## Reward Shaping via Diffusion Process in Reinforcement Learning

Peeyush Kumar

#### Abstract

Reinforcement Learning (RL) models have continually evolved to navigate the exploration - exploitation trade-off in uncertain Markov Decision Processes (MDPs). In this study, I leverage the principles of stochastic thermodynamics and system dynamics to explore reward shaping via diffusion processes. This provides an elegant framework as a way to think about exploration-exploitation trade-off. This article sheds light on relationships between information entropy, stochastic system dynamics, and their influences on entropy production. This exploration allows us to construct a dual-pronged framework that can be interpreted as either a maximum entropy program for deriving efficient policies or a modified cost optimization program accounting for informational costs and benefits. This work presents a novel perspective on the physical nature of information and its implications for online learning in MDPs, consequently providing a better understanding of information-oriented formulations in RL.

### 1 Introduction

In this article, I take inspiration from stochastic thermodynamics to derive a problem formulation for online learning in uncertain MDPs while grounded in system dynamics. The system balances the diffusion process with drif dynamics as a way to formulate the exploration-exploitation trade-off.

To this effect, I make an explicit link between the information entropy and the stochastic dynamics of a system coupled to an environment. I analyze various sources of entropy production: due to the decision-maker's uncertainty about the system-environment interaction characteristics; due to the stochastic nature of system dynamics; and the interaction of the decision maker's knowledge with system dynamics. This analysis provides a framework that can be formulated either as a maximum entropy program to derive efficient policies that balance the exploration and exploitation trade-off, or as a modified cost optimization program that includes informational costs and benefits.

### 1.1 Background

Markov decision processes (MDPs) are perhaps the most widely studied models of sequential decision problems under uncertainty [Puterman, 1994]. In this article, an MDP is described

by the tuple  $\mathcal{M} = (S, A, T, R, N)$ . Here, S is a finite set of states; A is a finite set of actions; T denotes a transition probability function of the form T(s'|s,a), for  $s,s' \in S$  and  $a \in A$ ; R denotes a reward function of the form R(s'|s,a), for  $s,s' \in S$  and  $a \in A$ ; and N denotes a finite planning horizon. This MDP models the following time-invariant, finite-horizon, sequential decision-making problem under uncertainty. A decision-maker observes the state  $s_t \in S$  of a system at the beginning of time-slot  $t \in \{1, 2, ..., N\}$  and then chooses an action  $a_t \in A$ . The system then stochastically evolves to a state  $s_{t+1} \in S$  by the beginning of slot t+1 with probability  $T(s_{t+1}|s_t,a_t)$ . As a result of this transition, the decision-maker collects a reward  $R(s_{t+1}|s_t,a_t)$ . This process of state observation, action selection, state evolution, and reward collection repeats until the end of slot N. A policy trajectory  $\pi = (\pi_1, \pi_2, ..., \pi_N)$  is a decision-rule that assigns actions  $\pi_t(s_t) \in A$  to states  $s_t \in S$ , for t = 1, 2, ..., N. Note that the set  $\mathcal{P}$  of such policy trajectories is finite. The decision-maker's objective is to find a policy trajectory  $\pi = (\pi_1, \pi_2, ..., \pi_N) \in \mathcal{P}$  that maximizes the expected reward

$$J_{\pi}(s_1) = E \left[ \sum_{t=1}^{N} R(s_{t+1}|s_t, \pi_t(s_t)) \right].$$

It is assumed for simplicity of notation that no terminal reward is earned at the end of slot N.

The transition probability function is often unknown to the decision-maker at the outset. This calls for online learning of transition probabilities while the system evolves. For instance, in medical treatment planning, a doctor might not know the uncertain dose-response function of an individual at the beginning of a treatment course, but may want to adaptively make drug selection and dosing decisions over the treatment course [Kotas and Ghate, 2016]. Similarly, a seller conducting a sequence of auctions may not know the bidder demand and willingness-to-pay distributions, but must adaptively make auction-design decision such as the minimum bid in each auction [Ghate, 2015]. Such problems fall under the broad framework of MDPs under imperfect information, and can be seen as Bayesian adaptive MDPs (BAMDPs) or partially observable MDPs (POMDPs) in some cases [Bertsekas, 2005, Dreyfus and Law, 1977, Krishnamurthy, 2016, Kumar, 1985, Kumar and Varaiya, 2016].

The challenge in any Bayesian learning approach is that there is no clear consensus on the actual problem that needs to be solved. Generally, we want to find a policy that maximizes cumulative reward while learning with uncertain or partial information. BAMDPs provide a classic formulation of this problem. But this formulation does not take into account the cost of information gain, and hence intuitively one can find better policies that leverage information gain while learning. The information theoretic methods developed so far rely on the heuristic idea of information ratio, which, I believe, is somewhat ad-hoc. In addition, this ratio does not give a strong insight into the global problem that is being solved. I find a relation between the optimization of reward and the cost of information that is embedded in the dynamics of the system and its interaction with the environment.

