

A conversion between utility and information

Pedro A. Ortega

Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, UK
peortega@dcc.uchile.cl

Daniel A. Braun

Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, UK
dab54@cam.ac.uk

Abstract

Rewards typically express desirabilities or preferences over a set of alternatives. Here we propose that rewards can be defined for any probability distribution based on three desiderata, namely that rewards should be real-valued, additive and order-preserving, where the latter implies that more probable events should also be more desirable. Our main result states that rewards are then uniquely determined by the negative information content. To analyze stochastic processes, we define the utility of a realization as its reward rate. Under this interpretation, we show that the expected utility of a stochastic process is its negative entropy rate. Furthermore, we apply our results to analyze agent-environment interactions. We show that the expected utility that will actually be achieved by the agent is given by the negative cross-entropy from the input-output (I/O) distribution of the coupled interaction system and the agent's I/O distribution. Thus, our results allow for an information-theoretic interpretation of the notion of utility and the characterization of agent-environment interactions in terms of entropy dynamics.

Keywords: Behavior, Utility, Entropy, Information

Introduction

Purposeful behavior typically occurs when an agent exhibits specific preferences over different states of the environment. Mathematically, these preferences can be formalized by the concept of a utility function that assigns a numerical value to each possible state such that states with higher utility correspond to states that are more desirable [Fishburn, 1982]. Behavior can then be understood as the attempt to increase one's utility. Accordingly, utility functions can be measured experimentally by observing an agent choosing between different options, as this way its preferences are revealed. Mathematical models of rational agency that are based on the notion of utility have been widely applied in behavioral economics, biology and artificial intelligence research [Russell and Norvig, 1995]. Typically, such rational agent models assume a distinct reward signal (or cost) that an agent is explicitly trying to optimize.

However, as an observer we might even attribute purposefulness to a system that does not have an explicit

reward signal, because the dynamics of the system itself reveal a preference structure, namely the preference over all possible paths through history. Since in most systems not all of the histories are equally likely, we might say that some histories are more probable than others because they are more desirable from the point of view of the system. Similarly, if we regard all possible interactions between a system and its environment, the behavior of the system can be conceived as a drive to generate desirable histories. This imposes a conceptual link between the probability of a history happening and the desirability of that history. In terms of agent design, the intuitive rationale is that agents should act in a way such that more desired histories are more probable. The same holds of course for the environment. Consequently, a competition arises between the agent and the environment, where both participants try to drive the dynamics of their interactions to their respective desired histories. In the following we want to show that this competition can be quantitatively assessed based on the entropy dynamics that govern the interactions between agent and environment.

Preliminaries

We introduce the following notation. A set is denoted by a calligraphic letter like \mathcal{X} and consists of *elements* or *symbols*. *Strings* are finite concatenations of symbols and *sequences* are infinite concatenations. The empty string is denoted by ϵ . \mathcal{X}^n denotes the set of strings of length n based on \mathcal{X} , and $\mathcal{X}^* \equiv \bigcup_{n \geq 0} \mathcal{X}^n$ is the set of finite strings. Furthermore, $\mathcal{X}^\infty \equiv \{x_1 x_2 \dots | x_i \in \mathcal{X} \text{ for all } i = 1, 2, \dots\}$ is defined as the set of one-way infinite sequences based on \mathcal{X} . For substrings, the following shorthand notation is used: a string that runs from index i to k is written as $x_{i:k} \equiv x_i x_{i+1} \dots x_{k-1} x_k$. Similarly, $x_{<i} \equiv x_1 x_2 \dots x_i$ is a string starting from the first index. By convention, $x_{i:j} \equiv \epsilon$ if $i > j$. All proofs can be found in the appendix.

Rewards

In order to derive utility functions for stochastic processes over finite alphabets, we construct a utility function from an auxiliary function that measures the de-

desirability of events, i.e. such that we can assign desirability values to every finite interval in a realization of the process. We call this auxiliary function the reward function. We impose three desiderata on a reward function. First, we want rewards to be mappings from events to reals numbers that indicate the degree of desirability of the events. Second, the reward of a joint event should be obtained by summing up the reward of the sub-events. For example, the “reward of drinking coffee and eating a croissant” should equal “the reward of drinking coffee” plus the “reward of having a croissant given the reward of drinking coffee”¹. This is the additivity requirement of the reward function. The last requirement that we impose for the reward function should capture the intuition suggested in the introduction, namely that more desirable events should also be more probable events given the expectations of the system. This is the consistency requirement.

