## Supporting Text

### The concurrent variable interval (VI) schedule

Matching behavior is commonly studied with reward schedules in which the probability of reward in a trial depends not only on the subject's current choice but also on past choices. In economic settings, it is usual for the return to diminish with frequency of choice and a common experimental paradigm exhibiting diminishing returns is the concurrent VI schedule (1). On each trial, the subject chooses between two targets. If the chosen target is baited with reward, the subject receives it, and the target becomes empty. An empty target is rebaited probabilistically, according to the toss of a biased coin. Once baited, a target remains baited until it is chosen. Rewards are binary and no more than a single reward can reside in each target. Therefore, the reward schedule has two parameters, the biases of the two coins used to bait the targets. These biases, or baiting probabilities, control whether a target is "rich" or "poor." A VI reward schedule has diminishing returns because a target is less likely to be baited if it has been chosen recently, as a consequence of the fact that reward persists at a target once the target is baited.

### Proof of Theorem 3

**Assumption 2.** $\mathbf{E}[N|A=1]$ *and* $\mathbf{E}[N|\mathrm{A}=2]$ *are different from* $\mathbf{E}[N]$.

**Assumption 3.** *The joint probability distribution of neural activity and choice is identical in each trial, that is, from trial to trial, each draw of* $(N_t, A_t)$ *is independent.*

**Assumption 4.** *Reward* $R_t$ *is independent of neural activity* $N_{t-\tau}$ *when conditioned on the choice* $A_{t-\tau}$.

**Assumption 5.** *There is a unique stationary distribution of the sequence of rewards if alternatives are chosen by tossing a biased coin* (see footnote ** in text).

**Theorem 3**. *Suppose that Assumptions 2, 3, 4 and 5 are satisfied, and define the expected reward* $U(p)=\mathbf{E}[R_t]$. *Then* $U'(p)=0$ *if and only if* $\sum_{\tau=0}^{\infty}\mathrm{Cov}\left[R_t,N_{t-\tau}\right]=0$ .

***Proof.*** Using *Assumptions* 3 and 5, it follows immediately from standard theories about gradient estimation for Markov decision processes (2) that

$$U'(p)=\sum_{\tau=0}^{\infty}\left(\mathbf{E}\left[R_t \mid A_{t-\tau}=1\right]-\mathbf{E}\left[R_t \mid A_{t-\tau}=2\right]\right) \qquad [6]$$

where expectations are taken with respect to the stationary distribution. Define $\delta N_{t-\tau}=N_{t-\tau}-\mathbf{E}\left[N_{t-\tau}\right]$. Then by construction $\mathbf{E}\left[\delta N_{t-\tau}\right]=0$, so that

$$\mathbf{E}\left[\delta N_{t-\tau} \mid A_{t-\tau}=1\right]\mathrm{Pr}[A_{t-\tau}=1]+\mathbf{E}\left[\delta N_{t-\tau} \mid A_{t-\tau}=2\right]\mathrm{Pr}\left[A_{t-\tau}=2\right]=0\,. \qquad [7]$$

*Assumption* 4 about conditional independence implies that

$$\begin{aligned}\mathrm{Cov}\left[R_t,N_{t-\tau}\right]&=\mathbf{E}\left[R_t\delta N_{t-\tau}\right]\\&=\mathbf{E}\left[R_t \mid A_{t-\tau}=1\right]\mathbf{E}\left[\delta N_{t-\tau} \mid A_{t-\tau}=1\right]\mathrm{Pr}\left[A_{t-\tau}=1\right]\\&\quad+\mathbf{E}\left[R_t \mid A_{t-\tau}=2\right]\mathbf{E}\left[\delta N_{t-\tau} \mid A_{t-\tau}=2\right]\mathrm{Pr}\left[A_{t-\tau}=2\right]\end{aligned} \qquad [8]$$

Substituting Eq. **7** into Eq. **8** yields

$$\begin{aligned}\mathrm{Cov}\left[R_t,N_{t-\tau}\right]=&\\&\mathbf{E}\left[\delta N_{t-\tau} \mid A_{t-\tau}=1\right]\mathrm{Pr}\left[A_{t-\tau}=1\right]\\&\cdot\left(\mathbf{E}\left[R_t \mid A_{t-\tau}=1\right]-\mathbf{E}\left[R_t \mid A_{t-\tau}=2\right]\right)\end{aligned} \qquad [9]$$

By *Assumption* 3, $\mathbf{E}[\delta N_{t-\tau} \mid A_{t-\tau}=1]=\mathbf{E}[\delta N_t \mid A_t=1]$ for all τ, so

$$\begin{aligned}\sum_{\tau=0}^{\infty}\mathrm{Cov}\left[R_t,N_{t-\tau}\right]=&\\&\mathbf{E}\left[\delta N_t \mid A_t=1\right]\mathrm{Pr}\left[A_t=1\right]\\&\cdot\sum_{\tau=0}^{\infty}\left(\mathbf{E}\left[R_t \mid A_{t-\tau}=1\right]-\mathbf{E}\left[R_t \mid A_{t-\tau}=2\right]\right)\end{aligned} \qquad [10]$$

By *Assumption* 2 the product $\mathbf{E}\left[\delta N_t \mid A_t=1\right]\mathrm{Pr}\left[A_t=1\right]$ is nonzero. Therefore from Eq. **6** $U'(p)=0$ if and only if $\sum_{\tau=0}^{\infty}\mathrm{Cov}\left[R_t,N_{t-\tau}\right]=0$ .

Thus, under *Assumptions* 2, 3, 4 and 5, maximizing takes place only if the infinite sum of the covariances of past neural activities and current reward vanishes.


1. Davison, M, McCarthy, D (1988) *The Matching Law: A Research Review* (Erlbaum, Hillsdale, NJ).
2. Baxter, J, Bartlett, PL (2001) *J Artif Intell Res* 15:319-350.