Reinforcement Learning China Summer School

RLChina 2021

Bayesian Brain

Prof. Jun Wang

UCL

August 17, 2021

Life and the universe

Life and the universe

Table of Contents

Table of Contents

What is “learning”?

- in biology, **learning** means
  a change of behaviour as a result of experience
  - in classical conditioning[1], animals can learn to identify a useful pattern in the environment by associating one stimulus with another:
    repeated given $\{$*ring-a-bell*, *food*$\}$, a dog will start to salivate (anticipate the upcoming of the food) when bell ringing again
- learned behaviours are **adaptive**, and thus are essential for animals to survive in the changing environment
  - e.g., they may learn not to eat certain foods if they have ever become ill after eating them
- more learned behaviours $\Longrightarrow$ more intelligent

---

[1] Ivan Petrovitch Pavlov and William Gantt. “Lectures on conditioned reflexes: Twenty-five years of objective study of the higher nervous activity (behaviour) of animals.”. In: (1928).

Learning is not limited to biological systems

- *machine learning* is to answer the question of how computer programmes can improve their performance through experience
  - past experience exists in various forms, typically including (1) human-labelled or unlabelled data sets, (2) past interactions with the environments, or (3) data collected from realistic simulators
- with experience, computers can be trained to identify the associations, spot the underlying patterns, and make accurate predictions and forecasts the futures
- as a field, machine learning studies fundamental theory and algorithms and provides a principled solution for the computational methods of learning

Three levels in any information processing system[2]

- **computational theory**
  - what is the problem in a generic manner?
  - what is the computing goal?
  - what is the logic of strategy behind?
- **representation and algorithm**
  - how can the identified computational problems be solved?
  - what are the inputs/outputs and the algorithm for the transformation?
- **hardware implementation**
  - how can the representation and algorithm be realised physically?
    - human: biological
    - AI: silicon using transistors

[2] David Marr. "Vision: A computational investigation into the human representation and processing of visual information". In: *Inc., New York, NY* 2.4.2 (1982).

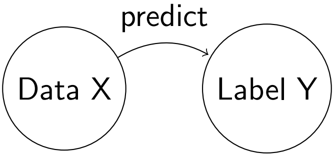
Human intelligent v.s. current AI[3]

Dseitpe the fcat taht the letetres in tehese wrdos are jmbuled, you are sitll albe to raed tehm, Bceasue the frsit and lsat ltertes are in the rghit palce. your bairn can use tohse ceus to fgiure out waht I'm syanig

研表究明,汉字序顺并不定一影阅响读，事证实明了当你看这完句话之后才发字现都乱是的

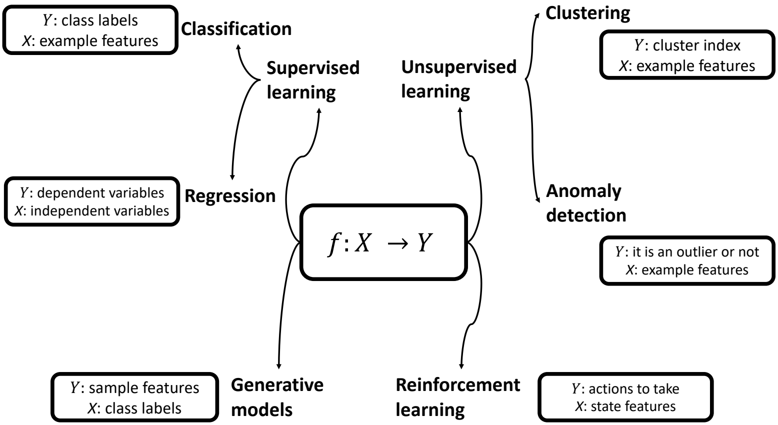
[3] Eliezer Sternberg. *NeuroLogic: The Brain's Hidden Rationale Behind Our Irrational Behavior*. Vintage, 2016.

Typical narrow machine learning



- mathematically, ml can loosely boil down to the question of finding a unknown function mapping $f : X \rightarrow Y$,
    - X is input feature space representing data points,
    - Y is label space representing the knowledge outputs
    - thus the mapping f represents a *knowledge discovery* process from a given data point $x \in X$ to the specific label $y \in Y$ associated with the data point:$y = f(x)$
- however, as f is not known a priori, the goal of the learning is to identify a hypothesis $h \in H$ from a predefined set H to approximate the unknown f, so that
    - the learned function h can *predict* the output variable $\hat{y}_{\text{new}} = h(x_{\text{new}})$ for a given new data point

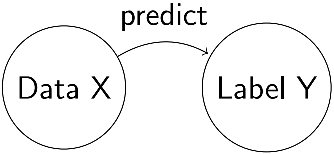
View narrow machine learning as function mapping



