

## 6.041 Final Review (Spring 2008)

**Final:** Closed-book, with THREE double-sided 8.5 x 11 formula sheets and calculator permitted. Please arrive early to find your seat before the prompt start at 9:00AM.

*Date:* Wednesday, May 21, 2008 9:00-12.00 PM

*Location:* Johnson Track (Upstairs)

*Content:* The Final exam will be comprehensive,  
but emphasize material not on Q1 and Q2.

Material covered after Quiz 2:

Lectures 16 - 24 (inclusive)

Textbook chapters: 4.5,4.6,5.1,  
6.1-6.4, NEW6.1-6.7(inclusive)

## Covariance and Correlation

- $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- If  $\text{Cov}(X, Y) = 0$ , we say  $X$  and  $Y$  are uncorrelated.
- $X$  independent of  $Y$  implies  $X$  and  $Y$  are uncorrelated. Reverse implication not true.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$
- Correlation coefficient defined as  $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$ ,  
 $-1 \leq \rho(X, Y) \leq 1$ .

## Bayesian Estimation vs. Classical Estimation

- In Bayesian Estimation, you have a joint distribution over the unknown parameter (a random variable) and the observation (a random variable).
- In Classical Estimation, you only have a parametrized distribution of the observation.

## Maximum A posteriori (MAP) Estimate

- This is a Bayesian Problem.
- Have a joint distribution over the unknown parameter ( $Y$ ) and the observation ( $X$ ). Specifically, you have the marginal distribution of  $Y$ , known as the a-priori distribution of  $Y$ . In addition, you have the conditional distribution of  $X$ , conditioned on  $Y$ .
- Using these two distributions, you can compute, using Bayes rule, the conditional distribution of  $Y$  given  $X$ , called the a-posteriori distribution of  $Y$ .
- Then, when an observation  $X=x$  is made, you pick that value  $y$  of  $Y$  which maximizes the a-posteriori distribution of  $Y$  as your estimate of  $Y$ .

## Conditional Expectation Estimator

- This is a Bayesian problem.
- Once again, you have the prior distribution on the parameter  $Y$  and the conditional distribution of  $X$  given  $Y$ .
- Using these and Bayes rule you compute the a-posteriori distribution of  $Y$ .
- When an observation  $X=x$  is made, you compute the mean of the posterior distribution of  $Y$  given  $X=x$  and use it as your estimate  $\hat{Y} = E[Y | X = x]$ .
- This makes sense only when  $Y$  is a continuous valued random variable because in the discrete case, the mean may not even be a value that  $Y$  takes.

## Least Squares Estimation

- The Conditional Expectation Estimator turns out to minimize the conditional Mean Squared Estimation error  $E[(\hat{Y} - Y)^2 | X = x]$  over a choice of all possible estimators. Hence it is also known as the Least Squares Estimator.
- Furthermore, the Conditional Expectation Estimator also minimizes the *unconditional* Mean Squared Estimation error  $E[(\hat{Y} - Y)^2]$  over all possible estimators of Y based on X.
- See page 244 for properties of the estimation error.

## Linear Least Squares Estimation

- While Conditional Expectation Estimator is optimal, its computation is often difficult.
- So we form a different, sub-optimal estimator that is easy to compute, a linear estimator.
- We choose coefficients to get the best linear estimator: the linear least squares estimator.
- Let  $Y$  be the r.v. to be estimated based on the observation  $X$ .
- We form the estimator:  $Y = aX + b$ .
- We compute its mean squared estimation error in terms of  $a$  and  $b$ .
- Then we choose  $a$  and  $b$  to minimize this error.
- This gives us  $Y = E[Y] + \rho \frac{\sigma_Y}{\sigma_X} (X - E[X])$  where  $\rho = \frac{\text{cov}(Y, X)}{\sigma_Y \sigma_X}$

- The resulting mean squared estimation error is  $(1 - \rho^2)\sigma_Y^2$ .
- This is a Bayesian situation since it involves the Covariance of X and Y.

