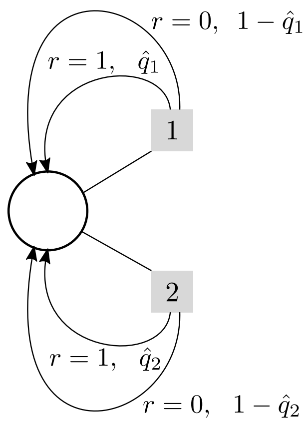
Two-armed Bernoulli bandit

A. Pezzotta*

(Dated: May 14, 2019)

STATEMENT OF THE PROBLEM

A MDP is defined as in Figure.



Only one state is present, that is just symbolic (we do not indicate it anywhere). Two actions are available, the *arms*, labeled 1 and 2. The arms give a stochastic reward R being a Bernoulli variable equal to 0 or 1 characterized by different probabilities $\hat{q}_1$ and $\hat{q}_2$,

$$B_{\hat{q}}(r) \equiv \begin{cases} \hat{q} & \text{for} \quad r = 1 \\ 1 - \hat{q} & \text{for} \quad r = 0 \end{cases}$$

so that

$$\text{Prob}\{R = r | a = j\} = B_{\hat{q}_j}(r) \ .$$

The goal of the agent is to find a policy π, that maximizes the expected discounted return

$$V_\pi = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \, r_t \right]$$

The agent, however, does not know what are the parameters $\hat{q}_1$ and $\hat{q}_2$ of the Bernoulli distributions. Therefore, they have to infer them *while acting*, and at every time step choose the best arm accordingly.

Bernoulli distribution and Beta conjugate prior

If θ is a set of parameters completely specifying a distribution, and x is the result of an experiment, the probability distribution over the parameters θ changes to

$$p'(\theta) = \frac{\ell(x|\theta)\, p(\theta)}{\int d\theta' \, \ell(x|\theta')\, p(\theta')} \ ,$$

* alberto.pezzotta@crick.ac.uk

after the experiment is performed. This is known as *Bayesian update*, and is a specific instance of *Bayes' theorem*. The probability p is the *prior*, i.e. the distribution that encodes knowledge about the parameters θ prior to the experiment. The quantity ℓ is the *likelihood*, i.e. the *model* of the observations, specifying how likely is the result of an experiment given one possible set of values of the parameters. The probability p' is the *posterior*, i.e. the distribution over the set of parameters with the added knowledge of the result of the experiment.

A convenient choice for the prior is a probability distribution which is *conjugate* to the likelihood. That is, belonging to a parametrized family of distribution and, when multiplied by the likelihood (and properly normalized), remaining in the same family. The prior which is conjugate to the Bernoulli distribution, where $\theta \equiv \hat{q} \in [0,1]$, is the *Beta distribution*, defined as

$$\mathrm{Beta}_{\alpha,\beta}(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \ ,$$

where B indicates the *Beta function*.

In general, the parameters of the prior/posterior (α and β, in this case) are referred to as *hyper-parameters*, while the arguments of the prior/posterior distribution (θ) are called *parameters*.

Multiplying $\mathrm{Beta}_{\alpha,\beta}$ times the Bernoulli distribution yields again a Beta distribution, with altered values of the hyper-parameters α and β. Indeed, applying Bayes theorem with $p(\theta) = \mathrm{Beta}_{\alpha,\beta}(\theta)$, $p'(\theta) = \mathrm{Beta}_{\alpha',\beta'}(\theta)$ and $\ell(x|\theta) = B_\theta(x)$, we have that

$$(\alpha', \beta') = \begin{cases} (\alpha+1,\, \beta) & \text{if} \quad x = 1 \ , \\ (\alpha,\, \beta+1) & \text{if} \quad x = 0 \ . \end{cases}$$

This gives a simple update rule for the hyper-parameters of the prior/posterior given the result of a Bernoulli trial.

POMDP for Bernoulli bandits

In the context of Bernoulli bandits, the only observations (experiments) are the rewards, and the likelihood ℓ is the Bernoulli distribution: it specifies the probability of the reward received by one arm given a possible estimate of its parameter $\hat{q}$.

In the framework of POMDP, prior and posterior are referred to as *beliefs*. These encompass all the information about the previous history of observations. The belief depends on time t, through that history, and will be denoted by b_t. In the present bandit problem the agent has partial information about the *laws* according to which the environment behaves, particularly, the probability distribution with which it yields rewards, $\mathrm{Prob}\{R = r | a = j\}$. Although it is known that the reward following each arm are a Bernoulli variables with possible values 0 and 1, the parameters are unknown. The belief is then defined over the space of these parameters [1], denoted $b = b(q_1,\, q_2)$.

