CPS260/BGT204.1 Algorithms in Computational Biology **October 9, 2003**

Lecture 13: Probabilistic and Stochastic Modeling

Lecturer: Pankaj K. Agarwal *Scribe: Tingting Jiang*

Before we go to the Hidden Markov Models (HMM) which Dr. Ron Parr will talk about next Thursday (Oct. 16, 2003), it is necessary to introduce the background of probability theory and statistics. Most materials of this lecture are from the book of R. Durbin [1].

13.1 Probability Distributions

13.1.1 Probability

Given a sample space Ω which is a set of elementary events, we can define the probability as a measure of how likely it is that some event will occur. For example, if we flip a coin, there are two possible outcomes: Head or Tail. So $\Omega = \{H, T\}$. In other words, if we view every event as the outcome of an experiment, the sample space includes all the possible outcomes of the experiments. And the probability tells us the chance of each outcome. Another example is tossing a die. The sample space is $\{1, 2, 3, 4, 5, 6\}$. If the die is not biased, the probability of each outcome is $\frac{1}{6}$.

However, some experiment cannot be repeated, such as "What's the probability that Bush will win the second run for the U.S. president?" But people still talk about it in probability. So probability can be interpreted in two ways:

- Count
- Belief

13.1.2 Continuous Distributions

It is said that a random variable X has a *continuous distribution* if there exists a nonnegative function f, defined on the real line, such that for any interval I,

$$Pr[X \in I] = \int_I f(x)\mathrm{d}x. \tag{13.1}$$

The function f is called the *probability density function* (abbreviated p.d.f.). Every p.d.f. must satisfy the following two requirements:

- $f(x) \geq 0$

- $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

A typical p.d.f. is sketched in Fig. 13.1. In this figure, the total area under the curve must be 1. Another

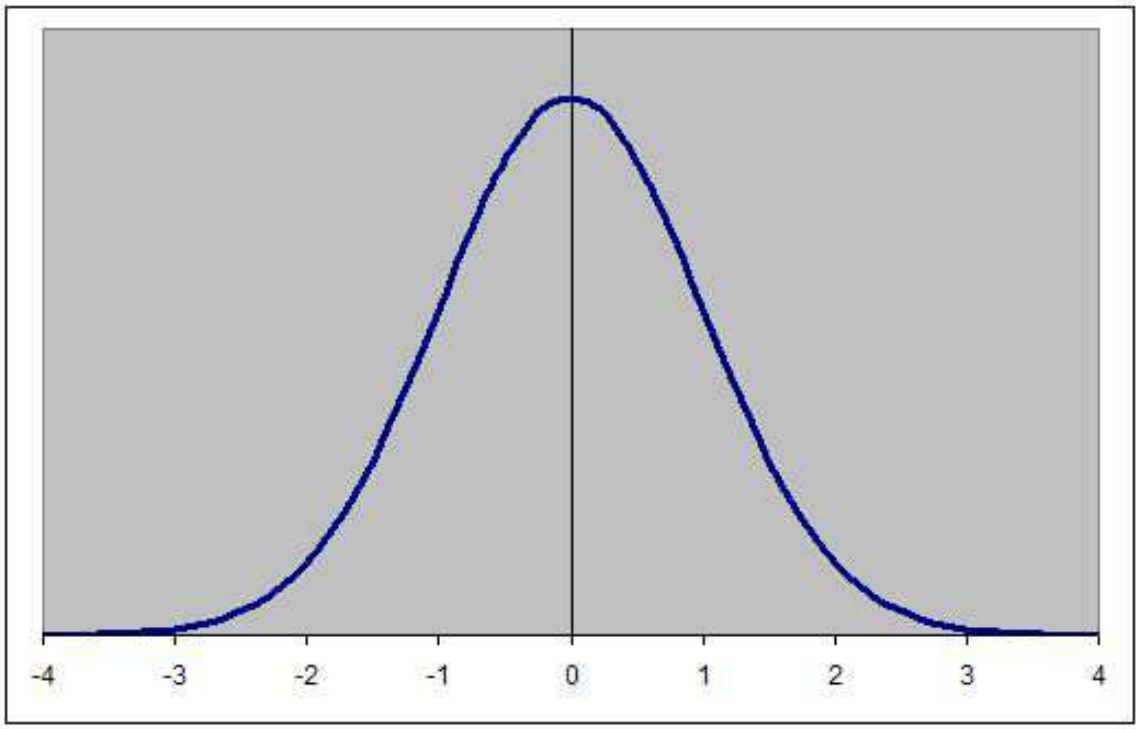

Figure 13.1: An example of a p.d.f.

interesting p.d.f. is the *delta function* (Fig. 13.2). When ϵ goes to zero, the value of the delta function will be infinite.

