Computational Biology

Lecture #3: Probability and Statistics

Bud Mishra

Professor of Computer Science, Mathematics, & Cell Biology

Sept 26 2005

Basic Probabilities

Random Variables

- ◊ A (discrete) random variable is a numerical quantity that in some experiment (involving randomness takes a value from some discrete set of possible values.

- ◊ More formally, these are measurable maps $X(\omega), \omega \in \Omega$, from a basic probability space (Ω, F, P) (= outcomes, a sigma field of subsets of Ω and probability measure P on F).

- ◊ *Events* … $\{\omega \in \Omega | X(\omega) = x_i\}$… same as $\{X = x_i\}$ [X assumes the value x_i.

Examples

- ◊ Example 1: Rolling of two six-sided dice. Random Variable might be the sum of the two numbers showing on the dice. The possible values of the random variable are 2, 3, …, 12.

- ◊ Example 2: Occurrence of a specific word *GAATTC* in a genome. Random Variable might be the number of occurrence of this word in a random genome of length 3×10^9. The possible values of the random variable are 0, 1, 2, …, 3×10^9.

Probability Distribution

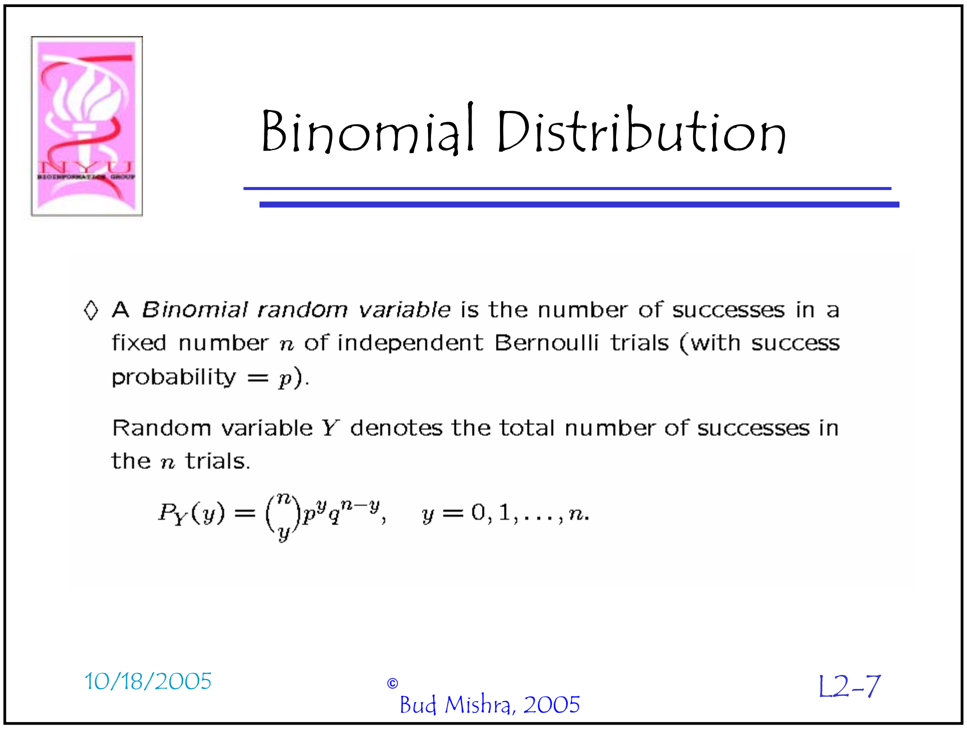
◊ The *probability distribution* of a discrete random variable Y is the set of values that this random variable can take, together with the set of associated probabilities.
Probabilities are numbers in the between zero and one inclusive that always add up to one when summed over all possible values of the random variable.

Bernoulli Trial

◊ A *Bernoulli trial* is a single trial with two possible outcomes: "success" & "failure." $P(\text{success}) = p$ and $P(\text{failure}) = 1 - p \equiv q$.

Random variable S takes the value -1 if the trial results in failure and $+1$ if it results in success.

$$P_S(s) = p^{(1+s)/2} q^{(1-s)/2}, \quad \text{s} = \text{-1}, +1.$$

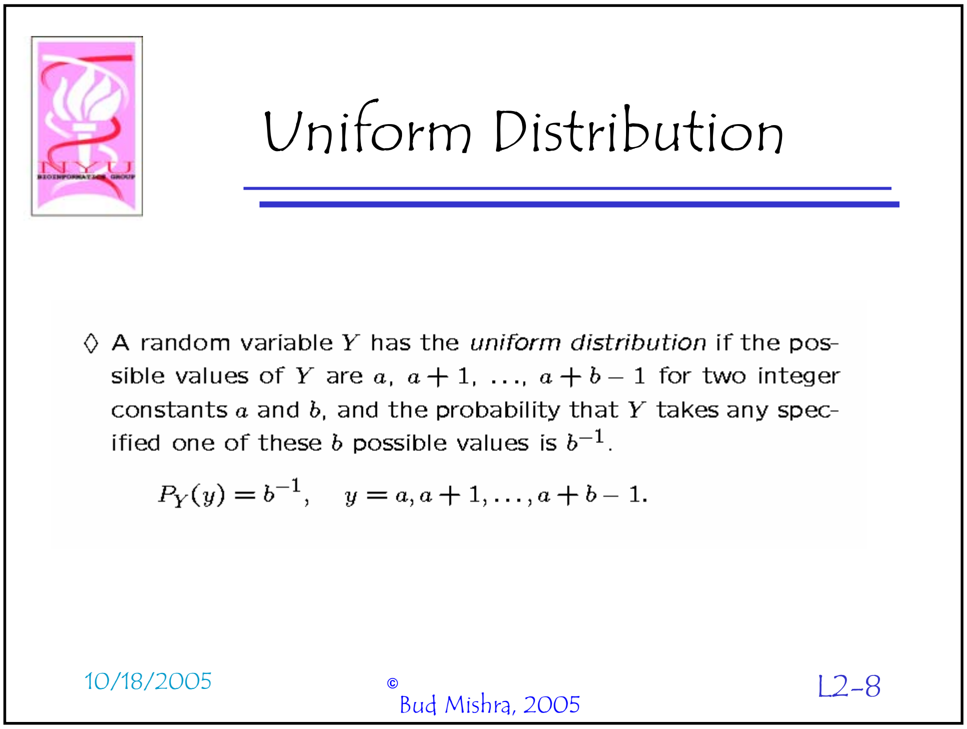


Binomial Distribution

◊ A *Binomial random variable* is the number of successes in a fixed number n of independent Bernoulli trials (with success probability $= p$).

