Engr207b Final Exam

There are 4 questions on this exam. You have 24 hours! The usual rules apply; you may use any software and any sources you like, but you must work on your own. You may not discuss the exam with anyone else, or work with anyone else, or communicate with anyone else about the exam. Good luck!

1. *Inventory system.* A warehouse can hold n widgets. We will use $n = 20$. We model the inventory level by a Markov chain, with state $x(t)$ equal to the current inventory level. At time $t = 0$, we have $x(0) = n$ and the warehouse is full.

   At time t, if the inventory level $x(t) \leq 2$, then the warehouse is refilled and so $x(t+1) = n$. *If we do not refill the warehouse*, then customers may buy products. With probability 0.2 we sell 1 widget, with probability 0.1 we sell 2 widgets, and with probability 0.7 we sell nothing. This demand is independent of the inventory level.

   We have two sensors installed in this warehouse. Sensor 1 returns y_1, which is either 1 or 2. If $x(t) > 6$, it always returns 1. If $x(t) \leq 6$ then it returns 2 with probability 0.8, and 1 otherwise. (If we interpret 1 as `no` and 2 as `yes`, then it has a 20% false-negative probability, and a zero false-positive probability.)

   Sensor 2 returns y_2, which is 2 if $x(t) \geq 12$ with a 20% false negative rate. If $x(t) < 12$ it returns 1 with a 20% false positive rate.

   We will work on the time interval $[0, T]$ where $T = 200$.

   (a) What is the transition matrix P for this markov chain?

   (b) What are the transition matrices for sensors y_1 and y_2?

   (c) Simulate the Markov chain to find $x(0), \ldots, x(T)$. Plot x as a function of time.

   (d) Compute the distribution of $x(t)$ at times $t = 50, 100, 150, 200$. Plot each of the four distributions as a bar graph.

   (e) You are given the sequences $y_1(t)$ and $y_2(t)$ for $t \in [0, T]$ in the data file `invsensors.txt` (cut and paste into your favorite language.) Use the Kalman filter to compute the posterior distribution $p_{t|t}$. At each t, represent $p_{t|t}$ by a vector in $\mathbb{R}^n$, so that its i'th component is

   $$(p_{t|t})_i = \text{Prob}(x(t) = i \mid y_1(0), \ldots, y_1(t), y_2(0), \ldots, y_2(t))$$

   Plot a bar graph of $p_{t|t}$ at each of the times $t = 15, 20, 25, 30, 35, 40, 45, 50$.

   Solution.

