Operant matching

Sebastian Seung

9.29 Lecture 6: February 24, 2004

1 The law of effect

Thorndike was a pioneer in the scientific study of animal learning. He devised the puzzle box as an experimental method. A special cage was constructed, such that a special sequence of actions was required to open its door. The animal was repeatedly placed in the cage, and the time to escape was tabulated. A graph of the escape time as a function of the number of trials was called a learning curve. Thorndike observed that the animals seemed to learn by a process of trial and error. By generating random actions while in the cage, occasionally they would happen to generate the specific sequence of actions that caused the door to open. This sequence became reinforced by success, so that it became more likely to recur in the future. By this process, the animal learned to reduce its escape time.

Thorndike sought to place psychology on an axiomatic foundation, and wrote down several laws of learning. One of them was the *law of effect*:

> Of several responses made to the same situation those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections to the situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.

The law of effect inspired further work on animal learning that eventually led to the behaviorist school of psychology. The study of learning from reinforcement was eventually formalized in the paradigm of operant or instrumental conditioning.

2 Variable-interval (VI) schedule

In operant conditioning, an animal is trained by making reward contingent on a desired behavior. In the laboratory, there are many methods of creating such contingency. B. F. Skinner and collaborators called these reinforcement schedules, and catalogued many of them.

In a variable-interval (VI) reinforcement schedule, rewards are made available at a target that is chosen by an animal. For example, pigeons choose a key by pecking at it. Some of these pecks are reinforced by food reward. In a VI schedule, the rewards have constant amplitude and are made available at the target at random times.

A target has two possible states: baited or unbaited. If the animal chooses an unbaited target, it receives nothing. If it chooses a baited target, it harvests the reward, and the target switches to the unbaited state. The target is rebaited after a random time interval t drawn from the exponential distribution $P(t) = \tau^{-1} e^{-t/\tau}$. The average time interval is set by the parameter τ.

Recall that the time intervals between events for a Poisson process are also exponentially distributed. In its use of an exponential distribution, the VI schedule has some similarity to a Poisson process. However, note that the baiting times are not Poisson, and neither are the harvesting times. It is the time interval between harvesting and rebaiting that is exponentially distributed.

Note that in the special case where the animal chooses the target very frequently, it will harvest each reward immediately after it becames available. Then the baiting times are approximately Poisson, as are the harvesting times.

3 Concurrent VI schedule

In a concurrent VI schedule, the animal is presented with a choice between multiple targets. Each target is baited according to a VI schedule. In general, the average rebaiting time is different for the targets. Suppose that there are two targets, and that the average rebaiting time is shorter for the first target than for the second. Then the first target is "rich," while the second target is "lean."

The schedules are all independent, except for one complication, a fixed changeover delay. When the animal switches from one target to another, choosing the new target does not result in harvesting reward during the changeover delay.

A changeover delay is natural and inevitable if the targets are placed far apart, so that the animal has to traverse some distance in order to switch targets. When the targets are close together, a changeover delay can be artificially imposed. Effectively, the changeover delay imposes a penalty on switching.

Without the changeover delay, animals tend to alternate between the targets, allocating their choices equally between them. When the changeover delay is larger than some threshold value, then the animal tends to prefer the rich target over the lean. A mathematical theory of such preference is given by the matching law. A typical value for the changeover delay is 1.5 seconds.

4 The matching law

Richard Herrnstein and collaborators trained pigeons using concurrent VI schedules. In such an experiment, the pecks of the pigeon at each key were recorded. Also, the rewards harvested by the pigeon were recorded. The number of pecks at each key was a measure of the pigeon's preference for that key. Therefore, the experiment made

it possible to study how the pigeon's preferences depended on how its choices were reinforced.

Let N and $\bar{N}$ be the number of pecks at the two targets, and H and $\bar{H}$ be the number of rewards harvested at the two targets. Herrnstein and collaborators found the following empirical law:

$$\frac{N}{N+\bar{N}} = \frac{H}{H+\bar{H}}$$

Alternatively, this can be expressed as

$$\frac{N}{\bar{N}} = \frac{H}{\bar{H}}$$

They called this the *matching law*, as the preferences of the pigeon "match" the rewards.

