Teoria: Sistemas de Processamento de Informação

Nestor Caticha

22 de março de 2013

Sumário

Entropia

> I wish to propose for the reader's favourable consideration a doctrine which may, I fear, appear wildly paradoxical and subversive. The doctrine in question is this: that it is undesirable to believe in a proposition when there is no ground whatever for supposing it true.
>
> Bertrand Russell, in *Sceptical Essays*

É inquestionável a capacidade de Russell de criar frases de efeito para começar um livro. Como discordar da frase acima? "É indesejável acreditar numa proposição quando não há nenhuma base para supó-la verdadeira". Parece óbvia, mas ele usa algumas centenas de páginas para mostrar que a doutrina é efetivamente alheia à forma de pensar e agir das pessoas em geral e por isso ele está justificado em chamá-la paradoxal e subversiva. Ainda assim, parece óbvia para quem quer uma descrição científica de alguma área de conhecimento. É interessante notar as consequências que ela tem. Russell as analisa do ponto de vista das relações humanas. Shannon, Jaynes e toda a escola de teoria de informação o fazem do ponto de vista das consequências matemáticas que essa doutrina tem quando devemos raciocinar sob a luz de informação incompleta. Esse é o tema deste capítulo.

Especificamente queremos encontrar um método para atribuir números às probabilidades satisfazendo vínculos impostos por informação dada. Este problema está relacionado à atualização das probabilidades para novos números devido à inclusão de nova informação. Novamente vamos pensar em casos particulares, que sendo suficientemente simples permitam expressar desejos de como deveria ser a teoria. Há várias formas de expressar informação e isso terá consequências sobre os métodos para atualizar probabilidades. Começamos descrevendo algo aparentemente simples, quando a informação não faz diferença entre as diferentes possibilidades. A partir dessa primeira atribuição continuamos em direção a um método geral. O resultado será que a cada distribuição de probabilidades será atribuido um número que indica quanta informação a mais seria necessário coletar para ter informação completa. Dentre aquelas distribuições que satisfazem os vínculos da informação conhecida como verdadeira, escolhemos a que é mais ignorante e portanto faz menos suposições não necessárias. A determinação desta distribuição requer o uso de algumas técnicas matemáticas. Nesse ponto o estudante deverá consultar algum livro de cálculo ou o Apêndice sobre o uso dos multiplicadores de Lagrange no cálculo de extremos sujeitos a vínculos. Nas últimas seções do capítulo começaremos a estudar sistemas físicos e fazer a conexão com a Termodinâmica.

Simetria

Um experimento é descrito pela informação contida em I_1= "Suponha que temos uma moeda com duas faces, que descrevemos pela variável $\sigma = \{\pm 1\}$. O valor $\sigma = 1$ está associado à cara e $\sigma = -1$ à coroa. Jogo a moeda para cima, bate no ventilador do teto, e cai num lugar onde não podemos no momento ver o resultado."

Consideremos o seguinte jogo J_1. Se $\sigma = 1$ você ganha e eu perco. Do contrário, eu ganho. Eu aposto um feijão. Quanto você estaria disposto a apostar?[1]. A resposta tem relação, para pessoas racionais, que não dependem do feijão para sobreviver [2] , com as probabilidades $P(\sigma = 1|IJ_1)$ e $P(\sigma = -1|IJ_1)$ que você atribui com base na informação I que inclui todo o que se sabe sobre a moeda e a forma como foi jogada.

[1] Jaynes não gosta de basear os fundamentos da teoria em algo tão vulgar como apostas por dinheiro. No entanto esperamos que qualquer noção *a priori* sobre apostas tenha evoluido por seleção natural onde as apostas amiúde não são por dinheiro mas sim pela própria vida.

[2] Este problema é talvez muito mais complicado pois não sabemos o que seja uma pessoa racional, mas simplesmente consideremos alguém que quer jogar e quer ganhar.

Suponhamos que o jogo fosse totalmente diferente, J_2: Introduzimos a nova variável $\sigma' = -\sigma$, voce ganha se $\sigma' = -1$. Quanto você estaria disposto a apostar contra o meu feijão? Note que temos uma informação que nos permite associar a este jogo probabilidades

$$\begin{aligned} P(\sigma' = 1|IJ_2) &= P(\sigma = -1|IJ_1) \\ P(\sigma' = -1|IJ_2) &= P(\sigma = 1|IJ_1). \end{aligned} \tag{1}$$

Isto é, podemos usar as equações acima para transformar o que se sabe de um jogo no outro e estas são válidas para qualquer I incluindo I_2 descrita abaixo.

Suponha agora que a informação que temos I_2, que difere de I por muito pouco, nos leve a ser **indiferentes** sobre que jogo está sendo jogado. Suponha que não faça diferença dado I_2 se é $\sigma = 1$ que está associada à cara ou se é $\sigma' = 1$ que o está. Isto é o "$+1$" esta pintado numa dos lados da moeda e "-1"está pintado no outro, mas não sabemos qual. É claro que as equações 1 aindam valem com I_2 no lugar I. Mas agora há uma simetria entre os jogos J_1 e J_2, que nos leva a concluir por consistência

$$\begin{aligned} P(\sigma' = 1|I_2J_2) &= P(\sigma = 1|I_2J_1) \\ P(\sigma' = -1|I_2J_2) &= P(\sigma = -1|I_2J_1). \end{aligned} \tag{2}$$

Note a sutíl diferença entre as equações 1 (escrita com I_2 em vez de I) e 2. Dado que $P(\sigma = 1|IJ_1) + P(\sigma = -1|IJ_1) = 1$, devemos concluir que $P(\sigma = 1|IJ_1) = 1/2$ e $P(\sigma = -1|IJ_1) = 1/2$.

Porque tantas voltas para chegar ao óbvio? Por vários motivos. Em primeiro lugar notamos que este não é o único exemplo onde usaremos simetria. A história da Física mostra muitas generalizações do uso de simetria para atribuir probabilidades ou definir a dinâmica, o que não é totalmente diferente, pois dinâmica vem das interações e

as interações estão relacionadas, como veremos adiante, com probabilidades condicionais e dependência. A idéia de analisar este caso simples deve-se a que as coisas vão ficar mais difíceis e é interessante se apoiar em casos simples.

Se tivessemos um dado de n faces, σ tomaria valores de 1 a n, teriamos chegado a $P(\sigma = i|I) = 1/n$, a distribuição uniforme. Note que esta atribuição tem a ver com a simetria da nossa informação sobre o experimento do dado e não é postulada *a priori*. Generalize o problema para n estados.

Este método de atribuição de probabilidades parece ter sido usado pela primeira vez por J. Bernuolli e posteriormente por Laplace. Recebe nomes como princípio da razão insuficiente ou da indiferença.

Incerteza, ignorância e Informação

Leia novamente a citação inicial deste capítulo. Suponha que sob dada informação I_A tenhamos que escolher entre diferentes distribuições de probabilidade que são igualmente compatíveis com os vínculos impostos pela informação I_A. Qual escolher? Para começar consideremos que temos só duas candidatas, $P_1(x_i)$ e $P_2(x_i)$. Suponha que $P_2(x_i)$, além de satisfazer os vínculos impostos por I_A também satisfaz àqueles impostos por informação adicional I_B, mas esta informação não é dada. Repetimos: só sabemos I_A. Qual das duas escolhemos?

Um critério poderia ser: "escolha P_2, pois pode ser que I_B seja verdadeiro." E se não for? Se você souber alguma coisa sobre a validade da asserção I_B claro que é bem vindo a usá-la. Mas se não souber, porque supor que é verdade? A escolha de P_2 é equivalente a assumir que I_B é verdadeira, mas isso não queremos fazer. Portanto parece natural que a escolha caia sobre P_1. Esta é, das distribuições compatíveis com os vínculos, aquela que faz menos suposições não garantidas pela informação disponível. Ao fazer esta escolha fazemos a escolha da distribuição que representa a maior ignorância possível. A outra escolha assume como certo informação que pode não ser verdadeira. É melhor não saber do que achar que se sabe algo que está errado [3].

[3] He who knows best knows how little he knows. Thomas Jefferson

Se você ainda insiste em escolher a outra distribuição, mesmo sem poder oferecer a veracidade de I_B como argumento, será por motivos não racionais e esse método, por mais que você o ache interessante, estará fora do alcance de nosso análise. Talvez a escolha seja por que você gostaria que I_B fosse verdade. Isso é ideologia, capricho ou dogma e não nos interessa discutir agora.

O método que procuramos - seguindo Shannon - procura atribuir a cada distribuição de probabilidade $P(x_i)$ uma *medida da ignorância* que essa distribuição representa sobre a variável x.

Se soubermos, por exemplo, que $x = 1$ não temos nenhuma igno-

rância. Se não soubermos nada sobre x, a não ser que está é uma variável que toma valores num conjunto discreto $\{x_i\}$ de n membros, então por simetria, $P(x_i) = 1/n$. Note que não é por símetria do sistema físico, mas porque a informação que temos não distingue as diferentes possibilidades i, que são simétricas quanto às preferências. Este é o princípio da razão insuficiente de Laplace, que Boltzmann seguiu, e que Jaynes discute de forma detalhada.

Máxima Entropia: Uma vez atribuida essa medida, que é chamada de *entropia,* tomaremos a distribuição com o maior valor possível dentre aquelas que satisfazem as restrições da informação conhecida. Esta é a distribuição mais incerta ou a que faz menos hipóteses e portanto está menos arriscada a fazer hipóteses incorretas.

Suponha que temos informação, na forma de um vínculo, no seguinte problema. I_A="Uma moeda indestrutível de três lados ($s = -1, 0$ ou 1) foi jogada muitas vezes, batendo no ventilador que gira no teto, e o valor médio $\overline{s}$ desse experimento é compatível com o valor 0."

É razoável atribuir valores $P(-1) = P(1) = 0$ e $P(0) = 1$? Se eu insistir que isso é razoável, a pergunta que você fará é: "porque foram eliminadas as possibilidades ± 1?" e ainda "o quê é que ele sabe que eu não sei?". Simplesmente dizer que eu não gosto de ± 1 não é argumento razoável. A mesma coisa pode ser dito para a escolha de outra distribuição. Porque esta e não aquela?

Como proceder? Procuramos um critério que dê a cada distribuição $P(x)$, uma medida da falta de informação para determinar o valor de x. Queremos um método geral. Em princípio não sabemos se é possível tal método. Novamente usaremos a idéia que se um método geral existe, então deve ser aplicável a casos particulares. Se tivermos um número suficientemente grande de casos pode ser que o método geral se mostre incompatível com esse grande número de casos particulares. Se não fizermos a tentativa não saberemos. Olharemos a seguir o caso de variáveis discretas. O de variáveis contínuas será feito posteriormente. Obteremos um método para atribuir probabilidades, números que caracterizam asserções com base em informações que é conhecido por Máxima Entropia. Este durante muito tempo deu a impressão que estamos frente a um método independente do teorema de Bayes. Enquanto inferência usando Bayes leva de uma distribuição *a priori* para uma posterior, o método de inferência baseado em entropia, proposto por Jaynes, seguindo Boltzmann, Gibbs e Shannon, pareceria atribuir probabilidades, e não atualizar probabilidades ante a disponibilidade de nova informação. Isto foi esclarecido posteriormente (ver AC), os dois métodos levam de um *a priori* para uma posterior. O problema surge devido a que a forma original de Jaynes levava de uma distribuição *a priori* particular, a uniforme obtida por razão insuficiente, para a posterior.