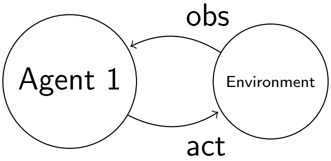
machine learning tasks are different in the experience available, the objective function, and the specific learning algorithms

Machine learning: reasoning under uncertainty

pattern recognition:



decision making (reinforcement learning):



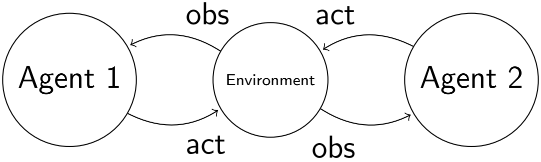
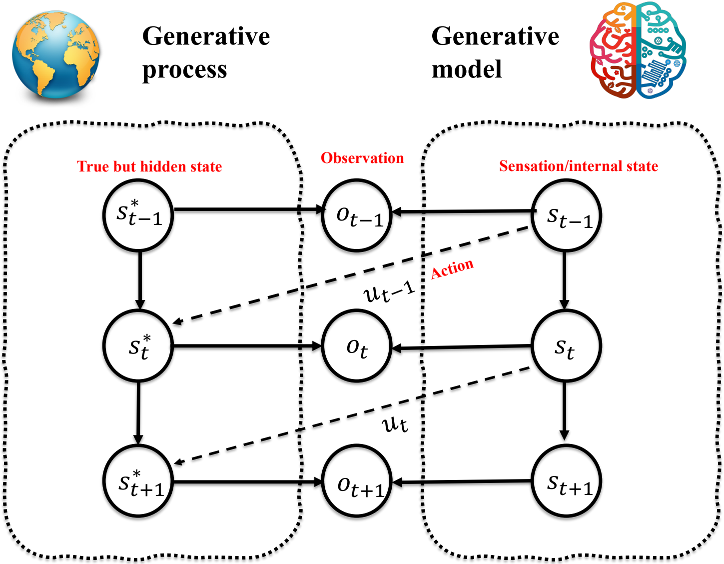


Table of Contents

A living system and its environment



Entropy

- *Entropy*: a measure of the average surprise (or uncertainty or disorder) of a random event sampled from a probability distribution or density
- definition: entropy of a random event X with possible outcomes $x_1, ..., x_n$:

$$H(p) = -\sum_i p(X = x_i) logp(X = x_i)$$

- low entropy means that, on average, the outcome is relatively predictable

Table of Contents

Variational Bayes[4]/Free energy principle[5]

- an agent builds a world model by having an internal representation of the world s from observation o as:

$$p(s|o) = \frac{p(o|s)p(s)}{\int_s p(o|s)p(s)\, dx}$$

- typically the denominator is intractable and one can approximate the posterior using a tractable function family $q(s) \in \mathbb{Q}$ by minimising a dis-similarity measure, e.g., KL Divergence:

$$KL(q(s)||p(s|o)) = \mathbb{E}_{q(s)}[\log \frac{q(s)}{p(s|o)}]$$

[4] Michael I Jordan et al. "An introduction to variational methods for graphical models". In: *Machine learning* 37.2 (1999), pp. 183–233.

[5] Karl Friston. "The free-energy principle: a unified brain theory?" In: *Nature reviews neuroscience* 11.2 (2010), pp. 127–138.

Variational Bayes[6]/Variational free energy[7]

- we then derive:

$$
\begin{aligned}
KL(q(s)||p(s|o)) &= \mathbb{E}_{q(s)}[\log \frac{q(s)}{p(s|o)}] \\
=&\mathbb{E}_{q(s)}[\log q(s) - \log p(s,o) + \log p(o)] \\
=&\mathbb{E}_{q(s)}[\log q(s) - \log p(s,o)] + \log p(o)
\end{aligned}
$$

- reorganising gives the measure of surprise (or negative model evidence or log marginal distribution):

$$
\begin{aligned}
-\log p(o) =&\mathbb{E}_{q(s)}[\log q(s) - \log p(s,o)] - D_{KL}[]q(s)||p(s|o)] \\
\leq&\mathbb{E}_{q(s)}[\log q(s) - \log p(s,o)] \equiv \mathbb{F}(q)
\end{aligned}
$$

where RHS is variational free energy (negative ELBO, Evidence Lower BOund)

[6] Michael I Jordan et al. “An introduction to variational methods for graphical models”. In: *Machine learning* 37.2 (1999), pp. 183–233.

[7] Karl Friston. “The free-energy principle: a unified brain theory?” In: *Nature reviews neuroscience* 11.2 (2010), pp. 127–138.