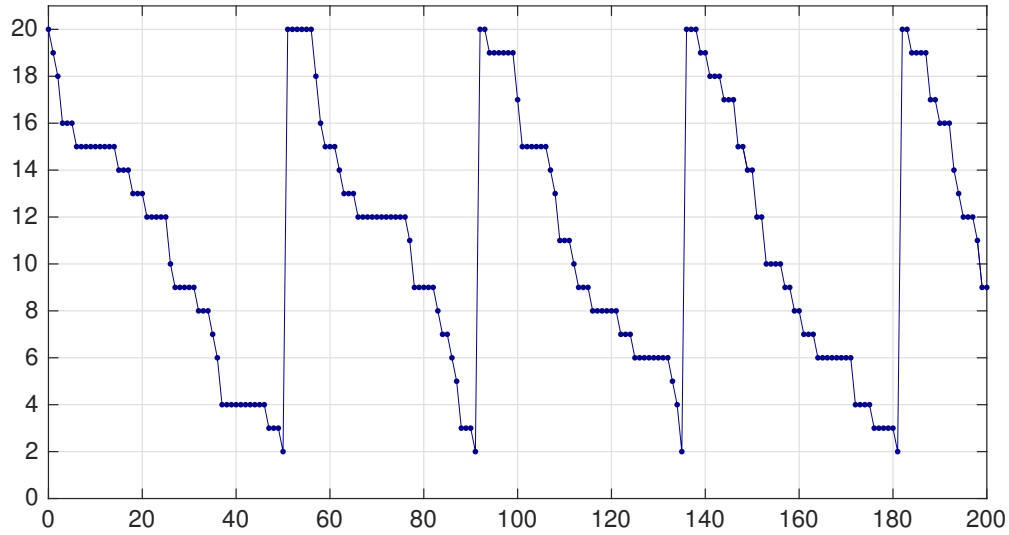
(a) The transition matrix is

$$P = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7
\end{bmatrix}$$

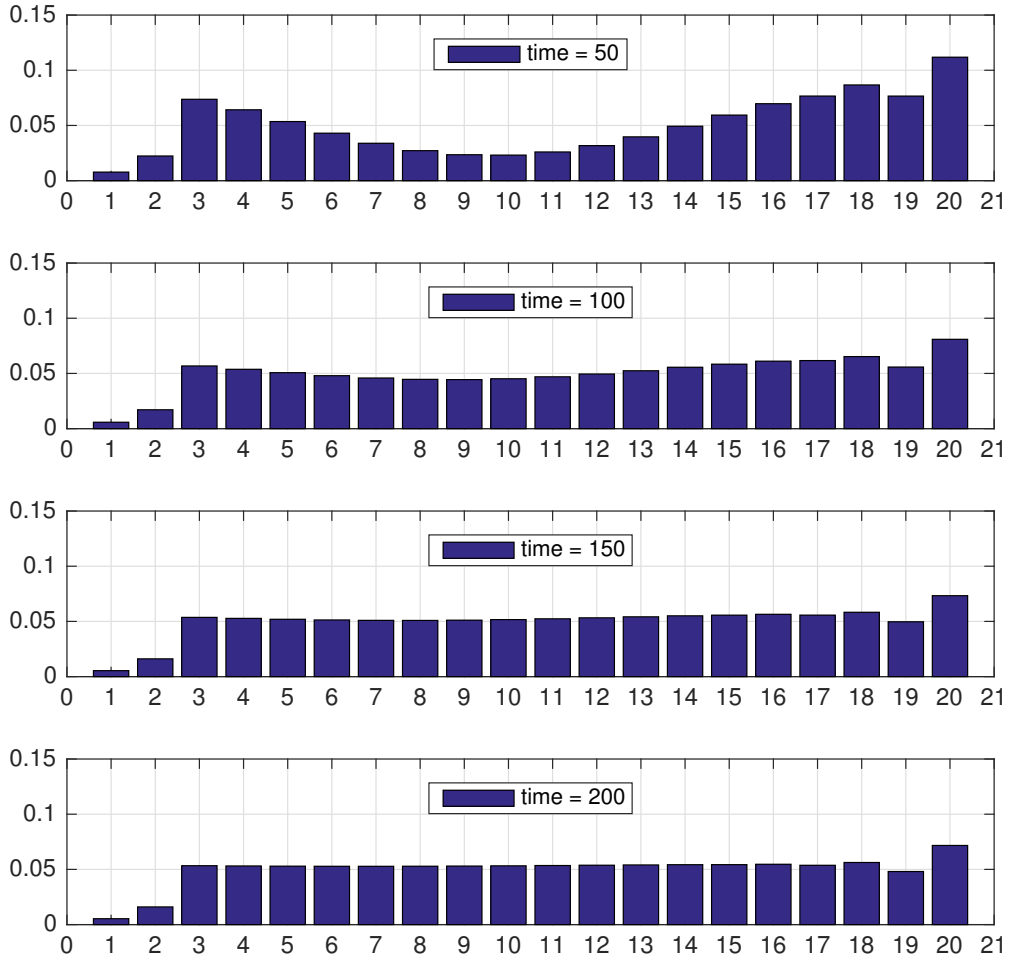
(b) We have

$$G_1 = \begin{bmatrix}
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0
\end{bmatrix} \qquad G_2 = \begin{bmatrix}
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.8 & 0.2 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8 \\
0.2 & 0.8
\end{bmatrix}$$

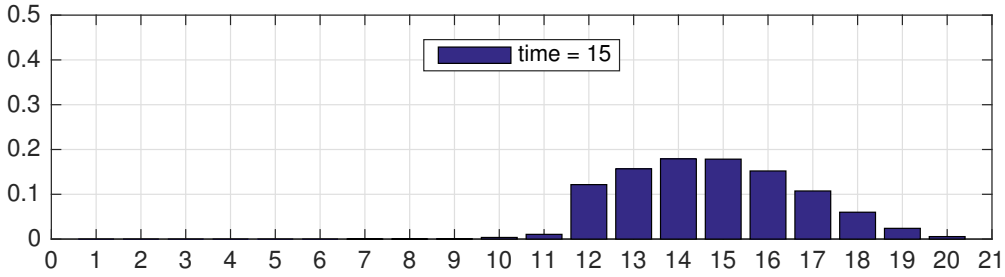
(c) A typical run is shown below.