Quais são os casos particulares que devemos adotar para este programa? A medida $H[P]$, a entropia, deve satisfazer, em primeiro lugar a transitividade, pois queremos escolher uma distribuição como sendo mais preferível que outra e se P_1 é preferida ante P_2 que por sua vez é preferida ante P_3, então P_1 deverá ser preferida ante P_3. Satisfazemos isto impondo que

- (S. 1) A cada distribuição de probabilidades associamos um número real $H[P]$

É tão razoável que é até difícil imaginar ante que críticas deveriamos defender este 'caso particular'. $H[P]$ é uma função de todo o conjunto de valores $\{p_i\}$. Para variáveis X que tomam valores reais, a entropia será um funcional da densidade $p(x)$. Isto é, a cada função $p(x)$ será atribuido um número.

No caso particular em que a informação é simétrica ante troca dos rótulos i, teremos $P(x_i) = 1/n$. Neste caso, por simplicidade notacional denotamos $H[1/n, 1/n, ..., 1/n] = F(n)$. Suponha que temos dois problemas. No primeiro $n = 2$ e no segundo $n = 10000$. Quanta informação falta em cada problema para determinar o valor de x? Não sabemos, mas é razoável supor que no segundo caso falta mais. Logo, a medida que buscamos deve satisfazer

- (S. 2) $F(n)$ é uma função crescente de n.

Um terceiro tipo de caso particular onde sabemos algo, ou achamos que se não fosse assim, algo estranho estaria acontencendo , é nosso velho amigo:

- (S. 3) Se há mais de uma forma de analizar uma situação, todas devem dar o mesmo resultado.

Isto ainda não é suficiente e é preciso colocar uma condição que é menos óbvia e que diz respeito às possíveis maneiras de analisar a incerteza frente a diferentes agrupamentos dos estados. Suponha que x possa ter n valores possíveis. As probabilidades p_i deverão ser atribuidas. Seja $\{m_g\}$ um conjunto de números inteiros positivos tal que $\sum_{g=1}^{N} m_g = n$ e que denotam o tamanho de agrupamentos de estados de x. Escolha o conjunto de valores $\{m_g\}$ e mantenha-o fixo durante a análise. Depois esolheremos outro conjunto e o manteremos fixo novamente. A probabilidade de que x tenha um valor dentro de um agrupamento g é $P_g = \sum_{i \in g} p_i$. Se for dado que está no grupo g, a probabilidade que esteja no estado i é $p(i|g) = \frac{p_i}{P_g}$. A incerteza associada à variável x pode ser medida em dois passos, o primeiro passo mede a incerteza de estar em um dos agrupamentos g, chamemos H_G esta entropia e o segundo, uma vez que está em g, sobre o estado i em particular esta entropia será chamada H_g.

A última imposição, que chamaremos de

- (S. 4) aditividade sob agrupamento,

diz que a entropia dos dois passos é aditiva e portanto será $H_G + \sum_{g=1}^{N} P_g H_g$ que é a entropia do primeiro passo mais a média das entropias da incerteza associada a cada agrupamento g.

Isto e a consistência do item (3) nos dá

$$H[p_i] = H_G[P_g] + \sum_{g=1}^{N} P_g H_g[p(i|g)] \tag{3}$$

Que isto é suficiente para determinar a forma funcional da entropia, será provado a seguir e será feito em dois passos, para cada um fazemos um dos dois tipos de agrupamento mencionados acima.

(Passo I) Começamos analisando o caso particular em que a informação é simétrica para todos os estados i e os agrupamentos são do mesmo tamanho $m_g = m = n/N$ para todos os N agrupamentos g. Então: $p_i = 1/n, P_g = 1/N$ e $p(i|g) = N/n = 1/m$

$$H[1/n...1/n] = H_G[1/N, ..., 1/N] + \sum_{g=1}^{N} \frac{1}{N} H_g[1/m, ...1/m] \tag{4}$$

como todos os termos lidam com entropias de distribuições uniformes, podemos introduzir a função monotônica desconhecida F:

$$F(n) = F(N) + \sum_{g=1}^{N} \frac{1}{N} F(m) \tag{5}$$

e, dado que $n = mN$, obtemos a equação funcional:

$$F(mN) = F(N) + F(m) \tag{6}$$

A solução óbvia é dada por

$$F(n) = k \log n. \tag{7}$$

Mas esta solução não é única a não ser que usemos o fato que $F(n)$ é monotônica pela imposição (2). A constante k não é muito importante neste estágio pois mede a escala das unidades, mudanças de k equivalem a mudanças na base do logarítmo que não alteram em nada a ordem de preferências de uma distribuição sobre outra. Mais para a frente veremos dentro da Mecânica Estatística, que tais mudanças equivalem a mudanças na escolha da escala de temperatura e nesse contexto poderá ser interpretada como k_B a constante de Boltzmann.

(Passo II) Ainda não obtivemos a forma de H no caso geral. Passamos a analizar o caso em que os tamanhos dos agrupamentos m_g são arbitrários, salvo o fato $\sum_{g=1}^{N} m_g = n$, mas ainda temos que $p_i = 1/n$ é uniforme. Temos então que a probabilidade $P_g = m_g/n$ é arbitrária,

mas a probabilidade condicional dentro de cada grupo é uniforme: $p(i|g) = 1/m_g$

$$H[1/n, ..., 1/n] = H_G[P_g] + \sum_{g=1}^{N} P_g H_g[1/m_g, ..., 1/m_g] \tag{8}$$

e substituimos a entropia da distribuição uniforme:

$$F(n) = H_G[P_g] + \sum_{g=1}^{N} P_g F(p(i|g)), \tag{9}$$

então a entropia da distribuição de probabilidades arbitrária P_g é dada por

$$H_G[P_g] = F(n) - \sum_{g=1}^{N} P_g F(m_g). \tag{10}$$

Introduzimos um 1 através de $1 = \sum_{g=1}^{N} P_g$, substituimos $F(n) = k \log n$

$$H_G[P_g] = (\sum_{g=1}^{N} P_g) k \log n - k \sum_{g=1}^{N} P_g \log(m_g) \tag{11}$$

$$H_G[P_g] = -k \sum_{g=1}^{N} P_g \log(m_g/n) \tag{12}$$

e finalmente

$$H_G[P_g] = -k \sum_{g=1}^{N} P_g \log(P_g) \tag{13}$$

que é a forma da entropia de Shannon, mas pode ser chamada de entropia de Y, onde Y é um subconjunto de nomes extraídos de { Shannon, Wienner, Boltzmann, Gibbs, Jaynes }, o que torna a vida dos historiadores da ciência mais interessante.

Deve ficar claro que não há nenhuma controvérsia quanto à utilidade deste formalismo. No entanto há muita controvérsia sobre a interpretação desta fórmula em geral, da necessidade dos *axiomas*, da suficiência, de se o nome de Boltzmann e Gibbs, associado à entropia termodinâmica deveria ser associado neste ponto a esta forma. Há também discussões atuais sobre o efeito acarretado pela mudança de um ou outro dos casos particulares que usamos. Sobre se estes deveriam ser chamados de axiomas [4]. Sobre como generalizar isto para o caso em que as variáveis x tomam valores em intervalos reais. Nesse caso a idéia de uniformidade fica relativa à escolha do sistema de coordenadas. A teoria fica muito mais interessante e a invariância antes transformações gerais de coordenadas leva as discussões para o contexto da geometria diferencial.

A contribuição de Boltzmann e Gibbs será brevemente discutida no resto do capítulo e não entraria aqui em um par de parágrafos. De

[4] Você pode perceber isso na exposição acima. Afinal procuramos fazer análise de casos onde não há informação completa para que deduções sejam possíveis. Os axiomas são úteis para deduções não para inferências.

Shannon e Wienner vem a idéia de discutir informação do ponto de vista de volume, quantidade com o objetivo de entender limitações para a transmissão em canais de comunicação. Como codificar uma mensagem, comprimindo-a e como, no outro lado do canal de comunicação, reobté-la. Não foi discutida a qualidade da informação ou a utilidade da informação: se fosse, toda a teoria não teria utilidade para descrever por exemplo meios como a televisão.... Isto não é totalmente piada. Se olharmos uma máquina de processamento de informação, como por exemplo um sistema nervoso de um animal, o conceito do valor da informação toma uma posição muito mais central. Há vantagens evolutivas em realizar inferência de uma forma frente a outra. Mais sobre tudo isso em aulas posteriores.

Uma contribuição importante de Jaynes foi a percepção que a maximização da entropia sujeita aos vínculos da informação leva a um método fundamental de indução, chamado por ele de MaxEnt. Outros nomes estão associados a extensão de idéias de entropia como método indutivo fundamental [5]. Em particular tem se conseguido nos últimos anos uma formulação muito mais satisfatória e menos ad hoc dos princípios por trás deste método.

[5] ver A.C

A generalização do método de máxima entropia para inferência em problemas com variáveis em variedades contínuas leva a que o método não é tanto para a atribuição de números mas de uma atualização frente à chegada de nova informação. De novo aparece uma densidade de probabilidades *a priori* $p_0(x)$ que era o melhor que podia ser feito antes de levar em conta a nova informação. A nova densidade de probabilidades $P(x)$ será determinada pelo máximo do funcional

$$H_G[P|p_0] = -k \int dP(x) \log(\frac{P(x)}{p_0(x)}) \tag{14}$$

sujeito aos vínculos impostos pela nova informação. A história desta entropia também é extensa. Uma mostra disto é que recebe vários nomes: entropia cruzada, (menos) distância (ou número) de Kullback-Leibler. Não é uma distância, pois por exemplo não é simétrica. No entanto se P e p_0 estiverem "muito perto" então sim poderemos introduzir uma distância entre densidades de probabilidade e a partir de ai um tensor métrico e uma geometria.

Exemplos

O problema variacional: variáveis discretas

Seja dado que a variável x toma um entre n valores e temos a informação que o valor médio de x é conhecido e dado por $< x >= \mu$. Qual é a distribuição de probabilidades de x que reflete a informação que temos e faz o menor número de hipóteses adicionais?