## Maximum Likelihood Estimation

- This is a Classical problem.
- You have a parameter  $T$  which can take a bunch of values.
- For each value that  $T$  takes, you have a different distribution of the observation r.v.  $X$ .
- Once an observation  $X=x$  is made, you evaluate each of these parametric distributions of  $X$  at that point.
- These evaluated quantities are known as the likelihoods of the corresponding values of the parameter.
- You pick that value of the parameter which has the largest likelihood as your estimate.

## Hypothesis Testing

- This is a Classical problem as defined.
- However, to solve it, it is 'seen' as a Bayesian problem and elements of the Bayesian framework are introduced. Once solved in this view, these Bayesian elements are discarded and the solution is 'massaged' to conform to the Classical view.
- We have two hypotheses,  $H_0$  and  $H_1$ .  $H_0$  is the null hypothesis.  $H_1$  is called the alternative hypothesis.
- Under each hypothesis, we have a parametric distribution of the observation r.v.  $X$ .
- We observe a value  $X=x$ . This gives us the likelihoods of the two hypotheses.
- Based on these likelihoods, we must either decide on  $H_0$  or reject  $H_0$  (and accept  $H_1$ ).

- So far everything is Classical.
- Now, we add some Bayesian MAP elements.
- We assume priors  $P(H_0)$  and  $P(H_1)$  on the two hypotheses. We also treat the parametric distribution of  $X$  as a conditional distribution of  $X$  given the hypothesis.
- We can now write down, using Bayes rule, the conditional distribution of the hypothesis given the observation  $X$ .
- We can then write our decision rule as picking  $H_0$  if the a posteriori probability of  $H_0$  given  $X=x$  is larger than the a posteriori probability of  $H_1$  given  $X=x$ .
- We get: Pick  $H_0$  if  $P(H_0 | X = x) > P(H_1 | X = x)$
- i.e. if  $\frac{P(X=x|H_0)P(H_0)}{P(X=x)} > \frac{P(X=x|H_1)P(H_1)}{P(X=x)}$
- i.e. if  $\frac{P(X=x|H_0)}{P(X=x|H_1)} > \frac{P(H_1)}{P(H_0)}$
- Now we get rid of the Bayesian MAP elements. We replace the

conditional distribution of  $X$  by the parametric distribution and we replace the ratio on the right by a constant  $\psi$ .

- So the decision rule finally becomes: Pick  $H_0$  if  $\frac{P(X=x;H_0)}{P(X=x;H_1)} > \psi$
- So the 'rejection' region can be defined as
$$R = \{x \mid \frac{P(X=x;H_0)}{P(X=x;H_1)} < \psi\}$$
- The probability of a False Rejection can be computed as
$$P(\text{False Rejection}) = \Pr(x \in R; H_0)$$
- The user of the decision rule usually specifies a desired  $P(\text{False Rejection})$ . Using this specification and the above equation, one can compute the threshold  $\psi$ .

### Neyman-Pearson Lemma

For a given  $P(\text{False Rejection})$ , the above Likelihood Ratio test offers the smallest  $P(\text{False Acceptance}) = \Pr(x \notin R; H_1)$

## Linear Regression

- We have data points  $(x_i, y_i)$ . Wish to find a relation between  $x$  and  $y$ .
- This problem, as stated, is not probabilistic.
- However, we will later put it into a probabilistic framework.
- When the data is plotted, it appears that there is a linear relation between  $x$  and  $y$ .
- So we propose to build a model  $y = \theta_0 + \theta_1 x$
- The problem now becomes how to choose  $\theta_0$  and  $\theta_1$ .
- We compute the residuals  $r_i = y_i - y(x_i)$  under this model and minimize the sum of their squares to obtain  $\theta_1$  and  $\theta_2$  in terms of the data points. See page 50 of the new Chapter for the formulas.

- Thus our linear model based on the data is complete.
- Now, we model the  $x_i$ 's as given numbers and  $Y_i = \theta_0 + \theta_1 x_i + W_i$  where  $W_i$ 's are zero-mean independent normal r.v's.
- Thus we have the distribution of each  $Y_i$ , parametrized by  $\theta = (\theta_0, \theta_1)$ . This gives us the distribution of  $Y = (Y_1, Y_2, \dots, Y_n)$ , parametrized by  $\theta$ .
- We now treat the  $y = (y_1, y_2, \dots, y_n)$  as the observation and obtain the likelihood function.
- We pick  $\theta$  to maximize the likelihood function.
- This in turn minimizes the sum of the squares of the residuals.
- So we end up again with the linear regression estimate of  $\theta$ .