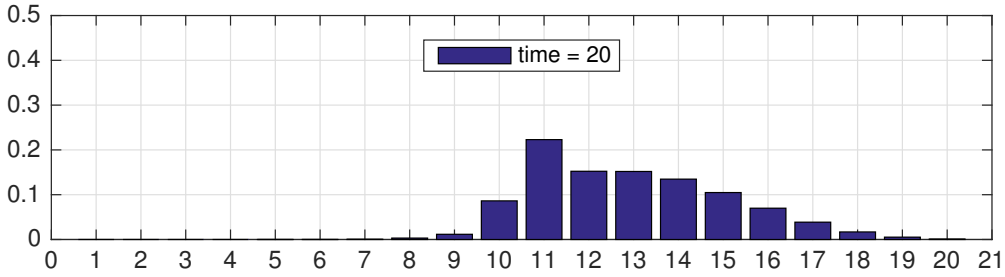


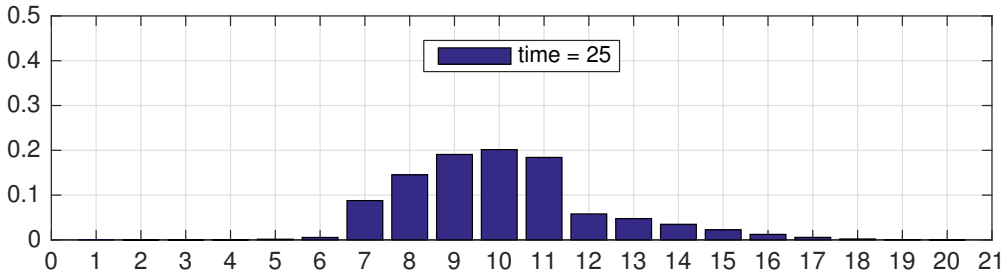
(d) The distributions are below.

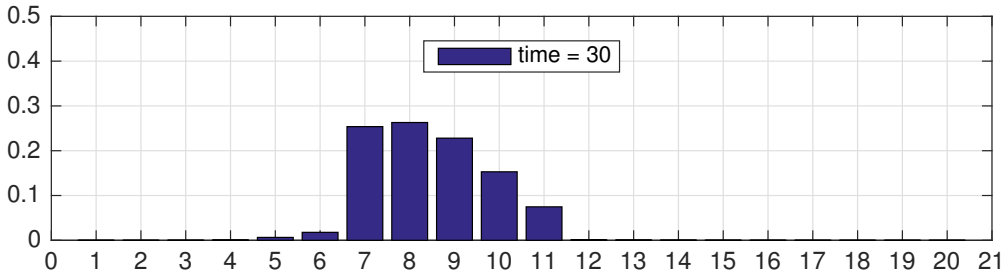


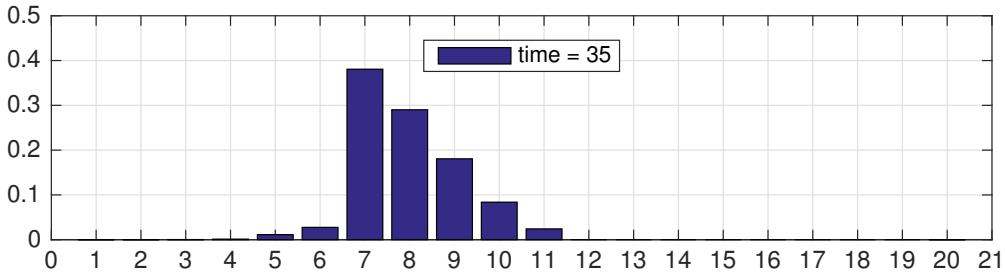
(e) The distributions are below.

