Shannon Entropy

參考資料：
http://blog.xuite.net/metafun/life/69851478-資訊的度量-+Information+Entropy

要了解 entropy 之前，我們先要知道資訊(Information)是甚麼，甚麼樣的訊息我們會視為資訊呢？我告訴你一天有 24 小時你會比較震驚，還是今年林書豪拿了 NBA 的 MVP 你會比較震驚，想當然爾一定是林書豪今年拿了 MVP，比較震驚代表甚麼事？代表這件事你獲得了比較多的資訊，一天有 24 小時你早就知道了，我告訴你等於沒有講，也代表這件事情發生的機率已經太高了，再發生一次，你也不會感到驚訝，不會認為你還有東西不知道，但是林書豪拿了 MVP，這機率實在太低了，如果你獲知了這件事，代表你現在認知的林書豪絕對和現在的拿 MVP 的林書豪有天壤之別，一定還有很多你不知道的事情，所以獲得這訊息你得到的資訊是非常巨大的，所以我們可以下一個小結論，**當你獲得了一個越無法預期的事情，換句話說，也就是發生機率越低，你獲得的訊息(Information)越多。**

Information 的單位(bit)

針對上面的結論，我們知道一件事情發生機率越小，獲得資訊越多，但我們要怎麼量化這件事情呢？我們使用這樣的公式

$-\log P(x)$

我們來慢慢感受這個公式想告訴我們甚麼，這個公式輸出的單位是 bit，可以這樣想，我們要問多少個 yes/no 的問題，才能得知到最後的答案，就好像是一個八格輪盤，結果已經出來了，但我們要問多少問題才能知道結果，假設這八格的編號分別為 1~8，結果為 3，我們可以先問是不是在 1~4 之間，再問是否再 1~2 之間，再問是否為 3，最後得知結果，所以我們共問了三個問題，這個問題也的確是最多需要問三個問題才能得到答案](log$_2$8=3)，所以若我們直接知道這個輪盤結果是 3，我們直接獲得 3 bits 的訊息量，總結一下，我們獲得一個訊息，背後的訊息量越大，bit 數就越大。

//如果還是不太懂 bits 的概念(因為有同學看完後說看不懂再來問我，所以我想可能是我的例子舉的不夠好，如果懂的話就不用看下一段了

　　我們應該都有看電影，小說被雷的經驗吧！為何我們不喜歡有人在我們還沒開始看電影、小說的時候直接跟我們說結局，因為如果我們知道結局，再看電影、小說，我們幾乎沒辦法得到我們想要得到的新資訊。再繼續看就沒有新鮮感，感覺就不如甚麼都不知道的時候看電影、小說的刺激感。因為我們獲得的資訊量就幾乎等同於是我們從頭到尾看整部電影、小說的資訊量，套回之前的例子，直接告訴你八格輪盤的結果，和你要重複問三次 yes/no 的問題是一樣的資訊量。

Expectation value of entropy

但是單一事件的資訊量大，就代表這個事件很重要嗎？顯然不是，像是上面 MVP 的例子，雖然我知道這項消息會很震驚，但是這種事情根本不可能發生，那這個事件根本就不重要。我們想要知道的是對整個宇集間取得一事件的平均資訊量為何。

e.g.

有一個骰子六面的點數為 1, 2, 3, 4, 5, 1000，發生機率為 $\sim\frac{1}{5},\sim\frac{1}{5},\sim\frac{1}{5},\sim\frac{1}{5},\sim\frac{1}{5},\frac{1}{10000000}$

