Decision theory

action $a \in A$ = set of possible actions

utility : $u(s,a) = -\ell(s,a)$

↑ situation ↑ loss

Expected utility given $\underline{I}$ on $s \in S'$

$$u(a,\underline{I}) := \langle u(s,a) \rangle_{(s|\underline{I})} = \int_{S'} ds\, P(s|\underline{I})\, u(s,a) =: -\ell(a,\underline{I})$$

optimal action given $\underline{I}$

$$a^*(\underline{I}) = \operatorname*{argmax}_a u(a,\underline{I}) = \operatorname*{argmin}_a \ell(a,\underline{I})$$

$u(I) =$ $u(a^*(\underline{I}),\underline{I})$

- quadratic loss
- linear loss
- delta losss

→ see script!

Optimal action

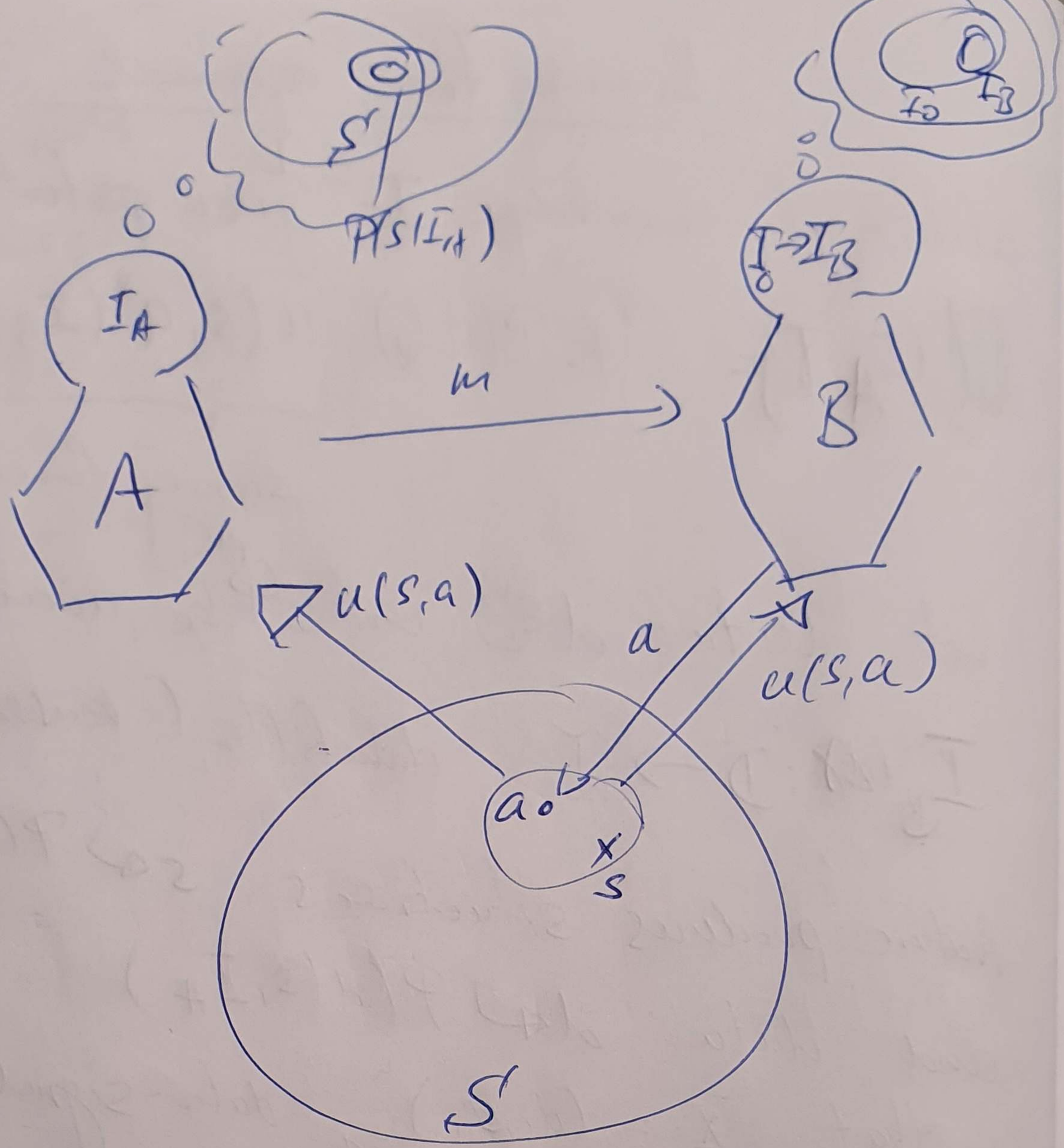
$$a_B^* = \underset{a \in A}{\operatorname{argmax}}\ u_B(a \mid I_B)$$

$$u_B(a \mid I) = \langle u(s \mid a) \rangle_{(s \mid I_B)}$$

assume: a^* is uniquely determined
$u(s,a)$ differentiable w.r.t. a

$$\Rightarrow \quad 0 = \frac{\partial u(a \mid I)}{\partial a} = \left\langle g(s,a) \right\rangle_{(s \mid I)}$$

