Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

A simple decision task in a social context: experiments, a model, and preliminary analyses of behavioral data

A. Nedic, D. Tomlin, P. Holmes, D.A. Prentice and J.D. Cohen

Abstract— **To investigate the influence of input from fellow group members in a constrained decision-making context, we consider a game in which subjects freely select one of two options (A or B), and are informed of the reward resulting from that choice following each trial. Rewards are computed based on the fraction x of past A choices by two functions $f_A(x)$, $f_B(x)$ (unknown to the subject) which intersect at a matching point $\bar{x}$ that does not generally represent globally optimal behavior. Playing individually, subjects typically remain close to the matching point, although some discover the optimum. We investigate the effects of additional feedback regarding the choices and reward scores of other players. We generalize a drift-diffusion model, commonly used to model individual decision making, to incorporate feedback from other players, study the resulting coupled stochastic differential equations, and compare the distributions of choices that they predict with those produced by a pool of subjects playing in groups of five without feedback and with feedback on other players' choices.**

I. INTRODUCTION

In an attempt to better understand and model collective decision making in small human groups, we have designed and are currently carrying out a highly constrained experiment that probes the manner in which limited input from group members influences individual choices. We have adapted and generalized an experimental paradigm of Montague et al. [1], [2] to a social context, allowing different types of feedback from group members to subjects playing a simple game: a "two-armed bandit" in which they select one of two alternatives on each trial, presumably basing their choices on the resulting rewards. We wish to understand how limited information regarding other players' rewards, or choices, or both, modifies individual behaviors.

In the game a deterministic rule, unknown to the subject, computes rewards based on his or her choice history, and, in the version of [1], [2], two different types of behavior are observed. A majority of subjects settle near a "matching point," at which both choices result in the same (moderate) reward, while a minority of "explorers" endure runs of low rewards and discover a global optimum that is substantially better than the matching strategy. This type of rule permits examination of whether subjects exploit a particular strategy in the long term, or explore different ones in an attempt to maximize their rewards. The *explore vs. exploit* question, central to studies of foraging in animal communities [3], is of increasing interest in cognitive neuroscience [4].

To describe the social context, here we generalize a commonly-used drift-diffusion (DD) model, in which evidence in favor of one choice over the other (a logarithmic likelihood ratio [5]) is integrated until a predetermined threshold is reached. We show that the model not only fits the behaviors of subjects playing alone, but, equipped with inter-subject feedback, also those of a group exchanging limited information. The DD model, and extensions of it, have been fitted to accuracy and reaction time data in numerous two-alternative forced-choice tasks [6], [7], [8]. For a recent review and derivations of DD processes from other, more complex, neurally-based models, see [9].

In §II we describe the reinforcement learning model of [1], [2], some of the games, and our extension to the group context. §III presents preliminary analyses of data and matches to the model, and we conclude in §IV. A related article maps a robot foraging task onto the two-armed bandit game and further explores the exploitation vs. exploration dichotomy [10], [11].

II. COGNITIVE CONTROL AND CHOICE WITH AND WITHOUT GROUP FEEDBACK

Here we describe the model for individual choices, the games and reward schedules, and an extension of the model to include feedback regarding other subjects' choices.

A. A model for individual choices

The simplest version of a DD process is described by the following stochastic differential equation:

$$dx = \alpha\, dt + \sigma\, dW\,; \quad x(0) = 0, \tag{1}$$

where α denotes the drift rate and $\sigma\, dW$ increments drawn from a Wiener process with standard deviation σ. The state variable $x(t)$ represents the integrated evidence in favor of choice A over choice B. On each trial the choice A or B is made when $x(t)$ first crosses either of the predetermined thresholds $\pm x_{th}$. As we show in §II-B, rewards are determined by the proportion of choices of alternative A, which, for (1), is governed by probability of choosing A:

$$P(\mathrm{A}) = \frac{1}{1+\exp(2\eta\theta)}, \text{ where } \eta = (\alpha/\sigma)^2,\ \theta = x_{th}/\alpha, \tag{2}$$

[12], [13], cf. [9]. Here η is the square of the signal-to-noise ratio (SNR, having the units of inverse time), and θ is the

This work was supported by the Air Force Office of Scientific Research under MURI 16.