則擲一次骰子平均可得到 $\frac{1}{5}\times(1+2+3+4+5)+1000\times\frac{1}{10000000}\sim 3$

就算 1000 那面點數比其他面還要大很多，若發生機率很小，對於期望值而言也幾乎是沒貢獻的

所以 Shannon 提出了資訊理論中 entropy 的概念，其實也就是 information 的期望值

$$H(X)=-\sum_{i=1}^{n}P(x_i)\log_2 P(x_i)$$

這個公式又被稱為熵，我們簡單跑幾組數據，我們更改骰子每個點數的機率看看

$$H(\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6})=-(\frac{1}{6}\log\frac{1}{6}+\frac{1}{6}\log\frac{1}{6}+\frac{1}{6}\log\frac{1}{6}+\frac{1}{6}\log\frac{1}{6}+\frac{1}{6}\log\frac{1}{6}+\frac{1}{6}\log\frac{1}{6})\text{=2.58}$$

$$H(\frac{2}{7},\frac{1}{7},\frac{1}{7},\frac{1}{7},\frac{1}{7},\frac{1}{7})=-(\frac{2}{7}\log\frac{2}{7}+\frac{1}{7}\log\frac{1}{7}+\frac{1}{7}\log\frac{1}{7}+\frac{1}{7}\log\frac{1}{7}+\frac{1}{7}\log\frac{1}{7}+\frac{1}{7}\log\frac{1}{7})\text{=2.52}$$

$$H(\frac{1}{21},\frac{2}{21},\frac{3}{21},\frac{4}{21},\frac{5}{21},\frac{6}{21})=-(\frac{1}{21}\log\frac{1}{21}+\frac{2}{21}\log\frac{2}{21}+\frac{3}{21}\log\frac{3}{21}+\frac{4}{21}\log\frac{4}{21}+\frac{5}{21}\log\frac{5}{21}+\frac{6}{21}\log\frac{6}{21})\text{=2.39}$$

$$H(\frac{1}{15},\frac{1}{15},\frac{1}{15},\frac{1}{15},\frac{1}{15},\frac{10}{15})=-(\frac{1}{15}\log\frac{1}{15}+\frac{1}{15}\log\frac{1}{15}+\frac{1}{15}\log\frac{1}{15}+\frac{1}{15}\log\frac{1}{15}+\frac{1}{15}\log\frac{1}{15}+\frac{10}{15}\log\frac{10}{15})\text{=1.69}$$

在這裡不做證明，等一下再證明，但我們可以感覺的到若 sample space 中的 outcome 機率分布的越平均，得到的 entropy 越大，也代表訊息越無法被預測，代表資料越亂，若我們得到一個越亂的資料，訊息含量越多！

e.g.

1. 一個公平的骰子，我們得知擲了一次的結果，此訊息有 $\text{-}\log_2^{\frac{1}{6}}=\log_2^{6}$ bits 的訊息量

若用公式顯示，則為

$$\text{-}(\frac{1}{6}\log_2^{\frac{1}{6}}+\frac{1}{6}\log_2^{\frac{1}{6}}+\frac{1}{6}\log_2^{\frac{1}{6}}+\frac{1}{6}\log_2^{\frac{1}{6}}+\frac{1}{6}\log_2^{\frac{1}{6}}+\frac{1}{6}\log_2^{\frac{1}{6}})=\log_2^{6}$$

2. 當我們得知了一個 32 格的輪盤的結果，此訊息有 $\text{-}\log_2^{\frac{1}{32}}=5$ bits 的訊息量

$$-(\frac{1}{32}\log_2^{\frac{1}{32}}\times 32)=5$$

3. 但是例如說 NBA 有 32 隊，理論上來說，我們知道冠軍球隊是哪隊的資訊量應該是 5 bits，但是有些球隊是奪冠大熱門，有些球隊是重建中球隊，奪冠機率不會是平均分布的，此時”奪冠隊伍”的預期資訊量就會不如 32 格輪盤

以現在 2017/9，美國權威 ESBN 預測明年的總冠軍的機率如下

隊伍	奪冠機率
金州勇士	58%
波士頓賽爾提克	12%
休士頓火箭	7%
聖安東尼奧馬刺	7%
克里夫蘭騎士	6%

資料來源:
http://www.appledaily.com.tw/realtimenews/article/local/20170921/1207911/ESPN 預測騎士奪冠機率將出現毀滅性崩壞

