An Adversarial Interpretation of Information-Theoretic Bounded Rationality

Pedro A. Ortega and **Daniel D. Lee**
School of Engineering and Applied Sciences
University of Pennsylvania
Philadelphia, PA 19104, USA
{ope,ddlee}@seas.upenn.edu

Abstract

Recently, there has been a growing interest in modeling planning with information constraints. Accordingly, an agent maximizes a regularized expected utility known as the free energy, where the regularizer is given by the information divergence from a prior to a posterior policy. While this approach can be justified in various ways, including from statistical mechanics and information theory, it is still unclear how it relates to decision-making against adversarial environments. This connection has previously been suggested in work relating the free energy to risk-sensitive control and to extensive form games. Here, we show that a single-agent free energy optimization is equivalent to a game between the agent and an imaginary adversary. The adversary can, by paying an exponential penalty, generate costs that diminish the decision maker's payoffs. It turns out that the optimal strategy of the adversary consists in choosing costs so as to render the decision maker indifferent among its choices, which is a definining property of a Nash equilibrium, thus tightening the connection between free energy optimization and game theory.

Keywords: bounded rationality, free energy, game theory, Legendre-Fenchel transform.

1 Introduction

Recently, there has been a renewed interest in *bounded rationality*, originally proposed by Herbert A. Simon as an alternative account to the model of perfect rationality in the face of complex decisions (Simon 1972). In artificial intelligence (AI), this development has been largely motivated by the intractability of exact planning in complex and uncertain environments (Papadimitriou and Tsitsiklis 1987) and the difficulty of finding domain-specific simplifications or approximations that would render planning tractable despite the latest theoretical advancements in understanding the AI problem (Hutter 2004). Newly proposed techniques based on Monte Carlo simulations such as Monte Carlo Tree Search (Coulom 2006; Browne et al. 2012), Thompson sampling (Strens 2000) and Free Energy (Kappen 2005b) have become the focus of much AI and reinforcement learning research due to their wide applicability and their unprecedented success in difficult problems such as computer Go (Gelly et al. 2006), universal planning (Veness et al. 2011), human-level video game playing (Mnih et al. 2013) and robot learning (Theodorou, Buchli, and Schaal 2010). Roughly, these efforts can be broadly classified into two groups: (a) randomized approximations to the perfect rationality model and (b) exact statistical-mechanical or information-theoretic (IT) approaches. The latter group is of special theoretical interest, as it rests on a model of bounded rationality[1] that yields randomized optimal policies as a consequence of resource-boundedness (Ortega and Braun 2013).

The difference between the perfectly rational and the IT bounded rational model lies in the choice of the objective function: the IT approach adds a regularization term in the form of a Kullback-Leibler divergence (KL-divergence) to the expected utility of the perfectly rational model. The resulting objective function goes under various names in the literature, such as *KL-control cost* and *free energy*, and has been motivated in numerous ways. The initial inspiration comes from the maximum entropy principle of statistical mechanics as a way to model stochastic control problems that are linearly-solvable (Fleming 1977; Kappen 2005a; Todorov 2006). The methods from robust control were also adopted by economists to characterize model misspecification (Hansen and Sargent 2008). Later, it was shown that the free energy can be derived from axioms that treat utilities and information as commensurable quantities (Ortega and Braun 2013). Then in robot reinforcement learning, researchers proposed the free energy as a way to control the information-loss resulting from policy updates (Peters, Mülling, and Altün 2010). Finally, the free energy has also been motivated from arguments that parallel rate-distortion theory and information geometry by showing that the optimal policy minimizes the decision complexity subject to a constraint on the expected utility (Tishby and Polani 2011). It is also worth mentioning that similar approaches arise in the context of neuroscience (Friston 2009) and in game the-

[1] There are several approaches to bounded rationality in the literature, most notably those coming from the field of behavioral economics (Rubinstein 1998), which put emphasis on the procedural elements of decision making. Here we have settled on the qualifier "information-theoretic" as a way to distinguish from these approaches.

ory (Wolpert 2004).

