

# A “MATCHIMIZING” THEORY OF CONSUMPTION

YONATAN LOEWENSTEIN<sup>1</sup>, DRAZEN PRELEC<sup>2</sup>, AND H. SEBASTIAN SEUNG<sup>1</sup>

<sup>1</sup>HOWARD HUGHES MEDICAL INSTITUTE, BRAIN AND COG. SCI. DEPT.,

AND <sup>2</sup> SLOAN SCHOOL OF MANAGEMENT

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

{YONATANL | DPRELEC | SEUNG}@MIT.EDU

ABSTRACT. It is intuitively plausible that some deviations from “rational” behavior are caused by an incomplete appreciation of the future consequences of present choices. We present a new theory of consumption that crystallizes this intuition, yet is simple enough to serve as a substitute for standard utility theory in microeconomic applications. The theory is based on a model in which hedonic rewards and choices are distributed over time, and hence is most appropriate for consumption of nondurable goods. The model can capture psychological phenomena such as satiation, because hedonic rewards depend on both present and past choices. Equilibrium behavior is defined using the expectations of reward conditioned on choice. Utility maximization results when a consumer fully takes into account long-term dependencies between reward and choice. If the accounting is not complete, then behavior may fail to maximize reward. In particular, if a consumer only takes into account the immediate dependence between reward and choice, then his behavior satisfies Herrnstein’s matching law. Maximizing and matching represent two extremes of a spectrum of “matchimizing” behaviors, each of which is specified by a temporal accounting function. In some cases, this function can be summarized by a single “rationality parameter” that determines where the consumer lies on the spectrum between matching and maximizing. A reinforcement learning algorithm is proposed for achieving the matchimizing equilibrium while adjusting overall expenditure to satisfy a budget constraint. Consumption rates are adjusted depending on the product of the reward and an eligibility trace that retains a memory of past actions, as well as whether the good is more or less expensive than average. Matchimizing is shown to be a steady state of the learning algorithm in the mean field approximation.

## 1. INTRODUCTION

The standard account of consumption found in textbooks is based on utility theory. Behavior is described by a consumption bundle, a vector composed of the amounts consumed of each good. A consumer’s preferences are described by a utility function, which maps a consumption bundle to a scalar. A consumer chooses the consumption bundle that maximizes the utility function.

Although widely accepted in economics, utility theory has been difficult to confirm using aggregate demand data [1, 2, 3]. At the same time, there is evidence from laboratory experiments that animals and humans do *not* maximize utility. Psychologists have quantified the behavior of subjects making choices that lead to rewards. For a variety of reinforcement schedules, behavior is well-approximated by the matching law, which will be explained in detail later[4]. Assuming that utility is the sum of rewards over time,<sup>1</sup>

---

*Date:* Draft version of March 4, 2007.

<sup>1</sup>While modern economists tend to prefer ordinal to cardinal utility, it is difficult to eliminate cardinal utility in contexts when utility must be aggregated over time.

matching is generically different from maximizing<sup>2</sup> [5]. Outside the laboratory, there are many behaviors that—at least on the surface—seem inconsistent with utility maximization. Some have argued that addictive behaviors are also a form of utility maximization [6] (i.e., that substance abuse is rational self-medication<sup>3</sup>). However, it has also been proposed that addiction is an unanticipated byproduct of matching behavior [7].

Intuitively, it is plausible that addiction and certain other types of “irrational” behaviors might involve neglect of the future consequences of choices, at least to some extent. Here we present a new theory of consumption that makes this intuition precise. The theory encompasses both matching and maximizing, but also contains a whole spectrum of “matchimizing” consumption behaviors between these two extremes.

Since our goal is to model partial neglect of future consequences, it is important to have a model of consumption over time. In our model, both individual choices and hedonic rewards are distributed over time. The notion of repeated choices and rewards is most appropriate for consumption of nondurable goods. Rewards depend both on current and past choices, which makes it possible to model psychological phenomena such as satiation.

We show that the necessary conditions for a utility maximum can be expressed in terms of the expectation of reward conditioned on choice, for all possible time lags between the two events. This expression formalizes the idea that an ideal consumer should take into account the future as well as the immediate consequences of choices.

However, it is reasonable to assume that in a complex situation an actual consumer might not fully take into account future consequences. Therefore, we will model actual behavior by altering the necessary condition for a utility maximum, allowing for graded accounting of different time lags between reward and choice. In the limit where the nonzero time lags are neglected completely, this reduces to the matching law studied by psychologists. For more general temporal accounting functions, one obtains a behavior that is a hybrid of matching and maximizing, which is the reason for the term “matchimizing.”

In general, a matchimizing consumer is described by an entire temporal accounting function. However, in two cases this function can be reduced to a single “rationality” parameter. One case is when the temporal accounting function takes on the same value for all nonzero time lags. The other case is when the reward is a permutation symmetric function of past choices. In both of these cases, the condition for matchimizing literally interpolates between matching and maximizing, and the “rationality” parameter controls this interpolation.

Since a matchimizer partially neglects the future consequences of choices, and this neglect is quantified by a temporal accounting function, it is tempting to invoke the concept of temporally discounted reward. However, it can be shown that a matchimizer is *not* a maximizer of temporally discounted reward, in general (see Appendix D).

Because our theory of consumption is based on the hedonic consequences of choices, it is natural to propose a dynamic theory of how the matchimizing equilibrium could be achieved by a reinforcement learning algorithm. The basic idea is that the consumer learns from reward by updating a set of consumption rates at each time step. The update rule is based on an eligibility trace, which retains a memory of its recent choices. If the eligibility

---

<sup>2</sup>There are special reinforcement schedules for which matching is equivalent to maximizing, but this is not generic.

<sup>3</sup>Testing utility theory is complicated by the fact that utility is not directly measurable (there is no “hedonimeter”). Economists have responded to this difficulty by creating the theory of revealed preference, which is based on ordinal rather than cardinal utility. In experimental studies of behavior like the ones mentioned above, all of the choices available to the subject affect the probability of being rewarded in a single common currency, such as food. Therefore it is natural to identify utility with the total amount of food received over the entire experiment.

trace remembers past choices very well, then maximizing results. If the eligibility trace forgets past choices immediately, then matching results. There is also a term in the update rule that depends on whether the price of the good is greater or less than average.

Finally, we conclude with a more philosophical discussion that explains how matchimizing is a challenge to the orthodoxy of marginalism, the idea that subjective value is determined by marginal utility. For a matchimizer, subjective value is given by an interpolation between average utility and differential utility. We explain why this is psychologically plausible for consumption that is distributed over time.

## 2. A MODEL OF HEDONIC REWARDS

The pleasure of eating an ice cream cone may depend on the time elapsed since ice cream was last consumed. Furthermore, the pleasure may depend on whether another good, such as chocolate cake, was consumed in the recent past. In general, the pleasure of consuming a good depends on past consumption events. Such history dependence will be modeled through the equation

$$(1) \quad R(t) = r(\mathbf{A}(t), \mathbf{A}(t-1), \dots, \mathbf{A}(t-W))$$

where  $R(t)$  is the hedonic reward at time  $t$ ,  $\mathbf{A}(t)$  is the consumer's choice at time  $t$ , and  $\mathbf{A}(t-1), \dots, \mathbf{A}(t-W)$  are the consumer's choices at  $W$  previous times. In Eq. (1), reward is assumed to be a deterministic function of choices. More generally, one could define reward as a random variable drawn from a probability distribution conditioned on choices. All of the mathematical results that will follow can easily be extended to this generalization.<sup>4</sup> In principle, the function  $r$  can be arbitrarily complex.<sup>5</sup> It might model the hedonic psychology of the consumer, capturing temporal effects such as satiation arising from repeated choice of the same good. It could also model the way in which the world provides rewards to the consumer.

## 3. A MODEL OF CHOICES

Each choice is assumed to correspond to selecting one good from a basket of  $N$  goods. The  $N$ -dimensional binary vector  $\mathbf{A}(t)$  is defined by<sup>6</sup>

$$A_i(t) = \begin{cases} 1, & \textit{ith good is consumed,} \\ 0, & \textit{otherwise.} \end{cases}$$

Note that choices are assumed to be mutually exclusive: a single good is consumed during each time step. The model can also allow consumption of nothing, by including a good in the basket that represents "consuming nothing" or "staying at home."

In general, human choices may have a very complex temporal structure. However, we will rely on a simple model in which successive choices are drawn at random from a

---

<sup>4</sup>Two further simplifications have been used to make the formalism easy to understand. First, time is modeled in discrete steps. Second, reward is assumed to depend on choices up to  $W$  time steps in the past, but not further. These simplifications are convenient, but not necessary. Appendix A describes how the model can be generalized to a Markov decision process, with an underlying state variable that contains information about the past. The state variable allows reward to depend on choices that were made arbitrarily far in the past. Furthermore, the Markov decision process can be generalized to continuous time, rather than discrete time.

