Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Adaptive decision making via entropy minimization

Armen E. Allahverdyan [a], Aram Galstyan [b], Ali E. Abbas [c], Zbigniew R. Struzik [d,e]

[a] *Yerevan Physics Institute, Alikhanian Brothers Street 2, Yerevan 375036, Armenia*
[b] *USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292, USA*
[c] *Industrial and Systems Engineering, University of Southern California, Los Angeles, USA*
[d] *RIKEN Brain Science Institute, 2-1 Hirosawa, Wako-shi, 351-0198, Japan*
[e] *Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan*

ARTICLE INFO

Article history:
Received 12 July 2018
Received in revised form 1 October 2018
Accepted 2 October 2018
Available online xxxx

Keywords:
Prior probability
Risk
Entropy maximization/minimization
Exploration–exploitation

ABSTRACT

An agent choosing between various actions tends to take the one with the lowest cost. But this choice is arguably too rigid (not adaptive) to be useful in complex situations, e.g., where exploration–exploitation trade-off is relevant in creative task solving or when stated preferences differ from revealed ones. Here we study an agent who is willing to sacrifice a fixed amount of expected utility for adaption. How can/ought our agent choose an optimal (in a technical sense) mixed action? We explore consequences of making this choice via entropy minimization, which is argued to be a specific example of risk-aversion. This recovers the ϵ-greedy probabilities known in reinforcement learning. We show that the entropy minimization leads to rudimentary forms of intelligent behavior: *(i)* the agent assigns a non-negligible probability to costly events; but it *(ii)* chooses with a sizable probability the action related to less cost (lesser of two evils) when confronted with two actions with comparable costs; *(iii)* the agent is subject to effects similar to cognitive dissonance and frustration. Neither of these features are shown by entropy maximization.

1. Introduction

Consider an agent who has to choose between a number of different actions $\mathcal{A}_1, \ldots, \mathcal{A}_n$. Before taking action, consequences of each action $\mathcal{A}_k$ are subjectively estimated to have the cost ε_k (or the utility $u_k = -\varepsilon_k$). The basic tenet of decision theory is that the agent ought to choose the action that minimizes the cost (or maximizes the utility) [1]. In terms of probabilities p_k for various actions ($p_k \geq 0$, $\sum_{k=1}^{n} p_k = 1$), this amounts to taking the action ℓ related to the least cost (if it exists and is available):

$$p_\ell = 1, \qquad p_{k \neq \ell} = 0, \qquad \varepsilon_\ell < \varepsilon_{k \neq \ell}, \qquad k = 1, \ldots, n. \tag{1}$$

There are, however, situations where ε_k may change after actions taken, also as a result of those actions.[1] Here is an example that points against choosing (1), and illustrates our problem. You got 100 eggs and several baskets, which seem to have different durabilities (i.e. utilities). The probability with which an action, i.e. a basket, is taken refers to the fraction of

E-mail addresses: armen.allahverdyan@gmail.com (A.E. Allahverdyan), galstyan@isi.edu (A. Galstyan), aliabbas@price.usc.edu (A.E. Abbas), z.r.struzik@p.u-tokyo.ac.jp (Z.R. Struzik).

[1] See section 2 for more details. We stress that we do not mean the delayed reward situation, where the utility is constant, but is discounted by some known factor, because the action is performed now, while its reward will come in future.

https://doi.org/10.1016/j.ijar.2018.10.001
0888-613X/© 2018 Published by Elsevier Inc.

eggs in it. Even if the most durable basket can seem to support all the eggs, it is not wise to put everything in one basket. First the durability of a basket can change due to very eggs inside of it. Second, the durability can change unexpectedly due to hindrances. Third, you loose the possibility to explore other baskets that may turn out to be more durable than you thought. Let us now mention several more, broadly defined situations, where "putting all eggs into one basket" is not good.

– In reinforcement learning the preliminary costs ε_k do change due to the actions taken [3]. Even if these changes are assumed to be predictable, the agent still needs to make several actions before enough experience is accumulated.

– The exploration–exploitation dilemma is known in adaptive (biological, organizational, social) systems; see [4,5] for reviews. Exploring possibilities that seem inferior from a local viewpoint may provide advantages in the long run. Exploitation (in the narrow sense) makes the choice that does seem optimal at the moment of choice. Broader exploitation scenarios do account for adaptivity, but still concentrate on the most useful possibilities [5].

