Designing Engaging Games Using Bayesian Optimization

Mohammad Khajah[⋆], **Brett D. Roads**[⋆], **Robert V. Lindsey**[⋆], **Yun-En Liu**[†], **Michael C. Mozer**[⋆]
[⋆] Department of Computer Science, University of Colorado, Boulder
[†] Department of Computer Science and Engineering, University of Washington

Abstract

We use Bayesian optimization methods to design games that maximize user engagement. Participants are paid to try a game for several minutes, at which point they can quit or continue to play voluntarily with no further compensation. Engagement is measured by both actual play duration and a projection users make about how long others will play. Using Gaussian process surrogate-based optimization, we conduct efficient experiments to identify game design characteristics that lead to maximal engagement. In this paper, we study a game requiring trajectory planning, the difficulty of which is determined by a three-dimensional continuous design space. Two of the design dimensions manipulate the game in a user-transparent manner (e.g., the spacing of obstacles), the third in a covert manner (subtle trajectory corrections). Converging results indicate that covert manipulation is significantly more effective in driving engagement, suggesting the critical role of a user's self-perception of competence.

1 Introduction

A recent surge of research has applied game-like mechanics to enhance engagement in domains such as personal health [6, 8], scientific discovery [9, 4], and education [5, 12, 11, 10]. Increased engagement should improve user experiences, data collection, and outcomes. Engagement can be readily quantified using electronic games. Using educational games as an example, one can measure the fraction of time students are attending to the screen [13], the rate of responses [1], the number of attempts to solve a problem, and a persistent focus on a single task [10]. All these measures relate to the user spending more time on task, which should ultimately yield better learning outcomes. The goal of our work is to design games that maximize engagement for a population of users via the manipulation of task difficulty or challenge. In the past, design decisions have been made with a designer's intuitions, A/B testing, or multiarm bandits. We use Bayesian optimization with Gaussian processes (GPs). In our application, the GP posterior represents a mapping from game designs to latent engagement states induced by a design. We must specify an observation model that characterizes the generative process by which an engagement state translates to a voluntary-play duration.

2 Bayesian Optimization

In this section, we identify an observation model that is robust to misspecification: we would like the model to work well even if real-world data—engagement as measured by the duration of play—are not distributed according to the model's assumptions. An observation model must have three properties to be suitable for representing play-duration distributions: (1) nonnegative support, (2) variance that increases with the mean, and (3) probability mass at zero to represent individuals who express no interest in voluntary play. To satisfy these three properties, our generative process assumes that play duration, denoted V, is given by $V = CT$, where $C|\pi \sim \text{Bernoulli}(\pi)$ is an individual's binary choice to continue playing or not and T is the duration of play if they continue. Criterion 1 rules out the popular ex-Gaussian density because it has nonzero probability for negative values. We tested four alternative distributional assumptions for T:

$$T \sim \text{Gamma}\left(\alpha, \frac{\alpha}{e^{f(\boldsymbol{x})}}\right) \quad T \sim \text{Weibull}\left(k, \frac{e^{f(\boldsymbol{x})}}{\Gamma\left(1+\frac{1}{k}\right)}\right)$$
$$T \sim \ln \mathcal{N}\left(f(\boldsymbol{x}) - \frac{\sigma^2}{2}, \sigma^2\right) \quad T \sim \text{Wald}\left(\lambda, e^{f(\boldsymbol{x})}\right)$$

where $\boldsymbol{x}$ is a game design and $f(\boldsymbol{x})$ is the latent valuation and has a GP prior. The first parameter of the Gamma, Weibull, and Wald distributions specify the *shape*, and the second parameter specifies the *rate*, *scale*, and *mean*, respectively. The two parameters of the log-Normal distribution specify the mean and variance, respectively. These four distributions all share the same mean, $e^{f(\boldsymbol{x})}$, but have different higher-order moments. To allow a design's valuation $f(\boldsymbol{x})$ to influence the choice C as well as the play duration T, we define $\text{logit}(\pi) \equiv \beta_0 + \beta_1 f(\boldsymbol{x})$. This general form includes design invariance as a special case ($\beta_1 = 0$).

We performed synthetic experiments with each of these four observation models. To evaluate robustness to misspecification, we evaluated each model using the same four models to simulate the underlying generative process (i.e., to generate synthetic data meant to represent human play durations). Synthetic data for these experiments were obtained by probing a valuation function, $f(\boldsymbol{x})$, that represents the engagement associated with a design $\boldsymbol{x}$. For $f(\boldsymbol{x})$, we used a mixture of two to four Gaussians with randomly drawn centers, spreads, and mixture coefficients, defined over a 2D design space. For examples, see Figure 1a. We generate synthetic observations by mapping the function value through the assumed generative process. The goal of Bayesian optimization is to recover the function optimum from synthetic data. We performed 100 replications of the simulated experiment, each with a different randomly drawn mixture of Gaussians and with $\beta_0 = 0$ and $\beta_1 = 1$. For the generative models, we need to assume values for the free parameters, and we used $\alpha = 2$, $k = 2$, $\sigma^2 = 1$ and $\lambda = 4$. (These parameters settings are used to generate the synthetic data and are not shared with the Bayesian optimization method.)