An important yet intriguing property of the free energy is that it is a universal aggregator of value. More precisely, it can implement the minimum, expectation, maximization, and all the "soft" aggregations in between by varying a single parameter. This has at least two important consequences. First, the free energy corresponds, in economic jargon, to the *certainty-equivalent*, i.e. the value a risk-sensitive decision-maker is thought to assign to an uncertain choice (van den Broek, Wiegerinck, and Kappen 2010). Second, it has been pointed out that decision trees are nested certainty-equivalent operations (Ortega and Braun 2012). This is important, as decision making in the face of an adversarial, an indifferent, and a cooperative environment, which have previously been treated as unrelated modeling assumptions, are actually instantiations of a single decision rule. Previous work explores this property in multi-agent settings, showing that solutions change from being risk-dominant to payoff-dominant (Kappen, Gómez, and Opper 2012).

Aim of this Work

Despite the identification of the free energy with the certainty-equivalent, the connection to adversarial environments is as yet not fully understood. Our work makes an important step into understanding this connection. By applying a Legendre-Fenchel transformation to the regularization term of the free energy, it is revealed that a single-agent free energy optimization can be thought of as representing a game between the agent and an imaginary adversary. Furthermore, this result explains stochastic policies as a strategy that guards the agent from adversarial reactions of the environment.

Structure of this Article

This article is organized into four sections. The aim of the next section (Section 2) is to familiarize the reader with the mathematical foundations underlying the planning problem in AI, and to give a basic introduction into IT bounded rationality necessary in order to contextualize the results of our work. Section 3 contains our central contribution. It first briefly reviews Legendre-Fenchel transformations and then applies them to the free energy functional in order to unveil the adversarial assumptions implicit in the objective function. Finally, Section 4 discusses the results and concludes.

2 Preliminaries

We review the abstract foundations of planning under expected utility theory and IT bounded rationality.

Variational Principles

The behavior of an agent is typically characterized in one of two ways: (a) by directly describing its policy, or (b) by specifying a *variational problem* that has the policy as its solution. While the former is a direct specification of the agent's actions under any contingency, the latter has the advantage that it provides an explicit (typically convex) objective function that the policy has to optimize. Thus, a variational principle has additional explanatory power: not only does it single out an optimal policy, but it also encodes a *preference relation* over the set of feasible policies. Crucially, the qualifier "optimal" only holds relative to the objective function and is by no means absolute: because given *any policy*, one can always engineer a variational principle that is extremized by it. In AI, virtually all planning algorithms are framed (either explicitly or implicitly) as *maximum expected utility* problems. This encompasses popular problem classes such as multi-armed bandits, Markov decision processes and partially observable Markov decision processes (Legg 2008, Chapter 3).

Sequential versus Single-Step Decisions

We briefly recall a basic theoretical result that will simplify our analysis. In planning, *sequential decision problems* can be rephrased as *single-step decision problems*: instead of letting the agent choose an action in each turn, one can equivalently let the agent choose a single *policy* in the beginning that it then must follow during its interactions with the environment[2]. This observation was first made in game theory, where it was proven that every *extensive form game* can always be re-expressed as a *normal form game* (Von Neumann and Morgenstern 1944). Consequently, we abstract away from the sequential nature of the general problem by limiting our discussion to single-step decisions involving a single action and observation. The price we pay is to hide the dynamical structure and to increase the complexity of the policy, but the mathematical results can stated more concisely.

(Subjective) Expected Utility

Expected utility (Von Neumann and Morgenstern 1944; Savage 1954) is the *de facto* standard variational principle in artificial intelligence (Russell and Norvig 2010). It is based on a *utility function*, that is, a real-valued mapping of the outcomes encoding the desirability of each possible realization of the interactions between the agent and the environment. The qualifiers "subjective" and "expected" in its name derive from the fact that the desirability of a stochastic realization is calculated as the expectation over the utilities of the particular realizations measured with respect to the subjective beliefs of the agent.

Formally, let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets, the former corresponding to the *set of actions* and the latter to the *set of observations*. A *realization* is a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Furthermore, let $U : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a utility function, such that $U(x, y)$ represents the desirability of the realization $(x, y) \in \mathcal{X} \times \mathcal{Y}$; and let $q(\cdot|\cdot)$ be a conditional probability distribution characterizing the environmental response where $q(y|x)$ represents the probability of the observation $y \in \mathcal{Y}$ given the action $x \in \mathcal{X}$. Then, the agent's optimal

[2] Note that this reduction encompasses policies that appear to change during its execution as a function of the history: the meta-policy controlling the changes is a compressed, yet sufficient description of the changing policy.