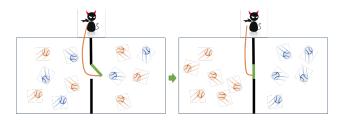


Figure 1: Maxwell's demon

### 2 Physical nature of information

In order to motivate the idea of the physical nature of information, I dive into the role of information in thermodynamics of gases. To guide the reader, a natural connection is to interpret particle configurations in gas systems as sample trajectories in stochastic system. Physicist Ludwig Boltzmann showed that with time, a system evolves towards lower states of energy, where the energy dispersed increases the entropy of the system due to the nature of statistics [Boltzmann, 1974]. As Wolchover [2017] commented, "There are many ways for energy to be spread among the particles in a system than concentrated in a few, so as particles move around and interact, they naturally tend toward states in which their energy is increasingly shared. This has been classically understood as the second law of thermodynamics. But Maxwell's letter [Maxwell, 1921] described a thought experiment in which an enlightened being, called Maxwell's demon (Figure 1), uses its knowledge to lower entropy and violate the second law. The demon knows the positions and velocities of every molecule in a container of gas. By partitioning the container and opening and closing a small door between the two chambers, the demon lets only fast-moving molecules enter one side, while allowing only slow molecules to go the other way. The demon's actions divide the gas into hot and cold, concentrating its energy and lowering its overall entropy. The once useless gas can now be put to work. This thought experiment lead to questions on how a law of nature could depend on one's knowledge of the positions and velocities of molecules. [This implies that second law of thermodynamics require a reinterpretation to include the subjective nature of information. Charles Bennett Bennett, 1987, building on work by Leo Szilard [Szilárd, 1976] and Rolf Landauer [Landauer, 1961], resolved the paradox by formally linking thermodynamics to the science of information. Bennett argued that the demon's knowledge is stored in its memory, and memory has to be erased, which takes work. [Landauer, 1961] calculated that at room temperature, it takes at least 2.9 zeptojoules of energy for a computer to erase one bit of stored information.) In other words, as the demon organizes the gas into hot and cold and lowers the gas's entropy, its brain burns energy and generates more than enough entropy to compensate. The overall entropy of the gas-demon system increases, satisfying the second law of thermodynamics. These findings revealed that, as Landauer put it, "Information is physical" [Landauer, 1991]. More information implies that more work can be extracted. Maxwell's demon can wring work out of a single-temperature gas because it has far more information than the average user."

This interaction of entropy and dynamics, capture by the second law, creates a strong foundation to analyze stochastic systems. There is a natural equivalence between stochastic thermodynamics and stochastic control theory. Any decision process can be modeled as a

classic control problem. Generally, the quantities which are of interest are averaged over trajectories of the system rather than sample path behaviors. Thermodynamics has provided an intuitive framework and solution about averaged entities on stochastic systems. I study this equivalence and bridge gaps in the existing literature on learning in MDPs. I develop an equivalent thermodynamic system and apply an information theoretic framework to find a formulation of the learning problem to compute good policies.

There has been some work in the literature to bridge this gap between control theory and stochastic thermodynamics. Brockett and Willems [1979] studies second law of thermodynamics from the point of view stochastic control theory. They compute a criterion which, when satisfied, permits one to assign a temperature to a stochastic system in a way that Carnot cycles become the optimal trajectories of optimal control problems. Propp [1985] also studied the connection between thermodynamic and Markovian systems. There, an input-output framework for thermodynamics was proposed, which allowed to introduce the notion of states, controls and response, thus drawing a connection between the two fields. There has also been a recent surge in understanding the field of stochastic thermodynamics to study Markovian processes at the trajectory level using statistical quantities [Seifert et al., 2011, Aurell et al., 2012. Saridis [1988] proposed a formulation that gives a generalized energy interpretation to the optimal control problem. This framework provides compatibility between the control problem and the information theoretic methodology for the intelligent control system using entropy as the common measure. A reformulation of the optimal control problem is based on the idea of expressing the design of the desirable control by the uncertainty of selecting a control law that minimized a given performance index.

### 3 Clairvoyant MDP: an information theoretic perspective

Consider the Bellman's equation for MDP  $M = \{S, A, T, R, N\}$ .

$$V^*(s) = \min_{a} \sum_{s'} T(s'|s, a) [R(s'|s, a) + V^*(s')]. \tag{1}$$

I consider an alternate formulation to this classical MDP, with a small loss of generality. Todorov [2009] proposed a linear problem where actions that are considered symbolic in the above formulation are replaced through making decisions over transition distributions. Therefore, the decision maker specifies a control dynamics distribution a(s'|s) = T(s'|s,a). This allows us to write an equivalent reward form as

$$q(s,a) = \ell(s) + \mathop{E}_{s' \sim a(\cdot|s)} \ln \left( \frac{a(s'|s)}{p(s'|s)} \right),$$

where the state cost  $\ell(s)$  is an arbitrary function encoding how undesirable different states are and p(s'|s) is an arbitrary transition distribution. Using this construction the Bellman's equation can be rewritten as:

$$V^{*}(s) = \min_{a} \left( \ell(s) + \mathop{E}_{s' \sim a(\cdot|s)} \left[ \ln \frac{a(s'|s)}{p(s'|s)} + V^{*}(s') \right] \right). \tag{2}$$

Now, I define the quantity  $G(s) = E_{s' \sim p(\cdot|s)} exp(-V^*(s'))$ . Therefore, through some algebraic manipulation, I get

$$\underset{s' \sim a(\cdot|s)}{E} \left[ \ln \frac{a(s'|s)}{p(s'|s)} + V^*(s') \right] = -\ln(G(s)) + \mathbb{KL} \left( a(\cdot|s) \left\| \frac{p(\cdot|s) \exp(-V^*(\cdot))}{G(s)} \right),$$

which gives

$$V^*(s) = \min_{a} \left[ \ell(s) - \ln(G(s)) + \mathbb{KL}\left(a(\cdot|s) || \frac{p(\cdot|s) \exp(V^*(\cdot))}{G(s)}\right) \right]. \tag{3}$$

An interesting observation is that the right hand side of the above function is minimized when the KL divergence is 0, which gives the optimality condition as

$$a^*(s'|s) = \frac{p(s'|s)\exp(-V^*(s'))}{G(s)}$$
(4)

$$= \frac{p(s'|s)\exp(-V^*(s'))}{\sum_{s'} p(s'|s)exp(-V^*(s'))}$$
(5)

Now consider the following Lemma [Theodorou and Todorov, 2012, Theodorou, 2015].

