JMLR: Workshop and Conference Proceedings vol:1–10, 2012 EWRL 2012

Free Energy and the Generalized Optimality Equations for Sequential Decision Making

Pedro A. Ortega PEDRO.ORTEGA@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics
Max Planck Institute for Intelligent Systems
ORAND S.A.

Daniel A. Braun DANIEL.BRAUN@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics
Max Planck Institute for Intelligent Systems

Editor: Marc Deisenroth, Csaba Szepesvari, Jan Peters

Abstract

The free energy functional has recently been proposed as a variational principle for bounded rational decision-making, since it instantiates a natural trade-off between utility gains and information processing costs that can be axiomatically derived. Here we apply the free energy principle to general decision trees that include both adversarial and stochastic environments. We derive generalized sequential optimality equations that not only include the Bellman optimality equations as a limit case, but also lead to well-known decision-rules such as Expectimax, Minimax and Expectiminimax. We show how these decision-rules can be derived from a single free energy principle that assigns a resource parameter to each node in the decision tree. These resource parameters express a concrete computational cost that can be measured as the amount of samples that are needed from the distribution that belongs to each node. The free energy principle therefore provides the normative basis for generalized optimality equations that account for both adversarial and stochastic environments.

Keywords: Foundations of AI, free energy, Bellman optimality equations, bounded rationality.

1. Introduction

Decision trees are a ubiquitous tool in decision theory and artificial intelligence research to represent a wide range of decision-making problems that include the classic reinforcement learning paradigm as well as competitive games (Osborne and Rubinstein, 1999; Russell and Norvig, 2010). Depending on the kind of system one is interacting with, there are different decision rules one has to apply—the most famous ones being *Expectimax*, *Minimax* and *Expectiminimax*—see Figure 1. When an agent interacts with a stochastic system, the agent chooses its decisions based on *Expectimax*. Essentially, Expectimax is the dynamic programming algorithm that solves the Bellman optimality equations, thereby recursively maximizing expected future reward in a sequential decision problem (Bellman, 1957).

In two-player zero-sum games where strictly competitive players make alternate moves, an agent should use the *Minimax* strategy. The motivation underlying minimax decisions is

that the agent wants to optimize the worst-case gain as a means of protecting itself against the potentially harmful decisions made by the adversary. Finally, there are games that mix the two previous interaction types. For instance, in Backgammon, the course of the game depends on the skill of the players and chance elements. In these cases, the agent bases its decisions on the *Expectiminimax* rule (Michie, 1966).

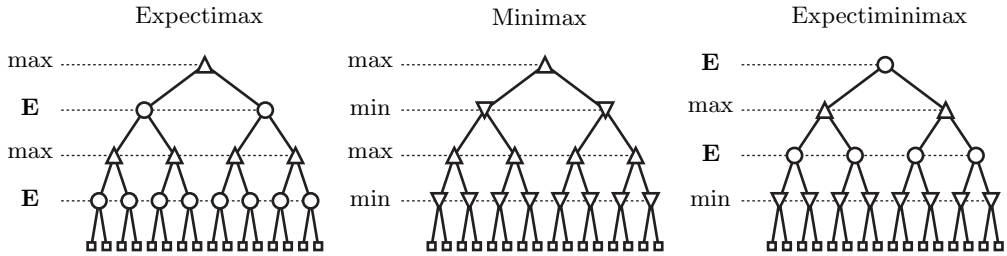

Figure 1: Illustration of Expectimax, Minimax and Expectiminimax in decision trees representing three different interaction scenarios. The internal nodes can be of three possible types: maximum ($\triangle$), minimum ($\triangledown$) and expectation ($\circ$). The optimal decision is calculated recursively using dynamic programming.

What is common to all of these decision-making schemes is that they presuppose a fully rational decision-maker that is able to compute all of the required operations with absolute precision. In contrast, a bounded rational decision-maker trades off expected utility gains against the cost of the required computations (Simon, 1984). Recently, the free energy has been suggested as a normative variational principle for such bounded rational decision-making that takes the computational effort into account (Ortega and Braun, 2011; Braun and Ortega, 2011; Ortega, 2011). This builds on previous work on efficient computation of optimal actions that trades off the benefits obtained from maximizing the utility function against the cost of changing the uncontrolled dynamics given by the environment (Kappen, 2005; Todorov, 2006, 2009; Kappen et al., 2012). The aim of this paper is to extend these results to generalized decision trees such that Expectimax, Minimax, Expectiminimax, and bounded rational acting can all be derived from a single optimization principle. Moreover, this framework leads to a natural measure of computational costs spent at each node of the decision tree. All the proofs are given in the appendix.