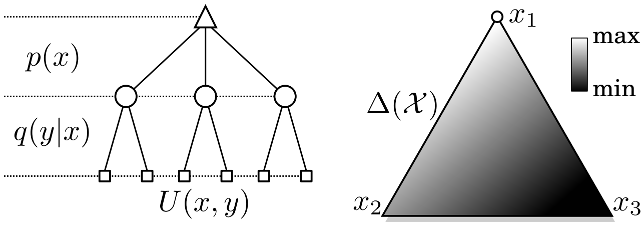

Figure 1: Expected Utility Theory. The decision problem (left) boils down to choosing a policy—a member of the simplex over actions $x \in \mathcal{X}$—maximizing the convex combination over conditional expected utilities $\mathsf{E}[U|x]$ (right).

policy $p \in \Delta(\mathcal{X})$ is chosen so as to maximize the functional

$$\begin{aligned}\mathsf{E}[U] &= \sum_x p(x)\mathsf{E}[U|x] \\ &= \sum_x p(x) \sum_y q(y|x)U(x,y). \qquad (1)\end{aligned}$$

This is illustrated in Figure 1. An *optimal policy* is any distribution p^* with no support over suboptimal actions, that is, $p^*(x) = 0$ whenever $\mathsf{E}[U|x] \leq \max_z \mathsf{E}[U|z]$. In particular, because the expected utility is linear in the policy probabilities, one can always choose a solution that is a vertex of the probability simplex $\Delta(\mathcal{X})$:

$$p^*(x) = \delta^x_{x^*} = \begin{cases} 1 & \text{if } x = x^* := \arg\max_x \mathsf{E}[U|x] \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

where δ is the Kronecker delta. Hence, there always exists a deterministic optimal policy.

Once the optimal policy has been chosen, the utility U becomes a well-defined random variable with probability distribution

$$\begin{aligned}\mathsf{P}(U = u) &= \mathsf{P}\big\{(x,y) : U(x,y) = u\big\} \\ &= \sum_{U^{-1}(u)} p^*(x)q(y|x). \qquad (3)\end{aligned}$$

Even though the optimal policy might be deterministic, the utility of the ensuing realization is in general stochastic.

Let us now consider the case when the environment is another agent. Formally, this means that the agent lacks the conditional probability distribution $q(\cdot|\cdot)$ over observations that it requires in order the evaluate the expected utilities. Instead, the agent possesses a second utility function $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ characterizing the desires of the environment. Game theory then invokes a *solution concept*, most notably the *Nash equilibrium*, in order to obtain the missing distribution $q(\cdot|\cdot)$ from U and V that renders the original description as a well-defined decision problem. Thus, conceptually, game theory starts from slightly weaker assumptions than expected utility theory. For simplicity, we restrict ourselves to two particular cases: (a) the fully adversarial case $U(x,y) = -V(x,y)$; and (b) the fully cooperative case $U(x,y) = V(x,y)$. The Nash equilibrium then yields the decision rules (Osborne and Rubinstein 1999):

$$p^* = \arg\max_p \sum_x p(x) \Big\{ \min_q q(y|x)U(x,y) \Big\}, \qquad (4)$$

$$p^* = \arg\max_p \sum_x p(x) \Big\{ \max_q q(y|x)U(x,y) \Big\} \qquad (5)$$

for the two cases respectively. Comparing these to (1) we immediately see that, for planning purposes, it is irrelevant whether the decision is made by the agent or not—what matters is the degree to which an outcome (be it an action or an observation) contributes to the agent's objective function (as encoded by one of the three possible aggregation operators). Thus, equations (1), (4) and (5) can be regarded as consisting of not one, but *two* nested "decision steps" of the form

$$\max_x\{U(x)\}, \quad \exp_p[U(X)] \quad \text{or} \quad \min_x\{U(x)\}.$$

We end this brief review by summarizing the main properties of expected utility:

1. *Linearity:* The objective function is linear in the policy.
2. *Existence of Deterministic Solution:* Although there are optimal stochastic policies, there always exists an equivalent deterministic optimal policy.
3. *Indifference to Higher-Order Moments:* Expected utility places constraints on the first, but not on the higher-moments of the probability distribution of the utility.
4. *Exhaustive Search:* Finding the optimal deterministic policy requires, in the worst-case, an exhaustive evaluation of all the expected utilities.

