4. Maximum Entropy

Coding a message *(repetition)*

Sender Alice: $p(s) = \mathcal{P}(s|A)$

Receiver Bob: $r(s) = \mathcal{P}(s|B) \xrightarrow{M} \mathcal{P}(s|B,M) = q(s)$ is determined by M

Requirements on action for optimal coding:

1. Locality
2. Properness
3. Coordinate invariance *(nice to have)*

$$\overset{1.\&2.}{\Rightarrow} \quad \text{minimize } \mathcal{S}(p,q) = -\int \mathrm{d}s\, p(s)\, \ln q(s) + \mathrm{const}(q) \text{ w.r.t. } q$$

$$\overset{3.}{\Rightarrow} \quad \text{minimize } D_{\mathrm{KL}}(p||q) = \mathcal{S}(p,q) - \mathcal{S}(p) = \int \mathrm{d}s\, p(s)\, \ln\left[p(s)/q(s)\right] \text{ w.r.t. } q$$

codes same message, as $\mathcal{S}(p) = \mathcal{S}(p,p) = \mathrm{const}(q)$

4.1 Decoding a Message

Decoding a message

Sender Alice: $p(s) = \mathcal{P}(s|A)$

Receiver Bob: $r(s) = \mathcal{P}(s|B) \xrightarrow{M} \mathcal{P}(s|B,M) = q(s)$ **to be** determined by M

Requirements on action for optimal decoding:

1. Locality
2. Coordinate independence of result
3. Separability
   Independent systems can be equally treated jointly as well as separately.

Maximum entropy principle *(to be proven)*
$\Rightarrow$ maximize entropy $\mathcal{S}[q|r] = -D_{\mathrm{KL}}(q||r) = -\int \mathrm{d}s\, q(s)\, \ln\left[q(s)/r(s)\right]$ w.r.t. q
$\Leftrightarrow$ minimize $D_{\mathrm{KL}}(q||r) = \int \mathrm{d}s\, q(s)\, \ln\left[q(s)/r(s)\right]$ w.r.t. q

Priniciple of Minimum Updating

Use of entropy:

- Optimal strategy for updating
- Set up probabilities
- General law to update information

Bayesian knowledge update of $M = \{d, \mathcal{P}(d|s, M)\}$:

$$\mathcal{P}(s|J) \quad \xrightarrow{M} \quad \mathcal{P}(s|J') = \mathcal{P}(s|J, M) = \frac{\mathcal{P}(d|s, M)\,\mathcal{P}(s|J)}{\mathcal{P}(d)}$$

Example: How does $\mathcal{P}(s|J, M)$ look like given $M = \{d, \langle f(s) \rangle_{(s|M)}\}$?

> **Principle of Minimum Updating (PMU):**
> Beliefs must be reviewed only to the extent required by the new information.

4.2 Maximum Entropy Principle

Entropy: *Measure for amount of information that forces a change in belief.*

Prior knowledge: $r(s) = \mathcal{P}(s|J)$
Posterior knowledge: $q(s) = \mathcal{P}(s|J')$

$$\begin{aligned} \text{relative entropy of } q \text{ w.r.t } r &= \mathcal{S}[q|r] \\ &= \text{negative information gain } r \to q \end{aligned}$$

Maximum Entropy Principle (MEP):
Updating from $r(s)$ to $q(s)$ given new information M should maximize the entropy $\mathcal{S}[q|r]$ under the constraints of M.

PMU + MEP define entropy as action principle to chose q. *(to be shown)*

1^{st} Design Criteria: Locality

Requirement on entropy:
if q_1 is to be preferred over q_2: $\mathcal{S}[q_1|r] > \mathcal{S}[q_2|r]$

Jaynes' criteria:
1. Locality: *Local Information has only local effects.*

If effect of M only on $\Omega' \subset \Omega = \{s\}$, then $\mathcal{P}(s|J, M,\, s \in \Omega \backslash \Omega') = \mathcal{P}(s|J,\, s \in \Omega \backslash \Omega')$

$\Rightarrow$ Non-overlapping domains of s have additive contibution to entropy

$$\mathcal{S}[q|r] \quad = \quad \int_\Omega ds\, F(q(s),\, r(s),\, s)$$

2^{nd} Design Criterium: Coordinate Invarinace

2. Coordinate invariance: *Chosen system of coordinates does not carry information.*

$m(s)$: some density function
$m'(t)$: transformed density function

$$\begin{aligned} m(s)\, ds &= m'(t)\, dt \\ m'(t) &= m(s(t)) \left| \frac{ds}{dt} \right| \end{aligned}$$

$$\Rightarrow \mathcal{S}[q|r] = \int ds\, m_1(s)\, F' \left(\frac{q(s)}{m_2(s)}, \frac{r(s)}{m_3(s)} \right)$$

