Probability in the Engineering and Informational Sciences, 4, 1990, 413-430. Printed in the U.S.A.

SEQUENTIAL ASSIGNMENT MATCH PROCESSES WITH ARRIVALS OF CANDIDATES AND OFFERS

ISRAEL DAVID
Department of Mathematics and Computer Science
Ben-Gurion University of the Negev
Beer-Sheva 84105, Israel

URI YECHIALI
Department of Statistics
The Raymond and Beverley Sackler Faculty of Exact Sciences
Tel Aviv University
Tel Aviv 69978, Israel

An infinite random stream of ordered pairs arrives sequentially in discrete time. A pair consists of a "candidate" and an "offer," each of which is either of type I (with probability p) or of type II (with probability $q = 1 - p$). Offers are to be assigned to candidates, yielding a reward $R > 0$ if they match in type, or a smaller reward $0 \le r \le R$ if not. An arriving candidate resides in the system until it is assigned, whereas an arriving offer is either assigned immediately to one of the waiting candidates or lost forever. We show that the optimal long-term average reward is R, independent of the population proportion p and the "second prize" r, and that the optimal average reward policy is to assign only a match. Optimal policies for discounted and finite horizon models are also derived.

1. INTRODUCTION

In recent studies (see David and Yechiali [3,4]), we introduced the notion of sequential assignment match processes (SAMPs) in which offers arriving in a ran-

 0269-9648/90 $5.00 + .00

dom stream are to be sequentially assigned to a group of waiting candidates. Each candidate as well as each offer is characterized by a random vector of attributes drawn from a known discrete-valued joint probability distribution function. Upon assignment of an offer to a candidate, their corresponding vectors of attributes are matched and the higher the compatibility, the larger the reward realized by this assignment.

The initial motivation and still a model application for such processes is the problem of optimal donor-recipient assignment in live organ transplants. The decision of whether to transplant an organ (e.g., a kidney) that becomes available depends on the degree of histocompatibility between the donor (offer) and the recipient (candidate). One relevant criterion for compatibility is the match level in the so-called HL-A antigen system. Basically, one counts the number of antigens of the donor that are not possessed by the recipient, and with each match level a value is associated, such as the odds for successful operation. This value is the "reward" of assigning a given offer to a waiting candidate. An important aspect of the problem is that new candidates for transplant join the waiting list while a live organ that is not assigned in a short period of time becomes unusable. Further description of the problem may be found in David and Yechiali [2].

In the aforementioned study [2], the focus was on a *single* candidate. We considered an appropriate time-dependent stopping problem and derived optimal assignment policies under various assumptions on the arrival process and on the decay properties of the lifetime distribution of the candidate. In the following studies [3,4], we considered the case of *many*, but *fixed*, numbers of candidates competing for the randomly arriving offers. Optimal assignment policies were derived, maximizing the total expected (discounted) reward for various models—both in discrete and continuous time. However, the models in [3,4] dealt with a single attribute distribution, resulting in only two possible match levels, "good" or "bad" (in the live organ transplant application, this means considering only one component in the antigen vector). Within the same limit on the number of match levels, it is the purpose of the present work to incorporate an additional factor, namely, allowing for an infinite incoming stream of candidates, alongside a parallel stream of offers. This extension brings the analysis one step closer to reality. Still, the model is somewhat restricted mainly because of the assumption to be made that candidates and offers arrive in pairs.

The model definition is thus the following: A joint incoming stream consisting of independent ordered pairs (offer, candidate) arrives sequentially in discrete time. The candidate and offer are, independently of each other, either of type I with probability p, or of type II with the complementary probability q. Offers are to be assigned to candidates, yielding a reward $R > 0$, if they match in type, or a smaller reward $0 \leq r \leq R$ if not. An arriving candidate resides in the system until it is assigned, whereas an arriving offer is either assigned immediately to one of the waiting candidates or lost forever.

The presentation becomes easier if we visualize the two types of attributes

as two different colors, say, blue and white. The decision maker earns a bigger reward for matching identical colors and a smaller reward for a bicolored assignment (blue offer to white candidate and vice versa). His objective is to maximize the expected total discounted, or long-run average, reward.

In Section 2, the average-reward problem is addressed. We show that the optimal long-run average-reward policy is to *assign only a match*, and that the optimal long-term average reward is R, *independent* of the population proportion p and the "second prize" r (as if each arrival results in a match). The optimal policy leads to an infinite-state space model, while for any $\epsilon > 0$, a finite-state ϵ near-optimal policy is also specified.

Next we study discounted models in finite and infinite horizon (Sections 3 and 4), where future rewards are discounted by a constant discount factor $0 < \alpha < 1$. As natural in some contexts of SAMPs, α may be thought of as the whole process's one-step survival probability. It is found that, for any $\alpha < 1$, the total discounted reward is maximized by a finite-state policy, whose control values are explicitly derived.

2. AVERAGE-REWARD CRITERION

In this section, our objective is to find an assignment policy that maximizes the long-run expected average reward per unit time. For any policy π, we denote by $r_\pi(n)$ the reward earned in stage (day) n. This is either 0 (rejection), r (assigning a mismatch), or R (assigning a perfect match). We let

$$\phi_\pi(s) = \liminf_{t\to\infty} \frac{E\left[\sum_{n=0}^{t-1} r_\pi(n) | \text{initial state} = s\right]}{t}, \tag{1}$$

calling a policy *average-reward optimal* if it maximizes Eq. (1) for all states s.

