熵之为熵（3）：信息熵背后的物理原理（下） 精选

已有 6943 次阅读 2016-7-19 10:16 | 个人分类:科研笔记 | 系统分类:科研笔记

【特别申明：本系列博文涉及学术原创，其他写作者在引用其中思想时，敬请注明博文的网址；转载时，也请注明来源和作者。另：本系列博文的发表，并不影响作者在其他学术刊物上以论文的形式再次发表。】

（1）

上篇博文已经介绍了信息熵和马尔科夫链的熵的基本概念。在此基础上，我来解释为什么信息熵的视角对重新思考热力学和统计物理中的熵至关重要。

至今为止，我们仅仅能在固体表面，通过原子探针，观察到原子尺度的几个粒子的运动情况，遑论对大量粒子同时进行针对每个粒子的行为的即时追踪。由此可知，我们现在的热力学和统计力学理论，是以体系的宏观行为测量结果为基础，反向推测-更确切地说，是猜测-粒子运动的情况的。

回顾历史，从卡诺循环建立时流行的热质说（https://en.wikipedia.org/wiki/Caloric_theory，https://en.wikipedia.org/wiki/Caloric_theory.），到麦克斯韦和玻尔兹曼与奥斯瓦尔德之间的关于世界基本构成的争论(https://en.wikipedia.org/wiki/Ludwig_Boltzmann,https://en.wikipedia.org/wiki/Wilhelm_Ostwald)，微观世界的形象一直到1905年佩林（https://en.wikipedia.org/wiki/Jean_Baptiste_Perrin）通过实验证明了电子可以从阴极发射，并在1908年测定了Brownian Motion的相关参数，大家才确切地接受世界本质上由微粒构成。到底是粒子还是连续的媒质，都需要如此多的争论，那么，单个粒子运动行为的统计，也不是件容易的事。我认为，不论前辈们如何建立微观世界的模型，说到底，关于这微粒每个个体的统计行为，不过是从大量粒子的群体行为反推猜测出来的。

从统计上看这个问题，这个反推相当的麻烦。因为不论单个粒子遵循什么行为，统计理论则告诉我们只要粒子数量足够多，其群体行为一般都会遵循高斯分布，而考虑到能量不能取负值的话，从下面几节的仔细分析，我们也可以认为，大量粒子构成的体系的能量一定是遵循Boltzmann分布的。因此，我们反推回去，认为单个粒子的动量、动量矩、磁矩等等，遵循高斯分布，能量遵循Boltzmann分布，是完全符合情理的。倒是不遵循这样的分布反而需要点特别的理由。比如费米分布，就需要泡利不相容原理的前提；比如波色-爱因斯坦分布，则需要所有粒子都可以同时占据同一个相格的假定。 而且在后面的博文中，我们还知道，实际上这两种分布依然是依靠Boltzemann分布律推导出来的。

但是，从信息论的视角看这个问题，如果熵总是使用$p_i \ln p_i$ pilnpi作为计算的基础，那么i的取值范围，在我们分析单个粒子运动行为的情况下，将对应着单个粒子可以取的状态的多少，最后必将影响系统熵值的计算。如果这些粒子的运动状态彼此独立，我们当然可以通过单个粒子的熵的简单叠加来处理问题。如果这些粒子的运动状态彼此相关，而系统的宏观行为表现稳定，容易想见，我们可以采用高阶的无吸收的齐时马尔科夫链作为模型来处理体系的熵的计算。

细心的读者一定会追问：“如果单个粒子的运动统计行为并不是真的遵循所谓的正态分布，那么我该如何预测真实的分布呢？我们凭什么相信诸如费米分布或者波色-爱因斯坦分布真的对应了电子或者光子的统计行为呢？”坦率地说，除了已知的实验现象和模型，我也不知道有什么理由相信这些分布。但是，Fisher的信息和Jaynes的原理则会告诉我们，我们现在相信的，就是最靠谱的。正是这个理由，也使我们相信，熵，或者说按照$p_i \ln p_i$ pilnpi定义的熵，具有广泛的实用性，而信息论为我们提供了足够的工具来来完成熵的计算。