Random variable Y denotes the total number of successes in the n trials.

$$P_Y(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, \ldots, n.$$



Uniform Distribution

◊ A random variable Y has the *uniform distribution* if the possible values of Y are a, $a+1$, ..., $a+b-1$ for two integer constants a and b, and the probability that Y takes any specified one of these b possible values is b^{-1}.

$$P_Y(y) = b^{-1}, \quad y = a, a+1, \ldots, a+b-1.$$

Geometric Distribution

- ◊ Suppose that a sequence of independent Bernoulli trials is conducted, each trial having probability p of success. The random variable of interest is the number Y of trials before but not including the first failure. The possible values of Y are 0, 1, 2, ...

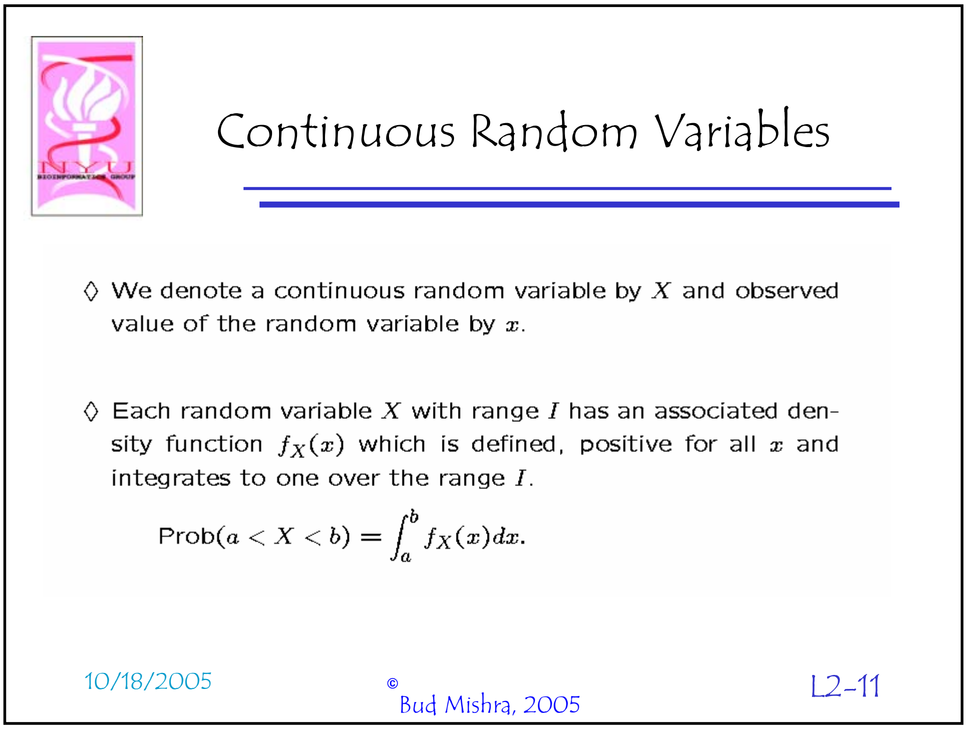
$$P_Y(y) = p^y q, \quad y = 0, 1, \ldots.$$

Poisson Distribution

- ◊ A random variable Y has a Poisson distribution (with parameter $\lambda > 0$) if

$$P_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, \ldots.$$

The Poisson distribution often arises as a limiting form of the binomial distribution.

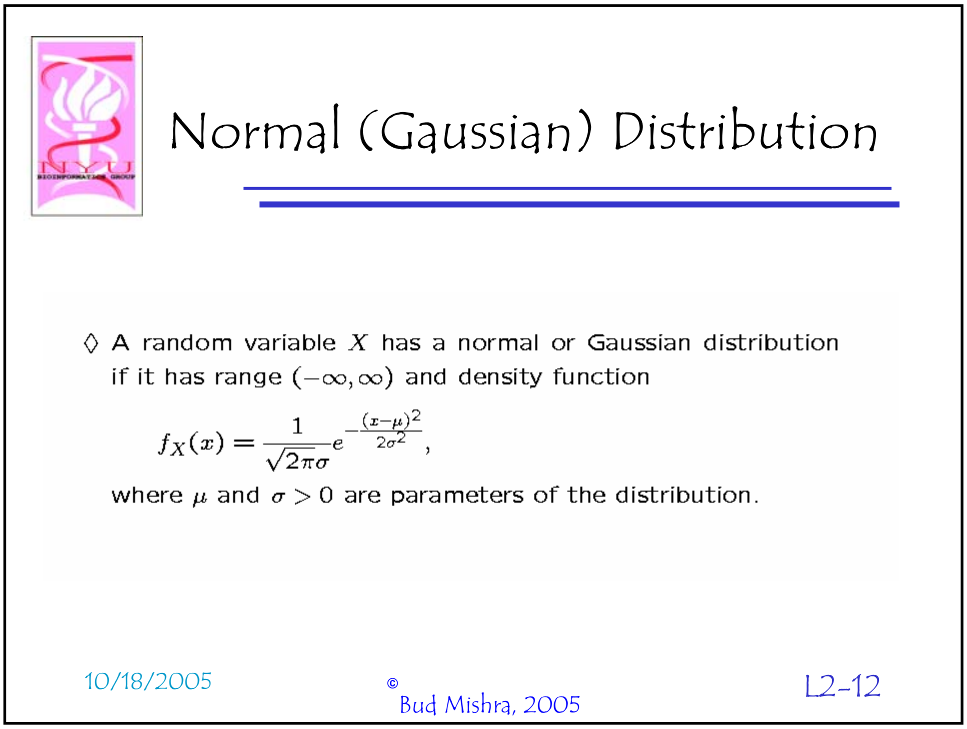


Continuous Random Variables

◊ We denote a continuous random variable by X and observed value of the random variable by x.

