A refresher on statistical mechanics

2025-03-21

This article provides a comprehensive refresher on fundamental concepts in statistical mechanics, drawing inspiration from Leonard Susskind's lectures. Beginning with probability theory and Shannon's information-theoretic definition of entropy, we establish the mathematical foundations that bridge information theory and thermodynamics. We explore the derivation of entropy formulas using both Shannon's axioms and combinatorial approaches with Stirling's approximation. The article presents the zeroth, first, and second laws of thermodynamics, with particular emphasis on the relationship between entropy, energy flow, and temperature in interacting systems. Using calculus of variations and Lagrange multipliers, we demonstrate how entropy maximization principles lead to the uniform and Boltzmann distributions. Throughout, we supplement theoretical discussions with visual representations and detailed derivations to provide intuitive understanding of these abstract concepts. This refresher serves as an accessible entry point for readers seeking to revisit or develop a deeper understanding of statistical mechanics and its connections to information theory.

blog: https://tetraquark.vercel.app/posts/refresher_statmech/

email: quarktetra@gmail.com

Throughout my education, I took two classes on thermodynamics: one undergraduate level and one grad level. I disliked it very much in both cases. I hated the second one so much mostly because of the professor teaching the class. It was a complete torture for me. It has been quite a while since then, and my bad memories faded away along with most of what I managed to learn in these two horrible classes. I needed a refresher, and I found that on Youtube in Susskind's lectures.

I want to clearly state that all the credit goes to Prof. Susskind, and I am just reproducing portions of it for my entertainment. I occasionally deviate from his notation for selfish reasons. I also include extra calculations, derivations, and plots to unpack some of the details skipped in his lectures.

If anybody wants to read through my notes, I suggest you do it in parallel with the videos since I am not including most of the verbal discussion and some introductory material in the lectures.

This is a work in progress, and I will keep updating the post.

Probability and entropy

Consider a random variable X with possible outcomes x. We will want to keep things simple first and consider the case of discrete case, for which there are finite number of outcoumes that we can labels as $\{x_1, x_2, \cdots, x_N\}$ where N is the number of possible outcomes. For example, if we are flipping a coin $N = 1$ and $\{x_1, x_2\} = \{\text{heads}, \text{tails}\}$; if we are rolling a die, $N = 6$ and $\{x_1, x_2, x_3, x_4, x_5, x_6\} = \{1, 2, 3, 4, 5, 6\}$. We can also assign a probability to observe the outcome x_i as follows:

$$p_i \equiv P(X = x_i), \tag{1}$$

with the unit normalization

$$\sum_{i=1}^{N} p_i \equiv 1. \tag{2}$$

The entropy associated with random variable is a measure of uncertainty in the outcome. Let us consider a uniform distribution as shown in Figure 1.

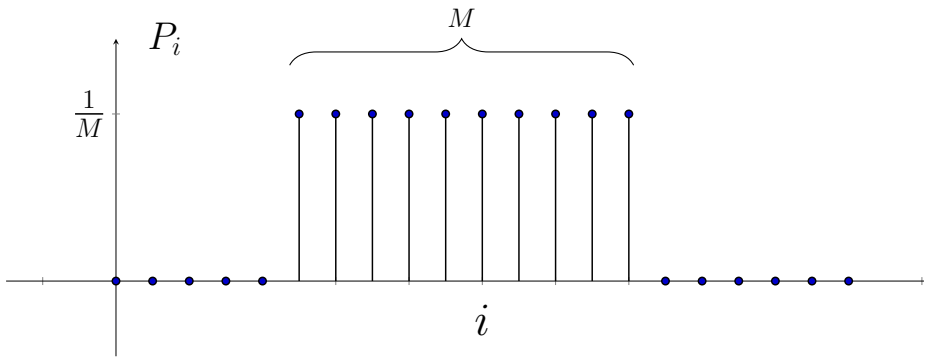

Figure 1: A probablity distribution with N states, M of which are possible with probability $\frac{1}{M}$.