We start out from a *probability space* $(\Omega, \mathcal{F}, \mathbf{Pr})$, where Ω is the sample space, \mathcal{F} is a dense σ -algebra and \mathbf{Pr} is a probability measure over \mathcal{F} . In this section, we use lowercase letters like x, y, z to denote the elements of the σ -algebra \mathcal{F} . Given a set \mathcal{X} , its complement is denote by complement \mathcal{X}^c and its powerset by $\mathcal{P}(\mathcal{X})$. The three desiderata can then be summarized as follows:

Definition 1 (Reward). Let $S = (\Omega, \mathcal{F}, \mathbf{Pr})$ be a probability space. A function \mathbf{r} is a *reward function* for S iff it has the following three properties:

1. *Real-valued*: for all $x, y \in \mathcal{F}$,

$$\mathbf{r}(x|y) \in \mathbb{R};$$

2. *Additivity*: for all $x, y, z \in \mathcal{F}$,

$$\mathbf{r}(x, y|z) = \mathbf{r}(x|z) + \mathbf{r}(y|x, z);$$

3. *Consistent*: for all $x, y, u, v \in \mathcal{F}$,

$$\mathbf{Pr}(x|u) > \mathbf{Pr}(y|v) \iff \mathbf{r}(x|u) > \mathbf{r}(y|v).$$

Furthermore, the unconditional reward is defined as $\mathbf{r}(x) \equiv \mathbf{r}(x|\Omega)$ for all $x \in \mathcal{F}$.

The following theorem shows that these three desiderata enforce a strict mapping rewards and probabilities. The only function that can express such a relationship is the logarithm.

Theorem 1. Let $S = (\Omega, \mathcal{F}, \mathbf{Pr})$ be a probability space. Then, a function \mathbf{r} is a reward function for S iff for all $x, y \in \mathcal{F}$

$$\mathbf{r}(x|y) = k \log \mathbf{Pr}(x|y),$$

where $k > 0$ is an arbitrary constant.

Notice that the constant k in the expression $\mathbf{r}(x|y) = k \log \mathbf{Pr}(x|y)$ merely determines the units in which we choose to measure rewards. Thus, the reward function

¹Note that the additivity property does not imply that the reward for two coffees is simply twice the reward for one coffee, as the reward for the second coffee will be conditioned on having had a first coffee already.

\mathbf{r} for a probability space $(\Omega, \mathcal{F}, \mathbf{Pr})$ is essentially unique. As a convention, we will assume natural logarithms and set the constant to $k = 1$, i.e. $\mathbf{r}(x|y) = \ln \mathbf{Pr}(x|y)$.

This result establishes a connection to information theory. It is immediately clear that the reward of an event is nothing more than its negative information content: the quantity $h(x) = -\mathbf{r}(x)$ is the Shannon information content of $x \in \mathcal{F}$ measured in *nats* [MacKay, 2003]. This means that we can interpret rewards as “negative surprise values”, and that “surprise values” constitute losses.

Proposition 1. Let \mathbf{r} be a reward function over a probability space $(\Omega, \mathcal{F}, \mathbf{Pr})$. Then, it has the following properties:

i. Let $x, y \in \mathcal{F}$. Then

$$-\infty = \mathbf{r}(\emptyset) \leq \mathbf{r}(x|y) \leq \mathbf{r}(\Omega) = 0.$$

ii. Let $x \in \mathcal{F}$ be an event. Then,

$$e^{\mathbf{r}(x^c)} = 1 - e^{\mathbf{r}(x)}.$$

iii. Let $z_1, z_2, \dots \in \mathcal{F}$ be a sequence of disjoint events with rewards $\mathbf{r}(z_1), \mathbf{r}(z_2), \dots$ and let $x = \bigcup_i z_i$. Then

$$e^{\mathbf{r}(x)} = \sum_i e^{\mathbf{r}(z_i)}.$$

The proof of this proposition is trivial and left to the reader. The first part sets the bounds for the values of rewards, and the two latter explain how to construct the rewards of events from known rewards using complement and countable union of disjoint events.

At a first glance, the fact that rewards take on complicated non-positive values might seem unnatural, as in many applications one would like to use numerical values drawn from arbitrary real intervals. Fortunately, given numerical values representing the desirabilities of events, there is always an affine transformation that converts them into rewards.

Theorem 2. Let Ω be a countable set, and let $d : \Omega \rightarrow (-\infty, a]$ be a mapping. Then, for every $\alpha > 0$, there is a probability space $(\Omega, \mathcal{P}(\Omega), \mathbf{Pr})$ with reward function \mathbf{r} such that:

1. for all $\omega \in \Omega$,

$$\mathbf{r}(\{\omega\}) \equiv \alpha d(\omega) + \beta,$$

$$\text{where } \beta \equiv -\ln \left(\sum_{\omega' \in \Omega} e^{\alpha d(\omega')} \right);$$

2. and for all $\omega, \omega' \in \Omega$,

$$d(\omega) > d(\omega') \iff \mathbf{r}(\{\omega\}) > \mathbf{r}(\{\omega'\}).$$

Note that Theorem 2 implies that the probability $\mathbf{Pr}(x)$ of any event x in the σ -algebra $\mathcal{P}(\Omega)$ generated by Ω is given by

$$\mathbf{Pr}(x) = \frac{\sum_{\omega \in x} e^{\alpha d(\omega)}}{\sum_{\omega \in \Omega} e^{\alpha d(\omega)}}.$$

Note that for singletons $\{\omega\}$, $\mathbf{Pr}(\{\omega\})$ is the Gibbs measure with negative energy $d(\omega)$ and temperature $\propto \frac{1}{\alpha}$. It is due to this analogy that we call the quantity $\frac{1}{\alpha} > 0$ the *temperature parameter* of the transformation.