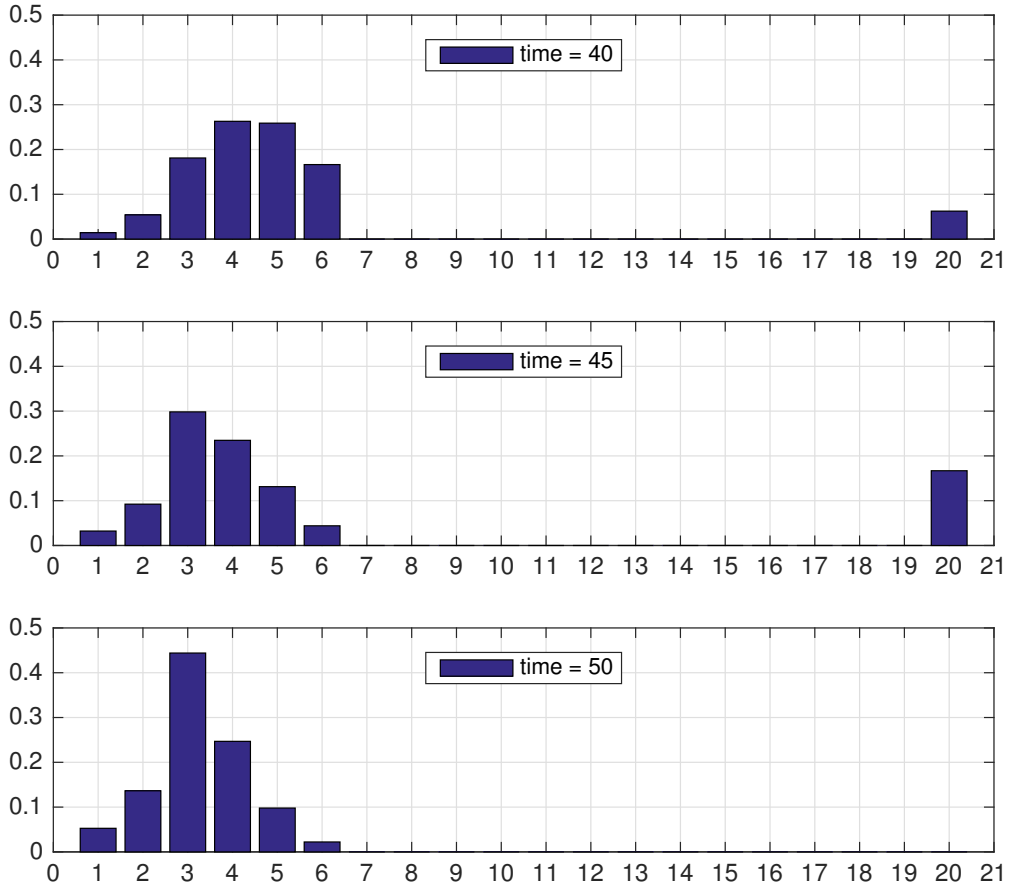

time = 15
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21


time = 20
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21


time = 25
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21


time = 30
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21


time = 35
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21



2. *Estimating production from serial numbers.* War is raging, and we (the good guys) have captured n enemy tanks. We want to estimate how many tanks the enemy has made in total. Call this quantity x. We don't have a very good idea of this, but we'll assume x has a uniform prior on $\{1, 2, \ldots, N\}$ with $N = 500$.

   The enemy has made a mistake. Following good manufacturing procedures, each tank is stamped with a serial number, and the serial numbers are assigned sequentially, with the first tank produced receiving serial number 1. We capture n tanks, and their serial numbers are $y_1, \ldots, y_n$. Note that we must have $y_i \neq y_j$ for $i \neq j$, since the serial numbers are unique. We are going to use this to estimate how many tanks they have produced.

   We have $n = 6$ and the serial numbers $y_1, \ldots, y_n$ of these tanks are $(3, 7, 11, 22, 28, 30)$.

   (Aside: this is widely reported as a true story from World War 2.)

   (a) Show that

$$\text{Prob}(y_1, \ldots, y_n \mid x) = \begin{cases} \binom{x}{n}^{-1} & \text{if } \max_i y_i \leq x \\ 0 & \text{otherwise} \end{cases}$$

   Here $\binom{a}{b}$ denotes the binomial coefficient $\frac{a!}{b!(a-b)!}$.

   (b) Find the MAP estimate of x given the measured $y_1, \ldots, y_n$.

   (c) Plot the posterior distribution of x given $y_1, \ldots, y_n$.

(d) Compute $\mathbb{E}(x \mid y_1, \ldots, y_n)$.

(e) Suppose that instead of receiving the full list of serial numbers from the battlefront, we only receive the maximum serial number, call it q, and the number of tanks captured. One can show that

$$\text{Prob}(q \mid x) = \begin{cases} \binom{q-1}{n-1}\binom{x}{n}^{-1} & \text{if } n \le q \le x \\ 0 & \text{otherwise} \end{cases}$$

Plot the posterior distribution of x given that $q = 30$ and $n = 6$. Give an intuitive explanation for the relationship between this answer and part 2c.

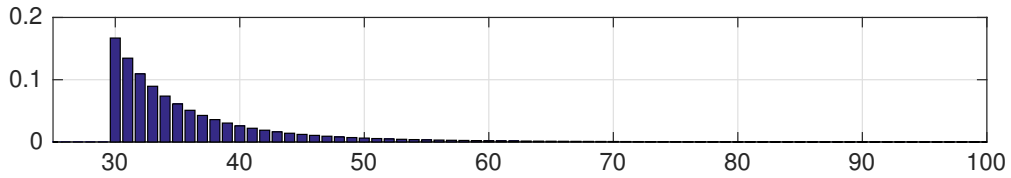
(f) Plot the posterior distribution of x given that $q = 30$ and $n = 10$.