A. Nedic is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. nedic@Princeton.EDU

D. Tomlin D.A. Prentice and J.D. Cohen are with the Department of Psychology, Princeton University, Princeton, NJ 08544, USA. dtomlin, predebb, jdc@Princeton.EDU

P. Holmes is with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA. pholmes@math.Princeton.EDU

threshold-to-drift ratio: i.e., the threshold passage time for the noise-free process $x(t) = \alpha t$. An explicit expression for the mean first passage time, corresponding to the average decision time, is also available [9], but we shall not need it here.

In the model developed by Egleman et al. [1], cf. [14] individual subjects maintain a fading memory of prior rewards in the expected rewards: weights w_A, w_B accorded to the alternatives that are updated as described below. The probability of choosing A is then determined by

$$P_1(\mathrm{A}) = \frac{1}{1 + \exp[-\mu(w_A - w_B)]}, \tag{3}$$

which is identical to (2) if we identify the steepness parameter μ with the ratio 2θ and the weight difference $w_A - w_B$ with η. The weight update rule, motivated by the role of dopamine neurons in coding for reward prediction error [15], [16] and by temporal difference learning theory [17], [18], proceeds as follows. If A is chosen on the nth trial, resulting in a reward r, the expected rewards are updated according to

$$w_A(n+1) = (1-\lambda)w_A(n) + \lambda r, \;\; w_B(n+1) = w_B(n); \tag{4}$$

if B is chosen, the roles of A and B in (4) are reversed. The *learning rate* λ determines the time scale on which the memory of previous choices decays: when $\lambda = 0$, no learning occurs; when $\lambda = 1$, memory of the chosen alternative is instantly erased.

In [14] a further time scale is added in the form of decaying eligibility traces (ETs) e_A (resp. e_B) that are incremented by 1 following a choice A (resp. B):

$$e_{A,B} \mapsto e_{A,B} \exp[-(t - t(n))/\tau], \tag{5}$$

where $t(n)$ denotes the time of the last update and τ is a further timescale. Now the weights are *both* updated by

$$\begin{aligned} w_A(n+1) &= w_A(n) + \lambda(r - w_*)e_A(n), \\ w_B(n+1) &= w_B(n) + \lambda(r - w_*)e_B(n), \end{aligned} \tag{6}$$

where $w_* = w_A$ (resp. w_B) if A (resp. B) was chosen. For large τ this rule converges to the simpler one above, since the ET for the choice not made decays immediately while the other is simply $e_*(n) = 1$. The model now has three parameters (μ, λ, τ). The reason for this modification is described in the next section.

B. Four gambling tasks

The two-armed bandit delivers rewards according to two schedules $f_A(x)$, $f_B(x)$ determined by the fraction $x \in [0, 1]$ of A choices (allocation to A) made over the past N trials. Clearly if f_A lies entirely above (or below) f_B the subject will rapidly deduce the better option and thereafter always choose A (or B); interesting cases occur when the curves cross at a *matching point(s)* $\bar{x}$, so called because the rewards are equal there.

At any point $x \in [0, 1]$ the average reward is

$$R(x) = x f_A(x) + (1 - x) f_b(x), \tag{7}$$

and rewards could clearly be maximized by following a hill-climbing algorithm expressible as a gradient dynamical system [19] with potential function $R(x)$:

$$\dot{x} = R'(x), \;\; \text{where } f' = df/dx, \tag{8}$$

since stable fixed points of (8) are (local) maxima of (7), see [2]. However, subjects cannot estimate the functions $f_{A,B}(x)$ from their trial-to-trial observations, especially if N is large; hence the more neurobiologically-plausible model of §II-A.