Utilities in Stochastic Processes

In this section, we consider a *stochastic process* \mathbf{Pr} over sequences $x_1x_2x_3\cdots$ in \mathcal{X}^∞ . We specify the process by assigning conditional probabilities $\mathbf{Pr}(x_t|x_{<t})$ to all finite strings $x_{<t} \in \mathcal{X}^*$. Note that the distribution $\mathbf{Pr}(x_{\leq t}) = \prod_{\tau=1}^t \mathbf{Pr}(x_\tau|x_{<\tau})$ for all $x_{\leq t} \in \mathcal{X}^*$ is normalized by construction. By the Kolmogorov extension theorem, it is guaranteed that there exists a unique probability space $S = (\mathcal{X}^\infty, \mathcal{F}, \mathbf{Pr})$. We therefore omit the reference to S and talk about the process \mathbf{Pr} .

The reward function \mathbf{r} derived in the previous section correctly expresses preference relations amongst different outcomes. However, in the context of random sequences, it has the downside that the reward of most sequences diverges. A sequence $x_1x_2x_3\cdots$ can be interpreted as a progressive refinement of a point event in \mathcal{F} , namely, the sequence of events $\epsilon \supset x_{\leq 1} \supset x_{\leq 2} \supset x_{\leq 3} \supset \cdots$. One can exploit the interpretation of the index as time to define a quantity that does not diverge. We define thus the utility as the reward rate of a sequence.

Definition 2 (Utility). Let \mathbf{r} be a reward function for the process \mathbf{Pr} . The utility of a string $x_{\leq t} \in \mathcal{X}^*$ is defined as

$$\mathbf{U}(x_{\leq t}) \equiv \frac{1}{t} \sum_{\tau=1}^t \mathbf{r}(x_\tau|x_{<\tau}),$$

and for a sequence $x = x_1x_2x_3\cdots \in \mathcal{X}^\infty$ it is defined as

$$\mathbf{U}(x) \equiv \lim_{t \rightarrow \infty} \mathbf{U}(x_{\leq t})$$

if this limit exists².

A utility function that is constructed according to Definition 2 has the following properties.

Proposition 2. Let \mathbf{U} be a utility function for a process \mathbf{Pr} . The following properties hold:

i. For all $x = x_1x_2\cdots \in \mathcal{X}^\infty$ and all $t, k \in \mathbb{N}$,

$$-\infty = \mathbf{U}(\lambda) \leq \mathbf{U}(x_{\leq t}) \leq \mathbf{U}(\epsilon) = 0,$$

where λ is any impossible string/sequence.

ii. For all $x_{\leq t} \in \mathcal{X}^*$,

$$\mathbf{Pr}(x_{\leq t}) = \exp\left(t \cdot \mathbf{U}(x_{\leq t})\right).$$

iii. For any $t \in \mathbb{N}$,

$$\mathbf{E}[\mathbf{U}(x_{\leq t})] = -\frac{1}{t} \mathbf{H}[\mathbf{Pr}(x_{\leq t})],$$

where \mathbf{H} is the entropy functional (see the appendix).

Part (i) provides trivial bounds on the utilities that directly carry over from the bounds on rewards. Part (ii) shows how the utility of a sequence determines its probability. Part (iii) implies that the expected utility of an interaction sequence is just its negative entropy rate.

²Strictly speaking, one could define the upper and lower rate $\mathbf{U}^+(x) \equiv \limsup_{t \rightarrow \infty} \mathbf{U}(x_{\leq t})$ and $\mathbf{U}^-(x) \equiv \liminf_{t \rightarrow \infty} \mathbf{U}(x_{\leq t})$ respectively, but we avoid this distinction for simplicity.

Utility in Coupled I/O systems

Let \mathcal{O} and \mathcal{A} be two finite sets, the first being the *set of observations* and the second being the *set of actions*. Using \mathcal{A} and \mathcal{O} , a set of interaction sequences is constructed. Define the *set of interactions* as $\mathcal{Z} \equiv \mathcal{A} \times \mathcal{O}$. A pair $(a, o) \in \mathcal{Z}$ is called an *interaction*. We underline symbols to glue them together as in $\underline{ao}_{<t} = a_1o_1a_2o_2\cdots a_t o_t$.

An *I/O system* \mathbf{Pr} is a probability distribution over interaction sequences \mathcal{Z}^∞ . \mathbf{Pr} is uniquely determined by the conditional probabilities

$$\mathbf{Pr}(a_t|\underline{ao}_{<t}), \quad \mathbf{Pr}(o_t|\underline{ao}_{<t}a_t)$$

for each $\underline{ao}_{<t} \in \mathcal{Z}^*$. However, the semantics of the probability distribution \mathbf{Pr} are only fully defined once it is coupled to another system. Note that an I/O system is formally equivalent to a stochastic process; hence one can construct a reward function \mathbf{r} for \mathbf{Pr} .