In the language of economics, the number of rewards harvested can be called "income." Then the matching law can be stated as:

> *Choices are in the same ratio as the incomes of the targets.*

For another statement in economic language, rewrite the equation as

$$\frac{H}{N} = \frac{\bar{H}}{\bar{N}}$$

The income from a target divided by the number of pecks can be called the "return," as in "return on investment." Then the matching law can be stated as:

> *The returns from the targets are equal.*

Note that the matching law is not interesting, unless the number of pecks far exceeds the number of rewards. To see why, suppose the opposite is true. Suppose that the VI schedules rebait the targets more quickly than the pigeons can peck. Then every peck would be reinforced, so that $H = N$ and $\bar{H} = \bar{N}$. The matching law would be trivially satisfied.

In a typical experiment on the matching law, the number of pecks might be one hundred times larger than the number of rewards. In this case, large deviations from matching are possible in principle, but empirically such deviations are not observed.

5 VI schedule for discrete trials

The VI schedule can be formulated for discrete trials. Rewards are made available at a target that is chosen by an animal in discrete trials.

A target has two possible states: baited or unbaited. If the animal chooses an unbaited target, it receives nothing. If it chooses a baited target, it harvests the reward, and the target switches to the unbaited state.

If a target is baited at the end of a trial, it remains baited for the next trial. If a target is unbaited at the end of a trial (either because reward was harvested, or because

the target was unbaited to begin with), then it is rebaited according to the toss of a coin with bias p.

This means that a target is rebaited after a number of trials n drawn from the geometrical distribution $P(n) = (1-p)^{n-1}p$. The average n is given by $1/p$. If $p = 1$, then the target is rebaited immediately, so that the average n is one. In the limit $p \to 0$ the time until rebaiting diverges to infinity.

6 Concurrent VI schedule for discrete trials

To implement a concurrent VI schedule for discrete trials, a changeover delay is needed. This is done by not allowing harvesting of reward in the first trial after an animal switches.

This kind of experiment has been implemented by Sugrue, Corrado, and Newsome. Green and red visual targets are presented to a monkey in repeated trials. The monkey chooses one of these targets by making a saccadic eye movement. The monkey is rewarded with juice. In the experiment, the VI schedule is nonstationary. That is, the baiting probability for both targets changes every few hundred trials. This requires the monkey to rapidly change its relative preferences for the targets.

7 A simple learning model

In the matching law, the choice probabilities are in the same ratio as the incomes. In the original research, the matching law was only applied to data aggregated from a single long experiment.

However, the matching law can also be applied to the immediate past to produce a simple learning model. In each trial, the model chooses between targets with probabilities set by the ratio of target incomes in the immediate past.

More formally, suppose there are two targets. The action taken in trial t is indicated by the binary variables a_t and $\bar{a}_t$, which satisfy $a_t + \bar{a}_t = 1$. After each action, the animal receives reward h_t. The incomes in the recent past can be computed by by

$$H_{t+1} = \beta H_t + (1-\beta) h_t a_t \quad (1)$$

$$\bar{H}_{t+1} = \beta \bar{H}_t + (1-\beta) h_t \bar{a}_t \quad (2)$$

This is like convolving the time series $h_t a_t$ by an exponential filter, where the discount factor β sets the time constant of the exponential.

Based on these income histories, the model chooses its action randomly, with odds given by

$$\frac{p_t}{\bar{p}_t} = \frac{H_t}{\bar{H}_t}$$

Here p_t is the probability that $a_t = 1$ and $\bar{p}_t$ the probability that $\bar{a}_t = 1$. These choice probabilities satisfy the constraint $p_t + \bar{p}_t = 1$.

Sugrue et al. found that monkey behavior could be modeled using a relatively short time constant of less than ten trials.

The model was earlier applied by Erev and Roth to a model of learning to play games.