**Lemma 1.** Consider distributions  $\mathbb{A}$  and  $\mathbb{P}$  defined on the same probability space with sample set  $\Omega$ , such that  $\mathbb{A}$  is absolutely continuous with respect to  $\mathbb{P}$ , and  $Q: \Omega \mapsto \mathbb{R}$  is a measurable function, then the following inequality holds

$$\frac{1}{\rho} \ln \left( E_{\mathbb{P}} \left[ e^{\rho Q(s)} \right] \right) \le E_{\mathbb{A}} [Q(s) + |\rho|^{-1} \mathbb{KL}(\mathbb{A}||\mathbb{P})],$$

where  $\rho \in \mathbb{R}^-$ .

*Proof.* The proof is reproduced here for completeness. It is a straightforward derivation from Jensen's inequality. Consider,

$$\ln\left(E\left[e^{\rho Q(s)}\right]\right) = \ln\sum_{s} p(s)e^{\left[\rho Q(s)\right]}$$

$$= \ln\left[\sum_{s} a(s)\frac{p(s)}{a(s)}\exp\left(\rho Q(s)\right)\right]$$

$$\stackrel{a}{\geq} \sum_{s} a(s)\ln\left[\frac{p(s)}{a(s)}\exp\left(\rho Q(s)\right)\right]$$

$$= \rho E\left[Q(s)\right] + \sum_{s} a(s)\ln\frac{p(s)}{a(s)}$$

$$= \rho\left(E\left[Q(s)\right] - \rho^{-1}\mathbb{KL}(\mathbb{A}||\mathbb{P})\right),$$

where inequality "a" follows from Jensen's inequality and concavity of the ln function. Now dividing both sides by  $\rho \in \mathbb{R}^-$  gives the required inequality.

Next, consider Equation (2), where I substitute  $Q(s') = \ell(s) + V^*(s')$ . Now using Lemma 1 with  $\rho = -1$ ,  $\mathbb{P} = p(s'|s)$ ,  $\mathbb{A} = a(s'|s)$ , I get

$$-\ln\left(\sum_{s'\sim p(\cdot|s)}\left[e^{-\ell(s)-V^*(s')}\right]\right) \leq \sum_{s'}a(s'|s)\left[\ell(s)+V^*(s')+\ln\frac{a(s)}{p(s)}\right],$$

which implies

$$-\ln\left(\mathop{E}_{s'\sim p(\cdot|s)}\left[e^{-\ell(s)-V^*(s')}\right]\right) = \min_{a} \sum_{s'} a(s'|s) \left[\ell(s) + V^*(s') + \ln\frac{a(s)}{p(s)}\right].$$

The right hand side of the above equation is the right hand side of the Bellman equation in Equation (2). Therefore

$$V^*(s) = -\ln \left( \underset{s' \sim p(\cdot|s)}{E} \left[ e^{-\ell(s) - V^*(s')} \right] \right).$$

This framework can also be used as an estimation framework where instead of minimizing the expected cumulative cost, the decision maker can maximize the KL divergence for a required performance. Therefore, the optimization problem in Equation (2) becomes

$$\min_{\mathbb{A}} \mathbb{KL}(\mathbb{A}||\mathbb{P}),$$

subject to

$$\sum_{s'} a(s'|s) = 1,$$

$$V(s) = K,$$

where K is the required performance. In the interest of providing interesting connection, I consider continuous optimization. Using the Lagrangian method, the optimization program reduces to

$$\mathcal{L} = \mathbb{KL}(\mathbb{A}||\mathbb{P}) + \mu(V(s) - K) + \lambda \left( \int_{s'} a(s'|s) ds' - 1 \right)$$
$$= -\int_{s'} a(s'|s) \left( \ln \frac{a(s'|s)}{p(s'|s)} + \mu V(s) + \lambda \right) ds' + \mu K + \lambda$$

Now, maximizing with respect to a(s'|s) gives

$$ln\frac{a^*(s'|s)}{p(s'|s)} + \mu V(s) + \lambda = 0,$$

which gives

$$a^*(s'|s) = \int exp(-muV(s) - \lambda)p(s'|s)ds'.$$

Substituting in the first constraint for  $\int a(s'|s)ds = 1$ , gives

$$\lambda = \ln \int p(s'|s) \exp(-\mu V(s)) ds'.$$

Substituting  $\lambda$  back and discretizing gives the optimal solution for  $a^*$  for a given level of performance. In the case where  $K = V^*(s)$ , this solution gives the optimal  $a^*$  as in Equation 5. This result is very similar to the one derived using HJB principle in the classic paper by Saridis [1988]. For the reader's convenience, I recall that result in Appendix 7.