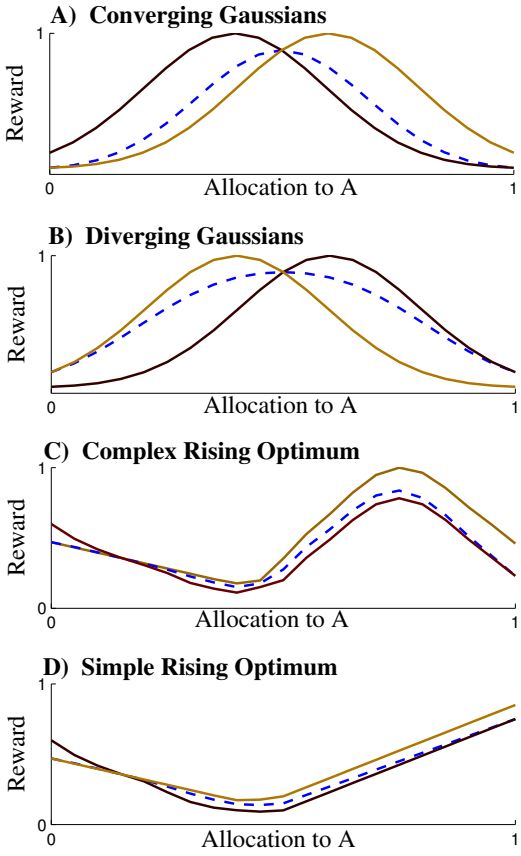

Fig. 1. Reward schedules f_A (black) and f_B (brown) for four of the tasks: (A) converging gaussians; (B) diverging gaussians; (C) complex rising optimum; (D) simple rising optimum. Dashed blue curves denote average rewards R(x).

We have developed six variants of the two-armed bandit game, some of which have unique optima while others have two. Four examples of the reward schedules are shown in Fig. 1. The converging and diverging gaussians each have unique global optima at 50% allocation to A, which are also their matching points. The third and fourth examples, modifications of the rising optimum task of [1], [2], have local maxima at 0% A's, near the matching point, *and* global optima at 75% and 100% respectively. (Reflections of the latter two, with local maxima at 100% A's, are also used.)

These games all generalize a "matching shoulders" task with linear schedules: Fig. 2. In this simpler case $f_A(x) = a_1 - b_1 x$ and $f_B(x) = a_2 + b_2 x$, with $a_j, b_j > 0$ and $b_1 + b_2 > a_1 - a_2 > 0$. The curves cross at $\bar{x} = (a_1 - a_2)/(b_1 + b_2)$ and the global maximum lies at $x_{\max} =$

$(a_1 - a_2 - b_2)/[2(b_1 + b_2)]$. This example shows that global (and even local) maxima need not lie at matching points. In cases where these points do not coincide, experiments have shown that under most circumstances human subjects tend to match rather than maximize [20], [21], [22]. In this study the allocation of A's is determined by averaging over the past $N = 20$ trials. A brief description of the experimental method is provided in Appendix V; further details and analyses will be reported in the psychological literature.

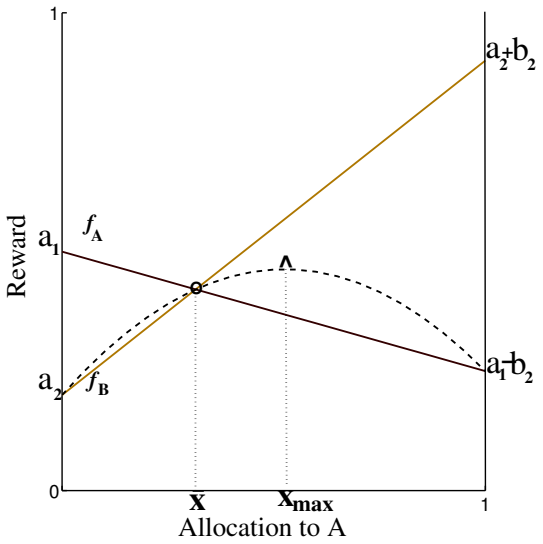

Fig. 2. The matching shoulders task with linear reward schedules, showing that optimal behavior (triangle) need not coincide with matching (circle); coincidence occurs if and only if $f_A(0) = f_B(1)$ [2, Appendix B].

In the converging gaussians task the optimum coincides with the matching point, and so we expect all subjects, whether playing alone or with group feedback, to hover around this point. Near it the reward schedules are well-approximated by the linear case of Fig 2, and recalling the weights accorded to prior rewards of §II-A, this "stable" behavior can be intuitively understood by noting that, if a subject is at the matching point and chooses, say, A, so that the fraction of A's in his history goes up, then his reward drops ($f_A(x) < f_A(\bar{x})$ for $x > \bar{x}$), prompting him to choose B on the next trial. Choice of B (reduction in the fraction of A's) following matching also results in a lower reward, again prompting reversal. Thus, choices tend to cycle around the matching point, and as we shall see in §III, human subjects do approximate this, their deviations providing a measure of noise and/or exploration in the decision making process.