If no new information ($\Omega = \Omega'$ and $M = \{\}$), we require $q = r$:

$$\Rightarrow S[q|r] = \int_\Omega ds\, q(s)\, F'' \left(\frac{q(s)}{r(s)} \right)$$

3rd Design Criteria: Separabillity of Independent Systems

3. Independence: *Independent systems can be equally treated jointly as well as separately.*
Two independent systems:

$$\begin{aligned} s &= (s_1, s_2) \\ r(s) &= r_1(s_1) r_2(s_2) \\ q(s) &= q_1(s_1) q_2(s_2) \end{aligned}$$

$$\begin{aligned} \Rightarrow \mathcal{S}[q|r] &= S[q_1|r_1] + S[q_2|r_2] \\ &= -\int ds\, q(s) \ln\left(\frac{q(s)}{r(s)}\right) \\ &= -D_{\mathrm{KL}}(\mathrm{q}||\mathrm{r}) \end{aligned}$$

Proof of Separability

Proof:

$$
\begin{aligned}
\mathcal{S}[q|r] &= -\int ds_1 \int ds_2\, q_1(s_1)q_2(s_2)\ln\left(\frac{q_1(s_1)q_2(s_2)}{r_1(s_1)r_2(s_2)}\right) \\
&= -\int ds_1 \int ds_2\, q_1(s_1)q_2(s_2)\left[\ln\left(\frac{q_1(s_1)}{r_1(s_1)}\right) + \ln\left(\frac{q_2(s_2)}{r_2(s_2)}\right)\right] \\
&= -\left[\int ds_1 q_1(s_1)\ln\left(\frac{q_1(s_1)}{r_1(s_1)}\right)\right]\cdot \underbrace{\int ds_2 q_2(s_2)}_{=1} \\
&\quad -\left[\int ds_2 q_2(s_2)\ln\left(\frac{q_2(s_2)}{r_2(s_2)}\right)\right]\cdot \underbrace{\int ds_1 q_1(s_1)}_{=1} \\
&= S[q_1|r_1] + S[q_2|r_2]\ \square
\end{aligned}
$$

4.3 Optimal Communication

Coding a message:

$$\text{Minimize} \quad D_{KL}(p||q) \quad \text{w.r.t. q}$$

Decoding a message:

$$\begin{aligned} \text{Minimize} \quad & D_{KL}(q||r) \quad \text{w.r.t. } q \\ \text{or Maximize} \quad & S[q|r] \quad \text{w.r.t. } q \end{aligned}$$

Optimal coding: *Choose message M that minimizes the expected surprise.*

$$M \quad = \underset{M}{\operatorname{argmin}} \operatorname{KL}(I, M) \quad = \underset{M}{\operatorname{argmin}} \left\langle \mathcal{H}(s|M) - \mathcal{H}(s|I) \right\rangle_{(s|I)}$$

4.3 Optimal Communication

Optimal decoding:
Choose posterior consistent with M that adds least amount of information.

$$
\begin{aligned}
q &= \operatorname*{argmin}_{q',\lambda} \left[D_{\mathrm{KL}}(q'||r) + \lambda M(q')\right] \\
J' &= \operatorname*{argmin}_{J'',\lambda} \left[\mathrm{KL}(J'', J) + \lambda M(\mathcal{P}(s|J''))\right]
\end{aligned}
$$

Optimal communication: *Optimal decoding of optimally coded message*

$$
\begin{aligned}
M &= \operatorname*{argmin}_{M} \mathrm{KL}(I, J'(J, M)) \\
&= \operatorname*{argmin}_{M} \mathrm{KL}\left(I, \operatorname*{argmin}_{J',\lambda} \left[\mathrm{KL}(J', J) + \lambda M(\mathcal{P}(s|J'))\right]\right)
\end{aligned}
$$

Real Communication

- sender emphasizes what she thinks is important
- ... communicates what she wants the receiver to believe
- ... makes wrong assumption about receiver's knowledge
- ... or ability to decode
- receiver distrusts sender
- ... makes wrong assumption about sender's intention
- ...