Queremos encontrar $\{P_i\}$ que maximize $H[P]$ sujeito a

$$\sum P_i = 1 \quad (15)$$
$$\sum x_i P_i = \mu \quad (16)$$

Introduzimos os vínculos através dos multiplicadores de Lagrange. Procuramos o máximo de

$$H[P] - \lambda_1\{\sum x_i P_i - \mu\} - \lambda_0\{\sum P_i - 1\} \quad (17)$$

então olhamos para o seguinte problema variacional (tomamos a liberdade de fazer $k = 1$):

$$\delta\left[-\sum P_i \log P_i - \lambda_1\{\sum x_i P_i - \mu\} - \lambda_0\{\sum P_i - 1\}\right] = 0 \quad (18)$$

O que significa o símbolo δ? Fazemos mudanças $P_i \to P_i + \delta P_i$, tratamos as variações δP_i como independentes , e finalmente escolhemos os multiplicadores de Lagrange para que os vínculos sejam satisfeitos. Temos três termos que olharemos separadamente:

$$(1) \quad \delta H[P] = H[P + \delta P] - H[P] = -\sum_i \delta P_i[\log P_i + 1] \quad (19)$$
$$(2) \quad \delta(-\lambda_1\{\sum_i x_i P_i - \mu\}) = -\lambda_1 \sum_i x_i \delta P_i \quad (20)$$
$$(3) \quad \delta(-\lambda_0\{\sum_i P_i - 1\} = -\lambda_0 \sum_i \delta P_i \quad (21)$$

Juntando os três termos:

$$-\sum_i \delta P_i\{\log P_i + 1 + \lambda_1 x_i + \lambda_0\} = 0 \quad (22)$$

Mas fazemos as variações δP_i independentes entre si, de tal forma que se a soma na equação acima for zero, o será termo a termo:

$$\log P_i + 1 + \lambda_1 x_i + \lambda_0 = 0 \quad (23)$$

e segue que

$$P_i = e^{-1-\lambda_0-\lambda_1 x_i} \quad (24)$$

Os vínculos determinam os multiplicadores de Lagrange. Definimos Z, que será uma grandeza central em todo desenvolvimento futuro, $Z = e^{1+\lambda_0}$, que fica determinado pelo vínculo da normalização:

$$Z = \sum_i e^{-\lambda_1 x_i} \quad (25)$$

A seguir notamos uma utilidade de Z, chamada de função de partição:

$$\frac{d \log Z}{d\lambda_1} = -\frac{1}{Z}\sum_i x_i e^{-\lambda_1 x_i} \quad (26)$$

de onde segue que o valor esperado de X é dado por

$$< x >= -\frac{d \log Z}{d\lambda_1} \tag{27}$$

Esta é uma equação para λ_1, dado que o resto é conhecido: os valores x_i que a variável X pode tomar, o o seu valor médio, que é o vínculo imposto nas distribuições $P(x)$.

Exemplo: Variável de dois estados

Uma variável binária de Bernoulli tem dois valores possíveis. A informação disponível não permite distinguir entre os dois estados. Qual é a distribuição de probabilidade $P(x)$? Já sabemos a resposta pelo princípio de razão insuficiente de Laplace. E pelo método de máxima entropia? Fazemos a variação da entropia, sujeita ao único vínculo de normalização:

$$\delta \left[-\sum P_i \log P_i - \lambda_0 \{ \sum P_i - 1 \} \right] = 0 \tag{28}$$

o que leva a $\log P_i = 1 + \lambda_0$ note que o resultado é que a probabilidade é independente do estado i, como esperado. Escolhendo o multiplicador de Lagrange, temos a resposta $P_1 = P_2 = 1/2$. (Você precisou fazer a conta?).

Se o vínculo $P_1 + P_2 = 1$ for explicitamente introduzido na expressão da entropia, teremos , fazendo $P_1 = p$ e $P_2 = 1 - p$

$$H(p) = -p \log p - (1-p) \log(1-p) \tag{29}$$

Desenhe a função e verifique que o máximo está em $p = 1/2$. Note que $\lim_{p \to 0^+} p \log p = 0$ e $\lim_{p \to 1^-} (1-p) \log(1-p) = 0$.

Não há muito o que fazer se for dada informação do tipo "o valor de $< x >= m$ ", pois não há mais que uma distribuição que satisfaz o vínculo.

3 estados

Suponha que seja dado "X toma valores 0, 1 ou 2 e o valor médio de x, é conhecido: $< x >= m$". Determine a distribuição de máxima entropia compatível com o vínculo. Faça o problema. Resultado

$$P_i = \frac{e^{-\lambda_1 x_i}}{Z(\lambda_1)} \tag{30}$$

com λ_1 determinado pela eq. 27 e Z pela eq. 25 . Portanto, chamando $u = exp(-\lambda_1)$,

$$Z = 1 + u + u^2 \tag{31}$$

e

$$P_i = \frac{u^i}{1+u+u^2} \tag{32}$$

$$m = \frac{u+2u^2}{1+u+u^2}$$

$$u^2(m-2) + u(m-1) + m = 0 \tag{33}$$

$$u = \frac{1-m \pm \sqrt{6m-3m^2+1}}{2(m-2)} \tag{34}$$

A equação 33 tem duas raizes. Discuta a escolha da solução correta. A figura 1 mostra o valor máximo da entropia H como função de m, e do lado esquerdo as probabilidades. Note que para o gráfico de $H[P]$ deveria ser desenhado num espaço tridimensional (bidimensional se o vínculo da normalização for incluido). Mas desenhamos o valor do máximo de $H[P]$ que é função de um único parâmetro.

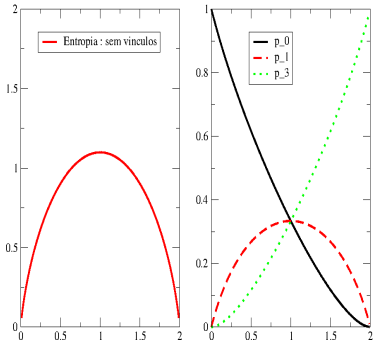

Figura 1: Esquerda: Máximo da entropia como função da média m. Direita: as probabilidades p_0, p_1 e p_2 que maximizam a entropia de Shannon como função da informação $< x >= m$.

Variáveis reais: a gaussiana

Queremos encontrar $P(x)$ que maximize $H[P]$ sujeito à informação I

$$\begin{aligned} \int P(x)dx &= 1 \\ \int xP(x)dx &= \mu \\ \int x^2P(x)dx &= \sigma^2+\mu^2 \end{aligned} \tag{35}$$

Isto é, qual é a densidade de probabilidade que devemos atribuir a uma variável quando sabemos que a sua média e variância tem valor finito dado? Novamente introduzimos os vínculos através dos multiplicadores de Lagrange. Procuramos o máximo de "$H[p]+$ vínculos", através do seguinte problema variacional:

$$\delta\left[-\int dxP(x)\log(\tfrac{P(x)}{p_0(x)} + \lambda_0\{\int P(x)dx - 1\} + \right. \tag{36}$$

$$\left.\lambda_1\{\int xP(x)dx - \mu\} + \lambda_2\{\int x^2P(x)dx - \sigma^2 - \mu^2\}\right] = 0 \tag{37}$$

O diferença importante com relação ao caso de variáveis discretas é que aparece a distribuição p_0. Supomos que neste caso o problema físico é tal que quando descrito usando a variável x, antes de levar em consideração a informação I [6] não temos nenhuma preferência, ou seja tomamos p_0 como constante, que pode ser esquecida. Novamente fazemos mudanças $P \to P + \delta P$, tratando as variações $\delta P(x)$ para valores de x diferentes, como independentes ,

$$\{1 + \log P(x) + \lambda_0 + \lambda_1 x + \lambda_2 x^2\}\,\delta P(x) = 0, \tag{38}$$

[6] Ou seja voltamos à entropia de Shannon. O que significa uma distribuição uniforme no eixo real inteiro? Veja Jeffreys para priors não normalizáveis.

lembrando que as variações são independentes, o termo entre chaves dever ser igual a zero, o que leva novamente a uma forma exponencial

$$P(x) = e^{1-\lambda_0-\lambda_1 x-\lambda_2 x^2} \tag{39}$$

Agora escolhemos os multiplicadores de Lagrange para que os vínculos sejam satisfeitos. Use $\int_{-\infty}^{\infty} exp(-y^2/2)dy/\sqrt{2\pi} = 1$, as propriedades das integrais ante mudanças de variáveis e mostre que

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{40}$$

O resultado é distribuição gaussiana. Vemos um dos motivos da utilidade de distribuições gaussianas em tantos lugares em análise de dados: se só soubermos os valores finitos da média e da variância, então o uso de uma gaussiana é forçada por máxima entropia, qualquer outra distribuição significaria o uso implícito de informação não conhecida.

Exemplo: Bayes e Máxima Entropia

Foram feitas N medidas da variável Y para diferentes valores da variável de contrôle X, o que constitui o conjunto de dados $D = \{(x_i, y_i\}_{i=1,...N}$. Supomos uma lei de Y em função de X, mas desconhecemos o parâmetro θ:

$$y = f_\theta(x)$$

e sabemos que há erros na medida dos valores de Y,que supomos independentes uns dos outros. Supomos ainda que as medidas de X não têm erros:

$$y_i = f_\theta(x_i) + \eta_i \tag{41}$$

Sabemos, pelo estudo do aparelho de medida, que a variância de η é finita e que não há desvios sistemáticos: $< \eta >= 0$ e $< \eta^2 >= \sigma^2 < \infty$. Supomos que temos um conhecimento prévio de θ codificado na distribuição a priori: $p_0(\theta)$, logo, pelo teorema de Bayes determinamos

$$P(\theta|(x_i, y_i\}, I) = \frac{p_0(\theta)P(\{y_i\}|\theta, \{x_i\}, I)}{P(D)}, \tag{42}$$

Qual é a expressão para a verossimilhança $P(\{y_i\}|\theta, \{x_i\}, I)$?

Note por indepêndencia que deve ser o produto $\prod_i P(y_i|\theta, x_i, I)$, logo só devemos nós preocupar com um termo.

Se não houvesse ruído η :

$$P(y_i|\theta, x_i, I) = \delta(y_i - f_\theta(x_i))$$

Devido à equação 41 podemos ver que

$$P(y_i|\theta, x_i, I) = P(\eta)$$

Dada a informação que temos, qual é $P(\eta)$?

Este é um caso para o método de máxima entropia. A distribuição que surge é a gaussiana e se o tivessemos dito desde o começo, poucos iriam reclamar.

Conclua o problema supondo que na região de relevância de θ a distribuição a prior $P(\theta)$ é uniforme. Obtenha então a posterior. Se olharmos para o valor θ_M que máximiza a posterior , neste caso a verossimilhança, teremos uma estimativa de *Máxima Verossimilhança*. Neste caso é o resultado de Mínimos Quadrados.

Entropia e convexidade

A divergência de Kullback-Leibler ou entropia cruzada entre duas distribuições de probabilidade que são não nulas no conjunto de n estados $\{i\}$ é definida por

$$K[p|q] = \sum_i p_i \log \frac{p_i}{q_i}. \tag{43}$$

Se considerarmos $q_i = 1/n$ a distribuição uniforme, termos

$$K[p|q] = \sum_i p_i \log(np_i), \tag{44}$$

$$K[p|q] = \log n - H[p_i], \tag{45}$$

Um número notável de resultados pode ser obtido a partir de considerações de convexidade do logaritmo, que leva por exemplo a

$$\log x \leq x - 1, \tag{46}$$

e que pode ser facilmente visto do gráfico das duas funções. Mas ver o gráfico não é suficiente, prove este resultado. A igualdade só ocorre se $x = 1$. A equação 46 usada na forma de $H[p]$ leva a

$$\begin{aligned} H[p] &= -\sum_i p_i \log(p_i), \\ &\geq -\sum_i p_i(p_i - 1) \\ &\geq 1 - \sum_i p_i^2 \geq 0, \end{aligned} \tag{47}$$

que a entropia é não negativa. Só será nula se $\sum_i p_i^2 = 1$, o que só pode ocorrer se $p_i = 0$ para todo i exceto para um estado, e.g. i', para o qual $p_{i'} = 1$, ou seja temos informação completa que o valor de X é $x_{i'}$.