The optimum and matching point also coincide at 50% for the diverging gaussians, but as we show in §III the weight change feedback algorithm of §II-A is typically *unstable* at this point. This can also be understood by a similar argument: the slopes are now reversed and, starting from the matching point, both A and B choices lead to *higher* rewards and further divergence, suggesting that players will not match in this game. Indeed, the task can be parameterized such that the matching point is a local minimum with two (equal) maxima on either side, leading to stronger, and ultimately more rewarding, instability, but here we keep the global optimum at the matching point. Moreover, since the reward schedules are symmetric about 50%, we expect this task to allow us to study "herding" behavior in the group context when choice feedback is supplied.

In [2, Appendix] it is proved that, if the weight difference $w_A - w_B$ employed in the choice probability (3) accurately reflects the difference in rewards $\Delta r(x) = f_A(x) - f_B(x)$ received on two succeeding trials, and $f'_A(\bar{x}) < 0 < f'_B(\bar{x})$, then the matching point $\bar{x}$ is locally asymptotically stable (one replaces $w_A - w_B$ by Δr in (3) and differentiates w.r.t x). A related Liapunov stability result appears in [10], [11].

Subject behavior in the rising optimum tasks (Fig. 1C) is markedly different. In the version of [1], [2] the net rewards curve rises monotonically with x: there is no local maximum at or near the matching point. This prompts some subjects to explore beyond the matching point at $\approx 20\%$ and discover the global optimum at 100% A's. In the present case, however, we select reward schedules with two locally-optimal behaviors, at 0 and 100% or 0 and 75%, neither coinciding with matching. These were chosen such that, playing alone, typical subjects (exploiters) should settle near matching, but a minority of explorers should discover, and mostly remain near, the 100% or 75% global optima [1].[1] This behavior prompted Bogacz et al. [14] to introduce the eligibility traces $e_{A,B}$, since the original model of Eqs. (3-4) could only capture matching behavior [1, Fig. 2], [2, Fig. 12]. With feedback, we conjecture that some exploiters will be prompted to discover the global optimum, or (in case of choice feedback) to follow the lead of the explorers.

C. Models for choice with group feedback

In the no-feedback condition subjects can still be modeled by independent DD processes, although the knowledge that they are in a competitive situation may require parameter modifications. Given explicit information regarding other players' choices and/or rewards, we *expect* parameters to be updated. When only choice feedback is provided it is not clear that other subjects, whose choices deviate significantly from one's own, are doing better or worse, while if only reward scores are provided, the strategies by which they are achieved remain mysterious. With this in mind, we model parameter updates as follows.

With choice feedback alone, we propose a majority rule. For the groups of five subjects used in our experiments, each individual increases his probability of choosing A (or B) by an amount determined by the fraction of A's (or B's) chosen by the other four players on the previous trial. Specifically, a decaying preference weight $u(n)$ is updated prior to each choice by the rule

$$u(n+1) = (1-\lambda)u(n) + \lambda \cdot \begin{cases} +1 & \text{if AAAB or AAAA,} \\ -1 & \text{if BBBA or BBBB,} \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

[1] In [2] these behaviors are called "conservative" and "risky."

TABLE I

PARAMETER VALUES AND L^2 FITTING ERRORS (ERR.) FOR THE DD MODEL WITHOUT ELIGIBILITY TRACE.

	c. gauss.	d. gauss.	c. & d. gauss.	complex	simple
λ	0.95	0.16	0.89	0.10	0.15
μ	2.6	2.5	2.7	11.05	10.90
Err.	0.071	0.114	0.084 & 0.126	0.081	0.165

and added to the weight difference in (3), replacing $w_A(n) - w_B(n)$ by $w_A(n) - w_B(n) + \nu(n)u(n)$, where $\nu(n)$ scales the influence accorded to the other players, additionally biasing the drift rate of the DD process. Eq. (9) maintains the update structure and learning rate λ of (4). Note that $\nu(n) \equiv \nu$ may provide a fitting parameter that remains fixed for a given player and game, or it may vary during a session, being determined as described in §III-B. In either case, one additional parameter describes the strength of feedback.