<sup>5</sup>An assumption about symmetry of the function  $r$  will be used later to prove one mathematical result (see Theorem 4), but our other results do not require this assumption.

<sup>6</sup>"A" is for action, used here as a synonym for choice.

probability distribution given by the vector  $\mathbf{p}$ , i.e.,  $\Pr[A_i(t) = 1] = p_i$ . The probabilities are nonnegative and satisfy the normalization constraint

$$(2) \quad \sum_i p_i = 1$$

This i.i.d. model of choice is admittedly simplistic. However, it should be noted that standard utility theory is even more simplistic, with no treatment of time whatsoever. The i.i.d. model is one of the simplest ways of introducing time into a model of consumer choice.<sup>7</sup>

We will also assume that the consumer has a budget, and would like to spend money at the rate  $m$ . This is an additional constraint on the consumption rates,

$$(3) \quad \sum_i \pi_i p_i = m$$

where  $\pi_i \geq 0$  is the price of good  $i$ .<sup>8</sup>

#### 4. MAXIMIZING

The consumer experiences hedonic rewards at a *utility rate*, defined as the time average

$$u(\mathbf{p}) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T R(t),$$

An equivalent definition utilizes an average over the choice probability distribution rather than an average over time,

$$(4) \quad u(\mathbf{p}) = \mathbf{E}[R(t)] = \mathbf{E}[r(\mathbf{A}(t), \mathbf{A}(t-1), \dots, \mathbf{A}(t-W)))]$$

Note that the result of this average is independent of  $t$ , because the probability distribution of the random variables  $\mathbf{A}(t), \mathbf{A}(t-1), \dots, \mathbf{A}(t-W)$  does not depend on  $t$ . In later sections, it will be convenient to define further quantities as averages over the choice probability distribution, but it should be kept in mind that they can also be defined using time averages.

Suppose that the consumer's goal is to maximize the utility rate  $u(\mathbf{p})$  with respect to  $\mathbf{p}$ , subject to the normalization constraint Eq. (2) and the budget constraint Eq. (3). Lagrange multiplier arguments provide a necessary condition for a utility maximum:<sup>9</sup>

$$(5) \quad \frac{\partial u}{\partial p_i} = \lambda \pi_i + \mu$$

The Lagrange multipliers  $\lambda$  and  $\mu$  are chosen so that the two constraints are satisfied. The necessary condition can be expressed in words through the statement that “differential utility” is a linear function of price.<sup>10</sup>

<sup>7</sup>In the continuous time limit, the i.i.d. model would become a Poisson process model in which the time intervals between consumption of a particular good are drawn from an exponential distribution.

<sup>8</sup>In the simulations presented below the price of “stay-at-home” good is zero. However, this is not necessary for our formalism.

<sup>9</sup>Strictly speaking, this is only true for the goods with nonzero consumption,  $p_i > 0$ . The full formalism requires the Karush-Kuhn-Tucker conditions, as discussed in Appendix B.

<sup>10</sup>In the standard theory, marginal utility is proportional to price. Here the extra additive constant  $\mu$  arises from the normalization constraint (2). The difference between differential and marginal utility will be discussed later.

## 5. MATCHING

Maximization is a type of behavior that is expected of consumers on the basis of utility theory. Here we mathematically define another type of behavior called *matching*. There are two motivations for considering matching. The first is empirical: it appears to describe human behavior better than maximizing in certain laboratory experiments [8, 9]. The second is theoretical: matching can be derived mathematically from the assumption that humans treat future rewards differently from immediate rewards. Both of these motivations will be discussed later.

To define matching, it is helpful to decompose the utility rate into the sum of *component utility rates*

$$u_i(\mathbf{p}) = \mathbf{E}[R(t)A_i(t)]$$

(It is straightforward to prove  $u = \sum_i u_i$  using the identity  $\sum_i A_i(t) = 1$ .) The *component utility rate*  $u_i$  can be called “the utility derived from good  $i$ ”, since it includes only the rewards that were received at times  $t$  when good  $i$  was consumed.<sup>11</sup>

We define “matching” behavior to be a set of consumption rates satisfying the condition

$$(6) \quad \frac{u_i}{p_i} = \lambda\pi_i + \mu$$

In the next section we show that in the case of goods with equal prices, Eq. (6) is reduced to the original definition of matching. This resembles the necessary condition (5) for maximizing, except that the differential utility  $\partial u / \partial p_i$  has been replaced by the average utility per choice of good  $i$ , defined by

$$(7) \quad \frac{u_i}{p_i} = \mathbf{E}[R(t) | A_i(t) = 1]$$

The right hand side is the expectation of  $R(t)$ , conditioned on choosing good  $i$  at time  $t$ . This formula, which relates the component utility rates to conditional expectations, will be used often below. The matching law can be summarized by the statement that the average utility is a linear function of price.<sup>12</sup>

## 6. EMPIRICAL EVIDENCE FOR MATCHING

For the special case of goods with equal prices, the budget constraint (3) becomes identical to the normalization constraint (2), and the matching law (6) reduces to

$$(8) \quad \frac{u_i}{p_i} = \text{const}$$

This is the original form of the matching law, which makes no reference to prices or money. It was introduced by psychologists to model data from experiments on animals and humans making repeated choices between alternatives [4]. They used a variety of experimental paradigms, which differ in how the choices are presented to the subject, and how rewards are administered.

---

<sup>11</sup>Attributing a reward only to the good that was just consumed may seem problematic. Strictly speaking, reward is also a function of past consumption events (see Eq. 1). However, it is psychologically plausible that consumers make such attributions. For example, introspection tells us that the pleasure of eating an ice cream cone is attributed to the ice cream cone, and not to the potato chips that were consumed an hour before. Note that decomposing utility into parts is unconventional. In the standard static theory, all rewards are aggregated into a single utility function. But in our theory, rewards and choices are events that occur in time, which is why it is possible to “pigeonhole” each reward to a single good.

<sup>12</sup>Like Eq. (5) this is only valid for those goods with nonzero consumption rate. The generalization to “corner” solutions with both zero and nonzero consumption rates is analogous to the Karush-Kuhn-Tucker conditions, and given in Appendix B.

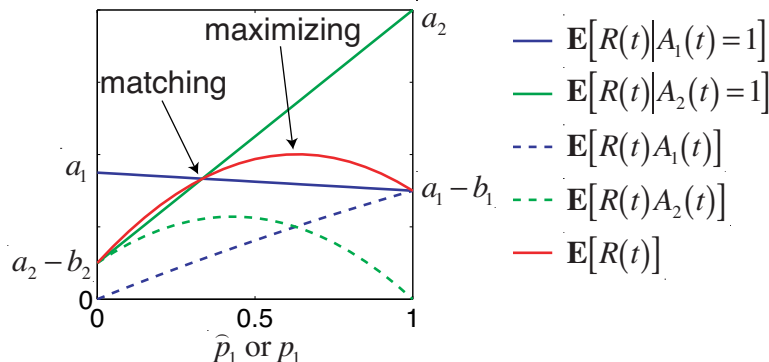


FIGURE 1. Reward schedule based on choice frequency. The blue and green lines specify the reward obtained from choosing alternatives “1” and “2” respectively, as functions of the frequency  $\hat{p}_1$  with which “1” was chosen in the past  $W$  trials. The parameters shown are  $a_1 = 0.35$ ,  $b_1 = 0.05$ ,  $a_2 = 0.8$ , and  $b_2 = 0.7$ . Depicted is the case where both alternatives have “diminishing returns,” i.e., the reward from choosing an alternative decreases with the frequency of choosing the alternative. Matching behavior is at the intersection of the blue and green lines. The dashed green and blue lines are the component utility rates of alternatives “1” and “2” respectively, obtained when the alternatives are chosen by tossing a coin with bias  $(p_1, p_2)$  and the red curve is the total utility rate, the sum of the two component utility rates. Maximizing behavior is at the peak of the red curve.

