

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 supô-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 conseqüências 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 conseqüências 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 também está relacionado à atualização de 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 conseqüências sobre os métodos para atualizar ou atribuir 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 atribuído 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.

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 certa informação que pode não ser verdadeira. É melhor não saber do que achar que se sabe algo que está errado ¹.

¹ 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 nossa 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 ignorância. 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 simetria 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 atribuída 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 $\bar{s}$ 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.

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

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

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

Pequenas mudanças na distribuição $\{p_i\}$ não devem levar a grandes mudanças em $H[P]$.

- (2) $H[p]$ deve ser contínua nos p_i .

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 $H[1/n, 1/n, \dots, 1/n] = F(n)$. 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

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

Outro caso particular onde sabemos algo pois achamos que se não fosse assim, algo estranho estaria acontecendo, é nosso velho amigo: Não queremos ser manifestamente inconsistentes, portanto o mais consistente possível é impor:

- (4) Se há mais de uma forma de analisar 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 um dado cúbico. Os estados possíveis do sistema, para simplificar, são rotulados pelo número de pontos na face que aponta para cima. Podemos agrupar em dois grupos e.g. de { pares ou ímpares }, ou de { maiores ou menores que 3.5 }, ou de { primos ou não-primos }. Outra forma poderia ser menos uniforme, e.g. de { maiores ou menores que 2.5 } ou de muitas outras maneiras, e.g. em mais de dois grupos.

Suponha que X possa ter n valores possíveis $\{x_i\}_{i=1\dots n}$. Estes *microestados* são mutuamente exclusivos. Queremos atribuir as probabilidades p_k . 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 escolheremos outro conjunto e o manteremos fixo novamente. Dentro de um grupo g os estados são indexados por i . Podemos escrever

$$k(i, g) = \sum_{g'}^{g-1} m_{g'} + i$$

e fixos os $\{m_g\}$, dado g e i obtemos k . A probabilidade de que X tenha um valor x_k é dada pela regra do produto em termos das probabilidades de estar em um grupo e de ter um índice i dado que está num grupo:

$$p_k = P(k) = P(i, g | \{m_g\}) = P(i | g \{m_g\}) P(g | \{m_g\})$$

e pela regra da soma, a probabilidade que esteja no grupo g , por $P(g | \{m_g\}) = \sum_{i \in g} P(i | g \{m_g\}) = P_g$ assim

$$P(i | g \{m_g\}) = \frac{P(k)}{P(g | \{m_g\})}.$$

Se for dado que está no grupo g , a probabilidade que esteja no estado i é $p(i | g) = \frac{p_{k(i,g)}}{P_g}$. 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

- (5) 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, junto com a consistência do item (3) nos dá

$$H[p_k] = H_G[P_g] + \sum_{g=1}^N P_g H_g[p(i|g)] \quad (10.1)$$

Que isto é suficiente para determinar a forma funcional da

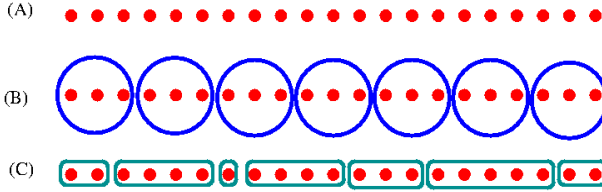


Figura 10.1: (A) Cada ponto (vermelho) representa um possível estado (mutuamente exclusivos) de uma variável. (B) Agrupamos em grupos com o mesmo tamanho. (C) Os agrupamentos são arbitrários.

entropia, será provado a seguir e será feito em dois passos, para cada um fazemos um dos dois tipos de agrupamento mencionados acima.

(I) Começamos analisando o caso particular (figura 10.1.B) 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 $p(i|g) = N/n = 1/m$

$$H[1/n \dots 1/n] = H_G[1/N, \dots, 1/N] + \sum_{g=1}^N \frac{1}{N} H_g[1/m, \dots, 1/m] \quad (10.2)$$

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

$$F(n) = F(N) + \sum_{g=1}^N \frac{1}{N} F(m) \quad (10.3)$$

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

$$F(mN) = F(N) + F(m) \quad (10.4)$$

A solução óbvia é dada por

$$F(n) = k \log n. \quad (10.5)$$

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 logaritmo 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.

