4.6 Different Flavors of Entropy

Prior knowledge I: $q(x) := \mathcal{P}(x|I)$

Updating information J: $d = \langle f(x) \rangle_{(x|J,I)} = \int dx\, f(x)\, P(x|J,\, I)$

Posterior knowledge: $p(x) := \mathcal{P}(x|J,\, I)$

- normalization: $\lambda \left(\int dx\, p(x) - 1 \right)$
- new information: $\mu \left(\int dx\, p(x) f(x) - d \right)$

$$\mathcal{S}[p|q,\, J] = \underbrace{\underbrace{-\left\{ \int dx\, p(x) \left[\ln \left(\frac{p(x)}{q(x)} \right) \right.\right.}_{\mathcal{S}[p|q]} \left.\left. - \lambda - \mu f(x) \right] \right\}}_{\mathcal{S}^*[p|q,J]} - \lambda - \mu d$$

constrained entropy:	$\mathcal{S}[p|q,\, J]$	maximized w.r.t p, λ, μ
amount of relative information:	$\mathcal{S}[p|q]$	of p w.r.t q in nits
auxiliary entropy:	$\mathcal{S}^*[p|q,\, J]$	maximized w.r.t p
		sloped by $\partial S^*/\partial\lambda \stackrel{!}{=} 1,\ \partial S^*/\partial\mu \stackrel{!}{=} d$

4.7 Information Gain by Maximizing the Entropy

Relative negative information gain: $\mathcal{S}[p|q] = -\int dx\, p(x) \ln\left(\frac{p(x)}{q(x)}\right)$

$$\begin{aligned}
\mathcal{S}[p|q] &= - \underbrace{\int dx\, \frac{q(x)e^{\mu f(x)}}{\mathcal{Z}(\mu)}}_{=1} \ln\left(\frac{e^{\mu f(x)}}{\mathcal{Z}(\mu)}\right) \\
& \qquad \int dx\, q(x) e^{\mu f(x)} = \mathcal{Z}(\mu) \\
&= - \left[\int dx\, \frac{q(x)e^{\mu f(x)}}{\mathcal{Z}(\mu)} \left(\mu f(x) - \ln \mathcal{Z}(\mu)\right)\right] \\
&= \ln \mathcal{Z}(\mu) - \mu \underbrace{\langle f(x)\rangle_{(x|J)}}_{=d} \\
&= \ln \mathcal{Z}(\mu) - \mu\, d
\end{aligned}$$

4.7 Information Gain by Maximizing the Entropy

Auxiliary entropy:

$$
\begin{aligned}
\mathcal{S}^*[p|q, J] &= \mathcal{S}[p|q] + \lambda + \mu \underbrace{\langle f(x) \rangle_{(x|J)}}_{=d} \\
&= \mathcal{S}[p|q] + \lambda + \mu d \\
&= \ln \mathcal{Z}(\mu) - \mu d + \lambda + \mu d \\
&= \ln \mathcal{Z}(\mu) + \lambda \\
Z(\mu) &= e^{1-\lambda} \\
\lambda &= 1 - \ln \mathcal{Z}(\mu) \\
\Rightarrow \mathcal{S}^*[p|q, J] &= 1
\end{aligned}
$$

4.7 Information Gain by Maximizing the Entropy

Constrained entropy:

$$\begin{aligned}\mathcal{S}[p|q, J] &= \mathcal{S}^*[p|q, J] - \lambda - \mu d \\ &= 1 - \lambda - \mu d \\ &= \ln \mathcal{Z}(\mu) - \mu d\end{aligned}$$

$\Rightarrow$ Information change at maximum of $\mathcal{S}[p|q, J]$:

$$\mathcal{S}[p|q, J] = \mathcal{S}[p|q] = \ln \mathcal{Z}(\mu) - \mu\, d$$

Maximizing the Constrained Entropy

Normalization: *Maximize w.r.t* λ

$$\frac{\partial \mathcal{S}[p|q, J]}{\partial \lambda} \overset{!}{=} 0$$

New information: *Maximize w.r.t* μ

$$\frac{\partial \mathcal{S}[p|q, J]}{\partial \mu} = \frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} - d \overset{!}{=} 0$$

$$\Rightarrow \frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} = d, \text{ Helmholtz free energy } \ln \mathcal{Z}(\mu)$$

Several constraints:

$$d = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}, f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \quad \Rightarrow \quad d^\dagger \mu := \sum_{i=1}^{n} \overline{d_i} \mu_i$$