Research on the phenomenon of operant matching started around 1960 [10, 4]. In the 1980s, Herrnstein, Prelec, and Vaughan [11] introduced an experimental paradigm that conformed to the reward model of Eq. (1). An example reward schedule for repeated choices between two alternatives is depicted in Figure 1. Depending on whether the subject chooses alternative “1” or “2” at time  $t$ , the subject receives a reward amount that is given by either the blue or green line, respectively. The location on the line is determined by the frequency with which the subject chose “1” over the preceding  $W$  time steps (horizontal axis). For instance, suppose that the subject has chosen alternative “1” for 100 percent of the preceding  $W$  trials. If the subject now chooses alternative “1” again, then reward  $a_1 - b_1$  is received. If the subject instead chooses alternative “2”, then reward  $a_2$  is received. Both alternatives exhibit “diminishing returns”: the more frequently the alternative is chosen, the less reward it yields when chosen. Figure 1 shows that matching and maximizing are distinct behaviors for this reward schedule. However, the difference in choice probability between the matching and maximizing solutions is not so large, and neither is the difference in utility rate.

Herrnstein, Prelec, and Vaughan showed how the parameters of the reward schedule can be changed so that the difference between matching and maximizing becomes very large. If  $b_2 = -b_1$  is negative, then Figure 2 results. The slope of the green line is reversed, and the lines become parallel. The reward for choosing alternative “1” is greater (by  $a_1 - a_2 - b_1$  units) than the reward for choosing alternative “2,” irrespective of choice history. However, rewards for both choices decline as a linear function of the percent choices of alternative 1, calculated over the previous  $W$  trials. Hence, choosing “1” exclusively yields  $a_1 - b_1$  while

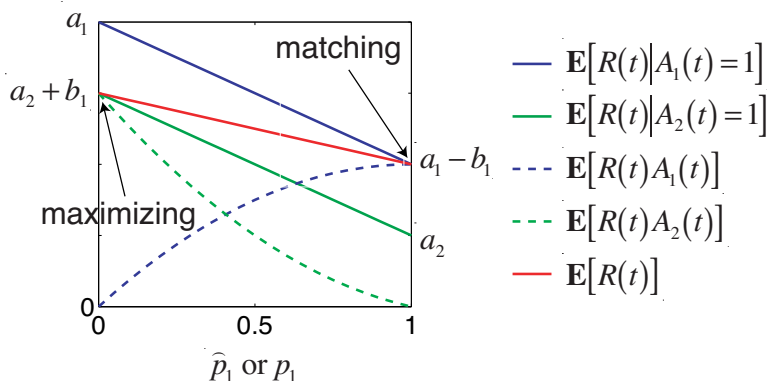


FIGURE 2. Reward schedule with a large separation between matching and maximizing. As in Figure 1, the blue and green lines are the rewards received from choosing alternatives “1” and “2” respectively, as a function of past frequency of choosing “1.” The parameters are chosen so that the blue and green lines are parallel, with the blue line always higher. Maximizing behavior is to choose alternative “2” exclusively, but matching behavior is to choose alternative “1” exclusively.

choosing “2” exclusively yields  $a_2 + b_1$ . Choosing a mixture of both “1” and “2” yields a total reward that interpolates between these two extremes (red line). Although maximizing behavior is to choose “1” exclusively, human subjects typically gravitate toward the corner that minimizes reward rate [8]. In other words, humans tend to generate matching behavior in this reward schedule, rather than maximizing. Note that here the maximizing and matching behaviors are on the corners, rather than in the interior. Revisions of the conditions (5) and (6) to accommodate corner solutions are given in Appendix B.

Not surprisingly, larger  $W$  values push subjects closer to the minimizing corner. However, even with  $W = 1$ , fewer than one-half of the subjects “solve” the problem and avoid alternative 1. Other procedural details also affect the distribution of choices. For example, providing a numerical rather than an analogue measure of reward improves performance [12], while stressing subjects with aversive images during the task depresses it [13].<sup>13</sup>

<sup>13</sup>Support for the matching law (8) originally came from animal studies based on choices and rewards occurring in continuous time, rather than discrete time, with concurrent variable-interval (VI) schedules of reinforcement. A VI schedule is programmed like a mailbox, depositing reward according to a response-independent timer. As with a mailbox, the next response after delivery collects the reward. The reward timer pauses while the schedule is baited, which creates some positive returns to increased response rate. These returns are vanishingly small, as the rates of responding are orders of magnitude greater than rates of reward [14]. The concurrent VI-VI experiment does not sharply distinguish between matching and maximizing, because any distribution of responding across the two alternatives yields close to the maximum overall reward rate. A better contrast is presented by concurrent variable-interval variable-ratio (VR) combinations. The VR is a one-armed bandit, rewarding each response with a fixed probability. On the VI, however, the probability of reward is, to a first approximation, inversely proportional to choice frequency. Hence, maximization would require steady work on the VR, with only an occasional check of the VI. However, matching (8) predicts that the subject will keep returning to the VI until the probability of reward there is reduced to the same level that is set by the VR side. The data present a somewhat mixed picture. One typically observes matching but with a slight bias toward the VR side, which is in the direction of maximizing, but the bias is far from what would be needed to maximize. This pattern holds for both animal [5, 15, 16] and human subjects [17]. Overall, performance falls between matching and maximizing, but is closer to the matching point.

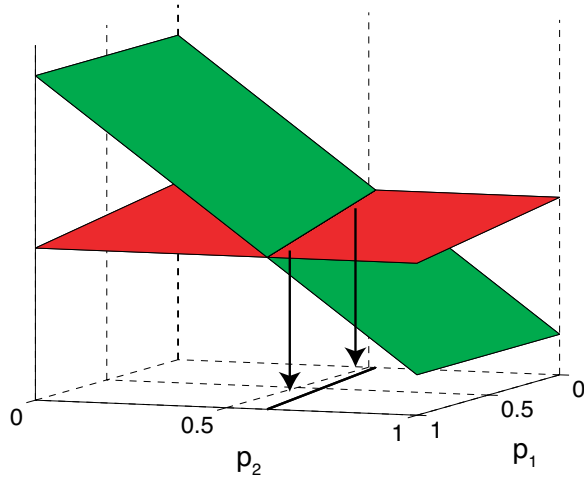


FIGURE 3. The matching set is defined as the solutions of Eq. (10). Shown is the example  $u_i/(\pi_i p_i) = (a_i - b_i p_i)/\pi_i$ , for which the graphs of  $u_1/(\pi_1 p_1)$  and  $u_2/(\pi_2 p_2)$  are planes, and intersect in a line. The projection of this line into the  $(p_1, p_2)$  plane is the matching set.

Experiments have also been performed with other reward schedules besides that of Figure 2, but yielded results that are more equivocal. Egelman et al. used a reward schedule with a long history dependence ( $W = 40$ ) and observed a bimodal distribution of choice frequencies, with some subjects close to matching and others intermediate between matching and maximizing [9]. Herrnstein et al. experimented with shorter and longer history dependence, and found variable results across subjects[12]. They also often observed behaviors that are intermediate between matching and maximizing. It is an interesting open question whether such experimental results can be explained by the theory to follow.

## 7. MATCHING AND PRICES

Figures 1 and 2 gave a graphical depiction of the original matching law (8). How does this change when matching is generalized to incorporate prices, as in Eq. (6)? Suppose that there exists a “stay at home” alternative “0” with zero immediate reward and zero price, so that  $u_0 = 0$  and  $\pi_0 = 0$  by construction. Then Eq. (6) for the “stay at home” alternative yields  $\mu = 0$ . For all other alternatives  $i > 0$ , the matching law Eq. (6) becomes

$$(9) \quad \frac{1}{\pi_i} \frac{u_i}{p_i} = \lambda$$

In words, the average utility per choice and per dollar is equal for all goods with positive price.

To consider a specific example, suppose that there are just three alternatives, two with positive price plus the “stay at home” alternative. Then the matching law takes the form

$$(10) \quad \frac{1}{\pi_1} \frac{u_1}{p_1} = \frac{1}{\pi_2} \frac{u_2}{p_2}$$

The solutions of this equation can be visualized by graphing  $u_1/(\pi_1 p_1)$  and  $u_2/(\pi_2 p_2)$  as functions of  $p_1$  and  $p_2$ . The intersection of these two graphs, when projected into the

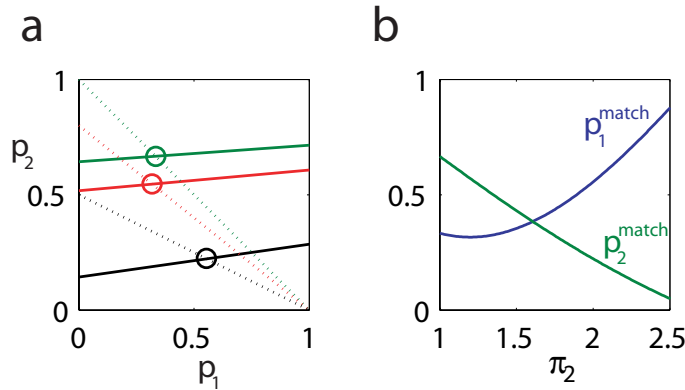


FIGURE 4. Effect of prices on matching behavior. (a) solid lines are the matching sets, for three prices of alternative “2”: green,  $\pi_2 = 1$ ; red,  $\pi_2 = 1.25$ ; and black,  $\pi_2 = 2$ . Dotted lines are budget constraints for the three prices. Circles are the matching solutions. (b) The dependence of the matching solution on the price of alternative “2”. Blue,  $p_1^{match}$ ; green,  $p_2^{match}$ . Reward schedule is as in Figure 1 with a “stay at home” alternative of zero price and immediate reward.  $m = 1$ .