For such a distribution, all the outcomes are equally likely and there are N of them. The entropy of such a distribution is defined as the logarithm of the number of possible outcomes, that is:

$$S = log(M). \tag{3}$$

For a generic distribution, we have to modify this definition. It will be modified as follows:

$$S = -\sum_i p_i \log(p_i). \tag{4}$$

We can quickly see that Eq. 4 reduces back to Eq. 3 when $p_i = \frac{1}{M}$:

$$S = -\sum_i p_i \log(p_i) = -\sum_i \frac{1}{M} \log\left(\frac{1}{M}\right) = -M\frac{1}{M}\log\left(\frac{1}{M}\right) = log(M). \tag{5}$$

In order to get to the bottom of this definition in Eq. 4, we have a couple of options, as discussed below.

Defining the entropy

This derivation comes from the father of the information theory, Claude Shannon in his ground breaking paper[1]. We define an information function, H, which takes in the probability distribution, i.e., $H = H(p_1, p_2, \cdots, p_N)$. Shannon requires the following features in H :

1. H should be continuous in the p_i.
2. If all p_i are equal, $p_i = 1/N$, then H should be a monotonically increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. The total information extracted from two events must be the sum of the information collected from each: $H(p \times q) = H(p) + H(q)$.

You can clearly see that the third requirement is begging for a *log* function, i.e. $H(p) = -log(p)$. Shannon shows that it is the only function that meets all of the requirements. Note that the negative sign in front of Eq. 4 makes sure the second requirement is satisfied. For multiple p_i, we simply sum over p_i's.

There is another way of getting the same answer by combinatorics. Let us assume that we have

Another way of getting the same result is by using the Stirling's approximation for factorial. Consider a stream of n bits consisting of 0's and 1's. If the probability of a bit being 1 is p, for large n, the average number of 1'a in such messages will be np, and the average number of 0's will be $n(1-p)$. We can easily estimate the number of different messages that can me constructed with these many 0's and 1's as $\binom{n}{np}$, and compute its log:

$$\begin{aligned} log\binom{n}{np} &= \frac{n!}{np!\, n(1-p)!} \simeq nlog(n) - n - np\, log(np) + np - n(1-p)\, log\left(n(1-p)\right) + n(1-p) \\ &= -n\left[p\, log(p) + (1-p)\, log(1-p)\right], \end{aligned} \tag{6}$$

where we used the Stirling approximation $log(n!) = n\, log(n) - n + \mathcal{O}(log(n))$.

Notes on Stirling's formula

Consider the following integral:

$$\int_0^\infty dx x^n e^{-x} = \left[(-1)^n \frac{d^n}{d\alpha^n} \int_0^\infty dx e^{-\alpha x}\right]_{\alpha=1} = \left[(-1)^n \frac{d^n}{d\alpha^n} \frac{1}{\alpha}\right]_{\alpha=1} = n!. \tag{7}$$

From this definition, we can do the following:

$$n! = \int_0^\infty dx x^n e^{-x} = \int_0^\infty dx e^{nln(x)-x}. \tag{8}$$

Let's take a close look at the function in the exponent:

$$u(x) \quad = \quad nln(x) - x, \tag{9}$$

as shown in Fig. @ref(fig:fplot). This function has its peak value at $x = n$. Note that this function appears in the exponent, under the integral. The dominant contribution to the integral will come from the domain around $x = n$. we can expand $u(x)$ around $x = n$:

$$\begin{aligned} u(x) &= nln(x) - x = nln(x - n + n) - x = nln(n[1 + \frac{x-n}{n}]) - x \\ &\simeq n\left(ln(n) + \frac{x-n}{n} - \frac{1}{2}\left[\frac{x-n}{n}\right]^2\right) - x = nln(n) - n - \frac{1}{2}\frac{(x-n)^2}{n} \equiv \tilde{u}(x) \end{aligned} \tag{10}$$

The original function and the approximated functions are plotted in **?@fig-fplot**.
n
From **?@fig-fplot**, we also notice that if we extended the x range to include negative values, the integral would not change much since $e^{\frac{1}{2}\frac{(x-n)^2}{n}}$ is rapidly decaying. Therefore we can change the lower limit of the integral from 0 to $-\infty$ to get:

$$\begin{aligned} n! &= \int_0^\infty dx x^n e^{-x} = \int_0^\infty dx e^{u(x)} \simeq \int_0^\infty dx e^{\tilde{u}(x)} = n^n e^{-n} \int_0^\infty dx e^{-\frac{1}{2}\frac{(x-n)^2}{n}} \\ &\simeq n^n e^{-n} \int_{-\infty}^\infty dx e^{-\frac{1}{2}\frac{(x-n)^2}{n}} = n^n e^{-n}\sqrt{2\pi n} = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n. \end{aligned} \tag{11}$$

?@fig-factplot shows the comparison of $n!$ with the Stirling's approximation given in Eq. 11.