Maximum Entropy Recipe

Maximum Entropy Recipe:

$q(x) = P(x|I)$, $J =$ "$\langle f(x) \rangle_{(x|J,I)} = d$", $p(x) = P(x|J, I) = ?$

1. calculate the partition sum: $\mathcal{Z}(\mu) = \int dx\, q(x) e^{\mu f(x)}$
2. determine μ: $\frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} \stackrel{!}{=} d$
3. assign p: $p(x) = \frac{q(x)\, e^{\mu^\dagger f(x)}}{\mathcal{Z}(\mu)}$
4. calculate the information gain: $\Delta \mathcal{I}[p|q] = -\mathcal{S}[p|q] = \mu\, d - \ln \mathcal{Z}(\mu)$

4.7.1 Coin Tossing Example

Prior information I: $x \in \{0, 1\}$
Prior knowledge: $q(x) := \mathcal{P}(x|I) = \frac{1}{2}$
Updating information J: $f = \langle x \rangle_{(x|J,I)}$
Posterior knowledge: $p(x) := \mathcal{P}(x|J, I) = ?$

1. calculate $\mathcal{Z}(\mu)$:

$$\mathcal{Z}(\mu) = \sum_{x \in \{0,1\}} q(x) e^{\mu x} = \frac{1}{2}(1 + e^{\mu})$$

2. determine μ:

$$\frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} = \frac{e^{\mu}}{1 + e^{\mu}} \stackrel{!}{=} f \quad \Rightarrow e^{\mu} = \frac{f}{1-f} \quad \Rightarrow \mu = \ln\left(\frac{f}{1-f}\right)$$

Insert in $\mathcal{Z}(\mu)$:

$$\mathcal{Z}(\mu) = \frac{1}{2}\left(1 + \frac{f}{1-f}\right) = \frac{1}{2(1-f)}$$

4.7.1 Coin Tossing Example

3. calculate $p(x) = P(x|J, I)$**:** $q(x) = 1/2$, $\mathcal{Z}(\mu) = \frac{1}{2(1-f)}$, $\mu = \ln\left(\frac{f}{1-f}\right)$

$$p(x) = \frac{q(x)\, e^{\mu x}}{\mathcal{Z}(\mu)} = \frac{1/2\, e^{x \ln(f/(1-f))}}{\frac{1}{2(1-f)}}$$

$$= (1-f)\left(\frac{f}{1-f}\right)^x = f^x\,(1-f)^{1-x} = P(x|f)$$

4. calculate the information gain $\Delta\mathcal{I}$**:**

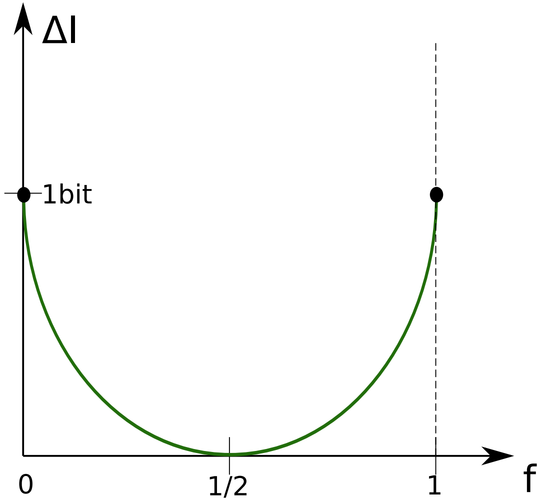
$$\Delta\mathcal{I}[p|q] = -\mathcal{S}[p|q] = \mu f - \ln \mathcal{Z}(\mu)$$

$$= f \ln\left(\frac{f}{1-f}\right) - \ln\left(\frac{1}{2(1-f)}\right)$$

$$= [\ln 2 + f \ln f + (1-f) \ln(1-f)] \text{ nits}$$

$$= [1 + f \log_2 f + (1-f) \log_2(1-f)] \text{ bits}$$

4.7.1 Coin Tossing Example



4.7.2 Positive Counts Example

Prior information I: $n \in \mathbb{N}$
Prior knowledge: $q(n) := \mathcal{P}(n|I) = \text{const.} \ = q$
Updating information J: $\langle n \rangle = \lambda$
Posterior knowledge: $p(n) := \mathcal{P}(n|J, I) = ?$