$(p_1, p_2)$  plane, will be called the *matching set*. It contains all matching behaviors consistent with a fixed set of prices and all possible expenditure rates  $m$ .<sup>14</sup> The intersection of the matching set with the budget constraint  $\pi_1 p_1 + \pi_2 p_2 = m$  yields matching solutions for a particular expenditure rate. These intersections are illustrated in Figure 3 for the special case where  $u_i/p_i$  is a linear function of  $p_i$  for all  $i$ .

Changing the price of an alternative has two effects. First, it changes the corresponding average utility per choice per dollar surface, which causes the matching set to shift. Second, it changes the location of the budget constraint. For example, the matching set for three prices of alternative “2” is presented in Figure 4a (solid lines, green,  $\pi_2 = 1$ ; red,  $\pi_2 = 1.25$ ; black,  $\pi_2 = 2$ ). The increase in the price of “2” shifts the matching set in the direction of less consumption of “2”. The budget constraint for the three prices is depicted here by three dotted lines. The larger the price of good “2”, the less can be consumed. The intersections of the solid and dotted lines (circles) are the matching solutions. The consumption rates of “1” and “2” are graphed as functions of  $\pi_2$  in Fig. 4b. For smaller  $\pi_2$  both rates decrease, i.e. the dominant effect of the price increase is an overall drop in consumption. For larger  $\pi_2$ , the consumption of “2” goes down while the consumption of “1” goes up, i.e., the dominant effect is a substitution of “1” for “2”. This suggests that in a theory of consumption based on matching, changes in demand involve both an income effect and a substitution effect. Both effects are formalized in standard utility theory using the Slutsky equation.

<sup>14</sup>It is not the entire  $(p_1, p_2)$  plane that is relevant, but the set of nonnegative  $p_1$  and  $p_2$  satisfying  $p_1 + p_2 \leq 1$ . The inequality follows from the fact that  $p_1 + p_2 = 1 - p_0$  by the normalization constraint, and the rate  $p_0$  of “staying at home” is nonnegative.

## 8. DIFFERENTIAL UTILITY

We have defined two notions of equilibrium behavior for a consumer: maximizing and matching. Maximizing is a basic tenet of utility theory, while evidence for matching comes from certain laboratory experiments. The matching law (6) differs from maximizing (5) only by the substitution of the average utility  $u_i/p_i$  for the differential utility  $\partial u/\partial p_i$ . This is a hint that there is some deeper mathematical relationship between the two. To understand this relationship, it will be helpful to derive a formula for the differential utility that involves the statistical dependence of rewards on choices.

The differential utility  $\partial u/\partial p_i$  was defined by differentiation of the utility rate  $u(\mathbf{p})$  without regard for the normalization constraint  $\sum_i p_i = 1$ .<sup>15</sup> Taking the derivative requires that the definition of utility rate be extended to all nonnegative vectors  $\mathbf{p}$ , not just those satisfying the normalization constraint. This extension is possible for the definition of utility rate given in Eq. (4). In the i.i.d. choice model, the probability of the action sequence  $\mathbf{a}(t), \mathbf{a}(t-1), \dots, \mathbf{a}(t-W)$  is  $\prod_{i=1}^N p_i^{\sum_{\tau=0}^W a_i(t-\tau)}$ . Therefore the utility rate takes the explicit form

$$(11) \quad u(\mathbf{p}) = \sum_{\mathbf{a}(t)} \cdots \sum_{\mathbf{a}(t-W)} r(\mathbf{a}(t), \mathbf{a}(t-1), \dots, \mathbf{a}(t-W)) \prod_{i=1}^N p_i^{\sum_{\tau=0}^W a_i(t-\tau)}$$

This quantity has an interpretation as an expectation value only for normalized probability vectors  $\mathbf{p}$ . However, it is also formally defined for vectors that are not normalized, so it can be differentiated to obtain  $\partial u/\partial p_i$ . The result of the calculation is given by the following theorem, which reveals that the differential utility involves the statistical dependence between reward and past choices.

**Theorem 1.** *Suppose that rewards are a function of the present action and the  $W$  past actions, actions are chosen i.i.d. from a probability distribution  $\mathbf{p}$ , and utility is defined as in Eq. (11). Then the differential utility is given by the sum of conditional expectations,*

$$(12) \quad \frac{\partial u}{\partial p_i} = \sum_{\tau=0}^W \mathbf{E}[R(t)|A_i(t-\tau) = 1]$$

*Proof.* Applying  $\partial/\partial p_i$  to Eq. (11) creates a factor  $\sum_{\tau=0}^W a_i(t-\tau)/p_i$  in the sum, so that

$$\frac{\partial u}{\partial p_i} = \sum_{\tau=0}^W \mathbf{E} \left[ R(t) \frac{a_i(t-\tau)}{p_i} \right]$$

This implies (12), since  $\mathbf{E}[R(t)|A_i(t-\tau) = 1] = \mathbf{E}[R(t)A_i(t-\tau)]/p_i$ . □

By the theorem, the differential utility is the sum of the conditional expectations  $\mathbf{E}[R(t)|A_i(t-\tau) = 1]$ , which measure the statistical dependency of reward on past choices. These expectations depend on the time lag  $\tau$  but not on the absolute time  $t$ , so that

$$\mathbf{E}[R(t)|A_i(t-\tau) = 1] = \mathbf{E}[R(t+\tau)|A_i(t) = 1]$$

---

<sup>15</sup>We use the term “differential utility” rather than “marginal utility”, because  $\partial u/\partial p_i$  does not have a direct economic interpretation. However, the difference of differential utilities  $\partial u/\partial p_i - \partial u/\partial p_j$  has an economic interpretation as the sensitivity of the utility rate to infinitesimal substitutions of good  $i$  for good  $j$ . According to standard utility theory, marginal utility is proportional to price at the utility maximum. Here differential utility is a linear function of price. The additive constant  $\mu$  arises because of the normalization constraint on the rates.

Therefore Eq. (12) can also be written as

$$(13) \quad \frac{\partial u}{\partial p_i} = \sum_{\tau=0}^W \mathbf{E}[R(t+\tau)|A_i(t)=1]$$

which involves the statistical dependency of future rewards on the present choice. Comparing Eqs. (12) and (13), we see that the differential utility can be written either in terms of the future or the past: statistical dependence between present reward and past choices, or between present choice and future rewards.

Substituting Eq. (13) into the necessary condition (5) for utility maximization yields

$$(14) \quad \sum_{\tau=0}^W \mathbf{E}[R(t+\tau)|A_i(t)=1] = \lambda\pi_i + \mu$$

A similar expression involving past choices can be derived by substituting Eq. (12) into Eq. (5). Either of these expressions allows the derivation of a sufficient condition for equivalence of matching and maximizing.

**Theorem 2.** *Suppose that  $W = 0$  in Eq. (14), i.e., the reward  $R(t)$  depends on the present choice  $\mathbf{A}(t)$  but not past choices. Then matching and maximizing are equivalent.*

*Proof.* If the reward doesn't depend on past choices, then the conditional expectations  $\mathbf{E}[R(t+\tau)|A_i(t)=1]$  for  $\tau > 0$  are all equal to  $\mathbf{E}[R(t)]$ . Since these terms of the sum (14) don't depend on  $i$ , they can be subsumed in the constant  $\mu$ . This implies that the maximizing condition (14) is equivalent to the matching condition (6).  $\square$

By this Theorem, matching and maximizing are equivalent for the multi-armed bandit (if Eq. 1 is generalized to the case of stochastic reward). This is because the component utilities are proportional to consumption rates, so that average utility is the same as differential utility.

## 9. MATCHIMIZING

According to Eq. (14), a consumer can maximize utility, given knowledge of the relationship between choices and rewards, as given by the conditional expectations  $\mathbf{E}[R(t+\tau)|A_i(t)=1]$ . For  $\tau = 0$ , this is the relationship between choice and immediate reward. For  $\tau > 0$ , this is the relationship between choice and future rewards. It is plausible that humans may neglect at least partially the future consequences of choices. Therefore we hypothesize that consumer behavior would be better described by (14) if  $\tau > 0$  were treated differently from  $\tau = 0$ ,

$$(15) \quad \mathbf{E}[R(t)|A_i(t)=1] + \sum_{\tau=1}^W w_\tau \mathbf{E}[R(t+\tau)|A_i(t)=1] = \lambda\pi_i + \mu$$

The coefficients  $w_\tau$  will collectively be called the *temporal accounting function*. We call this new description of consumer behavior “matchimizing,” for reasons that will be explained below.