With reward feedback alone, we suppose that if the maximum among the other players' rewards exceeds his own, an individual will be prompted to further explore the consequences of his choices. This can be achieved by reducing the steepness parameter μ of (3) to promote "random" choices and increase exploratory behavior. Alternatively the weight difference can be modified to promote a switch or, in case his reward exceeds all others, to reinforce his last choice. More complex rules can be envisaged for combined choice and reward feedback. Due to limited space, here we consider only choice feedback; other cases will be treated elsewhere.

III. SOME EXPERIMENTAL AND MODEL DATA

We present preliminary analyses of behavioral data from 5-player groups. Functional magnetic resonance (fMRI) imaging brain scan data from the same groups will be described in a subsequent publication.

A. Allocations without feedback: group behavior

We start by matching the DD model with drift rates on successive trials updated according to the rule (2) (without ETs), to data from individuals playing the games of Fig. 1 without feedback from other players. Fig. 3 shows histograms of choice allocations from the pooled data of 15, 20, 30 and 35 subjects respectively, all of whom played blocks of 150 trials. To determine initial allocations and hence rewards, subjects were given "seed" histories of $N = 20$ trials (unknown to them).

Model choice distributions were obtained from 15, 20, 30 or 35 consecutive runs of 150 trials, initialized by the starting allocations supplied to the individuals playing each game. Table I lists parameter values and fit errors. Here the learning rate λ and steepness parameter μ were estimated from data averaged across players, but fitted separately to each game. Constraining them to common values for the converging and diverging gaussians shows that the reward schedules, and the resulting trial-to-trial feedback of individual preformance, can lead to markedly different choice allocations *without* changes in DD parameters (fit errors increase modestly: see central column of Table I).

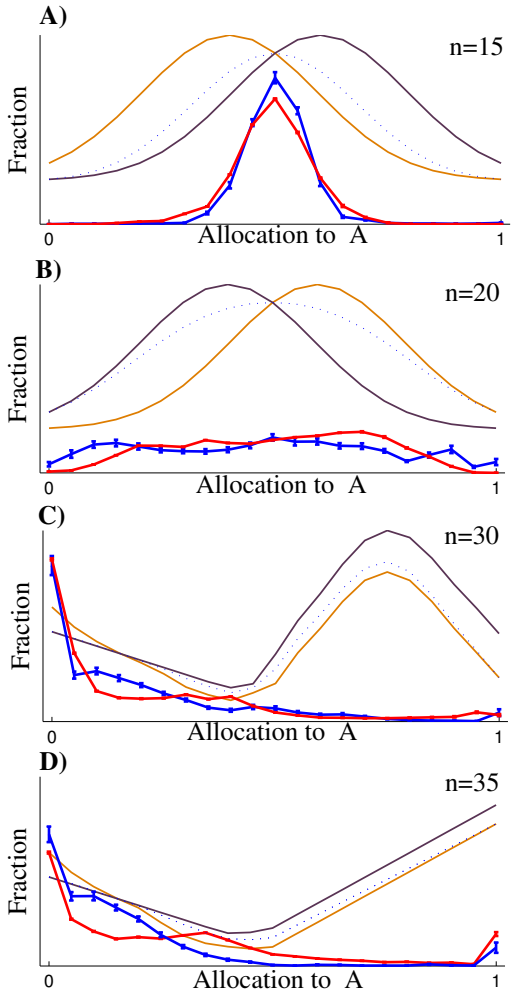

Fig. 3. Envelopes of choice distributions with standard error bars for the four tasks of Fig 1 (solid blue curves), compared with fits of the DD model without eligibility trace (solid red curves). Reward schedules are shown as faint curves. A) converging gaussians; B) diverging gaussians; C) complex rising optimum; D) simple rising optimum. Rising optimum games were also played in reflected versions with global optima to the left on the allocation axes; here data are combined and presented as if all individuals played the same versions.