Free energy principle[9]

- the VFE can be decomposed in three principal ways:

$$\begin{aligned}
\mathbb{F}(q) &= \mathbb{E}_{q(s)}[\log q(s) - \log p(s,o)] \\
&= \underbrace{\mathbb{E}_{q(s)}[\log q(s)]}_{\text{Negative Entropy}} - \underbrace{\mathbb{E}_{q(s)}[\log p(s,o)]}_{\text{Energy}} \\
&= \underbrace{\mathbb{E}_{q(s)}[\log p(o|s)]}_{\text{Accuracy}} + \underbrace{D_{KL}[q(s)|p(s)]}_{\text{Complexity}} \\
&= \underbrace{D_{KL}[q(s)\|p(s|o)]}_{\text{Posterior Divergence}} \quad \underbrace{-\log p(o)}_{\text{Negative Log Model Evidence}}
\end{aligned}$$

- free energy principle: the goal of a living system is to minimise the free energy $\mathbb{F}$ in order to avoid surprising observations (states) $\rightarrow$ maintain homeostasis[8] (thus remain alive)

[8] The process whereby an open or closed system regulates its internal environment to maintain its states within bounds.

[9] Karl Friston. "The free-energy principle: a unified brain theory?" In: *Nature reviews neuroscience* 11.2 (2010), pp. 127–138.

Table of Contents

Extension to handle dynamics and control

- the real world is dynamic and thus we extend the model to have an observation $o = \{o_1, ..., o_T\}$ and a hidden state at each point in time $s = \{s_1, ..., s_T\}$
- also add action policy $\pi = \{u_1, ...u_T\}$ which interacts with the environment by altering the next hidden state
- the generative model: 1) $p\left(s_{t+1} \mid s_t, u_t\right), t > 1$; $p(s_1)$ 2) $p(o_t|s_t)$
- the Expected Free Energy (EFE), from time τ until the time horizon T:
  $\mathcal{G} = \mathbb{E}_{q(o_{\tau:T}, s_{\tau:T}, \pi)}\left[\ln q\left(s_{\tau:T}, \pi\right) - \ln \tilde{p}\left(o_{\tau:T}, s_{\tau:T}\right)\right]$
- where an agent's goals are encoded as (subjective) desired distribution over observations $\tilde{p}(o_{\tau:T})$[10]; thus we have $\tilde{p}\left(o_\tau, s_\tau\right) \approx \tilde{p}\left(o_\tau\right) q\left(s_\tau \mid o_\tau\right)$

[10] in ML, an optimality variable is added to encode the desires

Expected free energy[12]

- a temporal mean-field factorisation[11] is assumed:
  - the variational function: $q\left(s_{\tau:T},\pi\right) \approx q(\pi)\prod_{\tau}^{T} q(s \mid \pi)$
  - the gen. model: $\tilde{p}\left(o_{\tau:T},s_{\tau:T}\right) \approx \prod_{t}^{T} \tilde{p}\left(o_{\tau}\right) q\left(s_{\tau} \mid o_{\tau}\right)$
- as a result, they are independent between time steps:

$$\mathbb{G} = \mathbb{E}_{q(o_{\tau:T},s_{\tau:T},\pi)}\left[\ln q\left(s_{\tau:T},\pi\right) - \ln \tilde{p}\left(o_{\tau:T},s_{\tau:T}\right)\right]$$

$$=\mathbb{E}_{q(o_{\tau:T},s_{\tau:T}\mid\pi)q(\pi)}\left[\ln q\left(s_{\tau:T} \mid \pi\right) + \ln q(\pi) - \ln \tilde{p}\left(o_{\tau:T},s_{\tau:T}\right)\right]$$

$$=\mathbb{E}_{q(\pi)}\left[\ln q(\pi) + \mathbb{E}_{q(o_{\tau:T},s_{\tau:T}\mid\pi)}\left[\sum_{\tau}^{T}\left[\ln q(s \mid \pi) - \ln \tilde{p}\left(o_{\tau},s_{\tau}\right)\right]\right]\right]$$

$$=\mathbb{E}_{q(\pi)}\left[\ln q(\pi) - \left(-\sum_{t}^{T}\mathbb{G}_{\tau}(\pi)\right)\right] = \mathrm{D}_{KL}\left[q(\pi) \| e^{-\sum_{t}^{T}\mathbb{G}_{\tau}(\pi)}\right]$$

<u>where $\mathbb{G}_{\tau}(\pi) = \mathbb{E}_{q(o_{\tau},s_{\tau}\mid\pi)}\left[\ln q\left(s_{\tau} \mid \pi\right) - \ln \tilde{p}\left(o_{\tau},s_{\tau}\right)\right]$</u>

[11] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[12] Beren Millidge, Alexander Tschantz, and Christopher L Buckley. "Whence the expected free energy?" In: *Neural Computation* 33.2 (2021), pp. 447–482.

