# 熵之为熵（2）：信息熵背后的"物理"原理（上） 精选

已有 6091 次阅读 2016-6-25 18:24 | 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）

这两天要准备学生的考试、答辩以及改考卷，所以没有太多的时间来写博客。没想到热力学突然在科学网上热起来了，各位朋友都开始谈论起来，从卡诺循环到不可逆过程，让我恨不得马上也来谈谈。

但是，读者诸君，按照我们的约定，我得先谈信息熵，以便我们有个信息的视角，来真正讨论热力学和统计物理。

大家一定会觉得我比较烦。比如我的一个同事就说:“要谈热力学或者统计物理，你就谈吧，干嘛老在信息论上折腾？我们学物理的，不懂，也不关心。你能不能别搞这么复杂？好的物理是straight foward的。”

读者诸君，并非我想把问题复杂化。实在是相关问题，有好多先贤前辈都讨论过。之所以不能让我满意，是因为他们要么就是搞信息的，对热力学不甚了了；要么是搞物理的，对信息论也未能深入考虑。比如，麦克斯韦妖，布里渊就讨论过，并且还变成了广泛共识；（https://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory，https://en.wikipedia.org/wiki/Maxwell%27s_demon）而Gibbs佯谬（https://en.wikipedia.org/wiki/Gibbs_paradox），科学网上张学文老师还专门思考过，在国内还有很多人赞同张老师的观点，他们还把热力学熵搞成的条件熵普适化。（必需说明，我是同意张老师对佯谬本身的意见的，但是我认为他的意见较初略，而后来别的学者的处理并不具有一般性，则是我不能完全同意的。）

作为一个工程师，我要求理论逻辑严密，模型实际可证，结果最好可测量，所以，我实在不能接受如此高大上的理论；所以，当我使用“物理”一词的时候，我的潜台词，总是对着模型可证和结果可测量去的。

故此，我非常愿意以一个工程师的身份，来介绍信息熵背后暗含的“物理”原理。

（2）

鉴于我已经不止一篇博文讨论过信息熵的基本含义，所以我这里就会直接分析相关问题。 如果读起来不甚了了的作者，请往前翻看我的博文。

概言之，信息熵的计算公式（这里我们以信息论的习惯，用H表示信息熵）为：

$$H = -\sum_{i=1}^{N} p_i \ln p_i \quad (2\text{-}1)$$

这里着重强调的是$-\ln p_i$一项，或者说$\ln(1/p_i)$一项，其代表的是出现第i个状态的信息量。其中$1/p_i$代表信息量的大小-显然，一种状态出现的概率越小，此状态一旦出现，带来的信息量就越大。而其取对数，则代表了关于信息量的最终换算。如果我们采用$\log_2(1/p_i)$,这个解释要简单得多-这代表表达一种状态至少需要的编码长度（表明采用两种符号，比如“0”和“1”，来编码时，需要多少“位”）。而剩下的$p_i$一项，则是表明出现某种编码的可能性。如此一来，熵就变成了表达体系出现各种状态的编码的平均码长。

观察公式（1），我们注意到如下几个特点：

（a）体系状态具有抽象性。也就是说，具体如何算着是体系的状态，跟你具体要探测什么量或者处理什么量有关。比如对于理想气体分子而言，你的统计是要考虑分子所处的几何位置，还是不考虑呢？这样的情况处理将会不同。对于二维Ising磁铁，所谓状态，你是考虑磁矩的方向呢，还是只考虑磁矩间的交换作用呢？这样的情况也有所不同。按照不同分类标准，计算出来的熵值也会不同。

（b）体系状态具有组合性。由（a），我们很容易想象，不同标准划分的状态可以彼此组合，成为新的体系状态的描述。比如假设有一个体系能够不断向外输出英文字母，你可以考察每输出一个字母的情况，然后将一个字母输出的情况作为体系的一种状态；也可以看一定长度的字母串来考虑体系的状态。当然，你还可以考虑每个字母的颜色，然后将之考虑进来，重新划分关于“状态”的标准。这一点尤为重要，也是我们要讨论马尔科夫链的重要理由，以下会有比较详细的分析。