States are represented by ordered pairs (i,j) denoting i white and j blue candidates waiting in the system ($0 \le i,j < \infty$). It is well known that for infinite-state average-reward models, optimal policies need not exist (see Ross [6]). However, in our case, we may conjecture that a stationary optimal policy does exist, giving a *gain* g, $r \le g \le R$, independent of the initial state. We may further expect our hypothetic optimal π to be *reasonable*, where a reasonable policy is defined by the following: (i) it assigns a match whenever possible, and (ii) if it assigns a mismatch when n_1 candidates (either all white or all blue) are present just prior to arrival, then it also does so for any larger number of such candidates $n > n_1$. (Although it might seem at first glance that there could be two different thresholds n_1 and n_2 for the two types of candidates, we show in the sequel that symmetry in the two types of attributes holds for all values of p.) A reasonable policy is of *order* k if k is the smallest number n_1 specified in (ii). Such a policy is denoted by π_k. The policy π_k induces a finite-state Markov reward process, as described by Howard [5].

For any state and every stage, there are four possible joint arrivals WW, WB, BW, and BB, where W stands for white, B stands for blue, the first let-

ter indicates the offer, and the second letter specifies the candidate. The corresponding probabilities of these events are pp, pq, qp, and qq, respectively. A few examples of one-step transitions and rewards for the resulting process under π_k are given below:

Present state	Arrival	Probability	Reward	Next state
(1,2)	WW	pp	R	(1,2)
(1,2)	WB	pq	R	(0,3)
(1,2)	BW	qp	R	(2,1)
(1,2)	BB	qq	R	(1,2)
(0,2)	WW	pp	R	(0,2)
(0,2)	WB	pq	0	(0,3)
(0,2)	BW	qp	R	(1,1)
(0,2)	BB	qq	R	(0,2)
$(0,k)$	WW	pp	R	$(0,k)$
$(0,k)$	WB	pq	r	$(0,k)$
$(0,k)$	BW	qp	R	$(1,k-1)$
$(0,k)$	BB	qq	R	$(0,k)$

Remark: It is evident that with no loss of generality, R could be taken as 1. Nevertheless, for the sake of clarity, we retain the R,r notation.

By grouping probabilities of arrivals leading to identical transitions, the process may be depicted as in Figure 1. The diagram illustrates that in the corresponding Markov chain, all states are *transient* except for those on the main diagonal. The diagonal states form an ergodic class with $k+1$ states $(i,k-i)$ and stationary limiting probabilities $\pi(i,k-i) \equiv \pi(i)$, $0 \le i \le k$. The transition probability matrix P for the ergodic class may now be deduced from Figure 1:

$$P = \begin{bmatrix} 1-pq & pq & 0 & & & \\ pq & 1-2pq & pq & & 0 & \\ 0 & pq & 1-2pq & pq & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & pq & 1-2pq & pq \\ & & & 0 & pq & 1-pq \end{bmatrix}.$$

P is a *doubly stochastic* matrix. As such, $\boldsymbol{\pi}P = \boldsymbol{\pi}$ implies $\boldsymbol{\pi} = (c,c,\ldots,c)$, and $\sum_{i=0}^{k} \pi(i) = 1$ implies $c = 1/(k+1)$. We conclude that under π_k, the system

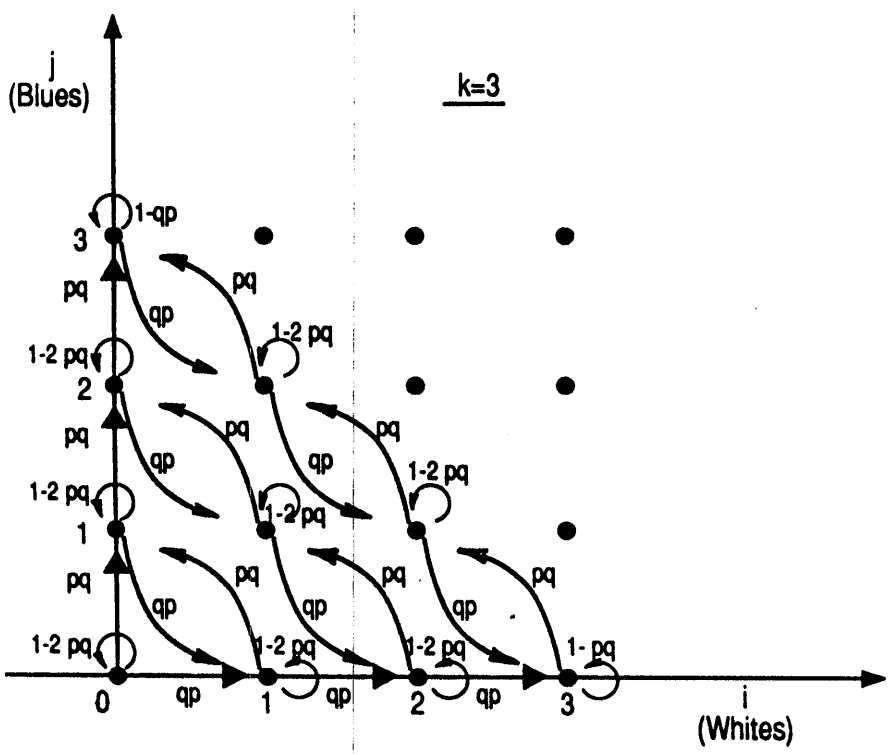

FIGURE 1. Flow diagram for the Markov process under π_k.

reaches an equilibrium in which there are always k waiting candidates, and all $k+1$ possible combinations of blues and whites are equally likely. If we let q_i be the immediate expected reward from state i, we have

$$q_0 = q_k = (1 - pq)R + pq \cdot r = R - pq(R - r)$$
$$q_i = R, \qquad 0 < i < k$$

and hence,

$$g = \sum_{i=0}^{k} \pi(i) q_i = \frac{2}{k+1} q_0 + \frac{k-1}{k+1} \cdot R = R - \frac{2pq}{k+1} (R - r) \tag{2}$$

is the average gain of the stationary reasonable policy of order k.