Expected free energy

- thus optimising $\mathbb{G}$ results in $q^*(\pi) = \mathrm{SoftMax}\Big(-\sum_t^T \mathbb{G}_\tau(\pi)\Big)$
- EFE at time τ, $\mathbb{G}_\tau$, can be further decomposed as:

$$
\begin{aligned}
&\mathbb{G}_\tau(\pi)\\
=&\mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln q\left(s_\tau \mid \pi\right) - \ln \tilde{p}\left(o_\tau, s_\tau\right)\right]\\
\approx&\mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln q\left(s_\tau \mid \pi\right) - \ln \tilde{p}\left(o_\tau\right) - \ln q\left(s_\tau \mid o_\tau\right)\right]\\
\approx&\underbrace{-\mathbb{E}_{q(o_\tau, \mid \pi)}\left[\ln \tilde{p}\left(o_\tau\right)\right]}_{\text{Extrinsic Value}} - \underbrace{\mathbb{E}_{q(o_\tau)} \mathrm{D}_{KL}\left[q\left(s_\tau \mid o_\tau\right) \| q\left(s_\tau \mid \pi\right)\right]}_{\text{Epistemic Value}}
\end{aligned}
$$

- thus, optimal policies are obtained by minimising the sum of the expected free energies
- EFE is estimated using the generative model to roll out predicted futures, and compute the EFE of those futures

Expected free energy

- similar to variational free energy, EFE can be also decomposed as:

$$
\begin{aligned}
&\mathbb{G}_\tau(\pi) \\
=&\mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln q\left(s_\tau \mid \pi\right)-\ln \tilde{p}\left(o_\tau, s_\tau\right)\right] \\
\approx&\mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln q\left(s_\tau \mid \pi\right)-\ln \tilde{p}\left(o_\tau\right)-\ln q\left(s_\tau \mid o_\tau\right)\right] \\
\approx&\underbrace{-\mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln \tilde{p}\left(o_\tau\right)\right]}_{\text{Extrinsic Value}}-\underbrace{\mathbb{E}_{q(o_\tau)} \mathrm{D}_{KL}\left[q\left(s_\tau \mid o_\tau\right) \| q\left(s_\tau \mid \pi\right)\right]}_{\text{Epistemic Value}}
\end{aligned}
$$

- note that an epistemic uncertainty refers to the deficiencies by a lack of knowledge or information; reducible with more data or a better model
- the second term measures the reduced entropy of s_τ when observed o_τ $\rightarrow$ max its value $\rightarrow$ we intend to choose the policy such that $H[q(s|\pi]$ is high and strong dependency between states and observation (aka $H[q(s|o]$ is low)

Expected free energy

- one can also decompose it into the following:

$$\mathbb{G}_\tau(\pi) = \mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln q\left(s_\tau \mid \pi\right) - \ln \tilde{p}\left(o_\tau, s_\tau\right)\right]$$
$$= \mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln q\left(s_\tau \mid \pi\right) - \ln \tilde{p}\left(o_\tau \mid s_\tau\right) - \ln p\left(s_\tau\right)\right]$$
$$= \underbrace{-\mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln \tilde{p}\left(o_\tau \mid s_\tau\right)\right]}_{\text{Accuracy}} + \underbrace{\mathbb{E}_{q(o_\tau \mid s_\tau)}\left[\mathrm{D}_{KL}\left[q\left(s_\tau \mid \pi\right) \| p\left(s_\tau\right)\right]\right]}_{\text{Complexity}}$$

- or this (typically used for computation):

$$\mathbb{G}_\tau(\pi) = \mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\left[\ln q\left(s_\tau \mid \pi\right) - \ln \tilde{p}\left(o_\tau\right) - \ln q\left(s_\tau \mid o_\tau\right)\right]$$
$$= \mathbb{E}_{q(o_\tau, s_\tau \mid \pi)}\Big[\cancel{\ln q\left(s_\tau \mid \pi\right)} - \ln \tilde{p}\left(o_\tau\right) - \ln p\left(o_\tau \mid s_\tau\right)$$
$$- \cancel{\ln q\left(s_\tau \mid \pi\right)} + \ln q\left(o_\tau\right)\Big]$$
$$= \underbrace{D_{KL}\left[q\left(o_t \mid \pi\right) \| \tilde{p}\left(o_t\right)\right]}_{\text{Expected Cost}} + \underbrace{\mathbb{E}_{q(s_t \mid \pi)}\Big[H\left[p\left(o_t \mid s_t\right)\right]\Big]}_{\text{Expected Ambiguity}}$$

Table of Contents

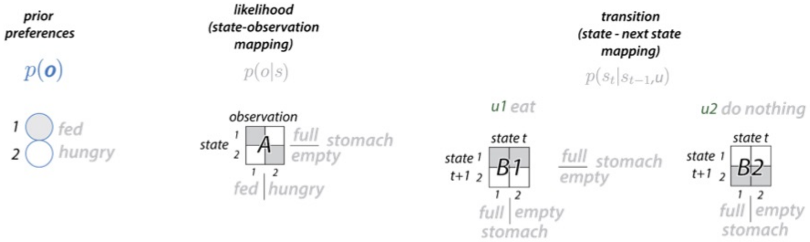
A simple example[13]