1. calculate $\mathcal{Z}(\mu)$:

$$\mathcal{Z}(\mu) = \sum_{n=0}^{\infty} q\, e^{\mu n} = q \sum_{n=0}^{\infty} [e^{\mu}]^n = \frac{q}{1 - e^{\mu}}$$

2. determine μ:

$$\frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} = \frac{\partial}{\partial \mu}[\ln q - \ln(1 - e^{\mu})] = -\frac{1}{1 - e^{\mu}}(-e^{\mu}) = \frac{e^{\mu}}{1 - e^{\mu}} \stackrel{!}{=} \lambda$$

$$\Rightarrow e^{\mu} = \frac{\lambda}{1 + \lambda}$$

4.7.2 Positive Counts Example

Insert in $\mathcal{Z}(\mu)$: $q(n) = q$, $\mathcal{Z}(\mu) = \frac{q}{1-e^{\mu}}$, $e^{\mu} = \frac{\lambda}{1+\lambda}$

$$\mathcal{Z}(\mu) = \frac{q}{1 - \frac{\lambda}{1+\lambda}} = (1+\lambda)\, q$$

3. calculate $p(n) = \mathcal{P}(n|J, I)$**:**

$$p(n) = \frac{q(n) e^{\mu\, n}}{\mathcal{Z}(\mu)} = \frac{q \cdot \left(\frac{\lambda}{1+\lambda}\right)^n}{q \cdot (1+\lambda)}$$

$$= \frac{1}{(1+\lambda)} \underbrace{\left(\frac{\lambda}{1+\lambda}\right)^n}_{=:\left(e^{-\mu'}\right)^n} = \frac{1}{(1+\lambda)} e^{-\mu'\, n}$$

$$\mu' = \ln(1+\lambda) - \ln\lambda = -\mu > 0$$

4.7.2 Positive Counts Example

Check of compliance with constraints:

- Normalization:

$$\sum_{n=0}^{\infty} p(n) = \sum_{n=0}^{\infty} \frac{1}{1+\lambda}\left(\frac{\lambda}{\lambda+1}\right)^n = \frac{1}{1+\lambda} \cdot \frac{1}{1-\frac{\lambda}{1+\lambda}} = \frac{1}{1+\lambda} \cdot (1+\lambda) = 1$$

- Average counts:

$$\sum_{n=0}^{\infty} n\, p(n) = \sum_{n=0}^{\infty} n \frac{1}{1+\lambda} \underbrace{\left(\frac{\lambda}{\lambda+1}\right)^n}_{=:y^n} = \frac{1}{1+\lambda} \sum_{n=0}^{\infty} y\, \partial_y y^n$$

$$= \frac{1}{1+\lambda} y\, \partial_y \sum_{n=0}^{\infty} y^n = \frac{y}{1+\lambda} \partial_y \frac{1}{1-y}$$

$$= \frac{y}{(1+\lambda)(1-y)^2} = \frac{\lambda}{(1+\lambda)^2 (1+\lambda)^{-2}} = \lambda \; \square$$

4.7.3 Many Small Count Processes

Prior information I: $N = \#$ of processes, $n = \sum_{i=1}^{N} n_i$ =total # of counts, $n_i \in \mathbb{N}$
Updating information J: $\langle n_i \rangle_{(n_i,J)} = \delta$ for all i, $\langle n \rangle_{(n,J)} = \lambda = \delta\, N$

Recap positive count examples:

$$\mathcal{P}(n_i|\delta = \langle n_i \rangle) = \frac{1}{1+\delta} \left(\frac{\delta}{1+\delta} \right)^{n_i}$$

Posterior calculation via marganalization over independent processes:

$$\mathcal{P}(n|\underbrace{\lambda,\, N}_{I'=J,I}) \quad = \quad \underbrace{\sum_{n_1=0}^{\infty} \ldots \sum_{n_N=0}^{\infty}}_{=:\sum_{\vec{n}=\vec{0}}^{\vec{\infty}}} \mathcal{P}(n,\, n_1,\, n_2\,, \ldots n_N|\lambda,\, N)$$