$$g(s,a) = \frac{\partial u(s,a)}{\partial a}$$



Alice: message m about $s \in S'$

Bob: update $I_0 \to I_B$

action a

both: experience utility $u(s,a)$

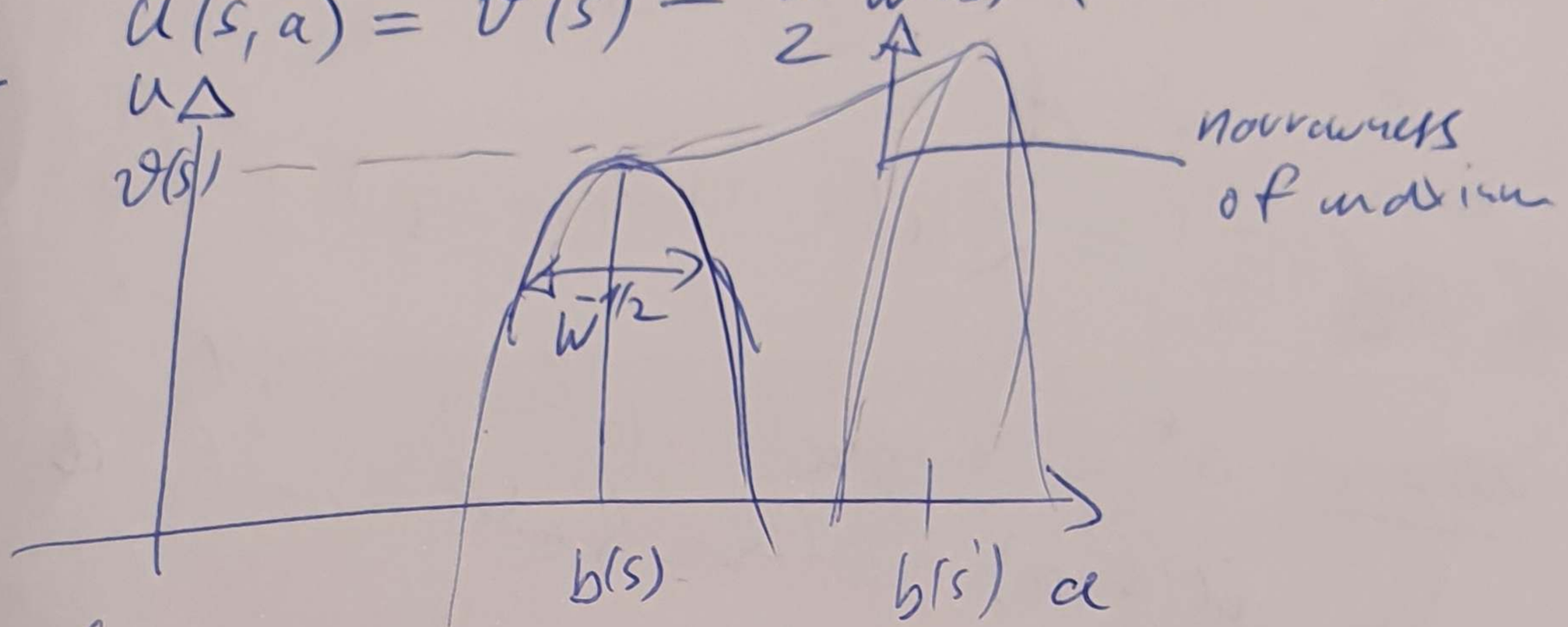
See: Eyal Lin et al 2024

Attention to Entropic Communication

Ann Phys. 2024, 536, 2300334

typical situation s has one optimal action $b(s)$

approx: $u(s,a) = v(s) - \frac{1}{2} w(s) (a - b(s))^2$



Bobs action

$$a_B^*(I_B) = \arg\max_a \langle u(s,a) \rangle_{(s|I_B)}$$

$$= a : \left[\left\langle \frac{\partial u(s,a)}{\partial a} \right\rangle_{(s|I_B)} = 0 \right]$$

such that

$$= a : \left[- \langle (a - b(s)) w(s) \rangle_{(s|I_B)} = 0 \right]$$

$$= a : \left[\langle b(s) w(s) \rangle_{(s|I_B)} = a \langle w(s) \rangle_{(s|I_B)} \right]$$

$$= \frac{\langle b(s) w(s) \rangle_{(s|I_B)}}{\langle w(s) \rangle_{(s|I_B)}} =: \langle b(s) \rangle^{(w)}_{(s|I_B)}$$

$$\langle b(s) \rangle^{(w)}_{(s|I)} := \int ds \, \mathcal{A}^{(w)}(s|I_B) \, b(s)$$

$$\mathcal{A}^{(w)}(s|I_B) := \frac{w(s) P(s|I_B)}{\int ds \, w(s) P(s|I_B)}$$

Attention!

