Principle of Maximum Entropy

i) Surprise (uncertainty): how surprised am I by what happened?

$H = -\ln(p) \Rightarrow H(s|I) = -\ln(P(s|I))$

Smaller $p \hookrightarrow$ more surprise.

Event at probability 0.5 happens $\hookrightarrow$ small surprise

event at probability 5×10^{-10} happens $\hookrightarrow$ ~~lot~~ much surprise!

Why ln? It makes surprise additive but also other reasons tied to additivity:

$H(s,d|I) = -\ln(P(s,d|I)) = -\ln(P(s|d,I)) - \ln(P(d|I))$

$= H(s|d,I) + H(d|I).$

Entropy: expected surprise $= -\sum_s P(s|I)\ln(P(s|I))$.

More entropy $\hookrightarrow$ more likely to be surprised.

ii) What is 1 bit?

Coin toss: $\begin{cases} \frac{1}{2} & H \\ \frac{1}{2} & T \end{cases}$ Say head occurs then surprise $= -\ln\left(\frac{1}{2}\right) = \ln(2)$ nits

$\frac{1 \text{ bits}}{\ln(2) \text{ nits}} \Rightarrow \ln(N) \text{ nits} = \frac{\ln(N)}{\ln(2)} \text{ bits} = \log_2(2) = 1 \text{ bit}.$

$= \log_2(N) \text{ bits}$

Expected surprise *before* learning outcome.

$$S[\{p_i\}] = -\sum_i p_i \ln(p_i) = \underbrace{-\tfrac{1}{2}\ln\left(\tfrac{1}{2}\right)}_{H} \underbrace{-\tfrac{1}{2}\ln\left(\tfrac{1}{2}\right)}_{T}$$

$$= \ln(2) = 1 \text{ bit}$$

H → $\frac{1}{2}$ prob, 1 bit

T → $\frac{1}{2}$ prob, 1 bit

Cross entropy: ~~both~~ How much does Alice expect Bob to be surprised.

$$S(A\|B) = -\int ds\, \underbrace{P(s|I_A)}_{\text{how likely Alice thinks } s \text{ is}} \underbrace{\ln\left(P(s|I_B)\right)}_{\text{how much she thinks Bob is surprised by this } s.}$$

Two cases: ~~Alice knows~~ Bob gives Alice a fair coin.

i) Bob gives Alice a fair coin. ~~Alice~~ Alice tosses coin and finds out "H".
What is ~~Bob's~~ Alice's expected surprise of Bob?

$$S(P_A\|P_B) = -1 \ln\left(\tfrac{1}{2}\right) - 0 \ln\left(\tfrac{1}{2}\right) = 1 \text{ bit.}$$

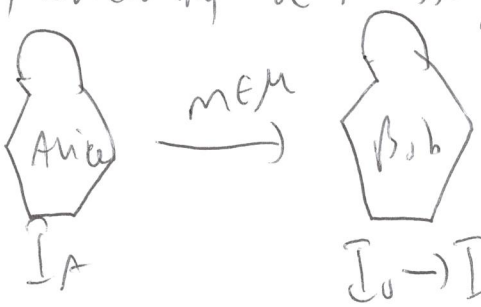
$P_{A,H}$ $P_{B,H}$ $P_{A,T}$ $P_{B,T}$

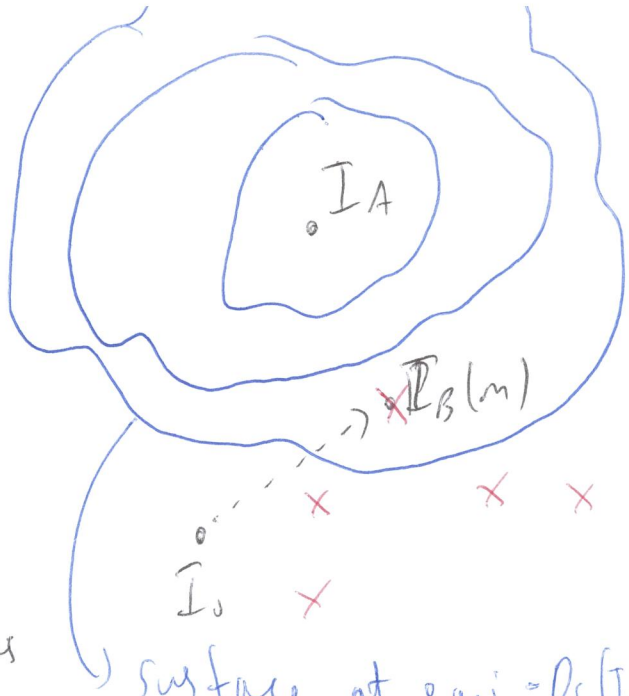
ii) Bob gives Alice a biased coin (both sides H).
~~Alice~~ Coin fairy replaces biased coin with fair one.
Alice tosses coin. What is Alice expected surprise at Bob?

