Path Integral Control and Bounded Rationality

Daniel A. Braun
Univ. Southern California
Los Angeles, USA
dab54@cam.ac.uk

Pedro A. Ortega
Univ. Cambridge
Cambridge, UK
peortega@dcc.uchile.cl

Evangelos Theodorou
Univ. Southern California
Los Angeles, USA
etheodor@usc.edu

Stefan Schaal
Univ. Southern California
Los Angeles, USA
sschaal@usc.edu

***Abstract*—Path integral methods [7], [15],[1] have recently been shown to be applicable to a very general class of optimal control problems. Here we examine the path integral formalism from a decision-theoretic point of view, since an optimal controller can always be regarded as an instance of a perfectly rational decision-maker that chooses its actions so as to maximize its expected utility [8]. The problem with perfect rationality is, however, that finding optimal actions is often very difficult due to prohibitive computational resource costs that are not taken into account. In contrast, a bounded rational decision-maker has only limited resources and therefore needs to strike some compromise between the desired utility and the required resource costs [14]. In particular, we suggest an information-theoretic measure of resource costs that can be derived axiomatically [11]. As a consequence we obtain a variational principle for choice probabilities that trades off maximizing a given utility criterion and avoiding resource costs that arise due to deviating from initially given default choice probabilities. The resulting bounded rational policies are in general probabilistic. We show that the solutions found by the path integral formalism are such bounded rational policies. Furthermore, we show that the same formalism generalizes to discrete control problems, leading to linearly solvable bounded rational control policies in the case of Markov systems. Importantly, Bellman's optimality principle is not presupposed by this variational principle, but it can be derived as a limit case. This suggests that the information-theoretic formalization of bounded rationality might serve as a general principle in control design that unifies a number of recently reported approximate optimal control methods both in the continuous and discrete domain.**

I. Introduction

In decision theory, a decision-maker is typically assumed to have preferences over a set of objects that are part of a choice set [8]. These objects can be simple monolithic objects (e.g. different fruits in a fruit bowl) or complex lotteries of processes with random outcomes (e.g. a sequence of roulette wheels with different outcomes and outcome probabilities). If the decision-makers' preferences fulfill certain axioms like completeness and transitivity then the preferences can be represented by a utility function over outcomes and the decision-maker's choices can be modeled as the maximization of the expected utility [18], [4]. Such decision-makers are called (perfectly) rational [12]. Optimal control is a particular instance of rational decision-making, where the choice set is given by all possible probability distributions over trajectories conditioned on different control laws.

While perfect rationality provides a sound normative basis for decision-making, the problem is that finding optimal actions can be a very difficult problem, because the search for the optimal action is itself associated with a cost that is not accounted for by the principle of maximum expected utility. Therefore, the concept of *bounded* rationality has been propounded to characterize decision-makers that have limited resources and cannot afford an unlimited search of the optimal action [14], [13]. Instead, bounded rational actors trade off the utility that an action achieves against the cost of finding the action. In the following we formalize bounded rational decision-making based on three axioms relating choice probabilities and utilities, which leads in general to probabilistic choice behavior that cannot be formalized by classic decision theory [11], [10], [9]. These choice probabilities also imply a variational principle that we use to derive bounded rational controllers. We show that the path integral approach [6], [7], [15] fits within this framework. Furthermore, we show that a related approach to discrete control [16] can also be explained by the same framework. The main contribution of the current paper is therefore not so much on the algorithmic level, but rather on the conceptual and mathematical level, providing a unifying framework for a number of recently published approximate optimal control methods.

II. Bounded Rationality

A. Resource Costs

A bounded rational decision-maker should not only consider the gain in expected utility of a particular choice, but also the resource costs entailed by this choice [14]. This raises the question of how to measure resources. Here we follow a thermodynamic argument [3] that allows measuring resource (or information) costs in physical systems in units of energy. The generality of the argument relies on the fact that ultimately any real agent has to be incarnated in a physical system, as any process of information processing must always be accompanied by a pertinent physical process [17]. In the following we conceive of information processing as changes in information states (i.e. ultimately changes of probability distributions), which consequently implies changes in physical states, such as flipping gates in a transistor, changing voltage on a microchip, or even changing location of a gas particle. Imagine, for example, that we use a gas particle in a box with volume V_i as an information processing system to represent a uniform probability distribution over a random variable with $p_i = \frac{1}{V_i}$. If we now want to update this probability to p_f, because we gained information $-\log p = -\log \frac{p_i}{p_f} > 0$, we have to reduce the original volume to $V_f = pV_i$. However, this

decrease in volume requires the work $W = -\int_{V_i}^{V_f} \frac{NkT}{V} dV = NkT \ln \frac{V_i}{V_f}$, where N is the number of gas molecules, k is the Boltzmann constant, and T is temperature. Thus, in this simple example we can compute the relation between the change in information state and the required work, that is $W = -\alpha \log p$, with $\alpha = \frac{kT}{\log e} > 0$ being the conversion factor between information and energy. Depending on the physical properties of the system this conversion factor is going to make information processing more or less expensive. In the next two sections, we derive a general expression of information costs for physical systems that implement optimal controllers. As such controllers optimize utility functions (or cost functions), we will first investigate the relation between information states and utilities and then show how information costs appear as an additional term in the utility in physically implemented controllers.

B. Conversion between Utility and Information

Consider a sample set Ω of possible outcomes and a σ-algebra $\mathcal{F}$ of measurable events between which a decision-maker can choose. We assume that the decision-maker can *freely choose* any probability measure $\mathbf{P}$ over the outcomes and that the decision-making process can consequently be modeled as a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. To formalize choice behavior, we postulate that a decision-maker (stochastically) prefers event A over event B if $\mathbf{P}(A) > \mathbf{P}(B)$. As in the case of classic decision-theory, we would like to express this preference with a desirability or utility function $\mathbf{U}$. We have previously proposed the following postulates [11], [10], [9].