Let \mathbf{P}, \mathbf{Q} be two I/O systems. An *interaction system* (\mathbf{P}, \mathbf{Q}) defines a *generative distribution* \mathbf{G} that describes the probabilities that actually govern the I/O stream once the two systems are coupled. \mathbf{G} is specified by the equations

$$\begin{aligned} \mathbf{G}(a_t|\underline{ao}_{<t}) &= \mathbf{P}(a_t|\underline{ao}_{<t}) \\ \mathbf{G}(o_t|\underline{ao}_{<t}a_t) &= \mathbf{Q}(o_t|\underline{ao}_{<t}a_t) \end{aligned}$$

valid for all $\underline{ao}_t \in \mathcal{Z}^*$. Here, \mathbf{G} is a stochastic process over \mathcal{Z}^∞ that models the true probability distribution over interaction sequences that arises by coupling two systems through their I/O streams. More specifically, for the system \mathbf{P} , $\mathbf{P}(a_t|\underline{ao}_{<t})$ is the probability of producing action $a_t \in \mathcal{A}$ given history $\underline{ao}_{<t}$ and $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$ is the predicted probability of the observation $o_t \in \mathcal{O}$ given history $\underline{ao}_{<t}a_t$. Hence, for \mathbf{P} , the sequence $o_1o_2\cdots$ is its input stream and the sequence $a_1a_2\cdots$ is its output stream. In contrast, the roles of actions and observations are reversed in the case of the system \mathbf{Q} . This model of interaction is very general in that it can accommodate many specific regimes of interaction. By convention, we call the system \mathbf{P} the *agent* and the system \mathbf{Q} the *environment*.

In the following we are interested in understanding the actual utilities that can be achieved by an agent \mathbf{P} once coupled to a particular environment \mathbf{Q} . Accordingly, we will compute expectations over functions of interaction sequences with respect to \mathbf{G} , since the generative distribution \mathbf{G} describes the actual interaction statistics of the two coupled I/O systems.

Theorem 3. Let (\mathbf{P}, \mathbf{Q}) be an interaction system. The expected rewards of \mathbf{G} , \mathbf{P} and \mathbf{Q} for the first t interactions are given by

$$\begin{aligned} \mathbf{E}[\mathbf{r}_{\mathbf{G}}(\underline{ao}_{\leq t})] &= -\mathbf{H}[\mathbf{P}(a_{\leq t}|o_{<t})] - \mathbf{H}[\mathbf{Q}(o_{\leq t}|a_{\leq t})], \\ \mathbf{E}[\mathbf{r}_{\mathbf{P}}(\underline{ao}_{\leq t})] &= -\mathbf{H}[\mathbf{P}(a_{\leq t}|o_{<t})] - \mathbf{H}[\mathbf{Q}(o_{\leq t}|a_{\leq t})] \\ &\quad - \mathbf{KL}[\mathbf{Q}(o_{\leq t}|a_{\leq t}) \parallel \mathbf{P}(o_{\leq t}|a_{\leq t})], \\ \mathbf{E}[\mathbf{r}_{\mathbf{Q}}(\underline{ao}_{\leq t})] &= -\mathbf{H}[\mathbf{P}(a_{\leq t}|o_{<t})] - \mathbf{H}[\mathbf{Q}(o_{\leq t}|a_{\leq t})] \\ &\quad - \mathbf{KL}[\mathbf{P}(a_{\leq t}|o_{<t}) \parallel \mathbf{Q}(a_{\leq t}|o_{<t})], \end{aligned}$$

where \mathbf{r}_G , \mathbf{r}_P and \mathbf{r}_Q are the reward functions for G , P and Q respectively. Note that H and KL are the entropy and the relative entropy functionals as defined in the appendix.

Accordingly, the interaction system's expected reward is given by the negative sum of the entropies produced by the agent's action generation probabilities and the environment's observation generation probabilities. The agent's (actual) expected reward is given by the negative cross-entropy between the generative distribution G and the agent's distribution P . The discrepancy between the agent's and the interaction system's expected reward is given by the relative entropy between the two probability distributions. Since the relative entropy is positive, one has $\mathbf{E}[\mathbf{r}_G(\underline{a}o_{\leq t})] \geq \mathbf{E}[\mathbf{r}_P(\underline{a}o_{\leq t})]$. This term implies that the better the environment is "modeled" by the agent, the better its performance will be. In other words: the agent has to recognize the structure of the environment to be able to exploit it. The designer can directly increase the agent's expected performance by controlling the first and the last term. The middle term is determined by the environment and only indirectly controllable. Importantly, the terms are in general coupled and not independent: changing one might affect another. For example, the first term suggests that less stochastic policies improve performance, which is oftentimes the case. However, in the case of a game with mixed Nash equilibria the overall reward can increase for a stochastic policy, which means that the first term is compensated for by the third term. Given the expected rewards, we can easily calculate the expected utilities in terms of entropy rates.