It is important to note that these properties only apply to expected utility, but not to game theory.

Free Energy

How does an agent make decisions when it cannot exhaustively evaluate all the alternatives? IT bounded rationality addresses this question by defining a new objective function that trades off utilities versus information costs. Planning is conceptualized as a process that transforms a prior policy into a posterior policy that incurs a cost measured in *utiles*. Importantly, it is assumed that the agent is not allowed itself to reason about the costs of this transformation[3]; rather, it tries to optimize the original expected utility but then runs out of resources. From the point of view of an external observer, this "interrupted" planning process will appear as if it were explicitly optimizing the free energy.

Formally, let $\mathcal{X}$ be a finite set corresponding to the *set of realizations* and $U : \mathcal{X} \to \mathbb{R}$ is the utility function. Furthermore, let $p_0 \in \Delta(\mathcal{X})$ be a prior policy and $\beta \in \mathbb{R}$ be a boundedness parameter. Then, the agent's optimal policy $p \in \Delta(\mathcal{X})$ is chosen so as to extremize the free energy functional

$$F_\beta[p] := \underbrace{\sum_x p(x)U(x)}_{\text{Expected Utility}} - \underbrace{\frac{1}{\beta}\sum_x p(x)\log\frac{p(x)}{p_0(x)}}_{\text{Information Costs}}, \qquad (6)$$

[3] The inability to reason about resource costs renders IT bounded rationality fundamentally different from the meta-reasoning approach of Stuart J. Russell (Russell 1995).

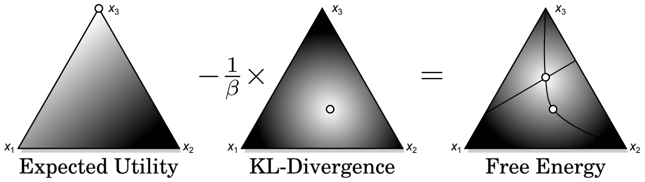

Figure 2: Free energy consists of the expected utility regularized by the KL-divergence between a given prior and posterior policy. The resulting objective function is convex and has an optimum p^* that is in general in the interior of the probability simplex $\Delta(\mathcal{X})$.

where it is seen that β controls the conversion from units of information to utiles (Figure 2). The optimal policy is given by the "softmax"-like distribution

$$p^*(x) = \frac{p_0(x)e^{\beta U(x)}}{\sum_x p_0(x)e^{\beta U(x)}} \tag{7}$$

which is known as the *equilibrium distribution*. Notice that β controls how much the agent is in control of the choice: a value $\beta \approx 0$ falls back to the prior policy and represents lack of influence; and values far away from zero yield either adversarial or cooperative choices depending on the sign of β. An important property of the equilibrium distribution is that it can be simulated exactly *without* evaluating all the utilities using Monte Carlo methods (Ortega, Braun, and Tishby 2014).

As we have mentioned before, the free energy (6) corresponds to the certainty-equivalent. This is seen as follows: the extremum of (6) given by

$$\frac{1}{\beta}\log Z_\beta, \qquad \text{where} \quad Z_\beta := \sum_x p_0(x)e^{\beta U(x)}, \tag{8}$$

but now seen *as a function of* β, can be thought of as an interpolation of the maximum, expectation and minimum operator, since

$$\begin{aligned} \tfrac{1}{\beta}\log Z_\beta &= \max_x\{U(x)\} && \beta \to +\infty \\ \tfrac{1}{\beta}\log Z_\beta &= \quad \mathbf{E}_{p_0}[U] && \beta \to \quad 0 \\ \tfrac{1}{\beta}\log Z_\beta &= \min_x\{U(x)\} && \beta \to -\infty. \end{aligned}$$

We end this brief review by summarizing the main properties of IT bounded rationality:

1. *Nonlinearity:* The objective function is nonlinear in the policy.
2. *Stochastic Solutions:* Optimal policies are inherently stochastic.
3. *Sensitive to Higher-Order Moments:* Because the free energy is nonlinear in the policy, it places constraints on the higher-order moments of the policy too. In particular, a Taylor expansion of the KL-divergence term reveals that the free energy is sensitive to *all* the moments of the policy.