Equation 4 has a summation since we assumed discrete distribution specified by index i. The

continuous version of it is straight forward to write:

$$S = -\int dx p(x) \log\left(p(x)\right), \tag{12}$$

where the integral is evaluated over the space $p(x)$ is defined. What kind of distribution would maximize the entropy? One can anticipate that it has to be the uniform distribution defined in a specific range. And we can easily prove that using calculus of variations with Lagrange multipliers. The Lagrange multiplier comes in to satisfy the normalization of the probability density, similar to Eq. 2

$$\int dx p(x) = 1. \tag{13}$$

In optimization problems with constraints, one tries to find the extrema of a function while satisfying the contraints imposed. Consider a function, $f(x, y)$, and assume we want to find the location (x_0, y_0) for which $f(x_0, y_0)$ assumes its maximum, and at the same time we want a constrain function to be satisfied: $g(x_0, y_0) = 0$. One can solve this problem using brute force:

- Require $f(x, y)|_{(x_0,y_0)} = 0$ and $g(x_0, y_0) = 0$.
- Solve these two equations with two unknowns.

Although it is technically possible to solve it this way, it may require us to invert complicated functions which might be hard to do. It gets even harder as we introduce more variables and constraints. We can do better than that!

Let us consider a contour curve of f, which is the pairs of numbers (x, y) for which $f(x, y) = k$. We want k to be as large as possible while satisfying $g(x_0, y_0) = 0$. To illustrate the method, let us take the following functions:

$$f(x, y) = y^2 - x^2, \quad g(x, y) = x^2 + y^2 - 1, \tag{14}$$

which are shown in **?@fig-lagplot**.

k

If there was no constraint, we would increase the value of k indefinitely. However, we are required to find a solution (x_0, y_0) that satisfies $g(x_0, y_0) = 0$, which means two curves have to pass through the point (x_0, y_0). As you tune the value of k, you realize that you can make the curves to intersect at different points. The optimal solution is the one at which two curves touch each other, and, for this particular example, we can graphically see that it happens at $k = 1$ and $(x_0, y_0) = (0, 1)$.

How do we solve this analytically though? Note that in this critical point, two curves are barely touching each other. More precisely, they are tangent to each other at that point, i.e.,

they have the same value and the same slope. Since the tangents are the same, the vector which is perpendicular to the tangents must be the same too. And that perpendicular vector is nothing but the gradient. Note that we are limiting ourselves to a two-dimensional problem for pedagogical reasons. The observation above holds for any dimension. Let's prove that gradient vector is indeed perpendicular to the curve. In a generic case, f can be a function of multiple variables: $f = f(x_1, x_2, \cdots, x_n)$ where $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is an n dimensional vector. The level surface of this function is composed of $\mathbf{x}$ values such that $f(\mathbf{x}_0) = k$, which defines an $n-1$ dimensional level surface. What we want to prove is that for any point on the level surface, $f(\mathbf{x}_0) = k$, the gradient of f, i.e., $f|_{\mathbf{x}_0}$ is perpendicular to the surface.

Let us take an arbitrary curve on this surface, $\mathbf{x}(t)$, parameterized by a parameter t, and assume it passes through $\mathbf{x}_0$ at $t = t_0$. On the surface $f(\mathbf{x}(t)) = f\left(x_1(t), x_2(t), \cdots, x_n(t)\right) = k$. Let's take the parametric derivative of f and apply the chain rule.

$$\frac{df}{dt} = 0 = \sum_{i=1}^{n} \left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}_0} \left.\frac{dx_i}{dt}\right|_{t_0} = \left. f\right|_{\mathbf{x}_0} \cdot \dot{\mathbf{x}}|_{t_0}, \tag{15}$$

where we defined $\dot{\mathbf{x}}|_{t_0} = \left.\frac{d\mathbf{x}(t)}{dt}\right|_{t_0}$, which is nothing but the tangent line. Therefore we conclude that the gradient is perpendicular to the tangent lines on the surface.