Solution.

(a) Every nonrepeating sequence $y_1, \ldots, y_n \in \{1, \ldots, n\}$ is equally likely. The number of such sequences is $\binom{x}{n}$. A sequence is valid if and only if all y_i are less than or equal to x.

(b) The map estimate is $n = 30$.

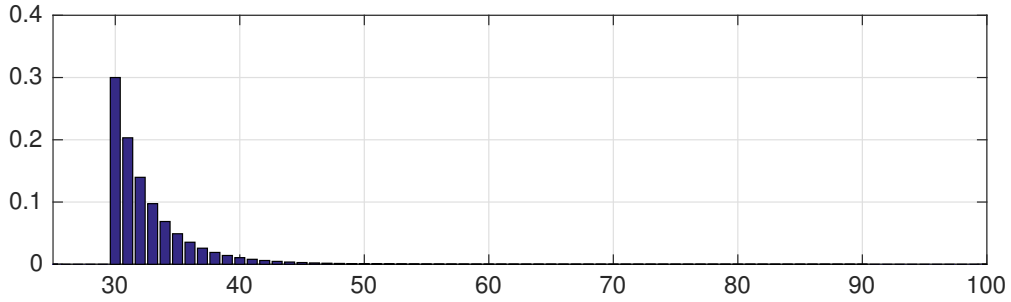
(c) The posterior $y_1, \ldots, y_n$ is below.



(d) The posterior mean is 36.11.

(e) This is the same plot as in part c. Because the largest tank serial number observed, and the number of serial numbers observed, are all that matter. The specific values of $y_1, \ldots, y_n$ do not enter into the distribution in part a), but the maximum y does. (We would say that $\max_i y_i$ is a *sufficient statistic*.)

(f) The posterior when $n = 10$ is below. After seeing more tanks, we know more about the total number produced, even though we have seen the same maximum serial number.



3. *Recursive sampling of a slowly-varying signal.*

The Gaussian discrete-time signal x is real-valued on the time interval $[1, n]$ (and so it is described by a vector $x \in \mathbb{R}^n$.) It has mean zero, and is slowly varying. We will have

$n = 100$. This slow variation shows up as a non-diagonal prior covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. The i, j entry of the prior covariance matrix is

$$\Sigma_{ij} = \exp(-(i-j)^2/100)$$

Note that the fact that this has the form of an exponential-squared is nothing to do with the Gaussian distribution, it's just a coincidence.

We makes noisy measurements of x at times $t_k = 20, 30, 40, 50, 60, 70$, given by

$$y_k = x(t_k) + w_k$$

where t_k is the sample time, and the measurement noises w_k are IID with $w_k \sim \mathcal{N}(0, 0.01)$. The measurements are