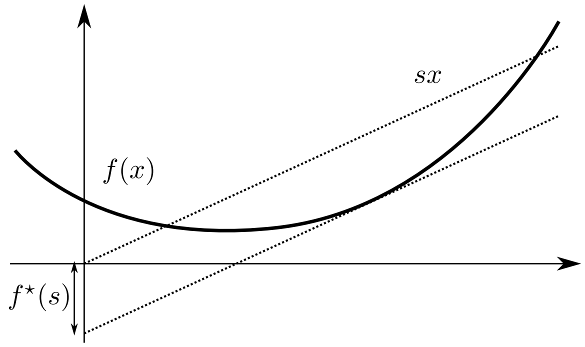

Figure 3: The Legendre-Fenchel Transform. Given a function $f(x)$, the convex conjugate $f^\star(s)$ corresponds to the intercept of a tangent line to the curve with slope s.

4. *Randomized Search:* Obtaining a decision from the optimal policy does not require the exhaustive evaluation of the utilities. Rather, it can be sampled using Monte Carlo techniques.

3 Adversarial Interpretation

Before presenting our main results, we give a brief definition of Legendre-Fenchel transforms.

Legendre-Fenchel Transforms

The *Legendre-Fenchel transformation* is a transformation that yields an alternative encoding of a given functional relationship. More precisely, the information contained in the function is expressed using the derivative as the independent variable. Because of this, Legendre-Fenchel transformations play an important role in thermodynamics and optimization. For a function $f(x) : \mathbb{R}^n \to \mathbb{R}$, it is defined by the variational formula

$$f^\star(s) = \sup_{x\in\mathcal{X}}\{\langle s, x\rangle - f(x)\}$$

where $f^\star(s) : \mathbb{R}^n \to \mathbb{R}$ is the *convex conjugate* and $\langle s, x\rangle$ is the inner product of two vectors s, x in $\mathbb{R}^n$. The following are common examples:

1. *Affine function:*

$$f(x) = ax - b \qquad f^\star(s) = \begin{cases} b & \text{if } s = a, \\ +\infty & \text{if } s \neq a. \end{cases}$$

2. *Power function:*

$$f(x) = \frac{1}{\alpha}|x|^\alpha \qquad f^\star(s) = \frac{1}{\alpha'}|s|^{\alpha'}$$

where $\frac{1}{\alpha} + \frac{1}{\alpha'} = 1$.

3. *Exponential function:*

$$f(x) = e^x \qquad f^\star(s) = \begin{cases} s\log s - s & \text{if } s > 0, \\ 0 & \text{if } s = 0, \\ +\infty & \text{if } s < 0. \end{cases}$$

We refer the reader to Touchette's article (Touchette 2005) for a brief introduction or Boyd and Vanderberghe's book (Boyd and Vandenberghe 2004) for a thorough treatment.

Unveiling the Adversarial Environment

Using a Legendre-Fenchel transformation, one can show that the regularization term of the free energy can be rewritten as a variational problem.

Lemma 1.

$$\begin{aligned}&-\frac{1}{\beta}\sum_x p(x)\log\frac{p(x)}{p_0(x)}\\ &=\min_C \sum_x -p(x)C(x) + p_0(x)e^{\beta C(x)} - \frac{1}{\beta}(\log\beta + 1)\end{aligned} \tag{9}$$

Proof. Since $n = |\mathcal{X}|$ is finite, $p(x), C(x)$ are vectors in $\mathbb{R}^n$. Essentially, the lemma says that the l.h.s. is the convex conjugate of the function

$$f(C) = -\sum_x p_0(x)e^{\beta C(x)} + \frac{1}{\beta}(\log\beta + 1),$$

because the r.h.s. can be re-expressed as

$$\min_C \sum_x -p(x)C(x) + f(C) = \max_C \sum_x p(x)C(x) - f(C).$$

Setting the derivative w.r.t. C to zero yields the optimality condition

$$p(x) = \beta p_0(x)e^{\beta C(x)}.$$

Isolating $C(x)$ and substituting into $\sum_x p(x)C(x) - f(C)$ proves the lemma. □

Consequently, the regularization term of the free energy encapsulates a second variational problem over an auxiliary vector C. In particular, the logarithmic form encodes an objective function that is exponential in C. Equipped with this lemma, we can state our main result.

Theorem 2. *The maximization of the free energy* (6) *is equivalent to the maximization of*

$$\min_C \underbrace{\sum_x p(x)[U(x) - C(x)]}_{\textit{Expected Net Utility}} + \underbrace{\sum_x p_0(x)e^{\beta C(x)}}_{\textit{Penalty of Adversary}} \tag{10}$$

w.r.t. the policy $p \in \Delta(\mathcal{X})$.