2. Free Energy

2.1. Equilibrium Distribution

In Ortega and Braun (2011) and in Ortega (2011) it was shown that a bounded rational decision-making problem can be formalized based on the *negative free energy difference* between two information processing states represented by two probability distributions P and Q. The decision process then transforms the initial choice probability Q into a final

choice probability P by taking into account the utility gains (or losses) and the transformation costs. This transformation process can be formalized as

$$P(x) = \frac{1}{Z} Q(x) e^{\alpha U(x)}, \qquad \text{where} \qquad Z = \sum_x Q(x) e^{\alpha U(x)}. \tag{1}$$

Accordingly, the choice pattern of the decision-maker is predicted by the *equilibrium distribution* P. Crucially, the probability distribution P extremizes the following functional (Callen, 1985; Keller, 1998):

Definition 1 (Negative Free Energy Difference) *Let Q be a probability distribution and let U be a real-valued utility function over the set $\mathcal{X}$. For any $\alpha \in \mathbb{R}$, define the* negative free energy difference $F_\alpha[P]$ *as*

$$F_\alpha[P] := \sum_x P(x) U(x) - \frac{1}{\alpha} \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{2}$$

The parameter α is called the inverse temperature.

Although strictly speaking, the functional $F_\alpha[P]$ corresponds to the negative free energy difference, we will refer to it as the "free energy" in the following for simplicity. When inserting the equilibrium distribution (1) into (2), the extremum of F_α yields:

$$\frac{1}{\alpha} \log \left(\sum_x Q(x) e^{\alpha U(x)} \right). \tag{3}$$

For different values of α, this extremum takes the following limits:

$$\begin{aligned}
\lim_{\alpha \to \infty} \tfrac{1}{\alpha} \log Z &= \max_x U(x) && \text{(maximum node)} \\
\lim_{\alpha \to 0} \tfrac{1}{\alpha} \log Z &= \sum_x Q(x) U(x) && \text{(chance node)} \\
\lim_{\alpha \to -\infty} \tfrac{1}{\alpha} \log Z &= \min_x U(x) && \text{(minimum node)}
\end{aligned}$$

The case $\alpha \to \infty$ corresponds to the perfectly rational agent, the case $\alpha \to 0$ corresponds to the expectation at a chance node and the case $\alpha \to -\infty$ anticipates the perfectly rational opponent. Therefore, the single expression $\frac{1}{\alpha} \log Z$ can represent the maximum, expectation and minimum depending on the value of α.

The inspection of (2) reveals that the free energy encapsulates a fundamental decision-theoretic trade-off: it corresponds to the expected utility, penalized—or regularized—by the information cost of transforming the base distribution Q into the final distribution P. The inverse temperature plays the role of the conversion factor between units of information and units of utility.

If we want to change the temperature α to β while keeping the equilibrium and reference distributions equal, then we need to change the corresponding utilities from U to V in a manner given by the following theorem. Temperature changes will be important for the application of the free energy principle to the general decision trees in Section 3.

Theorem 2 *Let P be the equilibrium distribution for a given inverse temperature α, utility function U and reference distribution Q. If the temperature changes to β while keeping P and Q fixed, then the utility function changes to*

$$V(x) = U(x) - \left(\tfrac{1}{\alpha} - \tfrac{1}{\beta}\right) \log \frac{P(x)}{Q(x)}.$$

2.2. Resource Costs

Consider the problem of picking the largest number in a sequence $U_0, U_1, U_2, \ldots$ of i.i.d. data, where each $U_i \in \mathcal{U}$ is drawn from a source with probability distribution M. After α draws the largest number will be given by $\max\{U_1, U_2, \ldots, U_\alpha\}$. Naturally, the larger the number of draws, the higher the chances of observing a large number.

Theorem 3 *Let $\mathcal{X}$ be a finite set. Let Q and M be strictly positive probability distributions over $\mathcal{X}$. Let α be a positive integer. Define M_α as the probability distribution over the maximum of α samples from M. Then, there are strictly positive constants δ and ξ depending only on M such that for all α,*

$$\left| \frac{Q(x)e^{\alpha U(x)}}{\sum_{x'} Q(x')e^{\alpha U(x')}} - M_\alpha(x) \right| \leq e^{-(\alpha-\xi)\delta}.$$

Consequently, one can interpret the inverse temperature as a resource parameter that determines how many samples are drawn to estimate the maximum. Note that the distribution M is arbitrary as long as it has the same support as Q. This interpretation can be extended to a negative α, by noting that $\alpha U(x) = (-\alpha)(-U(x))$, i.e. instead of the maximum we take the minimum of $-\alpha$ samples.