本篇博文，以理想气体分子体系的熵的计算为主线，我们首先将解释正态分布具有广泛适应性，以及Boltzemann分布对之能量分布有广泛适应性的缘由。然后我们来看看运动独立和不独立的体系的熵的计算，当然，除了介绍典型序列，我们又要使用马尔科夫链模型。最后我们来看Fisher信息、最大似然法和Jaynes的最大熵原理，以便对我们的处理手法更具数学上的信心。

以下，除非我们特别指明，“熵”是指信息熵。

（2）

对于频率派的统计学家(https://en.wikipedia.org/wiki/Frequentist_probability)而言，如果我们不断进行某个实验，且这些实验彼此间没有关联，即彼此独立，所谓概率，即是某个观察量的某个值出现的频率。这个概率的观点，体现在大数定律（https://en.wikipedia.org/wiki/Law_of_large_numbers）上，如下：

若有一组独立同分布（i.i.d,https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables）的随机变量{$X_i : i = 1, 2, \dots, n$ xi:i=1,2,...,n}，且$E[X_i] = \mu$ E[Xi]=μ,并定义，

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
X¯=1n∑i=1nXi

则有：

$$\overline{X} \to \mu \text{ when } n \to \infty$$
X¯→μ when n→∞

而考察$\overline{X}$ X¯的分布，如果X_i Xi存在方差σ^2 σ2,则我们得到中心极限定理（https://en.wikipedia.org/wiki/Central_limit_theorem），如下：

$$\overline{X} \xrightarrow{d} N[\mu, \frac{\sigma^2}{n}] \text{ when } n \to \infty$$
X¯→dN[μ,σ2n] when n→∞

其中箭头上的dd表示是$\overline{X}$ X¯"概率分布收敛于"（https://en.wikipedia.org/wiki/Convergence_of_random_variables#Convergence_in_distribution），$N[\mu, \frac{\sigma^2}{n}]$N[μ,σ2n]表示均值为μμ，方差为$\frac{\sigma^2}{n}$σ2n的高斯分布。

中心极限定理表明，只要你统计的随机变量彼此不关联或者关联程度较弱，从其加和后平均的随机变量$\overline{X}$ X¯来分析，我们总是得到高斯分布。故而反向猜测，不论体系内粒子的微观状态如何分布，我们认为其是高斯分布总是合理的。

比如，我们知道，麦克斯韦的理想气体的速度分布如下：

$$dN_{\nu_x,\nu_y,\nu_z} = N(m/2\pi kT)^{3/2}exp(-m(\nu_x^2+\nu_y^2+\nu_z^2)/2kT)d\nu_x d\nu_y d\nu_z$$ dNνx,νy,νz=N(m/2πkT)3/2exp(−m(νx2+νy2+νz2)/2kT)dνxdνydνz

（3-1）

其中$\{\nu_x, \nu_y, \nu_z\}$是一个气体分子在空间坐直角坐标的x、y、z轴上投影的速度分量，mm是一个气体分子的质量，kk是Boltzmann常数，N N是体系内气体分子的总数，TT是体系的温度，而dN_{ν_x,ν_y,ν_z}dNνx,νy,νz是速度在$\{\nu_x, \nu_y, \nu_z\}$νx,νy,νz到$\{\nu_x + d\nu_x, \nu_y + d\nu_y, \nu_z + d\nu_z\}$νx+dνx,νy+dνy,νz+dνz范围内的气体分子数目。

对比高斯分布的密度函数$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}exp(-\frac{x-\mu}{2\sigma^2})$f(x|μ,σ2)=12σ2πexp(−x−μ2σ2),即知单个粒子在x、y、z每个坐标上的速度概率分布是彼此独立的、均值为零、方差为kT/mkT/m的高斯分布。