Proof. The proof is an immediate consequence of Lemma 1:

$$\begin{aligned}&\arg\max_p\Big\{\sum_x p(x)U(x) - \frac{1}{\beta}\sum_x p(x)\log\frac{p(x)}{p_0(x)}\Big\}\\ &=\arg\max_p\Big\{\sum_x p(x)U(x) + \min_C\Big\{\sum_x -p(x)C(x)\\ &\qquad + p_0(x)e^{\beta C(x)} - \frac{1}{\beta}(\log\beta+1)\Big\}\Big\}\\ &=\arg\max_p\min_C\Big\{\sum_x p(x)[U(x) - C(x)] + \sum_x p_0(x)e^{\beta C(x)}\Big\}.\end{aligned}$$

□

This result can be interpreted as follows. When a single agent maximizes the free energy, it is implicitly assuming a situation where it is playing against an imaginary adversary. In this situation, the agent first chooses its policy $p \in \Delta(\mathcal{X})$, and then the adversary attempts to decrease the agent's net utilities by subtracting costs $C(x)$. However, the adversary cannot choose these costs arbitrarily; instead, it must pay exponential penalties. In fact, its optimal strategy can be directly calculated from Lemma 1.

Corollary 3. *The imaginary adversary's optimal strategy is given by*

$$C^*(x) = \frac{1}{\beta}\log\frac{p(x)}{\beta p_0(x)}. \tag{11}$$

Inspecting (11), we see that the optimal costs scale relatively with the agent's deviation from its prior probabilities. This leads to an interesting interaction between the agent's and the imaginary adversary's choices.

Indifference

Our second main result characterizes the solution to this adversarial setup. It turns out that the adversary's best strategy is to choose costs such that the agent's net payoffs are *uniform*.

Theorem 4. *The solution to*

$$\max_p \min_C \sum_x p(x)[U(x) - C(x)] + \sum_x p_0(x)e^{\beta C(x)}$$

has the property that for all $x \in \mathcal{X}$,

$$U(x) - C(x) = \textit{constant}.$$

Proof. We first note that we can exchange the maximum and minimum operations,

$$\begin{aligned}&\max_p \min_C \sum_x p(x)[U(x) - C(x)] + \sum_x p_0(x)e^{\beta C(x)}\\ &= \min_C \max_p \sum_x p(x)[U(x) - C(x)] + \sum_x p_0(x)e^{\beta C(x)}\end{aligned}$$

because there is no duality gap due to the concavity of the exponential function. We can thus maximize first w.r.t. the policy p. Define $\mathcal{X}^* \subset \mathcal{X}$ as the subset of elements maximizing the penalized expected utility, that is, for all $x^* \in \mathcal{X}$ and $x \in \mathcal{X}$,

$$U(x^*) - C(x^*) \geq U(x) - C(x). \tag{12}$$

Maximizing w.r.t. to p yields optimal probabilities $p^*(x)$ given by

$$p^*(x) = \begin{cases} q(x) & \text{if } x \in \mathcal{X}^*,\\ 0 & \text{otherwise,}\end{cases}$$

where q is any distribution over $\mathcal{X}^*$. Given this, the worst case costs $C^*(x)$ are

$$C^*(x) = \begin{cases} \frac{1}{\beta}\log\frac{q(x)}{\beta p_0(x)} & \text{if } x \in \mathcal{X}^*,\\ -\infty & \text{otherwise.}\end{cases} \tag{13}$$

However, if $\mathcal{X}^* \neq \mathcal{X}$, then we get a contradiction, since

$$U(x^*) - C^*(x^*) \not\geq U(x) - C^*(x)$$

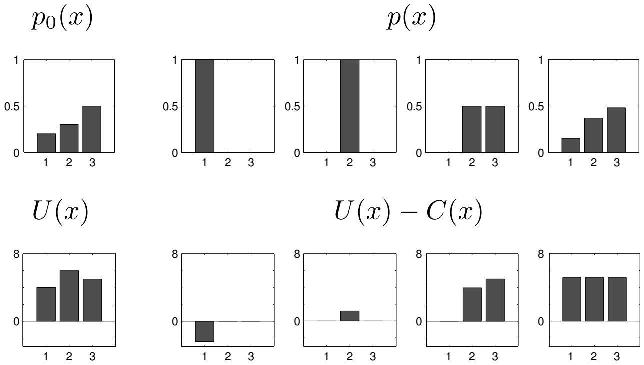

Figure 4: Given an agent's policy p, the environment chooses costs that change the net utilities $U(x) - C(x)$ of the realizations with support. The agent can protect itself from adversarial costs by randomization. The last column corresponds to the agent's optimal policy.