上述總共 90%，假設剩餘 27 隊平均分配奪冠機率，每隊共有 10/27=0.37%的奪冠機率
總冠軍這件事的預期資訊量為

$$-(0.58\log_2^{0.58}+0.12\log_2^{0.12}+0.07\log_2^{0.07}+0.07\log_2^{0.07}+0.06\log_2^{0.06}+0.0037\log_2^{0.0037}\times 27)=2.41$$ bits < 5

故**我們可以暫時獲得一個結論，若一個 outcome 發生是完全隨機的，假設有 n 種可能發生的 outcome，那麼每個 outcome 的機率為 1/n，這樣的情形會獲得最大的資訊量期望值。**

證明 Expectation value of entropy

話都你在說，根本就還沒有證明最大的資訊量期望值是落在各個 outcome 機率皆相同的情況下(也就是 uniform distribution)，那就來證明囉！

我們希望得到最大值的方程式為

$$H(X)=-\int_a^b P(x)\log_2 P(x)dx$$

(用連續函數來證，離散沒有好的數學性質，可將連續函數想成無限多個離散訊號就好)

而且我們還有一個限制是

$$\int_a^b P(x)dx=1$$

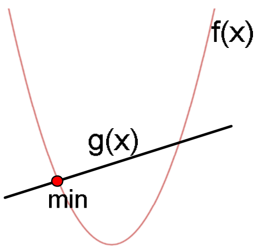
//這裡用 a,b 就只是代表定義域區間而已，像是銅板正面機率的 a,b 就是 0,1

骰子就是 1,6

而我們要使用 lagrange multiplier，使用時機就和現在一樣，我們希望最大化某一個函數，還附有些限制，lagrange multiplier 就是用來解決這些問題的。

Langrage multiplier

概念：



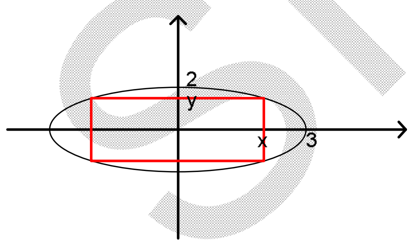
圖中 f(x)為欲找到最小值的函數，g(x)為限制函數，要求取的 x 點必須在 g(x)上，該紅點就是最小值

求解步驟：
步驟 1：將限制條件 g 及函數 f 取 gradiant
步驟 2：$\nabla f-\lambda\nabla g=0$ ，求解

note:
當要最大化時，步驟 2 為 $\nabla f+\lambda\nabla g=0$
當要最小化時，步驟 2 為 $\nabla f-\lambda\nabla g=0$

e.g.

求在$x>0, y>0$、$\frac{x^2}{9}+\frac{y^2}{4}=1$的限制條件下, $4xy$最大值時之x, y



紅色面積為 4xy

$g(x,y)=\frac{x^2}{9}+\frac{y^2}{16}, f(x,y)=4xy$

$L(x,y)=f(x,y)-\lambda g(x,y)=4xy-\lambda(\frac{x^2}{9}+\frac{y^2}{16})$

∇L中(如果不懂gradient的定義，請自己去搞清楚

$$\begin{cases}\frac{\partial \mathrm{L}}{\partial x}=4y-\lambda\frac{2x}{9}=0\\ \frac{\partial \mathrm{L}}{\partial y}=4x-\lambda\frac{y}{8}=0\end{cases}$$