而直接利用单个分子的能量$\varepsilon = m(\nu_x^2 + \nu_y^2 + \nu_z^2)/2$.=m(νx2+νy2+νz2)/2，通过（3-1）公式及基于球壳的多重积分，容易推知：

$$dN_\varepsilon = 2N\sqrt{\varepsilon/\pi}(kT)^{-3/2}exp(-\varepsilon/kT)d\varepsilon$$ dNε=2Nε/π(kT)−3/2exp(−ε/kT)dε (3-2)

（3-2）即理想气体的boltzemann分布。理想气体分布的推导可见https://en.wikipedia.org/wiki/Maxwell–Boltzmann_distribution。

从这个例子我们可以看出，只要以（3-1）为前提，推出（3-2）是自然而然的事。而由于物理学中熵本质上处理的是离散的状态，我们可以将一个气体分子可以在的状态写作是：

动量（速度）：$\{\nu_x, \nu_y, \nu_z\}$νx,νy,νz}到$\{\nu_x + d\nu_x, \nu_y + d\nu_y, \nu_z + d\nu_z\}$νx+dνx,νy+dνy,νz+dνz}；位置：$\{x, y, z\}$x,y,z}到$\{x + dx, y + dy, z + dz\}$x+dx,y+dy,z+dz}。那么这个由三维几何空间和三维动量（速度)空间构成的空间的六维空间称为"相空间"，而这里表示的六维空间中的一个微元称为一个"相格"。适当地划分或者规定$d\nu_x d\nu_y d\nu_z dxdydz$dνxdνydνzdxdydz的大小，我们就可以将整个相空间离散化，变成一个由一堆六维小方块堆积而成的空间。而每一个小方块就代表了一个气体分子可以取的一种状态,这里我们记为状态ii,而玻尔兹曼分布律，就表达为：

$$p_i \propto exp(-\varepsilon_i/kT)$$ pi∝exp(−εi/kT) (3-3)

（3）

是否有了公式（3-1）到（3-3），对于经典的理想气体分子构成的体系，就可以估算其分布呢？

先考察其动量。对于每个气体分子而言，其动量将遵循Boltzemann分布率，而其每个分子与环境或者其他分子间除了碰撞，我们也找不出其他因素来影响其行为。因此，我们没有任何理由假设各个分子的动量间有彼此约束的关系，故此，我们可以认为各个分子之动量作为随机变量，将彼此统计独立。这个假定与所有处理经典粒子的统计物理学家相同。

先考察一个单维的分子的速度分布，比如先考察 xx维的速度分布，将其按$d\nu_x$dνx离散化，即将$d\nu_x$dνx以有限大小的$\Delta\nu_x$ Δνx代替，并以$\nu_{x,i}$ νx,i表示第ii个$\Delta\nu_x$ Δνx区间的平均值，则有单个气体分子在x维占据第i个区间$[\nu_{x,i}, \nu_{x,i} + \Delta\nu_x]$[νx,i,νx,i+Δνx]的概率：

$$p_i = (\frac{m}{2\pi kT})^2 exp(-\frac{m\nu_{x,i}^2}{2kT})\Delta\nu_x$$ pi=(m2πkT)2exp(−mνx,i22kT)Δνx （3-4）

由（3-4）可以求出单个气体分子在速度空间x维的熵：

$$H_{\nu_x} = -\sum_i p_i \ln p_i = -\ln\frac{m^{1/2}\Delta\nu_x}{(2\pi kT)^{1/2}} + 1/2$$ Hνx=−∑ipilnpi=−lnm1/2Δνx(2πkT)1/2+1/2 (3-5)

而一个理想气体分子，不用考虑转动，其速度有三个维度，且这三个维度彼此统计独立，熵的形式相同。因此一个分子的速度的熵为各个维度熵的叠加，为：

$$H_\nu(1) = -\ln\frac{m^{3/2}\Delta\nu}{(2\pi kT)^{3/2}} + 3/2$$ Hν(1)=−lnm3/2Δν(2πkT)3/2+3/2 (3-6)