This exercise tells us that the gradients of the function we want to optimize is parallel to the gradient of the constraint function. That is:

$$\left. f\right|_{\mathbf{x}_0} = \lambda \left. g\right|_{\mathbf{x}_0}, \tag{16}$$

where the constant λ is the Lagrange multiplier. And keep in mind that we also need to satisfy $g(\mathbf{x}_0) = 0$ We can neatly combine these to requirements by defining a new function:

$$h(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x}), \tag{17}$$

which can be optimized by requiring

$$\left. h\right|_{\mathbf{x}_0} = 0, \quad \text{and} \quad \left.\frac{\partial h}{\partial \lambda}\right|_{\mathbf{x}_0} = 0. \tag{18}$$

The bottom line is that the constraint itself is mixed into the function that we want to optimize. The expression in 18 has equal number of equations and unknowns, so we can solve for $\mathbf{x}_0$ and λ.

You may also want to see this post for examples. Combining the condition with the target function to maximize, Eq. 12, we have the following integral to maximize:

$$I = -\int dx p(x) \log\left(p(x)\right) - \lambda\left(\int dx p(x) - 1\right) = -\int dx p(x)\left(\log\left(p(x)\right) + \lambda\right) + \lambda \tag{19}$$

Now move the function $p(x)$ to $p(x)+\delta(x)$, and require that $\delta I = 0$ for the $p(x)$ that maximizes I. This yields:

$$\delta I = -\int dx \delta p(x)\left[\log\left(p(x)\right) + 1 + \lambda\right] = 0. \tag{20}$$

Since $\delta p(x)$ is totally arbitrary, we need $\log\left(p(x)\right) + 1 + \lambda = 0$. Furthermore, as λ is just a constant, this shows that $p(x)$ is also a constant. Let's assume that we are interested in distributions defined in the range $[a, b]$. The normalization condition in Eq. 13 uniquely defines the value of the constant as $\frac{1}{b-a}$.

The laws of thermodynamics

So far we have been talking about probability distributions in the most abstract form. They can be anything: coin flip, bits in a message to be transmitted etc. Let's now switch to physical cases. In such cases, the probability distributions will be parameterized by a physical quantity, such as the average energy E.

$$S(E) = -\sum_i p_i \log\left(p_i\right), \tag{21}$$

where we are using the discrete index i and the sum. The average value of energy, E is the statistical average:

$$E = \sum_i p_i E_i, \tag{22}$$

where E_i is the energy values of the state i, and the probability of that state to be occupied is p_i. It is emphasize again that E is the average energy of the system, and it would have been more appropriate to denote it as $\bar{E}$ or $\langle E \rangle$, however, that would look very ugly in the equations. We will keep it as E and promise to remember that it is the mean value of the energy, not the energy of each level or particle. This might look recursive, but it will make more sense as we proceed.

In order to pack this law, we need to define the temperature and the energy flow. In order to do that, consider two systems which are held at temperatures T_B and T_A with $T_B > T_A$.

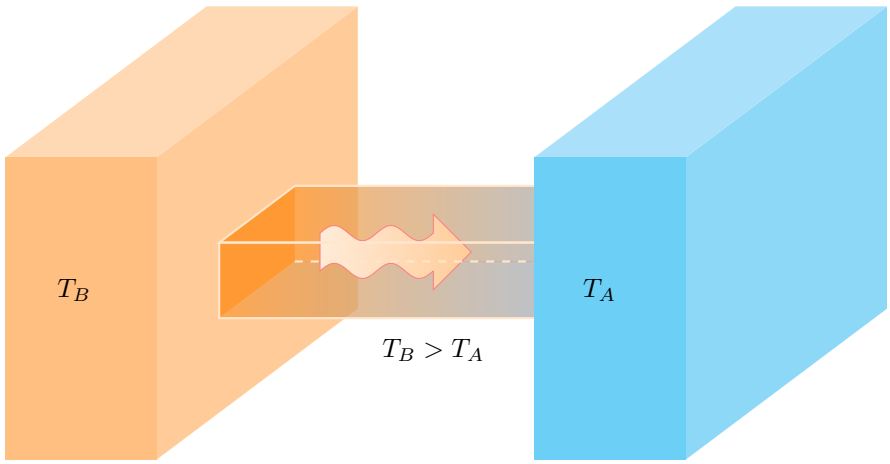

Figure 2: Two containers at temperatures T_B and T_A with $T_B > T_A$ are connected to exchange heat.