Learning from experience

utility of knowledge $I_B \in \mathcal{I}$ given perfect knowledge I_A

$$U_S(I_A, I_B) = \int ds\, P(s|I_A)\, \underbrace{u(s, a^*(I_B))}_{\text{scoring function}}$$

$s \in S$ S_e available to Σ

Let data $d \in D$ on

$\underline{T}_B : D \to \mathcal{I}$ data filter (= reasoning system)

Nature produces situations $s \hookleftarrow P(s|I_A)$

and data $d \hookleftarrow P(d(s, I_A))$

so that $\underline{X}_A = (d_i, s_i)_{i=1}^N$ data-signal pairs

are available

Utility of $\underline{T}_B$

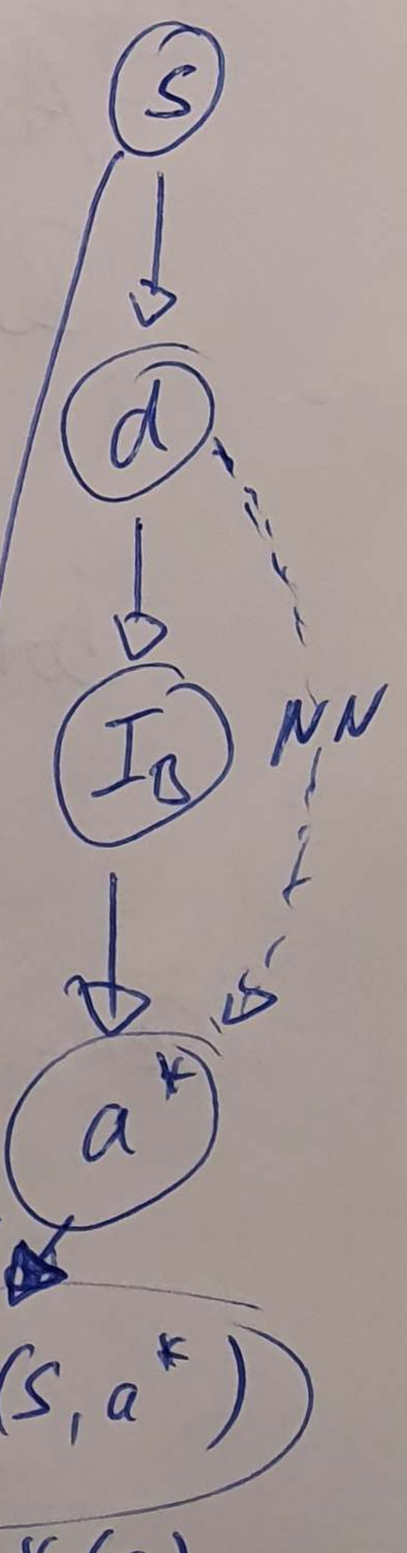
$$U_S(I_A, \underline{T}_B) = \Big\langle u(s, a^*(\underline{T}_B(d))) \Big\rangle_{d(s|I_A)}$$

$$\approx \frac{1}{N} u(s_i, a^*(I_B(d_i)))$$

optimal reasoning:

$$\underline{T}_B^* = \operatorname{argmax}_{I_B} U_S(I_A, I_B)$$

Pavlov-like learning: $a^*(d)$, abstract learning $I_B(d)$, $a^*(s)$



Achieved information gain (AIG)

I_A — m → Bob

Alice

$I_0 \to I_B$ Englert et al (2025)

Wanted: gain $G_S(I_A, I_B, I_0)$ that Alice uses to decide about which I_B she gives Bob

Axioms:

1) Additivity: if Alice knowledge is hierarchical

$$P(s|I_A) = \sum_{i=1}^{u} p_{A,i} P(s|I_{A,i})$$

and $I_A = (I_{A,i}, p_{Ai})_{i=1}^{n}$

Knowledge substates, their probabilities

then

$$G_S(I_A, I_B, I_0) = \sum_{i=1}^{u} p_{Ai} G_S(I_{Ai}, I_B, I_0)$$

this should also hold in the continuum

$$P(s|I_A) = \int dx\, p_{A,(x)} P(s|I_A(x))$$

$$\Rightarrow G_S(I_A, I_B, I_0) = \int dx\, p_A(x) G_S(I_A(x), I_B, I_0)$$

2) Locality: if Alice knows which $s' \in S$ happens, $I_A = I_{s=s'}$, $P(s|I_A) = \delta(s-s')$, gain should only depend on $P(s'|I_B)$, but not on $P(s''|I_B)$ for $s'' \neq s'$

3) Analytical: $G_{s'}(I_A, I_B, I_0) \in C^\infty(\mathcal{I})$ w.r.t I_B

4) Proper: $G_s(I_A, I_B, I_0)$ be maximal w.r.t I_B for $I_B = I_A$

5) Calibration: no update, no gain $G_s(I_A, I_0, I_0) = 0$

Derivation

decomposition of I_A into atomic beliefs by identifying $x = s$, $X = S$ such that