Definition. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A function $\mathbf{U} : \mathcal{F} \to \mathbb{R}$ is a utility function for $\mathbf{P}$ iff the conditional utility $\mathbf{U}(A|B) := \mathbf{U}(A \cap B) - \mathbf{U}(B)$ has the following three properties for all events $A, B, C, D \in \mathcal{F}$:

- real-valued: $\exists f, \mathbf{U}(A|B) = f\left(\mathbf{P}(A|B)\right) \in \mathbb{R}$
- additive: $\mathbf{U}(A \cap B|C) = \mathbf{U}(A|C) + \mathbf{U}(B|A \cap C)$
- monotonic: $\mathbf{P}(A|B) > \mathbf{P}(C|D)$ $\Leftrightarrow$ $\mathbf{U}(A|B) > \mathbf{U}(C|D)$

Note that the conditional utility measures differences between utilities, similar to scalar potentials in physics. Accordingly, the absolute values of utilities are irrelevant and the only thing that matters are utility differences. As the following theorem shows, these postulates enforce a unique mapping between the utility and the probability space.

Theorem. If a mapping f is such that $\mathbf{U}(A|B) = f(\mathbf{P}(A|B))$ for any probability space $(\Omega, \mathcal{F}, \mathbf{P})$, then f is of the form

$$f(\cdot) = \alpha \log(\cdot), \tag{1}$$

where $\alpha > 0$ is arbitrary strictly positive constant. The proof is provided elsewhere [11], [10], [9]. The constant α provides the conversion factor between utilities and information.

A probability measure $\mathbf{P}$ and a utility function $\mathbf{U}$ that satisfy (1) form a *conjugate pair*. Accordingly, given a utility $\mathbf{U}$ over a set of measurable events, its corresponding probability can be determined by the Gibbs measure with temperature α and energy levels $-\mathbf{U}(\omega)$, i.e. the measure given by

$$\mathbf{P}(A) = \frac{\sum_{\omega \in A} e^{\frac{1}{\alpha}\mathbf{U}(\omega)}}{\sum_{\omega \in \Omega} e^{\frac{1}{\alpha}\mathbf{U}(\omega)}} \tag{2}$$

for all $A \in \mathcal{F}$ and $\mathbf{U}(\omega) = \mathbf{U}(\{\omega\})$. As we will see further down, this log/exp-conversion between the probability and utility space also lies at the heart of the path integral formalism. It was originally proposed by Schrödinger and has been introduced into control theory by Fleming [5]. Here we derived it axiomatically from decision-theoretic considerations.

C. A Variational Principle for Bounded Rationality

It is well known in statistical physics that the Gibbs measure satisfies a variational problem in the free energy [2]. Since utilities correspond to negative energies, we can formulate a *free utility* that is maximized by a decision-maker that acts according to (2).

Theorem. Let X be a random variable with values in $\mathcal{X}$. Let $\mathbf{P}$ and $\mathbf{U}$ be a conjugate pair of probability measure and utility function over X. Define the free utility functional as

$$\mathbf{J}(\mathbf{Pr}; \mathbf{U}) = \sum_{x \in \mathcal{X}} \mathbf{Pr}(x)\mathbf{U}(x) - \alpha \sum_{x \in \mathcal{X}} \mathbf{Pr}(x) \log \mathbf{Pr}(x), \tag{3}$$

where $\mathbf{Pr}$ is an arbitrary probability measure over X. Then,

$$\mathbf{J}(\mathbf{Pr}; \mathbf{U}) \leq \mathbf{J}(\mathbf{P}; \mathbf{U}) = \mathbf{U}(\Omega).$$

The proof can be obtained by applying Jensen's inequality. By inserting the Gibbs measure $\mathbf{P}$ into the free utility equation $\mathbf{J}$, we find that the extremal of the free utility is given by the log-partition sum $\mathbf{J}(\mathbf{P}; \mathbf{U}) = \alpha \log \sum_{\omega \in \Omega} e^{\frac{1}{\alpha}\mathbf{U}(\omega)}$.

This variational principle also allows conceptualizing transformations of stochastic systems. Consider an initial system described by the conjugate pair $\mathbf{P}_i$ and $\mathbf{U}_i$. This system has free utility $\mathbf{J}_i(\mathbf{P}_i, \mathbf{U}_i)$. We now want to transform this initial system into another system by adding new constraints represented by the utility function $\mathbf{U}_*$. Then, the resulting utility function $\mathbf{U}_f$ is given by the sum

$$\mathbf{U}_f = \mathbf{U}_i + \mathbf{U}_*,$$

and the resulting system has the free utility $\mathbf{J}_f(\mathbf{P}_f, \mathbf{U}_f)$. The difference in free utility is

$$\mathbf{J}_f - \mathbf{J}_i = \sum_{x \in \mathcal{X}} \mathbf{P}_f(x)\mathbf{U}_*(x) - \alpha \sum_{x \in \mathcal{X}} \mathbf{P}_f(x) \log \frac{\mathbf{P}_f(x)}{\mathbf{P}_i(x)}. \tag{4}$$