Corollary 1. *Let (P, Q) be an interaction system. The expected utilities of G , P and Q are given by*

$$\begin{aligned}\mathbf{E}[\mathbf{U}_G] &= \mathbf{G}\mathbf{U}_P + \mathbf{G}\mathbf{U}_Q \\ \mathbf{E}[\mathbf{U}_P] &= \mathbf{G}\mathbf{U}_P + \mathbf{G}\mathbf{U}_Q + \mathbf{P}\mathbf{U}_P \\ \mathbf{E}[\mathbf{U}_Q] &= \mathbf{G}\mathbf{U}_P + \mathbf{G}\mathbf{U}_Q + \mathbf{P}\mathbf{U}_Q\end{aligned}$$

where $\mathbf{G}\mathbf{U}_P$, $\mathbf{G}\mathbf{U}_Q$ and $\mathbf{P}\mathbf{U}_P$ are entropy rates defined as

$$\begin{aligned}\mathbf{G}\mathbf{U}_P &\equiv -\frac{1}{t} \sum_{\tau=1}^t \mathbf{H}[\mathbf{P}(a_\tau | \underline{a}o_{<\tau})] \\ \mathbf{P}\mathbf{U}_P &\equiv -\frac{1}{t} \sum_{\tau=1}^t \mathbf{KL}[\mathbf{Q}(o_\tau | \underline{a}o_{<\tau} a_\tau) \| \mathbf{P}(o_\tau | \underline{a}o_{<\tau} a_\tau)] \\ \mathbf{G}\mathbf{U}_Q &\equiv -\frac{1}{t} \sum_{\tau=1}^t \mathbf{H}[\mathbf{Q}(o_\tau | \underline{a}o_{<\tau} a_\tau)] \\ \mathbf{P}\mathbf{U}_Q &\equiv -\frac{1}{t} \sum_{\tau=1}^t \mathbf{KL}[\mathbf{P}(a_\tau | \underline{a}o_{<\tau}) \| \mathbf{Q}(a_\tau | \underline{a}o_{<\tau})].\end{aligned}$$

This result is easily obtained by dividing the quantities in Theorem 3 by t and then applying the chain rule for entropies to break the rewards over full sequences into instantaneous rewards. Note that $\mathbf{G}\mathbf{U}_P$, $\mathbf{G}\mathbf{U}_Q$ are the contributions to the utility due the generation of interactions, and $\mathbf{P}\mathbf{U}_P$, $\mathbf{P}\mathbf{U}_Q$ are the contributions to the utility due to the prediction of interactions.

Examples

One of the most interesting aspects of the information-theoretic formulation of utility is that it can be applied both to control problems (where an agent acts in a non-adaptive environment) and to game theoretic problems (where two possibly adaptive agents interact). In the following we apply the proposed utility measures to two simple toy examples from these two areas. In the first example, an adaptive agent interacts with a biased coin (the non-adaptive agent) and tries to predict the next outcome of the coin toss, which is either 'Head' (H) or 'Tail' (T). In the second example two adaptive agents interact playing the *matching pennies* game. One player has to match her action with the other player (HH or TT), while the other player has to unmatched (TH or HT). All agents have the same sets of possible observations and actions which are the binary sets $\mathcal{O} = \{H, T\}$ and $\mathcal{A} = \{H, T\}$.

Example 1. The non-adaptive agent is a biased coin. Accordingly, the coin's action probability is given by its bias and was set to $\mathbf{Q}(o = H) = 0.9$. The coin does not have any biased expectations about its observations, so we set $\mathbf{Q}(a = H) = 0.5$. The adaptive agent is given by the Laplace agent whose expectations over observed coin tosses follows the predictive distribution $\mathbf{P}(o = H | t, n) = (n + 1)/(t + 2)$, where t is the number of coin tosses observed so far and n is the number of observed Heads. Based on this estimator the Laplace agent chooses its action deterministically according to $\mathbf{P}(a = H | t, n) = \Theta(\frac{n+1}{t+2} - \frac{1}{2})$, where $\Theta(\cdot)$ is the Heaviside step function. From these distributions the full probability over interaction sequences can be computed. Figure 1A shows the entropy dynamics for a typical single run. The Laplace agent learns the distribution of the coin tosses, i.e. the KL decreases to zero. The negative cross-entropy stabilizes at the value of the observation entropy that cannot be further reduced. The entropy dynamics of the coin do not show any modulation.

Example 2. The two agents are modeled based on *smooth fictitious play* [Fudenberg and Kreps, 1993]. Both players keep count of the empirical frequencies of Head and Tail respectively. Therefore, each player i stores the quantities $\kappa_i^{(1)} = n_i$ and $\kappa_i^{(2)} = t - n_i$ where t is the number of moves observed so far, n_1 is the number of Heads observed by Player 1 and n_2 is the number of Heads observed by Player 2. The probability distributions $\mathbf{P}(o = H | t, n_1) = \gamma_1$ and $\mathbf{Q}(a = H | t, n_2) = \gamma_2$ over inputs is given by these empirical frequencies through $\gamma_i = \kappa_i / \sum_j \kappa_j$. The action probabilities are computed according to a sigmoid best-response function $\mathbf{P}(a = H | t, n_1) = 1/(1 + \exp(-\alpha(\gamma_1 - 0.5)))$, and $\mathbf{Q}(o = H | t, n_2) = 1/(1 + \exp(-\alpha(0.5 - \gamma_2)))$ respectively in case of Player 2 that has to unmatched. This game has a well-known equilibrium solution that is a mixed strategy Nash equilibrium where both players act