（c）体系状态是离散的。从公式（1）标号i来看，我们很容易理解，熵的理论基础是建立在离散的基础上的。在实际中，我们总是要分析连续的信号，因此，确实存在所谓“连续熵”的概念。从香农的信息论出发，推出连续信道的编码定理及香农极限，是一个使用连续熵的典型例子。但是对物理学家而言，由于量子力学的存在，离散的概念对量子统计而言，是基础的要求。我们基本不需要和连续熵打交道，所以连续熵不是件重要的事情，我仅简介到此。

（d）所谓熵，无非是个对物理量$\ln(1/p_i)$ ln（1/pi）求平均的概念。从这个角度出发，熵并不具有比其它物理量更重要更根本的地位。但是在实际应用中，为什么它的位置如此重要呢？这也是本部分要讨论的重中之重，我们要通过马尔科夫链、统计中的大数定理以及Fisher信息和最大似然法来分析信息熵的中心地位。也会从这个角度出发，分析Jaynes的最大熵原理如此重要的原因。

（2）

在信息论中，总是要分析和刻画两个不同的体系之间的信息流通情况，因此会出现条件熵、平均互信息等等概念。而我们这里要将信息论的思想用到统计力学和热力学中间去。所以，我就会脱离开通信的术语，较抽象地来介绍这些概念。

设若有两个系统X和Y，X中出现一个状态$a_i (i = 1$ai(i=1 to N)N)的概率记为$p(a_i)$p(ai)，而Y中出现一个状态$b_j (j = 1$bj(j=1 to M)M)的概率记为$p(b_j)$p(bj)，而在X出现状态$a_i$ ai的情况下Y出现状态$b_j$ bj的条件概率记为$p(b_j/a_i)$ p(bj/ai)(一般的概率书中，条件和结果之间使用分隔记号是“|"，不是“/”，后面我们会看到这个分隔记号的好处。)，则根据公式（1），我们可以考虑将两个体系合起来得到的一个总的系统XY的熵为（以下，按照惯例，信息熵采用H(.)标记）：

$$H(XY) = -\sum_{i,j} p(a_i b_j) \ln p(a_i b_j)$$ H(XY)=−∑i,jp(aibj)lnp(aibj)

$$= -\sum_{i,j} p(a_i) p(b_j/a_i) (\ln p(a_i) + \ln p(b_j/a_i))$$ =−∑i,jp(ai)p(bj/ai)(lnp(ai)+lnp(bj/ai))

$$= -\sum_{i} p(a_i) \ln p(a_i) - \sum_{i,j} p(a_i b_j) \ln p(b_j/a_i)$$ =−∑ip(ai)lnp(ai)−∑i,jp(aibj)lnp(bj/ai)

$$= H(X) + H(Y/X)$$ =H(X)+H(Y/X) (2-2)

式中，我们使用了$H(Y/X)$ H(Y/X)来记$-\sum_{i,j} p(a_i b_j) \ln p(b_j/a_i)$ −∑i,jp(aibj)lnp(bj/ai),称其为**条件熵**，更准确地说是X条件下Y的条件熵。

依据同样的推理，我们还可以得到：

$$H(XY) = H(Y) + H(X/Y)$$ H(XY)=H(Y)+H(X/Y) (2-3)

根据（2-2）、（2-3），很容易想象下图：

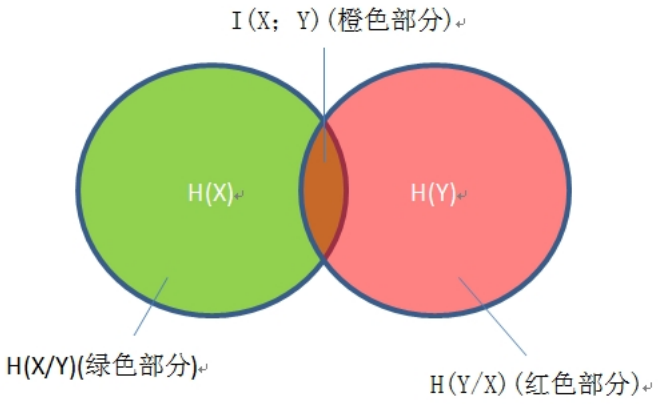


图1 条件熵与互熵示意图。

在图1中绿色圆圈是H(X)，红色圆圈是H(Y)，绿色圆圈除掉其与红圈交叠的部分是H(X/Y)，红色圈去掉绿圈的交叠部分是H(Y/X)。而交叠部分，用橙色表示，记为I(X;Y),称为**平均互信息**，更准确地说，是X与Y的平均互信息。