$p_A(x) := P(s=x|I_A)$, $I_A(x) = I_{s=x}$

$$P(s|I_A) = \int dx\, \underbrace{p_A(x)}_{P(s=x|I_A)} \underbrace{P(s|I_{s=x})}_{\delta(s-x)} = P(s|I_A) \checkmark$$

additive $\Rightarrow$ $G_s(I_A, I_B, I_0) = \int dx\, p_A(x)\, G_s(I_A(x), I_B, I_0)$

$$= \int_S ds'\, P(s'|I_A)\, G_{s'}(I_{s=s'}, I_B, I_0)$$

$$= \langle G_s(I_{s=s'}, I_B, I_0) \rangle_{(s'|I_A)}$$

Locality $\Rightarrow G_{S'}(I_{S=S'}, I_B, I_0) = g(S', P(S'|I_B), I_0)$ -§-

to be found function

$+ \lambda \left(1 - \int_{S'} ds'' P(S''|I_B)\right)$

Lagrange multiplier to ensure normalization of $P(S''|I_B)$

maximizing w.r.t $q(S) := P(S|I_B)$

Properness $0 = \left. \frac{\delta G_S(I_A, I_B, I_0)}{\delta P(S'|I_B)} \right|_{I_B = I_A}$ $\left| \frac{\partial G_S}{\partial \lambda} = 0 \Rightarrow \int ds'' P(S''|I_B) = 1 \right.$

properness: maximum at $I_B = I_A$!

$$= \frac{\delta}{\delta q(S')} \left\langle g(S, q(S), I_0) + \left(1 - \int_S ds'' q(S'')\right) \lambda \right\rangle_{(S|I_A)} \Big|_{q = P(\cdot|I_A)}$$

$$\leq \frac{\delta}{\delta q(S')} \left[\left\langle g(S, q(S), I_0) \right\rangle_{(S|I_A)} + \left(1 - \int_{S'} ds'' q(S'')\right) \lambda \right]_{q = P(\cdot|I_A)}$$

$$= \frac{\partial g(S', q(S'), I_0)}{\partial q(S')} P(S'|I_A) - \lambda \Big|_{q = P(\cdot|I_A)}$$

$$= \frac{\partial g(S', \tilde{q}, I_0)}{\partial \tilde{q}} \tilde{q} - \lambda \Big|_{\tilde{q} = P(S'|I_A)}$$

$$\Rightarrow \frac{\partial g(S', \tilde{q}, I_0)}{\partial \tilde{q}} = \frac{\lambda}{\tilde{q}} \Rightarrow g(S, \tilde{q}, I_0) = \lambda \ln \tilde{q} + C(S, I_B)$$

Calibration:

$$0 = G_S(I_A, I_0, I_0) = \langle g(s, P(s|I_0), I_0) \rangle_{(s|I_A)}$$

$$= \langle \lambda \ln P(s|I_0) + c(s, I_0) \rangle_{(s|I_A)}$$

$$\Rightarrow \langle c(s, I_0) \rangle_{(s|I_A)} = -\lambda \langle \ln P(s|I_0) \rangle$$

$$\Rightarrow G_S(I_A, I_B, I_0) = \langle g(s, P(s|I_B), I_0) \rangle_{(s|I_A)}$$

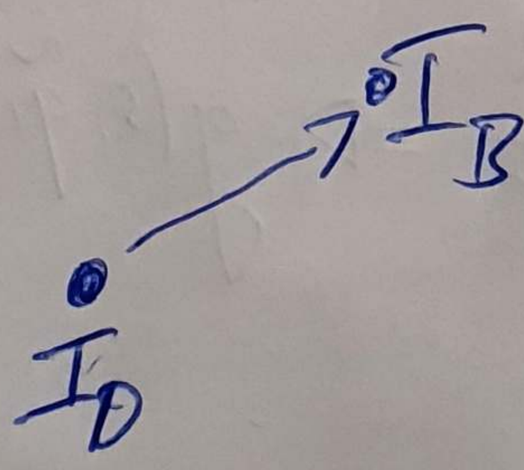
$$= \langle \lambda \ln P(s|I_B) + c(s, I_0) \rangle_{(s|I_A)}$$

$$= \lambda \langle \ln P(s|I_B) - \ln P(s|I_0) \rangle_{(s|I_A)}$$

$$G_S(I_A, I_B, I) = \lambda \int_S ds\, P(s|I_A) \ln \frac{P(s|I_B)}{P(s|I_0)}$$

$$=: D_S(I_A, I_B, I_0)$$

Achieved information gain for the update $I_0 \to I_B$ from the perspective of I_A



units: nits if $\ln$
bits if $\log_2$ is used

Properties of AIG

ideal AIG: $I_B = I_A$

$$D_S(I_A, I_A, I_0) = \int_{S'} ds\, P(s|I_A) \ln \frac{P(s|I_A)}{P(s|I_0)}$$

$$=: D_{S'}(I_A, I_0)$$

$$\equiv D_{KL}(P(s|I_A) \,\|\, P(s|I_B))$$

Kullback-Leibler Divergence or cross-entropy, $D_S(I_A, I_0) \geq 0$ if $I_A \neq I_0$

no update: $I_B = I_0$

$$D_S(I_A, I_0, I_0) = \int_{S'} ds\, P(s|I_A) \ln \underbrace{\frac{P(s|I_B)}{P(s|I_B)}}_{1} = 0$$