Se usarmos a mesma cota na expressão de $K[p|q]$ obteremos

$$\begin{aligned} K[p|q] &= -\sum_i p_i \log \frac{1}{np_i}, \\ &\geq -\sum_i p_i (\frac{1}{np_i} - 1) \\ &\geq \sum_i \frac{1}{n} - \sum_i p_i = 0, \end{aligned} \tag{48}$$

com a igualdade ocorrendo somente se $p_i = q_i$. Juntando as duas cotas com a relação entre K e H, equação 45, temos que

$$0 \leq H[p] \leq \log n, \tag{49}$$

com a igualdade à esquerda ocorrendo quando temos informação completa sobre X e à direita quando a informação é simétrica com respeito aos n estados.

Independência

Anteriormente vimos que a regra para análise do produto lógico ou conjunção em função das probabilidades de asserções mais simples é

$$P(X, Y|I) = P(X|YI)P(Y|I). \tag{50}$$

Suponha o caso em que a informação sobre Y não diz nada sobre X, isto é, sob a informação I, saber algo sobre Y não muda nada sobre as probabilidades que atribuimos a X , isto é X é independente de Y, ou seja $P(X|YI) = P(X|I)$. É trivial mostrar que então $P(Y|XI) = P(Y|I)$ e que $P(X, Y|I) = P(X|I)P(Y|I)$ a probabilidade conjunta se fatoriza. É interessante considerar $K[P(X, Y|I)|P(X|I)P(Y|I)$ que está relacionada à diferença de informação que temos quando temos a probabilidade conjunta e as marginais, $P(X|I) = \sum_Y P(X, Y|I)$ e $P(Y|I) = \sum_X P(X, Y|I)$.

$$\begin{aligned} K[P(X, Y|I)|P(X|I)P(Y|I)] &= \sum_i P(X, Y|I) \log \frac{P(X, Y|I)}{P(X|I)P(Y|I)} \\ = -H[P(X, Y|I)] &+ \sum_i P(X, Y|I) \log P(X|I) + \sum_i P(X, Y|I) \log P(Y|I) \\ = -H[P(X, Y|I)] &+ H[P(X|I)] + H[P(Y|I)] \end{aligned} \tag{51}$$

e como $K \geq 0$ obtemos que

$$H[P(X, Y|I)] \leq H[P(X|I)] + H[P(Y|I)]. \tag{52}$$

A equação 51 mostra que $K[P(X, Y|I)|P(X|I)P(Y|I)]$ mede a diferença entre a informação que falta se só olharmos para X e Y separadamente ou conjuntamente. A equação 52, que mostra a propriedade chamada

de subaditividade, mostra que a informação que falta quando as variáveis são analisadas conjuntamente é menor ou igual que quando o são em separado. A não ser que as variáveis sejam independentes a análise da distribuição conjunta incorpora mais informação ou seja, falta menos informação.

Exercício sobre independência: Variáveis Gaussianas correlacionadas

Considere a densidade conjunta de probabilidades de X e Y que tomam valores no eixo real x e y:

$$P(x,y|\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)}. \tag{53}$$

O parâmetro ρ está entre 0 e 1 e no que resta do exercício todas as probabilidads serão condicionadas a conhecimento de ρ que não será escrito explicitamente.

Calcule as distribuições marginais $P(x)$ e $P(y)$, e.g. integrando $P(x) = \int dy P(x,y)$. Mostre que são gaussianas com média nula e variância 1.

Encontre a distribuição condicional $P(x|y)$ (use o teorema de Bayes).

Calcule a entropia das marginais (gaussiana univariadas (sic)) Calcule a entropia da distribuição conjunta (gaussiana multivariada)

Discuta o significado de ρ como (A) correlação entre X e Y, (comece calculando $< xy >$) e (B) como medida da diferença de informação que temos nos seguintes dois casos: (a) dado $P(x,y)$ e (b) dados $P(x)$ e $P(y)$.

Use a sua linguagem preferida e trace as curvas de nível da distribuição conjunta para diferentes valores de ρ para mais uma forma de entender o papel de ρ.

Aqui acaba este exercício. O que segue é mais difícil, pode ser pulado, ou deixado até que seja dicutido quando estudarmos o problema de aprendizagem de máquinas usando idéias de mecânica estatística. Suponha que é dado o valor do sinal de y, i.e $\sigma_y = sinal(y)$. Calcule $P(\sigma_x|\sigma_y)$.

Suponha que queremos usar σ_x como estimador de σ_y. Calcule como função de ρ a probabilidade e_σ de fazer um *erro* , i.e. de errar o sinal de Y quando o sinal de X é usado no seu lugar ou que os sinais de X e Y sejam diferentes. A resposta $e_\sigma = \frac{1}{\pi}\arccos\rho$. Fica fácil de entender se fizermos um desenho. Trace duas linhas ξ_x e ξ_y que fazem um ângulo θ. Trace dois vetores perpendiculares 'as linhas $\vec{J_x}$ e $\vec{J_y}$. Também fazem um ângulo θ. Escolha uma direção qualquer, desenhe um vetor $\vec{S}$ do ponto de cruzamento nessa direção. A pergunta é qual é a probabilidade que a projeção desse vetor nas duas perpendiculares tenha o mesmo sinal. A resposta é $2\theta/2\pi$. Generalize o desenho par N dimensões. A resposta (deve ser provado) continua

sendo θ/π. Note que θ está relacionado ao produto escalar entre os dois vetores perpendiculares aos (hiper-)planos $N-1$ dimensionais. Portanto $\theta = \arccos \vec{J_x}.\vec{J_y}$. É uma consequência do teorema do limite central que as projeções $x = \vec{J_x}.\vec{S}$ e $y = \vec{J_y}.\vec{S}$ são variáveis cuja distribuição conjunta é dada por 53.

O discutido neste exemplo é extremamente útil em problemas de aprendizagem de máquina, onde queremos aprender uma direção desconhecida $\vec{J_y}$ só com informação sobre σ_y. O que se faz é construir uma aproximação $\vec{J_x}$ a partir de σ_y para vários exemplos $\vec{S}$. Isto é, a máquina aprende e a medida da qualidade da aprendizagem, $P(\sigma_x|\sigma_y)$ é chamado de erro de generalização, que cai quanto maior for a correlação ρ. Sobre isto, muito mais será dito em capítulos posteriores [7].

[7] Mas não neste curso introdutório.

Caso Geral para informação sobre valores esperados

Em geral podemos incluir vários vínculos

$$< g_k(x) >= G_k, \tag{54}$$

com $k = 1...K$ e chamando $1 = g_0$, o problema variacional será

$$\delta \left[-\sum P_i \log P_i - \sum_{k=0}^{K} \lambda_k \{ \sum_i g_k(x_i) P_i - \mu \} \right] = 0 \tag{55}$$

que leva a

$$-\sum_i \delta P_i \{ \log P_i + 1 + \sum_{k=0}^{K} \lambda_k g_k(x_i) \} = 0 \tag{56}$$

Como anteriormente, as variações são independentes e cada termo entre chaves é nulo, o que leva a

$$P_i = \frac{e^{-\sum_{k=1}^{K} \lambda_k g_k(x_i)}}{Z} \tag{57}$$

onde a função de partição $Z = \exp(-1-\lambda_0)$ toma a forma

$$Z = \sum_i e^{-\sum_{k=1}^{K} \lambda_k g_k(x_i)} \tag{58}$$

para satisfazer a normalização. Os multiplicadores de Lagrange remanescentes poderão ser calculados usando

$$< g_k >= G_k = -\left(\frac{\partial \log Z}{\partial \lambda_k} \right)_{\lambda_i, i \neq k} \tag{59}$$

exercício Mostre o resultado acima.

As equações acima formam um conjunto de equações não lineares que determinam os valores de $\{\lambda_k\}$ se os G_k forem conhecidos. No

entanto nos dá ainda mais. Suponha que os valores de G_k não sejam conhecidos, mas que só tenhamos o conhecimento que se os soubessemos teriamos determinado uma distribuição adequada para algum fim. Isso significa que se, em vez de G_k tivessemos acesso aos valores dos $\{\lambda_k\}$, então poderiamos calcular os G_k, e outros valores esperados de interesse teórico e experimental. Isto será possível quando tivermos uma interpretação quanto ao significado do λs. Isto será possível quando for feita a conexão com a termodinâmica. O valor máximo da entropia de Shannon será identificada com a entropia termodinâmica de Clausius e o multiplicador de Lagrange associado à energia do sistema terá as propriedades do inverso da temperatura. O fato que não sabemos o valor esperado da energia em nada nos atrapalha, mas indica que a temperatura é uma quantidade útil e seu conhecimento, adquirido através de mais uma medida serve para determinar, junto com outros parâmetros (e.g. volume, número de partículas) o estado de equilíbrio que pode de forma robusta ser estudado em diferentes laboratórios. Para poder fazer esta ligação precisamos ver mais algumas propriedades da distribuição de máxima entropia.

Conexão com a Termodinâmica: estrutura matemática

Transformações de Legendre

Denote o máximo da entropia de Shannon por S. Este valor foi encontrado a partir da informação dada, portanto é uma função dela, ou seja do conjunto $\{G_k\}$. Substituindo o valor da probabilidade que maximiza H mais vínculos, $p_i = \frac{\exp(-\sum_{k=1}^{K} \lambda_k g_k)}{Z}$ na forma $H = -\sum p_i \log p_i$ obtemos

$$S(G_1, G_2, ...G_K) = \log Z(\lambda_1, \lambda_2, ..., \lambda_3) + \sum_{k=1}^{K} \lambda_k G_k \tag{60}$$

Este resultado é importante porque mostra a relação análoga à relação termodinâmica $S = -\frac{F}{T} + \frac{E}{T}$, ou mais comumente escrita com $F = E - TS$. Na forma geral, mostra como podemos passar de uma descrição em termos do conjunto de variáveis $\{G_k\}$ para o conjunto $\{\lambda_k\}$.

Considerando uma pequena variação dos valores G_k , a partir do resultado acima podemos mostrar que:

$$\frac{\partial S(G_1, G_2, ...G_K)}{\partial G_i} = \sum_{j=1}^{K} \frac{\partial \log Z}{\partial \lambda_j} \frac{\partial \lambda_j}{\partial G_i} + \sum_{j=1}^{K} \frac{\partial \lambda_k}{\partial G_i} G_i + \lambda_i \tag{61}$$

Usando a equação 59, obtemos finalmente que

$$\lambda_i = \frac{\partial S(G_1, G_2, ...G_K)}{\partial G_i} \tag{62}$$

Um caso particular desta relação é $\frac{1}{T} = (\frac{\partial S}{\partial E})_{V,N}$ que veremos na seção **??**.