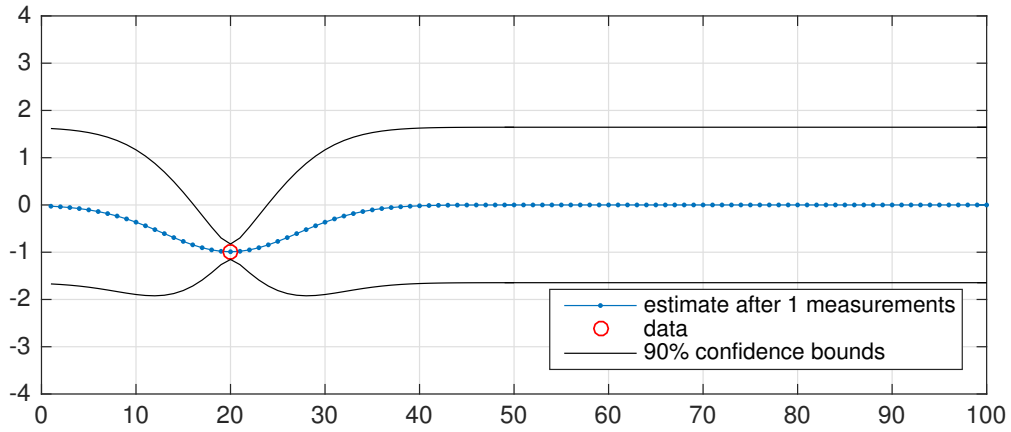
$$y = -1, -1.5, 0.4, 1, 2, 1$$

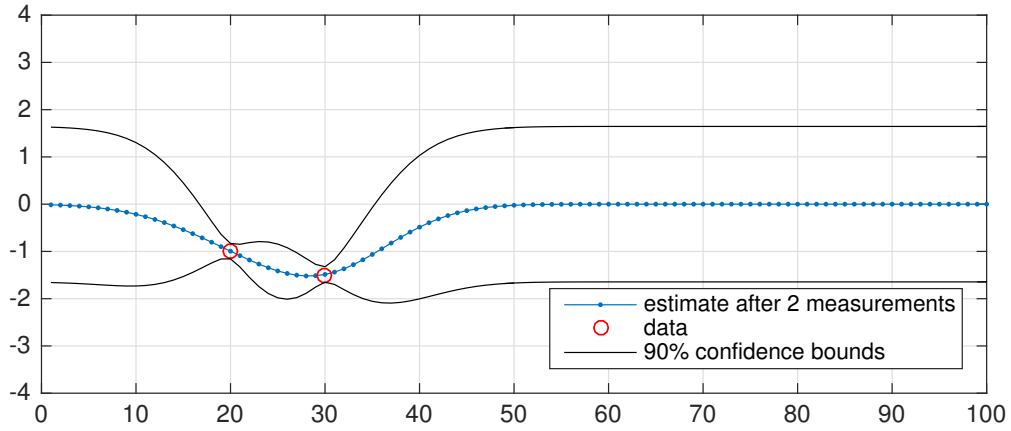
(a) What is the MMSE estimate of x after measuring $y(t_1)$? Plot the signal x as a function of time.

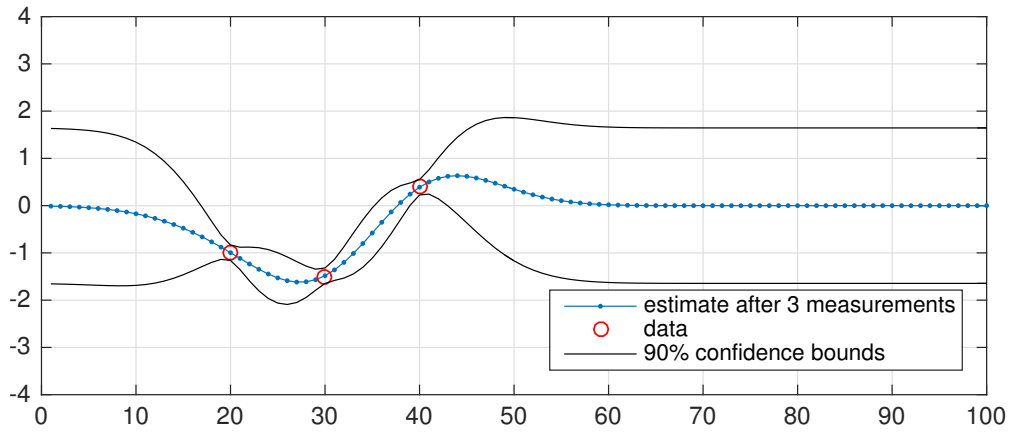
(b) At each time t you have a 90% confidence interval for $x(t)$. We can plot these intervals as two curves, an upper bound and a lower bound for x as a function of time. Plot these on top of your estimate.

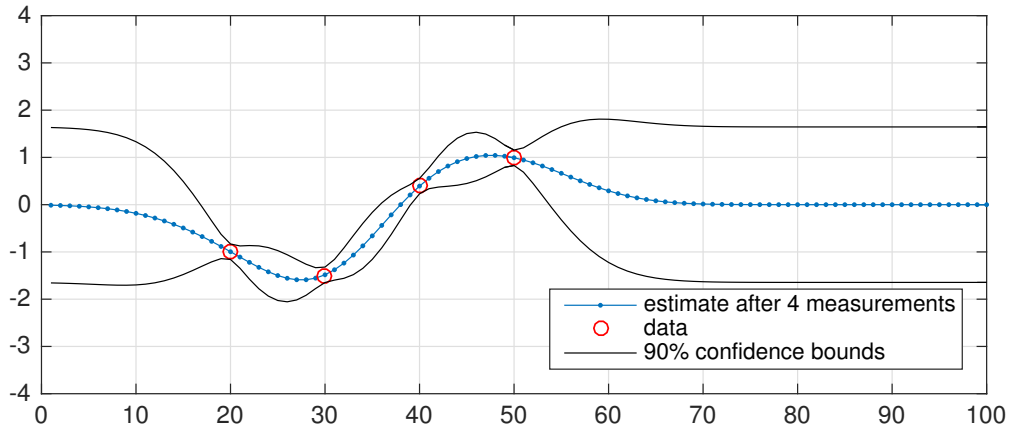
(c) Use recursive estimation to find the MMSE estimate $\mathbb{E}(x \mid y(t_1), \ldots, y(t_i))$ and repeat parts (a) and (b) above for each $k = 2, \ldots, 6$.

