Dynamic programming for optimizing evidence accumulation across correlated binary decision trials – Binary observations

Zachary P Kilpatrick & Krešimir Josić

July 29, 2020

The goal of this document is to develop Bellman equations for the accumulation of evidence across trials of a 2AFC task in which the correct choice on each trial is correlated with the previous. *Here we assume no feedback, and each observation is binary.*

Model. We start with a simple model in which an observer can either make a single observation ($a_n = 1$), on a trial or not make an observation ($a_n = 0$). The observer's state s_n is the log likelihood ratio (LLR) of the choice $+$ vs. $-$ being correct in trial n. An incorrect choice yields no reward. The reward is given at the end of the trial. Therefore, on each trial the observer's reward equals the expected reward given the state, *i.e.* the observer tries to optimize the expected reward on each trial by deciding whether or not to pay for information. Note that this is equivalent to the case of no feedback, as the reward does not affect the belief (state).

A correct choice yields a reward $r_n = 1$. However, the reward is uncertain, and the observer can only maximize the expected reward on each trial. Accumulating evidence on average increases the probability of reward $p(r_n = 1|s_n, a_n = 1) \geq p(r_n = 1|s_n, a_n = 0)$. A last important detail of this model is that the correct choice H_n in each trial switches according to a two state Markov process with transition rate α: $p(H_{n+1} \neq H_n) = \alpha$.

We will initially consider the case when $\alpha = 0$. This is related to the examples of infotaxis, as on each trial the observer chooses whether to gather information or not.

Action-conditional state/reward update. We need to derive an expression for $p(s_{n+1}, r|s_n, a_n)$ for the different belief states, and actions. We assume that the conditional distribution for each observation is given by $f_\pm(x_n)$ if $\pm$ is the true state (correct choice) in trial n. In the case of a binary observation we have

$$f_+(x) = \left\{ \begin{array}{cc} p, & x = 1, \\ 1-p, & x = -1, \end{array} \right. \qquad f_-(x) = \left\{ \begin{array}{cc} 1-p, & x = 1, \\ p, & x = -1. \end{array} \right.$$

Thus,

$$z(\pm 1) = \pm \log \frac{p}{1-p} \equiv \pm \Delta.$$

We then find that,

$$p(s_{n+1}|s_n, a_n = 1) = \left\{ \begin{array}{ll} \frac{p + (1-p)\mathrm{e}^{-s_n}}{1+\mathrm{e}^{-s_n}} = p_+(s_n), & \text{when } s_{n+1} = \log \frac{(1-\alpha)e^{s_n + \Delta} + \alpha}{\alpha e^{s_n+\Delta} + (1-\alpha)}, \\ \frac{p\mathrm{e}^{-s_n} + (1-p)}{1+\mathrm{e}^{-s_n}} = p_-(s_n), & \text{when } s_{n+1} = \log \frac{(1-\alpha)e^{s_n - \Delta} + \alpha}{\alpha e^{s_n-\Delta} + (1-\alpha)}. \\ 0 & \text{otherwise.} \end{array} \right. \tag{1}$$

where the belief increment is $\Delta = \log \frac{p}{1-p}$. If $a_n = 0$, then no evidence is accumulated in trial n. The evidence is still discounted between trials so that,

$$p(s_{n+1}|s_n, a_n = 0) = 1 \text{ if } s_{n+1} = \log \frac{(1-\alpha)\mathrm{e}^{s_n} + \alpha}{\alpha \mathrm{e}^{s_n} + (1-\alpha)}.$$

These equations follow from the conditional probability of each observation, *e.g.*

$$p(x_n = 1|s_n, a_n = 1) = \frac{p + (1-p)\mathrm{e}^{-s_n}}{1 + \mathrm{e}^{-s_n}},$$

while the discounting due to the change between trials gives

$$s_{n+1} = \log \frac{(1-\alpha)\mathrm{e}^{s'_n} + \alpha}{\alpha \mathrm{e}^{s'_n} + (1-\alpha)} = D(s'_n).$$

where s'_n is the evidence after the action, and s_{n+1} is the discounted information on the next trial.

To get the full joint probability of the reward and state we use

$$p(s_{n+1}, r_n|s_n, a_n) = p(r_n|s_{n+1}, s_n, a_n)p(s_{n+1}|s_n, a_n).$$

The conditional probability of reward depends only on the belief after choosing to make an observation, or choosing to stay with the previous information, that is after not making an observation. Since the belief after making an observation is a sufficient statistics for the state, the probability of reward if the agent chooses to make an observation is determined by the belief state $s_n + \Delta$ or $s_n - \Delta$. We could express this probability in terms of s_{n+1} as well, but that would just convolute things.

We assume that on every trial the observer makes the choice consistent with their belief, s_n, *i.e.* chooses the option -1 if the belief is negative. Whenever the observer chooses to make an observation, they pay a cost $c > 0$. Due to the symmetry we then have

$$p(r_n = 1 - c|s_{n+1}, s_n, a_n = 1) = \begin{cases} \frac{e^{|s_n+\Delta|}}{1+e^{|s_n+\Delta|}}, & \text{when } s_{n+1} = \log \frac{(1-\alpha)e^{s_n+\Delta}+\alpha}{\alpha e^{s_n+\Delta}+(1-\alpha)} = D(s_n + \Delta), \\ \frac{e^{|s_n-\Delta|}}{1+e^{|s_n-\Delta|}}, & \text{when } s_{n+1} = \log \frac{(1-\alpha)e^{s_n-\Delta}+\alpha}{\alpha e^{s_n-\Delta}+(1-\alpha)} = D(s_n - \Delta) \\ 0 & \text{otherwise,} \end{cases}$$