根据这个图，很容易推断：

$$I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X) = H(X) + H(Y) - H(XY)$$ I(X;Y)=H(X)−H(X/Y)=H(Y)−H(Y/X)=H(X)+H(Y)−H(XY) (2-4)

根据公式（2-2）和公式（2-4），有：

$$I(X;Y) = H(X) - H(X/Y)$$ I(X;Y)=H(X)−H(X/Y)

$$= -\sum_{i} p(a_i) \ln p(a_i) + \sum_{i,j} p(a_i b_j) \ln p(a_i/b_j)$$ =−∑ip(ai)lnp(ai)+∑i,jp(aibj)lnp(ai/bj)

$$= \sum_{i,j} p(a_i b_j) \ln (p(a_i/b_j)/p(a_i))$$ =∑i,jp(aibj)ln(p(ai/bj)/p(ai)) (2-5)

（3）

为了理解上面的各种基本概念，也是为顺利引进关于马尔科夫链的分析做准备，我们来看一个简单的例子。

比如YC写或者不写风花雪月，邢老人家会去捣乱或者不去捣乱。如果YC在一篇文章中写了风华雪月，我们记YC处于的状态为1，没写的话记YC的状态为0。YC出现1的概率为$p_{YC}(1)$pYC(1)=0.8,当然其出现0的概率就为$p_{YC}(0)$pYC(0)=0.2。再来看邢老人家，其去捣乱记为1，不去捣乱记为0。则其捣乱的条件概率分别为

$p_{XL}(1/YC1)$pXL(1/YC1)=0.99,$p_{XL}(1/YC0)$pXL(1/YC0)=0.7;当然其不去捣乱的条件概率也很容易算出来$p_{XL}(0/YC1)$pXL(0/YC1)=0.01,$p_{XL}(0/YC0)$pXL(0/YC0)=0.3。因此我们很容易算出以下各种熵来：

$$H(YC) = -0.8\ln(0.8) - 0.2\ln(0.2) = 0.500402(Nat/State) = 0.721928Bit/State$$ H(YC)=−0.8ln(0.8)−0.2ln(0.2)=0.500402(Nat/State)=0.7

$$H(XL/YC) = -0.8(0.99\ln(0.99) + 0.01\ln(0.01)) - 0.2(0.7\ln(0.7) + 0.3\ln(0.3))$$ H(XL/YC)=−0.8(0.99ln(0.99)+0.01ln(0.01))−0.2(0.7ln(0.7)+0.3ln(0.
$$= 0.166974(Nat/State) = 0.240893Bit/State$$

$$H((XL)(YC)) = H(YC) + H(XL/YC) = 0.667377(Nat/State) = 0.962821Bit/State$$ H((XL)(YC))=H(YC)+H(XL/YC)=0.667377(Nat/State)=0.962

$$H(XL) = -(0.8 \times 0.99 + 0.2 \times 0.7)\ln(0.8 \times 0.99 + 0.2 \times 0.7)$$ H(XL)=−(0.8×0.99+0.2×0.7)ln(0.8×0.99+0.2×0.7)−(
$$- (0.8 \times 0.01 + 0.2 \times 0.3)\ln(0.8 \times 0.01 + 0.2 \times 0.3) = 0.248435(Nat/State) = 0.358415Bit/State$$

$$I(XL;YC) = H(XL) - H(XL/YC) = 0.0814605(Nat/State) = 0.117523Bit/State$$ I(XL;YC)=H(XL)−H(XL/YC)=0.0814605(Nat/State)=0.117523Bi