### 4 Thermodynamics of information

This section provides a brief introduction on the relationship between information and thermodynamics. We consider a system M (such as a gas in a container) that is connected to
external reservoirs and other systems. Suppose the microstate of the system (for example,
the coordinates and momentum of particles of the gas) is given by x, and suppose that the
information gained as a result of measurement is denoted by m. This measurement is what
helps to prepare the state of the system. Let us denote a generic statistical state of the
system with  $\rho(x)$  (for example, the distribution of coordinate states and momentum of the
gas molecules). I assume that in state  $\rho(x)$  the system is in statistical equilibrium. Now
after making the measurement, the new state of the system in  $\rho(x|m)$ , which in general is
out of equilibrium. For example, in the context of the Sczilard's engine described in Section 2, after measurement the statistical state is confined to either the left or right half of
the box. Information drives the system away from equilibrium. The thermodynamics of
information allows us to reason about this scenario by associating an equivalent energy cost,
thus justifying this movement from equilibrium to a non-equilibrium state.

The most obvious entity that relates statistical entities to distributions is the entropy of the system. In this case, the non equilibrium entropy is defined using a scaled version of the Shannon Entropy as

$$S(\rho) = -\sum_{x} \rho(x) \ln \rho(x) = H(X),$$

where H(X) is the Shannon entropy. At equilibrium this entropy coincides with the cannonical entropy

$$\rho(x) = \exp^{-\beta E(x)} / Z,$$

where E(x) is the Hamiltonian of the system, and Z is the partition function, and  $\beta$  is the inverse temperature. Using this we recover the thermodynamic relationship between Free energy  $\mathcal{F}(\rho) = -\beta^{-1} \ln Z$ , and internal Energy E = E[H] and Entropy:  $\mathcal{F} = E - \beta^{-1}S$ . The free energy is interpreted as the amount of useful energy that can be used to extract work, taking in account all entropy related costs. The classical second law of thermodynamics for non equilibrium system, therefore, can be written as

$$\Delta S \ge 0 \implies W - \Delta \mathcal{F} \ge 0,$$
 (6)

where W is the average work done on the system.

The rest of the section evaluates the change in non-equilibrium free energy due to a measurement M. For this purpose the corresponding information gain is defined as

$$I(X; M) = H(X) - H(X|M).$$

Now, in the event that an external system changes the system parameter after an observation is made, results in work extracted from the system. The refined second law of thermodynamics then becomes

$$W - \Delta \mathcal{F} \ge -\beta^{-1} I(X; M) \tag{7}$$

An interesting observation is that ultimately, the information used to extract work during feedback was supplied as work by the external system during the measurement process.

# 5 Markovian systems and second law of thermodynamics

Now let's consider the second law of thermodynamics  $W \ge \Delta \mathcal{F}$  without feedback (Equation 6), and compare it with the Lemma 1

$$\frac{1}{\rho} \ln \left( E\left[ e^{\rho Q(s)} \right] \right) \leq E\left[ Q(s) + |\rho|^{-1} \mathbb{KL}(\mathbb{A}||\mathbb{P}) \right].$$

The quantity on the left hand side is the Free Energy change  $\Delta \mathcal{F}$  and the work done on the system is the expected cumulative cost given by the right hand side of the equation. Substituting the relevant entities for the MDP defined in Section 3 provides a bridge between MDP and the respective thermodynamic interpretation. Therefore, using the mathematical equivalence, the policy that minimizes work done (or maximum work extracted from the system) gives the optimal solution for the MDP.

The above results give sufficient evidence to explore equivalence between thermodynamic entities and Markov Decision Processes. In order to develop a learning framework in uncertain MDPs using information theoretic arguments, I develop the definition of thermodynamic quantities at the level of sample trajectories for Markovian system in the next section.

# 5.1 Second law of thermodynamics for a Markovian system in a heat bath

This section reviews the stochastic thermodynamics for Markovian Systems [Ito and Sagawa, 2016]. Stochastic Thermodynamics is a theoretical framework to define quantities such as work and heat at the level of sample trajectories.

Consider a system M that evolves stochastically. We assume a physical situation where system M is connected to a single heat bath at inverse temperature  $\beta$ . Also assume that the system M is driven by an external parameter  $\pi$  and the system is not subject to non-conservative forces. For simplicity, we will assume discrete time  $t_k$ ,  $k = \{1, 2, \dots, N\}$ , although, the mathematical setup does not force any assumption regarding the continuity of time. Let  $x_k$  be the state of the system at time  $t_k$ , and  $\pi_k$  be the external parameter of the system at time  $t_k$ . Let  $p(x_k|x_{k-1},\pi_k)$  be the conditional probability of state  $x_k$  under the past trajectory and external parameter  $\pi_k$ .

Now building on the thermodynamic principles, we define the Hamiltonian of the system as  $E(x_k, \pi_k)$ . The Hamiltonian change in the system is decomposed into 2 parts heat  $Q_k$  and work  $W_k$ . The heat absorbed by the system from heat bath at time  $t_k$  is defined as

$$Q_k = E(x_{k+1}, \pi_k) - E(x_k, \pi_k),$$

and the work done on the system M is defined as

$$W_k = E(x_k, \pi_k) - E(x_k, \pi_{k-1}).$$

For a given trajectory  $\{x_1, x_2, \dots, x_n\}$  the total heat is  $Q = \sum_{i=1}^{N-1} Q_k$  and total work is  $W = \sum_{i=1}^{N-1} W_k$ , where  $x_0$  is defined as a buffer state such that  $p(x_1|x_0, \pi) = 1$  for any  $\pi$ . This is done to impose consistency as it will become apparent later on.

Using the above definitions, one can easily show that  $\Delta E_k = Q_k + W_k$ , which is the first law of thermodynamics.