Let us state the first and second law:

- 1st law of thermodynamics: **Energy is conserved.**
- 2st law of thermodynamics: **Entropy always increases.**

The total entropy of the system is given by the sum of the entropies of the subsystems:

$$S = S_A + S_B. \tag{23}$$

The first law implies that changes in energies add up to zero:

$$dE_A + dE_B = 0. \tag{24}$$

The second law requires:

$$dS = dS_A + dS_B > 0 \tag{25}$$

The temperature of a system is defined in terms of the entropy function $S(E)$ as follows:

$$T \equiv \frac{dE}{dS}. \tag{26}$$

Inserting the definition from Eq. 26 into Eq. 24 we get:

$$dE_A + dE_B \quad = \quad 0 = T_A dS_A + T_B dS_B \implies dS_B = -\frac{T_A}{T_B} dS_A. \tag{27}$$

Putting this in the second law in Eq. 25, we get

The second law requires:

$$dS \;=\; dS_A + dS_B = \left(1 - \frac{T_A}{T_B}\right) dS_A = T_B(T_B - T_A)dS_A > 0. \tag{28}$$

Since $T_B > 0$ and $T_B > T_A$, we conclude that $dS_A > 0$, that is entropy increases. Also note that $T_A dS_A = dE_A > 0$, therefore the energy is flowing to container A from B. As T_B equalizes with T_A, the heat flow will stop and two containers will be in equilibrium. Therefore it is the temperature that determines the direction of the energy. One can extend the analysis above to a third container to state the 0^{th} law od thermodynamics:

- 0^{th} law of thermodynamics: **If two systems are both in thermal equilibrium with a third system, then they are in thermal equilibrium with each other.**

Occupation number

Consider a system with m states. An occupation number is the number of systems, n, occupying a given i^{th} state, and we will denote this number as n_i. Given m such states, i.e., $i \in \{1, 2, \cdots, m\}$, we are interested in finding the total number of possible ways to redistribute the systems among the states. This is illustrated in Figure 3.

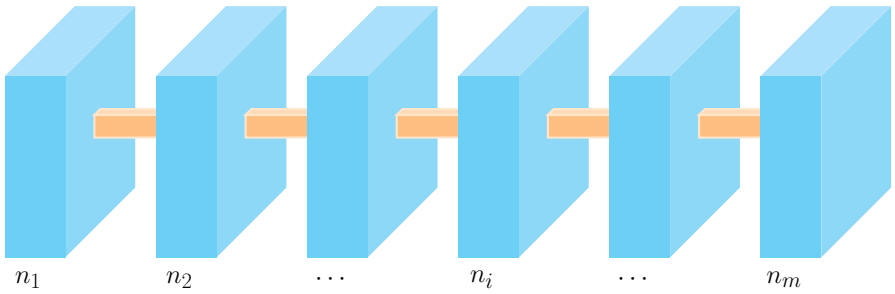

Figure 3: m boxes with given occupation numbers n_i.

We assume that the total number of occupation number is fixed, we will define it as N:

$$N \equiv \sum_{i=1}^{m} n_i. \tag{29}$$

For a randomly selected system, the probability of that system to be in state i is the ratio of the number of states in the i^{th} and the total number of states:

$$p_i = \frac{n_i}{N}, \tag{30}$$

which results in a normalized probability distribution:

$$\sum_{i=1}^{m} p_i = 1. \tag{31}$$

We also need to make sure that total energy is conserved:

$$\sum_{i=1}^{m} E_i n_i = N \sum_{i=1}^{m} E_i \frac{n_i}{N} = N \sum_{i=1}^{m} E_i p_i = NE, \tag{32}$$