例子中计算采用的自然对数ln，在信息论中，我们对使用自然对数来计算的，其对应的信息量单位为Nat，其对应的熵的单位为Nat/State,代表体系出现一个状态时体系平均输出的信息量。为了便于大家直觉上比较容易接受，我加入单位转换$1Nat = \log_2 e(Bit)$1Nat=log2e(Bit),来看看平均一个状态对应多少“位”信息。

通过这个例子，我们知道，原来YC八成爱写风花雪月，*H（YC）*的平均信息量只有约0.72“位”，但是邢老人家9成爱捣乱，所以*H（XL）*相当低，只有约0.36“位”；而*H((XL)(YC))*说明YC爱写风花雪月和邢老人家爱捣乱就是件比较确定的事，本来应该用2位来代表的两位朋友的状态，其实约0.96位就搞定了。当然YC的风华雪月程度和邢老人家的捣乱程度是有关联的，这通过*I（XL；CY）*反映出来，还算关系紧密，约0.12“位”。

（4）

马尔科夫链的概念由俄国数学家安德烈.马尔科夫（https://en.wikipedia.org/wiki/Andrey_Markov）提出。那么什么是马尔科夫链（https://en.wikipedia.org/wiki/Markov_chain）呢？

我们可以将马尔科夫链看做系统在一系列状态间的随机转移过程。马尔科夫链的特殊之处在于，体系从一种状态向另一种状态跳跃的概率-这称为转移概率-只依赖于体系当前所处的状态以及有限步跳转以前的状态，而与体系再往前所处的状态无关。如果一个体系向下一步跳转的概率仅依赖当前状态以及往前M-1步的状态，我们就称此马尔科夫链为M阶马尔科夫链。

我真正要讨论的是一个常用的马尔科夫链模型，它有以下特点：首先它是**齐时**（time-homoenous）的，也就是说，她的所有转移概率只与前面M-1步状态以及当前状态，而与这些状态处在何种时刻无关；其次它是无“吸收节点”的，或者说无“吸收状态”的，也就是说，它不会跳到一个这样的“吸收状态”，体系会永远停留与此，再也不能向其他状态跳转。

现在我们考虑为一个像理想气体或者Ising磁铁这样的体系的各种状态建立模型。容易理解，气体分子的速度在体系内随着时间或者空间位置变化出现了不同的大小或者方向分布，或者磁铁的磁矩会随时间或者空间位置的变化而有不同的能量状态和磁矩取向的变化。每个分子速度或者每个磁矩的取向各各不同，但是单独考察每个气体分子或者磁矩，我们发现每个气体分子的速度或者磁矩取向在体系内的每一个时刻或者位置的状态的信息熵都是一样的，而不同时间或者位置上的分子速度或者磁矩，由于气体分子的相互碰撞，或者磁矩之间的相互交换作用，在一定时间尺度或者位置范围内，彼此关联。这样，我们就可以用无吸收的“齐时的马尔科夫链”（https://en.wikipedia.org/wiki/Time-homogeneous_Markov_chain）来描述这些分子或者磁矩，其熵之状况如下图：