Now let us define the quantity  $p_B(x_k|x_{k+1},\pi_k)$  as the backward transition probability. In the absence of any non conservative the detailed balance [Seifert, 2005] is satisfied which gives

$$\frac{p(x_{k+1}|x_k,\pi_k)}{p_B(x_k|x_{k+1},\pi_k)} = e^{-\beta Q_k}.$$

Now, I define the stochastic entropy of the system as  $h(x_k) = -ln(x_k)$ . Therefore the *entropy* production is defined as the sum of stochastic entropy change in the system and the bath. The stochastic entropy change in the system is given by

$$\Delta h_k^M = h(x_{k+1}) - h(x_k).$$

The total stochastic entropy change therefore is given by

$$\Delta h^M = \ln \frac{p(x_1)}{p(x_N)}. (8)$$

The stochastic entropy change in the heat bath is given by the heat dissipation into the bath

$$\Delta h_k^{bath} = -\beta Q_k.$$

The total entropy change in the bath is given by

$$\Delta h^{bath} = ln \frac{p(x_N|x_{N-1}, \pi_{N-1})p(x_{N-1}|x_{N-2}, \pi_{N-2}) \dots p(x_2|x_1, \pi_1)}{p_B(x_1|x_2, \pi_1)p_B(x_2|x_3, \pi_2) \dots p(x_{N-1}|x_N, \pi_{N-1})}.$$
(9)

Therefore the entropy production  $\sigma$  is

$$\sigma = \ln \frac{p(x_N|x_{N-1}, \pi_{N-1})p(x_{N-1}|x_{N-2}, \pi_{N-2}) \dots p(x_2|x_1, \pi_1))p(x_1)}{p_B(x_1|x_2, \pi_1)p_B(x_2|x_3, \pi_2) \dots p(x_{N-1}|x_N, \pi_{N-1})p(x_N)}.$$

For brevity I define the trajectory of the system as  $O = \{x_1, x_2, \dots, x_N\}$ . Therefore the total entropy production becomes

$$\sigma = ln \frac{p(O)}{p_B(O)}.$$

Therefore, the entropy production is determined by the ratio of the probabilities of a trajectory and its time-reversal.

Simple algebraic calculation on this definition yields the second law of thermodynamics which states that

$$E[\sigma] \ge 0.$$

The equivalent stochastic energetics definition gives the form as in Equation (6)

$$W \ge \Delta \mathcal{F}$$
,

where  $\mathcal{F}(\lambda_k) = -\beta^{-1} \ln \sum_X \exp(-\beta E(x, \lambda_k))$ . This result can be derived using the integral fluctuation theorem and the arguments presented in Seifert [2005].

# 5.2 Second law of thermodynamics for a Markovian system in connection with an external entity

Here I consider the Markovian System M in contact with an external system D in addition to the heat bath. This external system, for instance can be the decision maker in the context of the MDP (More on this in the later sections). In particular, I state the generalized second law of thermodynamics, which states that the entropy production is bounded by the initial and final mutual information between M and D, and the transfer entropy from M to D.

Let's consider the states of the system D at time  $t_k$  be  $d_k$ . Therefore, the joint time evolution of system  $M \cup D$  is defined as  $\{(x_1, d_0), (x_2, d_1), \dots, (x_N, d_{N-1})\}$ . For brevity, I define  $pa(x_{k+1})$  as the parent of state  $x_{k+1}$  which is the set of all states which has a non zero transition probability to  $x_{k+1}$ , therefore  $pa(x_{k+1}) = \{x_k, d_{k-1}\}$ , such that  $p(x_{k+1}|x_k, d_{k-1}) > 0$ .

At the initial state I assume that  $pa(x_1) \subseteq D$ . The initial correlation between system S and D is then characterized by the mutual information between  $x_1$  and  $pa(s_1)$ . The corresponding stochastic mutual information is given by

$$I_{ini} = I(x_1; pa(x_1)).$$

Now, let's define  $an(x_{k+1})$  as the ancestors of  $x_k$  in the order that they were observed. Therefore  $an(x_{k+1}) = \{(x_1, d_0), (x_2, d_1), \dots, (x_k, d_{k-1})\}$ . The final correlation between system S and D is then characterized by the mutual information between  $x_N$  and  $an(x_N) \cap D$ .

$$I_{fin} = I(x_N; \{d_0, d_1, \dots, d_N\}).$$

Let's define another quantity  $pa(d_k)$  as the parent of  $d_k$  that corresponds to  $pa(d_k) = \{x_{k-1}, d_{k-1}\}$  Finally, I define the transfer entropy from M to D as

$$I_{tr}^{k} = I(d_{k}; pa(d_{k}) \cap M | d_{1}, d_{2}, \dots, d_{k-1}).$$

The total transfer entropy for the entire dynamics is therefore given by

$$\sum_{k=1}^{N} I_{tr}^k = I_{tr}.$$

By combining all the above informational content in the combined system, I define the total informational exchange as

$$\Theta = I_{fin} - I_{tr} - I_{ini}.$$

Now, as in the simple case in Section 5.1, I define the entropy production in system M and the heat bath, while in the presence of system D.