- suppose there is an agent interacts with an environment
- perception:
  - the environment has two hidden states $s \in \{1, 2\}$, e.g., *there is food in your stomach (1) or not (2)*
  - while s is NOT measurable, there is an observation $o = \{1, 2\}$, e.g., you feeling fed (1) or hungry (2)
  - assume $p(o|s)$ is given as a 2x2 likelihood matrix (called "A" matrix) which maps states to observations, e.g., if you fed, you have food and vice versa
- control:
  - transition prob. $p(s_t|s_{t-1}, u)$ maps the previous state to the next one, depending on action $u = \{u_1, u_2\}$
  - parameterised by a separate 2x2 transition matrix ("B") for each action, e.g., *either go get food ($u1$) or do nothing ($u2$); if $u1$, will have food in the next state, regardless of whether we have it now, and vice versa*

[13] https://medium.com/@solopchuk/tutorial-on-active-inference-30edcf50f5dc

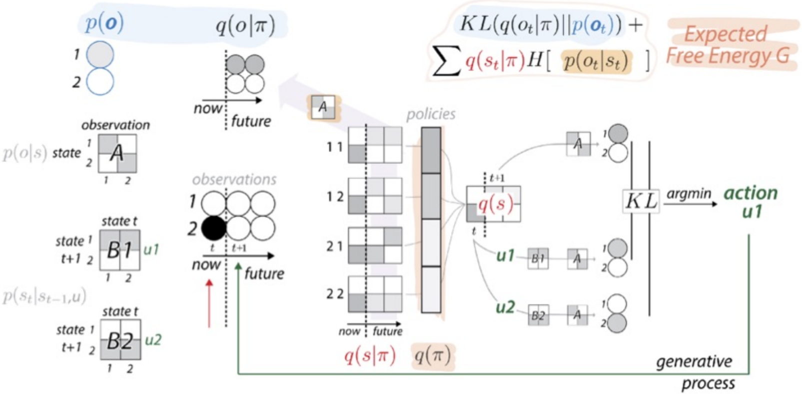
A simple example[14]

- preference (as a form of objective or reward)
  - we also have prior preferences $\tilde{p}(o)$, e.g., we like to be fed and not hungry, so we assign a higher probability to observation $o = 1$ fed. In other words, we express preferences over observations as probability $\tilde{p}(o)$



[14] https://medium.com/@solopchuk/tutorial-on-active-inference-30edcf50f5dc

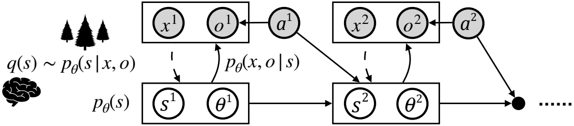
A simple example[15]



action selection: option 1) $q(s|\pi) \to q(o|\pi) \to \mathbb{G}(\pi) \to q(\pi)$
option 2 $u^*_{t+1} = \arg\max_u D_{KL}\left[AS_{t+1}||A(B(u))S_t\right]$, Bayesian model average method

[15] `https://medium.com/@solopchuk/tutorial-on-active-inference-30edcf50f5dc`

Unifying perception and control (Bayesian brain)[16]



- the latent state $\boldsymbol{s}$: true world configuration such as pixel assignment, the optimality o is a binary variable
- the perception model includes a bottom-up recognition model $q(\boldsymbol{s})$ and a top-down generative model $p(\boldsymbol{x}, o, \boldsymbol{s})$ (decomposed into the likelihood $p(\boldsymbol{x}, o|\boldsymbol{s})$ and the prior belief $p(\boldsymbol{s})$)
- the prior knowledge $\boldsymbol{\theta}$ represents the physical law of the environment (the property of each object)
- control is performed by taking an action a to change the environment state.

[16] Minne Li et al. "Joint Perception and Control as Inference with an Object-based Implementation". In: (2020).

Table of Contents

Bayesian decision principle

making decision under uncertainty can be interpreted as maximising expected utility in the face of uncertainty[17]

- the uncertainty is captured by the (hidden) *state of nature*: $z \in \mathbb{Z}$
- decisions (aka *actions*): $a \in \mathbb{A}$
- reward: $R(z, a)$
- the posterior distribution $p(z|o)$, where we can perform a statistical investigation to obtain information (denoted as $o = (o_1, o_2, ..., o_n) \in O$) about the nature state θ

(**Conditional Bayes Decision Principle**)

$$a^*|o = \arg\max_a E_{\theta \sim p(z|o)}[R(o, a)], \tag{1}$$

where the principle is the only fundamentally correct analysis[18].