◊ Each random variable X with range I has an associated density function $f_X(x)$ which is defined, positive for all x and integrates to one over the range I.

$$\text{Prob}(a < X < b) = \int_a^b f_X(x)dx.$$

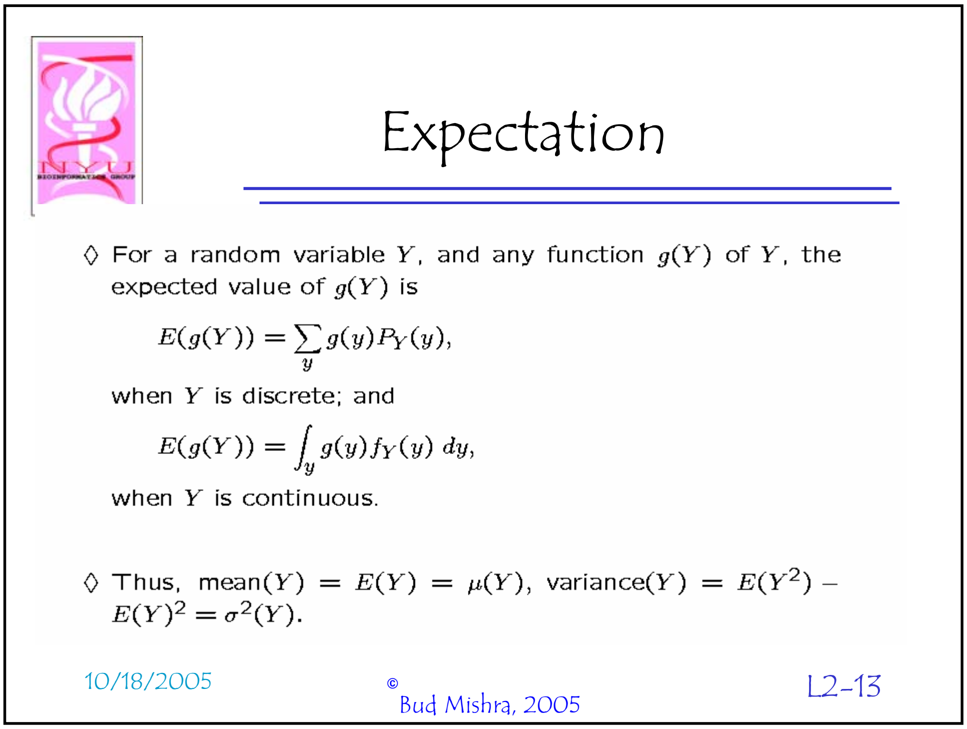


Normal (Gaussian) Distribution

◊ A random variable X has a normal or Gaussian distribution if it has range $(-\infty, \infty)$ and density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ and $\sigma > 0$ are parameters of the distribution.



Expectation

- For a random variable Y, and any function $g(Y)$ of Y, the expected value of $g(Y)$ is

$$E(g(Y)) = \sum_y g(y) P_Y(y),$$

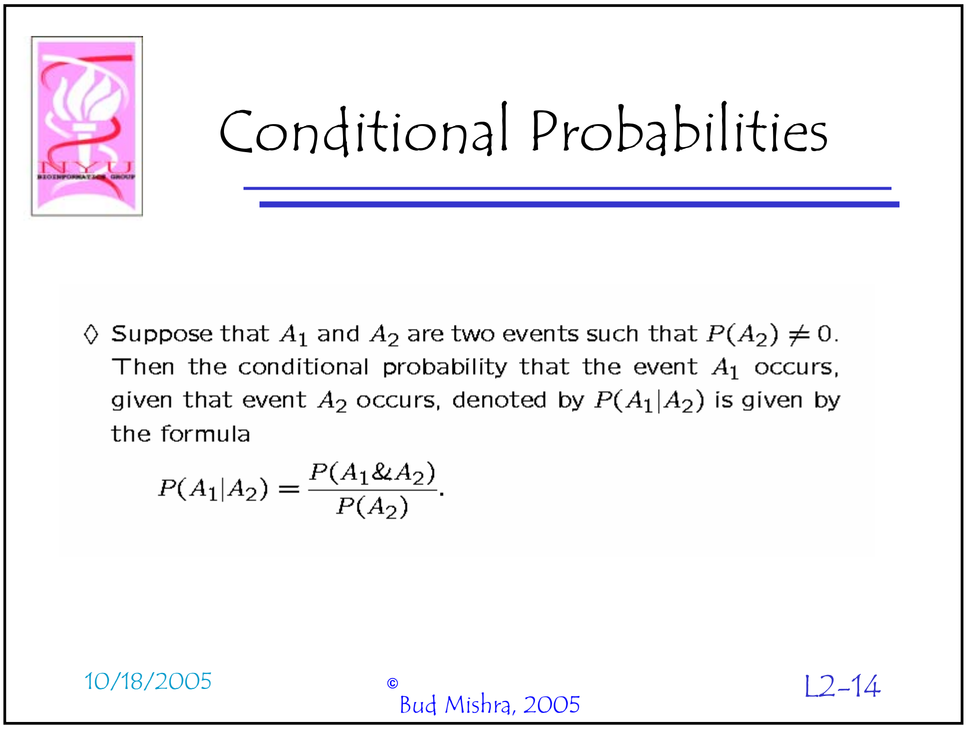
when Y is discrete; and

$$E(g(Y)) = \int_y g(y) f_Y(y)\, dy,$$

when Y is continuous.

- Thus, mean(Y) $= E(Y) = \mu(Y)$, variance(Y) $= E(Y^2) - E(Y)^2 = \sigma^2(Y)$.