In physical systems with constant α, this difference measures the amount of work necessary to change the state of the system from state i to state f. The first term of the equation measures the expected utility difference $\mathbf{U}_*(x)$ in the final state f, while the second term measures the information cost of transforming the probability distribution from state i to state f. These two terms can be interpreted as determinants of bounded rational decision-making in that they formalize a trade-off between an expected utility $\mathbf{U}_*$ (first term) and the information cost of transforming $\mathbf{P}_i$ into $\mathbf{P}_f$ (second

term). In this interpretation $\mathbf{P}_i$ represents an initial choice probability or policy, which includes the special case of the uniform distribution where the decision-maker has initially no preferences between the different choices. The probability $\mathbf{P}_f$ is the final choice probability that we are looking for since it considers the utility constraint $\mathbf{U}^*$ that we want to optimize. We can then formulate a variational principle for bounded rationality in the probabilities $\mathbf{P}_f(x)$

$$\arg\max_{\mathbf{P}_f(x)} \left(\sum_{x\in\mathcal{X}} \mathbf{P}_f(x)\mathbf{U}_*(x) - \alpha \sum_{x\in\mathcal{X}} \mathbf{P}_f(x) \log \frac{\mathbf{P}_f(x)}{\mathbf{P}_i(x)} \right). \tag{5}$$

The solution to this variational problem is given by

$$\mathbf{P}_f(x) \propto \mathbf{P}_i(x) \exp\left(\frac{1}{\alpha}\mathbf{U}_*(x)\right).$$

In particular, at very low temperature $\alpha \approx 0$, the maximum expected utility principle is recovered as

$$\mathbf{J}_f - \mathbf{J}_i \approx \sum_{x\in\mathcal{X}} \mathbf{P}_f(x)\mathbf{U}_*(x),$$

and hence resource costs are ignored in the choice of $\mathbf{P}_f$, leading to $\mathbf{P}_f \approx \delta_{x^*}(x)$, where $x^* = \arg\max_x \mathbf{U}_*(x)$. Similarly, at a high temperature, the difference is

$$\mathbf{J}_f - \mathbf{J}_i \approx -\alpha \sum_{x\in\mathcal{X}} \mathbf{P}_f(x) \log \frac{\mathbf{P}_f(x)}{\mathbf{P}_i(x)},$$

and hence only resource costs matter, leading to $\mathbf{P}_f \approx \mathbf{P}_i$.

D. Examples

In the following we discuss some toy examples to provide more intuition about bounded rationality in simple systems.

1) Information Costs in Control: Consider a binary potential well with two states L and R that have initial potentials $V_i(L) = V_i(R) = V_0$, such that a particle will assume either state with equal probability $P_i(L) = P_i(R) = \frac{1}{2}$. If we want to control the particle to stay in state L, for example, we can create the new potentials $V_f(L) = V_i(L)$ and $V_f(R) = V_i(R) + \Delta V$ with $\Delta V \geq 0$ so that we obtain the new state probabilities $P_f(L) = \frac{1}{1+e^{-\frac{1}{\alpha}\Delta V}}$ and $P_f(R) = \frac{1}{1+e^{\frac{1}{\alpha}\Delta V}}$. The work required for the state transition from i to f is given by the free utility difference that has the limit $-\alpha \log 2$ for $\Delta V \to \infty$ (a potential wall), as we gain at most 1 bit of information, whose equivalent we have to pay as work.

2) Precision Limits of Optimization: Consider the simple optimization problem of finding the maximum of the utility $U(x) = -\frac{1}{2}(x-\bar{x})^2$. Obviously, the optimal answer is $\mathbf{P}(x) = \delta(x-\bar{x})$. However, imagine that we want to implement this decision-making process in a physical system given by a particle in a potential $V_i(x) = -\frac{1}{2}(x-\bar{x})^2$ with the equilibrium distribution $\mathbf{P}_i(x) = \frac{1}{Z_i}\exp\left(-\frac{1}{2\alpha}(x-\bar{x})^2\right)$. As a consequence, we can measure x only with variance α. If we want to have a more precise measurement at this temperature, say $\mathbf{P}_f(x) = \frac{1}{Z}\exp\left(-\frac{1}{2\alpha}k(x-\bar{x})^2\right)$ with $k \geq 1$, then we could add the potential $V_*(x) = -\frac{1}{2}(k-1)(x-\bar{x})^2$ and measure the new equilibrium distribution $\mathbf{P}_f$. The difference in free utility between these two distributions is given by $\mathbf{J}_f - \mathbf{J}_i = -\frac{\alpha}{2}\log k$. We can see that this log term arises from information costs when we set $V^*(x) = 0$, that means we only consider fluctuations of the gas. As was the case for the ideal gas example of Section II-A, the free utility difference in fluctuating systems is exclusively given by the information cost. If the gas was to assume the distribution $\mathbf{P}_f(x) = \frac{1}{Z}\exp\left(-\frac{1}{2\alpha}k(x-\bar{x})^2\right)$ with $k \geq 1$ just by chance through random fluctuations, the free utility difference between this non-equilibrium distribution $\mathbf{P}_f$ and the equilibrium $\mathbf{P}_i$ is given by the relative entropy

$$-\alpha \int dx \mathbf{P}_f(x) \log \frac{\mathbf{P}_f(x)}{\mathbf{P}_i(x)} = \frac{\alpha}{2}\left(-\frac{1}{k} + 1 - \log k\right).$$

Thus, the term $-\frac{\alpha}{2}\log k < 0$ arises essentially as an information cost that implies that we have to spend work to get more information so we can increase precision. In fact, infinite precision would require infinite amount of work or energy resources as $-\frac{\alpha}{2}\log k \to -\infty$ for $k \to \infty$. The mathematically obvious solution $\mathbf{P}(x) = \delta(x-\bar{x})$ therefore turns out to be rather expensive. The only way to get a cheaper result, is by finding a physical process with low conversion factor α, as $\lim_{\alpha\to 0}\mathbf{P}_f(x) = \delta(x-\bar{x})$. However, according to the third law of thermodynamics the limit $\alpha = 0$ cannot be achieved, and therefore infinite precision required by perfect rationality remains elusive in real systems.