$\Rightarrow y=\pm 2\sqrt{2}$ (取正)

$$\Rightarrow\begin{cases}x=\frac{3}{\sqrt{2}}\\ y=2\sqrt{2}\end{cases}$$

那就開始來解最大的 entropy 囉！

沒有給任何 prior

最大化函數為$H(X)=-\int_a^b P(x)\log_2 P(x)dx$

限制條件為$\int_a^b P(x)dx=1$

故$lagrange\ multiplier$為$-\int_a^b P(x)\log_2 P(x)dx+\lambda(\int_a^b P(x)dx-1)$

這裡我們因為有變數$P(x_i)$，我們沒辦法做單純的梯度，所以必須用到變分法，簡單來說，就是一種將函數當成變數求極值的方法

以下是變分法的簡述，如果沒有興趣的人可以跳過直接看結論

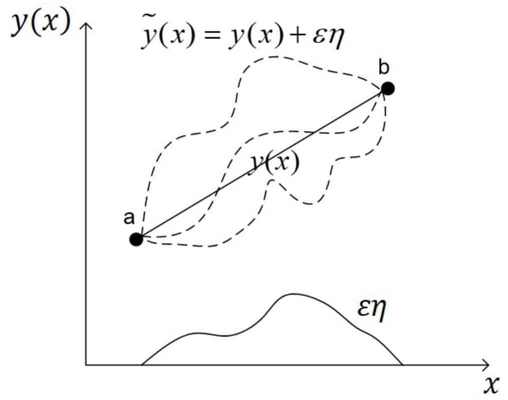
參考網站：https://zhuanlan.zhihu.com/p/20718489

- 泛函數

以前我們學過的函式是輸入為實數，輸出為實數，而泛函數為輸入為一個函式，輸出為實數，我們先將一些符號定義好，以便之後的討論
通常變分法處理的泛函數都是積分的形式，我們記為

$J=\int_a^b F(y,y';x)dx$

$F(y,y';x)$ 中的 $y,y';x$ 代表這裡的 y 和 y' 都是 x 的函數，這裡我們僅討論一階微分的泛函數



- 泛函數的極值

在邊界已訂(a,b)的情況下，我們可以產生無限多種從 a 到 b 的函數(就像上圖的實線和虛線，我們假設最佳(極值發生時)的函數為 $y(x)$，而不確定極值是否發生時的函數稱為 $\tilde{y}(x)$，我們要試圖找出 $\tilde{y}(x)$ 的極值 $y(x)$，而 $\tilde{y}(x) = y(x) + \varepsilon\eta(x)$，$\varepsilon\eta(x)$ 為 $\tilde{y}(x)$ 和 $y(x)$ 的差距，不是一個定值，當 $\tilde{y}(x)$ 不同，$\varepsilon\eta(x)$ 就會不同。而若我們給定 $\eta(x)$ (為了將邊界的差距設為 0，即 $\eta(a) = 0, \eta(b) = 0$)，我們可以藉由 ε 來產生無限多個 $\tilde{y}(x)$，$\eta(x)$ 可以想成差距的外型，ε 可以想成差距的外型固定後，對這個外型放大 ε 倍，而注意，當 ε=0 時，$\tilde{y}(x) = y(x)$。

泛函數 J 是泛函數 $\tilde{J}$ 的極值，我們可以寫出

$$\tilde{J} = \int_a^b F(y(x) + \varepsilon\eta(x), y'(x) + \varepsilon\eta'(x); x)dx$$

注意 $\tilde{J}$ 是一個 ε 的函數，因為在定值積分下，x 會消失成為實際的值，我們隨便找出一個函數 $\tilde{y}(x)$，此時我們將泛函數 $\tilde{J}$ 對 $\varepsilon = 0$ 做泰勒展開式，得

$$\begin{aligned}\tilde{J} &= \tilde{J}(\varepsilon = 0) + \frac{d\tilde{J}(\varepsilon = 0)}{d\varepsilon}(\varepsilon - 0) + \frac{1}{2!}\frac{d^2\tilde{J}(\varepsilon = 0)}{d\varepsilon^2}(\varepsilon - 0)^2 + \ldots \\ &= \tilde{J}_0 + \tilde{J}_1\varepsilon + \tilde{J}_2\varepsilon^2 + \ldots\end{aligned}$$