10/18/2005 © Bud Mishra, 2005 L2-13

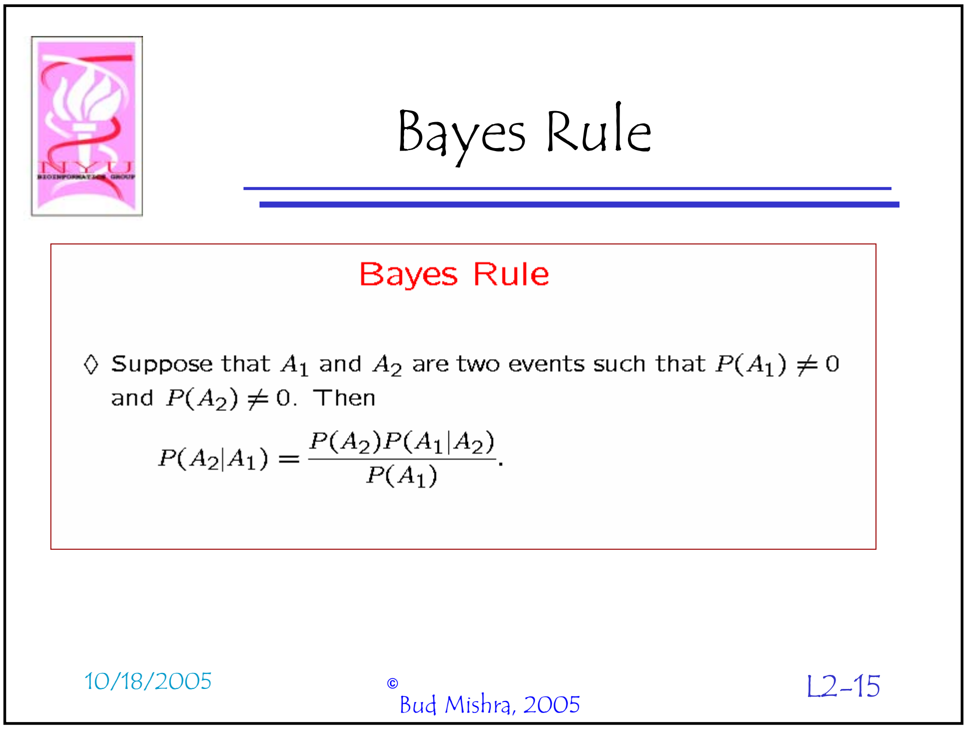


Conditional Probabilities

- Suppose that A_1 and A_2 are two events such that $P(A_2) \neq 0$. Then the conditional probability that the event A_1 occurs, given that event A_2 occurs, denoted by $P(A_1|A_2)$ is given by the formula

$$P(A_1|A_2) = \frac{P(A_1 \& A_2)}{P(A_2)}.$$

10/18/2005 © Bud Mishra, 2005 L2-14

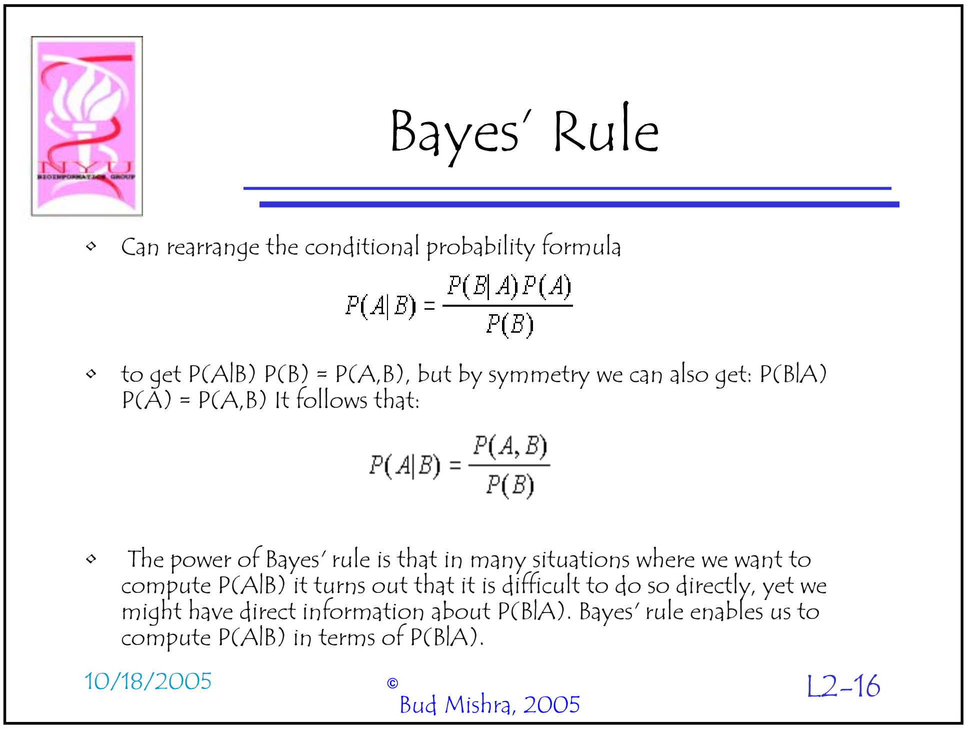


Bayes Rule

Bayes Rule

◊ Suppose that A_1 and A_2 are two events such that $P(A_1) \neq 0$ and $P(A_2) \neq 0$. Then

$$P(A_2|A_1) = \frac{P(A_2)P(A_1|A_2)}{P(A_1)}.$$



Bayes' Rule

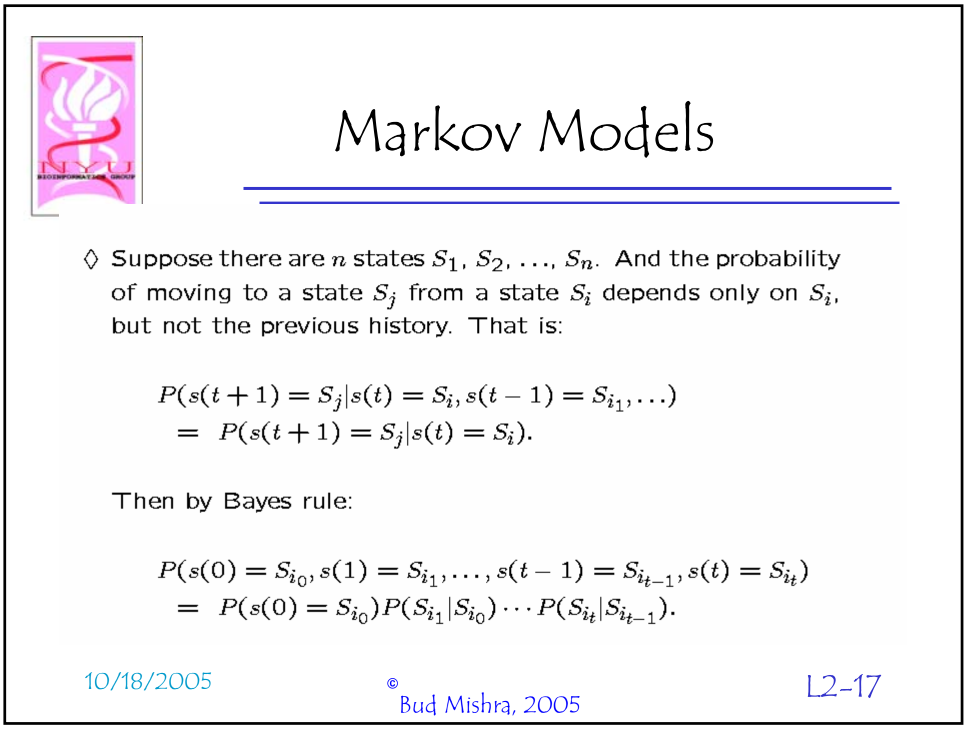
- Can rearrange the conditional probability formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- to get P(A|B) P(B) = P(A,B), but by symmetry we can also get: P(B|A) P(A) = P(A,B) It follows that:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

- The power of Bayes' rule is that in many situations where we want to compute P(A|B) it turns out that it is difficult to do so directly, yet we might have direct information about P(B|A). Bayes' rule enables us to compute P(A|B) in terms of P(B|A).