It is clearly that

$$Pr[X \in [x, x + \mathrm{d}x]] = f(x) \cdot \mathrm{d}x$$

and

$$Pr[X < x_0] = \int_{-\infty}^{x_0} f(x)\mathrm{d}x. \tag{13.2}$$

13.1.3 Discrete Distributions

So far we have discussed continuous probability distribution. However, in this course, we are more interested in discrete probability distribution. For instance, the sample space $\Omega = \{x_1, x_2, \ldots, x_n\}$ includes n items. During the process of the experiments, we choose one of the n items.

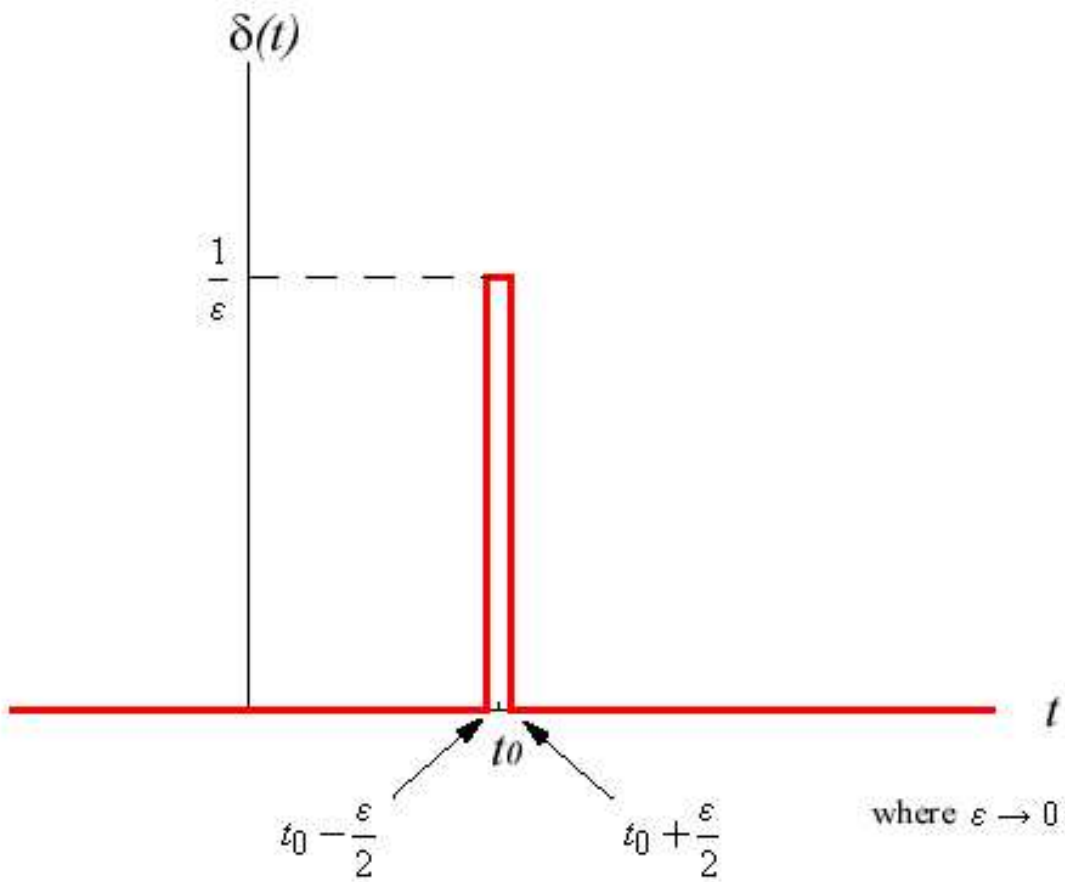

Figure 13.2: δ function.

The first discrete probability distribution we consider might be the simplest: *the binomial distribution*. It is defined on a finite set consisting of all the possible results of N tries of an experiment with a binary outcome, '0' or '1'. Consider flipping a coin. Let p is the probability of getting 'Head' and $1-p$ that of getting 'Tail', the probability that k out of the N tries yield "Head" is

$$Pr[k\text{'Heads' out of } N\} = \binom{N}{k} p^k (1-p)^{N-k}. \tag{13.3}$$

Define a random variable $X = 1$ when getting 'Head' otherwise $X = 0$. Then we can easily get the mean and variance of X as $\mu(X) = Np$ and $Var(X) = Np(1-p)$.