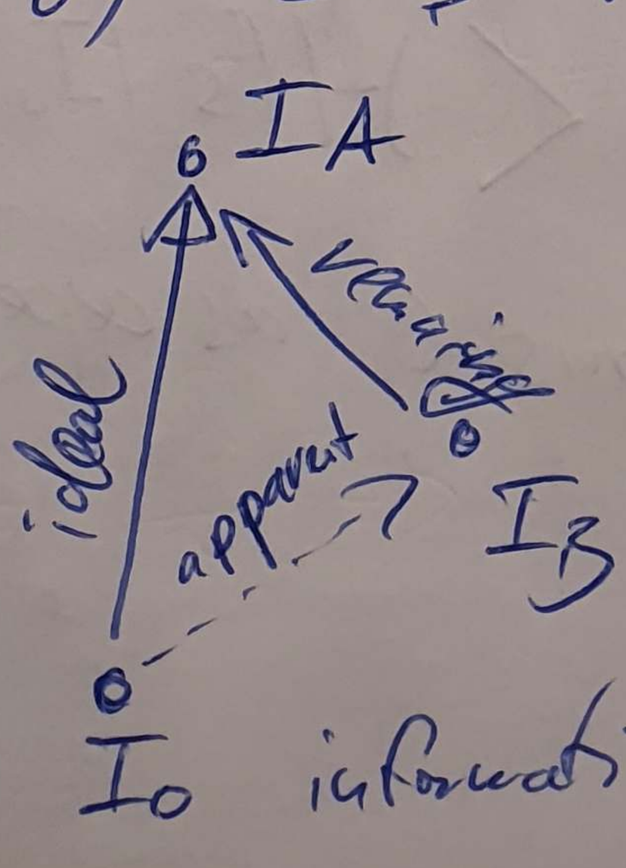
as requested by calibration. $\to 0$

relation to KL

$$D_S(I_A, I_B, I_0) = \int ds\, P(s|I_A) \left[\ln \frac{P(s|I_A)}{P(s|I_0)} - \ln \frac{P(s|I_A)}{P(s|I_B)} \right]$$

$$= D_S(I_A, I_0) - D_{S'}(I_A, I_B)$$

= (ideal-ranking) information gain



AIG for perfect knowledge (Alice is infinitely informed)

$$I_A = I_{S=s'}$$

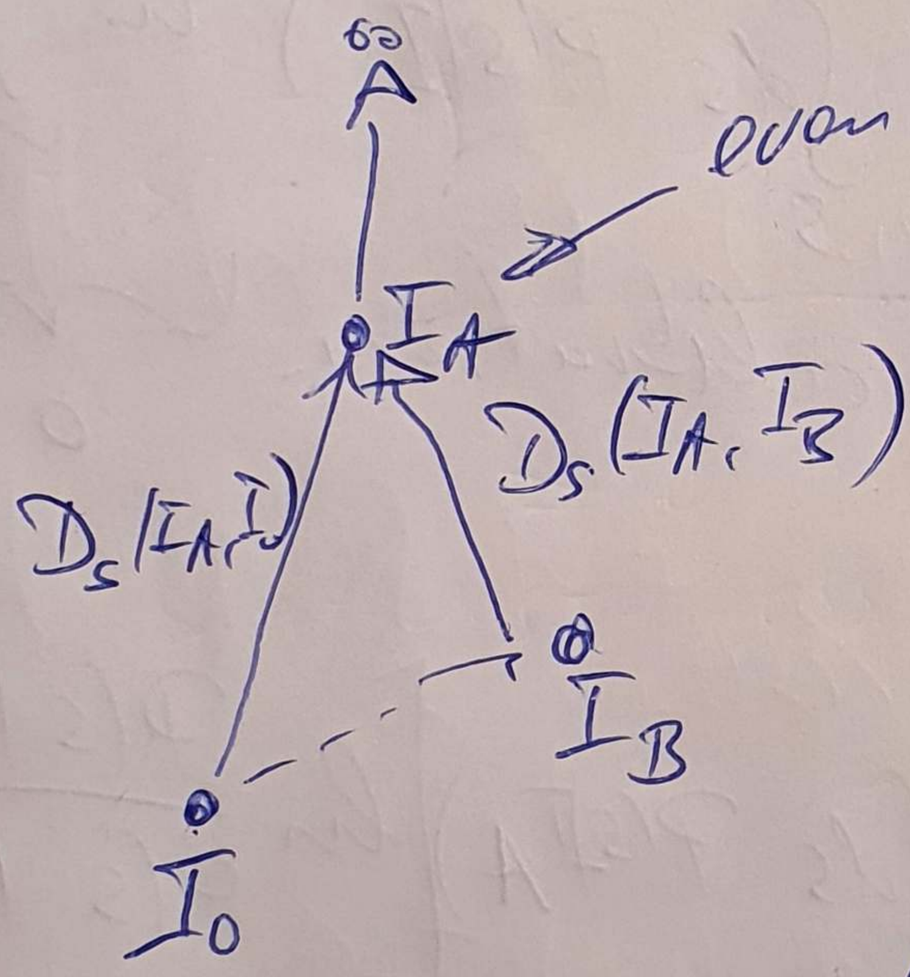
$$D_S(I_A, I_B, I_0) = \int_{S'} ds\ \delta(s-s') \ln \frac{P(s|I_B)}{P(s|I_0)}$$

$$= \ln \frac{P(s'|I_B)}{P(s'|I_0)}$$ well defined as long none of these is 0

However:

$$D_S(I_A, I_{B/0}) = \int ds\ \delta(s-s') \ln \frac{\delta(s-s')}{P(s|I_{B/0})}$$