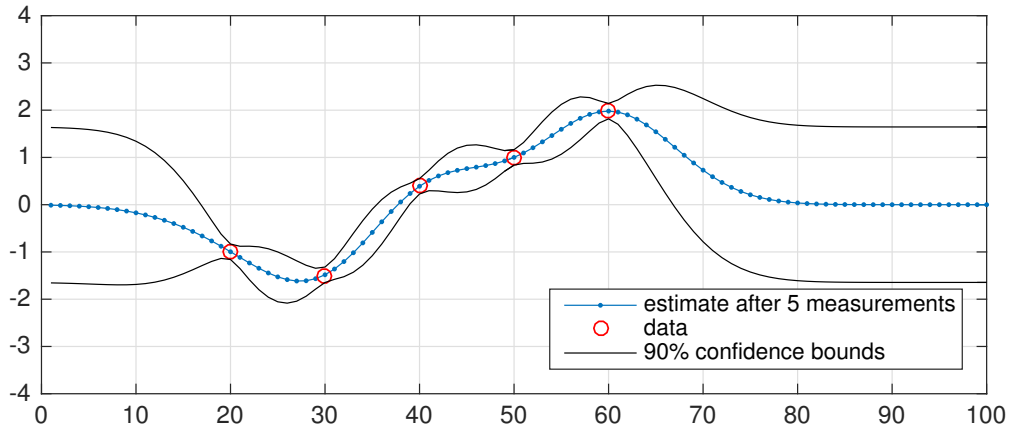
Solution. The plots are below.

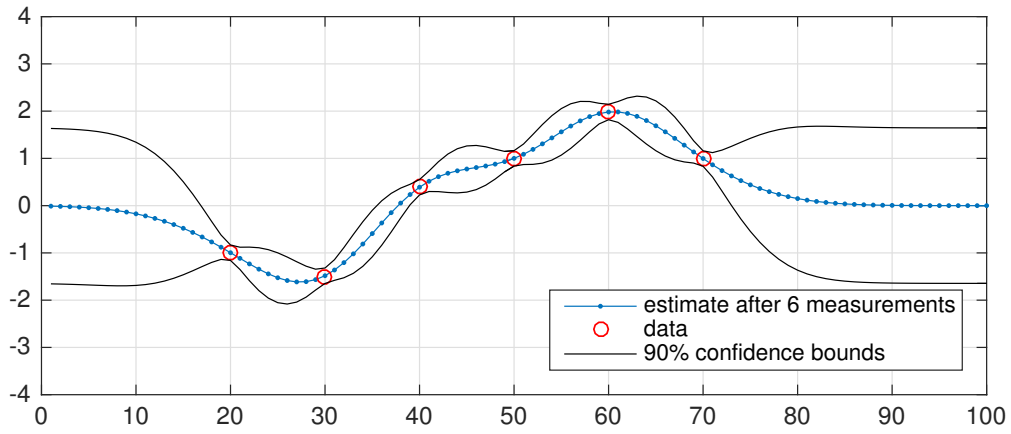





estimate after 3 measurements
data
90% confidence bounds


estimate after 4 measurements
data
90% confidence bounds


estimate after 5 measurements
data
90% confidence bounds



4. *Projecting a detection problem onto one dimension.*

   With high-dimensional data it is often useful to project the data into a lower dimensional space before performing detection. In this problem we will consider the simplest case where we project the data onto a line, i.e., a 1 dimensional space.

   Suppose we have hypotheses events X_1 and X_2. When X_i occurs, a measurement $y \in \mathbb{R}^m$ is generated whose distribution is Gaussian. We have

   $$y \mid X_j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

   We measure y and we would like to determine which of the two sources generated it. Ideally we would have the means μ_1 and μ_2 far apart, and the covariances Σ_1 and Σ_2 small, so that we would easily be able to distinguish the source.

   In this problem, we are going to project the measured y onto a line, given by the span of a vector $w \in \mathbb{R}^m$. We want to find w. We assume for convenience that $\|w\| = 1$. In the projected space, the squared distance between the projected means is

   $$s_{\text{between}} = \|w^\mathsf{T}(\mu_1 - \mu_2)\|^2$$

   This quantity is called the *between-class scatter*. It is how far apart the centers of the two clusters are.