Yet, the posterior can be difficult to obtain practically

[17] John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

[18] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

The duality between inference and control

- the filtering problem is shown to be the dual of the noise-free regulator (control) problem[19]

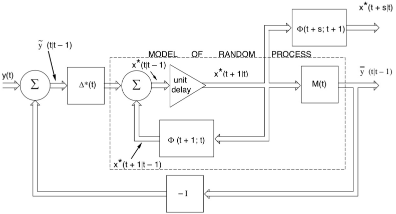


Figure: Optimal filter

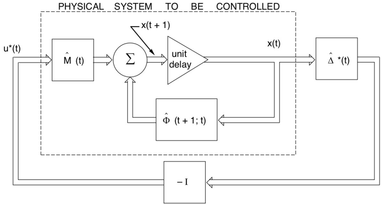
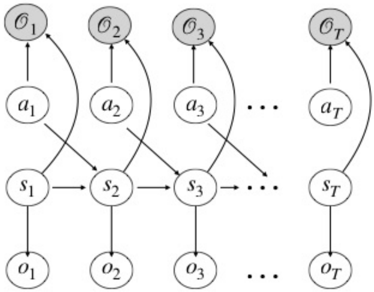


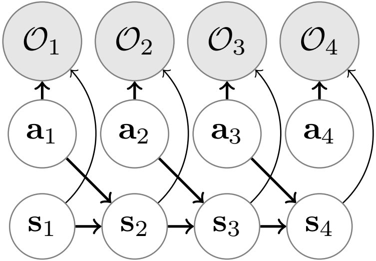
Figure: Optimal controller

[19] Rudolph Emil Kalman. "A new approach to linear filtering and prediction problems". In: (1960).

Control as inference:[20]



(a) states s are hidden



(b) states s are observable

- the idea: add an optimality variable $\mathcal{O}_t = 1$ and assume a biased probability with reward r:
  $p(\mathcal{O}_t = 1 \mid \mathrm{s}_t, \mathrm{a}_t) = \exp(r(\mathrm{s}_t, \mathrm{a}_t))$

[20] Sergey Levine. "Reinforcement learning and control as probabilistic inference: Tutorial and review". In: *arXiv preprint arXiv:1805.00909* (2018); Beren Millidge et al. "On the relationship between active inference and control as inference". In: *International Workshop on Active Inference*. Springer. 2020, pp. 3–11.

Control as inference:[21]

- let us look at an MDP model (states are observable)
- the evidence: $\mathcal{O}_t = 1$ for all $t \in \{1, ..., T\}$; thus ELBO is

$$\log p\left(\mathcal{O}_{1:T}\right) = \log \iint p\left(\mathcal{O}_{1:T}, \mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right) d\mathrm{s}_{1:T} d\mathrm{a}_{1:T}$$

$$= \log \iint p\left(\mathcal{O}_{1:T}, \mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right) \frac{q\left(\mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right)}{q\left(\mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right)} d\mathrm{s}_{1:T} d\mathrm{a}_{1:T}$$

$$= \log E_{(\mathrm{s}_{1:T}, \mathrm{a}_{1:T}) \sim q(\mathrm{s}_{1:T}, \mathrm{a}_{1:T})} \left[\frac{p\left(\mathcal{O}_{1:T}, \mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right)}{q\left(\mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right)}\right]$$

$$\geq E_{(\mathrm{s}_{1:T}, \mathrm{a}_{1:T}) \sim q(\mathrm{s}_{1:T}, \mathrm{a}_{1:T})} \left[\log p\left(\mathcal{O}_{1:T}, \mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right) - \log q\left(\mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right)\right]$$

[21] Sergey Levine. "Reinforcement learning and control as probabilistic inference: Tutorial and review". In: *arXiv preprint arXiv:1805.00909* (2018).

Control as inference:[22]

- we make use of the evidence $p\left(\mathcal{O}_{1:T}, \mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right) = \left[p\left(\mathrm{s}_1\right)\prod_{t=1}^{T} p\left(\mathrm{s}_{t+1} \mid \mathrm{s}_t, \mathrm{a}_t\right)\right] \exp\left(\sum_{t=1}^{T} r\left(\mathrm{s}_t, \mathrm{a}_t\right)\right)$ and
- factorise the variational distribution as $q(\tau) \equiv q\left(\mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right) = q\left(\mathrm{s}_1\right)\prod_{t=1}^{1} q\left(\mathrm{s}_{t+1} \mid \mathrm{s}_t, \mathrm{a}_t\right) q\left(\mathrm{a}_t \mid \mathrm{s}_t\right)$
- this leads to the final ELBO: $\log p\left(\mathcal{O}_{1:T}\right) \geq E_{q\left(\mathrm{s}_{1:T}, \mathrm{a}_{1:T}\right)}\left[\sum_{t=1}^{T} r\left(\mathrm{s}_t, \mathrm{a}_t\right) - \log q\left(\mathrm{a}_t \mid \mathrm{s}_t\right)\right]$
- where we define $q\left(\mathrm{s}_{t+1} \mid \mathrm{s}_t, \mathrm{a}_t\right) = p\left(\mathrm{s}_{t+1} \mid \mathrm{s}_t, \mathrm{a}_t\right)$

[22] Sergey Levine. "Reinforcement learning and control as probabilistic inference: Tutorial and review". In: *arXiv preprint arXiv:1805.00909* (2018).

