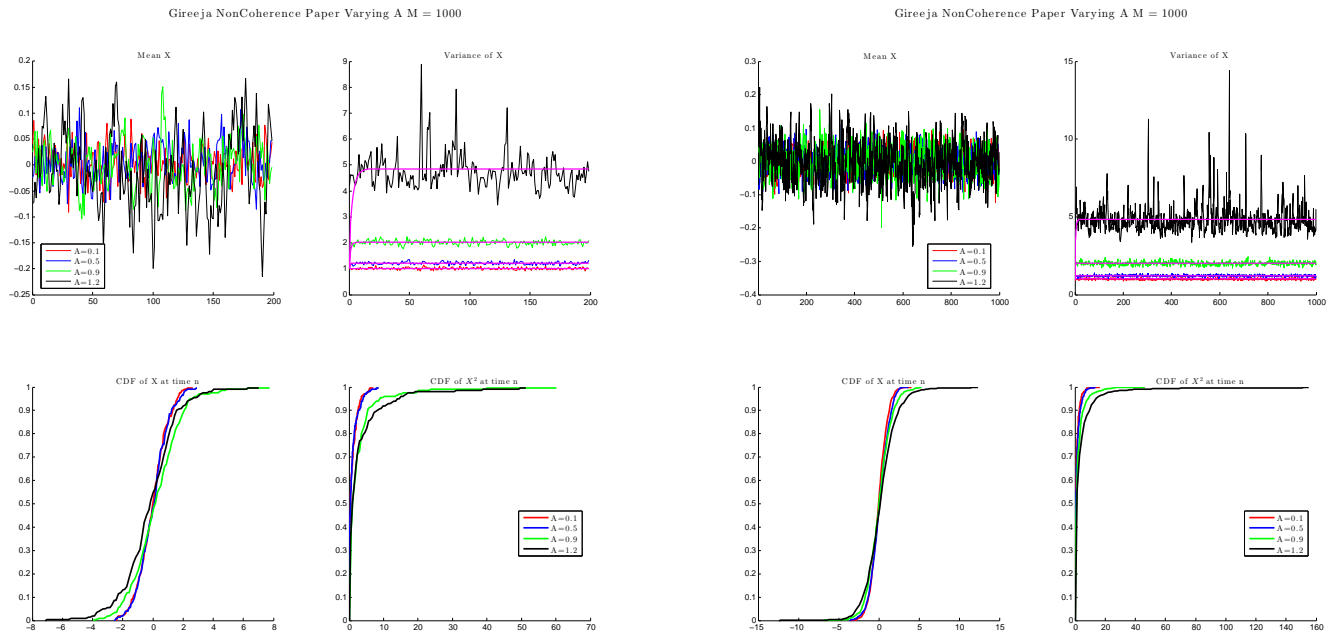


Gireeja NonCoherence Paper Varying A $M = 100$



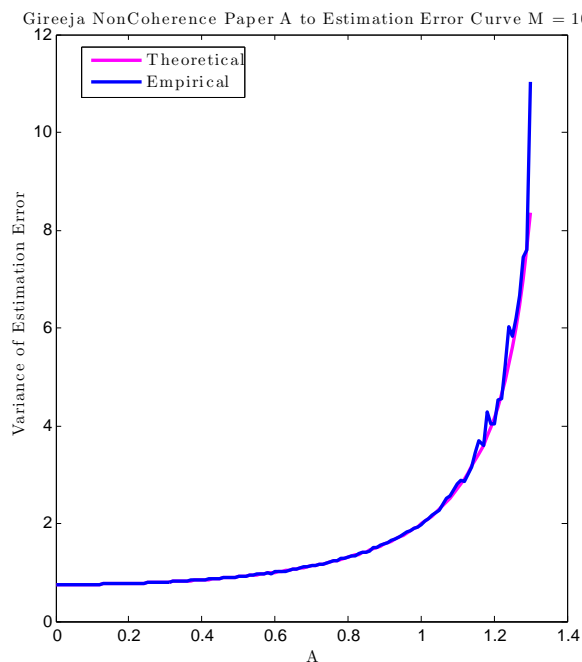
These figures are the implementation of Gireeja's Non-Coherence paper for estimation, not control. The figures above show that A affects the final bounded value of the estimation error variance. The figure to the left shows why A must be < 1 ; when A is close to 1 it takes longer to converge towards its asymptotic estimation error variance, and when $A \geq 1$ the system will explode.

13.6 NonCoherence Control: Gireeja



This figure is Gireeja's NonCoherence Paper Control System, and it shows the system is stabilized for varying levels of $A < 1.4$. Magenta represents the theoretical calculation for $\text{Var}(X)$, and this figure reflects that that calculation is accurate. Over time, $\text{Var}(X)$ quickly converges to $\text{Var}(X)^\infty$. In addition, the greater the A the higher the state variance, and this can be calculated.

