在看到一篇 2008 年预测变分推断和机器学习交叉领域趋势的论文后，我就知道不需要在这方面过于深入了。但仍需借对此的研究补充数学知识。

本文直接参考[1].

目录

根据 Intro 看，VI 用于估计贝叶斯学派提出的后验分布，优势是又快又不差。后验分布实际上是一个综述性质的迭代算法，对复杂问题很难计算。

从信息理论开始：什么是熵?

了解熵公式的由来为止。

报文的信息量

说实话之前了解 RNN for NLP 的时候已经 overview 了一下，但我发觉现代科普性的文章只会把东西变得更加云里雾里，所以还是从原始论文读起。

首要就是 Shanon 的 [2]. (其 intro 中提到的 Hartley 的 [3] 也可一读，只是晦涩些)

[2] 和 [3] 最开始提出的一个结论就是对于以某种特定编码编写的一段报文实例，在度量它的信息量时应该使用 $\log_s s^N = N$ 而不是 s^N (N 代表此报文实例所用的字符数，s 是报文所用编码的字符表大小)

这几乎就只是报文长度。Shannon 和 Hartley 都给出了很多理由——但都没有决定性的说服力。除了营销号的牵强解释和一些智者的公理化分析之外，并没有什么非此两者还非常自然非常严谨的 idea，我们只能看在它辉煌实绩的面子上勉强接受。

离散的信道容量

存在性/可计算性

定义式几乎只是在描述 bits/s，紧接着的一些存在性证明，我是说，证明是否对于任意编码都能计算出离散时间的信道容量，在我看来主要是帮你复习代数。(注意下面的 t 都是非负整数，因为这一章是离散语境)

隐含的假设是 $N(t)$ 的解析解为指数形式 $N(t) = CX^t$（注意这里 X 也是常数。虽然我们知道基于离散编码本身的性质，这并不是假设。而且我们还知道 $C = 0$），于是 $N(t)$ 的式子可以写成 $CX^t = CX^{t-t_1} + CX^{t-t_2} + \ldots + CX^{t-t_n}$，两边同除 CX^t 就是文中给出的特征方程。

根据我们证明存在性的目的，读者可能比较关注特征方程的解集形式，正好原文给了一个例子：

```
-- in lua shell
> 2^0.539
1.4529650495714
> mu = 2^0.539
> mu^-2 + mu^-4 + mu^-5 + mu^-7 + mu^-8 + mu^-10
0.99983286103003
```

好吧这个例子只是过过眼睛，直接分析：

0. 首先排除 simple case，$n = 1$ 时必然有 $X = 1$，存在非负解。
1. 考虑 $Z = X^{-1}$，构造函数 $g(Z) = Z^{t_1} + Z^{t_2} + \ldots + Z^{t_n} - 1$，目标是找 $g(x)$ 的非负零点
2. 由于 $n > 1$，有 $g(0) = -1, g(1) \geqslant 1$，而 $g(Z)$ 是个分析上性质非常好的函数，故根据介值定理，非负零点在 $(0,1)$ 之间

所以直接验证了上面给的数值例子，X 一定有解且值大于一。

自动机约束下的存在性/可计算性

这里的约束由确定性有限自动机（DFA）描述，而且一定要注意这里的 $\|$ 不是行列式！作者只是把自动机图的每条单向转移边用矩阵形式展示而已。

关于产生答案的过程，作者没有多解释，让我来补全：

实际上我们只是在计算时间 t 的可能性，在没有自动机约束时这种可能性可以直接写成 $N(t)$，但我们现在有约束，就要区分时间 t 时我们在自动机的哪个节点上，故可以写出：

$$N_1(t) + N_2(t) + \ldots = \text{一堆东西}$$

由于有上一节的基础，剩下的事情就很显然了。

PART I 的 2,3,4,5 小节

是可以跳过的

作者想要拓展基于字母出现的频率为其分配莫尔斯编码，以减少整体信道占用的想法。

于是提出不能仅仅看信源的统计特征，还要用随机过程来为信源建模。

当然这里读者就不要深究细节了，这部分是香农那篇文章难得的过时内容，现代的 NLP 研究已经向人们展示——大语言模型这样规模的东西才能为语言建模。

至于那些有关 ergodic 的随机过程的东西，你下一次在这篇文章里看到是在 PART III: MATHEMATICAL PRELIMINARIES

哈哈，Man，我还能说什么呢。

任意离散分布的熵

熵在文中最初的想法是用于度量一个 **任意随机过程** 的不确定性，或者也可以说是信息量吧，并量化为实数（众所周知，量化后就好办了，实数数值的优化算法研究有很多）。