(II) Ainda não obtivemos a forma de H no caso geral. Passamos a analisar o caso em que os tamanhos dos agrupamentos m_g são arbitrários (figura 10.1.C), salvo o fato $\sum_{g=1}^N m_g = n$, mas ainda temos que $p_k = 1/n$ é uniforme. Temos então que a probabilidade $P_g = m_g/n$ é arbitrária, mas a probabilidade condicional dentro de cada grupo é uniforme: $p(i|g) = 1/m_g$

$$H[1/n, \dots, 1/n] = H_G[P_g] + \sum_{g=1}^N P_g H_g[1/m_g, \dots, 1/m_g] \quad (10.6)$$

e substituímos a entropia da distribuição uniforme:

$$F(n) = H_G[P_g] + \sum_{g=1}^N P_g F(p(i|g)), \quad (10.7)$$

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

$$H_G[P_g] = F(n) - \sum_{g=1}^N P_g F(m_g). \quad (10.8)$$

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

$$H_G[P_g] = \left(\sum_{g=1}^N P_g \right) k \log n - k \sum_{g=1}^N P_g \log(m_g) \quad (10.9)$$

$$H_G[P_g] = -k \sum_{g=1}^N P_g \log(m_g/n) \quad (10.10)$$

e finalmente

$$H_G[P_g] = -k \sum_{g=1}^N P_g \log(P_g) \quad (10.11)$$

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, Wiener, 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. 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 discutida no próximo capítulo e não entraria aqui em um par de parágrafos. De Shannon

ing também
2da Guerra
a decifrar
teve menos
dencial. Ver

e Wiener ² vem idéia de discutir informação do ponto de vista de volume, quantidade com o objetivo de entender limitações para a transmissão em canais de comunicação. Isto inclui 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 ³.

³ Não cabe no curso introdutório de Mecânica Estatística

Um método variacional

O caminho para as aplicações está aberto. O método de escolha das distribuições de probabilidade é de maximizar a ignorância. De todas as distribuições compatíveis com os vínculos, escolhemos a que introduz a menor quantidade de informação adicional aos vínculos. Reduzimos a inferência a um método variacional. Isto não deve vir como uma surpresa, pois toda a Física é essencialmente redutível a métodos variacionais. A pergunta profunda é se esses métodos estão relacionados ou se são independentes. A resposta que está emergindo é que todos esses métodos estão relacionados. E isto transcende a Física e se estende por outras áreas. Uma contribuição importante de Jaynes foi a percepção que a maximização da entropia sujeita aos vínculos da informação leva a um método fundamental de inferência, chamado por ele de MaxEnt. Outros nomes estão associados a extensão de idéias de entropia como método de inferência fundamental ⁴. Em particular tem se conseguido nos últimos anos uma formulação muito mais satisfatória e menos ad hoc dos princípios por trás deste método. A generalização do método de máxima entropia para inferência em problemas com variáveis em variedades contínuas pode levar a problemas conceituais que serão atacados a seguir.

⁴ ver A.C

Entropia Relativa

As deduções aqui deveriam ser lidas após treinamento sobre métodos variacionais, que será abordado nos próximos capítulos. Portanto talvez seja recomendável pular esta seção numa primeira leitura. A entropia de Shannon vai ser usada para inferências em quase todo o resto do livro. Mas esta não é a última palavra em mecanismos de inferência. A utilidade de entropia de Shannon foi a de *atribuição* de probabilidades. Isso deve dar a impressão que do nada, as probabilidades são construídas através da imposição de certos vínculos sobre funções dos graus de liberdade. Dificilmente a inferência é feita a partir de nenhum conhecimento prévio. Para

qualquer situação é mais razoável que um estado de crenças seja modificado pela imposição de novos vínculos. Isso nos leva à definir Informação como tudo aquilo que leva a modificar as crenças codificadas em distribuições de probabilidades. Procuramos um mecanismo que sirva para *atualizar* as crenças a partir de uma crença *a priori*. A diferença importante pode ser vista no confronto de *atualizar* a partir de um estado de conhecimento *apriori* versus *atribuir* a partir do vácuo informacional.