3) One-Step Control: Assume we are given a system with initial state x_0 and can exert a control command u which is added to x_0 to achieve the final state $x = x_0 + u$. The target utility is $\mathbf{U}(x) = -\frac{1}{2}kx^2$, that is we want to control x to be close to zero. Furthermore, let the initial control policy be given by $\mathbf{P}_0(u) = \mathcal{N}(0,\sigma^2)$. Our aim is to find the bounded rational controller $\mathbf{P}(u)$. According to the variational principle (5), we can express $\mathbf{P}(u) \propto \mathbf{P}_0(u)e^{-\frac{1}{\alpha}\frac{k}{2}(x_0+u)^2}$, which results in the Gaussian distribution

$$\mathbf{P}(u) = \frac{1}{Z}\exp\left(-\frac{1}{2}\left(\frac{k}{\alpha}+\frac{1}{\sigma^2}\right)\left(u + \frac{x_0}{1+\frac{\alpha}{k\sigma^2}}\right)^2\right).$$

In the limit of perfect rationality $\alpha \to 0$, the controller becomes deterministic $\mathbf{P}(u) = \delta(u+x_0)$. Note that the same variational problem could have been formulated directly in x-space, since there is a direct mapping between x and u. The initial distribution is then $\mathbf{P}_0(x) = \mathcal{N}(x_0,\sigma^2)$, and the bounded rational solution is

$$\mathbf{P}(x) = \frac{1}{Z}\exp\left(-\frac{1}{2}\left(\frac{k}{\alpha}+\frac{1}{\sigma^2}\right)\left(x - \frac{x_0}{1+\frac{k\sigma^2}{\alpha}}\right)^2\right).$$

Both solutions are of course equivalent, as $\mathbf{P}(u)$ can be retrieved from $\mathbf{P}(x)$ by substituting x with $u + x_0$. It should be emphasized, however, that in both cases we should think about the stochasticity as arising from the control process, rather than assuming a classical deterministic controller with state noise.

4) One-Step Control with Constant Variance: In the previous example both mean and variance changed in the transformation from $\mathbf{P}_0(u)$ to $\mathbf{P}(u)$. Here we want to change only the mean $< u >$, and keep the variance $Var(u) = \sigma^2$ the same. We can incorporate this additional constraint into the variational principle (5) by directly assuming the form $\mathbf{P}(u) = \mathcal{N}(\mu, \sigma^2)$ for $\mathbf{P}_f$. We then insert $\mathbf{P}(u)$ into the variational functional (5) and get a parametric optimization problem in μ:

$$\begin{aligned}\mu^* &= \arg\max_\mu \left\{ -\int du \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}} \frac{1}{2} k(x_0+u)^2 \right. \\ &- \left. \alpha \int du \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}} \log \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{u^2}{\sigma^2}}} \right\}\end{aligned}$$

Consequently, we have

$$\mu^* = \arg\min_\mu \left\{ \frac{1}{2} k \left(\sigma^2 + (\mu + x_0)^2\right) + \alpha \frac{\mu^2}{2\sigma^2} \right\}$$

which results in $\mu^* = -\frac{x0}{1+\frac{\alpha}{k\sigma^2}}$. The result is identical to the mean $< u >= \int du \mathbf{P}(u) u$ obtained in the previous example. Similarly, the mean control signal $< u >$ can also be obtained from $\mathbf{P}(x)$ as $< u >= \int dx \mathbf{P}(x) u(x)$, where $u(x) = x - x_0$. Since $u(x)$ is linear in x_0, another way to retrieve the same result is by taking the derivative of the log partition sum with respect to x_0. To this end, we note that $Z = \int dx e^{-\frac{1}{\alpha} S(x)}$, with $S(x) = \frac{1}{2}\alpha \frac{(x-x_0)^2}{\sigma^2} + \frac{1}{2} k x^2$ and

$$\frac{\partial}{\partial x_0} \log Z = -\frac{1}{\alpha Z} \int dx e^{-\frac{1}{\alpha} S(x)} \frac{\partial S}{\partial x_0}.$$

The derivative of $S(x)$ with respect to the initial state x_0 is $\frac{\partial S}{\partial x_0} = -\frac{\alpha u(x)}{\sigma^2}$. Consequently, we can write the log partition sum as an expectation in $u(x)$

$$\frac{\partial}{\partial x_0} \log Z = \frac{1}{\sigma^2} \int dx u(x) \mathbf{P}(x) = \frac{< u >}{\sigma^2}$$

with $\mathbf{P}(x) = \frac{e^{-\frac{1}{\alpha} S(x)}}{\int dx' e^{-\frac{1}{\alpha} S(x')}}$. We will use a similar line of thought later for the path integrals.

III. Path Integral Control

A. Variational Principle for Paths

In the case of space- and time-continuous control, the objects of decision-making are system trajectories or paths $\mathfrak{x}$ made up of trajectory points $x(t)$ with $t_0 \leq t \leq T$. The utility of paths is given by functionals $\mathbf{U}[\mathfrak{x}]$. Similarly, the probability distributions $\mathbf{P}[\mathfrak{x}]$ become functionals over paths. The variational principle is then formulated over paths as well. We are looking for the distribution $\mathbf{P}[\mathfrak{x}]$ that maximizes the functional

$$\int D\mathfrak{x} \mathbf{P}[\mathfrak{x}] \mathbf{U}[\mathfrak{x}] - \alpha \int D\mathfrak{x} \mathbf{P}[\mathfrak{x}] \log \frac{\mathbf{P}[\mathfrak{x}]}{\mathbf{P}_0[\mathfrak{x}]}, \tag{6}$$

where $\mathbf{P}_0[\mathfrak{x}]$ corresponds to an initially given distribution. The integrals are to be understood as path integrals. As discussed in the previous section, the solution to this variational problem is given by

$$\mathbf{P}[\mathfrak{x}] \propto \mathbf{P}_0[\mathfrak{x}] \exp\left(\frac{1}{\alpha} \mathbf{U}[\mathfrak{x}]\right). \tag{7}$$