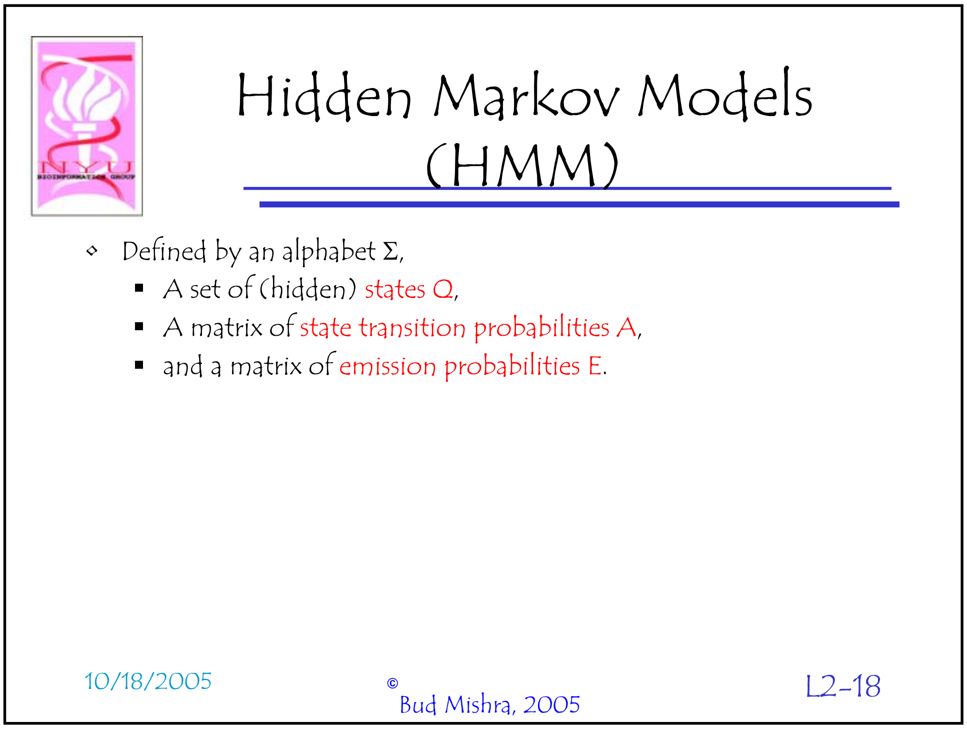
Markov Models

- Suppose there are n states S_1, S_2, …, S_n. And the probability of moving to a state S_j from a state S_i depends only on S_i, but not the previous history. That is:

$$P(s(t+1) = S_j | s(t) = S_i, s(t-1) = S_{i_1}, \ldots)$$
$$= \ P(s(t+1) = S_j | s(t) = S_i).$$

Then by Bayes rule:

$$P(s(0) = S_{i_0}, s(1) = S_{i_1}, \ldots, s(t-1) = S_{i_{t-1}}, s(t) = S_{i_t})$$
$$= \ P(s(0) = S_{i_0}) P(S_{i_1} | S_{i_0}) \cdots P(S_{i_t} | S_{i_{t-1}}).$$



Hidden Markov Models (HMM)

- Defined by an alphabet Σ,
  - A set of (hidden) states Q,
  - A matrix of state transition probabilities A,
  - and a matrix of emission probabilities E.

States

- Σ = An alphabet of symbols
- Q = A set of states that emit symbols from the alphabet Σ
- A = (a_{kl}) = |Q| £ |Q| matrix of state transition probabilities
- E = $(e_K(B))$ = |Q| £ |Σ| matrix of emission probabilities

10/18/2005

© Bud Mishra, 2005

L2-19

A Path in the HMM

- $\pi = \pi_1\ \pi_2 \cdots \pi_n$

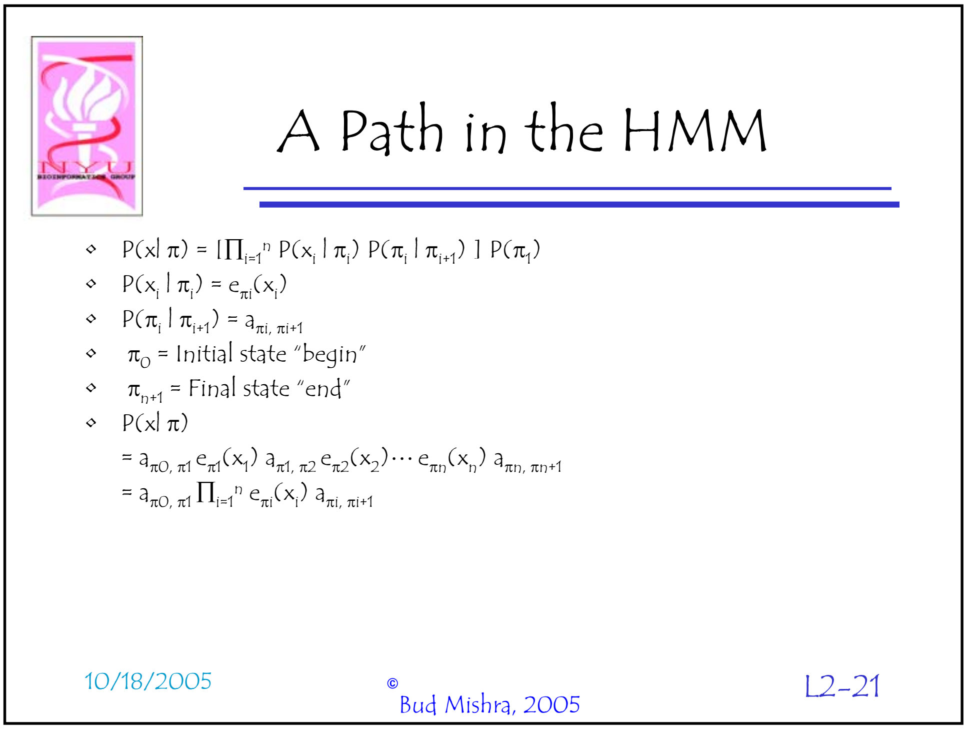
= a sequence of states 2 Q^* in the hidden Markov model M