and

$$p(r_n = -c|s_{n+1}, s_n, a_n = 1) = \begin{cases} \frac{1}{1+e^{|s_n+\Delta|}}, & \text{when } s_{n+1} = \log \frac{(1-\alpha)e^{s_n+\Delta}+\alpha}{\alpha e^{s_n+\Delta}+(1-\alpha)} = D(s_n + \Delta), \\ \frac{1}{1+e^{|s_n-\Delta|}}, & \text{when } s_{n+1} = \log \frac{(1-\alpha)e^{|s_n-\Delta|}+\alpha}{\alpha e^{s_n-\Delta}+(1-\alpha)} = D(s_n - \Delta). \\ 0 & \text{otherwise.} \end{cases}$$

The case when the observer chooses not to make an observation is equivalent, with s_n replacing $s_n \pm \Delta$, and $c = 0$, since there is no price for information,

$$p(r_n = 1|s_{n+1}, s_n, a_n = 0) = \frac{e^{|s_n|}}{1 + e^{|s_n|}}, \text{ when } s_{n+1} = D(s_n),$$
$$p(r_n = 0|s_{n+1}, s_n, a_n = 0) = \frac{1}{1 + e^{|s_n|}}, \text{ when } s_{n+1} = D(s_n),$$

with the probability equal to 0 in all other cases.

Note the relation between this set of equations for the reward and Eq. (1). This is because the choice to make an observation will either increase the belief to $s_n + \Delta$, or decrease the belief to $s_n - \Delta$ which then gives the probability of the reward upon making the decision. Only after the choice is made is the belief discounted to take the value s_{n+1}.

Special case when $\alpha = 0$**.** When there is no change, then we expect that once sufficient information is obtained the observer will not pay the price for more information. We expect this to happen when the cost of acquiring information, c, is greater than the expected increase in payoff due to the obtained information. If the observer does not acquire new information at step n, they will not in the future, so the expected payoff is then $E[r_n + \ldots + r_N|s_n, a_n = 0] = (N - n)E[r_n|s_n, a_n = 0]$ where N is the total number of trials.

I don't see how to get a lower bound on $E[r_n|s_n, a_n = 1]$ since this involves stopping at any intermediate point, which is exactly what dynamic programming does.

Value function maximization. The value function for a given policy, $\pi(a|s)$ can be obtained by

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + v_\pi(s')]$$

since we are assuming no discounting. Note that the states have a different form than above: Each state, s in the is a pair of the form (s_n, n) and transitions to $(s_{n+1}, n+1)$. This is the way to keep track of a fixed number of trials.

Since only two actions are possible, and we assume that the mapping from states to actions is deterministic, we get the following (the middle equation right here abuses notation a bit, the last line is correct):

$$\begin{aligned}
v_\pi(s) &= \sum_{s',r,\pi(s)=0} p(s',r|s,a=0)[r + v_\pi(s')] + \sum_{s',r,\pi(s)=1} p(s',r|s,a=1)[r + v_\pi(s')] \\
&= \sum_{\pi(s)=0} \left(\frac{e^{|s|}}{1+e^{|s|}}[1 + v_\pi(D'(s))] + \frac{1}{1+e^{|s_n|}} v_\pi(D'(s)) \right) \\
&+ \sum_{\pi(s)=1} \left(\frac{e^{|s_n+\Delta|}}{1+e^{|s_n+\Delta|}} p_+(s_n)[1 - c + v_\pi(D'(s+\Delta))] + \frac{1}{1+e^{|s_n+\Delta|}}(-c + v_\pi(D'(s+\Delta)) \right) \\
&+ \sum_{\pi(s)=1} \left(\frac{e^{|s_n-\Delta|}}{1+e^{|s_n-\Delta|}} p_-(s_n)[1 - c + v_\pi(D'(s-\Delta))] + \frac{1}{1+e^{|s_n-\Delta|}}(-c + v_\pi(D'(s-\Delta)) \right) \\
&= \begin{cases} \frac{e^{|s_n|}}{1+e^{|s_n|}} + v_\pi(D'(s)) & \text{when } \pi(s) = 0 \\ p_+(s_n)\left(\frac{e^{|s_n+\Delta|}}{1+e^{|s_n+\Delta|}} - c + v_\pi(D'(s+\Delta))\right) + p_-(s_n)\left(\frac{e^{|s_n-\Delta|}}{1+e^{|s_n-\Delta|}} - c + v_\pi(D'(s+\Delta))\right) & \text{when } \pi(s) = 1 \end{cases}
\end{aligned}$$

where we use $D'(s_n) = (D(s_n), n+1)$, and abused notation in the exponential by assuming that e^s is the exponential of the first term (the belief) in the state pair.

We can use the Bellman equation to define the optimal solution, as the one that satisfies

$$v_{\pi^*}(s) = \max_\pi \begin{cases} \frac{e^{|s_n|}}{1+e^{|s_n|}} + v_{\pi^*}(D'(s)) & \text{when } \pi(s \\ p_+(s_n)\left(\frac{e^{|s_n+\Delta|}}{1+e^{|s_n+\Delta|}} - c + v_{\pi^*}(D'(s+\Delta))\right) + p_-(s_n)\left(\frac{e^{|s_n-\Delta|}}{1+e^{|s_n-\Delta|}} - c + v_{\pi^*}(D'(s+\Delta))\right) & \text{when } \pi(s \end{cases}$$