In the following we will assume that the path cost can be obtained by summing up an instantaneous utility and a final utility, that is

$$\mathbf{U}[\mathfrak{x}] := -\int_{t_0}^{T} dt q(x(t), t) - \alpha \log \psi(x(T)).$$

The normalization factor in (7) is then given by the path integral

$$\Psi = \int D\mathfrak{x} \mathbf{P}_0[\mathfrak{x}] e^{-\frac{1}{\alpha}\int dt q(x(t),t)} \psi(x(T)). \tag{8}$$

If the initial distribution $\mathbf{P}_0[\mathfrak{x}]$ is given by a diffusion process $dx = \mu(x,t)dt + \sqrt{\alpha}\sigma(x,t)dw$, then we can write the partition sum Z as the path integral

$$Z = \int D\mathfrak{x} e^{-\frac{1}{\alpha}\int dt \left(\frac{(\dot{x}-\mu(x,t))^2}{2\sigma^2(x,t)} + q(x(t),t)\right)} \psi(x(T)).$$

In fact the partition sum can be formulated for any starting position $x(t_1)$ of the path $\mathfrak{x}_{t_1:T}$ with $t_0 \leq t_1 \leq T$ as

$$Z(x(t_1), t_1) = \int D\mathfrak{x}_{t_1:T} e^{-\frac{1}{\alpha}\int_{t_1}^{T} dt \left(\frac{(\dot{x}-\mu(x,t))^2}{2\sigma^2(x,t)} + q(x(t),t)\right)} \psi(x(T)).$$

We will need the partition sum in the following when computing expectation values of the bounded rational controller.

B. Problem Formulation

The optimal control problem addressed by the path integral formalism has the following state dynamics

$$\dot{x} = f(x,t) + g(x)u,$$

utility function

$$\mathbf{U}(x,t) = -q(x,t),$$

and terminal cost

$$\phi(x(T)) = \alpha \log(\psi(x(T))),$$

where x is the state vector, f the nonlinear state transition function, g the control gain, u the control signal, q the state cost, and T the time horizon.

For reasons of mathematical tractability, we will restrict our admissible controllers to the set of diffusion processes of the form $du = \mu(x,t)dt + \sqrt{\alpha}\sigma dw$ with known diffusion constant $\sqrt{\alpha}\sigma$, but unknown drift μ. We can then use the partition sum to compute the drift as the expectation over all possible controls (see example above). Since there is a linear mapping between u and x, we also have a diffusion in x given by $dx = f(x,t)dt + g(x)du$, where the dynamics f play the role of a baseline drift and du is a diffusion process with drift μ and diffusion $\sqrt{\alpha}\sigma$. In total, we then get the diffusion process

$$dx = \Big(f(x,t) + g(x)\mu(x,t)\Big)dt + \sqrt{\alpha}\sigma g(x) dw \tag{9}$$

with drift $f(x,t)+g(x)\mu(x,t)$ and a state-dependent diffusion constant $\sqrt{\alpha}\sigma g(x)$. Thus, the bounded rational control problem consists of finding the optimal drift μ of a stochastic controller realized by a diffusion process.

C. Mean Controls

As discussed in the previous example, we can use the partition sum to compute the mean drift in controls. If we discretize the trajectories $\mathfrak{x}$ into N equidistant points $x(t_j)$ with $j=1,\ldots,N$, we can write the partition sum $Z(x,t_i)$ as

$$Z(x,t_i) = \lim_{\Delta t\to 0}\int D\mathfrak{x}_{t_i:T}\exp\left(-\frac{1}{\alpha}S[\mathfrak{x}_{t_i:T}]\right)$$

with $\Delta t = \frac{T}{N}$ and

$$\begin{aligned} S[\mathfrak{x}_{t_i:T}] &= \phi(x(T)) + \sum_{j=i}^{N-1} q(x_{t_j})\Delta t \\ &+ \frac{1}{2}\sum_{j=i}^{N-1}\frac{\left(\frac{x_{t_{j+1}}-x_{t_j}}{\Delta t} - f_{t_j}\right)^2}{g_{t_j}\sigma^2 g_{t_j}}. \end{aligned}$$

The derivative of the log partition sum with respect to the initial state x_{t_i} is then given by

$$\nabla_{x_{t_i}}\log Z = \lim_{\Delta t\to 0} -\frac{1}{\alpha}\int D\mathfrak{x}_{t_i:T}\mathbf{P}[\mathfrak{x}_{t_i:T}]\nabla_{x_{t_i}}S[\mathfrak{x}_{t_i:T}]$$

with

$$\mathbf{P}[\mathfrak{x}_{t_i:T}] = \frac{\exp\left(-\frac{1}{\alpha}S[\mathfrak{x}_{t_i:T}]\right)}{\int D\mathfrak{x}_{t_i:T}\exp\left(-\frac{1}{\alpha}S[\mathfrak{x}_{t_i:T}]\right)}$$

and

$$\nabla_{x_{t_i}}S[\mathfrak{x}_{t_i:T}] = -\frac{1}{2\Delta t}\zeta_{t_i} - \frac{1}{2}\beta_{t_i}\nabla_{x_{t_i}}f_{t_i} + \frac{1}{2\Delta t}\beta_{t_i}\nabla_{x_{t_i}}\zeta_{t_i}$$

with

$$\begin{aligned} \beta_{t_i} &= x_{t_{i+1}} - x_{t_i} - f_{t_i}\Delta t \\ \zeta_{t_i} &= \frac{\beta_{t_i}}{g_{t_i}\sigma^2 g_{t_i}}. \end{aligned}$$