Equation (2) leads to an interesting result: By definition, the limiting policy of the π_k's when k tends to infinity, call it π^*, assigns only a perfect match. This is clearly a reasonable policy of order ∞. As will be justified formally at the end of this section, its gain is the limit in k of Eq. (2), namely, R. Hence, no matter what the mismatch reward and population proportion are, the achiev-

able gain is as if there were a perfect match every day. We write this conclusion as the following.

THEOREM 1: *For any* $0 \le p \le 1$ *and* $0 \le r \le R$, *the average reward for the matching problem is maximized by the persistent stationary policy that assigns only perfect matches. The optimal gain is the perfect-match reward,* R.

The optimal policy π^* induces a two-dimensional random walk on the first quadrant (see Figure 1). All states here are transient, and the system expands forever. One might speculate that, under π^*, a moment arrives after which only perfect matchings take place (many candidates of each type are present). This is wrong. On the contrary, with probability 1 we shall face a mismatch (a point on one of the axes in Figure 1) sometime in the future. π^* treats such an event with persistency, namely, enlarging the number of waiting candidates. These events become rarer and rarer, probabilistically not effecting the ratio in Eq. (1).

Theorem 1 practically says that we may always ensure maximal gain, but at the expense of infinite patience and infinite resources (waiting place, information capacity, etc.). This raises the question, what if, as in reality, resources are limited to handling at most k candidates? The answer is given in the theorem below.

THEOREM 2: *For a finite-state space with at most* k *candidates allowed, the reasonable policy of order* k, π_k, *is average-reward optimal.*

PROOF: It should be noted that once on a diagonal (see Figure 1), one cannot move to a lower diagonal. Assume inductively that for any waiting capacity $n < k$, the optimal policy is the reasonable policy π_n (which is obvious for $n = 0$). By Eq. (2), the gain for π_n is increasing in n, and hence, π_k is better than π_n, i.e., the optimal policy stays on the diagonal with states $S_i \equiv (i, k-i)$. It is left to show that π_k is better than any other policy on the kth diagonal.

We use Howard's improvement routine (see Howard [5]). The value-determination equations are

$$g + V_i = q_i^{\pi_k} + \sum_{j=0}^{k} P_{ij}^{\pi_k} V_j, \qquad 0 \le i \le k.$$

These equations take the form

$$\begin{aligned}
g + V_0 &= q_0 + (1 - pq)V_0 + pqV_1 \\
g + V_1 &= R + pqV_0 + (1 - 2pq)V_1 + pqV_2 \\
&\vdots \\
g + V_{k-1} &= R + pqV_{k-2} + (1 - 2pq)V_{k-1} + pqV_k \\
g + V_k &= q_0 + pqV_{k-1} + (1 - pq)V_k.
\end{aligned} \tag{3}$$

By letting $h = (R - g)/pq$, we can write the inner $k - 1$ equations as

$$\begin{aligned} V_2 &= 2V_1 - V_0 - h \\ V_3 &= 2V_2 - V_1 - h \\ &\vdots \\ V_k &= 2V_{k-1} - V_{k-2} - h, \end{aligned}$$

admitting a symmetric solution for the relative values

$$V_i = \frac{i(k-i)}{2} h, \qquad 0 \le i \le k. \tag{4}$$

Substituting Eq. (4) in the first equation of (3) and using $g = R - pqh$, we obtain

$$h = \frac{2(R-r)}{k+1}, \tag{5}$$

as was formerly given in Eq. (2).

We now attempt a policy improvement. In Table 1 we illustrate the transitions and compute immediate expected rewards for several modes of action. Note that the term "action" has been used in connection with the *actual* rejection or assignment of an offer upon arrival of a pair. By "mode of action," we mean a local policy specifying *before* arrival the action to be taken in each of the four possible joint arrivals. Since for each state S_i $(0 < i < k)$ on the diagonal associated with π_k, only the actions of matching or mismatching are allowed, there are $2^4 = 16$ modes of action. For state S_0 (S_k) where white (blue) candidate assignment might not be possible, only 2^2 modes of action exist. The modes of action presented in Table 1 are denoted by R, the "reasonable" one; WR, the one that assigns offers only to white candidates; BR, the one that assigns offers only to blue candidates; and IR, unreasonable at all, the one that assigns a white candidate to a blue offer and vice versa.

Now, take $0 < i < k$. To show that π_k is optimal, it suffices to establish that, for each mode of action a,

$$g + V_i \ge q_i^a + P_{i,i-1}^a V_{i-1} + P_{ii}^a V_i + P_{i,i+1}^a V_{i+1}$$

where q_i^a and P_{ij}^a are the corresponding one-step expected-reward and transition probabilities, respectively. Substituting $g = R - pqh$ in the above inequality, using the fact that $P_{i,i-1}^a + P_{ii}^a + P_{i,i+1}^a = 1$, and utilizing Eqs. (4) and (5), we see that the above is equivalent to

$$q_i^a + \frac{R-r}{k+1}\{(2i-k)(P_{i,i-1}^a - P_{i,i+1}^a) + [P_{ii}^a - (1-2pq)]\} \le R. \tag{6}$$

For the reasonable policy, $P_{i,i-1}^a = P_{i,i+1}^a$, $P_{ii}^a = 1 - 2pq$, and $q_i^a = R$, so that Eq. (6) holds as an equality. Relation (6) also holds as an equality for all modes