任意随机过程有多复杂？读者可能还没意识到这个问题。回顾我们对编码信息量的讨论，时间步 t 时可能的编码数量为 s^t，再举一个类似生成函数的例子，也就是类似 $(a+b+c)(d+e+f)(g+h+i)$ 这样的随机过程，它们实际上都可以被类似乘法的东西描述。

而后面三条要求的第三条就展示了一种“不可乘”的随机过程。

继续，说那三条。

对 $H(X)$ 提出的三条要求如今看来比较显然：

1. 要求是连续函数。这个太基本了，事实的微小误差的确可以导致概率分布的突变，但概率分布的微小误差一定不能导致 $H(X)$ 的突变，$H(X)$ 应该是稳定而平滑的。**这里的 idea 是：两个看起来很像的概率分布实际上差距就是很小。**
2. 假设了一种不确定性最高的场景。实际上我们也可以轻易扩展这个要求：不论 n 多大，只要有 $p_i=1$ 且 $p_j=0 \mid j\in[0,n]\wedge j\neq i$，$H(X)$ 应取得最低的数值。**这里只是划定了上下界，实际没有太多解析意义，后续的公式推导也只是把这个当作入手点**
3. 给出了一个 $H(\frac{1}{2},\frac{1}{3},\frac{1}{6})=H(\frac{1}{2},\frac{1}{2})+\frac{1}{2}H(\frac{2}{3},\frac{1}{3})+\frac{1}{2}H(1,0,0,0,...)$，由于任意随机过程很可能不是“可乘的随机过程”，所以这里并没有给出类似 $H(X_{[0,1]})=H\left(X_{[0,\frac{1}{2}]}X_{[\frac{1}{2},1]}\right)=H\left(X_{[0,\frac{1}{2}]}\right)+H\left(X_{[\frac{1}{2},1]}\right)$ 的东西。**但 main idea 是不变的，这里还是想要采用类似 $\log$ 函数的东西，让决策分支带来的 $H(X)$ 累计在数值上体现为相加而不是相乘**

作者把公式的推理放到了附录二（在第 28,29 页左右），所以我们直接翻过去看。

首先这里 $A(t^n)=nA(t)$ 之类的结果是不用疑虑的，符合上面第三条。（作者第一次看疑虑了一瞬间，但验算后发现正确）

总能找到一个 m 使得 $s^m\leqslant t^n<s^{m+1}$，也是个比较显然的事实。（提示点在于，这里为什么不用 $\leqslant s^{m+1}$ 呢）

后续的数值分析虽然过程容易理解，但萌新真的没了解过这方面哈哈，总之结果还是很惊艳的 $|\frac{A(t)}{A(s)}-\frac{\log t}{\log s}|<2\varepsilon$

如果说 $A(n)$ 的解析满足了条件 2，那么接下来针对条件 3 的构造则更令人惊艳。

首先展示一下更容易理解的公式（我验算过了，读者放心用）：$A(\sum n_i)=H(p_1,...,p_n)+\sum p_iA(n_i)$，它在描述一个条件三中提到过的那种两个时间步的随机过程，要么说这是个天才的构造呢。读者按照条件三推导一下就自然能理解。

关于香农其它在熵方面的工作

文中额外涉及的就是联合熵和条件熵，以及连续分布的熵了。剩下的都是通信相关成果。

KL 散度

至于 KL 散度，这里有一些定性的简短说明，以思维实验的形式给出。

whynot

一个想法是对于这样一个度量 $0\leqslant\sum_{x\in X}|p(x)-q(x)|\leqslant 2$，令 $\sum_{x\in X}|p(x)-q(x)|=2$ 时达到 D 的上界，换句话说此时 $|p(x)-q(x)|=\max(p(x),q(x))$，且 $\min(p(x),q(x))=0$

而这个度量的问题是，有个绝对值，不好对公式操作。这个绝对值甚至是必须的，因为计算 $\sum_{x\in X}p(x)-q(x)$ 只会得到零。

或许这里我们可以回过头来看看 KL 散度的公式：$\sum_{x\in p(X)} p(x)\log\left(\frac{p(x)}{q(x)}\right)$，它在两个分布完全一样时给出 0 的散度，而在两个分布完全不同时（此时 $\sum_{x\in X}|p(x)-q(x)|=2$）不可避免地要涉及 $\log(0)=-\infty$，且在此时散度总是给出 $+\infty$

可以将此处使用 $\log\left(\frac{p(x)}{q(x)}\right)$ 而不是 $\frac{p(x)}{q(x)}$ 的原因解释为 $\log\left(\frac{p(x)}{q(x)}\right)$ 可以分解为 $\log p(x)-\log q(x)$ 以避免 $\frac{p(x)}{0}$ 的情况出现，毕竟你可以选择把 $\log 0$ 解释成 $\lim_{x\to 0}\log(x)$ 的同时坚决反对 $\frac{1}{0}$ 这样情况的存在，但这属于代数系统的 corner case，并不能真正说服我们。

why?