To perform Bayesian optimization, we require an *active-selection policy* that determines where in design space to probe next. The *probability of improvement* and *expected improvement* policies are popular heuristics in the Bayesian optimization literature. Both policies balance exploration and exploitation without additional tuning parameters. However, since the variance increases with the mean in our observation models, both policies tend to degenerate to pure exploitation. Instead, we chose Thompson sampling [3], which is not susceptible to this degeneracy. For each replication of the simulated experiment, we ran 40 active selection rounds with 5 observations (simulated subjects) per round. The GP used the squared exponential Automatic-Relevance-Determination (ARD) kernel whose hyperparameters were inferred by slice sampling.

For each combination of the four distributions as observation model and for each combination of the four distributions as generative model, we ran the battery of 100 experiment replications each with 200 simulated subjects. The simulation results are summarized in Figure 1b. By two measures of performance, the log-Normal distribution is most robust to incorrect assumptions about the underlying generative process. We use this observation model in the human studies that follow.

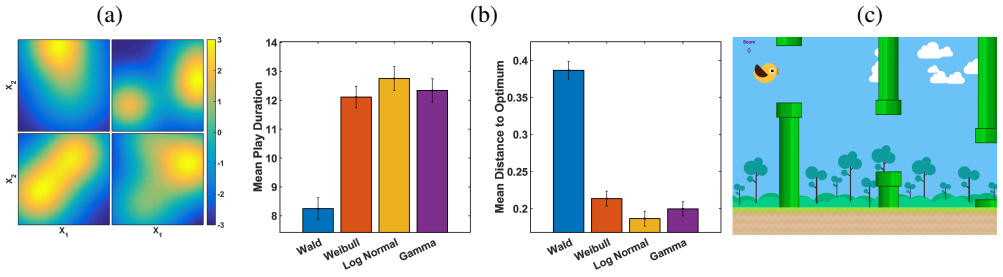

Figure 1: (a) Examples of the 2D functions used for generating synthetic data. (b) Results of synthetic experiment. The left and right plots depict the mean function value (higher is better) and the mean distance to the true optimum (lower is better) for various observation models. Results are averaged over four different generative-process models, 100 replications of each simulation, the last 10 trials per replication. Error bars indicate ± 1 standard error. (c) Flappy Bird: The player flaps bird's wings to keep it aloft and to avoid hitting pipes.

3 Experiments

Our human studies were conducted using Amazon's Mechanical Turk platform. Given evidence from earlier studies that Turk participants are willing to voluntarily commit time to activities that they find engaging, we devised a method for measuring *voluntary time on activity* or *VTA*. In each of our experiments, participants are required to play a game for sixty seconds. During the mandatory play period, a clock displaying remaining time is displayed. When the mandatory play period ends, the clock is replaced by a button that allows the participant to terminate the game and and receive full compensation. Participants are informed that they can continue playing with no further compensation. The *experiential* VTA is measured as the lag between the button appearance and the button press. After clicking the button, participants were told their experiential VTA and asked to enter how long they expected *other* players to voluntarily play. We call this measure the *projected* VTA and we use it exclusively in this work as the measure of engagement because our earlier pilot studies showed that it is a smoother version of the experiential VTA.

A specific research question we address in this experiment is whether *covert* manipulation of difficulty is more effective in engaging users than *overt* manipulations—those of which users are fully aware and to which they can attribute causal effects. Covert manipulations can be used to make a task more difficult than the user believes, but also to make a task easier than the user believes.

We studied Flappy Bird, a simple popular trajectory-planning game whose objective is to keep a bird in the air by flapping its wings to resist gravity and avoid hitting vertical pipes or the top/bottom of the screen (Figure 1). We manipulated two overt factors affecting game difficulty—the horizontal spacing between pipes and the vertical gap between pipes—as well as one covert factor, which we refer to as the *assistance*. Assistance acts as a force that, when the wings are flapped, steers the bird toward the gap between the next pair of pipes. The assistance level can be adjusted to range from no assistance whatsoever to essentially a guarantee that nearly any action taken by the player will result in success. For moderate levels of assistance, the manipulation can be quite subtle. In informal testing, players were unaware that the game dynamics were modulating to guide them along.

We conducted two studies with Flappy Bird. In the first study, we tested 958 participants. Each participant was assigned to a random point in the three dimensional, continuous design space. The large number of participants in this *random-assignment* experiment enabled us to fit an accurate model that characterizes the relationship between the game design and latent engagement. In the second study, we ran the experiment again from scratch and tested 201 participants. Participants were assigned to designs chosen by Thompson sampling. We seeded active-selection by assigning the first 55 participants to a Sobol-generated set of random points in design space. Here we included a short questionnaire about the participant's experience in the game. The questionnaire consisted of 6 true/false items with each item phrased such that "true" corresponds to an engaging game. Four phrases in the questionnaire were taken from the Game Engagement Questionnaire [2].