Os resultados 60 e 62 mostram como podemos descrever o sistema em termos dos G_ks ou dos λ_ks, ou seja $S(\{G_k\}$ e $\log Z(\{\lambda_k\})$ são obtidos um do outro através de uma transformada de Legendre.

exercício Suponha $K = 1$ e $< g_1 >= E$, a energia de um sistema com graus de liberdade X e estados $\{x_i\}$ e Hamiltoniano g_1. (A) Mostre que $F = E - TS$, onde $F = -T \log Z$ e $T = 1/\lambda_1$. (B) Mostre que $\frac{1}{T} = (\frac{\partial S}{\partial E})_{V,N}$.

Este exercício mostra a utilidade de tudo isto para a termodinâmica. Mas ainda há muita física para explicar. Neste ponto deve ser entendido como um exercício formal.

Relações de Maxwell

Podemos continuar estudando as derivadas parciais superiores e obter a estrutura das relações de Maxwell. Olhemos para as segundas derivadas, a partir das derivadas da equação 62:

$$\frac{\partial \lambda_i}{\partial G_j} = \frac{\partial^2 S(G_1, G_2, ...G_K)}{\partial G_i \partial G_j} = \frac{\partial^2 S(G_1, G_2, ...G_K)}{\partial G_j \partial G_i} = \frac{\partial \lambda_j}{\partial G_i} \tag{63}$$

e da equação 59

$$\frac{\partial G_k}{\partial \lambda_i} = -\frac{\partial^2 \log Z}{\partial \lambda_k \partial \lambda_i} = \frac{\partial G_i}{\partial \lambda_k} \tag{64}$$

Aqui é interessante fazer uma pausa e relembrar a enorme variedade de resultados de interesse experimental que seguem das relações de Maxwell (e.g. Callen). Do ponto de vista matemático não passam da hipótese que as derivadas mistas são contínuas, essencialmente uma trivialidade. Do ponto de vista físico, justificam a teoria e a escolha dos graus de liberdade $\{x_i\}$ e a energia para descrever o equilíbrio termodinâmico do sistema.

Suponha que para um dado sistema as predições falhem no campo experimental. O que fazer? Podemos jogar fora o método de máxima entropia ou podemos pensar que talvez nos tenhamos enganado ao escolher os graus de liberdade para descrever o problema, escolher de outra forma e começar a calcular novamente para comparar com os resultados experimentais. Historicamente há precedentes importantes que nos levam a pensar na segunda possibilidade com grande atenção. Afinal foi a falta de concordância entre a termodinâmica obtida da Mecânica Estatística de Gibbs e Boltzmann, com seus métodos similares ao exposto acima, calculada a partir da Mecânica Clássica e o Eletromagnetismo Clássico [8] e a experiência (e.g. corpo negro, calor específico de sólidos a baixas temperaturas), que levou às primeiras idéias da Mecânica Quântica. O que foi abandonado não foi a forma

[8] Na época não era necessário o adjetivo clássico.

consistente de calcular, introduzida pela Mecânica Estatística, mas a forma errada de descrever os graus de liberdade da física clássica. A lição disso é que se uma aplicação dos métodos entrópicos levarem a problemas, não devemos jogar fora o método, mas desconfiar da maneira como foi aplicado.

Suponha que jogamos com um dado comum que Gabriel ganhou de brinde como lembrancinha de uma festa de aniversário. Suponha que eu acredite que o dado tem sete faces e que a informação que tenho sobre elas seja simétrica. Espero que a probabilidade de cada uma seja $1/7$. Você acredita que eu terei sucesso no jogo? Por exemplo, apostamos no seguinte jogo. Se sair, na próxima jogada do dado o número sete eu ganho, se sair o seis, perco. Se o resultado for outro número, o dado é jogado novamente. Não há pior erro que a enumeração errada dos estados acessíveis ao sistema. Dados de um jogo realizado mostram que eu perdi. Talvez seja hora de mudar minha maneira de descrever os estados possíveis do dado. O Eletromagnetismo Clássico fez uma aposta semelhante e perdeu o jogo ao descrever o espectro de cavidade do corpo negro. As mudanças necessárias, feitas inicialmente por Planck começaram a mostrar a maneira correta, que finalmente levou à Mecânica Quântica.

Conexão com a Termodinâmica: Micro e Macroestados

O uso de uma palavra num novo contexto requer que o proponente justifique seu novo uso. Aparentemente von Neumann teria dito a Shannon que poderia usar a palavra entropia para denotar a sua medida de informação por dois motivos. Primeiro porque a forma $-\sum p_i \log p_i$ aparece em trabalhos de Gibbs e também de Boltzmann e é efetivamente uma entropia. Segundo, porque já que ninguém sabe exatamente o que é entropia, não seria possível rebater o seu uso do nome.

O objetivo principal da Mecânica Estatística é obter as propriedades termodinâmicas de sistemas físicos. Aparentemente a primeira e mais importante informação que devemos ter sobre o sistema se refere às propriedades microscópicas e devemos então indagar sobre quais são os constituintes microscópicos e como interagem, isto é, caracterizar os graus de liberdade. Na linguagem da Mecânica Estatística, isto é definir quais são os *microestados* do sistema. Por exemplo, "este sistema esta feito de tal tipo de molécula", "estas interagem entre si de tal forma". Por que parar em moléculas? Elas são feitas de átomos. Devemos descrever as posições dos átomos. Deve ser necessário entender e caracterizar os estados eletrônicos. E os núcleos? Claro, são feitos de néutrons, prótons. O estado do núcleo é certamente informação microscópica e deve ser incorporada. Os quarks que formam os núcleons, são eles microscópicos? Certamente e então devemos

nos perguntar se a estrutura do quark também deverá ser levada em conta. Ai há dúvidas. Em alguns modelos são fundamentais, em outros são compostos. Como decidir? A coisa está ficando exagerada. Esta discussão levaria a pensar que iremos atribuir probabilidades a microestados especificados pelos valores de

$$\{qp\}_{mol}\{qp\}_{atom}\{qp\}_{elet}\{qp\}_{nucl}\{qp\}_{prot}\{qp\}_{neutr}\{qp\}_{quark}\{qp\}_{gluon}\{qp\}? \tag{65}$$

As escalas de energias em que os efeitos de cada tipo de estrutura se faz sentir são muito diferentes. Se houver necessidade de uma especificação tão detalhada dos graus de liberdade, não seria possível obter a termodinâmica. Do ponto de vista histórico vemos que não seria possível obter nenhum resultado no século 19, pois então o conhecimento da estrutura da matéria estava longe do que é hoje. Também não devemos esperar nenhum resultado hoje, pois certamente o conhecimento que temos é incompleto e deverá mudar. A especificação dos graus de liberdade relevantes é o principal problema em Mecânica Estatística, pois ela determinará as perguntas que queremos responder e "nada restringe tanto uma resposta como a própria pergunta"[9].

[9] A.C.

Devemos parar e começar de novo. Qual é a pergunta que queremos responder? A resposta inclui questões como determinar quais são as propriedades termodinâmicas numa certa escala de temperatura. Imaginemos duas caixas de paredes rígidas, impermeáveis e isolantes que não permitem trocas de energia, também chamadas adiabáticas, com volumes, número de moles e energia iguais em dois laboratórios diferentes. Queremos caracterizar o estado dos sistemas de forma que *certas* perguntas tenham a mesma resposta. A velocidade da partícula 17 será a mesma nas duas caixas? Certamente essa não é uma das perguntas que queremos que tenham a mesma resposta. Queremos saber por exemplo, o calor específico ou quanta energia é necessária para elevar a temperatura de 14,5C até 15,5C de um grama de fluido. Não parece razoável que seja necessário, nesse regime de temperaturas, que os detalhes dos estados dos quarks tenham alguma influência. Devemos incluí-los, mesmo que não tenham importância? Se a resposta fosse sim como poderiamos justificar a inclusão do descrição dos graus de liberdade que ainda não sabemos como descrever, ou será que já sabemos tudo sobre a estrutura microscópica da matéria? ou como foi que no século 19, sem saber nada sobre a estrutura atómica foi possível a obtenção dos primeiros resultados?

Suponha que jogamos N moedas e queremos saber por exemplo o número de caras para cima. A resposta que você dará é baseada na distribuição binomial $P(m|N) = N!/(m!(N-m)!p^m q^{N-m})$. Suponha que agora é dada mais informação. Cada moeda tem um rosto, e para cada moeda é dado em que direção aponta o nariz. Temos dois graus

de liberdade para cada moeda, um o valor de $\sigma = \pm 1$, que indica se é cara ou coroa. Outro, a direção do nariz, está entre θ e $\theta + d\theta$.

A probabilidade da ocorrência de um dado resultado, para moedas independentes é

$$P(\sigma_1 = +1, ...\sigma_N = -1, \theta_1, ..., \theta_N) \prod_i d\theta_i = p...qP(\theta_1)...P(\theta_N) \prod_i d\theta_i \tag{66}$$

Queremos a probabilidade que o número de caras seja m. Seja n o de coroas. Então $m + n = N$ e $m - n = \sum_i \sigma_i$ o que leva a $m = (N + \sum_i \sigma_i)/2$. De acordo com o capítulo 2 esperamos que

$$\begin{aligned} P(m|N\{p, q, P(\theta)\}) &= \sum_{\sigma_1,....\sigma_N} \int \prod_i d\theta_i P(\sigma_1, ...\sigma_N, \theta_1, ..., \theta_N) \delta_{m,(N+\sum_i \sigma_i)/2} \\ &= \frac{N!}{m!(N-m)!} p^m q^{N-m} \end{aligned} \tag{67}$$

As moedas são claramente distinguíveis entre si. Mas $P(m|N\{p, q, P(\theta)\})$ tem fatoriais para eliminar a contagem de combinações mais de uma vez. Isto não indica que as moedas são idênticas? A pergunta que queremos responder, sobre estados de $\sum_{i=1,N} \sigma_i$ é independente do valor de θ e a distinguibilidade (?) das moedas é irrelevante. [10]. Qual é a entropia que devemos atribuir a esse sistema? A resposta, $-\sum p_i \log p_i$ depende explicitamente do que chamamos de estado i. Se as perguntas que quisermos responder não tratam de θ, então duas moedas com mesmo valor de σ são indistinguíveis.