查看 KL 散度第一次被提出的论文[4]，那时候就已经有测度论了……实际上作者直接给出了 $\log\left(\frac{p(x)}{q(x)}\right)$，并未说明来源。

但至少我们知道了不应该从 cross-entropy 入手。

没有变化量（$\Delta=Q-P$）的熵这回事，所以从熵的变化量的角度看一下 P 和 Q 两个不同的分布：

$$\delta_1+\delta_2+\ldots+\delta_n=0$$

$$q_i=p_i+\delta_i$$

$$\text{Q 的熵}=\sum_i(p(x_i)+\delta_i)\log(p(x_i)+\delta_i)=\sum_i p(x_i)\log(p(x_i)+\delta_i)+\delta_i\log(p(x_i)+\delta_i)$$

那么就有熵的变化量 $H(Q)-H(P)$ ：

$$\sum_i p(x_i)\log\left(\frac{p(x_i)+\delta_i}{p(x_i)}\right)+\delta_i\log(p(x_i)+\delta_i)$$

虽然这玩意实际上只是把 $\sum Q\log Q-P\log Q+P\log Q-P\log P=\sum Q\log Q-P\log P$ 以更拗口的方式表示了一下，但也让我们的目光转向了 $P\log Q$ 和 $Q\log P$ 这两者。

实际上，现有的约束不足以将这个熵的差分进行定量分析，但回到我们原本的目的，**找出衡量两种分布区别的方法**，或许不需要完整分析这整个公式，而是分成两部分：

$$\sum Q\log Q-P\log Q=\sum(Q-P)\log Q$$

$$\sum P\log Q-P\log P=\sum P\log\frac{Q}{P}$$

实际上这两种均有其意义，后者是 KL 散度的形式，前者虽没有标准名称，但问了 LLM，它编了一些若有若无的东西出来。总之，对 KL 散度的追溯就告一段落。

Forward vs. Reverse KL

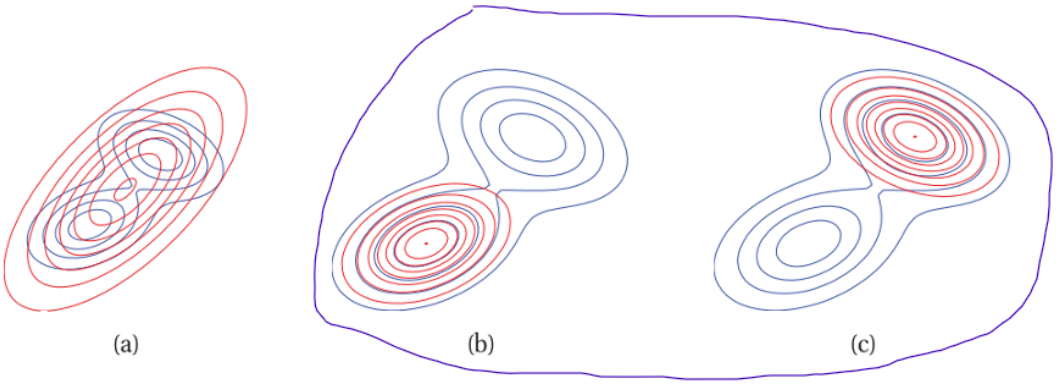

Figure 2: Figure illustrating forward vs reverse KL-divergence on a bimodal distribution. The blue and the red contours epresent the actual probability density, p, and the unimodal approximation, q, respectively. The left panel shows he forward KL-divergence minimization where q tends to *cover* p. The centre and the right panels show the reverse KL-divergence minimization where q locks on to one of the two modes.