Qual deve ser o mecanismo de mudar crenças? Há em geral muitas distribuições de probabilidade que satisfazem um certo conjunto de vínculos. A idéia é fazer um ranking de todas elas. Isto significa atribuir um número a cada distribuição de probabilidades e escolher aquela que maximiza o ranking de preferências. Ou seja devemos construir um funcional que atribui um número a cada distribuição. Claro que esse número deve depender no estado de crença antes da imposição dos novos vínculos. Como escolher o funcional?

Novamente procuramos casos simples onde sabemos algo.

Impomos como desejo que o funcional dê a resposta esperada neles. O método de obtenção do funcional é o de eliminação de todos os funcionais que falham em satisfazer esses casos simples. Resta algum funcional após este processo? Poderia ser que não, mas não é o caso. Chamaremos este funcional de Entropia Relativa $S[p||q]$, entre a distribuição p candidata a ser escolhida pelo processo de inferência e a distribuição *a priori* q . Também pode ser chamado de (menos) a divergência de Kullback-Leibler (KL). Por economia, será simplesmente chamada entropia. A menos de pequenos detalhes, a dedução segue A. Caticha de forma simplificada.

Desiderata Entrópica

Os desejos expostos a seguir são considerados em adição aos do Capítulo 1. Portanto a teoria de probabilidades é a estrutura matemática adequada. Mas agora faremos algumas demandas adicionais. Pedimos ao leitor que encontre argumentos contra os seguintes desejos para uma teoria de inferência:

- DE_1 : **Generalidade**: O método deve ser Universal.

Queremos lidar com informação incompleta e o algoritmo de procedimento deve ser o mesmo independente de qual o tipo de problema de inferência que estamos tratando. A descrição do problema, através da escolha dos graus de liberdade apropriados e dos vínculos impostos levará em conta a natureza explícita do problema.

- DE_2 : **Parcimônia** A atualização deve ser minimalista.

De todas as distribuições $p(x)$ que satisfazem o vínculo devemos escolher a que menos mude a distribuição *a priori* $q(x)$. Claro que o que quer dizer mínimo deve ser definido. Mas há um caso simples em que mínimo é fácil de decidir. No caso em que não há nova

informação, então a menor mudança é não fazer nada. Nosso método deve ser tal que $p(x) = q(x)$ após a incorporação de nenhuma informação. Parece muito trivial mas será útil.

- DE_3 : **Localidade** no espaço de configurações.

Quando o espaço de configurações pode ser dividido em duas partes disjuntas, informação que diz respeito somente a uma das partes, ao ser incorporada, não deve alterar a atribuição de probabilidades às configurações da outra.

- DE_4 : **Invariância**: A escolha dos rótulos dos graus de liberdade é uma convenção: não deve alterar o resultado da Inferência.

Para sistemas de variáveis que tomam valores num sub espaço de $\mathbb{R}^N$, a escolha do sistema de coordenadas não deve interferir no resultado do problema. Assim o funcional deve ser invariante ante mudanças (contínuas diferenciáveis) do sistema de coordenadas.

- DE_5 : **Independência**: Sistemas independentes devem ser descritos da mesma forma quando estudados separadamente ou em conjunto.

Há situações em que sabemos que há sistemas independentes. Consideremos uma garrafa de café e a galáxia X9. Ao fazer previsões sobre a termodinâmica do café podemos (i) tratar o sistema de café sozinho, ou (ii) incluir os graus de liberdade da galáxia X9 além dos do café, e marginalizar a distribuição, integrando sobre as configurações da galáxia. Os dois métodos devem dar o mesmo resultado sob pena de ao realizar medidas experimentais, poder distinguir entre os casos em que a galáxia foi ou não incluída. Note que poderíamos, caso dessem diferentes, fazer astronomia olhando para uma garrafa de café.

DE_3 : Localidade

Começamos com as consequências de DE_3 . As configurações de um sistema formam o conjunto $\chi = \{x_1, x_2, \dots, x_n\}$. Podemos separar em dois conjuntos $i \in A$ e $j \in B$, tal que $A \cap B = \emptyset$ e $A \cup B = \chi$, ou seja mutuamente exclusivos e exaustivos. Um caso simples é quando a informação é dada através dos seguintes vínculos