[10] As moedas são distintas mas são tratadas de forma que mostra que não faz diferença se a moeda que tem $\sigma = 1$ tem um nariz que aponta numa ou outra direção Voltaremos a falar destas moedas ao discutir gases quânticos

A lição do parágrafo anterior está em que a resposta depende da pergunta. O tipo de pergunta feita determina quais são os graus de liberdade que devem ser usados para descrever o sistema. O leitor terá percebido que se perguntarmos sobre m, nem o valor de θ nem o estado eletrônico do material que compõe a moeda, que tinhamos esquecido de considerar, interessa. É comum neste ponto introduzir a noção de microestado, embora fique claro que sua definição é um tanto quanto vaga. Um microestado das moedas poderia ser descrito pelos valoresde $\{(\theta_i, \sigma_i)\}_{i=1,N}$ se as perguntas que queremos responder tem algoa ver com a orientação dos marizes. Pode ser que haja um campo magnético e os narizes tendam a apontar em alguma direção privilegiada. Mas o microestado que nós interessa quando queremos responder sobre o número total de caras é determinado simplesmente por $\{\sigma_i\}_{i=1,N}$. A probabilidade do microestado de interesse $P(\{\sigma_i)\}_{i=1,N})$ é obtida por marginalização dos θ_i

$$P(\{\sigma_i\}_{i=1,N}|I) = \int \prod_i d\theta_i P(\{(\theta_i, \sigma_i)\}_{i=1,N}|I) \tag{68}$$

e também por marginalização de outras variáveis irrelevantes para a pergunta presente. Este um tipo de agrupamento: todos os valores de

θ são agrupados e a probabilidade total do estado é obtido pela soma dos elementos do grupo.

Toda vez que falamos num microestado tomamos como óbvio que as variáveis em escalas menores foram agrupadas. É difícil exagerar a importância desta idéia, se não fosse assim não haveria como falar de nenhum sistema, pois não seriamos capazes de acabar com a sequência de escalas de forma satisfatória.

O estado descrito por $P(m|N)$ chamamos de macroestado, mas este também depende da pergunta que fazemos. Pode ser que hajam agrupamentos intermediários que podem ser chamados de mesoestados.

O que vai definir o que chamamos de macroestado é a pergunta que queremos responder. Perguntas em física são experiências. O macroestado é escolhido pelo arranjo experimental que inclui fixar alguns parâmetros e a medida de outros. No mesmo sistema podemos mudar o que é fixo e o que se mede. A informação que usaremos para determinar as probabilidades é o próprio valor das quantidades que são fixas pelo experimental durante a experiência.

Conexão com a Termodinâmica: Física

O Ensemble Microcanônico

A importância do conceito de energia é devida à sua conservação. O célebre teorema de Noether mostra a relação entre simetrias e leis de conservação. Se sob certas condições o sistema tem uma simetria contínua, concluimos que numa experiência realizada nessas condições haverá uma quantidade conservada. Suponha um sistema isolado de tal maneira que as propriedades são invariantes antes translações temporais, a energia será conservada. A preparação do arranjo experimental é feita com o objetivo de responder certas perguntas. Isso define o macroestado. Pode ser que não saibamos como fixá-lo, mas essencialmente esse é o objetivo de nosso estudo, encontrar quais as condições exprimentais, ou quais os vínculos impostos ao sistema , para que o o macroestado seja um macroestado específico. O problema consiste agora em descrever de forma correta os microestados. Mas o que é um microestado? Suponhamos que o experimento será feito numa certa região de energias e que nessa faixa os graus de liberdade relevantes sejam descritos por $\{q_i, p_i\}$ e que a dinâmica, seja em termos dessas variáveis relevantes, descrita por um Hamiltoniano $\mathcal{H}(q_i, p_i)$. Suponhamos então para ser específicos que colocamos o sistema em condições que sua energia é E fixa. Fixar E num valor determinado é muito difícil experimentalmente. Talvez seja melhor dizer que sua energia é fixa entre E e $E + \delta E$. Para alguns sistemas, fluídos simples, é também necessário fixar o número de partículas N e o volume do

sistema V. Devemos caracterizar os microestados, por exemplo determinar se número. Mas o que é um microestado? Suponhamos que o experimento será feito numa certa região de energias e que nessa faixa os graus de liberdade relevantes sejam descritos por $\{q_i, p_i\}$, o espaço de fases, e que a dinâmica, seja em termos dessas variáveis relevantes, descritos por um Hamiltoniano $\mathcal{H}(q_i, p_i)$. Podemos associar à esses microestados densidades de probabilidades $P(q_i, p_i)$ usando a informação disponível e o método de máxima entropia

$$H[P] = -k \int_{E \leq \mathcal{H}(q_i,p_i) \leq E+\delta E,\, V} \prod_i dq_i dp_i P(q_i, p_i) \log P(q_i, p_i) \tag{69}$$

sujeito unicamente ao vínculo de normalização. Tomaremos $k = 1$. O resultado de maximizar a entropia de Shannon quando a informação sobre os microestados é indiferente é simplesmente que a distribuição é uniforme sobre os microestados permitidos:

$$P(q_i, p_i) = c \tag{70}$$

como esperado pelo princípio da razão insuficiente.

O fluido simples, introduzido por Gibbs, é um sistema onde o estado termodinâmico é determinado por algumas poucas variáveis. Suponhamos um gás isolado numa caixa de paredes adiabáticas. O experimental pode controlar separdamente a energia E, o volume V e o número de constituintes (e.g moléculas) N. A constante c na equação 70 é determinada pela imposição do vínculo de normalização,

$$P(q_i, p_i) = \frac{1}{\Omega(E, N, V)} \tag{71}$$

onde $\Omega(E, N, V)$ é o volume no espaço de fases e para variáveis contínuas é

$$\Omega(E, N, V) = \int_{E \leq \mathcal{H}(q_i,p_i) \leq E+\delta E} \prod_i dq_i dp_i. \tag{72}$$

Uma vez determinada a probabilidade de cada microestado substituimos de volta na expressão da entropia de Shannon para obter o valor máximo da entropia:

$$H_{extr}[P] = \int_{E \leq \mathcal{H}(q_i,p_i) \leq E+\delta E} \prod_i dq_i dp_i \frac{1}{\Omega(E, N, V)} \log \Omega(E, N, V) \tag{73}$$

e o extremo da entropia é

$$S(E, N, V) = \log \Omega(E, N, V). \tag{74}$$

Esta expressão foi escrita primeiramente por Planck, que a chamou de entropia de Boltzmann, por estar implícita nos seus escritos. Expressão equivalente está gravada no seu túmulo ($S = k \log W$).

Não é incomum que os livros de Mecânica Estatística comecem por *postular* que todos os microestados com energia E são igualmente prováveis. A conexão entre a Mecânica e a Termodinâmica é também *postulada* através da equação 74.

Ainda não está claro porque S, o máximo da entropia de Shannon, está relacionada (é) a entropia de Clausius, feito isso iremos descobrir as relações entre quantidades mecânicas e termodinâmicas. E então teremos tipicamente, para avançar em Física Estatística, que olhar o problema da determinação de $\Omega(E, N, V)$ ou a quantidade equivalente em cada sistema e situação experimental. Também veremos que certas situações experimentais são melhor descritas por vínculos que vem na forma de estipulação de valores esperados de certas quantidades. A percepção que há interesse teórico e experimental em descrever vínculos experimentais diferente de simplesmente impor que a energia ou o volume ou o número de partículas é fixo, é devida a Gibbs e à sua teoria dos *ensembles*. Differentes tipos de vínculos experimentais levam a resultados diferentes, cada situação, chamada de *ensemble* recebe um nome específico. O que acabamos de ver, onde a energia do sistema sistema é fixa, porque o sistema está isolado, é chamado de *ensemble* microcanônico. Outros *ensembles* são obtidos por extensões da análise deste caso.

Conexão com a Termodinâmica: T, P, μ

Temperatura

O mesmo sistema isolado, numa caixa de volume V é dividido em dois subsistemas, o sistema 1 de graus de liberdade $\{q_i, p_i\}$, N_1 partículas e volume V_1 e o sistema 2 de graus de liberdade $\{Q_i, P_i\}$, N_2 partículas e volume V_2 . Isto pode ser feito pela inclusão de uma parede ideal que não permite a passagem de átomos de um lado para outro (impermeável) e rígida, não há variações de volume. A separação é feita mantendo

$$E_1 + E_2 = E \tag{75}$$
$$V_1 + V_2 = V, \tag{76}$$
$$N_1 + N_2 = N, \tag{77}$$

tal que todos os termos das equações 76 e 77 conhecidos enquanto só o lado direito da equação 75 é conhecida . Como é feita a distribuição de energia entre os dois subsistemas? Suponhamos que o Hamiltoniano é bem descrito por

$$\mathcal{H}(\{q_i, p_i, Q_i, P_i\}) = \mathcal{H}_1(\{q_i, p_i\}) + \mathcal{H}_2(\{Q_i, P_i\}) \tag{78}$$

onde desprezamos um termo de interação que é muito pequeno e talvez so dependa da superficie da região de contato entre os dois sistemas. Supomos as interações de curto alcance ou se de longo alcance blindadas por cargas opostas. A energia do sistema se dividirá de forma ainda desconhecida entre os dois subsistemas, mas uma vez dividida entre eles, cada subsistema terá uma distribuição uniforme, da mesma forma que no caso da seção anterior. Uma vez estabelecido o equilíbrio cada sistema terá uma energia, um volume e um número de partículas fixos, análogo ao caso já visto. Desprezadas as interações inter sistema, esperamos que as variáveis $\{q_i, p_i\}$ e $\{Q_i, P_i\}$ sejam independentes e a distribuição de probabilidades seja um produto: $P(\{q_i, p_i, Q_i, P_i\}) = P(\{q_i, p_i\})P(\{Q_i, P_i\})$, assim

$$\begin{aligned} H[P(\{q_i, p_i, Q_i, P_i\})] &= -\int_{E_1 \leq \mathcal{H}_1(q_i,p_i) \leq E_1+\delta E_1, E_2 \leq \mathcal{H}_2(Q_i,P_i,) \leq E_2+\delta E_2} \\ &\times \quad P(\{q_i, p_i, Q_i, P_i\}) \log P(\{q_i, p_i, Q_i, P_i\}) \prod_i dq_i dp_i dQ_i dP_i \\ &= -\int_{E_1 \leq \mathcal{H}_1(q_i,p_i) \leq E_1+\delta E_1} \prod_i dq_i dp_i P(\{q_i, p_i\}) \log P(\{q_i, p_i\}) \\ &- \int_{E_2 \leq \mathcal{H}_2(Q_i,P_i,) \leq E_2+\delta E_2} \prod_i dQ_i dP_i P(\{Q_i, P_i\}) \log P(\{Q_i, P_i\}) \\ H[P(\{q_i, p_i, Q_i, P_i\})] &= H[P(\{q_i, p_i\})] + H[(P\{Q_i, P_i\})] \end{aligned} \tag{79}$$

Note que na mesma hipótese sobre o hamiltoniano do sistema e de divisão de energia, o volume no espaço de fase se fatoriza

$$\begin{aligned} \Omega(E, N, V) &= \int_{E_1 \leq \mathcal{H}_1(\{q_i,p_i\}) \leq E_1+\delta E_1,\, E_2 \leq \mathcal{H}_2(\{Q_i,P_i\}) \leq E_2+\delta E_2} \prod_i dq_i dp_i dQ_i dP_i \\ &= \int_{E_1 \leq \mathcal{H}_1(\{q_i,p_i\}) \leq E_1+\delta E_1} \prod_i dq_i dp_i \int_{E_2 \leq \mathcal{H}_2(\{Q_i,P_i\}) \leq E_2+\delta E_2} \prod_i dQ_i dP_i \\ &= \Omega(E_1, N_1, V_1)\Omega(E_2, N_2, V_2), \end{aligned} \tag{80}$$

e os máximos de entropia de cada um dos subsistemas, para uma dada divisão de energia $E = E_1 + E_2$ serão dados pela análise anterior

$$S(E_1, E_2, N_1, N_2, V_1, V_2) = S(E_1, N_1, V_1) + S(E_2, N_2, V_2) \tag{81}$$

onde $S(E_1, N_1, V_1,) = \log \Omega(E_1, N_1, V_1)$ e forma análoga para o sistema 2. Este resultado parece muito bom até o ponto que percebemos que não sabemos a divisão de energias, só o vínculo que $E = E_1 + E_2$.