而因為 $\tilde{J}(\varepsilon = 0) = J$，可以改寫上式為

$$\tilde{J} - J = \tilde{J}_1\varepsilon + \tilde{J}_2\varepsilon^2 + \ldots$$

而一階變分定義為

$$\delta J = \tilde{J}_1 \varepsilon = \frac{d\tilde{J}(\varepsilon=0)}{d\varepsilon}\varepsilon$$

二階變分以此類推，而在一階變分=0 時會達到該”函式的函式”的極值，而若再加上$\varepsilon = 0$是”最佳函數的極值，故

$$\begin{aligned}
0 = \delta J &= \frac{d\tilde{J}(\varepsilon=0)}{d\varepsilon}\varepsilon \\
&= \frac{d}{d\varepsilon}\int_a^b \tilde{F}(\varepsilon=0)dx \cdot \varepsilon \\
&= \int_a^b \frac{d\tilde{F}(\varepsilon=0)}{d\varepsilon}dx \cdot \varepsilon \\
&= \int_a^b \frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}}\frac{d\tilde{y}}{d\varepsilon} + \frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}'}\frac{d\tilde{y}'}{d\varepsilon}dx \cdot \varepsilon \\
&= \int_a^b \frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}}\varepsilon\eta(x) + \frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}'}\varepsilon\eta'(x)dx \quad (\because y = \tilde{y} + \varepsilon\eta(x)) \\
&= \int_a^b \frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}}\delta y + \frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}'}\delta y' dx \quad (\because \delta y = y - \tilde{y} = \varepsilon\eta(x))
\end{aligned}$$

對第二項做分部積分，得

$$= \int_a^b \frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}}\delta y - \left(\frac{d}{dx}\frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}'}\right)\delta y dx + \frac{d\tilde{F}(\varepsilon=0)}{d\tilde{y}'}\delta y\Big|_a^b$$

當$\varepsilon = 0$時，$\tilde{y} = y, \tilde{F} = F$ 故

$$= \int_a^b \left(\frac{dF}{dy} - \frac{d}{dx}\frac{dF}{dy'}\right)\delta y dx + \frac{dF}{dy'}\delta y\Big|_a^b$$

第三項$\frac{dF}{dy'}\delta y\Big|_a^b = 0$是邊界條件，由於$\delta y$ 可能是任意值，故若需要保證$\delta J = 0$ ，須

$$\frac{dF}{dy} - \frac{d}{dx}\frac{dF}{dy'} = 0$$

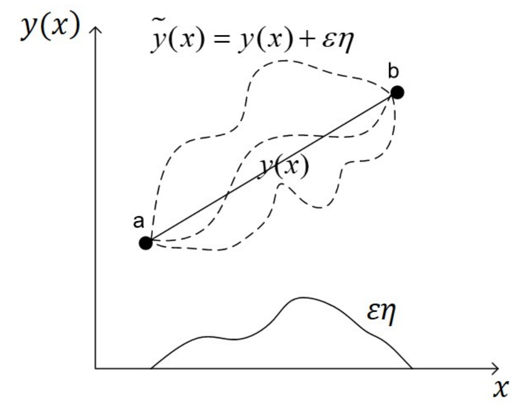
上式是經典的 Euler-Lagrange 方程式

而邊界條件部分，若$\frac{dF}{dy'}$在 a 及 b 上皆為 0，此為 Natural Boundary Condition，若δy 在 a 和 b 上皆為 0，也就是在 a、b 上 y 皆為確切的值，稱為 Essential Boundary condition。

- 例題

e.g.