式中$\Delta\nu = \Delta\nu_x\Delta\nu_y\Delta\nu_z$ Δν=ΔνxΔνyΔνz。最后考虑N个气体分子，显然其运动是彼此独立的，故有：

$$H_\nu(N) = -N\ln\frac{m^{3/2}\Delta\nu}{(2\pi kT)^{3/2}} + 3N/2$$ Hν(N)=−Nlnm3/2Δν(2πkT)3/2+3N/2 (3-7)

现在来理想气体分子考虑几何空间分布的熵。

当我们考察经典粒子的几何空间分布的时候，我们能够清楚察觉到我们原有的经典粒子的模型明显和实际情况有出入：当一个理想气体分子占据了一个小的区域的时候，另外一个理想气体分子要在此区域存在的概率明显会降低，因为两个气体分子会有很高的概率对撞，然后离开，所以当我们拿个相机不断给气体分子照相的时候，绝大多数情况下，我们看到的分子们是趋向于均匀分布的。从统计上看，当一个分子占据了一个小区域时，另一个分子在此区域出现的概率就大大降低。为了建立简单易处理的模型，我们可以认为，某个气体分子占据空间的一个微元的时候，这个微元就不能被其他气体分子占据，这个结果将直接破坏关于描述分子位置的随机变量的统计独立性。这个假定，我几乎没有看到有与我相同的，只有李政道的统计物理学讲义中提到过类似的模型，他认为当气体分子作为一个一个微小的粒子，大多数情况下，彼此间的距离将大于其德布罗意波长，因而是彼此可区分的，且在空间将占据不同位置。

我们用两种方式来分析气体分子几何空间分布的熵。

第一种方式比较传统。假定在体积V中有N个气体分子，体积元大小为ΔV ΔV。则我们认为N个分子分布于$V/\Delta V$ V/ΔV个体积元中，故总的分布方式为$C^N_{V/\Delta V}$ CV/ΔVN，且每种方式出现概率相等，故有气体分子几何空间分布的熵为：

$$H_V(N) = \ln C^N_{V/\Delta V}$$ HV(N)=lnCV/ΔVN

$$\approx N\ln\frac{V}{\Delta V N} + N$$ ≈NlnVΔVN+N (3-8)

推导过程中，使用了Stirling公式；又认为气体分子足够少，而有$\Delta V N/V$ ΔVN/V是小量；最后的结果，使用了$\Delta V N^2/V \approx 0$ ΔVN2/V≈0。

第二种方式，则很像量子统计的一般处理。我们认为共有$V/\Delta V$ V/ΔV个体积元，N个粒子。故每个体积元被一个粒子占据的概率为$\Delta V N/V$ ΔVN/V,空着的概率为$1-\Delta V N/V$ 1−ΔVN/V，则一个体积元的熵为：

$$H_V(Cell) = -(1-\Delta V N/V)\ln(1-\Delta V N/V) - (\Delta V N/V)\ln(\Delta V N/V)$$ HV(Cell)=−(1−ΔVN/V)ln(1−ΔVN/V)−(ΔVN/V)ln(ΔVN/V)

$$\approx (\Delta V/V)(N\ln\frac{V}{\Delta V N} + N)$$ ≈(ΔV/V)(NlnVΔVN+N) (3-9)

而每个体积元彼此是否被占据是彼此统计独立的，故有：

$$H_V(N) = (V/\Delta V)H_V(Cell) \approx N\ln\frac{V}{\Delta V N} + N$$ HV(N)=(V/ΔV)HV(Cell)≈NlnVΔVN+N (3-10)

在第二种推导过程中，我们没有使用Stirling公式，处理显得简洁得多。

一个气体分子的速度分布和其所在位置是统计无关的，因此结合公式（3-7）和（3-9），我们得到整个体系的熵：

$$H_s(N,V) = H_v(N) + H_V(N)$$ Hs(N,V)=Hv(N)+HV(N)

$$= N\ln(V/N) + \frac{3}{2}N\ln T + \frac{3}{2}N\ln\frac{2\pi k}{m(\Delta\nu\Delta V)^{2/3}} + \frac{5}{2}N$$ =Nln(V/N)+32NlnT+32Nln2πkm(ΔνΔV)2/3+52N (3-11)