In the limit of a large number of events ($N \to \infty$), a binomial distribution becomes a Gaussian (see Fig. 13.3). The standard Gaussian distribution is in the form:

$$N(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2}/2\sigma^2. \tag{13.4}$$

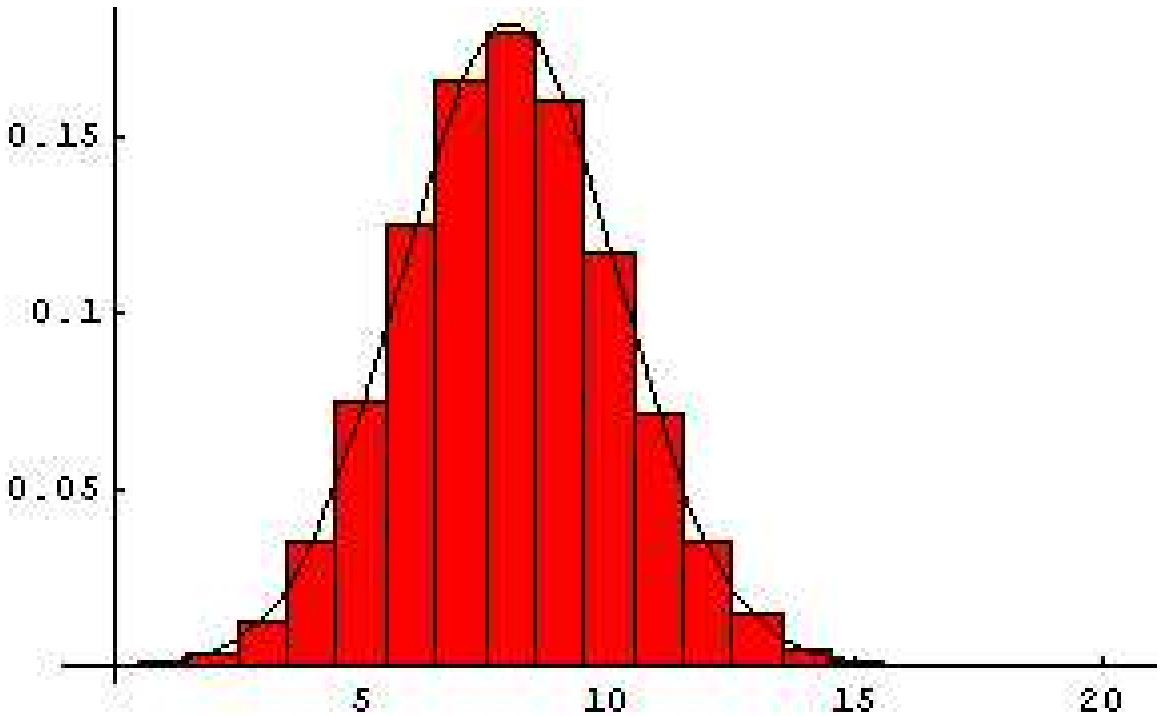

Figure 13.3: The limit for large N of a binomial tends to a Gaussian.

13.1.4 Conditional Distributions

In many cases, we are not only interested in one event, but also multiple events. Let A and B be two different events. The probability that both A and B happen is equal to :

$$Pr[A \cap B] = Pr[A|B] \cdot Pr[B] \tag{13.5}$$

For example, consider an experiment where we have a fair coin and a biased coin which always gets 'Head'. We randomly chose one of them and then toss it. Define two events as follows:

- A: getting 'Head'
- B: the biased coin was picked.

Then we have

$$\begin{gathered} Pr[B] = 1/2 \\ Pr[A|B] = 1 \\ Pr[A|\bar{B}] = 1/2 \\ Pr[A] = Pr[A|B] \cdot Pr[B] + Pr[A|\bar{B}] \cdot Pr[\bar{B}] = 1 \cdot (1/2) + (1/2) \cdot (1/2) = 3/4 \end{gathered}$$

where $Pr(\bar{B})$ denotes the probability that event B does not occur It is clear that

$$Pr(A \cap B) = Pr(A|B)Pr(B) = Pr(B|A)Pr(A) \tag{13.6}$$

Then we have

$$\boxed{Pr(A|B) = \frac{Pr(A)Pr(B|A)}{Pr(B)}}$$

which is the well-known ***Bayes' Theorem***.

Now let's do another experiment. We choose one of the two coins and toss it twice. Define two events as follows:

- A: the biased coin was picked
- B: the coin shows 'Head' both times.