3. General Decision Trees

A generalized decision tree is a tree where each node corresponds to a possible interaction history $x_{\leq t} \in \mathcal{X}^t$, where t is smaller or equal than some fixed horizon T, and where edges connect two consecutive interaction histories. Furthermore, every node $x_{\leq t}$ has an associated inverse temperature $\beta(x_{\leq t})$; and every transition has a base probability $Q(x_t|x_{<t})$ of moving from state $x_{<t}$ to state $x_{\leq t} = x_{<t}x_t$ representing the stochastic law the interactions follow when it is not controlled, and an immediate reward $R(x_t|x_{<t})$. The objective of the agent is to make decisions such that the sum $\sum_{t=1}^{T} R(x_t|x_{<t})$ is maximized subject to the temperature constraints.

3.1. Free Energy for General Decision Trees

The free energy principle is stated above for one decision variable x. If x represents a tuple of (possibly dependent) random variables $x_1, \ldots, x_T$, then the free energy principle can be applied in a straightforward manner to the corresponding tree. However, all nodes of the tree will have the same inverse temperature assigned to them and, therefore, the same amount of computational resources will be spent at each node of the tree. This allows for

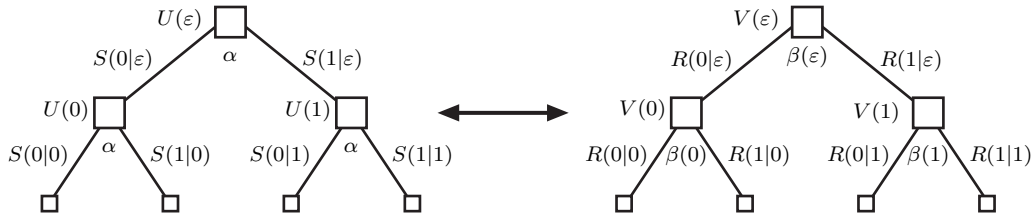

Figure 2: The free energy formalism can only be applied in a straightforward manner to trees with uniform resource allocation (left). In order to apply it to general trees that have different resource parameters at each node (right), we need to transform the utilities as described in (4) to preserve the equilibrium distribution.

example deriving the formalisms of path integral control and KL control (Todorov, 2009; Braun and Ortega, 2011; Kappen et al., 2012).

In the case of general decision trees the assumption of uniform temperatures has to be relaxed (Figure 2). In general, we can then dedicate different amounts of computational resources to each node of the tree. However, this requires a translation between a tree with a single temperature and to a tree with different temperatures. This translation can be achieved using Theorem 2. Define a reward as the change in utility of two subsequent nodes. Then, the rewards of the resulting decision tree are given by

$$R(x_t|x_{<t}) := \big[V(x_{\leq t}) - V(x_{<t})\big] = \big[U(x_{\leq t}) - U(x_{<t})\big] - \Big(\tfrac{1}{\alpha} - \tfrac{1}{\beta(x_{<t})}\Big) \log \frac{P(x_t|x_{<t})}{Q(x_t|x_{<t})}. \quad (4)$$

This allows introducing a collection of node-specific (not necessarily time-specific) inverse temperatures $\beta(x_{<t})$, allowing for a greater degree of flexibility in the representation of information costs. The next theorem states the connection between the free energy and the general decision tree formulation.

Theorem 4 *The free energy of the whole trajectory can be rewritten in terms of rewards:*

$$\begin{aligned} F_\alpha[P] &= \sum_{x_{\leq T}} P(x_{\leq T}) \Big\{ U(x_{\leq T}) - \frac{1}{\alpha} \log \frac{P(x_{\leq T})}{Q(x_{\leq T})} \Big\} \\ &= U(\varepsilon) + \sum_{x_{\leq T}} P(x_{\leq T}) \sum_{t=1}^{T} \Big\{ R(x_t|x_{<t}) - \frac{1}{\beta(x_{<t})} \log \frac{P(x_t|x_{<t})}{Q(x_t|x_{<t})} \Big\}. \end{aligned} \quad (5)$$