$$P_A = \sum_{x_i \in A} p_i, \quad (10.12)$$

$$P_B = \sum_{x_j \in B} p_j, \quad (10.13)$$

$$\text{com } P_A + P_B = 1, \quad (10.14)$$

$$G = \sum_{x_j \in B} p_j g(x_j), \quad (10.15)$$

onde $p_k = p(x_k)$. Maximizar a entropia sujeita a estes vínculos leva a

$$0 = \delta \left[S - \lambda_1 \left(\sum_A p(x_i) - P_A \right) - \lambda_2 \left(\sum_B p(x_j) - P_B \right) - \lambda_3 \left(\sum_B p(x_j) g(x_j) - G \right) \right]$$

para variações independentes de quaisquer dos p . Temos n equações:

$$\frac{\partial S}{\partial p_i} = \lambda_1 \quad (10.16)$$

$$\frac{\partial S}{\partial p_j} = \lambda_2 + \lambda_3 g(x_j). \quad (10.17)$$

$$(10.18)$$

Em geral teríamos, para $k \in A$ ou $k \in B$, chamando $\frac{\partial S}{\partial p_k} = f_k$, a dependência

$$\frac{\partial S}{\partial p_k} = f_k(p_1, \dots, p_n, q_1, \dots, q_n). \quad (10.19)$$

Comparando as equações 10.16 e 10.19 vemos que

$$\lambda_1 = f_i(p_1, \dots, p_n, q_1, \dots, q_n) \quad (10.20)$$

$$= f_i(p_i, q_i). \quad (10.21)$$

A segunda equação decorre de que mudando a informação que altera p_j do conjunto B , poderíamos mudar o lado direito sem alterar o multiplicador de Lagrange que é constante. Ainda mais, podemos considerar todas as possíveis divisões de χ em subconjuntos exaustivos e mutuamente exclusivos. Eliminariamos a dependência em todos os índices diferentes de i . Assim

$$\frac{\partial S}{\partial p_k} = f_k(p_k, q_k) \quad (10.22)$$

Integrando, vemos que a forma geral do funcional é

$$S = \sum F_k(p_k, q_k)$$

onde F_k com $f_k = \frac{\partial F_k}{\partial p_k}$ são funções ainda desconhecidas. A generalização para variáveis que tomam valores reais:

$$S = \int F(p(x), q(x), x) dx. \quad (10.23)$$

DE₄: Invariância

Vamos começar com um problema em uma dimensão. Uma densidade de probabilidades de uma variável X , $m(x)$ satisfaz

$$\int m(x) dx = 1$$

Podemos mudar variáveis $x \rightarrow x'$ e a densidade mudará $m(x) \rightarrow m'(x')$. A probabilidade de uma região dx em x deve ser preservada, portanto

$$m(x) dx = m'(x') dx'$$

e isso deve valer para a transformação de qualquer densidade. Em mais de uma dimensão teremos

$$m(x) = m'(x') J(x')$$

onde $J(x') = \det \left| \frac{\partial x'}{\partial x} \right|$ é o Jacobiano da transformação $x \rightarrow x'$.

A estratégia a seguir é a seguinte: fazemos a inferência no sistema de coordenadas x , obtendo um resultado. A seguir, a fazemos no referencial x' . Transformamos os x' de volta a x e comparamos. Os dois resultados deveriam dar igual. Mas não ocorre a não ser que imponhamos condições sobre a função F que aparece na equação 10.23. Isso restringe ainda mais os funcionais sobreviventes. Antes de prosseguir, facilita introduzir uma densidade $m(x)$ que por agora está à nossa disposição escolher. A motivação é que a razão de duas densidades tem propriedades simples de transformação

$$\frac{p(x)}{m(x)} = \frac{p'(x')}{m'(x')}, \quad (10.24)$$

pois o Jacobiano se cancela ⁵. Assim consideramos que a equação 10.23 pode ser escrita

⁵ Cabe ressaltar que deveríamos ter cuidado de afirmar que as regiões em que as diferentes distribuições se anulam são as mesmas e são excluídas destas considerações.