Table 1. Four Possible Modes of Action at States S_i, $0 < i < k$

State	Mode of action	Arrival	Probability	Reward	Next state
S_i	R	WW	p^2	R	S_i
		WB	pq	R	S_{i-1}
		BW	qp	R	S_{i+1}
		BB	q^2	R	S_i
Transition probabilities	$P_{i,i-1} = pq$	$P_{ii} = 1 - 2pq$	$P_{i,i+1} = pq$	One-step expected reward	$q_i = R$
S_i	WR	WW	p^2	R	S_i
		WB	pq	R	S_{i-1}
		BW	qp	r	S_i
		BB	q^2	r	S_{i-1}
	$P_{i,i-1} = q$	$P_{ii} = p$	$P_{i,i+1} = 0$		$q_i = pR + qr$
S_i	BR	WW	p^2	r	S_{i+1}
		WB	pq	r	S_i
		BW	qp	R	S_{i+1}
		BB	q^2	R	S_i
	$P_{i,i-1} = 0$	$P_{ii} = q$	$P_{i,i+1} = p$		$q_i = qR + pr$
S_i	IR	WW	p^2	r	S_{i+1}
		WB	pq	r	S_i
		BW	qp	r	S_i
		BB	q^2	r	S_{i-1}
	$P_{i,i-1} = q^2$	$P_{ii} = 2pq$	$P_{i,i+1} = p^2$		$q_i = r$

of action if $r = R$. As the left-hand side of Eq. (6) is linear in r, it is only left to show for all a that (6) holds for $r = 0$. One may further use the fact that P_{ij}^a is independent of i, so that the LHS of Eq. (6) is maximized for $i = k$ when $P_{i,i-1}^a \geq P_{i,i+1}^a$ and for $i = 1$ when $P_{i,i-1}^a \leq P_{i,i+1}^a$. By proper substitution of the values of the transition probabilities and the q_i^a's (see, for example, Table 1), condition (6) holds.

For the "corner" states S_0 and S_k, similar but simpler calculations show that the reasonable policy π_k is optimal. ■

Remark: Theorem 2 shows that the optimal policy selects at each state the mode of action that results in the highest immediate expected reward, i.e., it is "myopic," or "one-stage look ahead."

As was previously claimed, the optimal policy π^* requires an *infinite-state space.* If we are content with ϵ near-optimality, a *finite* waiting room will suffice. Its minimal size is given below.

COROLLARY 1: *For any* $\epsilon > 0$, *the reasonable finite-state stationary policy* $\pi_{k(\epsilon)}$ *is* ϵ*-optimal, where*

$$k(\epsilon) = \frac{2pq}{\epsilon}(R - r) - 1.$$

PROOF: By Theorem 1, $g(\pi^*) = R$. The best one can achieve with a waiting room of capacity k is, by Theorem 2, the gain g of π_k given by Eq. (2). Now substitute $k(\epsilon)$ in Eq. (2) to acquire $g[\pi_{k(\epsilon)}] = R - \epsilon$. ■

We conclude our discussion of the average-reward criterion with a formal proof of Theorem 1.

PROOF OF THEOREM 1: Since R is the best gain a policy can earn, it is enough to show that the policy stated in the theorem, π^*, achieves this gain. For the random walk that this policy induces on the first quadrant, we first compute the expected time the process stays on any diagonal. Let T_i $(0 \le i \le k)$ be the expected time the process stays on diagonal k before jumping to diagonal $k + 1$, starting from state S_i. Then, the T_i's satisfy the recursive equations

$$\begin{aligned}
T_0 &= 1 + (1 - 2pq)T_0 + pqT_1 \\
T_1 &= 1 + pqT_0 + (1 - 2pq)T_1 + pqT_2 \\
&\vdots \\
T_{k-1} &= 1 + pqT_{k-2} + (1 - 2pq)T_{k-1} + pqT_k \\
T_k &= 1 + pqT_{k-1} + (1 - 2pq)T_k
\end{aligned}$$

(see Figure 1 again).

If we set $h = 1/pq$ and rearrange terms, the inner $k - 1$ equations of the above set become identical to those we wrote for the relative values V_i in the proof of Theorem 2. They have, therefore, the same solution up to an additive constant (as always in Howard's equations); call it $c(k)$. It follows [see Eq. (4)] that $T_i = c(k) + i(k - i)/2pq$. Substituting in the first and last equations for the T_i's, we find that $T_0 = T_k = c(k) = (k + 1)/2pq$. Thus, the expected time for jumping from a diagonal to the next one starting at either one of the two "corner" states, S_0 or S_k, is proportional to the number of states in the diagonal.

Now, let $N(t)$ be the process counting the number of these jumps in $(0, t)$. Then,

$$\phi_{\pi^*}(\cdot) = \liminf_{t \to \infty} \left\{ \frac{[t - EN(t)]}{t} \right\} R.$$

It suffices to show that $\lim_{t\to\infty} E[N(t)]/t = 0$. If we let $X_1, X_2, \ldots$ be the inter-occurrence times of the jumps, we have shown that $E[X_n] = cn$ (for some

$c > 0$), increasing without bound. The result then follows in view of the elementary renewal theorem applied to nonidentically distributed renewal times (see Cox [2]).

3. FINITE-HORIZON DISCOUNTED MODELS

We turn now to study discounted models. In this section, we analyze the finite-horizon case. States will be denoted by quadruplets (m,n,W,N) and (m,n,B,N), where m is the number of white candidates at hand (possibly including the one who has just arrived), n is the same for blues, W or B refers to the type of offer that has just arrived, and N is the number of stages to go (including the present one). It is found helpful to introduce an auxiliary notation of triplets (m,n,N), indicating the state of the system "a moment before the arrival of a pair." As usual, $V(s)$ denotes the optimal total expected discounted reward from state s onward, enabling us to use the same symbol V for both triplet and quadruplet states.