$= \infty$ not well defined



even if $I_A \to I_{s=s'}$ (∞ information)

AIG stays finite

Surprisal $\mathcal{H}(\cdot|\cdot) := -\ln P(\cdot|\cdot)$

$$D_S(I_A, I_B, I_0) = \left\langle \mathcal{H}(s|I_0) - \mathcal{H}(s|I_B) \right\rangle_{(s|I_A)}$$

Bbi reduction in surprisal as expected by Alice

Separability

be $S' = S_1 \otimes S_2$, $s = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$ and

B's knowledge I_0, I_B separable,

$$P(s|I_{0/B}) = P(s_1|I_{0/B})\, P(s_2|I_{0/B})$$

$$\Rightarrow D_{S'}(I_A, I_B, I_0) = \left\langle \ln \frac{P(s_1|I_B)\, P(s_2|I_B)}{P(s_1|I_0)\, P(s_2|I_0)} \right\rangle_{(s|I_A)}$$

$$= \left\langle \ln \frac{P(s_1|I_B)}{P(s_1|I_0)} \right\rangle_{(s|I_A)} + \left\langle \ln \frac{P(s_2|I_B)}{P(s_2|I_0)} \right\rangle_{(s|I_A)}$$

$$= D_{S'_{s_1}}(I_A, I_B, I_0) + D_{S_{2}}(I_A, I_B, I_0)$$

Antisymmetry

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

Learning a bit

$S = \{head, tail\}$, $I_A = I_{s=head} = head$

$I_0 : P(s|I_0) = 1/2 \longrightarrow I_B = I_A$

$$D_S(I_A, I_A, I_0) = \sum_{s \in S} P(s|I_A) \log_2 \frac{P(s|I_A)}{P(s|I_0)}$$

bit

$$= \underbrace{P(head|head)}_{1} \log_2 \frac{P(head|head)}{1/2} + \underbrace{P(tail|head)}_{0} \log_2 \frac{0}{1/2}$$

$$1 = \log_2 2 = 1 \quad \checkmark$$

Cognitive operations

~~Info~~communicating $A \xrightarrow{m} B$
$I_A \quad I_o \to I_B$

$D_S(I_A, I_B, I_o)$ is AIG of communication act

memorizing

$I_o = I_A$ Alice's initial memory
I_A = perfect update
I_B = imperfect memory update (future self)

$D_S(I_A, I_B, I_o)$ is AIG of memory update

$= D_S(I_A, I_B, I_A) = \underbrace{D_S(I_A, I_A)}_{=0} - \underbrace{D_S(I_A, I_B)}_{\geq 0}$

≤ 0

memorizing is at best lossless!

inference

prior $P(s|I_o)$
message data $d \longrightarrow$ ideal update to $P(s|\underbrace{d, I_o}_{I_A})$
imperfect update $P(s|I_B)$

$D_S((d, I_o), I_B, I_o)$ is AIG of approximative inference.

Cognitive fidelity:

$$CF(I_A, I_B, I_0) := \frac{D_S(I_A, I_B, I_0)}{D_S(I_A, I_A, I_0)}$$

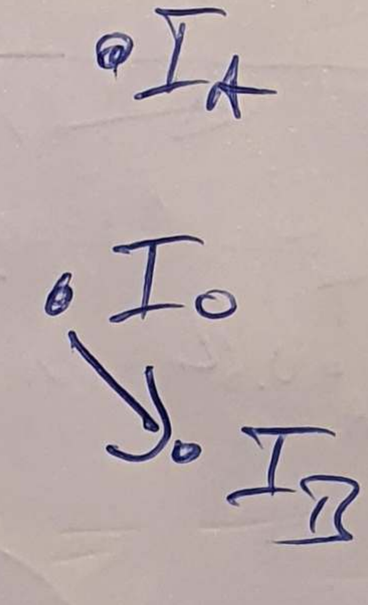
$$= \frac{D_S(I_A, I_0) - D_S(I_A, I_B)}{D_S(I_A, I_0)}$$