图 1 前向 KL 散度(a) vs 反向 KL 散度(b,c)

可以看到由于前向 KL 散度把 P（原始分布）有而 Q（近似分布）无的情况惩罚值设的太大，所以了现出了一种倾向于变成原分布超集的行为，后向类似。

由于限制了近似分布 Q 为单峰，才能为我们展示这两种 KL 散度的“缺陷”。

本质上是 KL 散度不好避免 $\lim_{n\to 0}\log n=+\infty$ 这种东西的出现。

ELBO

Evidence Lower BOund，ELBO，实际上代表了一种简化 KL 散度优化器的方法。

这里提前补充一些知识，KL 散度优化器实际上要做的是求一个近似分布 $Q(Z)$，要求最小化 $\mathcal{D}_{KL}(Q(Z)\|P(Z|X))$，这里的 $P(Z\|X)$ 是后验分布，也就是贝叶斯学派的迭代算法收敛出来的那个分布。

但这个 KL 散度优化器根本没有意义，毕竟都算出后验分布了就不需要一个近似的分布 Q 了

继续补充知识，贝叶斯迭代算法长这样：

$$\text{what I wanted is } P(Z|X)$$
$$\forall z\in Z$$
$$p(z|X)=\frac{p(X|z)}{p(X)}p(z)$$

这里的迭代初始给定一个先验 $p(Z)$，然后根据公式算出后验 $p(Z|X)$，每次迭代将上一步的后验 $p(Z\mid X)$ 作为这一次的先验 $p(Z)$，算出新的后验。

这种迭代预计会让 $P(Z|X)$ 收敛，只要在 X 中加入新数据就可以继续迭代。

要理解它为什么生效，重点是 $\frac{p(X|Z)}{p(X)}$。这其中 $P(X|Z)$ 是似然关于 Z 的分布很好理解，而这里的 $p(X)=\int\limits_{z\in Z}p(X|z)p(z)\mathrm{d}z$，学名叫**边缘似然**，其实就是 $\mathbb{E}_{Z\sim P(Z)}\left[P(X|Z)\right]$，它边缘掉的是 Z 维度。

所以 $\frac{p(X|Z)}{p(X)}$ 实际上是在迭代中激励那些似然超出平均值的 z，很好理解。

补充完知识，继续说 ELBO，它源于一个等式 $\mathcal{D}_{KL}(Q(Z)\,\|\,P(Z|X)) = \log p(x) - \mathrm{ELBO}(Q)$，这里的 $\log p(x)$ 就是边缘似然，所以这里面的数值关系就很显然了：

$$\begin{gathered}\log p(x) = \mathcal{D}_{KL}(Q(Z)\,\|\,P(Z|X)) + \mathrm{ELBO}(Q)\\ \mathcal{D}_{KL}(Q(Z)\,\|\,P(Z|X)) \geqslant 0\\ \log p(x) \geqslant \mathrm{ELBO}(Q)\end{gathered}$$

所以为什么叫 Lower Bound，下界，是因为我们基于上面的数值关系，可以**将最小化 KL 散度的行为转化为最大化 ELBO**，而 ELBO 正好是边缘似然的下界。

接下来好好看看 ELBO(Q) 的公式：

$$\begin{aligned}\mathrm{ELBO}(Q) &= \mathbb{E}_{z\sim Q}\big[\log p(x|z)p(z) - \log q(z)\big]\\ &= \mathbb{E}_{z\sim Q}\big[\log p(x|z)\big] - \mathcal{D}_{KL}(Q(Z)\,\|\,P(Z))\end{aligned}$$

这里又多出一个散度，但这个散度是可以在迭代过程中顺便计算的，ELBO 的意义就在于此。

散度项前面是交叉边缘对数似然（我自己起的名字）。

假设我们迭代进行最大化 ELBO 的过程，负的散度项让 Q 不要与先验偏离太远，而交叉边缘对数似然鼓励 Q 的分布更加关注 $P(X|Z)$ 中高似然的部分，所以这会是一个健全的迭代算法。

VAE

D. P. Kingma and M. Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

参考文献

[1] A. Ganguly 和 S. W. F. Earp, 《An Introduction to Variational Inference》. [在线]. 载于: https://arxiv.org/abs/2108.13083

[2] C. E. Shannon, 《A mathematical theory of communication》, *The Bell System Technical Journal*, 卷 27, 期 3, 页 379–423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[3] R. V. L. Hartley, 《Transmission of information》, *Bell System Technical Journal*, 卷 7, 页 535–563, 1928, [在线]. 载于: https://api.semanticscholar.org/CorpusID:109872947

[4] S. Kullback 和 R. A. Leibler, 《On Information and Sufficiency》, *Annals of Mathematical Statistics*, 卷 22, 页 79–86, 1951, [在线]. 载于: https://api.semanticscholar.org/CorpusID:120349231