   The other quantity of interest is the *within-class scatter*. This is

   $$s_{\text{within}} = \text{var}(w^\mathsf{T} y \mid X_1) + \text{var}(w^\mathsf{T} y \mid X_2)$$

   It is the sum of the mean-square distances from the mean in each cluster. The ratio of the two is a *signal-to-noise* ratio for the projected data.

   (a) Show that

   $$\frac{s_{\text{between}}}{s_{\text{within}}} = \frac{w^\mathsf{T} A w}{w^\mathsf{T} B w}$$

   for some matrices A and B. Find expressions for A and B.

   (b) Give an expression for the optimal w that maximizes this ratio. (Hint: a change of coordinates makes this easier.)

(c) Suppose

$$\mu_1 = \begin{bmatrix} -3 \\ 0 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 1 & 3 \\ 3 & 10 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Find the optimal w.

(d) Plot the 90% confidence ellipsoids for $y \mid X_i$.

(e) On the same figure, plot the line w onto which the data should be projected.

(f) Suppose the prior is $\text{Prob}(X_i) = \frac{1}{2}$. On the same figure, plot the decision boundaries for the MAP estimator.

(g) The projected decision problem requires distinguishing between two 1-dimensional Gaussians. Compute the probability of error of the MAP estimator for this problem.

Solution.

(a) We have

$$\begin{aligned} s_{\text{between}} &= \|w^\mathsf{T}(\mu_1 - \mu_2)\|^2 \\ &= w^\mathsf{T}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\mathsf{T} w \end{aligned}$$

and so set $A = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\mathsf{T}$. We also have

$$\begin{aligned} s_{\text{within}} &= \text{var}(w^\mathsf{T} y \mid X_1) + \text{var}(w^\mathsf{T} y \mid X_2) \\ &= w^\mathsf{T}\Sigma_1 w + w^\mathsf{T}\Sigma_2 w \end{aligned}$$

and so $B = \Sigma_1 + \Sigma_2$.

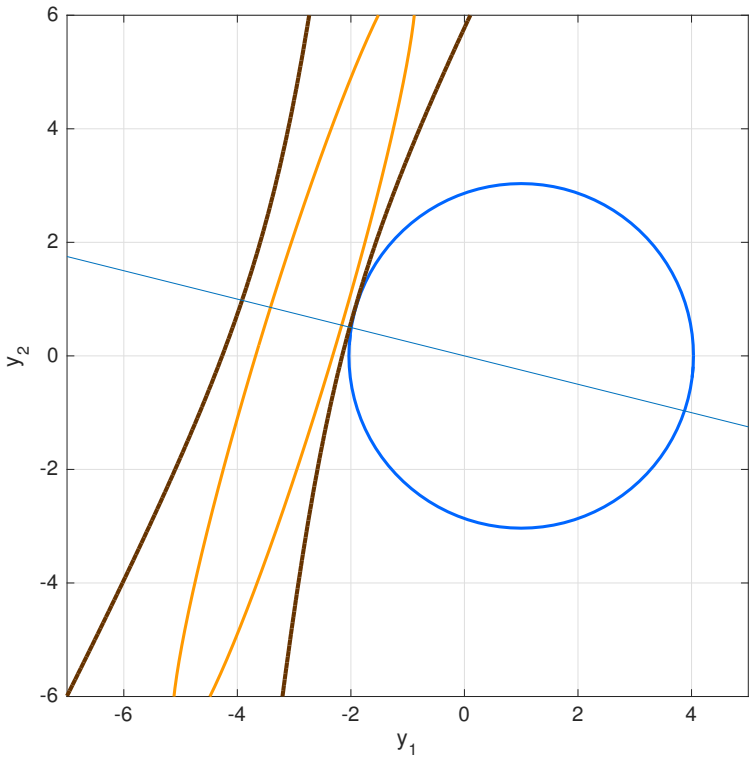
(b) To maximize this, let $q = B^{\frac{1}{2}} w$, then we have

$$\max_q \frac{q^\mathsf{T} Z q}{\|q\|^2}$$

and since Z is symmetric and positive semidefinite, the optimal q is the eigenvector of Z corresponding to the largest eigenvalue. Inverting this transformation gives the optimal w.

(c) The optimal is $w = (-0.97, 0.24)$.

(d) The plot is below.



(e) The plot is above.

(f) Really, the plot is above.

(g) The probability of error is 0.011.