The optimality equations may now be written simultaneously as follows:

$$\begin{aligned} V(m,n,N) = {} & p^2 V(m+1,n,W,N) + q^2 V(m,n+1,B,N) \\ & + pqV(m,n+1,W,N) + qpV(m+1,n,B,N). \end{aligned} \tag{7}$$

[In the right-hand side of Eq. (7), the first term refers to the arrival of a white offer and a white candidate, and so forth. As before, when writing arrival probabilities, the left letter will refer to the offer and the right one to the candidate.]

$$\begin{aligned} V(m,n,W,N) = \max\{ & R + \alpha V(m-1,n,N-1), \\ & r + \alpha V(m,n-1,N-1), \alpha V(m,n,N-1)\}. \end{aligned} \tag{8}$$

(The first term refers to assigning the white offer to a white candidate, the second to assigning a mismatch, and the third to not assigning at all. It is assumed here that $m,n > 0$. If $m = 0$ or $n = 0$, the appropriate term is omitted from the equation.) An equation similar to (8) holds for $V(m,n,B,N)$.

As initial conditions, we have $V(m,n,0) = 0$ for all m and n. The function V is thus uniquely determined, and clearly, a (finite-stage) optimal strategy exists for any $0 \le \alpha \le 1$. The following list of properties pertains to the optimal reward V. They all have intuitive explanations and can be formally proved by the use of the optimality equations. We avoid the technicalities and mathematical inductions involved.

1. For any (m,n,N), each value function V appearing in Eqs. (7) and (8) is *symmetric* around $p = 1/2$, where it attains its minimum. This readily follows if one is willing to accept at this stage the intuitive fact that it is always optimal to assign a perfect match when possible. In such a case, combining Eqs. (7) and (8) results in

$$V(m,n,N) = (p^2 + q^2)[R + \alpha V(m,n,N-1)] + pq[V(m,n+1,W,N) + V(m+1,n,B,N)]$$

which is interchangeable in p and q. Thus, inductively, all value functions are symmetric in p.

2. There is a *symmetry in blue and white* for all values of p

$$V(n,m,N) = V(m,n,N),$$
$$V(n,m,W,N) = V(m,n,B,N).$$

The above follows from property 1, since interchanging p and q is equivalent to interchanging the colors of all members involved.

3. All V's are monotone-increasing in m,n,N,α,r, and R.
4. (In what follows, all "white" formulations have "blue" counterparts.)
   (a) $V(m,\cdot,W,N) = R + \alpha V(m-1,\cdot,N-1)$ for any $m \geq 1$.
   (b) If $V(0,n,W,N) = r + \alpha V(0,n-1,N-1)$ for $r = r^*$, then the same equality holds for any $r \geq r^*$.
   (c) If $V(0,n^*,W,N) = r + \alpha V(0,n^*-1,N-1)$, then the same equality holds for any $n \geq n^*$. Property 4(a) follows since it is optimal to assign a good match when available. 4(b) is true since if it is optimal to assign a mismatch and receive the second reward r^*, then it is all the same so when receiving a higher second reward, $r > r^*$. If it is optimal to assign a mismatch to one of n^* unicolored candidates, it means that there are "enough of them" for the future, and thus, it is also true for any $n \geq n^*$ [4(c)].

From the above properties we obtain the following.

THEOREM 3: *For a finite-horizon model, the optimal policy is of the following form: Always assign a perfect match when available. In states* $(0,n,W,N)$ *or* $(n,0,B,N)$, *assign according to a control-limit (threshold)* $r^*_{n,N}$, *i.e., assign a mismatch if and only if* $r > r^*_{n,N}$.

Naturally, $r^*_{n,N}$ is a function of R,p,α, and N. $r^*_{n,N} \uparrow N$, and $r^*_{n,N} \downarrow n$ [by property 4(c)].

The calculation of the control limits $r^*_{n,N}$ is a straightforward manipulation of Eqs. (7) and (8) and demonstrates the properties stated above. For example, starting with $V(0,0,0) = 0$, we obtain

$$V(0,0,1) = p^2R + q^2R + pqr + qpr = (1 - 2pq)R + 2pqr \equiv \xi_p. \quad (9)$$

ξ_p is the expected reward achieved by matching the individuals of a random pair. Clearly, ξ_p is a parabola, symmetric around its minimum at $p = 1/2$. Indeed, $p = q = 1/2$ means maximum heterogeneity in the population and, thus, the least prospect for well-matched pairs.

Table 2. Finite-Horizon Optimal Control Limits

n	N	$r^*_{n,N}$
$n=1$	$N\geq 2$	$\dfrac{pq\alpha[1+\alpha+\ldots+\alpha^{N-2}]}{1+pq\alpha[1+\alpha+\ldots+\alpha^{N-2}]}\cdot R$
$n=2$	$N=3$	$\dfrac{(pq\alpha)^2}{1+(pq\alpha)^2}\cdot R$
$n=2$	$N=4$	$\dfrac{(pq\alpha)^2[1+\alpha(2-3pq)]}{1+(pq\alpha)^2[1+\alpha(2-3pq)]}\cdot R$
$n=2$	$N=5$	$\dfrac{(pq\alpha)^2\{1+\alpha(2-3pq)+\alpha^2[2-3pq+(1-3pq)^2]\}}{1+(pq\alpha)^2\{1+\alpha(2-3pq)+\alpha^2[2-3pq+(1-3pq)^2]\}}\cdot R$
Any	$N=n+1$	$\dfrac{(pq\alpha)^n}{1+(pq\alpha)^n}\cdot R$

The method of direct computation becomes laborious and no general formula emerges for the $r^*_{n,N}$'s. Still, we present some results in Table 2.