If we take the limit $\Delta t\to 0$ of $\nabla_{x_{t_i}}S[\mathfrak{x}_{t_i:T}]$ we find

$$\lim_{\Delta t\to 0}\nabla_{x_{t_i}}S[\mathfrak{x}_{t_i:T}] = -\frac{\lim_{\Delta t\to 0}\frac{x_{t_{i+1}}-x_{t_i}}{\Delta t} - f_{t_i}}{g_{t_i}\sigma^2 g_{t_i}}.$$

We realize, that we can express this in terms of the control $u(x_{t_{i+1}}, x_{t_i})$ as

$$\lim_{\Delta t\to 0}\nabla_{x_{t_i}}S[\mathfrak{x}_{t_i:T}] = -\frac{u(x_{t_{i+1}},x_{t_i})}{\sigma^2 g_{t_i}}.$$

Consequently, we can express the derivative of the log partition sum as an expectation value over the controls

$$\nabla_{x_t}\log Z = \frac{1}{\alpha\sigma^2 g(x)}\int \mathbf{P}[\mathfrak{x}_{t:T}]u[\mathfrak{x}_{t:T}],$$

where $u[\mathfrak{x}_{t:T}]$ is the control trajectory associated with the state path $\mathfrak{x}_{t:T}$. Finally, we can express the mean controls in terms of the log partition sum

$$< u_t >= \sigma^2 g(x)\alpha\ \nabla_{x_t}\log Z.$$

Thus, the bounded rational controller has mean drift $< u_t >$ and variance $\alpha\sigma^2$.

D. Connection to Hamilton-Jacobi-Bellman equation

As outlined above, the perfectly rational limit of the bounded rational controller can be obtained for $\alpha\to 0$. In the present case, this leads to diverging controls as there are no control costs in the problem statement and therefore the best strategy is to apply infinite control. However, there is also another way to link the above framework to traditional optimal control based on the Hamilton-Jacobi-Bellman equation. To this end, we realize that the normalization factor of (8) can be written as a partial differential equation if $\mathbf{P}_0[\mathfrak{x}]$ is given by a diffusion process. In the case of the above problem statement, the diffusion process of the initial controller is given by (9) with $\mu=0$. Following the Feynman-Kac formula, we then realize that the path integral for $\Psi(x,t)$ can be expressed as the partial differential equation:

$$\frac{\partial\Psi}{\partial t} + f(x,t)\frac{\partial\Psi}{\partial x} + \frac{1}{2}g(x)\sigma^2 g(x)\frac{\partial^2\Psi}{\partial x^2} = q(x,t)\Psi$$

with the boundary condition $\Psi(x,T)=\psi(x(T))$. Dividing by Ψ and introducing the transform

$$V = -\alpha\log\Psi$$

one notes that

$$\begin{aligned} \frac{\partial\Psi}{\partial t}\frac{1}{\Psi} &= -\frac{1}{\alpha}\frac{\partial V}{\partial t} \\ \frac{\partial\Psi}{\partial x}\frac{1}{\Psi} &= -\frac{1}{\alpha}\frac{\partial V}{\partial x} \\ \frac{\partial^2\Psi}{\partial x^2}\frac{1}{\Psi} &= -\frac{1}{\alpha}\frac{\partial^2 V}{\partial x^2} + \frac{1}{\alpha^2}\left(\frac{\partial V}{\partial x}\right)^2. \end{aligned}$$

This leads to the new partial differential equation

$$\begin{aligned} -\frac{\partial V}{\partial t} &= q(x,t) + f(x,t)\frac{\partial V}{\partial x} + \frac{1}{2}g^2(x,t)\sigma^2\frac{\partial^2 V}{\partial x^2} \\ &- \frac{1}{\alpha}\frac{1}{2}g^2(x,t)\sigma^2\left(\frac{\partial V}{\partial x}\right)^2 \end{aligned} \tag{10}$$

This differential equation can be compared to the Hamilton-Jacobi-Bellman equation for the same system with control costs $c(x,u) = q(x) + \frac{1}{2}uRu$. The Hamilton-Jacobi-Bellman equation is then given by

$$\begin{aligned} -\frac{\partial V}{\partial t} &= \min_u\left(q(x,t) + \frac{1}{2}uRu + f(x,t)\frac{\partial V}{\partial x}\right. \\ &+ \left.\frac{1}{2}g^2(x,t)\sigma^2\frac{\partial^2 V}{\partial x^2}\right) \end{aligned}$$

Since the left hand side is quadratic in u, we can solve the minimization problem analytically by

$$u(x_t) = -R^{-1}g(x)\nabla_x V(x,t).$$

The Hamilton-Jacobi-Bellman equation then becomes a diffusion equation similar to (10)

$$\begin{aligned} -\frac{\partial V}{\partial t} &= q(x,t) + f(x,t)\frac{\partial V}{\partial x} + \frac{1}{2}g^2(x,t)\sigma^2\frac{\partial^2 V}{\partial x^2} \\ &- \frac{1}{2}g(x,t)R^{-1}g(x,t)\left(\frac{\partial V}{\partial x}\right)^2. \end{aligned} \tag{11}$$

The two equations (11) and (10) are in fact identical if we require that

$$\frac{\sigma^2}{\alpha} = R^{-1}.$$

This requirement implies that we interpret the "information costs" of the bounded rational controller as a "control cost". Mathematically, this interpretation works out for quadratic control cost functions and Gaussian control distributions, because the information cost between different Gaussian controllers with different means but same variance turns out to be quadratic. As outlined above information costs are measured by the relative entropy between initial and desired controls $\mathbf{P}_0(u)$ and $\mathbf{P}(u)$. If these two distributions are Gaussian with the same variance and means $\mu_1 = \mu$ and $\mu_2 = 0$ respectively, then the information cost is

$$\alpha \int du \mathbf{P}(u) \log \frac{\mathbf{P}(u)}{\mathbf{P}_0(u)} = \frac{1}{2}\alpha \frac{\mu^2}{\sigma^2}.$$

Thus, the relationship $\frac{\sigma^2}{\alpha} = R^{-1}$ implies that the information cost is interpreted as a control cost.