O que podemos afirmar sobre a partição de energias? A única forma que temos de escolher as distribuiçoes de probabilidade é através da determinação do máximo da entropia e impomos que a entropia do sistema conjunto seja máximo. Basta olhar para o valor da entropia conjunta para uma dada divisão e dentre essas escolher o valor de E_1 que leva ao máximo valor:

$$S(E_1, E_2, N_1, N_2, V_1, V_2) = \log \Omega(E_1, N_1, V_1)\Omega(E_2, N_2, V_2),. \tag{82}$$

$$S(E_1, E_2, N_1, N_2, V_1, V_2) = S(E_1, N_1, V_1,) + S(E_2, N_2, V_2) \tag{83}$$

portanto a maximização da energia sujeita ao vínculo $E = E_1 + E_2$ dará

$$0 = \frac{\partial S(E_1, E_2, N_1, N_2, V_1, V_2)}{\partial E_1} = \frac{\partial S(E_1, N_1, V_1,)}{\partial E_1} + \frac{\partial S(E_2, N_2, V_2)}{\partial E_1} \tag{84}$$

$$= \frac{\partial S(E_1, N_1, V_1,)}{\partial E_1} - \frac{\partial S(E_2, N_2, V_2)}{\partial E_2} \tag{85}$$

$$\left(\frac{\partial S(E_1, N_1, V_1)}{\partial E_1}\right)_{V_1, N_1} = \left(\frac{\partial S(E_2, N_2, V_2)}{\partial E_2}\right)_{V_2, N_2}. \tag{86}$$

Na última equação escrevemos de forma explícita o que é mantido constante ao tomar as derivadas parciais. Reconhecemos a relação que define a temperatura em termodinâmica,

$$\left(\frac{\partial S(E, N, V)}{\partial E}\right)_{V, N} = \frac{1}{T} \tag{87}$$

e reescrevemos a equação 86 como $1/T_1 = 1/T_2$, os sistemas 1 e 2 estarão em equilíbrio quando suas temperaturas forem iguais.

Pressão

Agora e na próxima subseção olharemos para situações experimentais onde um vínculo é relaxado. O sistema conjunto será descrito por uma nova distribuição de probabilidades obtida por maximização da entropia. O vínculo que liberamos agora diz respeito à rigidez da parede que separa o volume V em V_1 e V_2. Ainda temos

$$E_1 + E_2 = E \tag{88}$$

$$V_1 + V_2 = V, \tag{89}$$

$$N_1 + N_2 = N, \tag{90}$$

mas do lado esquerdo, só os termos da equação 90 são conhecidos. Novamente procuramos o máximo da equação

$$S(E_1, E_2, N_1, N_2, V_1, V_2) = S(E_1, N_1, V_1,) + S(E_2, N_2, V_2) \tag{91}$$

mas agora devemos satisfazer duas condições:

$$\left(\frac{\partial S(E_1, E_2, N_1, N_2, V_1, V_2)}{\partial E_1}\right)_{V_1, N_1, N_2} = 0; \quad \left(\frac{\partial S(E_1, E_2, N_1, N_2, V_1, V_2)}{\partial V_1}\right)_{E_1, N_1, N_2} = 0 \tag{92}$$

que levam a

$$\left(\frac{\partial S(E_1, N_1, V_1)}{\partial E_1}\right)_{V_1, N_1} = \left(\frac{\partial S(E_2, N_2, V_2)}{\partial E_2}\right)_{V_2, N_2}. \tag{93}$$

e

$$\left(\frac{\partial S(E_1, N_1, V_1)}{\partial V_1}\right)_{V_1,N_1} = \left(\frac{\partial S(E_2, N_2, V_2)}{\partial V_2}\right)_{V_2,N_2}. \tag{94}$$

que significam respectivamente que

$$\frac{1}{T_1} = \frac{1}{T_2} \; e \; \frac{P_1}{T_1} = \frac{P_2}{T_2} \tag{95}$$

no equilíbrio, não só as temperaturas serão iguais, mas também as pressões, a parede mudará de posição até que as pressões de ambos os lados se igualem.

Potencial químico

Consideremos a situação em que a parede que separa os dois subsistemas é rígida mas não impermeável. Só os termos da equação 89 são determinados pelo arranjo experimental. Podemos supor que o sistema é composto por vários tipos de moléculas

$$N_1^a + N_2^a = N^a \tag{96}$$
$$N_1^b + N_2^b = N^b \tag{97}$$
$$N_1^c + N_2^c = N^c \tag{98}$$

mas a parede só permite a passagem de moléculas do tipo a, sendo os termos da equações 97 e 98 conhecidos e fixos. Obteremos, por maximização da entropia do sistema conjunto que

$$\left(\frac{\partial S(E_1, N_1^a, N_1^b, N_1^c, V_1)}{\partial N_1^a}\right)_{E_1,N_1^b,N_1^c,V_1} = \left(\frac{\partial S(E_2, N_2^a, N_2^b, N_2^c, V_2)}{\partial N_2^a}\right)_{E_2,N_2^b,N_2^c,V_2}. \tag{99}$$

e lembrando da definição termodinâmica do potencial químico $\mu(E, V, N)$:

$$-\frac{\mu}{T} = \left(\frac{\partial S(E, N, V)}{\partial N}\right)_{E,V} \tag{100}$$

$$\frac{1}{T_1} = \frac{1}{T_2} \; e \; -\frac{\mu_1^a}{T_1} = -\frac{\mu_2^a}{T_2} \tag{101}$$

e o equilíbrio só é obtido quando a temperatura e o potencial químico da espécie a são iguais dos dois lados da parede.

Exercício

Discuta o problema que resulta ao considerar subsistemas separados por uma parede que permita mudança de volume V mas não de E: a parede adiabática móvel.

Conexão com a Termodinâmica

As condições de equilíbrio que vimos nas seções anteriores: igualdade de temperatura, pressão ou potencial químico, permitem fazer o pulo e dizer que o extremo da entropia de Shannon, sob os vínculos adequados, isto é, a entropia de Boltzmann-Gibbs é a entropia termodinâmica de Clausius. Não é por acaso que a forma funcional da entropia como função de (E, V, N) - os parâmetros chamados *extensivos*- recebe o nome de relação fundamental em termodinâmica. Se a tivermos, qualquer quantidade de interesse termodinâmico pode ser calculado. Temos uma relação fundamental $S(E, V, N) = \log \Omega$, onde o número Ω de microestados relevantes para o conjunto de experiências em questão pode ser calculado a partir do conhecimento do hamiltoniano do sistema. Temos, então a conexão da Termodinâmica com a Mecânica, seja ela Clássica ou Quântica. O trabalho que segue, no que diz respeito às aplicações consiste em calcular a entropia. É comum que este cálculo não seja trivial. Muitas das técnicas de aproximação que serão desenvolvidas requerem uma compreensão profunda do problema. Os capítulos que seguem darão exemplos de aplicações.

O formalismo apresentado traz um paralelo total com a termodinâmica. A determinação do estado de equilíbrio em termodinâmica é feito a partir de um princípio de extremo. Ao liberar o sistema de um vínculo, ou ao impor um novo vínculo, um novo estado de equilíbrio resulta. A escolha dentro de todos os possíveis estados compatíveis com os vínculos é feito pela maximizaçao da entropia, expressa pela relação fundamental. Note que os outros estados compatíveis com os vínculos impostos, não ocorrem. Mas cada tal estado poderá ocorrer se for feita a imposição de algum novo vínculo específico adicional. A escolha do estado termodinâmico é feita dentre aqueles que satisfazem os vínculos que sabemos existirem e não mais que isso.

Qual é a justificativa por trás desse princípio de extremo? Uma possibilidade é que o estudante de Física ao chegar a esse ponto esteja acostumado a princípios de extremo, que terá visto em Mecânica Clássica ou Quântica, Eletromagnetismo e suas aplicações. Devido a esse costume talvez não ousará discutir a possibilidade de mais um. A conexão com a teoria de informação nos traz a justificativa do princípio de extremo. O estado escolhido, não pelo sistema, mas por nós, é aquele que faz menos hipóteses não justificadas. Pela primeira vez um princípio de extremo está sendo apresentado sem dizer que essa é a forma que funciona a natureza. Neste caso esta é a forma que funciona a maneira de fazer previsões o mais honestas possíveis.

Neste ponto o estudante de Física pode voltar a argumentar: "Porque o que eu sei (vínculos impostos) tem alguma influência sobre o que o experimental mede no laboratório?". A resposta é simples. Não

tem nenhuma influência! Mas aquilo que *sabemos* tem uma influência direta sobre as *previsões* que fazemos. Se a informação que temos não for boa, suficiente, relevante...as previsões teóricas sobre as experiências, serão igualmente ruins. Voltamos ao debate entre subjetivo e objetivo? Só mais uma vez. O método pode parecer subjetivo ao ser olhado do ponto de vista que as previsões que eu faço dependem do que sei. Como poderia ser diferente? Deveriam depender do que não sei? A Mecânica Estatística fornece um conjunto de regras, se duas pessoas tiverem a mesma informação sobre o sistema e aplicarem as regras de forma adequada, farão previsões iguais. Nada pode ser mais objetivo que isso. Se eles tiverem informações diferentes, muito possívelmente farão previsões diferentes e o resultado experimental servirá de juiz. Se a contagem de estados for feita usando a informação dada pela Física Clássica poderá haver erros ao comparar com a experiência. Repetimos, os primeiros resultados da Mecânica Quântica foram obtidos de forma a que as previsões da Mecânica Estatística refletissem os resultados experimentais no problema do Corpo Negro. Não foi mudada a ME mas sim a informação usada.

Os parâmetros extensivos escalam com o tamanho do sistema. Isto está claro para E, V e N. Devemos provar a extensividade da entropia. Isto é postulado na termodinâmica. Neste trabalho também o é, isso ocorreu no momento que colocamos aditividade ante agrupamentos. Há possibilidade de mudar este "postulado" e obter outra termodinâmica? Este é um tópico de grande discussão no momento mas que não nos interessa aqui [11].