$$\begin{aligned} S &= \int F(p(x), q(x), x) dx = \int \frac{1}{m(x)} F\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x\right) m(x) dx \\ &= \int \Phi\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x\right) m(x) dx, \end{aligned} \quad (10.25)$$

onde introduzimos a ainda desconhecida função

$$\Phi(u, v, m, x) = \frac{1}{m} F(um, vm, x).$$

Queremos mudar as coordenadas de forma geral, mas é conveniente se restringir a um caso fácil. O vínculo imposto será

$$\int p(x) a(x) = A,$$

onde a função $a(x)$ é um escalar, o que significa que $a(x) \rightarrow a'(x') = a(x)$, não muda ante transformações de coordenadas. A normalização é um exemplo de vínculo escalar, portanto não é preciso considerá-lo separadamente. O cálculo variacional $\delta(S + \text{vínculos}) = 0$, leva a

$$\begin{aligned} 0 &= \frac{\delta}{\delta p(x)} \left(S - \lambda \int p(x) a(x) dx \right), \\ &= \Phi\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x\right) - \lambda a(x), \end{aligned} \quad (10.26)$$

onde

$$\dot{\Phi}(u, v, m, x) = \frac{d}{du} \Phi(u, v, m, x).$$

Analogamente, começando no sistema de coordenadas x' , chegamos a

$$0 = \dot{\Phi}\left(\frac{p'(x')}{m'(x')}, \frac{q'(x')}{m'(x')}, m'(x'), x'\right) - \lambda' a'(x'). \quad (10.27)$$

Usando a equação 10.24, esta última equação pode ser reescrita como

$$0 = \dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m'(x'), x'\right) - \lambda' a'(x'). \quad (10.28)$$

Para vínculos escalares, segue que, a razão

$$\frac{\dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x\right)}{\dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m'(x'), x'\right)} = \frac{\lambda}{\lambda'} \quad (10.29)$$

é uma constante, pois os multiplicadores de Lagrange não dependem de x nem x' . Como isto vale para qualquer transformação contínua e diferenciável de $x' = \phi(x)$, podemos olhar para diferentes escolhas de ϕ e eliminar candidatos. Primeiro, olhamos para transformações com Jacobiano igual a 1 em todo o espaço. Segue que $m(x) = m'(x')$, portanto a única diferença entre o numerador e o denominador é no quarto argumento. Como a razão é constante, $\dot{\Phi}$ não pode depender do quarto argumento x ou x' . Agora consideramos, novamente com Jacobiano igual a 1, a função $\phi(x) = x, x \in \bar{D}$ e $\phi(x) \neq x, x \in D$. Se $x \in \bar{D}$, a razão é 1:

$$\begin{aligned} 1 &= \frac{\dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x)\right)}{\dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m'(x')\right)} \\ &= \frac{\lambda}{\lambda'} \end{aligned}$$

Mas para $x \in D$, não muda a razão. Portanto $\dot{\Phi}$ não pode depender agora do seu terceiro argumento e

$$1 = \frac{\dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}\right)}{\dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}\right)} \quad (10.30)$$

é trivialmente satisfeita. Segue que o funcional só pode ter esta estrutura:

$$S = \int \Phi\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}\right) m(x) dx. \quad (10.31)$$

DE₂ : Parcimônia

Temos à nossa disposição este intruso $m(x)$. Podemos escolher qualquer densidade? Impondo o desejo de parcimônia veremos que não é possível essa liberdade. Estamos interessados em não fazer nada. O aluno não deve se entusiasmar desnecessariamente. Se não chega informação, não deveria mudar nossa escolha de $p(x)$.

Portanto a solução do problema variacional

$$\begin{aligned} \frac{\delta}{\delta p(x)} \left(S - \lambda \int p(x) dx \right) &= 0, \\ \dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}\right) &= \lambda, \end{aligned} \quad (10.32)$$

deve levar à escolha $p(x) = q(x)$. Mas nem todo funcional do tipo 10.31 leva a esse resultado. Isto restringe mais a família de funcionais. O problema 10.32 é uma equação diferencial ordinária

$$\begin{aligned} \dot{\Phi}(u, v) &= \lambda \\ \Phi(u, v) &= \lambda u + c(v) \end{aligned}$$

portanto, se seguissemos este caminho teríamos

$$\begin{aligned} S &= \int \left(\lambda \frac{p(x)}{m(x)} + c \left(\frac{q(x)}{m(x)} \right) \right) m(x) dx \\ &= \lambda \int p(x) dx + \text{const}, \end{aligned}$$

o que é uma trivialidade, pois a normalização diz que isto é uma constante e toda distribuição normalizada teria a mesma preferência. Portanto escolher a função Φ leva a um beco sem saída. Mas há outra saída para satisfazer a equação 10.32, basta escolher $m(x) = q(x)$, neste caso