for all $x \notin \mathcal{X}^*$, violating (12). Hence, it must be that for all $x \in \mathcal{X}$,

$$U(x) - C(x) = \text{constant},$$

concluding the proof. □

To get a better understanding on the meaning of this result, it is helpful to consider an example. Figure 4 illustrates four choices of policies and the corresponding optimal adversarial costs. Here it is seen that an agent can protect itself by spreading the probability mass of its policy over many realizations.

4 Discussion

Starting from the free energy functional, we have shown how to construct an alternative adversarial interpretation that constitutes an equivalent problem. Conceptually, our findings can be summarized as follows:

1. A regularization of the expected utility encodes assumpions about deviations from the expected utility.
2. The Legendre-Fenchel transformation reinterprets the free energy as a game against an environmental adversary.
3. In this transformation, it is seen that the regularization is equivalent to the penalization of adversarial costs.
4. Stochastic policies guard against adversarial costs. Expected utility alone is linear in the policy and thus encodes deterministic optimal policies.

Indifference and Nash Equilibrium

Our results establish an interesting relation to game theory. Theorem 4 and (11) immediately imply

$$U(x) - C^*(x) = U(x) - \frac{1}{\beta} \log \frac{p(x)}{\beta p_0(x)} = \text{constant}$$

for all $x \in \mathcal{X}$. This turns out to be an known characterization of the equilibrium distribution (7). However, in light of our present results, it acquires a different twist; it fits with the well-known result that a Nash equilibrium is a strategy profile such that each player chooses a (mixed) strategy that renders the others players indifferent to their choices (Osborne and Rubinstein 1999).

Other Regularizers

The presented method is general and can also be applied to other regularizers in order to make the assumptions about the imaginary adversary explicit. For instance, the following list enumerates some examples.

1. *Expected Utility:* Because the regularization term is null, the resulting adversary's objective function is
$$\max_p \min_C \sum_x p(x)[U(x) - C(x)] - \delta_0^{C(x)}.$$
Here we see that a null regularization implies an adversary devoid of power, i.e. one that cannot alter the utilities chosen by the agent.
2. *Power Function:* If the regularizer is a power function, then we have the dual power function
$$\max_p \min_C \sum_x p(x)[U(x) - C(x)] + |C(x)|^\alpha.$$
This case is interesting because it encompases ridge and lasso as special cases.
3. *Modern Portfolio Theory:* One notable case that is widely used in practice is Modern Portfolio Theory. Here, an investor trades off asset returns versus portfolio risk, encoded into a regularization term that is quadratic in the policy (Markowitz 1952). The transformed objective function is given by
$$\max_p \min_C \sum_x p(x)[U(x) - C(x)] + \frac{1}{2}\lambda C^T \Sigma C.$$

Conclusions

Our central result shows how to extract the adversarial cost function implicitly assumed in the agent's objective function by means of an appropriately chosen Legendre transformation. Conversely, one can also start from the cost function of an adversarial environment and reexpress it as a regularized optimization. While our main motivation was to apply this to the free energy functional, the transformation is general and works also in other well-known frameworks such as in modern portfolio theory. The immediate application is that we can switch between the two representations and pick the one that is easier to solve. This suggests novel algorithms for solving the single-agent planning problem based on ideas from differential game theory (Dockner et al. 2001) and convex programming (Zinkevich 2003). Furthermore, this also suggests that agents that randomize their decisions, are essentially expressing their information limitations by protecting themselves from undesired outcomes. As such, we believe this is a basic result that opens up a series of additional questions regarding the nature of stochastic policies, such as the conditions under which they arise and how they encode risk sensitivity, which need to be further explored in the future.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions for improving this manuscript. This study was funded by grants from the U.S. National Science Foundation, Office of Naval Research and Department of Transportation.

References

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge Univeristy Press.