Fig. 3 shows that the model captures both the "stable" behavior of players in the converging gaussians task, whose allocations remain close to optimal at 50%, as well as their much more diffuse exploratory behavior in the diverging gaussians task, although it underestimates choice fractions near allocations of 0 and 100%. The model captures the allocation distribution of the complex rising optimum slightly better than that of the simple rising optimum. In both cases it reproduces the peak at 0% A's (the local maximum) and approximates the small upticks near 100% A's, but it overestimates allocations to A immediately to the right of the matching point, where rewards are lower, and underestimates them to its left, where rewards are higher. This may be due to fitting to data averaged across all players rather than fitting to individuals and averaging across models.

The model also exhibits trial-to-trial dynamics similar to that of typical subjects. Allocations to A as a function of trial number for the model show that it qualitatively reproduces

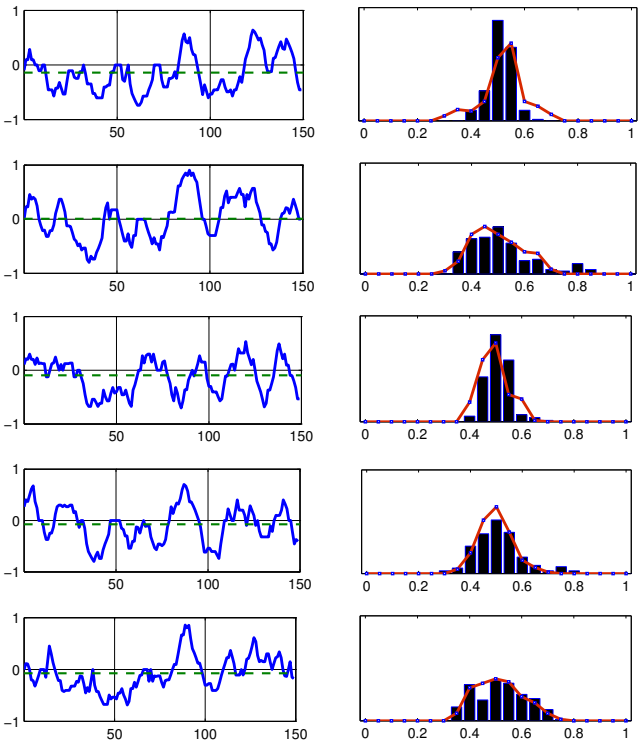

Fig. 4. Left : Windowed correlation sequences for individuals of group 6 playing the converging gaussians task with choice feedback. Right : DD model fits to session-averaged allocation data with choice feedback (9) when corresponding scaled correlation sequence is used as individual's $\nu(n)$); mean correlation values are shown for each individual's sequence as dashed lines.

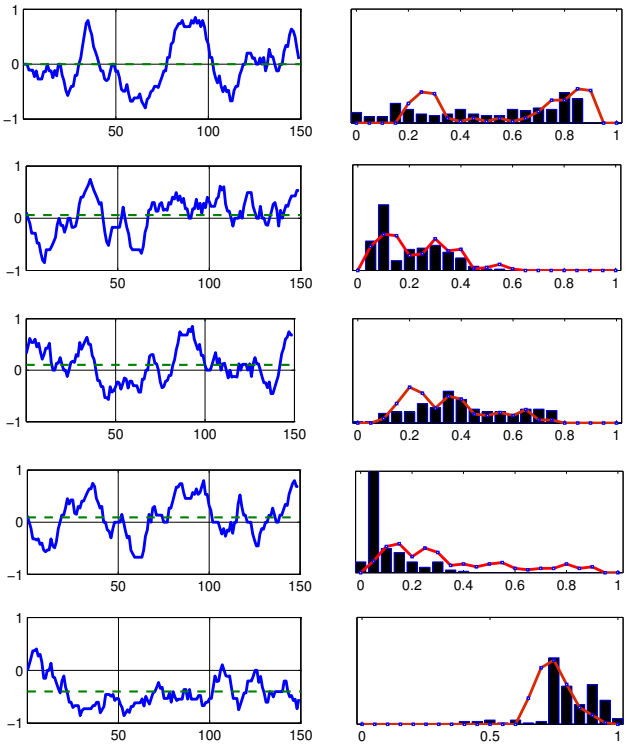

Fig. 5. Left : Windowed correlation sequences for individuals of group 3 playing the diverging gaussians task with choice feedback. Right : DD model fits to session-averaged allocation data with choice feedback (9) when corresponding scaled correlation sequence is used as individual's $\nu(n)$); mean correlation values are shown for each individual's sequence as dashed lines.

rapid cycling around the matching point in the converging gaussians task, and the considerably slower, larger amplitude cycling seen for diverging gaussians. It even captures some features of the rising optimum sequences, occasionally reproducing the discovery, or subsequent abandonment of the global optimum (data not shown here).