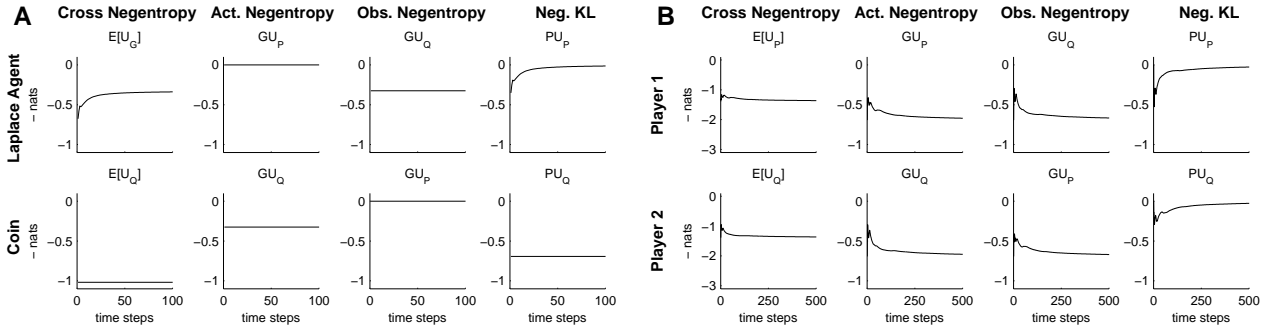


Figure 1: (A) Entropy dynamics of a Laplace agent interacting with a coin of bias 0.9. The Laplace agent learns to predict the coin’s behavior as can be seen in the decrease of the KL-divergence and the cross entropy. Since the Laplace agent acts deterministically its action entropy is always zero. Its observation entropy equals the action entropy of the coin. The coin does not change its behavior, which can be seen from the flat entropy curves. (B) Entropy dynamics of two adaptive agents playing matching pennies. Both agents follow smooth fictitious play. They converge to uniform random policies, which means that their action negentropies converge to $\log(2)$. Both agents learn the probability distribution of the other agent, as can be seen in the decrease of the KL-divergences.

randomly. Both action and observation entropies converge to the value $\log(2)$. Interestingly, the information-theoretic utility as computed by the cross-entropy takes the action entropy into account. Compare Figure 1B.

Conclusion

Based on three simple desiderata we propose that rewards can be measured in terms of information content and that, consequently, the entropy satisfies properties characteristic of a utility function. Previous theoretical studies have reported structural similarities between entropy and utility functions, see e.g. [Candea et al., 2001], and recently, relative entropy has even been proposed as a measure of utility in control systems [Todorov, 2009, Kappen et al., 2009, Ortega and Braun, 2008]. The contribution of this paper is to derive axiomatically a precise relation between rewards and information value and to apply it to coupled I/O systems.

The utility functions that we have derived can be conceptualized as *path utilities*, because they assign a utility value to an entire history. This is very similar to the path integral formulation in quantum mechanics where the utility of a path is determined by the classic action integral and the probability of a path is also obtain by taking the exponential of this ‘utility’ [Feynman and Hibbs, 1965]. In particular, we obtain the (cumulative time-averaged) cross entropy as a utility function when an agent is coupled to an environment. This utility function not only takes into account the KL-divergence as a measure of learning, but also the action entropy. This is interesting, because in most control problems controllers are designed to be deterministic (e.g. optimal control theory) in response to a known and stationary environment. If, however, the environment is not stationary and in fact adaptive as well, then it is a well-known result from game theory

that optimal strategies might be randomized. The utility function that we are proposing might indeed allow quantifying a trade-off between reducing the KL and reducing the action entropy. In the future it will therefore be interesting to investigate this utility function in more complex interaction systems.

Appendix

Entropy functionals

Entropy: Let \mathbf{Pr} be a probability distribution over $\mathcal{X} \times \mathcal{Y}$. Define the (average conditional) entropy [Shannon, 1948] as

$$\mathbf{H}[\mathbf{Pr}(x|y)] \equiv - \sum_{x,y} \mathbf{Pr}(x,y) \ln \mathbf{Pr}(x|y).$$

Relative Entropy: Let \mathbf{Pr}_1 and \mathbf{Pr}_2 be two probability distributions over $\mathcal{X} \times \mathcal{Y}$. Define the (average conditional) relative entropy [Kullback and Leibler, 1951] as

$$\mathbf{KL}[\mathbf{Pr}_1(x|y) \parallel \mathbf{Pr}_2(x|y)] \equiv \sum_{x,y} \mathbf{Pr}_1(x,y) \ln \frac{\mathbf{Pr}_1(x|y)}{\mathbf{Pr}_2(x|y)}.$$