$$S(P_A\|P_B) = -\tfrac{1}{2}\ln(1) - \tfrac{1}{2}\ln(0) = \infty \text{ bits!}$$

iii) Communication:

a) Encoding a message





surface of equi-$D_S(I_A, I_B)$ as a function of I_B. Not a metric space! Still surface can be drawn.

Alice choose $m \in M$ that maximises ~~for~~ the achieved information gain:

$$D_S(I_A, I_B(m), I_0) = D_S(I_A, I_0) - D_S(I_A, I_B(m)).$$

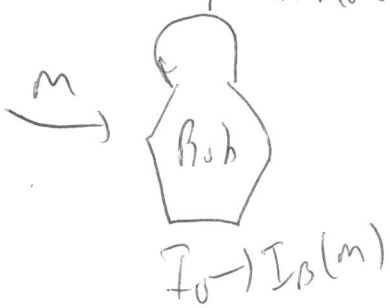
~~what~~ by

$$m = \underset{m \in M}{\operatorname{argmax}} \left[D_S(I_A, I_0) - D_S(I_A, I_B(m')) \right]$$

$$= \underset{m' \in M}{\operatorname{argmin}} \left(D_S(I_A, I_B(m')) \right).$$

As $D_S(I_A, I_0)$ does not depend on I_B, only second term matters here.

b) Decoding a message: No I_A! There is no longer an ideal probability distribution



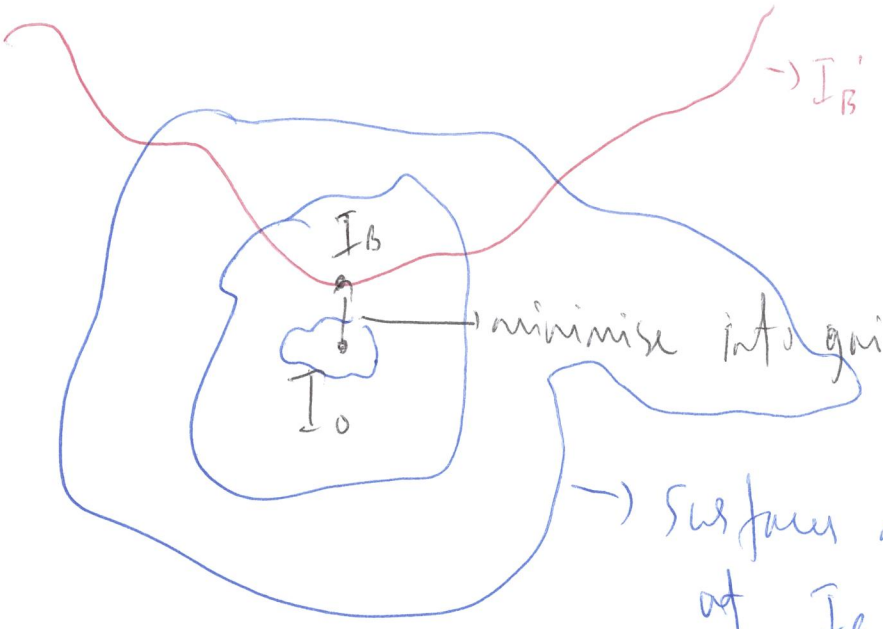
Bob assumes I_B is correct.

Minimise information gain from I_0 consistent with m — Max Entropy principle

$$I_B = \underset{I_B'}{\operatorname{argmin}} \left(D_S(I_B', I_0) + \lambda(m) \right)$$

Constraints from m as Lagrange multiplier terms

Minimise spurious information — Max Ent principle.



→ I_B' consistent with constraints m.