We conclude this section with a few remarks:

1. Again, R emerges naturally just as a scale factor in the expressions for the controls $r^*_{n,N}$.
2. As far as optimal policies are concerned, the probabilities p and q always appear in product form as pq. Thus, one may replace this product with a single symbol, signifying (half) the mismatch probability, relaxing the independence assumption between arrivals of offers and candidates.
3. For finite-horizon models, the total *undiscounted* reward is also meaningful. For these cases, Table 2 supplies some specific results, e.g.,

$$r^*_{1,N}=\frac{(N-1)pq}{1+(N-1)pq}R.$$

4. INFINITE-HORIZON DISCOUNTED MODELS

We consider now the infinite-horizon discounted case. We use the same scheme of states and the V notation, where the horizon N is omitted. The V's relate to an optimal policy π that certainly exists for a denumerable-state-discounted dynamic programming model. Furthermore, the value functions $V(\cdot,\cdot,\cdot,N)$ are successive approximations of the expected discounted total reward function V^π, where the convergence is uniform in the space of states (see Ross [6]). $V^\pi=V$ inherits all qualitative properties discussed in the previous section, e.g.,

a perfect match is always assigned and there is symmetry between the blue and white colors. Naturally, the states of interest are $(0, k, W)$ or $(k,0,B)$ where a decision must be made whether to assign a mismatch or to reject the offer. We have

$$V(0,k,W) = \max\{r + \alpha V(0,k-1), \alpha V(0,k)\} \tag{10}$$

for any $k \geq 1$. As in the finite-horizon case, the max in Eq. (10) is chosen according to a control limit r_k^* on r. That is, in state $(0,k,W)$, one optimally assigns a mismatch if and only if $r > r_k^*$. Also, by property 4(c) in Section 3, the controls are monotone-decreasing in k. We summarize in the theorem below.

THEOREM 4: *The infinite-horizon discounted-reward optimal policy is of the following structure*:

(i) *Always assign a perfect match.*

(ii) *Assign a mismatch according to a set of controls*

$$r_1^* \geq r_2^* \geq \cdots \geq r_{k-1}^* \geq r_k^* \geq \cdots$$

on r and according to k, the number of mismatching candidates.

We may expect that $\lim_{k\to\infty} r_k^* = 0$ because with an infinite number of candidates, it would be optimal to "give up" one and obtain any positive second reward r. Indeed, by monotonicity and boundedness (by R), $V(0,k)$ converges in k. So, for any $r = \epsilon > 0$, $\lim_{k\to\infty} V(0,k) = \lim_{k\to\infty} V(0,k-1)$, implying that $r \geq \alpha[V(0,k) - V(0,k-1)]$ for some k, or by Eq. (10), $r_k^* < \epsilon$.

We wish now to find the explicit values of the controls r_k^*'s, thus completely determining the optimal assignment policy. One approach is to apply a limiting procedure to the finite-horizon controls. For example, by using Table 2, we find

$$r_1^* = \lim_{N\to\infty} r_{1,N}^* = \frac{pq\alpha[1/(1-\alpha)]\cdot R}{1 + pq\alpha[1/(1-\alpha)]} = \frac{pq\alpha R}{1 - \alpha(1-pq)}. \tag{11}$$

However, since the task of completing Table 2 is not an appealing direction, we proceed in a different way.

4.1 Back to the Reasonable Policies

From Theorem 4, it follows that the optimal policy is *reasonable* of some order k^*. Recall that we defined a reasonable policy as one that assigns a perfect match whenever available and satisfies property 4(c) in the previous section. If $k+1$ is the least number of candidates (including the arriving one) among which we assign a mismatching offer, then that policy, denoted by π_k, is said to be "*reasonable of order k*." Clearly, under π_k, exactly k candidates reside in the system in a steady state, if we assume that initially there are no more than k can-

didates. Thus, π_k leads to a finite-state Markov chain. We have (see Ross [6], Proposition 2.4)

$$V^{\pi_k}(0,k+1,W) = r + \alpha V^{\pi_k}(0,k). \tag{12}$$

Now, for $k \geq 1$ and $i = 0$,

$$\begin{aligned} V^{\pi_k}(0,k) &= (1-2pq)\,[R + \alpha V^{\pi_k}(0,k)] + pqV^{\pi_k}(0,k+1,W) \\ &\quad + qpV^{\pi_k}(1,k,B) \\ &= (1-2pq)\,[R + \alpha V^{\pi_k}(0,k)] + pq[r + \alpha V^{\pi_k}(0,k)] \\ &\quad + qp[R + \alpha V^{\pi_k}(1,k-1)]. \end{aligned} \tag{13}$$

For $1 \leq i < k$,

$$\begin{aligned} V^{\pi_k}(i,k-i) &= (1-2pq)\,[R + \alpha V^{\pi_k}(i,k-i)] \\ &\quad + pq[R + \alpha V^{\pi_k}(i-1,k-i+1)] \\ &\quad + qp[R + \alpha V^{\pi_k}(i+1,k-i-1)], \end{aligned} \tag{14}$$

and finally, for $i = k$,

$$\begin{aligned} V^{\pi_k}(k,0) &= (1-2pq)\,[R + \alpha V^{\pi_k}(k,0)] + pq[R + \alpha V^{\pi_k}(k-1,1)] \\ &\quad + qp[r + \alpha V^{\pi_k}(k,0)]. \end{aligned} \tag{15}$$