$$= 1 - \frac{D_S(I_A, I_B)}{D_S(I_A, I_0)} \quad \in [-\infty, 1]$$

$CF(I_A, I_A, I_0) = 1$ for perfect update

$CF(I_A, I_0, I_0) = 0$ for no update

$CF(I_A, I_B, I_0) < 0$ if $D_S(I_A, I_0) < D_S(I_A, I_B)$



Cognitive efficiency

$$CE(I_A, I_B, I_0) = \frac{D_S(I_A, I_B, I_0)}{C(I_A, I_B, I_0)}$$

costs

units: bits/Euro, Energy, time, ---

AIG for Bernoulli experiment

$S = \{0, 1\}$ $\quad p_{A,B,0} := P(S=0 \mid I_{A,B,C})$

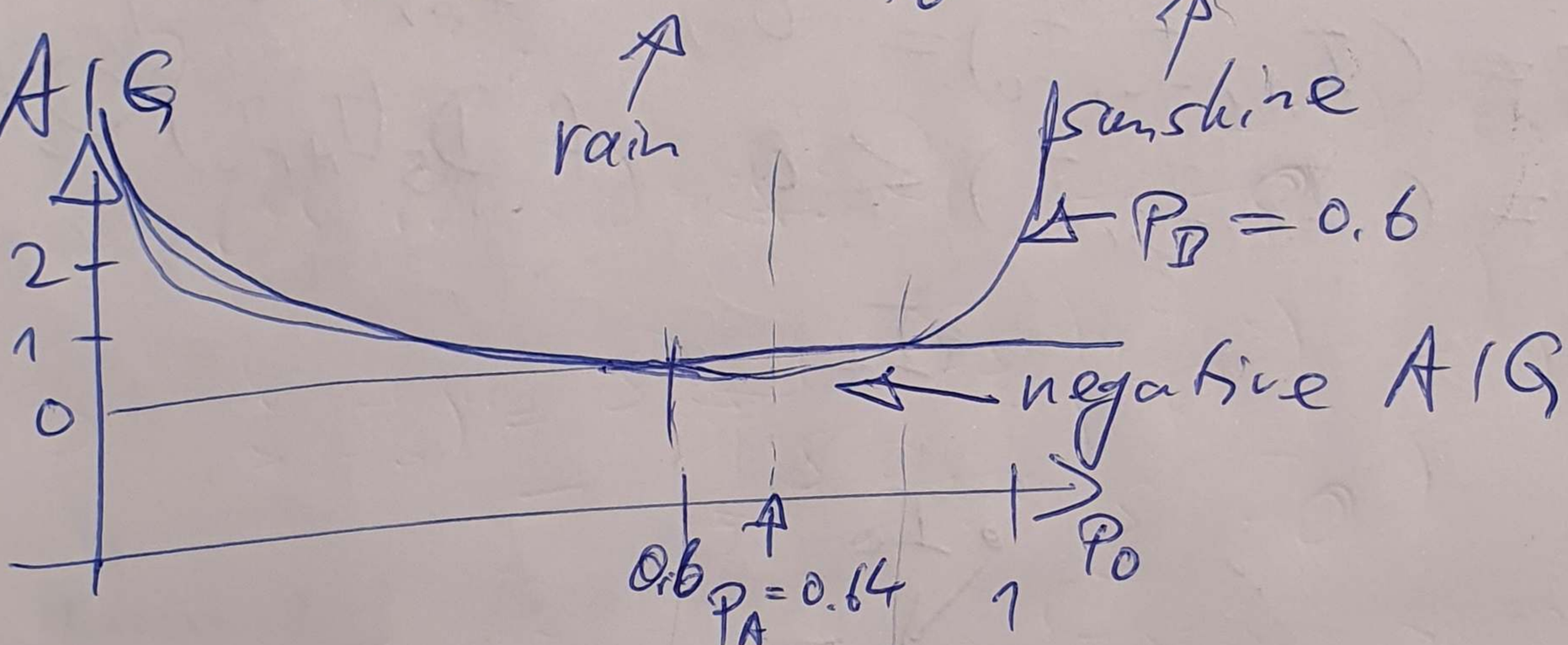
rain sunshine

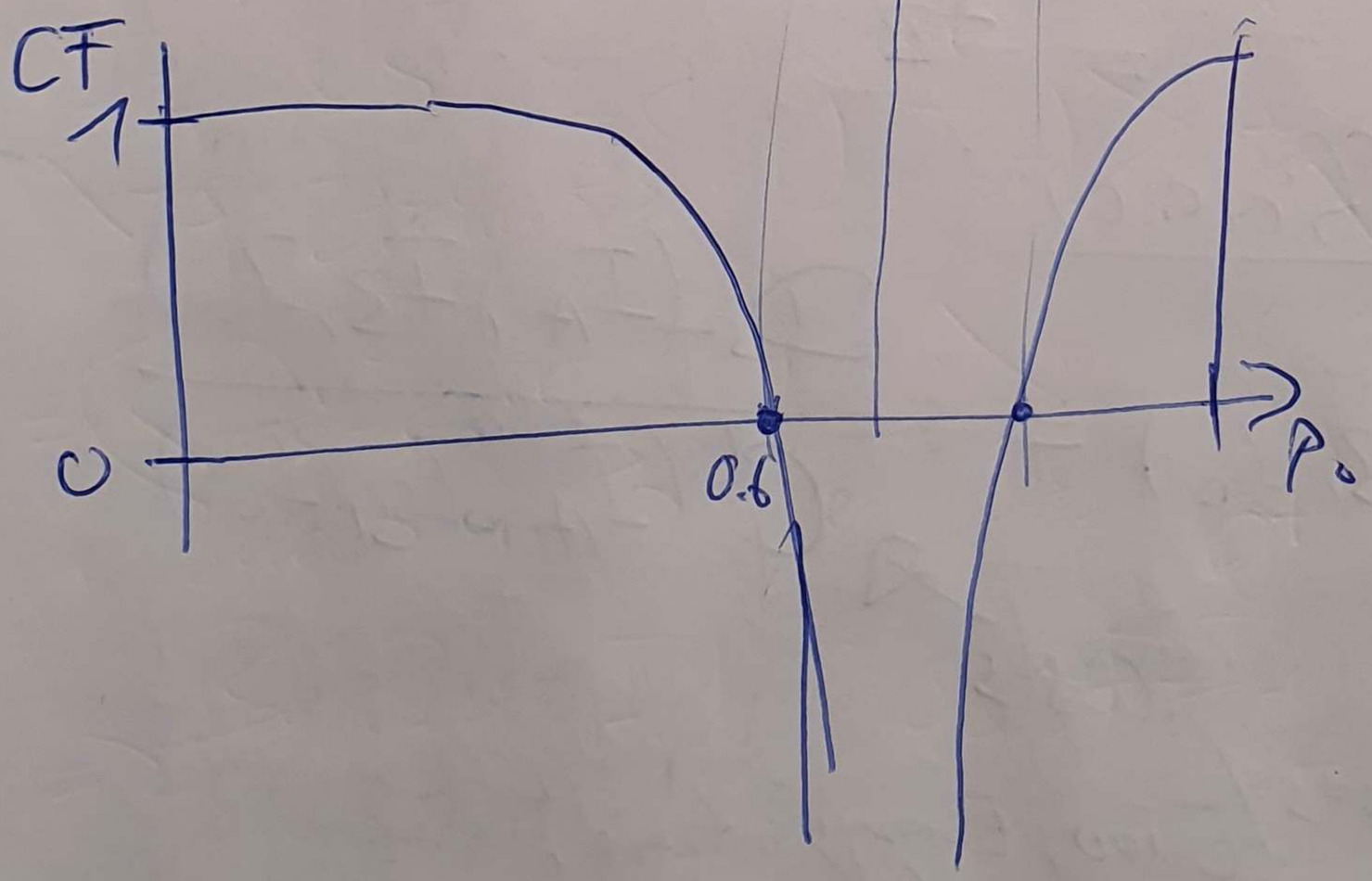
Alice = weather reporter

Bob = audience

Should A update B when she can only communicate ~~[illegible]~~ $\cdot p_A$ rounded to 10%?

$$D_S(I_A, I_B, I_0) = p_A \ln \frac{p_B}{p_0} + (1-p_A) \ln \frac{1-p_B}{1-p_0}$$





expected AIG from generative model - 16

Scenario: update after receiving data d

prior $P(s|I_0)$

posterior $P(s|d,I_0) =: P(s|I_A)$ ideal update

approx posterior $P(s|I_B(d),I_0)$

how to get ~~AIG~~ calculate AIG for $I_B(d)$ when $P(s|d,I_0)$ is intractable?

Assume $(d_i,s_i)_i \hookleftarrow P(d,s|I_0)$ are available via $s_i \hookleftarrow P(s_i|I_0)$ and $d_i \hookleftarrow P(d|s_i,I_0)$, then expected AIG of posterior knowledge $I_{S=s_i}$ averaged over i