## 10. INTERPOLATING BETWEEN TWO ENDS OF A SPECTRUM

As an extreme case, suppose that the consumer ignores completely the conditional expectations at nonzero time lag,  $w_\tau = 0$  for all  $\tau > 0$ , Fig. 5a. Then matchimizing reduces to the condition  $\mathbf{E}[R(t)|A_i(t) = 1] = \lambda\pi_i + \mu$ , which is equivalent to the matching law (6) by Eq. (7). On the other hand, if nonzero time lags are considered by the consumer as equal in importance to zero time lag,  $w_\tau = 1$  for all  $\tau$ , Fig. 5b, then matchimizing is equivalent to the maximizing condition (5).

Therefore, matchimizing includes maximizing and matching as special cases. Furthermore, matchimizing also includes behaviors that are a hybrid of maximizing and matching, in a sense that is made precise by the following theorem.

**Theorem 3.** *Suppose that  $w_\tau = \alpha$  for all  $\tau > 0$  with  $0 < \alpha < 1$ , and consider an arbitrary reward function (1). Then matchimizing (15) is equivalent to the condition*

$$(16) \quad (1 - \alpha) \frac{u_i}{p_i} + \alpha \frac{\partial u}{\partial p_i} = \lambda\pi_i + \mu$$

*Proof.* Using Eq. (7) and the assumption about  $w_\tau$  (Fig. 5c), Eq. (15) becomes

$$\frac{u_i}{p_i} + \alpha \sum_{\tau=1}^W \mathbf{E}[R(t + \tau)|A_i(t) = 1] = \lambda\pi_i + \mu$$

From Eq. (13) and Eq. (7) it follows that

$$\frac{\partial u}{\partial p_i} - \frac{u_i}{p_i} = \sum_{\tau=1}^W \mathbf{E}[R(t + \tau)|A_i(t) = 1]$$

Combining these two equations yields Eq. (16). □

As  $\alpha$  varies from 0 to 1, matchimizing ranges from matching to maximizing. For intermediate values of  $\alpha$ , matchimizing is a hybrid behavior in which a linear interpolation between average utility and differential utility appears.

The accounting function  $w_\tau$  treated all nonzero time lags  $\tau = 1, \dots, W$  the same way, but zero time lag differently. This is reminiscent of the “beta-delta” model that has been proposed for temporal discounting of reward, with  $\delta = 1$  [18]. However, a matchimizer is *not* the same as a maximizer of temporally discounted reward, as is explained in Appendix D.

More complicated accounting functions (Fig. 5d) do not generally lead to the simple linear interpolation form of (16), unless some further assumption can be made about the structure of the reward function.

**Theorem 4.** *Suppose that the reward function (1) is symmetric under permutations of the past actions  $\mathbf{A}(t - 1), \dots, \mathbf{A}(t - W)$ , and consider an arbitrary accounting function  $w_\tau$ . Then matchimizing (15) is equivalent to the linear interpolation form (16), with*

$$\alpha = \frac{1}{W} \sum_{t=1}^W w_\tau$$

*Proof.* Use the fact that the conditional expectations  $\mathbf{E}[R(t)|A_i(t - \tau) = 1] = \mathbf{E}[R(t + \tau)|A_i(t) = 1]$  are equal for all  $\tau = 1, \dots, W$ , given the assumption of symmetry of the reward function. Therefore,  $\sum_{\tau=1}^W w_\tau \mathbf{E}[R(t + \tau)|A_i(t) = 1] = \alpha \left( \frac{\partial u}{\partial p_i} - \frac{u_i}{p_i} \right)$  □

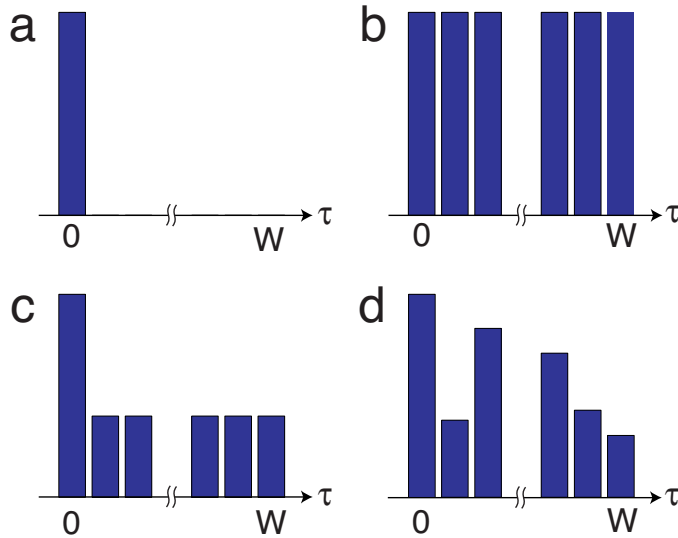


FIGURE 5. Temporal accounting function  $w_\tau$  for various cases considered in the text, with the convention that  $w_0 = 1$ , and with maximum time lag  $W$ . (a) Matching (b) Maximizing (c) Theorem 3 (d) Theorem 4

## 11. REINFORCEMENT LEARNING

As defined above in Eqs. (5), (6), and (15), matching, maximizing, and matchimizing are equilibrium concepts. No account has been given of how a consumer might arrive at these equilibrium behaviors. Here we turn to an account based on ideas from reinforcement learning, in which an agent is assumed to learn from the consequences of its actions on rewards [19, 20, 21]. Reinforcement learning is natural because the equilibrium behaviors of interest can be written using the conditional expectation  $\mathbf{E}[R(t)|A_i(t-\tau) = 1]$ , which measures the dependence of reward on choices.

In a reinforcement learning model, the consumer is assumed to maintain a set of consumption rates  $p_i$ . After every reward, the consumer updates the consumption rates. A desirable property for such an update rule is that it converge to an equilibrium behavior, such as matching, maximizing, or matchimizing.

The equilibrium concept of matchimizing is more general than the reinforcement learning model that will follow. It might also be possible to achieve the matchimizing equilibrium with other types of learning models besides reinforcement learning [22]. It should be stressed that reinforcement learning models are just one type of dynamical model compatible with matchimizing.

## 12. LEARNING MATCHING

We first tackle the issue of a reinforcement learning algorithm for matching. This is the simplest case, and later will be generalized to matchimizing. Consider the update rules

$$(17) \quad \Delta p_i(t) = \eta_p [R(t) (A_i(t) - p_i(t)) - \lambda(t)(\pi_i - c(t))p_i(t)]$$

$$(18) \quad \Delta \lambda(t) = \eta_\lambda \left( \sum_i \pi_i A_i(t) - m \right)$$

where  $c(t) = \sum_i \pi_i p_i(t)$  is the expenditure rate, and  $\eta_p$  and  $\eta_\lambda$  are parameters that control the rate of learning. The first update rule (17) governs the consumption rates. It is straightforward to show that this rule conserves the normalization constraint  $\sum_i p_i = 1$ . The update rule (18) governs the Lagrange multiplier  $\lambda$ .

The update rules (17) and (18) are stochastic, since they contain the random variables  $A_i(t)$ . A deterministic approximation to the update rules is obtained by replacing the right hand sides of (17) and (18) with their expectation values,

$$(19) \quad \Delta p_i \approx \eta_p p_i (\mathbf{E}[R(t)|A_i(t) = 1] - \mathbf{E}[R(t)] - \lambda(\pi_i - c))$$

$$(20) \quad \Delta \lambda \approx \eta_\lambda (c - m)$$

This is called the *mean field approximation*. It is often true that a stochastic dynamics behaves qualitatively the same as its mean field approximation. However, there can sometimes be important differences.

A steady state of the mean field approximation satisfies the budget constraint  $c = m$ , as well as

$$\begin{aligned} 0 &= p_i (\mathbf{E}[R(t)|A_i(t) = 1] - \mathbf{E}[R(t)] - \lambda(\pi_i - m)) \\ &= p_i \left( \frac{u_i}{p_i} - u - \lambda(\pi_i - m) \right) \end{aligned}$$

for every  $i$ . Thus, the goods with nonzero consumption,  $p_i > 0$ , obey the matching law Eq. (6), where  $\mu = u - \lambda m$ . Therefore matching is a steady state of the mean field approximation to the stochastic dynamics (17) and (18). This suggests (but does not prove) that the stochastic dynamics has a stationary state that is similar to matching.