where E is the average energy. While keeping the occupation numbers fixed, we can shuffle systems around to create different configurations. For N systems, we get $N!$ shufflings. However, we should remove the overcounting within the states with n_i as the occupation number. Therefore the total number of combination to create such as system is:

$$C = \frac{N!}{\prod_{i=1}^{m} n_i!}. \tag{33}$$

We can now take the log of C and use the Stirling approximation: $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$:

$$\begin{aligned} log(C) &= log(N!) - \sum_{i=1}^{m} log(n_i!) = Nlog(N) - N - \sum_{i=1}^{m} nilog(n_i) + \sum_{i=1}^{m} n_i + \mathcal{O}(1) \\ &= -N \sum_{i=1}^{m} \frac{ni}{N} log\left(\frac{ni}{N}\right) = -N \sum_{i=1}^{m} p_i log(p_i). \end{aligned} \tag{34}$$

We have shown earlier that this expression is maximized when p_i are equally likely, which was the case for the Shannon entropy. However, it is very different for this case since we have an additional constraint now, as described in Eqs. 29 and 31. We will multiply these constraints with Lagrange multipliers which we will call α and β, and subtract them from the original function in Eq. 34. Therefore the combined function becomes:

$$\begin{aligned} h(p_i, \alpha, \beta) &= -N \sum_{i=1}^{m} p_i log(p_i) - \alpha \left(\sum_{i=1}^{m} n_i - N \right) - \beta \left(N \sum_{i=1}^{m} E_i p_i - NE \right) \\ &= -N \left\{ \sum_{i=1}^{m} p_i log(p_i) - \alpha \left(\sum_{i=1}^{m} p_i - 1 \right) - \beta \left(\sum_{i=1}^{m} E_i p_i - E \right) \right\} \end{aligned} \tag{35}$$

Note that overall factors, such as the factor N in Eq. 35, do not affect the optimization. Now we just do the math:

$$\frac{\partial}{\partial p_j} h(p, \alpha, \beta) = 0 = -log(p_j) - 1 - \alpha - \beta E_j \tag{36}$$

which implies

$$p_j = e^{-(1+\alpha+\beta E_j)} = e^{-(1+\alpha)} e^{-\beta E_j} \equiv \frac{e^{-\beta E_j}}{\mathcal{Z}}, \tag{37}$$

where

$$\mathcal{Z} \equiv e^{1+\alpha}. \tag{38}$$

$\mathcal{Z}$ is referred to as the partition function, and one can think of it as the normalization factor. We can see that by imposing the normalization condition in Eq. 31:

$$\sum_{i=1}^{m} p_i = 1 = \sum_{i=1}^{m} \frac{e^{-\beta E_i}}{\mathcal{Z}}, \tag{39}$$

which results in:

$$\mathcal{Z} = \sum_{i=1}^{m} e^{-\beta E_i}. \tag{40}$$

We can figure out the relation between $\mathcal{Z}$ and E by imposing the conservation of energy constraints in Eq. 32:

$$\begin{aligned} E &= \sum_{i=1}^{m} E_i p_i = \sum_{i=1}^{m} E_i \frac{e^{-\beta E_i}}{\mathcal{Z}} = \frac{1}{\mathcal{Z}} \sum_{i=1}^{m} E_i e^{-\beta E_i} = \frac{1}{\mathcal{Z}} \left(-\frac{\partial}{\partial \beta}\right) \left[\sum_{i=1}^{m} e^{-\beta E_i}\right] \\ &= -\frac{1}{\mathcal{Z}} \frac{\partial \mathcal{Z}}{\partial \beta} = -\frac{\partial log \mathcal{Z}}{\partial \beta}. \end{aligned} \tag{41}$$

We can now compute the entropy:

$$\begin{aligned} S &= -N \sum_{i=1}^{m} p_i log(p_i) = -N \sum_{i=1}^{m} \left[\frac{e^{-\beta E_i}}{\mathcal{Z}} log\left(\frac{e^{-\beta E_i}}{\mathcal{Z}}\right)\right] \\ &= \beta E + log(\mathcal{Z}). \end{aligned} \tag{42}$$

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948 [Online]. Available: %22https://ieeexplore.ieee.org/document/6773024%22. [Accessed: 22-Apr-2003]