我們可以藉由變分法證明兩點的最近路線為直線



我們隨意假設一條路徑 $\tilde{y}(x)$，其路徑上的一小段令為 ds，

$$(ds)^2 = (dx)^2 + (dy)^2 = (1 + y'^2)(dx)^2$$
$$\Rightarrow ds = \sqrt{1 + y'^2}\, dx$$

則該路徑長為

$$\int_a^b ds = \int_a^b \sqrt{1 + y'^2}\, dx$$

對比 Euler-Lagrange 方程式，$F = \sqrt{1 + y'^2}$，且 $y(a), y(b)$ 皆為固定值(起點終點固定)，符合 Essential Boundary condition

代入 Euler-Lagrange 方程式中為

$$0 = \frac{dF}{dy} - \frac{d}{dx}\frac{dF}{dy'} = 0 - \frac{d}{dx}\frac{y'}{\sqrt{1 + y'^2}} = -\frac{d}{dx}\frac{y'}{\sqrt{1 + y'^2}}$$
$$\Rightarrow \frac{y'}{\sqrt{1 + y'^2}} = C$$
$$\Rightarrow y' = \frac{C^2}{1 - C^2}$$

故斜率從頭到尾皆為定值，為一個直線

總之，結論就是如果符合邊界條件 $\frac{dF}{dy'}$ 在 a 及 b 上皆為 0，或在 a、b 上 y 皆為確切的值，則可以使

用 Euler-Lagrange 方程式$\frac{dF}{dy}-\frac{d}{dx}\frac{dF}{dy'}=0$得到極值

那開始來解最大的 entropy 的 P(x)了！原式為

$$-\int_a^b P(x)\log_2 P(x)dx+\lambda(\int_a^b P(x)dx-1)$$

我們整理一下上式成為

$$J=-\int_a^b P(x)\log_2 P(x)-\lambda P(x)dx-\lambda=-\int_a^b F(1,P,\frac{dP}{dx})dx-\lambda$$

會影響J中$P(x)$的最大值的只有$P(x)$，λ並不會影響，所以我們先省略最後一項

$$J=-\int_a^b P(x)\log_2 P(x)+\lambda P(x)dx=-\int_a^b F(1,P,P')dx$$

而因為 F 中沒有P'，則$\left.\frac{dF}{dy'}\right|_{x=a}=\left.\frac{dF}{dy'}\right|_{x=b}=0$，符合邊界條件

而根據 Euler-Lagrange 方程式

$$\begin{aligned}
0&=\frac{dF}{dy}-\frac{d}{dx}\frac{dF}{dy'}\\
&=\frac{dF}{dP(x)}-\frac{d}{dx}\frac{dF}{dP'(x)}\\
&=-\log P(x)-1+\lambda-0\\
\Rightarrow P(x)&=e^{\lambda-1}
\end{aligned}$$

而因為$\int_a^b P(x)dx=1$

$$\Rightarrow \int_a^b e^{\lambda-1}dx=(a-b)e^{\lambda-1}=1\Rightarrow P(x)=e^{\lambda-1}=\frac{1}{a-b}$$

故跟我們之前討論的一樣，假設有 a-b 個 outcome，發生最大 entropy 的 P(x)為$\frac{1}{a-b}$

給定 mean

通常我們得到許多數據時，也可以得到這些數據的 mean，如果我將限制條件再加上一個

$\int_0^\infty xP(x)dx=\mu$ //0 和∞都只是為了結果好看而已，沒有甚麼為什麼

最大的 entropy 的 P(x)分布會變成甚麼樣子呢？

統整一下問題：

最大化$-\int_a^b P(x_i)\log_2 P(x_i)dx$

限制條件：

$$\int_0^\infty P(x)dx = 1$$

$$\int_0^\infty xP(x)dx = \mu$$

我們用剛剛用過的 Lagrange multiplier

$$-\int_0^\infty P(x)\log_2 P(x)dx + \lambda_0(\int_0^\infty P(x)dx - 1) + \lambda_1(\int_0^\infty xP(x)dx - \mu)$$
$$= \int_0^\infty -P(x)\log_2 P(x) + \lambda_0 P(x) + \lambda_1 xP(x)dx - \lambda_0 - \lambda_1\mu$$