Browne, C.; Powley, E.; Whitehouse, D.; Lucas, S.; Cowling, P.; Rohlfshagen, P.; Travener, S.; Perez, D.; Samothrakis, S.; and Colton, S. 2012. A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in Games* 4(1).

Coulom, R. 2006. Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. In *Computer and Games*.

Dockner, E.; Jorgensen, S.; Long, N.; and Sorger, G. 2001. *Differential Games in Economics and Management Science*. Cambridge University Press.

Fleming, W. 1977. Exit Probabilities and Optimal Stochastic Control. *Applied Mathematics and Optimization* 4:329–346.

Friston, K. 2009. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Science* 13:293–301.

Gelly, S.; Wang, Y.; Munos, R.; and Teytaud, O. 2006. Modification of UCT with Patterns in Monte-Carlo Go. Technical report, Inst. Nat. Rech. Inform. Auto. (INRIA).

Hansen, L., and Sargent, T. 2008. *Robustness*. Princeton: Princeton University Press.

Hutter, M. 2004. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer.

Kappen, H.; Gómez, V.; and Opper, M. 2012. Optimal control as a graphical model inference problem. *Machine Learning* 1:1–11.

Kappen, H. 2005a. A linear theory for control of non-linear stochastic systems. *Physical Review Letters* 95:200201.

Kappen, H. 2005b. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*.

Legg, S. 2008. *Machine Super Intelligence*. Ph.D. Dissertation, Department of Informatics, University of Lugano.

Markowitz, H. 1952. Portfolio Selection. *The Journal of Finance* 7:77–91.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing Atari with Deep Reinforcement Learning. *ArXiv* (1312.5602).

Ortega, P., and Braun, D. 2012. Free Energy and the Generalized Optimality Equations for Sequential Decision Making. In *European Workshop on Reinforcement Learning (EWRL'10)*.

Ortega, P. A., and Braun, D. A. 2013. Thermodynamics as a Theory of Decision-Making with Information Processing Costs. *Proceedings of the Royal Society A 20120683*.

Ortega, P.; Braun, D.; and Tishby, N. 2014. Monte Carlo Methods for Exact & Efficient Solution of the Generalized Optimality Equations. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Osborne, M., and Rubinstein, A. 1999. *A Course in Game Theory*. MIT Press.

Papadimitriou, C., and Tsitsiklis, J. 1987. The Complexity of Markov Decision Processes. *Mathematics of Operations Research* 12(3):441–450.

Peters, J.; Mülling, K.; and Altün, Y. 2010. Relative entropy policy search. In *AAAI*.

Rubinstein, A. 1998. *Modeling Bounded Rationality*. Cambridge, MA: MIT Press.

Russell, S., and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 3rd edition edition.

Russell, S. 1995. Rationality and Intelligence. In Mellish, C., ed., *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 950–957. San Francisco: Morgan Kaufmann.

Savage, L. 1954. *The Foundations of Statistics*. New York: John Wiley and Sons.

Simon, H. 1972. Theories of Bounded Rationality. In Radner, C., and Radner, R., eds., *Decision and Organization*. Amsterdam: North Holland Publ. 161–176.

Strens, M. 2000. A Bayesian Framework for Reinforcement Learning. In *ICML*.

Theodorou, E.; Buchli, J.; and Schaal, S. 2010. A generalized path integral approach to reinforcement learning. *Journal of Machine Learning Research* 11:3137–3181.

Tishby, N., and Polani, D. 2011. *Perception-Action Cycle*. Springer New York. chapter Information Theory of Decisions and Actions, 601–636.

Todorov, E. 2006. Linearly solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19, 1369–1376.

Touchette, H. 2005. Legendre-Fenchel Transforms in a Nutshell. Technical report, Rockefeller University.

van den Broek, B.; Wiegerinck, W.; and Kappen, H. 2010. Risk Sensitive Path Integral Control. In *UAI*, 615–622.

Veness, J.; Ng, M.; Hutter, M.; Uther, W.; and Silver, D. 2011. A Monte-Carlo AIXI Approximation. *Journal of Artificial Intelligence Research* 40:95–142.

Von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wolpert, D. 2004. *Complex Engineering Systems*. Perseus Books. chapter Information theory - the bridge connecting bounded rational game theory and statistical physics.

Zinkevich, M. 2003. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In *ICML*, 928–936.