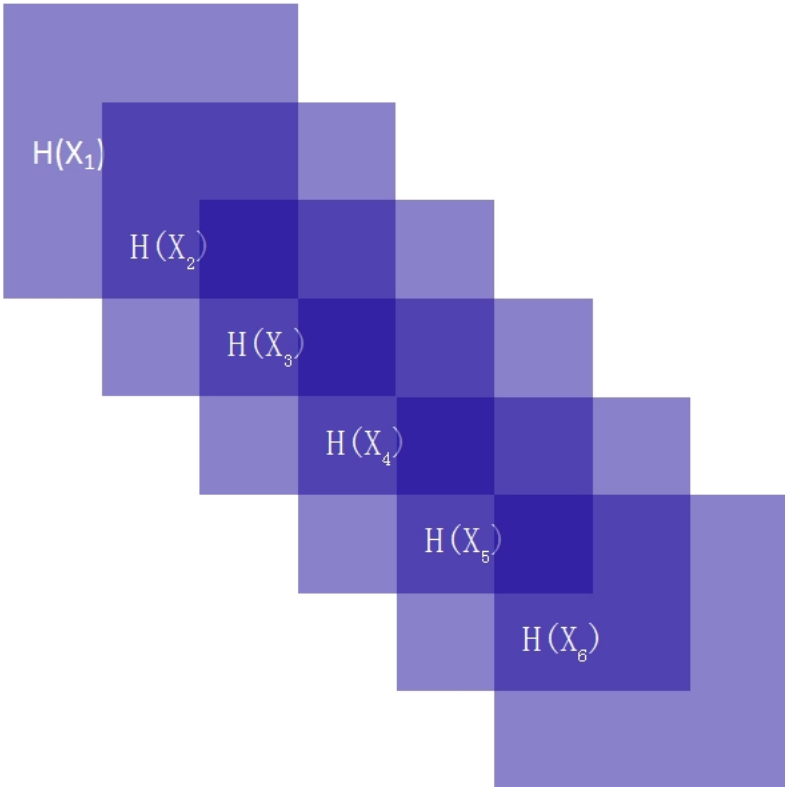


图2 马尔科夫链的熵的示意图

在图中，我展示了一个2阶马尔科夫链的熵的局部情况。图中共展示了六个连续的状态-对于时间，这可能意味着系统在相继的六个时间点上的状态；对于空间，这可能意味着系统在六个相邻间隔上的局部的状态-的熵$H(X_1)$ H(X1)到$H(X_6)$ H(X6)的彼此情况。在图中，大的浅蓝色方块（其包含了正蓝色和深蓝色方块）代表了$X_1$X1到$X_6$X6各自的熵，而正蓝色方块（包含了深蓝色方块）代表不同状态熵两两相交的互熵，而深蓝色方块则代表了三种状态熵相交而形成的互熵。从图中，我们可以看到：（a）$H(X_1)$ H(X1)到$H(X_6)$ H(X6)彼此大小是相等的；（b）一种状态只与其前、后两种状态相关，因此$H(X_1)$ H(X1)与$H(X_4)$ H(X4)无交叠，$H(X_2)$ H(X2)与$H(X_5)$ H(X5)无交叠，如此等等；（C）每增加一个状态$X_i$ Xi，并将所有状态合并起来作为系统的新的表达状态的话（比如，在图中考虑考虑$H(X_1X_2...X_6)$H(X1X2...X6)），那么系统增加的熵为$H(X_i/H_{i-1}H_{i-2})$ H(Xi/Hi−1Hi−2)。

在直觉上，我们容易理解，所谓系统的熵，对于平衡态下的热力学体系，当体系趋于足够大时，我们会用如$H(X_i/H_{i-1}H_{i-2})$ H(Xi/Hi−1Hi−2)这样的条件熵来表达体系每增加单位大小-不论是增加多一个分子或者磁矩，还是多增加一个单位体积空间-而相应增加的熵。通过这样的方式，我们足以在理论上来处理相互关联的系统的热力学熵和统计力学熵。

另外，我们也可以看出选用符号"/"的好处：图形上看，这个斜杠表示条件熵是指从自己的熵中划去属于条件的那部分熵而剩下来的熵。

（5）

在本部分的下部，我们将讨论大数定理、典型序列、Fisher信息和最大似然法，再检讨Jaynes的最大熵原理，为我们讨论热力学做准备。

当然，按照我的惯例，在结束本篇博文前，我们得八卦一下历史。

据说，在信息论完整提出前，香农（https://en.wikipedia.org/wiki/Claude_Shannon）经常拜访冯.诺依曼（https://en.wikipedia.org/wiki/John_von_Neumann），正是冯.诺依曼的建议，才有“熵”的定名：

The theory was in excellent shape, except that he needed a good name for "missing information"."Why don’t you call it entropy", von Neumann suggested. "In the first place, a mathematical development very much like yours already exists in Boltzmann's statistical mechanics, and in the second place, no one understands entropy very well, so in any discussion you will be in a position of advantage.

（Avery, John (2003). *Information Theory and Evolution*. World Scientific. ISBN 981-238-400-6.，https://en.wikipedia.org/wiki/History_of_entropy）