使用變分法，一樣的因為 F 中沒有P'，則$\left.\frac{dF}{dy'}\right|_{x=a} = \left.\frac{dF}{dy'}\right|_{x=b} = 0$，符合邊界條件，可以使用-Euler-Lagrange 方程式

$$F = -P(x)\log_2 P(x) + \lambda_0 P(x) + \lambda_1 xP(x)$$
$$0 = \frac{dF}{dy} - \frac{d}{dx}\frac{dF}{dy'}$$
$$= \frac{dF}{dP(x)} - \frac{d}{dx}\frac{dF}{dP'(x)}$$
$$= -\log P(x) - 1 + \lambda_0 + \lambda_1 x$$
$$\Rightarrow \log P(x) = \lambda_0 + \lambda_1 x - 1$$
$$\Rightarrow P(x) = e^{\lambda_0 + \lambda_1 x - 1}$$

代入限制條件中

第一個限制條件($\int_0^\infty P(x)dx = 1$)

$$\int_0^\infty P(x)dx = \int_0^\infty e^{\lambda_0 + \lambda_1 x - 1}dx = e^{\lambda_0 - 1}\int_0^\infty e^{\lambda_1 x}dx = \frac{e^{\lambda_0 - 1}}{\lambda_1} e^{\lambda_1 x}\Big|_0^\infty = \frac{e^{\lambda_0 - 1}}{\lambda_1}(e^{\lambda_1 \infty} - 1) = 1$$

若$\lambda_1 > 0, \because \lambda_0, \lambda_1 \in R$　此等式必不成立($\infty = 1$)

$$\therefore \lambda_1 <= 0, e^{\lambda_1 \infty} = 0 \Rightarrow \frac{e^{\lambda_0 - 1}}{\lambda_1}(e^{\lambda_1 \infty} - 1) = -\frac{e^{\lambda_0 - 1}}{\lambda_1} = 1 \Rightarrow \lambda_1 = -e^{\lambda_0 - 1}$$

第二個限制條件($\int_0^\infty xP(x)dx = \mu$)

$$\int_0^\infty xP(x)dx = \int_0^\infty xe^{\lambda_0 + \lambda_1 x - 1}dx = e^{\lambda_0 - 1}\int_0^\infty xe^{\lambda_1 x}dx$$

需要使用分部積分

微	積
x	$e^{\lambda_1 x}$
1	$\frac{1}{\lambda_1}e^{\lambda_1 x}$

$$= e^{\lambda_0-1}\frac{x}{\lambda_1}e^{\lambda_1 x}\Big|_0^{\infty} - e^{\lambda_0-1}\int_0^{\infty}\frac{1}{\lambda_1}e^{\lambda_1 x}dx = (0-0) - \frac{1}{\lambda_1}\int_0^{\infty}e^{\lambda_0+\lambda_1 x-1}dx = -\frac{1}{\lambda_1} = \mu \Rightarrow \lambda_1 = \frac{-1}{\mu}$$

$$\because \lambda_1 = -e^{\lambda_0-1}, \lambda_1 = \frac{-1}{\mu} \Rightarrow P(x){=}e^{\lambda_0+\lambda_1 x-1} = -\lambda_1 e^{\lambda_1 x} = \frac{1}{\mu}e^{\frac{-1}{\mu}x}$$

故若我們給予 mean，entropy 最大期望值的 P(x)分部為一 exponential 函數

給定 mean 和 variance

如果我們又得到 variance 呢？
就不做推導了，之後會用另一個概念來推導，其結果會是一個 Gauss distribution，所以只要我們給予 mean 和 variance 其最大 entropy 的 P(x)分布為 Gauss distribution

統整

給予條件(prior)	最大 entropy 的 P(x)分佈
無	uniform distribution($\frac{1}{a-b}$)
$\int_a^b xP(x)dx = \mu$	$\frac{1}{\mu}e^{\frac{-1}{\mu}x}$
$\int_0^{\infty} xP(x)dx = \mu$ $\int_0^{\infty} x^2P(x)dx = \sigma^2$	Gauss distribution