Let  $\mathcal{B}_{k+1} \subseteq D$  define the set of states in D that effect  $x_{k+1}$ , therefore  $\mathcal{B}_{k+1} = \{d_{k-1}\}$ . Now  $p(x_{k+1}|x_k,\mathcal{B}_{k+1})$  describes the transition probability from  $x_k$  to  $x_{k+1}$  under the condition that the states of D that affect M are given by  $\mathcal{B}_{k+1}$ . We then define the backward transition probability as  $P_B(x_k|x_{k+1},\mathcal{B}_{k+1})$ . Following the definition of entropy change in the heat bath from time k to k+1 as in Equation (9) is given by:

$$\Delta s^{bath} = \sum_{k} x_{k}^{bath}$$

$$= \sum_{k} \ln \frac{p(x_{k+1}|x_{k}, B_{k+1})}{p_{B}(x_{k}|x_{k+1}, B_{k+1})}.$$

The total entropy change in the system M is similar to Equation (8)

$$\Delta s^{sys} = \ln \frac{p(x_1)}{p(x_N)}.$$

The total entropy production is therefore,

$$\sigma = \ln \frac{p(x_1)}{p(x_N)} \prod_k \frac{p(x_{k+1}|x_k, B_{k+1})}{p_B(x_k|x_{k+1}, B_{k+1})}.$$

Now, we can write the refined second law of thermodynamics, through some algebraic manipulation it can be shown that

$$E[\sigma] \geq \Theta.$$

Using the integral fluctuation theorem and theory of stochastic energetics, this result can be restated as in Equation (7)

$$W - \Delta \mathcal{F} \ge -\beta^{-1}\Theta \tag{10}$$

# 6 MDP with uncertainty: a stochastic thermodynamics perspective

The framework in the previous section provides a way to model the effect of information gain in MDPs with uncertainty with the objective of maximizing the work that can be extracted out of the system. The system M considered in the previous section is the system that is acting in the real environment, the system D is the decision maker, who changes some parameter of the system M in order to achieve the required objective. Both these systems are suspended in a "heat bath" to account for the part of the work that is dissipated and cannot be used for any useful work. The thermodynamic framework allows us to define the objective of the optimization program when the MDPs have model uncertainty. To be consistent, the uncertainty in the MDPs is assumed to be completely reflected through the uncertainty in the transition probabilities. In this section, I propose 2 different perspectives of how the system D interacts with system M: a) the first perspective is where system D directly maintains a distribution over the policies and changes this distribution based on feedback; b) the second perspective is where the system D maintains a distribution over a parameter of the transition distribution and adapts this based on feedback in order to find a good policy.

### 6.1 MDP with distribution over policies

Consider an MDP  $M = \{S, A, T, R, N\}^1$  and the decision maker  $D = \{\pi\}$ . The decision maker maintains a probability distribution  $\nu_k(\pi|s_k)$  over policies  $\pi$  at every time step  $t_k$  in state  $s_k$ . The probability distribution is updated based on feedback. This setup is analogous to the thermodynamic setup described in Section 5.2. In a standard MDP, the objective is to minimize the expected cumulative cost

$$V^{\pi}(s_t) = \sum_{k=t}^{N-1} E[c(s_k, \pi(s_k))],$$

where the expectation is taken over  $\{s_k, \pi(s_k)\}$ , in terms of the classical discrete MDP  $c(s_k, \pi(s_k)) = E_{s_{k+1} \sim T(\cdot|s_k, \pi(s_k))}[R(s_{k+1}|s_k, \pi(s_k))].$ 

As in Section 3, the MDP problem for finding a policy to achieve the maximum *performance* can be formulated as either a maximum entropy optimization program or the classical expected cost optimization. In will start by formulating an expected cost optimization program using the Second Law of Thermodynamics. Equation 10 can be written as

$$W + \beta^{-1}\Theta \ge \mathcal{F}.$$

Note that the free energy  $\mathcal{F}$  is the amount of useful energy, and the infimum of the left hand side will give the most amount of net work that can be extracted out of the system. Therefore, the optimization program becomes

<sup>&</sup>lt;sup>1</sup>Please note that for the purpose of this discussion I will consider R as the cost function (rather than the reward function)

$$\min_{\nu_t; t=1:N} W + \beta^{-1}\Theta,$$

where  $W = \sum_{k=1}^{N-1} E[c(s_k, \pi(s_k))]$ ,  $\Theta = I_{fin} - I_{tr} - I_{ini}$ , and  $\nu_t = p(\pi_t | s_t, \pi_{t-1})$ . From previous section  $x_k = \{s_k\}$ , and  $d_k = \pi_k$ . For the classical MDP  $p(s_1) = \delta(s_1 - s_{init})$ , and  $pa(s_1) = \emptyset$ . Therefore,

$$I_{ini} = I(s_1; pa(s_1)) = 0.$$

The final information correlation is given by

$$I_{fin} = I(s_N; \pi_1, \dots, \pi_{N-1}),$$

note that  $p(\pi_1, \dots, \pi_{N-1}) = \prod_{i=2}^{N-1} p(\pi_i | \pi_{i-1}).$ 

$$I_{tr}^{k} = I(\pi_{k}; s_{k-1} | \pi_{1}, \dots, \pi_{k-1}) = I(\pi_{k}; s_{k-1} | \pi_{k-1})$$

Therefore the optimization program becomes

$$\min_{\nu_t:t=1,\cdots,N} \left( \sum_{k=1}^{N-1} \left( E[c(s_k,\pi(s_k))] - \beta^{-1}I(\pi_{k+1};s_k|\pi_k) \right) + \beta^{-1}I(s_N;\pi_1,\cdots,\pi_{N-1}) \right).$$

For the case,  $I_{fin} = 0$ , the solution to the resulting optimization program is discussed in Tanaka et al. [2017].