Proof of Theorem 1

Proof. Let the function g be such that $g(\mathbf{Pr}(x)) = \mathbf{r}(x)$. Let $x_1, x_2, \dots, x_n \in \mathcal{F}$ be a sequence of events, such that $\mathbf{Pr}(x_1) = \mathbf{Pr}(x_i|x_{<i}) > 0$ for all $i = 2, \dots, n$. We have $\mathbf{Pr}(x_1, \dots, x_n) = \prod_i \mathbf{Pr}(x_i|x_{<i}) = \mathbf{Pr}(x_1)^n$. Since $\mathbf{Pr}(x) > \mathbf{Pr}(x') \Leftrightarrow \mathbf{r}(x) > \mathbf{r}(x')$ for any $x, x' \in \mathcal{F}$, then $\mathbf{Pr}(x) = \mathbf{Pr}(x') \Leftrightarrow \mathbf{r}(x) = \mathbf{r}(x')$, and thus $\mathbf{Pr}(x_1) = \mathbf{Pr}(x_i|x_{<i}) \Leftrightarrow \mathbf{r}(x_1) = \mathbf{r}(x_i|x_{<i})$ for all $i = 2, \dots, n$. This means, $\mathbf{r}(x_1, \dots, x_n) = n\mathbf{r}(x_1)$. But $g(\mathbf{Pr}(x_1, \dots, x_n)) = \mathbf{r}(x_1, \dots, x_n)$, and hence $g(\mathbf{Pr}(x_1)^n) = n\mathbf{r}(x_1)$. Similarly, for a second sequence of events $y_1, y_2, \dots, y_m \in \mathcal{F}$ with $\mathbf{Pr}(y_1) = \mathbf{Pr}(y_i|y_{<i}) > 0$ for all $i = 1, \dots, m$, we have $g(\mathbf{Pr}(y_1)^m) = m\mathbf{r}(y_1)$.

The rest of the argument parallels Shannon's entropy theorem [Shannon, 1948]. Define $p = \Pr(x_1)$ and $q = \Pr(y_1)$. Choose n arbitrarily high to satisfy $q^m \leq p^n < q^{m+1}$. Taking the logarithm, and dividing by $n \log q$ one obtains

$$\frac{m}{n} \leq \frac{\log p}{\log q} < \frac{m}{n} + \frac{1}{n} \Leftrightarrow \left| \frac{m}{n} - \frac{\log p}{\log q} \right| < \varepsilon,$$

where $\varepsilon > 0$ is arbitrarily small. Similarly, using $g(p^n) = n g(p)$ and the monotonicity of g , we can write $m g(q) \leq n g(p) < (m+1) g(q)$ and thus

$$\frac{m}{n} \leq \frac{g(p)}{g(q)} < \frac{m}{n} + \frac{1}{n} \Leftrightarrow \left| \frac{m}{n} - \frac{g(p)}{g(q)} \right| < \varepsilon,$$

where $\varepsilon > 0$ is arbitrarily small. Combining these two inequalities, one gets

$$\left| \frac{\log p}{\log q} - \frac{g(p)}{g(q)} \right| < 2\varepsilon,$$

which, fixing q , gives $\mathbf{r}(p) = g(p) = k \log p$, where $k > 0$. This holds for any $x_1 \in \mathcal{F}$ with $\Pr(x_1) > 0$. \square

Proof of Theorem 2

Proof. For all $\omega, \omega' \in \Omega$, $d(\omega) > d(\omega') \Leftrightarrow \alpha d(\omega) + \beta > \alpha d(\omega') + \beta \Leftrightarrow \mathbf{r}(\{\omega\}) > \mathbf{r}(\{\omega'\})$ because the affine transformation is positive. Now, the induced probability over $\mathcal{P}(\Omega)$ has atoms $\{\omega\}$ with probabilities $\Pr(\{\omega\}) = e^{\mathbf{r}(\{\omega\})} \geq 0$ and is normalized:

$$\sum_{\omega \in \Omega} e^{\mathbf{r}(\{\omega\})} = \sum_{\omega \in \Omega} e^{\alpha d(\{\omega\}) + \beta} = \frac{\sum_{\omega \in \Omega} e^{\alpha d(\omega)}}{\sum_{\omega \in \Omega} e^{\alpha d(\omega)}} = 1.$$

Since knowing $\Pr(\{\omega\})$ for all $\omega \in \Omega$ determines the measure for the whole field $\mathcal{P}(\Omega)$, $(\Omega, \mathcal{P}(\Omega), \Pr)$ is a probability space. \square

Proof of Proposition 2

Proof. (i) Since $-\infty < \mathbf{r}(x_\tau | x_{<\tau}) \leq 0$ for all τ , then $-\infty < \frac{1}{t} \sum_{\tau=1}^t \mathbf{r}(x_\tau | x_{<\tau}) = \mathbf{U}(x_{\leq t}) \leq 0$ for all t . (ii) Write $\Pr(x_{\leq t})$ as

$$\begin{aligned} \Pr(x_{\leq t}) &= \prod_{\tau=1}^t \Pr(x_\tau | x_{<\tau}) = \prod_{\tau=1}^t \exp(\mathbf{r}(x_\tau | x_{<\tau})) \\ &= \exp\left(\sum_{\tau=1}^t \mathbf{r}(x_\tau | x_{<\tau})\right) = \exp(t \cdot \mathbf{U}(x_{\leq t})). \end{aligned}$$