Based upon the belief at time t, b_t (the prior), the agents chooses the next action, a_t (pulling either arm, j), according to a policy π. The bandit gives a reward r_t (the result of the experiment), according to which the agent updates its belief through Bayes' rule, obtaining b_{t+1} (the posterior):

$$b_{t+1}(q_1, q_2) = \frac{\ell(r_{t+1}|q_1,\, q_2,\, a_t)\ b_t(q_1, q_2)}{f_t(r_{t+1},\, a_t)} \ ,$$

with $a_t \sim \pi(\cdot|b_t)$, and where we indicated

$$f_t(r,a) = \int dq_1 dq_2\, \ell(r|q_1, q_2, a)\, b_t(q_1, q_2) \ .$$

Mapping to an MDP in hyper-parameter space

If we assume that the two parameters are independent, i.e. the belief is factorized into single-arm beliefs,

$$b(q_1, q_2) = b^1(q_1)\, b^2(q_2) \ ,$$

[1] We indicate with $\hat{\ }$ the *true* parameters, and without the argument of the belief.

and we choose b^j to be Beta distributions with hyper-parameters α_j and β_j, since we have

$$\ell(r|q_1, q_2, a = j) = B_{q_j}(r) \ ,$$

Bayes' rule translates into the following update for the hyper-parameters:

$$\begin{aligned}(\alpha_1,\, \beta_1,\, \alpha_2,\, \beta_2) \mapsto & \mathbb{I}(a_t = 1) \Big[\mathbb{I}(r_{t+1} = 1) \big(\alpha_1 + 1,\, \beta_1,\, \alpha_2,\, \beta_2\big) + \mathbb{I}(r_{t+1} = 0) \big(\alpha_1,\, \beta_1 + 1,\, \alpha_2,\, \beta_2\big)\Big] \\ & + \mathbb{I}(a_t = 2) \Big[\mathbb{I}(r_{t+1} = 1) \big(\alpha_1,\, \beta_1,\, \alpha_2 + 1,\, \beta_2\big) + \mathbb{I}(r_{t+1} = 0) \big(\alpha_1,\, \beta_1,\, \alpha_2,\, \beta_2 + 1\big)\Big] \ .\end{aligned}$$

For instance, an initial prior $b^j_{t=0}$ uniform corresponds to initial values $\alpha_j = \beta_j = 1$. Therefore, if up to time t the arm j has been chosen t_j times, yielding n_j times reward 1 and $m_j = t_j - n_j$ times reward 0, then

$$\alpha_j = n_j + 1 \quad \text{and} \quad \beta_j = m_j + 1 \ .$$

The Bayesian update is equivalent to a random walk on a 4-dimensional lattice, with points identified by the 4-tuples with the numbers of wins and losses per each arm, (n_1, m_1, n_2, m_2). Each of these lattice points defines the state of a Markov process.

Therefore, with the choice of the Beta prior the POMDP in which the parameters of the Bernoulli distribution of rewards are unknown, transforms into a MDP in which the states –the possible combinations of hyper-parameters– are *known*.

The random walk starts from $n_j = m_j = 0$, and always move towards the nearest-neighbouring lattice points with increasing values of n and m.

The action, at time t, is chosen as $a_t \sim \pi(\cdot|s_t)$.

The reward that the agent gets by pulling arm a_t is stochastic. In this formulation, we replace the stochastic reward by its expected value over the current belief. If the state at time t is $s_t = (n_1, m_1, n_2, m_2)$, by choosing action j, the agent gets a reward

$$\begin{aligned} r_t = r\big(s_t, a_t = j\big) \equiv \langle q_j \rangle = \int_0^1 dq\, b^j(q)\, q = \int_0^1 dq\, q\, \frac{q^{n_j}(1-q)^{m_j}}{B(n_j+1,\, m_j+1)} = \frac{B(n_j+2,\, m_j+1)}{B(n_j+1,\, m_j+1)} \\ = \frac{n_j+1}{n_j+m_j+2} \ . \end{aligned}$$

In the last equality, we use the fact that the Beta function can be expressed in terms of the Gamma function,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \ ,$$

and that the latter satisfies

$$\Gamma(z+1) = z\,\Gamma(z) \ .$$

The next state visited is $s_{t+1} \sim p(\cdot|s_t, a_t)$, as in

$$s_{t+1} = \begin{cases} (n_1+1,\, m_1,\, n_2,\, m_2) & \text{w.p.} \ \ \langle q_1 \rangle = \dfrac{n_1+1}{n_1+m_1+2} \\ (n_1,\, m_1+1,\, n_2,\, m_2) & \text{w.p.} \ \ \langle 1-q_1 \rangle = \dfrac{m_1+1}{n_1+m_1+2} \end{cases} \quad \text{if} \ \ a_t = 1 \ ,$$

and

$$s_{t+1} = \begin{cases} (n_1,\, m_1,\, n_2+1,\, m_2) & \text{w.p.} \ \ \langle q_2 \rangle = \dfrac{n_2+1}{n_2+m_2+2} \\ (n_1,\, m_1,\, n_2,\, m_2+1) & \text{w.p.} \ \ \langle 1-q_2 \rangle = \dfrac{m_2+1}{n_2+m_2+2} \end{cases} \quad \text{if} \ \ a_t = 2 \ .$$

The transition probabilities are given as the expected value, over the belief specified by s_t, of the probability of winning, q_j (or losing, $1 - q_j$) when choosing arm j.

The goal of the agent is then to find the policy π that maximizes

$$V_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t\, r_t \Big| s_0 = s\right]\,.$$

The Bellman equation for the MDP in this hyper-parameter space writes

$$V^*(s) = \max_{a\in\{1,2\}} \sum_{s'} p(s'|s,a)\big[r(s,a) + \gamma\, V^*(s')\big]$$

where p and r are defined above. The optimal policy is

$$a_t = \operatorname*{argmax}_{a\in\{1,2\}} \sum_{s'} p(s'|s,a)\big[r(s,a) + \gamma\, V^*(s')\big]$$