- x 2 Σ^* = sequence generated by the path π, determined by the model M
- $P(x|\ \pi) = P(\pi_1)[\ \prod_{i=1}^{n} P(x_i \mid \pi_i)\ P(\pi_i \mid \pi_{i+1})\ ]$

10/18/2005

© Bud Mishra, 2005

L2-20

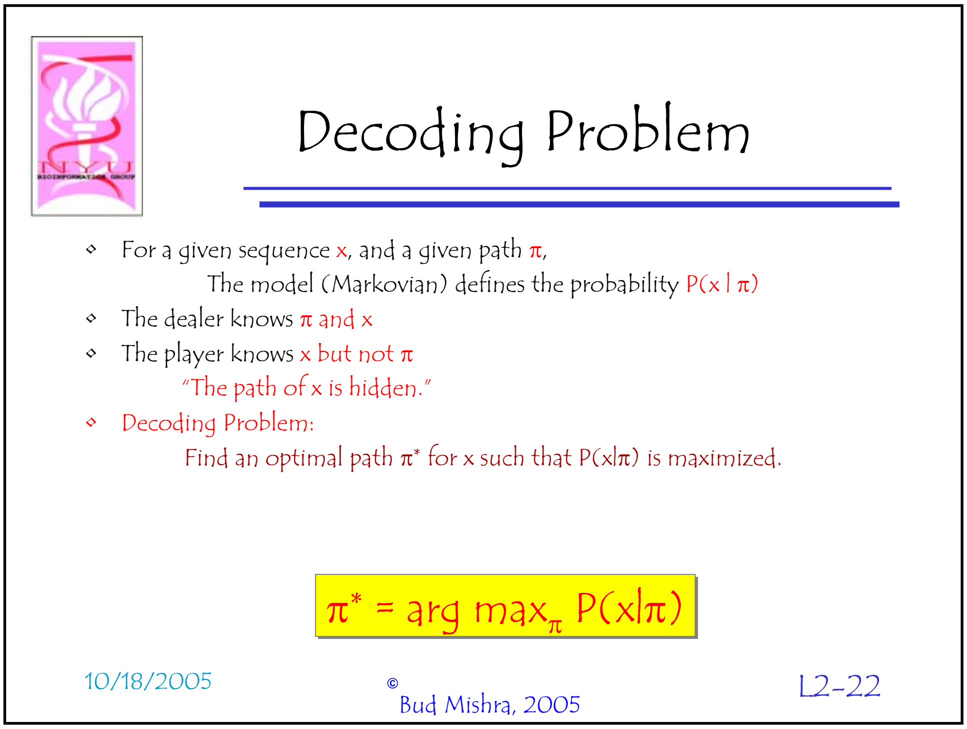


A Path in the HMM

- $P(x| \pi) = [\prod_{i=1}^{n} P(x_i \mid \pi_i)\, P(\pi_i \mid \pi_{i+1})\,]\, P(\pi_1)$
- $P(x_i \mid \pi_i) = e_{\pi i}(x_i)$
- $P(\pi_i \mid \pi_{i+1}) = a_{\pi i,\, \pi i+1}$
- π_0 = Initial state "begin"
- π_{n+1} = Final state "end"
- $P(x| \pi)$

  $= a_{\pi 0,\, \pi 1}\, e_{\pi 1}(x_1)\, a_{\pi 1,\, \pi 2}\, e_{\pi 2}(x_2) \cdots e_{\pi n}(x_n)\, a_{\pi n,\, \pi n+1}$

  $= a_{\pi 0,\, \pi 1} \prod_{i=1}^{n} e_{\pi i}(x_i)\, a_{\pi i,\, \pi i+1}$



Decoding Problem

- For a given sequence x, and a given path π,

  The model (Markovian) defines the probability P(x | π)
- The dealer knows π and x
- The player knows x but not π

  "The path of x is hidden."
- Decoding Problem:

  Find an optimal path π* for x such that P(x|π) is maximized.

$$\pi^* = \arg\max_{\pi} P(x|\pi)$$

Dynamic Programming Approach

- Principle of Optimality:
- Optimal path for the (i+1)-prefix of x

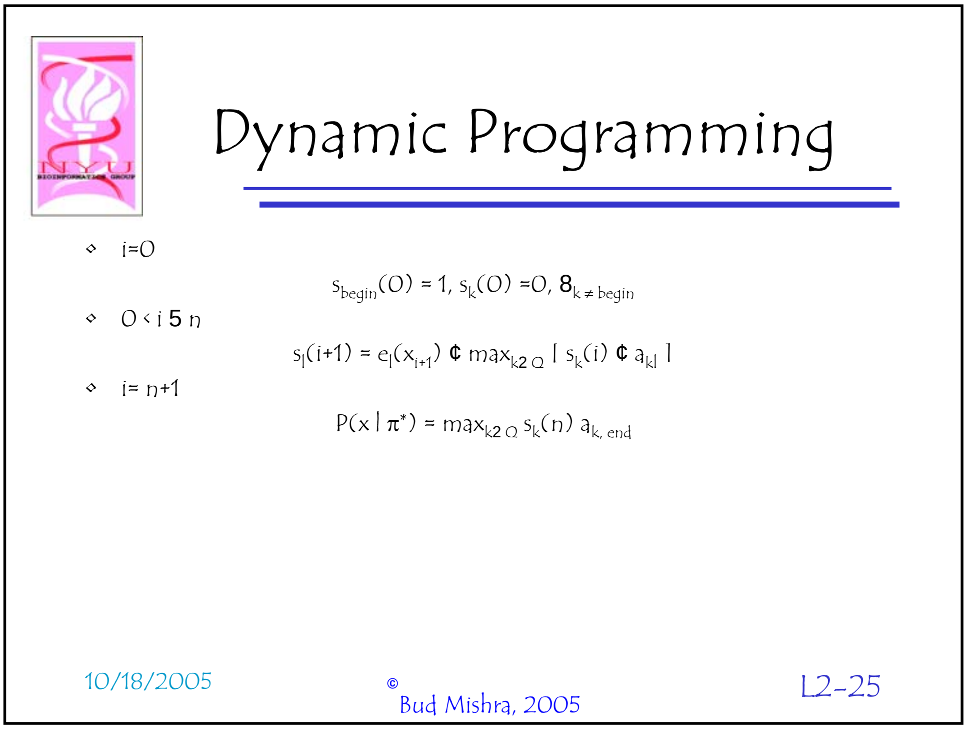
  $x_1 \cdots x_{i+1}$
- uses a path for an i-prefix of x that is optimal among the paths ending in an (unknown) state $\pi_i = k\ 2\ Q$