(iii) $\mathbf{E}[\mathbf{U}(x_{\leq t})] = \sum_{x_{\leq t}} \Pr(x_{\leq t}) \mathbf{U}(x_{\leq t}) = \sum_{x_{\leq t}} \Pr(x_{\leq t}) \frac{1}{t} \mathbf{r}(x_{\leq t}) = -\frac{1}{t} \mathbf{H}[\Pr(x_{\leq t})]$, where we have applied (ii) in the second equality and $\mathbf{r}(\cdot) = \ln(\Pr(\cdot))$ in the third equality. \square

Proof of Theorem 3

Proof. This proof is done by straightforward calculation. First note that

$$\begin{aligned} \mathbf{G}(\underline{a} \underline{o}_{\leq t}) &= \prod_{\tau=1}^t \mathbf{P}(a_\tau | \underline{a} \underline{o}_{<\tau}) \mathbf{Q}(o_\tau | \underline{a} \underline{o}_{<\tau} a_\tau) \\ &= \mathbf{P}(a_{\leq t} | o_{<t}) \mathbf{Q}(o_{\leq t} | a_{\leq t}), \end{aligned}$$

which is obtained by applying multiple times the chain rule for probabilities and noting that the probability of a symbol is fully determined by the previous symbols. Similarly $\mathbf{P}(\underline{a} \underline{o}_{\leq t}) = \mathbf{P}(a_{\leq t} | o_{<t}) \mathbf{P}(o_{\leq t} | a_{\leq t})$ is obtained. We calculate here $\mathbf{E}[\mathbf{r}_{\mathbf{P}}(\underline{a} \underline{o}_{\leq t})]$. The calculation for $\mathbf{E}[\mathbf{r}_{\mathbf{G}}(\underline{a} \underline{o}_{\leq t})]$ and $\mathbf{E}[\mathbf{r}_{\mathbf{Q}}(\underline{a} \underline{o}_{\leq t})]$ are omitted because they are analogous.

$$\begin{aligned} \mathbf{E}[\mathbf{r}_{\mathbf{P}}(\underline{a} \underline{o}_{\leq t})] &\stackrel{(a)}{=} \sum_{\underline{a} \underline{o}_{\leq t}} \mathbf{G}(\underline{a} \underline{o}_{\leq t}) \ln \mathbf{P}(\underline{a} \underline{o}_{\leq t}) \\ &\stackrel{(b)}{=} \sum_{\underline{a} \underline{o}_{\leq t}} \mathbf{G}(\underline{a} \underline{o}_{\leq t}) \left(\ln \mathbf{P}(a_{\leq t} | o_{<t}) + \ln \mathbf{P}(o_{\leq t} | a_{\leq t}) \right) \\ &\stackrel{(c)}{=} \sum_{\underline{a} \underline{o}_{\leq t}} \mathbf{G}(\underline{a} \underline{o}_{\leq t}) \left(\ln \mathbf{P}(a_{\leq t} | o_{<t}) + \ln \mathbf{P}(o_{\leq t} | a_{\leq t}) \right. \\ &\quad \left. + \ln \mathbf{Q}(o_{\leq t} | a_{\leq t}) - \ln \mathbf{Q}(o_{\leq t} | a_{\leq t}) \right) \\ &\stackrel{(d)}{=} -\mathbf{H}[\mathbf{P}(a_{\leq t} | o_{<t})] - \mathbf{H}[\mathbf{Q}(o_{\leq t} | a_{\leq t})] \\ &\quad - \mathbf{KL}[\mathbf{Q}(o_{\leq t} | a_{\leq t}) \| \mathbf{P}(o_{\leq t} | a_{\leq t})]. \end{aligned}$$

Equality (a) follows from the definition of expectations and the relation between rewards and probabilities. In (b) we separate the term in the logarithm into the action and observation part. In (c) we add and subtract the term $\mathbf{Q}(o_{\leq t} | a_{\leq t})$ in the logarithm. Equality (d) follows from the algebraic manipulation of the terms and from identifying the entropy terms, noting that $\mathbf{G}(\underline{a} \underline{o}_{\leq t}) = \mathbf{P}(a_{\leq t} | o_{<t}) \mathbf{Q}(o_{\leq t} | a_{\leq t})$. \square

References

- J.C. Candeal, J.R. De Miguel, E. Induráin, and G.B. Mehta. Utility and entropy. *Economic Theory*, 17:233–238, 2001.
- R.P. Feynman and A.R. Hibbs. *Quantum Mechanics and Path Integrals*. McGraw-Hill, 1965.
- P.C. Fishburn. *The Foundations of Expected Utility*. D. Reidel Publishing, Dordrecht, 1982.
- D. Fudenberg and D.M. Kreps. Learning mixed equilibria. *Games and Economic Behavior*, 5:320–367, 1993.
- B. Kappen, V. Gomez, and M. Opper. Optimal control as a graphical model inference problem. *arXiv:0901.0633*, 2009.
- S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, mar 1951. ISSN 0003-4851.
- D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- P.A. Ortega and D.A. Braun. A minimum relative entropy principle for learning and acting. *arXiv:0810.3605*, 2008.
- S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 1st edition edition, 1995.
- C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, Jul and Oct 1948.
- E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences U.S.A.*, 106: 11478–11483, 2009.