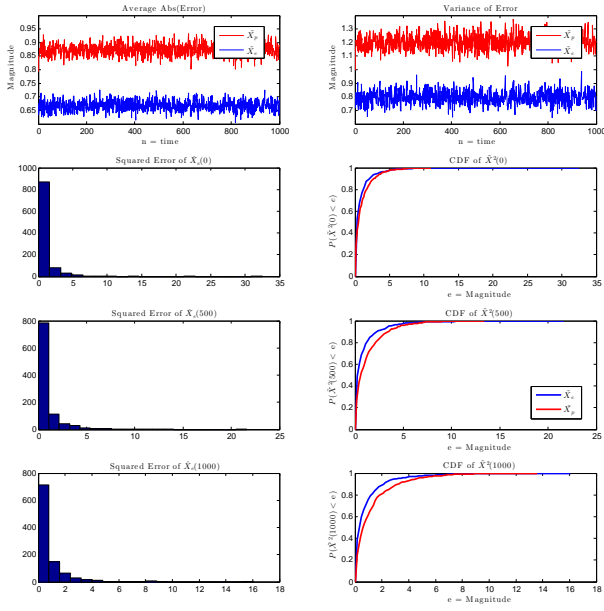
13.7 Varying A, NonCoherence Control, Gireeja



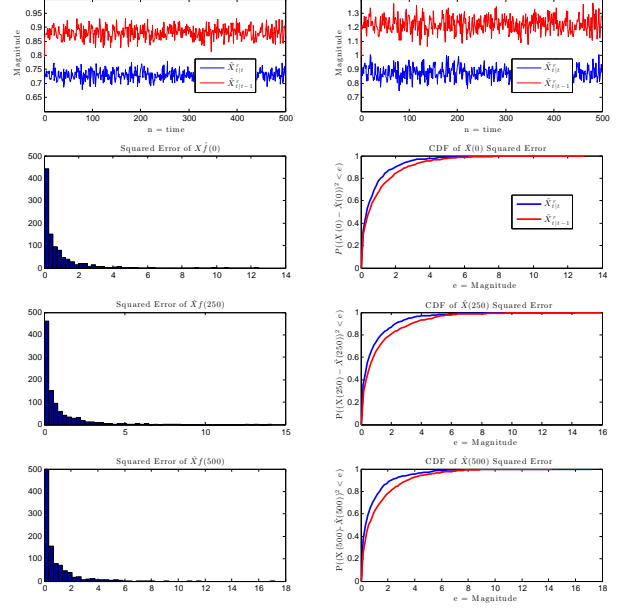
THIS NEEDS TO BE UPDATED!! In this figure, blue represents the asymptotic empirical variance for levels of A 0:0.01:1.3, while magenta represents the asymptotic theoretical variance. These values are $\text{mean}(\text{mean}())$, or averaged over both trials and timesteps. The curve appears roughly exponential, or at the very least when $A > 1$ the curve rises sharply. Based on Gireeja's paper, for these values the curve should asymptote at $\sqrt{2} = 1.414$, which is reflected in the curve. Since this curve was generated over $n = 250$ timesteps and $M = 1000$ trials, the spikes in the blue curve should *not* be attributed to randomness. Multiple runs could confirm whether those spikes are significant or not.

13.8 Rajasekaran vs Schenato: Prediction vs Estimation

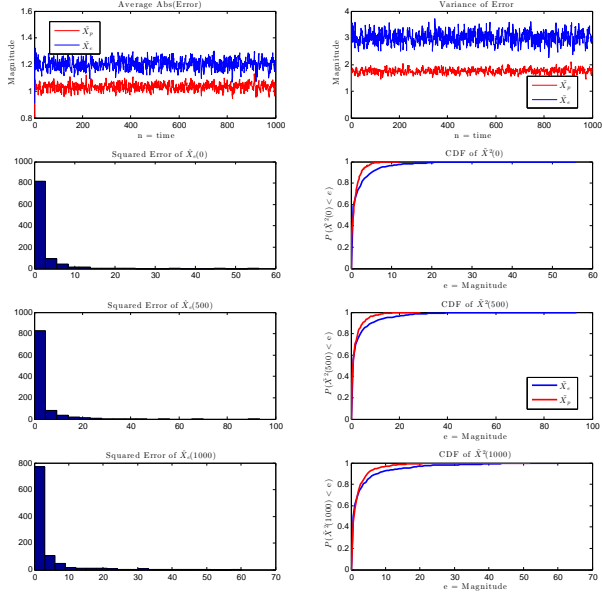
A = 0.5; V = 1; W = 1; U = 1; M = 1000



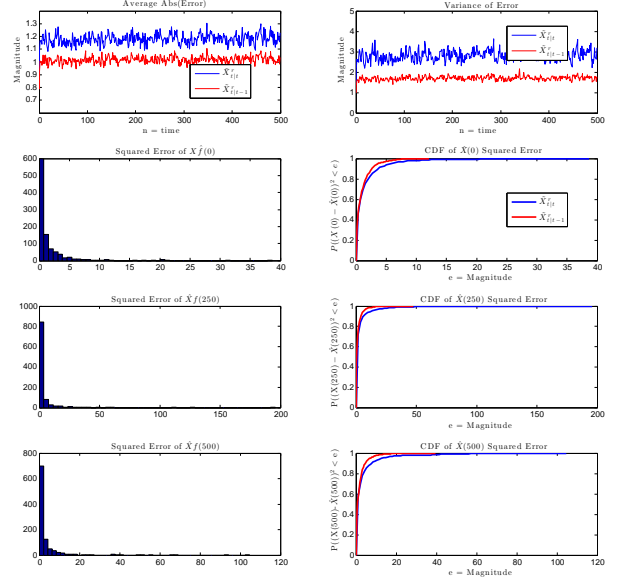
A = 0.5; V = 1; W = 1; P = 1; M = 1000



A = 0.5; V = 100; W = 1; U = 100; M = 1000



A = 0.5; V = 1; W = 1; P = 0.01; M = 1000



Left = Rajasekaran's Kalman Filter, Right = Schenato's Receiver Estimators.
In all four figures, red represents prediction and blue represents estimation.