## Classifying Estimators

- We have some parameter to be estimated  $Y$  and we have some observation  $X$ .
- We form the estimator  $g(X)$ .
- The estimator is **unbiased** if  $E[g(X)] = Y$ .
- The estimator is **asymptotically unbiased** if  $\lim_{n \rightarrow \infty} E[g(X_1, X_2, \dots, X_n)] = Y$ .
- The estimator is **consistent** if  $g(X_1, X_2, \dots, X_n)$  converges to  $Y$  in probability.

## Bernoulli Process

Bernoulli process is a sequence  $X_1, X_2 \dots$  of **independent** Bernoulli random variables with

$$\mathbf{P}(X_i = 1) = p$$

$$\mathbf{P}(X_i = 0) = 1 - p$$

## Memoryless property

For any given time  $n$ , the sequence  $X_{n+1}, X_{n+2} \dots$  is also a Bernoulli process, and is **independent** from  $X_1, X_2 \dots X_n$ .

## Fresh-Start

Every arrival restarts the process.

## Important RV associated with Bernoulli Processes

- **First arrival** : The time to first arrival (T) is a **geometric** RV

$$p_T(t) = (1 - p)^{t-1}p, t = 1, 2 \dots$$

- **Number of arrivals**: The number of arrivals (K) in  $n$  trials is a **binomial** RV

$$p_K(k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0, 1 \dots n$$

Note: (n-fixed, k-random)

- $K^{th}$  **arrival**: The time to the  $K^{th}$  arrival  $Y_K$  is a **Pascal** RV

$$p_{Y_K}(t) = \binom{t-1}{k-1} p^k (1 - p)^{(t-k)}$$

Note: (k-fixed, t-random)

## Alternate description of the Bernoulli Process

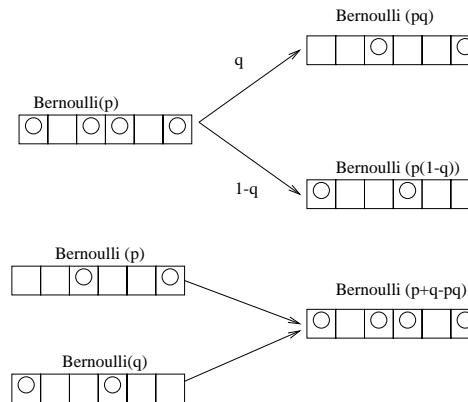
- Start with a sequence of **independent geometric** RVs  $T_1, T_2, \dots$ , with common parameter  $p$ .
- Record success(arrival) at times,  $T_1, T_1 + T_2, T_1 + T_2 + T_3, \dots$
- $K^{th}$  arrival time  $Y_k$  is the sum of the first  $k$  inter-arrival times

$$Y_k = T_1 + T_2 \dots T_k$$

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1 + T_2 \dots T_k] = \frac{k}{p}$$

$$\text{var}(Y_k) = \text{var}(T_1 + T_2 \dots T_k) = \frac{k(1-p)}{p^2}$$

## Splitting and Merging of Bernoulli Processes



- If arrivals from a Bernoulli process are split into two processes with probability  $q$  and  $(1-q)$ , each process is a Bernoulli process with parameters  $pq$  and  $p(1-q)$  respectively.
- Conversely, if we merge two independent Bernoulli processes with parameters  $p$  and  $q$ , we get a Bernoulli process with parameter  $(p+q-pq)$ .