Conditional entropy

我們已經知道 H(X)的意義，也就是得知一個 X 中一個 x 元素所需要資訊的位元數期望值，而 H(X,Y)就是我們得到(x,y)所需要資訊的位元數期望值，就像是我們擲一枚硬幣和擲一個骰子後得到的訊息量。那麼若我們已經有了 X 的一個元素的位元期望值的資訊，我們還需要再得到多少位元才足夠得

到(x,y)的資訊量呢？這和條件機率的意思很像，就將這樣的概念表示為 H(Y|X)

$$
\begin{aligned}
H(Y|X) &= -\sum_i P(x_i)H(Y|X=x_i) \\
&= -\sum_i P(x_i)\sum_j P(y_j|x_i)\log(y_j|x_i) \\
&= -\sum_i\sum_j P(x_i)P(y_j|x_i)\log(y_j|x_i) \\
&= -\sum_i\sum_j P(x_i,y_j)\log(y_j|x_i) = -\sum_i\sum_j P(x_i,y_j)\log P(y_j|x_i) \\
&= -\sum_i\sum_j P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)} = \sum_i\sum_j P(x_i,y_j)\log\frac{P(x_i)}{P(x_i,y_j)}
\end{aligned}
$$

relative entropy (KL divergence)

參考資料：
http://kuanchen-blog.logdown.com/posts/333763#content8
https://zh.wikipedia.org/wiki/相对熵

先讓我們介紹 cross entropy，再來介紹 relative entropy，假設有兩個 sample space: P,Q，P 和 Q 各自擁有其 entropy 最佳的 P(x)分佈，cross entropy 就是將其他人的 P(x)分佈套用在自己身上，看看這樣的資訊量有多少 bits，我們將這樣概念以數學形式定義

$$H_p(q) = \sum_i p(x_i)\log(\frac{1}{q(x_i)})$$

上面式子代表 Q 用 P 的 P(x)分佈計算 Q 的資訊 bits 期望值。

當然，若別人的分佈和自己的最佳分佈不同，其得出的 entropy bits 應該較小，不過，有趣的是，$H_p(q)$和$H_q(p)$有時並不會相等，換句話說，P 用 Q 的最佳分布和 Q 用 P 的最佳分布所得到的資訊量是不一樣的。

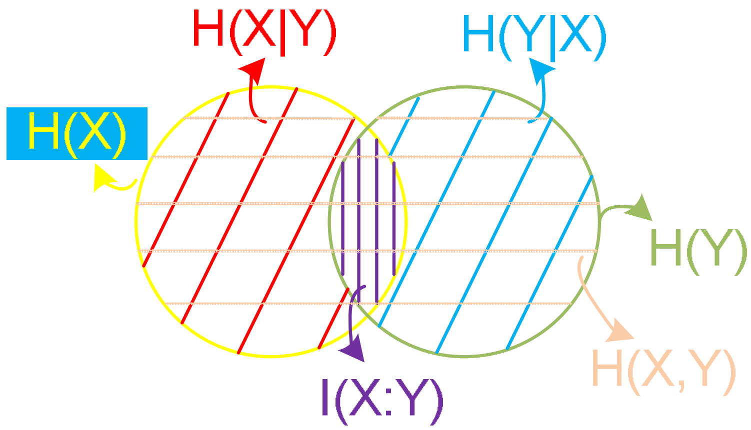
而通常我們有興趣的是$H_p(q)$和$H(p)$的差，這樣的差就稱為 relative entropy，也代表兩個分布的距離，數學定義如下

$$KL(p\|q) = H_p(q) - H(p) = \sum_i p(x_i)\log(\frac{1}{q(x_i)}) - \sum_i p(x_i)\log(\frac{1}{p(x_i)}) = \sum_i p(x_i)\log(\frac{p(x_i)}{q(x_i)})$$

但在這裡會有個大問題，KL divergence，代表兩個分布的距離，p 到 q 的距離居然和 q 到 q 的距離是不一樣的。

Mutual information(互資訊)

可想成是兩個資訊中有重疊的部分，形式為 I(X:Y)



$$I(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$