$$\Phi\left(\frac{p(x)}{q(x)}, \frac{q(x)}{q(x)}\right) = \lambda, \quad (10.33)$$

que é constante de forma trivial quando $p(x) = q(x)$. Novamente temos um avanço, restringindo ainda mais as estruturas possíveis:

$$S = \int \Phi\left(\frac{p(x)}{q(x)}\right) q(x) dx. \quad (10.34)$$

DE₅ : Independência

Este último desejo será usado para determinar a forma explícita de Φ na equação 10.34. Além da normalização, para o sistema composto pela garrafa de café, com graus de liberdade x_1 e a galáxia X9 com graus de liberdade x_2 , temos os seguintes vínculos

$$\begin{aligned} \int p_1(x_1) f_1(x_1) dx_1 &= F_1 \\ \int p_2(x_2) f_2(x_2) dx_2 &= F_2. \end{aligned}$$

Tratando os dois sistemas de forma independente temos que $p_1(x_1)$ e $p_2(x_2)$ são dados pelas soluções de

$$\Phi\left(\frac{p_1(x_1)}{q_1(x_1)}\right) = \lambda_{01} + \lambda_1 f_1(x_1) \text{ e } \Phi\left(\frac{p_2(x_2)}{q_2(x_2)}\right) = \lambda_{02} + \lambda_2 f_2(x_2) \quad (10.35)$$

Tratando os sistemas juntos, o problema variacional é

$$\begin{aligned} 0 &= \frac{\delta}{\delta p(x_1, x_2)} \left\{ S - \mu_0 \left(\int p(x_1, x_2) dx - 1 \right) \right. \\ &\quad - \mu_1 \left(\int f_1(x_1) p(x_1, x_2) dx_1 dx_2 - F_1 \right) \\ &\quad \left. - \mu_2 \left(\int f_2(x_2) p(x_1, x_2) dx_1 dx_2 - F_2 \right) \right\}, \end{aligned}$$

$$\Phi\left(\frac{p(x_1, x_2)}{q_1(x_1) q_2(x_2)}\right) = \mu_0 + \mu_1 f_1(x_1) + \mu_2 f_2(x_2) \quad (10.36)$$

Impomos que a solução deve satisfazer $p(x_1, x_2) = p_1(x_1) p_2(x_2)$ e definimos $y = \frac{p_1(x_1) p_2(x_2)}{q_1(x_1) q_2(x_2)}$. Assim

$$\Phi(y) = \mu_0 + \mu_1 f_1(x_1) + \mu_2 f_2(x_2). \quad (10.37)$$

Derivamos duas vezes, primeiro com respeito a x_2

$$\ddot{\Phi}(y) \frac{\partial y}{\partial x_2} = (\mu_2 f_2(x_2))', \quad (10.38)$$

e depois com respeito a x_1

$$\ddot{\Phi}(y) \frac{\partial y}{\partial x_1} \frac{\partial y}{\partial x_2} + \dot{\Phi}(y) \frac{\partial^2 y}{\partial x_1 \partial x_2} = 0. \quad (10.39)$$

Isto parece complicado mas é bem simples. Note que, com a linha indicando a derivada com respeito a x_1 ou x_2

$$\begin{aligned} \frac{\partial y}{\partial x_1} &= \left(\frac{p_1(x_1)}{q_1(x_1)} \right)' \frac{p_2(x_2)}{q_2(x_2)} \\ \frac{\partial y}{\partial x_2} &= \frac{p_1(x_1)}{q_1(x_1)} \left(\frac{p_2(x_2)}{q_2(x_2)} \right)' \\ \frac{\partial^2 y}{\partial x_1 \partial x_2} &= \left(\frac{p_1(x_1)}{q_1(x_1)} \right)' \left(\frac{p_2(x_2)}{q_2(x_2)} \right)'. \end{aligned} \quad (10.40)$$

Multiplique as duas primeiras e divida pela terceira equação, segue que $\Phi(y)$ satisfaz a equação diferencial ordinária:

$$y \ddot{\Phi}(y) + \dot{\Phi}(y) = 0. \quad (10.41)$$

A solução geral contém três constantes arbitrárias

$$\Phi(y) = Ay \log y + By + C. \quad (10.42)$$

A família de funcionais entropia é reduzida a

$$S[p||q] = A \int p(x) \log \frac{p(x)}{q(x)} dx + B \int p(x) dx + C \int q(x) dx.$$