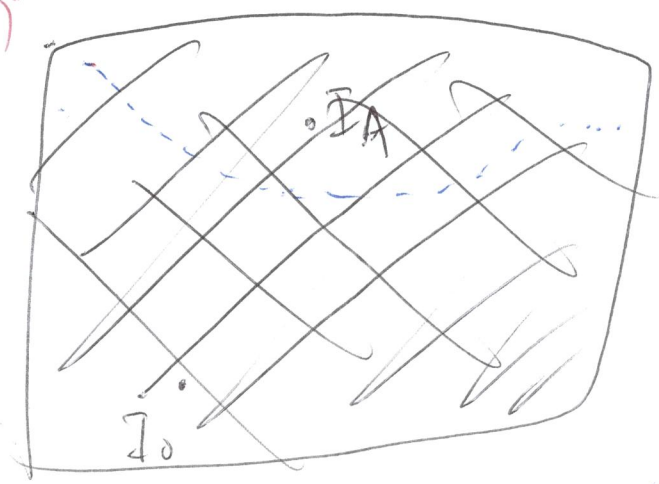
→ Surfaces of $D_S(I_B, I_0)$ are a function of I_B - Again, not a metric space.

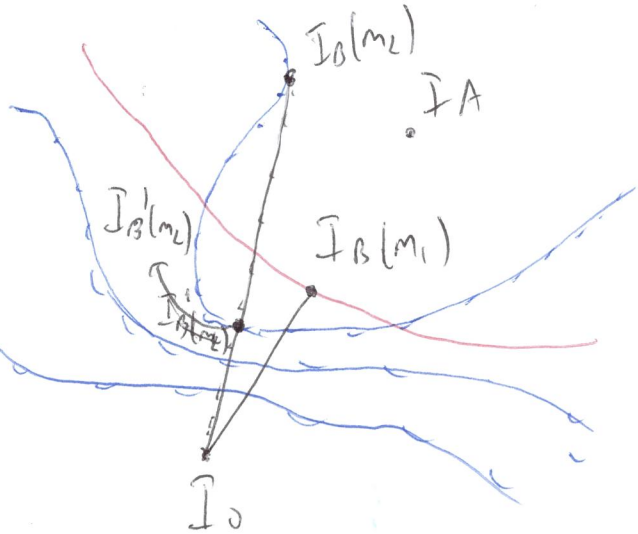
c) Sending a message to a cognizant agent:

$$m = \underset{m' \in \mathcal{M}}{\operatorname{argmin}} \left[D_S\left(I_A, \underbrace{\underset{I_B'}{\operatorname{argmin}} \left[D_S\left(I_B'(m'), I_0\right)\right]}_{\text{(i)}}\right)\right] \quad \text{(ii)}$$

(iii)

minimization (i) must be done by Alice first before (ii)





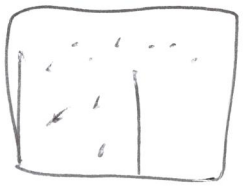
While $I_B(m_2)$ is closer to I_A than $I_B(m_1)$, Bob would choose $I_B'(m_2)$ based on MaxEnt principle so Alice should choose message m_1 over m_2 as Bob is cognizant!

iv) Max Ent principle : Action principle.

Newtonian mechanics ↪ Lagrangian / Hamiltonian mechanics

force drives dynamics

minimise energy in gravitational potential



pressure drives gas

maximize free energy in a thermal state

v) Soft data Constraint

$\langle f(s) \rangle = d$

Maximize $-D_S(I_B | I_0)$ with $\langle f(s) \rangle_{p(s|I_B)} \overset{!}{=} d$.

$p(s|I_B) =: p(s)$ $\quad p(s|I_A) =: q(s)$

$$\Rightarrow \max \left[-\int ds\, p(s) \ln\left(\frac{p(s)}{q(s)}\right) + \lambda \left(\int ds\, p(s) - 1 \right) + \mu \left(\int ds\, f(s) p(s) - d \right) \right]$$

Constraints

analyticity $\longrightarrow$ $0 \overset{!}{=} \frac{\delta [\dots]}{\delta p(s)}$

$$= -\ln\left(\frac{p(s)}{q(s)}\right) - 1 + \lambda + \mu f(s) \quad \Rightarrow \quad p(s) = q(s)\, e^{\lambda - 1 + \mu f(s)}$$