IV. Discrete Control

In the following we show that the same principles can be carried over to discrete control. The variational principle is then formulated over sequences of random variables $x_{1:T} = x_1 x_2 \dots x_T$, where T is the time horizon. The utility of these sequences is given by a function $\mathbf{U}(x_{1:T})$. We are looking for the distribution $\mathbf{P}(x_{1:T})$ that maximizes the functional

$$\sum_{x_{1:T}} \mathbf{P}(x_{1:T})\mathbf{U}(x_{1:T}) - \alpha \sum_{x_{1:T}} \mathbf{P}(x_{1:T}) \log \frac{\mathbf{P}(x_{1:T})}{\mathbf{P}_0(x_{1:T})},$$

where $\mathbf{P}_0(x_{1:T})$ corresponds again to an initially given distribution, which includes the uniform as a special case. As previously, the solution is given by

$$\mathbf{P}(x_{1:T}) \propto \mathbf{P}_0(x_{1:T}) \exp\left(\frac{1}{\alpha}\mathbf{U}(x_{1:T})\right).$$

In the following we will assume that we are dealing with Markov systems

$$\begin{aligned} \mathbf{P}(x_{1:T}) &= \prod_{i=1}^{T} P(x_i|x_{i-1}) \\ \mathbf{P}_0(x_{1:T}) &= \prod_{i=1}^{T} P_0(x_i|x_{i-1}) \end{aligned}$$

and that the utilities are given by costs that are additive and state-dependent

$$\mathbf{U}(x_{1:T}) = -\sum_{i=1}^{T} q(x_i).$$

The variational principle can then be written in the following recursive form

$$\begin{aligned} & \sum_{x_1} P(x_1) \Bigg[-q(x_1) - \alpha \log \frac{P(x_1)}{P_0(x_1)} \\ + & \sum_{x_2} P(x_2|x_1) \Bigg[-q(x_2) - \alpha \log \frac{P(x_2|x_1)}{P_0(x_2|x_1)} \\ + & \cdots \\ + & \sum_{x_T} P(x_T|x_{T-1}) \Bigg[-q(x_T) - \alpha \log \frac{P(x_T|x_{T-1})}{P_0(x_T|x_{T-1})} \Bigg] \dots \Bigg]. \end{aligned}$$

Consequently, the innermost variational problem can be solved independently of the other ones, resulting in

$$P(x_T|x_{T-1}) = \frac{1}{\Psi_T} P_0(x_T|x_{T-1}) e^{-\frac{1}{\alpha}q(x_T)},$$

where Ψ_T is the normalization factor that is a function of x_{T-1}. If we re-insert this solution into the above maximization problem, we can then solve the variational problem for x_{T-1}, which results in

$$P(x_{T-1}|x_{T-2}) = \frac{1}{\Psi_{T-1}} P_0(x_{T-1}|x_{T-2}) e^{-\frac{1}{\alpha}q(x_{T-1}) + \log \Psi_T}.$$

We can proceed likewise for all the other probabilities until we reach $P(x_1)$. This procedure also imposes a recursive relationship between the normalization factors

$$\Psi_t = \sum_{x_t} P_0(x_t|x_{t-1}) e^{-\frac{1}{\alpha}q(x_t)} \Psi_{t+1}. \tag{12}$$

Iterating these recursive equations in Ψ fully determines the probabilities $P(x_t|x_{t-1})$ for all t and therefore solves the optimization problem.

Instead of iterating Ψ_t we can also define the quantity $z_T = \exp(-\frac{1}{\alpha}q(x_T))$ and $z_{T-1} = \exp(-\frac{1}{\alpha}q(x_{T-1}) + \log \Psi_T)$ and so forth, to get the recursion

$$z_{t-1} = e^{-\frac{1}{\alpha}q(x_{t-1})} \sum_{x_t} P_0(x_t|x_{t-1}) z_t$$

These z-values can be used to obtain the final probabilities through

$$P(x_t|x_{t-1}) = \frac{P_0(x_t|x_{t-1}) z_t}{\sum_{x_t} P_0(x_t|x_{t-1}) z_t}.$$

This z-iteration was suggested in [16] to solve MDPs. Since we deal with Markov systems, both recursions can also be formulated in matrix form, which allows solving the optimization problem through methods of linear algebra.

In the perfectly rational limit $\alpha \to 0$ we can recover the Bellman optimality equation from (12). Again we start with the final time step and take the transform $V_t = -\alpha \log \Psi_t$ which results in

$$V_T = -\alpha \log \sum_{x_T} P_0(x_T|x_{T-1}) e^{-\frac{1}{\alpha}q(x_T)}$$

and the limit

$$V_T^* = \lim_{\alpha \to 0} V_T = \min_{x_T} q(x_T).$$

Analogously, we get in the preceding time step

$$V_{T-1} = -\alpha \log \sum_{x_{T-1}} P_0(x_{T-1}|x_{T-2}) e^{-\frac{1}{\alpha}(q(x_{T-1})+V_T)}$$

with the limit

$$V^*_{T-1} = \lim_{\alpha \to 0} V_{T-1} = \min_{x_{T-1}} \left(q(x_{T-1}) + V^*_T\right).$$

Thus, we retrieve from (12) the general Bellman recursion

$$V^*_t = \min_{x_t} \left(q(x_t) + V^*_{t+1}\right)$$

in the perfectly rational limit $\alpha \to 0$. Accordingly, the perfectly rational action probabilities are given by

$$P(x_t|x_{t-1}) = \delta(x_t - x^*_t)$$

with

$$x^*_t = \arg\min_{x_t} \left(q(x_t) + V^*_{t+1}\right)$$

and the boundary condition $x^*_T = \arg\min_{x_T} q(x_T)$.