We wish to determine $Pr(A|B)$. Since $Pr(A) = 1/2, Pr(B|A) = 1, Pr(\bar{A}) = 1/2$ and $Pr(B|\bar{A}) = 1/4$. Hence, by Bayes' Theorem

$$Pr(A|B) = \frac{(1/2) \cdot 1}{(1/2) \cdot 1 + (1/2)(1/4)} = 4/5.$$

13.2 Entropy

An *entropy* is a measure of the average uncertainty of an outcome. Given a random variable X with probabilities $P(x_i)$ for a discrete set of n events $x_1, \ldots, x_n$, the Shannon entropy is defined by

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i) \tag{13.7}$$

In this definition, $P(x_i) \log_2 P(x_i) = 0$ if $P(x_i) = 0$. Intuitively, the entropy tells us the number of bits we need to encode the information. The entropy is maximized when all the $P(x_i)$ are equal ($P(x_i) = 1/n$). The maximum value of entropy is given by

$$H = -\sum_i \frac{1}{n} \log_2 \frac{1}{n}$$

If we are certain of the outcome of a sample from the distribution i.e. $P(x_i) = 1$ for one i and the other $P(x_j) = 0$, the entropy is zero. The unit of entropy is called "bit".

Example: Entropy of random DNA
If each symbol $\{A, C, G, T\}$ of a DNA sequence occurs equiprobably ($p_a = 1/4$) then the entropy per DNA symbol is $-\Sigma_a p_a \log_2 p_a = 2$ bits. But we observe that G is paired to C and A is paired to T. So the actual entropy is lower.

13.3 Relative Entropy

For two distributions P and Q the relative entropy(also known as the Kullback-Leibler 'distance') is defined by

$$H(P||Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)} \tag{13.8}$$

Let $\beta(x_i) = \log \frac{P(x_i)}{Q(x_i)}$. $\beta(x_i)$ in fact represents the logarithm of the likelihood ratio. For example, we have one unbiased coin and one biased coin with turns 'Head' with probability 0.9. Then we have the following table:

Distribution	Head	Tail
Q (unbiased)	1/2	1/2
P (biased)	0.9	0.1
$\beta(x)$	$\log 1.8$	$\log 0.2$

From the above table, we can see that $\beta(x)$ entry for 'Head' is positive while the one for 'Tail' is negative. That means it is more likely to get 'Head' from the distribution P than Q. Generally speaking, the positive entries $\beta(x_i)$ imply more likely the event x_i will occur and the negative entries $\beta(x_i)$ imply less likely the event x_i will occur.

13.4 Inference

Once a type of model is chosen, the parameters of the model have to be *inferred* from the data. For example, we may model the outcome of rolling a die with a multinomial distribution. Suppose the number of observations yielding i is $n_i (i = 1, 2, \cdots, 6)$. We don't know if it is a fair die, so we need to estimate the parameters of the multinomial distribution, i.e. the probability θ_i of getting i in a throw of the die. There are different strategies used for inference. Two of them are described here.

13.4.1 Maximum Likelihood

Let M be the model for which we wish to infer the parameters $\theta = \theta_i$ from set of data D. The maximum likelihood criterion maximizes $Pr(D|\theta, M)$ i.e

$$\theta^{\text{ML}} = \underset{\theta}{\text{argmax}}\, Pr(D \mid \theta, M)$$

where the $\text{argmax}_x f(x)$ gets the value of x for which $f(x)$ is maximized.

For example, given a set of points (x_i, y_i) $(i = 1, \ldots, n)$ in the plane, we wish to find a straight line $y = ax + b$ which can approximate these points with highest probability. Here, we choose the linear model $y = ax + b$, (x_i, y_i) are the data and we wish to estimate two parameters a and b.

13.4.2 Posterior Probability Distribution

By the Bayes' Theorem, we have

$$Pr(\theta \mid D, M) = \frac{Pr(D \mid \theta, M) Pr(\theta \mid M)}{Pr(D \mid M)} \tag{13.9}$$

where $Pr(\theta|M)$ is the prior probability which is to be chosen in some reasonable manner and $Pr(D|\theta, M)$ is the posterior probability for the parameters given the data and the model. If we want a specific set of parameters values for the model, we might be guided by analogy with ML and choose the *maximum posteriori probability* (MAP) estimate,

$$\theta^{\text{MAP}} = \underset{\theta}{\text{argmax}}\, Pr(D \mid \theta, M) Pr(\theta \mid M).$$

Note that we ignore the data prior $P(D \mid M)$, because it does not depend on the parameter θ.

13.5 Markov Chains

A Markov chain is a special type of stochastic process. Given a set of states $\Sigma = \{s_1, \ldots, s_n\}$, at any time t, the current state X_t only depends only on the state X_{t-1}. We can define the transition probability from s_j to s_i as

$$P_{ij} = Pr[X_t = s_i \mid X_{t-1} = s_j].$$

It is easy to show that

$$\Sigma_i P_{ij} = 1.$$

Markov chain can be used to analyze the phenomenon of C_pG islands.

References

[1] R. Durbin, S. Eddy, A. Krogh and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.