Control as inference:[23]

- without using temporal mean-field factorisation as Active Inference, one can derive Value Iteration similar to a standard RL solution
- the ELBO can be decomposed recursively as:
$$E_{q(\mathrm{s}_t,\mathrm{a}_t)}\left[r\left(\mathrm{s}_t,\mathrm{a}_t\right)-\log\pi\left(\mathrm{a}_t\mid\mathrm{s}_t\right)\right]+$$
$$E_{q(\mathrm{s}_t,\mathrm{a}_t)}\left[E_{\mathrm{s}_{t+1}\sim p(\mathrm{s}_{t+1}\mid\mathrm{s}_t,\mathrm{a}_t)}\left[V\left(\mathrm{s}_{t+1}\right)\right]\right]=$$
$$E_{q(\mathrm{s}_t)}\left[-D_{\mathrm{KL}}\left(\pi\left(\mathrm{a}_t\mid\mathrm{s}_t\right)\|\frac{1}{\exp(V(\mathrm{s}_t))}\exp\left(Q\left(\mathrm{s}_t,\mathrm{a}_t\right)\right)\right)+V\left(\mathrm{s}_t\right)\right]$$
- where we define:
$$Q\left(\mathrm{s}_t,\mathrm{a}_t\right)\equiv r\left(\mathrm{s}_t,\mathrm{a}_t\right)+E_{\mathrm{s}_{t+1}\sim p(\mathrm{s}_{t+1}\mid\mathrm{s}_t,\mathrm{a}_t)}\left[V\left(\mathrm{s}_{t+1}\right)\right]$$
$$V\left(\mathrm{s}_t\right)\equiv\log\int_{\mathcal{A}}\exp\left(Q\left(\mathrm{s}_t,\mathrm{a}_t\right)\right)d\mathrm{a}_t$$

$$\rightarrow\pi\left(\mathrm{a}_t\mid\mathrm{s}_t\right)=\exp\left(Q\left(\mathrm{s}_t,\mathrm{a}_t\right)-V\left(\mathrm{s}_t\right)\right)$$

[23] Sergey Levine. "Reinforcement learning and control as probabilistic inference: Tutorial and review". In: *arXiv preprint arXiv:1805.00909* (2018).

Table of Contents

Agent modelling with maximum entropy objective[24]

- each agent pursues the maximal cumulative reward

$$\max \quad \eta^i(\pi_\theta) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R^i(s_t, a_t^i, a_t^{-i})\right], \qquad (2)$$

  with actions (a_t^i, a_t^{-i}) sampled from policy $(\pi^i_{\theta^i}, \pi^{-i}_{\theta^{-i}})$
- a strategy profile $(\pi^{1*}, \ldots, \pi^{n*})$ reaches optimum when:

$$\begin{aligned} &\mathbb{E}_{s\sim p_s, a_t^{i*}\sim\pi^{i*}, a_t^{-i*}\sim\pi^{-i*}}\left[\sum_{t=1}^{\infty} \gamma^t R^i(s_t, a_t^{i*}, a_t^{-i*})\right] \\ &\geq \mathbb{E}_{s\sim p_s, a_t^{i}\sim\pi^{i}, a_t^{-i}\sim\pi^{-i}}\left[\sum_{t=1}^{\infty} \gamma^t R^i(s_t, a_t^{i}, a_t^{-i})\right] \qquad (3) \\ &\forall \pi \in \Pi, i \in (1 \ldots n), \end{aligned}$$

  where $\pi = (\pi^i, \pi^{-i})$ and agent i's optimal policy is π^{i*}

[24] Zheng Tian et al. "A regularized opponent model with maximum entropy objective". In: *IJCAI* (2019).