– In creative problem solving there are conceptually simple tasks which are nevertheless not easy to solve in practice because solving them via the least cost (implied by the statement of the problem and/or the previous experience of the solver) is a dead end [6–8]. This *Einstellung* effect is one of the main hindrances of human creativity [6–8]. Creative tasks can be solved only if (subjectively) less probable ways are looked at [6,7].

How to assign prior probabilities to avoid the strictly deterministic (1)? Such probabilities should hold a natural constraint that actions related to higher cost are getting smaller probabilities. Two ad hoc solutions are especially simple: one can take into account only the second-best action, or take all non-best actions with the same (small) probability. In reinforcement learning the latter prior probability is known as the ϵ-greedy [3]. It is preferable to have a regular method of choosing non-deterministic probabilities, which will reflect people's attitudes towards the decision making in an uncertain situation, and which will include the above ad hoc solutions as particular cases.

Here we explore the possibility of defining the prior probabilities via risk minimization (or maximization); see [9,10] for reviews on the notion of risk and its various interpretations. We assume that the agent first decides how much average utility $E - \min_k[\varepsilon_k]$ he invests into exploration by going into nonoptimal—in the sense of not holding (1)—behavior. We employ the notion of risk in a specific context, namely when comparing the behavior of agents having the same utilities for various actions and the same value of E. We argue below that maximizing (minimizing) risk in this specific situation can be done via maximizing (minimizing) the entropy $-\sum_{k=1}^{n} p_k \ln p_k$. People demonstrate both risk minimization (aversion) and maximization (seeking) [12,25], though the risk in those situations is a less specific (and more difficult to describe) notion—first because it involves agents having different utilities for same actions, and second because it involves a difference between the monetary value (gain or loss) and its utility.

Our results show that there are important behavioral differences between entropy-minimizing and entropy-maximizing agents. They are seen for at least three different actions (and the same E). The entropy-minimizing agent implements risk-aversion by weighting the least-cost action more, but he also assigns a non-negligible probability for the high-cost action—whereas the entropy-maximizing agent ignores it. The extent to which the high-cost action is accounted for by the entropy-minimizing agent depends on the amount of utility invested into exploration: investing more utility leads to assigning less probability. As we argue below, this closely relates with the notion of cognitive dissonance [60,61]. Another feature is frustration: due to competing local minima of entropy, the entropy-minimizing agent can abruptly change the action probabilities as a result of a small change of E. Also, when confronted with two actions with different, but comparable costs, the entropy-minimizing agent tends to select the one with a smaller cost (chooses the lesser of two evils), while the entropy-maximizing agent simply does not distinguish between them. The important point is that for a risk-minimizing agent (which does a constrained minimization of a concave function in a convex domain) choosing the probabilities of actions means selecting between several local minima. In contrast, the risk-seeking agent always has a unique and well-defined probabilistic solution that results from minimizing a convex function [13]. We relate the above features of the entropy-minimizing agent with a rudimentary form of intelligence (see Section 7).

The remainder of this paper is organized as follows. Section 2 explains the statement of our problem. Section 3 discusses stochastic dominance, risk, majorization and its relation with entropy. In particular, Section 3.4 provides general remarks and references on entropy optimization. The reader who agrees from the outset with entropy as a measure of risk (and uncertainty) can consult Sections 2 and 3 very briefly. The next two sections, 4 and 5, present details of (resp.) entropy minimization and maximization. The latter may seem to be standard, but Section 5 still contains salient points that are frequently overlooked. Section 6 compares entropy maximization and minimization scenarios from the viewpoint of the agent's behavior. We summarize in Section 7.

2. Statement of the problem

2.1. Costs and constraints

Let us explain the above problem. An agent faces different actions $\{\mathcal{A}_k\}_{k=1}^{n}$ with (resp.) costs $\{\varepsilon_k\}_{k=1}^{n}$. These costs are subjective estimates made by the agent before choosing an action about future consequences of those actions. After making several actions, the agent can change his estimates also as a result of the actions taken. However, before making actions he does not know relative to which specific way the costs will change. In such an agnostic situation, the agent faces two normative demands—he should behave according to $\{\varepsilon_k\}_{k=1}^{n}$, but he also should also explore all actions. Hence he decides