This translation allows applying the free energy principle to each node with a different resource parameter $\beta(x_{<t})$. By writing out the sum in (5), one realizes that this free energy has a nested structure where the latest time step forms the innermost variational problem and all other variational problems of the previous time steps can be solved recursively by working backwards in time. This then leads to the following solution:

Theorem 5 *The solution to the free energy in terms of rewards is given by*

$$P(x_t|x_{<t}) = \frac{1}{Z(x_{<t})} Q(x_t|x_{<t}) \exp\Big\{\beta(x_{<t})\big[R(x_t|x_{<t}) + \frac{1}{\beta(x_{\leq t})} \log Z(x_{\leq t})\big]\Big\},$$

where $Z(x_{\leq T}) = 1$ *and where for all* $t < T$

$$Z(x_{<t}) = \sum_{x_t} Q(x_t|x_{<t}) \exp\Big\{\beta(x_{<t})\big[R(x_t|x_{<t}) + \frac{1}{\beta(x_{\leq t})} \log Z(x_{\leq t})\big]\Big\}.$$

3.2. Generalized Optimality Equations

Theorem 5 together with the properties of the free energy extremum (3) suggest the following definition.

Definition 6 (Generalized Optimality Equations)

$$V(x_{<t}) = \frac{1}{\beta(x_{<t})} \log\Big\{\sum_{x_t} Q(x_t|x_{<t}) \exp\Big\{\beta(x_{<t})\big[R(x_t|x_{<t}) + V(x_{\leq t})\big]\Big\}\Big\}.$$

By virtue of our previous analysis, this equation tells us how to recursively calculate the *value function* (i.e. the utility of each node) given the computational resources allocated in each node.

It is immediately clear that the tree kinds of decision trees mentioned in the introduction are special cases of general decision trees. In particular, the three classical operators are obtained as limit cases:

$$V(x_{<t}) = \begin{cases} \max\limits_{x_t}\{R(x_t|x_{<t}) + V(x_{\leq t})\} & \text{if } \beta(x_{<t}) = \infty, \\ \mathbf{E}\{R(x_t|x_{<t}) + V(x_{\leq t})\} & \text{if } \beta(x_{<t}) = 0, \\ \min\limits_{x_t}\{R(x_t|x_{<t}) + V(x_{\leq t})\} & \text{if } \beta(x_{<t}) = -\infty. \end{cases}$$

The familiar Bellman optimality equations for stochastic systems are obtained by considering an agent decision node followed by a random decision node:

$$\begin{aligned} V(x_{<t}) &= \max_{x_t}\Big\{R(x_t|x_{<t}) + V(x_{\leq t})\Big\} \\ &= \max_{x_t}\Big\{R(x_t|x_{<t}) + \mathbf{E}\big[R(x_{t+1}|x_{\leq t}) + V(x_{\leq t+1})\big]\Big\}. \end{aligned}$$

4. Discussions & Conclusions

Bounded rational decision-making schemes based on the free energy generalize classic decision-making schemes by taking into account information processing costs measured by the Kullback-Leibler divergence (Wolpert, 2004; Todorov, 2009; Peters et al., 2010; Ortega and Braun, 2011; Kappen et al., 2012). Ultimately, these costs are determined by Lagrange multiplier constraints given by the inverse temperature playing the role of a resource parameter. Here we generalize this approach to general decision trees where each node can have a different resource allocation. Consequently, we obtain generalized optimality equations

for sequential decision-making that include the well-known Bellman optimality equation as well as Expectimax-, Minimax- and Expectiminimax-decision rules depending on the limit values of the resource parameters. The resource parameters themselves are amenable to interesting computational, statistical and economic interpretations. In the first sense they measure the number of samples needed from a distribution before applying the max operator and therefore correspond directly to computational effort. In the second sense they reflect the confidence of the estimate of the maximum and therefore they can also express risk attitudes. Finally, the resource parameters reflect the control an agent has over a random variable. These different ramifications need to be explored further in the future.