## Summary

	Bernoulli
Inter-arrival time	Geometric
Number of arrivals	Binomial
$K^{th}$ arrival	Pascal

## Markov Chain

A Markov chain is defined by a set of states,  $S \in \{1, \dots, m\}$ , transition pairs,  $(i, j)$ , and transition probabilities,  $p_{ij}$ . Let  $X_n$  be the state after  $n$  transitions, then

$$p_{ij} = \mathbf{P}(X_{n+1} = j | X_n = i, X_{n-1}, \dots, X_0) = \mathbf{P}(X_{n+1} = j | X_n = i)$$

## Modeling Steps

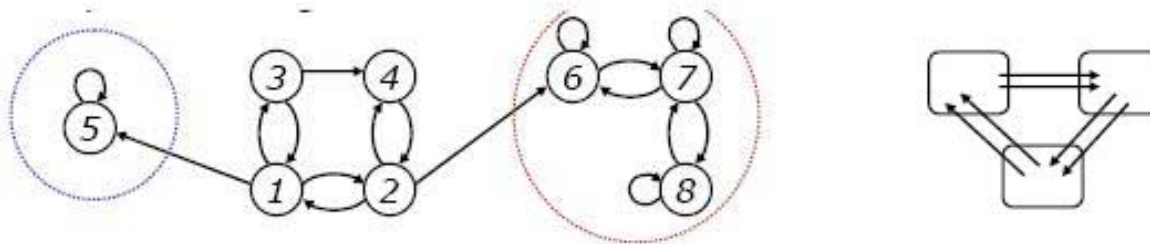
Identify possible states, mark possible transitions, and record transition probabilities. Careful in defining *states* - too broadly defined may make the problem harder to solve.

## n-Step Transition Probabilities

$$\begin{aligned} r_{ij}(n) &= \mathbf{P}(X_n = j | X_0 = i) \\ &= \sum_k r_{ik}(n-1)p_{kj} \end{aligned}$$

## Classification of States

- State  $i$  is **recurrent** if starting from  $i$ , and from *wherever* you can go, there is a way to get back to  $i$ . If a state is not recurrent, it is called **transient**. A chain can have one or more recurrent classes. See example below (left).
- **Periodic States:** States can be grouped into disjoint subset with transitions only between subsets - see example below (right)



## Markov Chain: Steady-State Probabilities

- Do the  $r_{ij}(n)$  converge to some  $\pi_j$  independent of the starting state  $i$ . The answer is yes if there is only one recurrent class in the chain and the class is not periodic.
- Two key equations:

$$\pi_j = \sum_k \pi_k p_{kj}$$
$$\sum_j \pi_j = 1$$

- Note that

$$\pi_j = 0 \quad \text{for all transient states } j$$
$$\pi_j > 0 \quad \text{for all recurrent states } j$$

## Frequency Interpretation of Steady State Probabilities

For a chain with a single aperiodic recurrent class

- Long-run frequency of being in state  $j$  is  $\pi_j$
- Frequency of transitions  $j \rightarrow k$  is  $\pi_j p_{jk}$

## Birth-Death Processes

Interesting and important special case. Nothing conceptually new but review the results on pg. 333-336 of the book for exam.

## Absorbing States

A recurrent state  $k$  is absorbing if  $p_{kk} = 1$  and  $p_{kj} = 0$  for all  $j \neq k$ .

## Absorption Probabilities

For a fixed absorbing state  $s$ , let  $a_i$  be the probability of eventually reaching state  $s$  starting from  $i$ .  $a_i$  are the unique solution to

$$\begin{aligned} a_s &= 1, \\ a_i &= 0, && \text{for all absorbing } i \neq s \\ a_i &= \sum_j p_{ij} a_j, && \text{for all transient } i \end{aligned}$$

## Expected Time to Absorption

Measures the number of steps until a recurrent class is reached.

The expected absorption time for state  $i$  is

$$\begin{aligned}\mu_i &= \mathbf{E}[\text{number of transitions until absorption starting from } i] \\ &= \mathbf{E}[\min\{n \text{ such that } X_n \text{ is recurrent}\} \mid X_0 = i]\end{aligned}$$

The expected time to absorption,  $\mu_1, \dots, \mu_n$ , are the unique solution to the following system of equations

$$\begin{aligned}\mu_i &= 0, & \text{for all absorbing states } i \\ \mu_i &= 1 + \sum_j p_{ij} \mu_j, & \text{for all transient states } i\end{aligned}$$

**Now for Some Problems...**