To investigate this possibility, we have performed numerical simulations of the stochastic learning dynamics of Eqs. (17) and (18). We used a reward schedule like the linear one discussed in Section 6 and Appendix C, except there is also a third “stay at home” alternative with zero immediate reward and zero price. The results of these simulations are illustrated in Fig. 6, where we plot the probability of choosing alternatives “1,” “2,” and “stay at home” (red, black and blue solid lines respectively). Initially, the probabilities of the three alternatives are equal, violating the budget constraints. With time, the dynamics converges to the expected matching behavior with the budget constraint satisfied. At time  $t = 10^4$ , the price of alternative “2” is decreased. This results in increased consumption of that alternative at the expense of the “stay at home” alternative. In order to estimate the quality of the mean field approximation, the probabilities of choice, as calculated from Eqs. (19) and (20) is depicted by the dotted lines. Deviations from the mean field approximation are expected to increase if the learning rate parameters,  $\eta_p$  and  $\eta_\lambda$ , are increased. Also, if the simulation is run for very long times, the choice probabilities can become “stuck” at a “corner” solution at which one alternative is chosen exclusively. This is an absorbing state for the stochastic dynamics, although it is not a steady state of the mean field approximation. A detailed comparison of the mean field approximation and the stochastic dynamics is out of the scope of the present paper, and will be discussed elsewhere.

### 13. RELATION TO OTHER LEARNING MODELS

The learning update rules (17) and (18) are related to a number of other previous learning models. If all prices are equal, then  $\pi_i = c(t)$ , reducing Eq. (17) to

$$\Delta p_i = \eta_p R(t) [A_i(t) - p_i(t)]$$

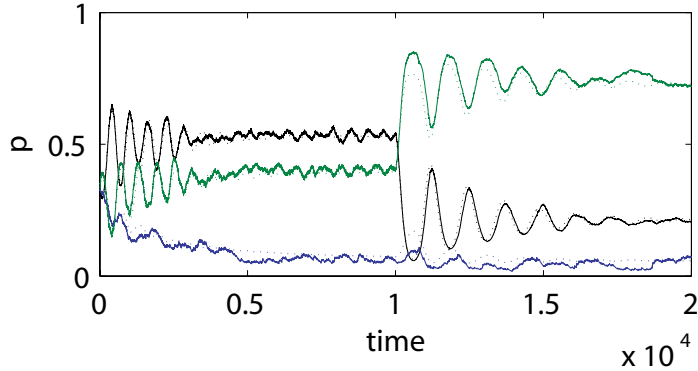


FIGURE 6. Numerical simulations of the stochastic learning dynamics of Eqs. (17) and (18) for a reward schedule that is linear in choice frequencies. There are three alternatives: “1” (blue), “2” (green), and “stay at home” (black) with zero immediate reward and zero price. The dynamics converges to the matching law with prices, Eq. (6). At time  $t = 10^4$ , the price of alternative “2” is decreased, so its consumption increases. The average utilities were  $u_1/p_1 = 0.35 - 0.05p_1$  and  $u_2/p_2 = 0.8 - 0.7p_2$ . The price of alternative “1” was  $\pi_1 = 1.5$ . The price of alternative “2” was  $\pi_2 = 2.25$  for the first half of the simulation, and  $\pi_2 = 1.25$  for the second half. The length of the history dependence was  $W = 3$ . The learning rate parameters were  $\eta_p = \eta_\lambda = 0.01$ . The desired expenditure rate was  $m = 1$ . The initial conditions were  $p_i = 1/3$  and  $\lambda = 0$ .

This model is known as the linear reward-inaction ( $L_{R-I}$ ) algorithm to computer scientists [23] and Cross’ learning model to economists [20, 24]. The steady state of its mean field approximation is the original matching law (8) without prices. It is similar to the melioration model of Herrnstein and Prelec [8].

Therefore, Eq. (17) can be viewed as the generalization of the  $L_{R-I}$  algorithm to prices. Prices enter through the second term on the right hand side of Eq. (17). Assuming that the sign of the Lagrange multiplier  $\lambda$  is positive, this term biases behavior in the direction of choosing goods with prices that are lower than the average expenditure  $c$ . This bias is larger when the value of  $\lambda$  is larger.

The value of  $\lambda$  is determined by Eq. (18). According to the mean field approximation given in the previous section, when the actual expenditure  $c$  is larger than the desired expenditure  $m$ , the value of  $\lambda$  increases to bias the learning towards less expensive goods. The opposite is true for  $c < m$ . Thus Eqs. (17) and (18) allow subjects to learn to match while also conforming to the budget constraint.

The general idea of a dynamical system for the consumption bundle and the Lagrange multiplier for the budget constraint was proposed by Arrow and collaborators as a model for utility maximization by a consumer[25].

#### 14. LEARNING MATCHIMIZING

The learning rules for matching contained the reward and choice at equal times. To generalize them to matchimizing, the learning rules must contain past choices as well as

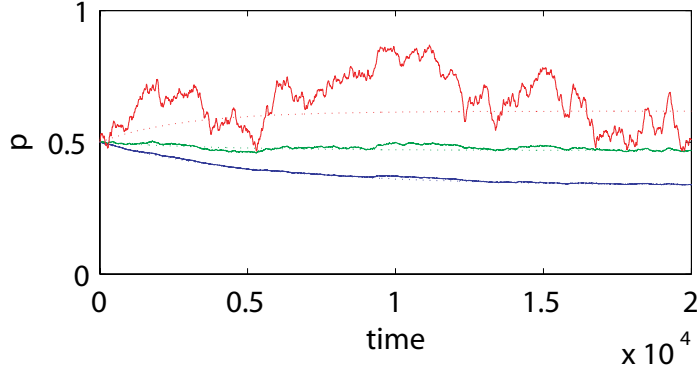


FIGURE 7. Learning to matchimize. Numerical simulations of the probability of choosing alternative “1” in the stochastic learning dynamics of Eq. (21) (solid line), and its mean field approximation, Eq. (22) (dotted line) for the reward schedule described in Figure (1).  $\gamma = 0$ , blue;  $\gamma = 0.5$ , green;  $\gamma = 0.95$ , red. The learning rate was  $\eta_p = 0.001$ . The initial conditions were  $p_1 = 0.5$ .

the present choice. This is done by defining an “eligibility trace”

$$E_i(t) = \gamma E_i(t-1) + A_i(t) - p_i(t)$$

which “remembers” previous choices. The update rule for the consumption rates involves this eligibility trace,

$$(21) \quad \Delta p_i(t) = \eta_p [R(t)E_i(t) - \lambda(t)(\pi_i - c(t))p_i(t)]$$

The update rule for  $\lambda$  remains the same, Eq. (18). The mean field approximation to Eq. 21 is

$$(22) \quad \Delta p_i \approx \eta_p p_i \left( \sum_{\tau=0}^{\infty} \gamma^\tau (\mathbf{E}[R(t)|A_i(t-\tau) = 1] - \mathbf{E}[R(t)]) - \lambda(\pi_i - c) \right)$$

The steady state of the mean field approximation is matchimizing with an exponential accounting function  $w_\tau = \gamma^\tau$ .

In order to study the validity of the mean field approximation to the stochastic dynamics, we have performed numerical simulations of the stochastic dynamics, Eq. (21) and the deterministic approximation, Eq. (22). We used the reward schedule of Figure 1, disregarding the “stay at home” alternative and assuming that the two alternatives have equal prices (Figure 7). When  $\gamma = 0$  (blue), the matchimizing learning equations are the  $L_{R-I}$  equations, and mean field dynamics (dotted line) converges to the matching solution. When the learning rate  $\eta_p$  is sufficiently small, the deviations of the stochastic dynamics (solid line) from the deterministic mean field approximation are small. As the value of  $\gamma$  is increased (green,  $\gamma = 0.5$ ; red,  $\gamma = 0.95$ ), the dynamics of the mean field approximation converges to values that are closer to the maximizing value (dotted lines). Note that the larger the value of  $\gamma$ , the larger the deviations of the stochastic dynamics from the mean field approximation (solid lines). A detailed comparison of the mean field approximation and the stochastic dynamics is out of the scope of the present paper, and will be discussed elsewhere.