将整个体系的熵乘以Boltzmann常数，并按照量子统计的要求，认为一个相格的大小由不确定性原理确定，即$m^3\Delta\nu\Delta V = h^3$ m3ΔνΔV=h3，则整个体系的统计力学熵为：

$$S = kH_s(N,V)$$ S=kHs(N,V)

$$= k(N\ln(V/N) + \frac{3}{2}N\ln T + \frac{3}{2}N\ln\frac{2\pi km}{h^2} + \frac{5}{2}N)$$ =k(Nln(V/N)+32NlnT+32Nln2πkmh2+52N) (3-12)

（3-12）即我们常用的理想气体分子的统计力学熵计算公式。

从我们的推导很容易看出，我们将不再会有吉布斯佯谬，这件事后述。

(4)

表面上看，公式（3-8）和（3-9）所依据的推导模型没有区别。但是，仔细思考，我们会发现，第二种办法中的粒子数并不固定，我们只是说平均来看，这个体系里有N个粒子。回顾我们整个的关于理想气体分子体系的推导过程，我们应该留意到，实际上我们的体系能量也是不固定的，只是由T来确定了平均能量。对于熟悉统计物理的朋友， 可以看到我们实际上是使用了正则系综和巨正则系综。但是我们的处理过程是从正态分布到Boltzmann分布律，其出发点和整个统计物理的出发点完全不同。那么是什么数学原理保证了这个过程的合理性呢？或者更具体地问，为什么第二种办法不需要使用Stirling公式呢？

这里我们要谈到εε-典型序列（https://en.wikipedia.org/wiki/Typical_set），如下：

若有序列 $x_1, x_2, \ldots, x_n$x1,x2,...,xn，其中序列的每个元素的状态都是从只有有限种状态的随机变量X中抽取按概率抽取，那么如果 $\exp(-n[H(X)+\varepsilon] \le Pr(x_1, x_2, \ldots, x_n) \le \exp(-n[H(X)-\varepsilon]$ exp(−n[H(X)+ε]≤Pr(x1,x2,...,xn)≤exp(−n[H(X)−ε]，则此序列称为X的εε-典型序列。

从数学上可以证明，当$n \to \infty$ n→∞，则$\varepsilon \to 0$ε→0，且典型序列的集合出现的概率趋于1，而每个序列出现的概率趋于$1/m$1/m,其中m是典型序列集合中典型序列的个数。这个性质称为典型序列的渐进等分割性（https://en.wikipedia.org/wiki/Asymptotic_equipartition_property）。

我们使用的两种处理了理想气体分子的几何空间分布方法，都是针对粒子的大量行为的，所以自然地体现了典型序列的特征。第一种方法，使用了古典的概率技术，需要计算排列组合，但是，通过Stirling公式，约去了那些非典型的序列，最后当然得到了典型序列的结果；第二种方法，一开始就避免了求阶乘的问题，所以不需要Stirling公式来保证非典型序列的约减，但是也带来了粒子数N不固定的问题，但是由于典型序列最后的出现概率一致，也就说明其出现粒子数一致，所以也避免了那些不固定于粒子数N的非典型序列。

有了典型序列及其性质，我们很容易明白，当面对大量粒子的群体性行为的时候，如果这些粒子的行为是彼此统计独立的或者近独立的，我们并不需要知道具体的分布，仅仅知道了熵$H(X)$ H(X)，我们就能通过等概率出现的大量典型序列来统计粒子的行为；当然，如果我们知道了熵$H(X)$ H(X)，我们也可以反向推知单个粒子运动和位置的概率分布情况。

Boltzmann建立统计物理基石的年代，离信息论的建立还隔着半个多世纪。正是Boltzmann等概率的假设，天才地预测了粒子群体的行为。不论Boltzmann的假设有何种局限，他的H定理受到多少责难，他依然是现代物理的真正揭幕人，是我心目中真正的物理英雄。对我而言，他的思想，是超越牛顿和爱因斯坦的。我现在做的事，无非是揣测，在当今的年代，他会怎么做。