B. Allocations with choice feedback: individual behaviors

In Figs. 4 and 5 we show that the preference weight (9) proposed in §II-C provides reasonable fits for individual players of the gaussian games in the choice feedback condition. To better investigate the effect of the time-varying influence coefficient $\nu(n)$ we fixed the parameters λ and μ at the following values obtained from fits to each groups' distribution: $\mu = 2.4$, $\lambda = 0.65$ and $\mu = 3.8$, $\lambda = 0.55$ respectively for groups 6 and 3 playing the converging and diverging gaussian games. Time-varying sequences $\tilde{\nu}_j(n)$ were obtained for each individual by comparing their choices to that of the majority among the other four players. This was done using sliding windows of length 10 and correlating segments of the individual's choice sequence with those of the majority, under a lag of one trial. The resulting sequences are shown in the left columns of Figs. 4 and 5. Finally the $\tilde{\nu}_j(n)$'s were multiplied by a scaling factor $\hat{\nu}_j$ that was adjusted to obtain the influence sequence $\nu_j(n) = \hat{\nu}_j\tilde{\nu}_j(n)$ that best fits each individual's allocation histogram. The parameter $\hat{\nu}_j$ characterizes the individual's response to feedback.

The individual correlation sequences show that players within the same group exhibit a variety of different behaviors. Some have consistently low or high correlations with the majority while others appear to follow their group only intermittently during the session. The right columns of Figs. 4 and 5 also show that the DD model with feedback modulated by the correlation sequence can reproduce individual differences in players' allocation histograms quite well (e.g., note the low and high A (left and right) biases of players 2, 4 and 5 in Fig. 5).

Thus, when DD models play the games with information from human players, they can capture changing feedback dynamics and the resulting choice allocations. We have also verified that when DD model players receive information about other players' choices, the model produces cycling and "transfer of allegiance" behaviors similar to those of human groups (data not shown here). These studies suggest that simple feedback rules such as (9) can capture aspects of both group dynamics and individual behavior. We are currently analyzing and fitting data from other feedback conditions and from the more challenging rising optimum tasks, which will provide more stringent tests of the models.

IV. DISCUSSION AND CONCLUSIONS

The behavioral games and mathematical models described in this article provide a window into the dynamics of

collective decision making, or "social choice," in human and animal groups. We have designed and parameterized games that can be played alone and in groups, and proposed simple feedback rules to couple previous reinforcement learning and DD choice models of single players [1]. Our preliminary analysis of behavioral data, along with model fitting, confirms earlier work on the matching and rising optimum tasks [2], [14]. More strikingly, in the case of choice feedback, it shows that the addition of a time-dependent preference weight based on the actions of fellow players can capture some aspects of group behavior and dynamics. The ET extension (5-6) of [14] was not used in the present work; it could presumably further improve fit quality, particularly in the rising optimum tasks. Results on stability and equilibrium distributions for the model, both with and without group feedback, remain to be determined (cf. [10], [11]). Furthermore, the influence sequences $\nu_j(n)$ can be incorporated in analysis of fMRI data collected during the experiment. Variations in susceptibility to social pressure both between and within subjects may reveal brain structures that process social feedback or regulate its influence on behavior.

The DD model (1) is a continuum limit of the discrete sequential probability ratio test [9], which is optimal in that it delivers a decision of guaranteed accuracy with the smallest possible number of samples [23]. This observation links the choice model of [1], [2] with a mainstay of statistical decision theory, suggesting that related methods in signal processing such as change detection theory might also be useful in modeling human social behavior.

V. APPENDIX: EXPERIMENTAL METHOD

Subjects, recruited in the Houston area via website and fliers, were given consent forms, instructed regarding the experiment, and led to scanners. An experiment in which groups of five players exchange information during the task was performed. In the group experiment, six different reward-based decision-making games were played, in which each subject chose by pressing one of two buttons (A or B), receiving a points reward after each choice. Rewards were determined as explained in the main text, but the rule was not explained to participants, who were simply told to accumulate as many points as possible. Each subject played each game for a session containing 150 choices (2.5 seconds inter-trial interval, synchronized across group members), after which a screen indicated the start of a new game. Games were presented in a randomized order.