Dynamic Programming Approach

- $s_k(i)$ = The probability of the most probable path for the i-prefix ending in state k.

$$8_{k\,2\,Q}\ 8_{1\,5\,i\,5\,n}$$

$$s_l(i+1) = e_l(x_{i+1}).\ \max_{k2\,Q}\ [s_k(i)\ .\ a_{kl}]$$



Dynamic Programming

- i=0

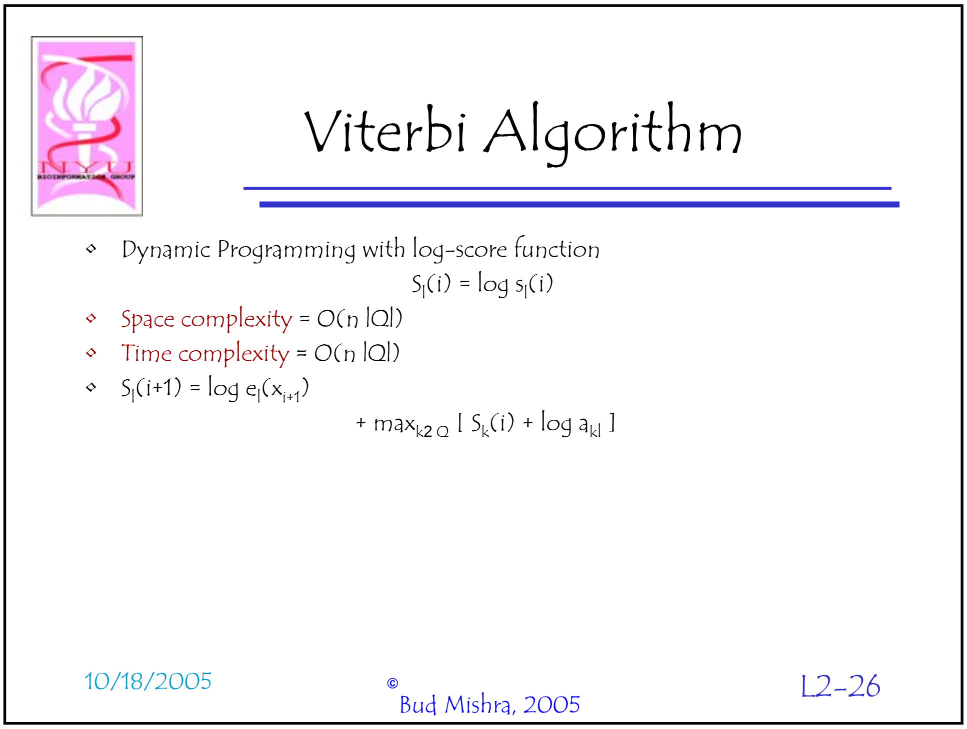
$$s_{begin}(0) = 1, s_k(0) = 0, 8_{k \neq begin}$$