Appendix A. Proofs

A.1. Proof of Theorem 2

Proof

$$\sum_x P(x)U(x) - \frac{1}{\alpha}\sum_x P(x)\log\frac{P(x)}{Q(x)} = \sum_x P(x)V(x) - \frac{1}{\beta}\sum_x P(x)\log\frac{P(x)}{Q(x)}$$

Since the equilibrium and reference distributions $P(x)$ and $Q(x)$ are constant but arbitrarily chosen, it must be that

$$U(x) - \frac{1}{\alpha}\log\frac{P(x)}{Q(x)} = V(x) - \frac{1}{\beta}\log\frac{P(x)}{Q(x)}.$$

Hence,

$$V(x) = U(x) - \left(\tfrac{1}{\alpha} - \tfrac{1}{\beta}\right)\log\frac{P(x)}{Q(x)}.$$

■

A.2. Proof of Theorem 3

Proof Let $x_1, x_2, \ldots, x_N$ be the ordering of $\mathcal{X}$ such that $U(x_1), U(x_2), \ldots, U(x_N)$. It is well known that the distribution over the maximum of α samples is equal to $F_\alpha(x) = F(x)^\alpha$, where F is the cumulative distribution $F(x_n) = \sum_{k\leq n} M(x_k)$. Defining $F(x_0) := 0$, one has $M_\alpha(x_n) = F(x_n)^\alpha - F(x_{n-1})^\alpha$. Hence, the probability can be bounded as $0 \leq M_\alpha(x_n) \leq F(x_n)^\alpha$, or

$$0 \leq M_\alpha(x_n) \leq e^{-\alpha\gamma(x_n)}, \tag{6}$$

if we use $F(x_n) = e^{-\gamma(x_n)}$ where $\gamma(x_n) \geq 0$. The Boltzmann distribution can be bounded as

$$0 \leq \frac{Q(x_n)e^{\alpha U(x_n)}}{\sum_k Q(x_k)e^{\alpha U(x_k)}} \leq \frac{Q(x_n)e^{\alpha U(x_n)}}{Q(x_N)e^{\alpha U(x_N)}}.$$

The upper bound is obtained by dropping all the summands in the expectation but the largest. In exponential form, the bounds are written as

$$0 \leq \frac{Q(x_n)e^{\alpha U(x_n)}}{\sum_k Q(x_k)e^{\alpha U(x_k)}} \leq e^{-\alpha\delta(x_n)+c(x_n)}, \tag{7}$$

where $\delta(x_n) := U(x_N) - U(x_n)$, $c(x_n) := -\log Q(x_N) + \log Q(x_n)$. Note that $\delta(x_n)$ is positive. Subtracting the inequalities (6) from (7) yields

$$-e^{-\alpha\gamma(x_n)} \leq \frac{Q(x_n)e^{\alpha U(x_n)}}{\sum_k Q(x_k)e^{\alpha U(x_k)}} - M_\alpha(x_n) \leq e^{-\alpha\delta(x_n)+c(x_n)}.$$

Choosing $\xi(x_n) = c(x_n)/\delta(x_n) \geq 0$ allows rewriting the upper bound and changing the lower bound to

$$-e^{-(\alpha-\xi(x_n))\gamma(x_n)} \leq \frac{Q(x_n)e^{\alpha U(x_n)}}{\sum_k Q(x_k)e^{\alpha U(x_k)}} - M_\alpha(x_n) \leq e^{-(\alpha-\xi(x_n))\delta(x_n)}.$$

Finally, choosing $\xi := \max_n\{\xi(x_n)\}$ and $\delta = \max\{\max_n\{\delta(x_n)\}, \max_n\{\gamma(x_n)\}\}$ yields the bounds of the theorem

$$-e^{-(\alpha-\xi)\delta} \leq \frac{Q(x_n)e^{\alpha U(x_n)}}{\sum_k Q(x_k)e^{\alpha U(x_k)}} - M_\alpha(x_n) \leq e^{-(\alpha-\xi)\delta}.$$

■

A.3. Proof of Theorem 4

Proof The free energy of the whole trajectory with inverse temperature α is given by

$$\sum_{x_{\leq T}} P(x_{\leq T})\left\{U(x_{\leq T}) - \frac{1}{\alpha}\log\frac{P(x_{\leq T})}{Q(x_{\leq T})}\right\}.$$

Using a telescopic sum $\sum_{t=1}^{T}(a_t - a_{t-1}) = a_T - a_0$ for the utilities yields

$$U(\varepsilon) + \sum_{x_{\leq T}} P(x_{\leq T})\sum_{t=1}^{T}\left\{\left[U(x_{\leq t}) - U(x_{<t})\right] - \frac{1}{\alpha}\log\frac{P(x_t|x_{<t})}{Q(x_t|x_{<t})}\right\}.$$

Using the definition of rewards (4), one gets the result

$$U(\varepsilon) + \sum_{x_{\leq T}} P(x_{\leq T})\sum_{t=1}^{T}\left\{R(x_t|x_{<t}) - \frac{1}{\beta(x_{<t})}\log\frac{P(x_t|x_{<t})}{Q(x_t|x_{<t})}\right\}.$$