In case $k = 0$, we get a single equation,

$$\begin{aligned} V^{\pi_0}(0,0) &= (1-2pq)\,[R + \alpha V^{\pi_0}(0,0)] + pq[r + \alpha V^{\pi_0}(0,0)] \\ &\quad + qp[r + \alpha V^{\pi_0}(0,0)] \\ &= \xi_p + \alpha V^{\pi_0}(0,0), \end{aligned}$$

yielding

$$V^{\pi_0}(0,0) = \frac{\xi_p}{(1-\alpha)}. \tag{16}$$

Indeed, if we start at (0,0), π_0 always matches the randomly coming pair, so its total expected discounted reward is $\xi_p + \alpha\xi_p + \alpha^2\xi_p + \cdots = \xi_p/(1-\alpha)$. Using Eqs. (12) and (16), we write

$$\begin{aligned} V^{\pi_0}(0,1,W) &= r + \alpha V^{\pi_0}(0,0) = r + \frac{\alpha\xi_p}{(1-\alpha)} \\ &= \frac{(1-2pq)\alpha}{1-\alpha} R + \frac{1-(1-2pq)\alpha}{1-\alpha} r \end{aligned} \tag{17}$$

which is *linear* in r.

The set of equations (13), (14), and (15) may be written in matrix form

$$A\mathbf{x} = \mathbf{b}(r) \tag{18}$$

where $x_i = V^{\pi_k}(i, k-i)$, $0 \le i \le k$,

$$A = \begin{bmatrix} 1-(1-pq)\alpha & -pq\alpha & & & & & \\ -pq\alpha & 1-(1-2pq)\alpha & -pq\alpha & & & & \\ & -pq\alpha & 1-(1-2pq)\alpha & \ddots & & & \\ & & \ddots & \ddots & -pq\alpha & & \\ & & & -pq\alpha & 1-(1-2pq)\alpha & -pq\alpha & \\ & & & & -pq\alpha & 1-(1-2pq)\alpha & -pq\alpha \\ & & & & & -pq\alpha & 1-(1-pq)\alpha \end{bmatrix},$$

and $\mathbf{b}(r)$ is a vector function of r

$$\mathbf{b}(r) \equiv \mathbf{b} = \begin{bmatrix} R - pq(R-r) \\ R \\ \vdots \\ R \\ R - pq(R-r) \end{bmatrix}.$$

One can see that (except for $p = \alpha = 1$, or $q = \alpha = 1$) A is a regular matrix. Furthermore, a symmetry [well expected by the interpretation of the x_i's as $V^{\pi_k}(i, k-i)$] holds: $x_0 = x_k$, $x_1 = x_{k-1}$, etc. To see this formally, consider the matrix

$$J = \begin{bmatrix} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \\ & & & 0 \end{bmatrix}.$$

The operator J reverses a row or column vector, and $J^2 = I$. It is easy to see that $J\mathbf{b} = \mathbf{b}$ and $JAJ = A$. Thus, using Eq. (18), we have

$$A(J\mathbf{x}) = JAJ(J\mathbf{x}) = JAJ^2\mathbf{x} = JA\mathbf{x} = J\mathbf{b} = \mathbf{b},$$

and hence, $J\mathbf{x}$ is a solution of Eq. (18). Since A is regular, $J\mathbf{x} = \mathbf{x}$, i.e., $x_i = x_{k-i}$. The above symmetry reduces the number of unknowns to $1 + [k/2]$. Also, as r appears (linearly) only in the right-hand side of Eq. (18), each x_i is *linear in r*.

Now we develop an explicit expression for $V^{\pi_k}(0, k+1, W)$. Considering Eq. (12), we may write $V^{\pi_k}(0, k+1, W) = r + \alpha x_0(r)$. By the linearity of x_i (and in particular x_0) in r, it suffices to calculate the values of $V^{\pi_k}(0, k+1, W)$

at the two points $r = R$ and $r = 0$. Write the matrix A as $(\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)$ and write $A(r) = (\mathbf{b}, \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k)$. If $r = R$, then $\mathbf{b}^T = (R, R, \ldots, R)$. Also, by adding all the columns of A onto the first column, we get a matrix $\hat{A} = (\mathbf{1} - \boldsymbol{\alpha}, \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k)$ where $(\mathbf{1} - \boldsymbol{\alpha})^T = (1 - \alpha, 1 - \alpha, \ldots, 1 - \alpha)$. It follows that for $r = R$,

$$V^{\pi k}(0,k) = x_0(R) = \frac{\det A(R)}{\det A} = \frac{\det A(R)}{\det \hat{A}} = \frac{R}{1-\alpha}.$$

This result is natural since, for $r = R$, one always gets a reward R (either by a match or mismatch), and the total discounted reward is $R + \alpha R + \alpha^2 R + \cdots = R/(1-\alpha)$. It follows that $V^{\pi k}(0, k+1, W) = R + \alpha R/(1-\alpha) = R/(1-\alpha)$. For $r = 0$, $V^{\pi k}(0, k+1, W) = 0 + \alpha x_0(0) \equiv y_0^k$, where

$$y_0^k = \alpha \frac{\det A(0)}{\det A}. \tag{19}$$

To summarize, $V^{\pi k}(0, k+1, W)$ as a function of r displays a straight line stretching from the point $(0, y_0^k)$ to the point $[R, R/(1-\alpha)]$. (Note that the second point is common for all k.) We give specific results for $k = 0, 1, 2$.