A constante C pode ser tomada igual a zero, pois simplesmente adiciona uma constante sem mudar preferências. A constante B também pode ser tomada zero, pois a normalização de $p(x)$ será sempre um vínculo. Finalmente e demonstrando otimismo os físicos tomam $A < 0$ negativo para que o processo variacional seja o de maximização. Podemos tomar $A > 0$ e minimizar o que é chamado de divergência de Kullback-Leibler. A escolha de $A = -1$ é simplesmente uma escolha de unidades da entropia. Como veremos isto levará a que a temperatura e a energia sejam medidos nas mesmas unidades. O resultado final é dado por

$$S[p||q] = - \int p(x) \log \frac{p(x)}{q(x)} dx \quad (10.43)$$

É fácil mostrar que os resultados $p_1(x_1)p_2(x_2)$ usando a equação 10.35 e $p(x_1, x_2)$ da equação 10.36, coincidem.

Exercício Para variáveis discretas mostre que $H_{SJ} = - \sum_i p_i \log p_i$ difere por uma constante de $S[p||q] = - \sum_i p_i \log \frac{p_i}{q_i}$, se $q_i = 1/N$ é

constante e portanto o valor das distribuições que as extremizam coincidem.

Isto permite concluir que a entropia de Shannon faz referência a uma distribuição *a priori*, a distribuição uniforme. O fato de não aparentar de forma explícita não quer dizer que não esteja lá. É claro que o ponto onde a distribuição uniforme faz a sua entrada no raciocínio é quando a função $F(n)$ crescente é introduzida.

Exercício Suponha $\mathcal{F}$ uma família paramétrica de densidades, e dois membros $p_1 = p(x|\theta_1)$ e $p_2 = p(x|\theta_2)$ de $\mathcal{F}$. Suponha que os parâmetros θ estejam em algum subconjunto de $\mathbb{R}^K$. Mostre que em geral $S[p_1||p_2] \neq S[p_2||p_1]$, mas para valores pequenos da distância Euclideana, $|\theta_1 - \theta_2|$ há simetria de S ante $\theta_1 \leftrightarrow \theta_2$. Isso permite definir uma distância entre duas densidades e uma geometria Riemanniana no espaço dos θ , que agora passa a ser mais que um simples subconjunto de $\mathbb{R}^K$, mas o que é conhecido como uma variedade estatística. Considere $\theta_2 = \theta_1 + d\theta$ e expanda a entropia até segunda ordem em $d\theta$. Dessa forma é possível definir o tensor métrico $g_{\mu\nu}$. Encontre uma expressão para o tensor métrico que é conhecido pelo nome de Fisher-Rao. Encontre a distância entre duas gaussianas multivariada com a mesma covariância e médias diferentes.

Apêndice: Multiplicadores de Lagrange

Uma montanha é descrita pela altura $z = f(x, y)$, com x e y as coordenadas de um ponto de altura z e f um função suficientemente bem comportada, tem um máximo que pode ser obtido igualando as derivadas parciais de f a zero. Suponha que uma estrada cruza a superfície da montanha e as coordenadas da estrada são descritas por $\phi(x, y) = c$. Qual é a altura máxima de um carro que viaja pela estrada?

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^*) caísse 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 :

$$F_\lambda(x, y) = f(x, y) + \lambda(\phi(x, y) - c) \quad (10.44)$$

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 ⁶. Resolvemos o sistema

⁶ a não ser que não seja....

$$\frac{\partial F_\lambda}{\partial x} = 0 \quad (10.45)$$

$$\frac{\partial F_\lambda}{\partial y} = 0, \quad (10.46)$$

onde λ ainda não foi especificado. A resposta depende do 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. Substituímos 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 quando se satisfaz o vínculo.

Agora procure um livro de cálculo e preencha os detalhes. Discuta também como lidar com casos em que o extremo está na borda de C . Há vínculos que são representados por desigualdades, os nomes de Kuhn e Tucker estão associados a esta extensão. Em muitos casos isto pode útil mas não no curso introdutório.