## 15. DISCUSSION

Equation (16) could be regarded as the central result of this paper, and is worth repeating here

$$(1 - \alpha) \frac{u_i}{p_i} + \alpha \frac{\partial u}{\partial p_i} = \lambda \pi_i + \mu$$

This equation includes both maximizing ( $\alpha = 1$ ) and matching ( $\alpha = 0$ ) as special cases. The equation was derived theoretically by using a model of consumption over time to write the necessary condition (14) for maximizing behavior in terms of the conditional expectation of reward on previous actions. If the necessary condition is modified so that future rewards are weighted differently than immediate reward, then the matchimizing condition (15) results. For two special cases, we proved that matchimizing reduces to the simple equation above, which involves a linear interpolation between average utility and differential utility. The interpolation parameter  $\alpha$  adjusts the balance of future rewards and immediate reward in the consumer's calculations. The linear interpolation form of matchimizing is potentially useful for reformulating microeconomic theory to account for deviations from "rational behavior," with a parameter  $\alpha$  that serves as a "rationality index."

The interpolation form of matchimizing can also be written as

$$V_i = \lambda \pi_i + \mu$$

where the *subjective value* of good  $i$  is a linear interpolation between average utility and differential utility,

$$V_i = (1 - \alpha) \frac{u_i}{p_i} + \alpha \frac{\partial u}{\partial p_i}$$

The marginalist revolution in economics was the intellectual movement that established the idea that subjective value is proportional to marginal utility. We see here that the interpolation form of matchimizing can be interpreted as a new theory of subjective value.

To better understand the underpinnings of this new theory, it is helpful to derive a formula for the difference between the differential utility and average utility. The utility rate  $u$  can be written as the sum of components  $\sum_j u_j$ , or

$$u = \sum_j p_j \frac{u_j}{p_j}$$

Taking the derivative and applying the product rule yields

$$(23) \quad \frac{\partial u}{\partial p_i} = \frac{u_i}{p_i} + \sum_j p_j \frac{\partial}{\partial p_i} \left( \frac{u_j}{p_j} \right)$$

This means that the differential utility is equal to the average utility, plus a term that depends on the change in the average utilities of all the goods.

To interpret the second term, let's first consider the case of a single good, and a consumer who habitually purchases it at some rate  $p$ . The utility derived from the good is assumed to be the sum of hedonic rewards over time, as sketched in Fig. 8. If averaged over a long time period, utility is experienced at a rate  $u$ . Over short time periods like the one shown, temporal structure is evident. Consumption events are separated by time intervals of length  $1/p$ . The hedonic reward from one consumption event is  $u/p$ , represented here by the area of the shaded rectangle, which is consistent with an average utility rate of  $u$ . The sensitivity of the utility rate to small changes in consumption, or differential utility

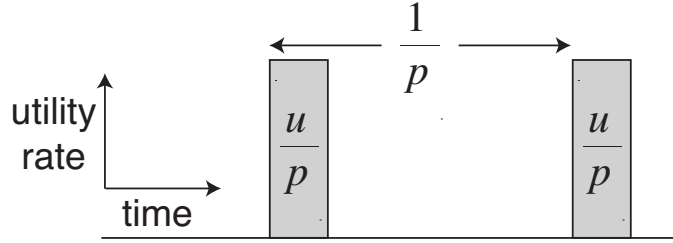


FIGURE 8. A temporal stream of hedonic rewards.

as we call it, is equal to the average utility plus a residual term,

$$(24) \quad \frac{du}{dp} = \frac{u}{p} + p \frac{d}{dp} \left( \frac{u}{p} \right)$$

This is because an increase in the rate of consumption affects the utility rate in two ways. The primary effect is an increase in the number of units consumed, which leads to the first term. The secondary effect is a change in the hedonic reward per consumption event, which gives rise to the second term in the equation. If  $u$  is linear in  $p$ , i.e., if the hedonic reward per consumption event is independent of consumption rate, the residual second term vanishes, and the differential and average utilities are equal. In general, however, one expects the hedonic reward to be a function of  $p$ . When hedonic reward per consumption event diminishes with consumption rate, the residual term in the equation above is negative, which also means that the differential utility is less than the average utility.

The psychological assumption underlying matchimizing is that the first and second terms of Eq. (24) are quite different in terms of their mental accessibility. The ratio  $u/p$  is simply the hedonic reward per consumption event, and is easily accessible, because it is attached to a single consumption event. In contrast, the second term measures how the hedonic reward per unit will decrease if the consumption rate changes. To appreciate this term, a consumer would need to understand exactly the interaction between two consumption events separated in time.

Using Eq. (23), subjective value can be written as

$$(25) \quad V_i = \frac{u_i}{p_i} + \alpha \sum_j p_j \frac{\partial}{\partial p_i} \left( \frac{u_j}{p_j} \right)$$

The parameter  $\alpha$  quantifies how much the consumer can take into account the dependence of hedonic rewards on consumption rate. If  $\alpha = 1$ , the consumer behaves rationally according to utility theory, in that the subjective value of a good is given by the differential utility. But  $\alpha < 1$  creates an ‘‘internality,’’ [12], which is to say, an unaccounted causal relationship between choices and total reward.

The presence of the internality is a necessary but not a sufficient condition for suboptimal behavior. If the second term in Eq. (25) is the same across all goods, then the first order conditions for matching and maximizing coincide, and neglecting the second term has no effect. If, however, the second term varies greatly across goods (and especially if it changes sign), then matchimizing can lead to grossly suboptimal choices. In particular, it can promote the worst of all possible choice distributions, as shown by Figure 2.

## APPENDIX A. MARKOV DECISION PROCESS

In the main text, we introduced the model (1) of history dependent rewards. This model had the limitation that reward depends on choices up to  $W$  time steps in the past, but no further. More generally, one can model the rewards as arising from a Markov decision process (MDP). The state of the MDP at time  $t$  is  $\mathbf{S}(t)$ . This state variable contains information about the past, and allows history-dependence. The action  $\mathbf{A}(t)$  controls the probability of transition from state  $\mathbf{S}(t)$  to state  $\mathbf{S}(t+1)$ .

$$\dots \rightarrow \mathbf{S}(t-1) \xrightarrow{\mathbf{A}(t-1)} \mathbf{S}(t) \xrightarrow{\mathbf{A}(t)} \mathbf{S}(t+1) \rightarrow \dots$$

The reward received at time  $t$  is a function of the state and the action,  $r(\mathbf{S}(t), \mathbf{A}(t))$ , and is written for short as  $R(t)$ . The model (1) is a special case of an MDP where the state variable  $\mathbf{S}(t)$  consists of the  $W$  past choices  $\mathbf{A}(t-1), \dots, \mathbf{A}(t-W)$ .

If the actions of an MDP are chosen i.i.d., it becomes an ordinary Markov process. Under reasonable assumptions, the probability distribution of  $\mathbf{S}(t)$  becomes independent of the initial condition  $\mathbf{S}(1)$  as  $t \rightarrow \infty$ . In other words, the Markov process “forgets” its initial condition, converging to a unique stationary distribution. In that case, the Markov process is said to be *ergodic*. Using methods described in [19], the mathematical results derived using the simple model (1) can be generalized to the more complex MDP model provided that ergodicity can be assumed. This assumption is useful, because it implies that time averaging is equivalent to averaging over the stationary distribution.

## APPENDIX B. ALLOWING ZERO CONSUMPTION RATES

In the main text, the necessary condition (5) was for the special case of a utility maximum at which all consumption rates are positive. To allow for the case where some consumption rates are zero, one can use the Karush-Kuhn-Tucker conditions

$$p_i \geq 0 \text{ and } \frac{\partial u}{\partial p_i} = \lambda \pi_i + \mu$$

or

$$p_i = 0 \text{ and } \frac{\partial u}{\partial p_i} \leq \lambda \pi_i + \mu$$

for all  $i$ . In words, the differential utility is linear in price for every good with nonzero consumption, and less for all goods with zero consumption.

Similarly, one can generalize the matching law to include zero consumption rates.

$$p_i \geq 0 \text{ and } \frac{u_i}{p_i} = \lambda \pi_i + \mu$$

or

$$p_i = 0 \text{ and } \frac{u_i}{p_i} \leq \lambda \pi_i + \mu$$

for all  $i$ . The average utility is linear in price for every good with nonzero consumption, and less for all goods with zero consumption.