■

A.4. Proof of Theorem 5

Proof The inner sum of the free energy

$$U(\varepsilon) + \sum_{x_{\leq T}} P(x_{\leq T})\sum_{t=1}^{T}\left\{R(x_t|x_{<t}) - \frac{1}{\beta(x_{<t})}\log\frac{P(x_t|x_{<t})}{Q(x_t|x_{<t})}\right\}.$$

can be expanded as

$$\begin{aligned}
U(\varepsilon) + \sum_{x_1} P(x_1)\Bigg\{ & R(x_1) - \frac{1}{\beta(\varepsilon)}\log\frac{P(x_1)}{Q(x_1)} \\
& + \sum_{x_2} P(x_2|x_1)\Bigg\{R(x_2|x_1) - \frac{1}{\beta(x_1)}\log\frac{P(x_2|x_1)}{Q(x_2|x_1)} \\
& + \cdots \\
& + \sum_{x_T} P(x_T|x_{<T})\left\{R(x_T|x_{<T}) - \frac{1}{\beta(x_{<T})}\log\frac{P(x_T|x_{<T})}{Q(x_T|x_{<T})}\right\}\cdots\Bigg\}\Bigg\}.
\end{aligned}$$

This can be solved by induction, starting with the innermost sums and then recursively solving the outer sums. The innermost sums

$$\sum_{x_T} P(x_T|x_{<T})\Big\{R(x_T|x_{<T}) - \frac{1}{\beta(x_{<T})}\log\frac{P(x_T|x_{<T})}{Q(x_T|x_{<T})}\Big\}$$

are maximized when

$$P(x_T|x_{<T}) = \frac{1}{Z(x_{<T})}Q(x_T|x_{<T})\exp\Big\{\beta(x_{<T})R(x_T|x_{<T})\Big\}.$$

This seen by noting that for probabilities p_i and positive numbers $r_i > 0$, the quantity $\sum_i p_i \log(p_i/r_i)$ is minimized by choosing $p_i = \frac{1}{Z}r_i$, where $Z = \sum_i r_i$ is just a normalizing constant. Substituting this solution yields the outer sums

$$\sum_{x_t} P(x_t|x_{<t})\Big\{R(x_t|x_{<t}) - \frac{1}{\beta(x_{<t})}\log\frac{P(x_t|x_{<t})}{Q(x_t|x_{<t})} + \frac{1}{\beta(x_{\leq t})}\log Z(x_{\leq t})\Big\}$$

where

$$Z(x_{<t}) = \sum_{x_t} Q(x_t|x_{<t})\exp\Big\{\beta(x_{<t})\big[R(x_t|x_{<t}) + \frac{1}{\beta(x_{\leq t})}\log Z(x_{\leq t})\big]\Big\}.$$

These sums are then maximized by choosing

$$P(x_t|x_{<t}) = \frac{1}{Z(x_{<t})}Q(x_t|x_{<t})\exp\Big\{\beta(x_{<t})\big[R(x_t|x_{<t}) + \frac{1}{\beta(x_{\leq t})}\log Z(x_{\leq t})\big]\Big\}.$$

■

References

R.E. Bellman. Dynamic programming, 1957.

D. A. Braun and P. A. Ortega. Path integral control and bounded rationality. In *IEEE Symposium on adaptive dynamic programming and reinforcement learning*, pages 202–209, 2011.

H.B. Callen. *Thermodynamics and an introduction to thermostatistics*. John Wiley & Sons, New York, 1985.

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

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

G. Keller. *Equilibrium States in Ergodic Theory*. London Mathematical Society Student Texts. Cambridge Univeristy Press, 1998.

D. Michie. Game-playing and game-learning automata. *Advances in Programming & Non-Numerical Computation*, pages 183–200, 1966.

P. Ortega. *A unified framework for resource-bounded autonomous agents interacting with unknown environments*. PhD thesis, Department of Engineering, University of Cambridge, UK, 2011.

P.A. Ortega and D.A. Braun. Information, utility and bounded rationality. In *Lecture notes on artificial intelligence*, volume 6830, pages 269–274, 2011.

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

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

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

H. Simon. *Models of Bounded Rationality*. MIT Press, Cambridge, MA, 1984.

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

E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences U.S.A.*, 106:11478–11483, 2009.

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