The above problem can reformulated as a maximum entropy principle, which translates to

$$\max_{\nu_t:t=1,\cdots,N} \sum_{k=1}^{N-1} I(\pi_{k+1}; s_k | \pi_k) - I(s_N; \pi_1, \cdots, \pi_{N-1})$$

subject to

$$\sum_{k} \nu_t(\pi_k) = 1 \ \forall k,$$

$$\sum_{k=1}^{N-1} E[c(s_k, \pi_k)] = K,$$

where K is the required performance. When  $K = V^*$ , the resulting policy is the optimal policy with respect to the cost based optimization program.

### 6.2 MDP with parametric uncertainty

In this case the decision maker D maintains a distribution over the parameter of the system. The state of the system D is denoted by  $\lambda_k$  at time  $t_k$ . Again, the specific informational correlations are given by

$$I_{ini} = I(s_1; pa(s_1)) = 0,$$

$$I_{fin} = I(s_N; \lambda_1, \dots, \lambda_{N-1}),$$

and

$$I_{tr}^{k} = I(\lambda_{k}; s_{k-1}|\lambda_{1}, \dots, \lambda_{k-1}) = I_{tr}^{k} = I(\lambda_{k}; s_{k-1}|\lambda_{k-1}).$$

The optimization program becomes

$$\min_{\nu_t: t=1,\cdots,N} \left( \sum_{k=1}^{N-1} \left( E[c(s_k, \pi(s_k))] - \beta^{-1} I(\lambda_{k+1}; s_k | \lambda_k) \right) + \beta^{-1} I(s_N; \lambda_1, \cdots, \lambda_{N-1}) \right),$$

where  $\nu_t = p(\pi_t|s_t)$ , and the distribution over  $\lambda$  is updated using Bayesian learning. As in the previous section this can also be formulated as a maximum entropy framework.

### 7 Discussion and future work

This article provides a framework for formulating an optimization program for solving uncertain MDPs built from fundamental principles of system dynamics and information theory. The exact formulation of the optimization program depends on the specific nature of interaction between the decision maker and the system to be controlled. Sections 6.1 and 6.2 provide optimization program for 2 different scenarios. Given these formulation, we can use many of the techniques for optimization (including the Bellman's principle) to solve for a solution. This will be a future work. An important discussion point is the entity  $\beta$  in the above equations. Thermodynamically,  $\beta$  capture the inverse temperature (with a scaling constant). The temperature is a property of the heat bath and assumed to be constant throughout the dynamic process. In the context of a decision process, the temperature is a property of the decision process and can be estimated. A good way to estimate temperature will be to find an equilibrium solution and solve it inversely to get the temperature. For instance, given a MDP  $M = \{S, A, T, R, N\}$  one can chose the starting state for which there a solution is known apriori and that can be used to estimate the temperature of the decision process. In the event we do not have access to this knowledge, the temperature can be considered a pseudo state and a new MDP can be defined  $M' = \{\{S, \beta\}, A, T', R', N\}$ .

Another important point is that the above framework works when certain conditions on the underlying Markovian process is satisfied. One sufficient condition, as discussed in Section 5, is the detailed balance equation, which implies reversibility of the Markovian system. We know that is not a necessary condition, in fact, it can be shown that the results still hold for non-reversible Langevin dynamics. Additional research is required to state and prove the necessary and sufficient conditions for this framework to hold.

In conclusion, this work opens up avenues for further research in employing information theoretic arguments to learning in MDPs with model uncertainty. This work explicitly models information content and system dynamics for MDPs. I provide a framework to formulate the optimality criterion for MDPs with model uncertainty. Hopefully, future work can extend the rich theory of MDPs to learn and make good decisions in the situations of information uncertainty.

## **Appendix**

### A Maximum Entropy Principle: A Control Theoretic Approach

A classic paper by Saridis [1988] derives a maximum entropy framework for Control systems. I present this here as it is interesting to see connections without using the definitions from Thermodynamics.

Consider a generic decision system formulated in a classic control theoretic framework. Assume that the dynamics of the system are deterministic for simplicity. Then the dynamics are given by

$$\dot{x}(t) = f(x, u, t), \quad x(t_0) = x_0$$

and the associated cost function is

$$V^*(x_0, u, t_0) = \int_{t_0}^T L(x, u, t) dt; \quad L(x, u, t) > 0.$$

Here,  $x(t) \in X$  is a *n*-dimensional state vector and  $u(t) : X \times T$  is the *m* dimensional feedback control law. The solution is to find a control law  $u_k(x,t)$  such that the value function *V* will take a value *K* such that  $V_{min} \le K < \infty$ .

$$V^*(x_0, t_0|u_K(x(t), t) = K$$

This satisfies the Hamilton-Jacobi-Bellman equation

$$\frac{\partial V}{\partial t} + \frac{\partial V^T}{\partial x} f(x, u_k, t) + L(x, u_k, t) = 0.$$

In order to formulate the problem in entropy terms we consider the decision-maker's uncertainty of selecting the proper control from the set of admissible controls to satisfy the value function requirement to equal K ( $V_{min}$  in the case of optimal control). This may be expressed as a condition that the expected value of V equals K:

$$E_{u \sim p(u)}[V^*(x_0, t_0; u(x, t))] = K.$$

The expected value of V is taken over the set of admissible controls U, over which a probability density p(u) is assumed to express the uncertainty of selecting the proper control. The corresponding entropy can then be expressed as

$$H(u,p) = -\int_{U} p(u) \ln p(u) du.$$

According to Jaynes principle [Jaynes, 1957], the least biased estimate possible on the given information is given by the probability distribution p(u) that maximizes the above entropy H(u, p). Following the method of Lagrange, define

$$I = H(u) - \mu(E[V] - K) - \lambda(\int p(u)du - 1).$$

Using calculus of variation to maximize I with respect to the distribution p(u) yields

$$\ln(p) + 1 + \mu V + \lambda = 0.$$

Therefore,

$$p(u) = \exp^{-1-\lambda - \mu V^*(x_0, u(t), t)},$$

and the entropy with maximum information is given by

$$H(u) = 1 + \lambda + \mu E[V^*(x_0, u(x), t)].$$

For optimality, a control policy u is computed that minimizes the above entropy. This is, therefore, a max-min problem.