## APPENDIX C. REWARD SCHEDULE BASED ON PAST CHOICE FREQUENCIES

In the paradigm introduced by Herrnstein, Prelec, and Vaughan [11], reward at time  $t$  is a function of the present choice  $\mathbf{A}(t)$ , as well as the frequency of choices in the past  $W$  time steps

$$\hat{p}_i(t) = \frac{1}{W} \sum_{\tau=1}^W A_i(t-\tau)$$

This reward schedule is a special case of (1), and was later utilized by Montague and collaborators [9, 26]. In particular, we will consider the case of two alternatives (“1” and “2”), and a linear function of the choice frequencies,  $R(t) = A_1(t)[a_1 - b_1\hat{p}_1(t)] + A_2(t)[a_2 - b_2\hat{p}_2(t)]$ . The reward function can also be split into two cases,

$$R(t) = \begin{cases} a_1 - b_1\hat{p}_1, & \text{if } A_1(t) = 1, \\ a_2 - b_2\hat{p}_2, & \text{if } A_2(t) = 1. \end{cases}$$

If the subject chooses between alternatives 1 and 2 using i.i.d. draws from the probability distribution  $(p_1, p_2)$ , then the conditional averages of reward are the same linear functions of  $p_1$ ,

$$\begin{aligned} \mathbf{E}[R(t)|A_1(t) = 1] &= a_1 - b_1p_1 \\ \mathbf{E}[R(t)|A_2(t) = 1] &= a_2 - b_2p_2 \end{aligned}$$

Therefore, this experimental paradigm maps nicely onto our mathematical formalism, since the frequency  $\hat{p}_1$  and the probability  $p_1$  become interchangeable. The lines are graphed in Figure 1, for the case where  $b_1 > 0$  and  $b_2 > 0$ . The dashed lines in the Figure are

$$\begin{aligned} \mathbf{E}[R(t)A_1(t)] &= p_1(a_1 - b_1p_1) \\ \mathbf{E}[R(t)A_2(t)] &= p_2(a_2 - b_2p_2) \end{aligned}$$

The sum of these two quantities is the total utility rate

$$\mathbf{E}[R(t)] = a_1p_1 + a_2p_2 - b_1p_1^2 - b_2p_2^2$$

drawn in red. The maximum of the total utility rate is attained at

$$p_{1,max} = \frac{b_2 + (a_1 - a_2)/2}{b_1 + b_2}$$

while the matching point (intersection of the blue and green lines) is

$$p_{1,match} = \frac{b_2 + a_1 - a_2}{b_1 + b_2}$$

Matching and maximizing are the same if  $a_1 = a_2$ , a special case called “matching shoulders,” but generically they are distinct. Some experiments with human subjects have used the “matching shoulders” case where matching and maximizing are the same [9, 26].

#### APPENDIX D. TEMPORALLY DISCOUNTED REWARD

The temporal accounting function  $w_\tau$  in Eq. (15) is reminiscent of the temporal discount function used by economists. The goal of this section is to show that matchimizing can indeed be formulated in terms of temporally discounted reward, but this does *not* mean that a matchimizer is a maximizer of temporally discounted reward. If the order of the expectation and the weighted sum is reversed, the matchimizing condition (15) takes the form

$$(26) \quad \mathbf{E}[D(t)|A_i(t) = 1] = \lambda\pi_i + \mu$$

where the random variable

$$(27) \quad D(t) = R(t) + \sum_{\tau=1}^W w_\tau R(t + \tau)$$

is the *temporally discounted reward*. In economics, it is standard to choose an exponential discount function,  $w_\tau = \gamma^\tau$  and allow an infinite time horizon  $W \rightarrow \infty$ . It has also been proposed that hyperbolic discount functions lead to a better description of human behavior [18].

Superficially, it seems that temporal discounting should have a strong effect on a utility-maximizing consumer. But surprisingly, the expectation values of the discounted reward and the reward are the same, up to a proportionality constant<sup>16</sup>

$$\mathbf{E}[D(t)] = \left(1 + \sum_{\tau=1}^W w_\tau\right) \mathbf{E}[R(t)]$$

This means that maximizing the average of temporally discounted reward is no different from maximizing the average of reward.

Therefore, a matchimizer is generically distinct from a maximizer of temporally discounted reward. However, matchimizing can be explained as a kind of “erroneous” maximization using the following theorem.

**Theorem 5.** *Suppose that the choice  $\mathbf{A}(t)$  is drawn from the probability distribution  $\mathbf{p}$ , while all other choices are drawn from the distribution  $\mathbf{p}'$ . Now define the discounted utility rate  $d(\mathbf{p}, \mathbf{p}') = \mathbf{E}_{\mathbf{p}, \mathbf{p}'}[D(t)]$ . Then the matchimizing condition (15) is equivalent to*

$$\left. \frac{\partial d}{\partial p_i} \right|_{\mathbf{p}'=\mathbf{p}} = \lambda\pi_i + \mu$$

*Proof.* By the same arguments as in Theorem 1,

$$\frac{\partial d}{\partial p_i} = \mathbf{E} \left[ D(t) \frac{A_i(t)}{p_i} \right] = \mathbf{E}[D(t) | A_i(t) = 1]$$

Substituting Eq. (27) yields the matchimizing condition (26). □

According to the Theorem, a matchimizer tries to maximize discounted reward with respect to consumption rates, but only calculates the derivative of the objective function with respect to the probabilities of the current choice. A matchimizer makes the error of neglecting the dependence of past and future choices on the probabilities.

## REFERENCES

- [1] H. Schultz. Interrelations of Demand. *The Journal of Political Economy*, 41(4):468–512, 1933.
- [2] AP Barten. Evidence on the Slutsky Conditions for Demand Equations. *The Review of Economics and Statistics*, 49(1):77–84, 1967.
- [3] HA Keuzenkamp and AP Barten. Rejection without falsification On the history of testing the homogeneity condition in the theory of consumer demand. *Journal of Econometrics*, 67(1):103–127, 1995.
- [4] H. Rachlin and D.I. Laibson, editors. *The matching law: papers in psychology and economics*. Harvard University Press, 1997.
- [5] RJ Herrnstein and GM Heyman. Is matching compatible with reinforcement maximization on concurrent variable interval variable ratio? *J Exp Anal Behav*, 31:209–223, 1979.
- [6] G.S. Becker and K.M. Murphy. A Theory of Rational Addiction. *The Journal of Political Economy*, 96(4):675–700, 1988.
- [7] R.J. Herrnstein and D. Prelec. A theory of addiction. In G. Loewenstein and J. Elster, editors, *Choice over time*. Russell Sage Press, 1992.
- [8] R.J. Herrnstein and D. Prelec. Melioration: A Theory of Distributed Choice. *The Journal of Economic Perspectives*, 5(3):137–156, 1991.

---

<sup>16</sup>Only when conditioned on choices, as in Eq. (26), does the expectation value of  $D(t)$  become different from the expectation value of  $R(t)$ .

- [9] D.M. Egelman, C. Person, and P.R. Montague. A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience*, 10(5):623–630, 1998.
- [10] R J Herrnstein. Relative and absolute strength of response as a function of frequency of reinforcement. *J Exp Anal Behav*, 4:267–272, Jul 1961.
- [11] RJ Herrnstein, D. Prelec, and W. Vaughan, Jr. An intra-personal prisoners’ dilemma. In *IX Symposium on the Quantitative Analysis of Behavior*, Harvard University, 1986.
- [12] R.J. Herrnstein, G.F. Loewenstein, D. Prelec, and W. Vaughan Jr. Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making*, 6(3):149–85, 1993.
- [13] J.R. Gray. A Bias Toward Short-Term Thinking in Threat-Related Negative Emotional States. *Personality and Social Psychology Bulletin*, 25(1):65, 1999.
- [14] D. Prelec. Matching, maximizing, and the hyperbolic reinforcement feedback function. *Psychological Review*, 84:189–230, 1982.
- [15] G. Heyman and RJ Herrnstein. More on concurrent interval-ratio schedules: A replication and review. *Journal of the Experimental Analysis of Behavior*, 46:331–51, 1986.
- [16] S.A. Vyse and T.W. Belke. Maximizing versus matching on concurrent variable-interval schedules. *J Exp Anal Behav*, 58(2):325–334, 1992.
- [17] HI Savastano and E. Fantino. Human choice in concurrent ratio-interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior*, 61:453–463, 1994.
- [18] D. Laibson. Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics*, 112(2):443–477, 1997.
- [19] J. Baxter and P.L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15(4):319–350, 2001.
- [20] J.G. Cross. A Stochastic Learning Model of Economic Behavior. *The Quarterly Journal of Economics*, 87(2):239–266, 1973.
- [21] I. Erev and A.E. Roth. Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*, 88(4):848–881, 1998.
- [22] D. Fudenberg and D.K. Levine. *The theory of learning in games*. MIT Press, 1999.
- [23] K.S. Narendra and M.A.L. Thathachar. *Learning automata: an introduction*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1989.
- [24] T. Borgers and R. Sarin. Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.
- [25] K.J. Arrow. *Studies in linear and non-linear programming*. 1958.
- [26] J Li, S M McClure, B King-Casas, and P Read Montague. Policy adjustment in a dynamic economic game. *PLoS ONE*, 1, 2006.