There were four information conditions in group games: *alone*, in which subjects only received reward feedback on their own choices; *reward*, in which points earned by other group members after each choice were displayed; *choice*, in which subjects could see other group members' choices, and *both*, in which points earned and choices were displayed. Subjects were shown the type of information being presented, and feedback conditions remained constant over blocks of choices. Each group played under each condition at least once, and the conditions played by the various groups were balanced across the total number of groups.

After completion, subjects were debriefed and compensated according to their point totals, following guidelines set by the Princeton University and Baylor College of Medicine Institutional Review Panels.

VI. ACKNOWLEDGMENTS

This work was supported by AFOSR FA9550-07-1-0528 under the Multidisciplinary University Research Initiative. Experiments were conducted in the Human Neuroimaging Lab at Baylor College of Medicine.

REFERENCES

[1] D.M. Egelman, C. Person, and P.R. Montague. A computational role for dopamine delivery in human decision-making. *J. Cogn. Neurosci.*, 10:623–630, 1998.

[2] P.R. Montague and G.S. Berns. Neural economics and the biological substrates of valuation. *Neuron*, 36:265–284, 2002.

[3] J.R. Krebs, A. Kacelnik, and P Taylor. Tests of optimal sampling by foraging great tits. *Nature*, 275:27–31, 1978.

[4] J.D. Cohen, S.M. McClure, and A.J. Yu. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Phil. Trans. Roy. Soc. Lond. B.*, 362:933–942, 2007.

[5] J.I. Gold and M.N. Shadlen. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36:299–308, 2002.

[6] R. Ratcliff. A theory of memory retrieval. *Psych. Rev.*, 85:59–108, 1978.

[7] R. Ratcliff, T. Van Zandt, and G. McKoon. Connectionist and diffusion models of reaction time. *Psych. Rev.*, 106 (2):261–300, 1999.

[8] P.L. Smith and R. Ratcliff. Psychology and neurobiology of simple decisions. *Trends in Neurosci.*, 27 (3):161–168, 2004.

[9] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J.D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two alternative forced choice tasks. *Psych. Rev.*, 113(4):700–765, 2006.

[10] M. Cao, A. Stewart, and N.E. Leonard. Integrating human and robot decision-making dynamics with feedback: Models and convergence analysis. In *Proc. 47th IEEE Conf. on Decision and Control*, 2008.

[11] M. Cao, A. Stewart, and N.E. Leonard. Convergence in human decision-making dynamics and integration of humans with robots. Systems and Control Letters, submitted, 2008.

[12] C.W. Gardiner. *Handbook of Stochastic Methods, Second Edition*. Springer, New York, 1985.

[13] J.R. Busemeyer and J.T. Townsend. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100:432–459, 1993.

[14] R. Bogacz, S.M. McClure, J. Li, J.D. Cohen, and P.R. Montague. Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, 1153:111–121, 2007.

[15] P.R. Montague, P. Dayan, and T.J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, 16:1936–1947, 1996.

[16] J.N. Reynolds and J. Wickens. Dopamine-dependent plasticity of corticostriatal synapses. . *Neural Netw.*, 15:507–521, 2002.

[17] R. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44, 1988.

[18] R.S. Sutton and A.G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, MA, 1998.

[19] J. Guckenheimer and P.J. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer-Verlag, New York, 1983.

[20] R.J. Herrnstein. Melioration as behavioral dynamism. In M.L. Commons, R.J. Herrnstein, and H. Rachlin, editors, *Quantitative Analyses of Behavior. Matching and Maximizing Accounts, Vol II*. Ballinger Publishing Co., Cambridge, MA, 1982.

[21] R.J. Herrnstein. Rational choice theory: Necessary but not sufficient. *American Psychologist*, 45:356–367, 1990.

[22] R.J. Herrnstein. Experiments on stable suboptimality in individual behavior. *AEA Papers and Proceedings*, 81:360–364, 1991.

[23] A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *Ann. Math. Statist.*, 19:326–339, 1948.