V. Iterative Control

Bounded rational controllers have also an interesting interpretation when used for iterative control. In particular, the path integral approach has recently been extended for iterative control [15]. Consider a static control problem with initial control $\mathbf{P}_0(x)$ and control cost $q(x)$. The bounded rational controller at temperature α is then given as

$$\mathbf{P}(x) = \frac{\mathbf{P}_0(x) e^{-\frac{1}{\alpha}q(x)}}{\int dx' \mathbf{P}_0(x') e^{-\frac{1}{\alpha}q(x')}}.$$

This expression looks very much like an inference step in Bayesian inference, where we identify the prior $\mathbf{P}_0(x)$, the likelihood model $e^{-\frac{1}{\alpha}q(x)}$ and the posterior $\mathbf{P}(x)$, normalized by the partition function. If we now use the posterior control of one iteration as the prior of the next iteration, we get something that is very similar to Bayesian inference, where each x corresponds to a hypothesis. Over the subsequent iterations the hypotheses x compete for probability mass, where the x that have a lower-than-average cost are favored. A similar Bayesian rule has been recently proposed for adaptive control in [10], where different control modes compete for expression.

VI. Conclusion

In the present paper we have proposed a physically inspired notion of bounded rationality as a unifying framework for a number of recently published approximate optimal control methods in the continuous [6], [7], [15] and the discrete domain [16]. The proposed bounded rational controllers maximize a free utility functional and implicitly trade-off desired utilities against required resource costs. The resource costs are interpreted as information costs in physically embedded systems, since changing physically implemented information states comes with a thermodynamic cost measured in energy units. This change in information state is always measured with respect to an initial baseline policy, which includes the special case of a baseline policy that attributes equal probability mass to all possible actions. Thus, bounded rational controllers are in general stochastic. In the limit of negligible information costs the classical expected utility principle can be recovered, which includes classical stochastic optimal control as a special case.

We could show that the path integral formalism [6] and a recently published approximate optimal control method in the discrete domain [16] can be conceptualized as bounded rational controllers. Importantly, these controllers can be derived from a free utility principle without invoking the Hamilton-Jacobi-Bellman equation or the Bellman optimality equations in the discrete case.

Previously, the path integral control formalism has been derived from the Hamilton-Jacobi-Bellman equation. One of the important steps in the path integral control framework is the transformation of the Hamilton-Jacobi-Bellman equation into a linear and second order partial differential equation (which is equivalent to the backward Kolmogorov partial differential equation), via the use of 1) a logarithmic transformation of the value function and 2) the assumption that the variance of the noise scales with the control cost. Even though the connection between the control cost and the noise makes sense from a control theoretic point of view, by allowing control authority in cases of high variance, this assumption is imposed so that the Hamilton-Jacobi-Bellman equation can be linearized.

In contrast, the relation between noise and control cost falls out naturally in our derivation. Thus, the path integral control formalism can be interpreted as bounded rational control, where the "control cost" is equated with the "information cost" of the bounded rational controller. The result is a stochastic controller that controls a deterministic system, which also explains why the noise in the path integral formalism has to be in the control space. Such stochastic controllers are not optimal in the traditional sense of optimal control, but require new optimality principles that allow for stochastic control. The suggested framework of bounded rationality provides such an optimality principle.

References

[1] Jonas Buchli, Evangelos Theodorou, Freek Stulp, and Stefan Schaal. Variable impedance control - a reinforcement learning approach. In *Robotics: Science and Systems Conference (RSS)*, 2010.

[2] H.B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. John Wiley & Sons, 2nd edition, 1985.

[3] R. P. Feynman. *The Feynman Lectures on Computation*. Addison-Wesley, 1996.

[4] P.C. Fishburn. *The Foundations of Expected Utility*. D. Reidel Publishing, Dordrecht, 1982.

[5] W. Fleming. Exit probabilities and optimal stochastic control. *Applied Mathematics and Optimization*, 4:329–346, 1978.

[6] B. Kappen. A linear theory for control of non-linear stochastic systems. *Physical Review Letters*, 95:200201, 2005.

[7] B. Kappen, V. Gomez, and M. Opper. Optimal control as a graphical model inference problem. *arXiv:0901.0633v2*, 2010.

[8] D.M. Kreps. *Notes on the Theory of Choice*. Westview Press, 1988.

[9] P. Ortega. *A unified framework for resource-bounded autonomous agents interacting with unknown environments*. PhD thesis, 2011.

[10] P. A. Ortega and D. A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

[11] P.A. Ortega and D.A. Braun. A conversion between utility and information. In *Proceedings of the third conference on artificial general intelligence*, pages 115–120. Atlantis Press, 2010.
[12] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 3rd edition, 2010.
[13] S.J. Russell. Rationality and intelligence. In Chris Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 950–957, San Francisco, 1995. Morgan Kaufmann.
[14] H. Simon. *Models of Bounded Rationality*. MIT Press, 1982.
[15] E. Theodorou, J. Buchli, and S. Schaal. A generalized path integral approach to reinforcement learning. *Journal of Machine Learning Research*, 11:3137–3181, 2010.
[16] E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences U.S.A.*, 106:11478–11483, 2009.
[17] M. Tribus and E.C. McIrvine. Energy and information. *Scientific American*, 225:179–188, 1971.
[18] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.