$\Rightarrow$ really, really complicated mess, but interesting psychology

Corrective strategies:

1. **Robust communication:** Send raw data instead of interpretation
2. **Question:** Request necessary, unambiguous information.
3. **Reputation system:** Remember & reward honest and informative communications

4.4 Maximum Entropy with Hard Data Constraints

Prior knowledge I: $q(d,s) := \mathcal{P}(d,s|I)$
Updating information J: $d = d^*$
Posterior knowledge: $p(d,\, s) := \mathcal{P}(d,\, s|I,\, J) = \delta_{d,d^*}\, \underbrace{\mathcal{P}(s|IJ)}_{=:p(s)}$

Constrained Entropy:

$$\begin{aligned}
\mathcal{S}^*[p|q] &= -\int ds \sum_d p(d,\, s) \left[\ln\left(\frac{p(d,\, s)}{q(d,\, s)}\right) - \lambda\right] \\
&= -\int ds\, p(s) \left[\ln\left(\frac{p(s)}{q(d^*,\, s)}\right) - \lambda\right] \\
0 \ln 0 &= \lim_{\epsilon \to 0} \epsilon \ln \epsilon = 0
\end{aligned}$$

Lagrange multiplier λ enforces normalization of $p(s)$: $\frac{\partial \mathcal{S}^*}{\partial \lambda} = \int ds\, \sum_d p(d,s) \stackrel{!}{=} 1$

Hard Data Constraints

$$\text{Entropy: } \mathcal{S}[p|q] = -\int ds\, p(s) \left[\ln\left(\frac{p(s)}{q(d^*,\, s)} \right) - \lambda \right]$$

$$\text{Maximizing: } \frac{\delta \mathcal{S}[p|q]}{\delta p(s')} = -\ln\left(\frac{p(s')}{q(d^*,\, s')} \right) + \lambda - \underbrace{\frac{p(s')}{p(s')}}_{=1} \stackrel{!}{=} 0 \Rightarrow p(s) = q(d^*,\, s) \cdot e^{\lambda - 1}$$

$$\text{Normalization: } \frac{\partial \mathcal{S}^*}{\partial \lambda} = \int ds\, p(s) = e^{\lambda - 1} \underbrace{\int ds\, q(d^*,\, s)}_{\mathcal{Z}(d^*)} \stackrel{!}{=} 1 \;\Rightarrow e^{\lambda - 1} = \frac{1}{\mathcal{Z}(d^*)}$$

$$P(s|I,\, J) = p(s) = \frac{q(d^*,\, s)}{\mathcal{Z}(d^*)} = \frac{P(d^*,\, s|I)}{\int ds\, P(d^*, s|I)} = P(s|d^*, I)$$

$\Rightarrow$Maximum entropy embraces and extends Bayes updating!

4.5 Maximum Entropy with Soft Data Constraints

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: $p(x) := \mathcal{P}(x|\underbrace{I,\, J}_{I'})$

Constrained Entropy:

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

Normalization:

$$\frac{\partial \mathcal{S}^*}{\partial \lambda} \;=\; \int dx\, p(x) = \langle 1 \rangle_{(x|I,\,J)} \overset{!}{=} 1$$

New information:

$$\frac{\partial \mathcal{S}^*}{\partial \mu} \;=\; \int dx\, p(x) f(x) = \langle f(x) \rangle_{(x|I,\,J)} \overset{!}{=} d$$

Soft Data Constraints

Maximizing the Entropy:

$$\begin{aligned}\mathcal{S}^*[p|q] &= -\int dx\, p(x)\left[\ln\left(\frac{p(x)}{q(x)}\right) - \lambda - \mu\, f(x)\right]\\ \frac{\delta\mathcal{S}^*}{\delta p(x)} &= -\ln\left(\frac{p(x)}{q(x)}\right) + \lambda + \mu\, f(x) - \frac{p(x)}{p(x)} \stackrel{!}{=} 0\end{aligned}$$

$$\boxed{\begin{aligned}\Rightarrow p(x) &= q(x)e^{\lambda-1}e^{\mu\, f(x)} = \frac{q(x)}{\mathcal{Z}(\mu)}e^{\mu\, f(x)}\\ \mathcal{Z}(\mu) &= \int dx\, q(x)e^{\mu\, f(x)}\end{aligned}}$$

$\Rightarrow$Choose Lagrange multiplier μ s.t.:

$$d \stackrel{!}{=} \langle f(x)\rangle_{(x|J)} = \frac{\int dx f(x)q(x)e^{\mu\, f(x)}}{\mathcal{Z}(\mu)} = \frac{1}{\mathcal{Z}(\mu)}\frac{\partial\mathcal{Z}(\mu)}{\partial\mu} = \frac{\partial \ln \mathcal{Z}(\mu)}{\partial\mu}$$