For $k = 0$, we have from Eq. (17)

$$y_0^0 = \frac{(1-2pq)\alpha}{1-\alpha} \cdot R. \tag{20}$$

For $k = 1$, with

$$A = \begin{bmatrix} 1-(1-pq)\alpha & -pq\alpha \\ -pq\alpha & 1-(1-pq)\alpha \end{bmatrix} \quad \text{and} \quad \mathbf{b}(0) = \begin{bmatrix} R - pqR \\ R - pqR \end{bmatrix},$$

we immediately find $x_0 = x_1 = (1-pq)R/(1-\alpha)$, and thus,

$$y_0^1 = \frac{(1-pq)\alpha}{1-\alpha} R.$$

Stretching a line between $(0, y_0^1)$ and $[R, R/(1-\alpha)]$, we get

$$V^{\pi_1}(0,2,W) = \frac{(1-pq)\alpha}{1-\alpha} R + \frac{1-(1-pq)\alpha}{1-\alpha} r. \tag{21}$$

The case $k = 2$ is treated in a similar way, using the symmetry $x_0 = x_2$. For $r = 0$,

$$x_0 = \frac{(1-pq)(1-\alpha) + (3-2pq)pq\alpha}{(1-\alpha)[1-(1-3pq)\alpha]} R$$

which determines, as before, y_0^2 and $V^{\pi_2}(0,3,W)$.

4.2. Constructing Optimal π

Considering again the optimal value functions, one could envision a picture of the $V(0,k,W)$'s, put together under the same scale as functions of r for any k. This may be based on Theorem 4 and Eq. (10), while using the monotonicity of $V(0,k,W)$ in r and in k [see property (3) in Section 3]. Each function $V(0,k,W), k \geq 1$, is partitioned into two: To the right of r_k^*, it equals $r + \alpha V(0,k-1)$ [the first alternative to be taken in Eq. (10)], and to the left of r_k^*, it equals $\alpha V(0,k)$ [the second alternative in Eq. (10)]. By the nature of the π_k's and definition of the r_k^*'s, we conclude that to the right of r_k^*, $V(0,k,W) = V^{\pi_{k-1}}(0,k,W)$. That is, the respective parts of the $V(0,k,W)$'s ending at the point $[R, R/(1-\alpha)]$ are straight-line segments. The above observations suggest a method for determining the r_k^*'s: Take, for instance, $k = 1$. At r_1^*, two line segments of $V(0,1,W)$ intersect. To the right of r_1^* the line segment is $V^{\pi_0}(0,1,W)$. To the left of r_1^* it is $\alpha V(0,1)$, which, by considering $V(0,2,W)$ for $r_2^* \leq r \leq r_1^*$, may be written as

$$\alpha V(0,1) = V(0,2,W) - r = V^{\pi_1}(0,2,W) - r$$
$$= \frac{(1-pq)\alpha}{1-\alpha} R + \left[\frac{1-(1-pq)\alpha}{1-\alpha} - 1\right] r = \frac{(1-pq)\alpha}{1-\alpha} R + \frac{pq\alpha}{1-\alpha} r \tag{22}$$

[see Eq. (21)]. We now equate Eq. (22) with (17) to obtain

$$r_1^* = \frac{pq\alpha R}{1-\alpha(1-pq)},$$

exactly as anticipated in Eq. (11). Any other r_k^* is found independently in the same manner.

The general result is given in the next theorem.

THEOREM 5: *Let y_0^k be defined by Eqs.* (19) *and* (20) *for $k \geq 0$. Given a mismatch reward r, perfect-match reward R, type I population-proportion p, and discount factor α, the optimal total expected discounted reward policy is to assign a perfect match whenever possible, and to assign a mismatch if and only if*

$$r \geq r_k^*$$

where k is the number of candidates at hand, and

$$r_k^* = \frac{(y_0^k - y_0^{k-1})R}{R + (y_0^k - y_0^{k-1})}. \tag{23}$$

PROOF: By means of Theorem 4 and the previous discussion, for any $k \geq 1$,

$$V(0,k,W) = V^{\pi_{k-1}}(0,k,W) = y_0^{k-1} + \left(\frac{1}{1-\alpha} - \frac{y_0^{k-1}}{R}\right) r, \tag{24}$$

where the first equality holds for $r \geq r_k^*$, and the last expression is the one connecting the points $(0, y_0^{k-1})$ and $[R, R/(1-\alpha)]$. Similarly, for $r_{k+1}^* \leq r \leq r_k^*$, we get

$$V(0,k,W) = \alpha V(0,k) = V^{\pi_k}(0,k+1,W) - r = y_0^k + \left(\frac{1}{1-\alpha} - \frac{y_0^k}{R} - 1\right) r. \tag{25}$$

Equating Eqs. (24) and (25) gives the desired expression, Eq. (23), for r_k^*. ■

Practically, given the rewards (R,r) in a specific problem, the optimal waiting capacity k is determined by $k = \min\{n : r_n^* \leq r\}$. It is found that, although the undiscounted average-reward assignment problem is not maximized by *any* finite-state matching policy, the discounted problem is maximized for any $\alpha < 1$ by a finite-state reasonable policy π_k, where k is extracted from Eq. (23).

References

1. Cox, D.R. (1962). *Renewal theory.* London: Methuen.
2. David, I. & Yechiali, U. (1985). A time-dependent stopping problem with application to live-organ transplants. *Operations Research* 33: 491-504.
3. David, I. & Yechiali, U. (1989). Discrete-time finite-state sequential assignment match processes. Technical Report, Dept. of Statistics, Tel Aviv University.
4. David, I. & Yechiali, U. (1989). Continuous-time finite-state sequential assignment match processes. Technical Report, Dept. of Statistics, Tel Aviv University.
5. Howard, R.A. (1960). *Dynamic programming and Markov processes.* Cambridge, Mass.: MIT Press.
6. Ross, S.M. (1983). *Introduction to stochastic dynamic programming.* New York: Academic Press.