4.7.3 Many Small Count Processes

$$\mathcal{P}(n|I') = \sum_{\vec{n}=\vec{0}}^{\vec{\infty}} \mathcal{P}(n|n_1,\, n_2, \ldots n_N,\, I')\mathcal{P}(n_1|I')\mathcal{P}(n_2|I')\ldots\mathcal{P}(n_N|I')$$

$$= \sum_{\vec{n}=\vec{0}}^{\vec{\infty}} \delta_{n,\, \sum_{i=0}^{N} n_i} \frac{1}{1+\delta}\left(\frac{\delta}{\delta+1}\right)^{n_1} \cdots \frac{1}{1+\delta}\left(\frac{\delta}{\delta+1}\right)^{n_N}$$

$$= \left(\frac{1}{1+\delta}\right)^N \sum_{\vec{n}=\vec{0}}^{\vec{\infty}} \delta_{n,\, \sum_{i=0}^{N} n_i} \left(\frac{\delta}{1+\delta}\right)^{\sum_{i=0}^{N} n_i}$$

- $\binom{n+N-1}{n} = \frac{(n+N-1)!}{n!(N-1)!}$ possibilities to distribute n counts on N processes, as there are $N-1$ process separators to be but between the n counts

$$\mathcal{P}(n|I') = \left(\frac{1}{1+\delta}\right)^N \frac{(n+N-1)!}{n!(N-1)!} \left(\frac{\delta}{1+\delta}\right)^n$$

4.7.3 Many Small Count Processes

$$\begin{aligned}
\mathcal{P}(n|I') &= \left(\frac{1}{1+\delta}\right)^N \frac{(n+N-1)!}{n!(N-1)!}\left(\frac{\delta}{1+\delta}\right)^n \\
&= \frac{\delta^n}{(1+\delta)^{N+n}} \cdot \frac{(n+N-1)!}{n!(N-1)!} = \frac{(\lambda/N)^n}{(1+\lambda/N)^{N+n}} \cdot \frac{(n+N-1)!}{n!(N-1)!} \\
&= \frac{\lambda^n}{n!} \cdot \left(1+\frac{\lambda}{N}\right)^{-N} \left(1+\frac{\lambda}{N}\right)^{-n} \frac{(n+N-1)!}{(N-1)!N^n}
\end{aligned}$$

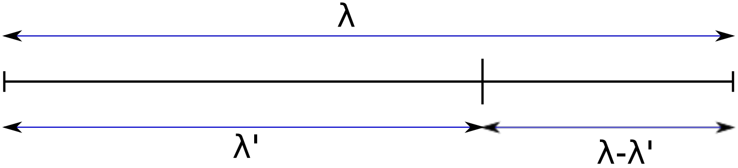
Infinite number of processes: $N \to \infty$, $\lambda =$ fixed, $\delta = \lambda/N \to 0$

$$\Rightarrow \quad \mathcal{P}(n|\lambda, N \to \infty) = \frac{\lambda^n}{n!} \underbrace{\left(1+\frac{\lambda}{N}\right)^{-N}}_{\to e^{-\lambda}} \underbrace{\left(1+\frac{\lambda}{N}\right)^{-n}}_{\to 1} \underbrace{\frac{(n+N-1)!}{(N-1)!N^n}}_{\to 1} = \frac{\lambda^n e^{-\lambda}}{n!}$$

The Poisson Distribution

$$\mathcal{P}(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

- Poisson distribution is divisibel:



$$\mathcal{P}(n|\lambda) = \sum_{m=0}^{n} \mathcal{P}(m|\lambda')\,\mathcal{P}(n-m|\lambda-\lambda')$$

The Poisson Distribution

- If $\mathcal{P}(m|\lambda')$ and $\mathcal{P}(n-m|\lambda-\lambda')$ are Poisson distributions
  $\Rightarrow \mathcal{P}(n|\lambda)$ is Poisson distribution

Proof:

$$\begin{aligned}
\sum_{m=o}^{n} \mathcal{P}(m|\lambda')\mathcal{P}(n-m|\lambda-\lambda') &= \sum_{m=0}^{n} \frac{\lambda'^m e^{-\lambda'}}{m!} \frac{(\lambda-\lambda')^{n-m} e^{-\lambda+\lambda'}}{(n-m)!} \\
&= \frac{e^{-\lambda}}{n!} \sum_{m=0}^{n} \frac{n!}{m!(n-m)!} \lambda'^m (\lambda-\lambda')^{n-m} \\
&= \frac{e^{-\lambda}}{n!} (\lambda' + (\lambda-\lambda'))^n \\
&= \frac{e^{-\lambda}}{n!} \lambda^n = \mathcal{P}(n|\lambda) \square
\end{aligned}$$

End