- 0 < i 5 n

$$s_l(i+1) = e_l(x_{i+1}) ¢ \max_{k2 Q} [s_k(i) ¢ a_{kl}]$$

- i= n+1

$$P(x | \pi^*) = \max_{k2 Q} s_k(n) a_{k, end}$$



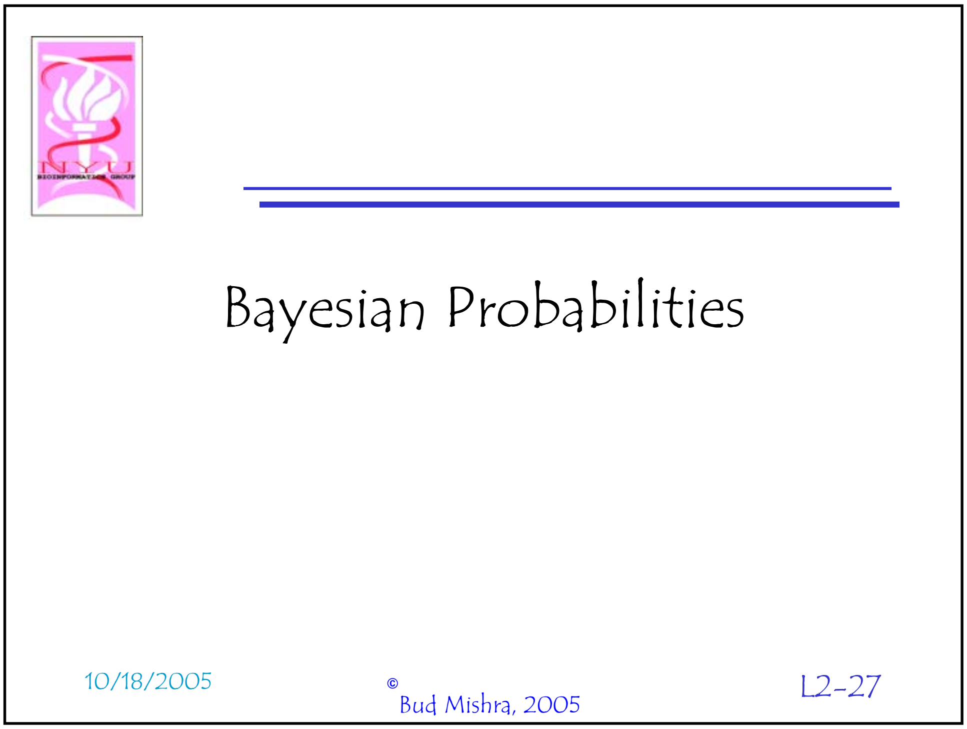
Viterbi Algorithm

- Dynamic Programming with log-score function

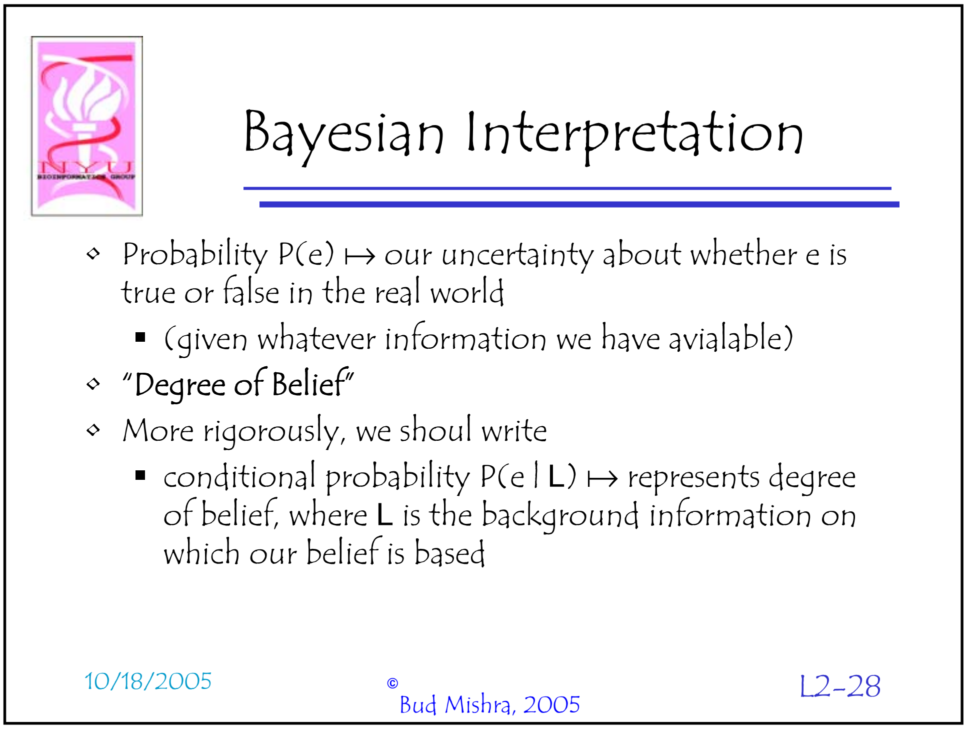
$$S_l(i) = \log s_l(i)$$

- Space complexity = O(n |Q|)
- Time complexity = O(n |Q|)
- $S_l(i+1) = \log e_l(x_{i+1})$

$$+ \max_{k2 Q} [S_k(i) + \log a_{kl}]$$



Bayesian Probabilities



Bayesian Interpretation

- Probability P(e) ↦ our uncertainty about whether e is true or false in the real world
  - (given whatever information we have avialable)
- "Degree of Belief"
- More rigorously, we shoul write
  - conditional probability P(e | **L**) ↦ represents degree of belief, where **L** is the background information on which our belief is based

Probability as a Dynamic Entity

- Update the "degree of belief" as more data arrives:
- Bayes Theorem: P(e | **D**) = P(**D** | e) P(e)/P(**D**)
- *Prior Probability*: P(e) is your belief in the event e before you see any data at all
- *Posterior*: P(e | **D**) is the updated posterior belief in e given the observed data.
- *Likelihood*: P(**D** | e) ↦ probability of the data under the assumption e.
- Posterior is proportional to the prior.

Dynamics

- P(e | $\mathbf{D}_1$, $\mathbf{D}_2$) = P($\mathbf{D}_2$ | e, $\mathbf{D}_1$) P(e | $\mathbf{D}_1$)/ P($\mathbf{D}_2$ | $\mathbf{D}_1$)
- *Important Observation:*
- The effects of prior diminish as the number of data points increases.
- *The Law of Large Number:*
- With large number of data points, Bayesian and frequentist viewpoints become indistinguishable.

Parameter Estimation

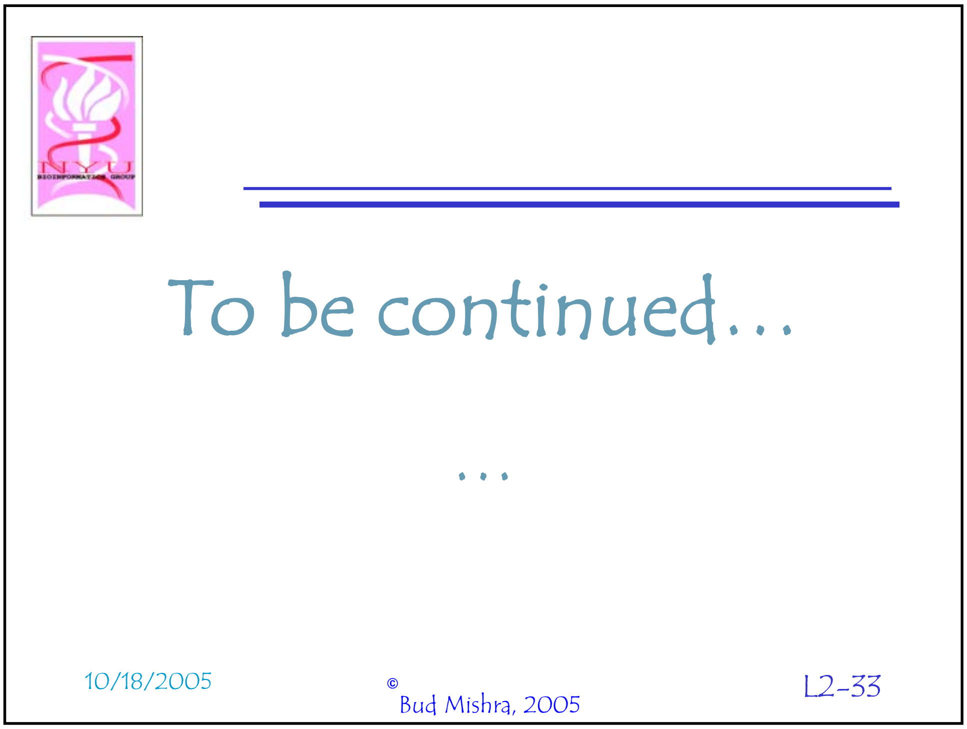
- Functional form for a model M
  - Depends on parameters Θ
  - Best estimation for Θ?
- Typically our parameters Θ are a set of real-valued numbers
  - Both prior P(Θ) and the posterior P(Θ | D) are defining probability density functions

Maximum A Posteriori (MAP)

- Find the set of parameters Θ
  - maximizing the posterior P(Θ | D) or minimizing a score -log P(Θ | D)
  - E'(Θ) = -log P(Θ | D)

    = -log P(D | Θ) – log P(Θ) + logP(D)
  - Same as minimizing E(Θ) = -log P(D | Θ) – log P(Θ)
  - If the prior P(Θ) is uniform over the entire parameter space (uninformative):

    Minimize E_L(Θ) = -log P(D | Θ)
  - *Maximum likelihood solution*



To be continued…

…

10/18/2005

© Bud Mishra, 2005

L2-33