Saridis [1988] generalizes this analysis in the presence of dynamical uncertainty. Consider that  $y \in Y$  is the observation on the state x. It is essentially shown that the entropy H(u) can be decomposed in three parts as

$$H(u) = H(u|y) + H(y) - H(y|u),$$

where the associated probabilities are given by

$$p(u|y) = e^{-1-\lambda-\mu W(u(y),t)}$$
(11)

$$p(y) = e^{-\rho - \nu \int_0^T ||y - x||^2 dt}$$
(12)

$$p(y|u) = p(u|y)p(y)/p(u).$$
(13)

Here,  $W(u(y),t) = E_{x_0,w(t)}\{V^*(x_0,u(x,t),w,\nu,t_0)\}$ , and  $\rho,\nu$  are appropriate constants for the entropy estimation of H(y) based on Jayne's principle.

In case of parametric uncertainty, when

$$\dot{x} = f(x, u, \lambda, w, t)$$

when y are the observations

$$H(y) = H(u|y,\lambda) + H(y|\lambda) + H(\lambda) - H(y,\lambda|u).$$

An interesting observation is that entropy in a stochastic control system is decoupled into 4 different parts which can be individually computed.

### References

- E. Aurell, K. Gawędzki, C. Mejía-Monasterio, R. Mohayaee, and P. Muratore-Ginanneschi. Refined second law of thermodynamics for fast random processes. *Journal of statistical physics*, 147(3):487–505, 2012.
- C. H. Bennett. Demons, engines and the second law. Scientific American, 257(5):108–116, 1987.

- D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Nashua, NH, USA, 3rd edition, 2005.
- L. Boltzmann. The second law of thermodynamics. In *Theoretical physics and philosophical problems*, pages 13–32. Springer, 1974.
- R. Brockett and J. Willems. Stochastic control and the second law of thermodynamics. In Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on, volume 17, pages 1007–1011. IEEE, 1979.
- S. E. Dreyfus and A. M. Law. *The Art and Theory of Dynamic Programming*. Academic Press, New York, NY, USA, 1st edition, 1977.
- A. Ghate. Optimal minimum bids and inventory scrapping in sequential, single-unit, vickrey auctions with demand learning. *European Journal of Operational Research*, 245(2):555–570, 2015.
- S. Ito and T. Sagawa. Information flow and entropy production on bayesian networks. *Mathematical Foundations and Applications of Graph Entropy*, 3:2, 2016.
- E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
- J. Kotas and A. Ghate. Response-guided dosing for rheumatoid arthritis. *IIE Transactions on Healthcare Systems Engineering*, 6(1):1–21, 2016.
- V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, Cambridge, United Kingdom, 1st edition, 2016.
- P. R. Kumar. A survey of some results in stochastic adaptive control. SIAM Journal on Control and Optimization, 23(3):329–380, 1985.
- P. R. Kumar and P. P. Varaiya. Stochastic Systems: Estimation, Identification, and Adaptive Control. SIAM, Philadelphia, PA, USA, 2016.
- R. Landauer. Irreversibility and heat generation in the computing process. *IBM journal of research and development*, 5(3):183–191, 1961.
- R. Landauer. Information is physical. *Physics Today*, 44(5):23–29, 1991.
- J. C. Maxwell. Theory of heat. Longmans, 1921.
- M. B. Propp. The thermodynamic properties of Markov processes. PhD thesis, Massachusetts Institute of Technology, 1985.
- M. L. Puterman. Markov decision processes: Discrete stochastic dynamic programming. John Wiley and Sons, New York, NY, USA, 1994.
- G. N. Saridis. Entropy formulation of optimal and adaptive control. *IEEE Transactions on Automatic Control*, 33(8):713–721, 1988.

- U. Seifert. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Physical review letters*, 95(4):040602, 2005.
- U. Seifert, P. L. Garrido, J. Marro, and F. de los Santos. Stochastic thermodynamics: An introduction. In *AIP Conference Proceedings*, volume 1332, pages 56–76, 2011.
- L. Szilárd. On entropy reduction in a thermodynamic system by interference by intelligent subjects. *Zhurnal Physik*, 53, 1976.
- T. Tanaka, H. Sandberg, and M. Skoglund. Finite state markov decision processes with transfer entropy costs. arXiv preprint arXiv:1708.09096, 2017.
- E. A. Theodorou. Nonlinear stochastic control and information theoretic dualities: Connections, interdependencies and thermodynamic interpretations. *Entropy*, 17(5):3352–3375, 2015.
- E. A. Theodorou and E. Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *Decision and Control (CDC)*, 2012 IEEE 51st Annual Conference on, pages 1466–1473. IEEE, 2012.
- E. Todorov. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483, 2009.
- N. Wolchover. The quantum thermodynamics revolution, 2017. Accessed: 2017-05-05.