Agent modelling with maximum entropy objective

- a lower bound on the likelihood of optimality of agent i:

$$\log P(\mathcal{O}^i_{1:T} = 1 | \mathcal{O}^{-i}_{1:T} = 1) \geq \sum_t \mathbb{E}_{(s_t, a^i_t, a^{-i}_t) \sim q} [R^i(s_t, a^i_t, a^{-i}_t)$$

$$+ H(\pi(a^i_t | s_t, a^{-i}_t)) - D_{\mathrm{KL}}(\rho(a^{-i}_t | s_t) || P(a^{-i}_t | s_t))] \tag{4}$$

$$= \sum_t \mathbb{E}_{s_t} [\underbrace{\mathbb{E}_{a^i_t \sim \pi, a^{-i}_t \sim \rho} [R^i(s_t, a^i_t, a^{-i}_t) + H(\pi(a^i_t | s_t, a^{-i}_t))]}_{\text{MEO}}$$

$$- \underbrace{\mathbb{E}_{a^{-i}_t \sim \rho} [D_{\mathrm{KL}}(\rho(a^{-i}_t | s_t) || P(a^{-i}_t | s_t))]]}_{\text{Regulariser of } \rho}. \tag{5}$$

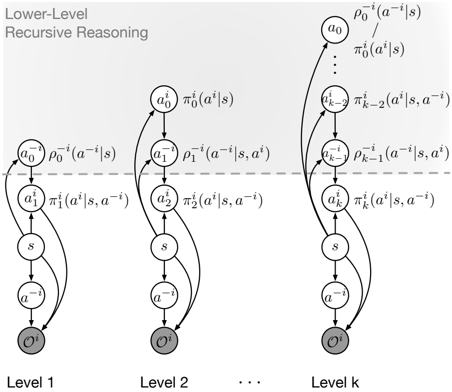
  - $\rho(a^{-i}_t | s_t, o^{-i}_t = 1)$ is agent i's opponent model
  - $\pi(a^i_t | s_t, a^{-i}_t, o^i_t = 1, o^{-i}_t = 1)$ is the agent i's conditional policy at optimum $(o^i_t = o^{-i}_t = 1)$ and
  - $P(a^{-i}_t | s_t, o^{-i}_t = 1)$ is the prior of opponent model
  - the prior $P(a^{-i}_t | s_t, o^{-i}_t = 1)$ is estimated empirically
  - we drop $(o^i_t = 1, o^{-i}_t = 1)$ in π, ρ and $P(a^{-i}_t | s_t)$

Recursive reasoning[25]

- Example: in the "beauty contest" game, players are asked to pick numbers from 0 to 100, and the player whose number is closest to 2/3 of the average wins a prize

[25] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. "A cognitive hierarchy model of games". In: *The Quarterly Journal of Economics* 119.3 (2004), pp. 861–898.

Multi-agent generalized recursive reasoning[26]



- unobserved opponent policies are approximated by ρ^{-i}
- agent i recursively reasons about opponents (grey area)
- in the recursion, agents with higher-level beliefs take the best response to the lower-level thinkers' actions.

[26]Ying Wen et al. "Modelling Bounded Rationality in Multi-Agent Interactions by Generalized Recursive Reasoning". In: *IJCAI* (2020).

Table of Contents

Remarks

- AI: reasoning under uncertainty
- a learning system is an information system
- pattern recognition and decision making problems (include multiagent learning) can be unified and modelled by probabilistic inference such as Variational Bayes (e.g., active- inference or inference-as-control)
- as such information theory can be directly utilised
- however, we call for new research on information theory that deals with "intrinsic information exchange"

References I

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. “Variational inference: A review for statisticians”. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. “A cognitive hierarchy model of games”. In: *The Quarterly Journal of Economics* 119.3 (2004), pp. 861–898.

References II

Karl Friston. “The free-energy principle: a unified brain theory?” In: *Nature reviews neuroscience* 11.2 (2010), pp. 127–138.

Michael I Jordan et al. “An introduction to variational methods for graphical models”. In: *Machine learning* 37.2 (1999), pp. 183–233.

Rudolph Emil Kalman. “A new approach to linear filtering and prediction problems”. In: (1960).

Sergey Levine. “Reinforcement learning and control as probabilistic inference: Tutorial and review”. In: *arXiv preprint arXiv:1805.00909* (2018).

References III

- Minne Li et al. "Joint Perception and Control as Inference with an Object-based Implementation". In: (2020).
- David Marr. "Vision: A computational investigation into the human representation and processing of visual information". In: *Inc., New York, NY* 2.4.2 (1982).
- Beren Millidge et al. "On the relationship between active inference and control as inference". In: *International Workshop on Active Inference*. Springer. 2020, pp. 3–11.

References IV

Beren Millidge, Alexander Tschantz, and Christopher L Buckley. "Whence the expected free energy?" In: *Neural Computation* 33.2 (2021), pp. 447–482.

Ivan Petrovitch Pavlov and William Gantt. "Lectures on conditioned reflexes: Twenty-five years of objective study of the higher nervous activity (behaviour) of animals.". In: (1928).

Eliezer Sternberg. *NeuroLogic: The Brain's Hidden Rationale Behind Our Irrational Behavior*. Vintage, 2016.

Zheng Tian et al. "A regularized opponent model with maximum entropy objective". In: *IJCAI* (2019).

References V

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

Ying Wen et al. "Modelling Bounded Rationality in Multi-Agent Interactions by Generalized Recursive Reasoning". In: *IJCAI* (2020).