（5）

如果一个体系的粒子间的相互作用破坏了粒子运动行为的相互独立性，那么我们应该想办法变换空间或者模型来使这些行为描述独立，比如通过谐振子引入来讨论固体的原子间震动，那里我们就是统计谐振子，或者说统计声子，来处理相关问题。

但是，对于有的问题，比如体系的磁学性能分析，或者比如引入了范德瓦尔兹力（https://en.wikipedia.org/wiki/Van_der_Waals_force）的气体分子体系分析，统计独立的条件很难引入。因此，我在后面准备用马尔科夫链处理。现在，我不加证明地引入关于马尔科夫链的典型序列（在后面具体要用到时我再给出证明）：

若序列$x_1, x_2, \ldots x_n$x1,x2,...xn是m阶马尔科夫序列集合的一个样本，而且m阶马尔科夫链的转移概率为 $Pr(X_{m+1} /X_{1} X_{2} \ldots X_{m})$ Pr(X_{m+1} /X_{1} X_{2} ...X_{m})，则此序列是此马尔科夫序列集合的εε-典型序列的条件是：

$$\exp(-n[H(X_{m+1} /X_{1} X_{2} \ldots X_{m})+\varepsilon] \le Pr(x_1, x_2, \ldots, x_n) \le \exp(-n[H(X_{m+1} /X_{1} X_{2} \ldots X_{m})-\varepsilon]$$

exp(−n[H(X_{m+1} /X_{1} X_{2} ...X_{m})+ε]≤Pr(x1,x2,...,xn)≤exp(−n[H(X_{m+1} /X_{1} X_{2} ...X_{m})−ε]

这个典型序列的性质与前面的典型序列的性质是相同的。

有关马尔科夫链的典型序列陈述于此，以备后用。

(6)

至此，我们已经从信息论角度，阐明了熵的重要地位。但是，读者难免会对粒子个体的统计行为的推断产生疑问。

而Jaynes的最大熵原理（https://en.wikipedia.org/wiki/Principle_of_maximum_entropy）就告诉我们，在满足已知约束的条件下，满足信息熵最大的概率分布，即是我们得到的与真实情况最接近的分布。

为什么会这样呢？

我们知道，统计上常用的最大似然法（https://en.wikipedia.org/wiki/Maximum_likelihood_estimation），就是先通过约束条件假定一个分布，然后通过取对数的办法，来求取分布的相关参数的。数学上已经证明，实际上最大似然法的本质，就是使分布的熵在约束条件下取极大值。

早在1925年，Fisher（https://en.wikipedia.org/wiki/Ronald_Fisher）提出Fisher信息，就是最大似然法使用和推广的一个最重要的里程碑。在Fisher的统计理论在生物上广泛成功的应用，为最大似然法的实际有效性提供了最好的证明。

而相同的方式，在统计物理上，我们也是经常使用的。

因此，这些成功的案例，让我们应该对最大熵原理抱有信心。虽然，对最大熵原理的争议，几十年未曾停止，但是，在我看来，那是因为，应用者未能察觉对象的重要约束条件而造成的结果。

熟悉统计物理的读者，应该清楚地知道，Boltzmann最可几分布的求取，就是使体系在粒子数和能量固定的情况下，通过拉格朗日乘子法，使体系的熵最大。因此，Boltzmann分布实际上和最大熵原理完全契合的。这将使我们使用Boltzmann分布律，更具数学上的信心。

（7）

本篇博文从信息论阐明了信息熵应用于群体粒子的行为的重要地位，我们也成功地推出了理想气体体系的熵。大数定律以及由此可证明的典型序列性质和最大似然法，从数学保证了这些推断的合理性。这种处理方式，都可以用最大熵原理予以概括。

在博文写作过程中，得到文克玲老师和其他两位网友的热心帮助，特此致谢。