A equação 78 mostra sob que condições ocorre a extensividade. Consideremos um sistema num estado (E, V, N) com interações de curto alcance. Imaginemos, sem fazer, uma separação em dois sistemas 1 e 2 como nas seções anteriores. Façamos a separação em K partes

$$S(E, V, N) = \sum_{i=1,K} S(E_i, V_i, N_i) \tag{102}$$

com $E = \sum_{i=1,K} E_i$, $V = \sum_{i=1,K} V_i$ e $N = \sum_{i=1,K} N_i$. Se os E_i, V_i e N_i das partes forem iguais ,teremos

$$S(E, V, N) = KS(E/K, V/K, N/K) \tag{103}$$

Podemos estender isto para números reais em geral e não só para inteiros e concluimos que se E, V e N forem proporcionalmente mudados a entropia também o será e chamamos essa propriedade de extensividade. Além dos parâmetros extensivos temos os intensivos T, P e μ identificados pelo papel que desempenham na determinação do estado de equilíbrio de dois sistemas em contato entre si, isolados do resto do mundo. Os parâmetros intensivos, que são derivadas parciais de quantidades extensivas com respeito a outras extensivas, não mu-

[11] Faremos apenas um breve comentário. A entropia cruzada pode ser usada para definir uma distância, ao menos para pequenas diferenças. Há muitas formas de introduzir distância. Há algumas preferíveis a outras. A forma de Kullback-Leibler é a única invariante ante Markov embeddings (Ver Cenkov) Se há um princípio geral de inferência, então entropias não extensivas podem levar a resultados no mínimo estranhos ao lidar com sistemas independentes. Previsões de medidas termodinâmicas feitas sobre um sistema na terra sob a hipótese que existe a estrela A dariam resultados diferentes sob a hipótese que não existe a estrela. Fazendo a medida poderiamos decidir se existe ou não a estrela. Aparentemente este tipo de argumento não convence todos os pesquisadores e ainda há cálidas discussões. (inserir referencias)

dam com a escala. Seja T a temperatura de um sistema. Consideremos que é composto de duas partes, S, E, V e N de cada parte muda mas a temperatura é a mesma.

Para pequenas mudanças nos parâmetros extensivos, a entropia ($S = \log \Omega$) muda de acordo com

$$dS(E,V,N) = \left(\frac{\partial S}{\partial E}\right)_{VN} dE + \left(\frac{\partial S}{\partial V}\right)_{EN} dV + \left(\frac{\partial S}{\partial N}\right)_{EV} d\mu \qquad (104)$$

e usando as definições dos parâmetros intensivos

$$dS(E,V,N) = \frac{1}{T}dE + \frac{P}{T}dV - \frac{\mu}{T}d\mu \qquad (105)$$

Ensemble Canônico

Há vezes em que uma experiência termodinâmica é feita numa situação diferente daquela em que o sistema está dentro de paredes isolantes, rígidas e imperméaveis. Por exemplo podemos olhar uma transformação a temperatura constante, a pressão constante ou a potencial químico constante. Queremos investigar o que significa isto em termos de vínculos de infromação. Devemos olhar de novo a seção . Obtivemos resultados gerais para casos onde o valor esperado de certas quantidades é conhecido. O ensemble canônico de Gibbs é obtido quando o sistema não está isolado, mas a temperatura é fixa. Como implementar o vínculo informacional que a temperatura é fixa? Usamos a equação 62, que nos dá indícios de como proseguir.

Suponhamos que não fixamos a temperatura mas sim o valor esperado da energia. Como isso é feito numa experiência? Ficará claro a seguir. É dado que o valor esperado da energia

$$< \mathcal{H} >= \int \prod_i dq_i dp_i P(q_i, p_i) \mathcal{H}(q_i, p_i) \qquad (106)$$

é fixo, podemos supor que não o conhecemos, mas tem um valor definido. Se soubessemos esse valor poderiamos aplicar os resultados da seção e obter a distribuição - de Boltzmann, de Boltzmann-Gibbs, canônica - em analogia à equação 57, que agora toma a forma

$$P(\{q_i, p_i\}) = \frac{e^{-\beta \mathcal{H}}}{Z} \qquad (107)$$

onde o multiplicador de Lagrange que chamamos β em lugar de λ, pode ser obtido da equação 62

$$\beta = \frac{\partial S(E,V,N)}{\partial E} \qquad (108)$$

e usamos a notação $E =< \mathcal{H} >$. Mas a derivada da entropia com respeito à energia é o inverso da temperatura. Portanto identificamos

β com o inverso da temperatura:

$$\beta = \frac{1}{T} \tag{109}$$

O curioso deste resultado é que ao fazer previsões para uma experiência onde conhecemos o multiplicador de Lagrange, mas não o valor esperado, agimos da mesma forma que se soubessemos o valor esperado. Saber o valor do multiplicador de Lagrange, do ponto de vista da informação, é equivalente a saber o valor esperado da energia.

A função de partição Z adquire uma interpretação importante. Note que a equação 25 toma a forma

$$Z = \int \prod_i dq_i dp_i e^{-\beta \mathcal{H}} \tag{110}$$

e a equação 59 fica assim

$$E = < \mathcal{H} > = -\frac{\partial \log Z}{\partial \beta} \tag{111}$$

A expressão para a entropia (o extremo), que obtemos substituindo a distribuição canônica no funcional de entropia de Shannon nos dá

$$\begin{aligned} S(E,V,N) &= -\int \prod_i dq_i dp_i P(q_i,p_i) \log P(q_i,p_i) && (112)\\ &= -\int \prod_i dq_i dp_i \frac{e^{-\beta\mathcal{H}}}{Z} \log\left(\frac{e^{-\beta\mathcal{H}}}{Z}\right) && (113)\\ &= -\int \prod_i dq_i dp_i \frac{e^{-\beta\mathcal{H}}}{Z} (-\beta\mathcal{H} - \log Z) && (114)\\ &= \beta < \mathcal{H} > -\beta \log Z && (115)\\ &= E - \beta \log Z. && (116) \end{aligned}$$

Rearranjando obtemos $-\log Z/\beta = E - TS$, ou seja podemos identificar o logaritmo da função de partição e a energia livre de Gibbs

$$\begin{aligned} F(T,V,N) &= -\frac{1}{\beta} \log Z(\beta,V,N) && (117)\\ F &= E - TS && (118) \end{aligned}$$

Comparação da Entropia nos diferentes Ensembles

Encontramos duas formulações, o ensemble microcanônico e o canônico. O que tem em comum e o que tem de diferente? Esperamos que as entropias encontradas sejam as mesmas? A resposta é sim e não. Não de forma trivial, a informação é diferente, a situação experimental sob análise. é diferente e portando o valor numérico também deverá sé-lo. No entanto ao descrever sistemas físicos que incluem um número muito grande de graus de liberdade os valores sob condições "iguais" e as derivadas ou seja , as equações de estado, serão iguais.

Exercício: Ensemble Grande Canônico

Considere a situação em os vínculos de informação sobre a experiência são o valor esperado da energia e do número de partículas. A notação que usamos é que $P(q_i, p_i, n) = P(n)P(\{q_i, p_i\}|n)$ denota a probabilidade atribuimos a que o sistema tenha n particulas e elas estejam no estado $\{q_i, p_i\}$. Obviamente dado um valor de n o número de variáveis depende de n, e.g. poderia ser $6n$ se o sistema estiver em três dimensões. O equivalente aos vínculos $< g_k >$ da equação 54 tomam a forma

- $\sum_{n=1}^{\infty} \int P(q_i, p_i, n) \prod dq_i dp_i = 1$
- $\sum_{n=1}^{\infty} \int \mathcal{H}(\{q_i, p_i\}, n) P(q_i, p_i, n) \prod dq_i dp_i =< \mathcal{H} >= E$
- $\sum_{n=1}^{\infty} \int nP(q_i, p_i, n) \prod dq_i dp_i =< n >= N$

Encontre a distribuição de probabilidades chamada grande canônica ou grã canônica que maximiza o funcional de entropia sob os vínculos acima. Mostre que o multiplcador de Lagrange para o vínculo sobre o número de partículas está relacionado ao potencial químico $\lambda = \mu/T$, veja a equação 100.

Apêndice: Multiplicadores de Lagrange

O problema de encontrar pontos extremos ou só estacionários de funções sujeitos a vínculos é muito vasto. Damos algumas idéias básicas sem preocupação com rigor, para lembrar o estudante de técnicas que deveriam ser vistas em Cálculo 2 ou curso equivalente.

Seja o problema

- P_{livre}: Queremos encontrar um ponto (x^*, y^*) dentro de uma certa região C no plano real onde uma função $f(x, y)$ tem localmente um valor estacionário.

Fácil, tome as derivadas parciais e resolva o sistema $\partial_x f = 0$, $\partial_y f = 0$.

Queremos, a seguir resolver um caso mais difícil.

- P_{vinc}: Suponha que não procuramos o resultado em qualquer lugar de C, mas especificamente queremos um ponto estacionário entre aqueles pontos que satisfazem $\phi(x, y) = c$, que supomos descreva uma curva no plano que está parcialmente contida em C e chamaremos γ.

A solução do parágrafo anterior dificilmente nos dá a resposta pois seria uma coincidência se (x^*, y^*) caisse encima dessa curva.

A solução a esta classe de problema foi proposta por Lagrange. Considere a classe de funções $F_\lambda(x, y)$ que dependem do chamado

multiplicador de Lagrange λ:

$$F_\lambda(x,y) = f(x,y) + \lambda(\phi(x,y) - c) \tag{119}$$

Note que se o ponto de coordenadas x e y estiver na curva γ então F_λ e f tem o mesmo valor. Repetindo: F_λ e f tem o mesmo valor se o vínculo que " x e y estão sobre γ" for respeitado.

Consideremos o P_{livre} mas para a função $F_\lambda(x,y)$. O problema é novamente simples [12]. Resolvemos o sistema

[12] a não ser que não seja....

$$\frac{\partial F_\lambda}{\partial x} = 0 \tag{120}$$

$$\frac{\partial F_\lambda}{\partial y} = 0, \tag{121}$$

onde λ ainda não foi especificado. A resposta depende de valor escolhido para λ, isto é define uma curva ρ, parametrizada por λ : $(x^*(\lambda), y^*(\lambda))$ onde F_λ é extremo . Agora voltamos ao problema P_{vinc}. Da resposta à dupla pergunta "onde f é máximo?" e "onde o vínculo é satisfeito?" , quando as duas são respondidas simultaneamente, decorre a solução. Substituimos a primeira por "onde F é máximo?" (resposta: em ρ) junto com a afirmação "$f = F_\lambda$ sob a condição de estar em γ". Segue que queremos encontrar o cruzamento de γ e ρ. Basta escolhermos $\lambda = \lambda_*$ tal que $\phi\,(x^*(\lambda_*), y^*(\lambda_*)) = c$, o resultado é um extremo para f e satisfaz o vínculo.

Agora procure um livro de cálculo e prencha os detalhes. Discuta também como lidar com casos em que o extremos está na borda de C. Há vínculos que são representados por desigualdades, os nomes de Kuhn e Tucker estão associados a esta extensão.