$$\left\langle D_S[I_{S=s_i}, I_B(d_i), I_0)\right\rangle_i \approx \left\langle D_S(I_{S=s'}, I_B(d'), I_0)\right\rangle_{(d',s'|I_0)}$$

$$= \left\langle \left\langle D_S(I_{S=s'}, I_B(d'), I_0)\right\rangle_{(s'|d',I_0)}\right\rangle_{(d'|I_0)}$$

$$= \left\langle \int ds' P(s'|d,I_0) \int ds\, \delta(s-s') \underbrace{I_A(d')}_{\text{on } P(s|I_B(d))} \right\rangle_{(d|I_0)}$$

$$= \left\langle \int ds' P(s'|d,I_0) \ln \frac{P(s'|I_B(d))}{P(s'|I_0)} \right\rangle_{(d|I_0)}$$

$$= \left\langle D_S(I_A, I_B(d), I_0)\right\rangle_{(d|I_0)}$$

Recipe to get expected AIG of an cognitive operation $I_B(d)$:

draw samples $s_i \leftarrow P(s_i | I_o)$

$d_i \leftarrow P(d_i | s_i, I_o)$

$$\langle D_S(\underbrace{I_A}_{(d, I_o)}, I_B(d), I_o) \rangle_{(d|I_o)} \approx \sum_i \ln \frac{P(s_i | I_B(d_i))}{P(s_i | I_o)}_i$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[H(s_i | I_o) - H(s_i | I_B(d_i)) \right]$$

why does this work? product rule!

$$P(d, s | I_o) = P(d | s, \tilde{I_o}) P(s | I_o)$$
$$= P(s | d, I_o) P(d | I_o)$$

If AIG is of value to a cognitive system and evolution shaped it, it will aim to maximize it, which means there is a selection force pushing $I_B(d)$ towards the Bayesian optimum $I_A(d) = (I_o, d)$