Among participants in the active-selection study, the mean experiential VTA is 10 sec, with SD 42 sec and range 0–298 sec; 20% of participants chose to play beyond the requirement. The mean projected VTA is 23 sec, with standard deviation 33 sec and range 0–199 sec; 84% of participants projected that others would continue playing beyond the requirement.

Figures 2a and 2b show the model posterior mean VTA over the three dimensional design space in the random-assignment and active-selection studies, respectively. The reassuring finding is that the two independent studies yield very similar outcomes: the optimal design identified by the two studies is in almost exactly the same point in design space (the red squres in the Figures). The random-assignment study should yield reliable results due to the relatively large number of participants tested. The active-selection study is far more efficient in its use of participants, due to intelligent selection of where to explore in design space. In both studies, the peak design is predicted to obtain a VTA of 30 seconds—an increase of 50% of the time on task. Because Turk workers are paid by the task, this time increase reduces the pay rate by two thirds, a fairly clear indication of engagement.

The Figures indicate that engagement is sensitive to each dimension the design space with not much hint of an interaction across the dimensions. Notably, with no covert assistance—the leftmost array in each Figure—the other two overt difficulty dimensions have little or no impact on engagement, and are not sufficient to motivate participants to continue playing voluntarily. Thus, we conclude that covert assistance is key to engaging our participants. Consistent with the hypothesis that participants

need to be unaware of the assistance, the experiments show that engagement is poor with maximum assistance—the rightmost array in each Figure. With maximum assistance, the manipulation causes the bird to appear to be pulled into the gap, and this is therefore no longer covert in nature.

To obtain further converging evidence in support of the optimum identified in Figures 2a and 2b, we fitted a Gaussian process model to questionnaire scores. We defined the score as the number of 'true' responses made by the participant. The higher the score, the higher the engagement because we phrased questionnaire items such that an affirmative response indicated engagement. We used Gaussian process regression with a binomial observation model to fit the scores. Figure 2c shows the model posterior mean score over the three dimensional design space. The notable result here is that the posterior mean score looks similar to the posteriors from the random-assignment and active-selection studies. More importantly, the predicted design optima, denoted by red squares, lie close to one another in Figures 2a, 2b and 2c. The consistency across studies and across response measures provides converging evidence that increase our confidence in the experiment outcomes, and also provide support for the appropriateness of using VTA as measure of engagement in place of a more traditional questionnaire.

4 Discussion

We have applied Bayesian optimization with a suitable generative theory to the problem of designing software to engage users. A key component of the research described in this article is our exploration of candidate generative theories, and a contribution of our work is the specification of a theory that is robust to misspecification, i.e., robust to the possibility that humans behave differently than the theory suggests. We collected behavioral and self-reported measures of engagement and obtained converging evidence from these two different measures. We also showed that covert manipulation of game dynamics had the most impact on engagement which we believe is due to players attributing in-game success to their own competence. In future research, we plan to conduct longer-term usage studies and to apply Bayesian optimization to specific users rather than populations.

5 Acknowledgments

This research was supported by NSF grants SES-1461535, SBE-0542013, and SMA-1041755.

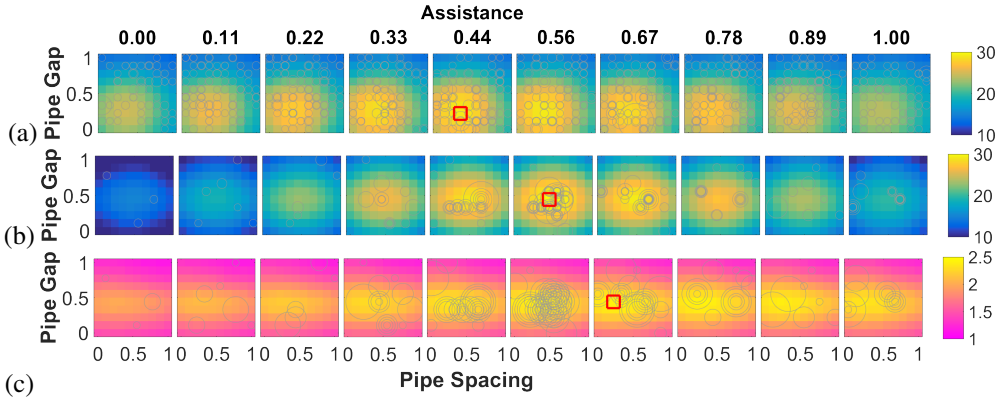

Figure 2: Bayesian model fits of VTA (in seconds) over the Flappy Bird design space for (a) the random-assignment and (b) active-selection studies. Each array corresponds to a fixed level of assistance, from no assistance (level 0) to maximal (level 1). Each array depicts model-fit VTA across the range of horizontal spacings between pipes (x axis) and vertical gaps (y axis). Pipe gap and pipe spacing is calibrated such that a level of 0 is a challenging game and 1 is readily handled by a novice. The circles correspond to observations with the radii indicating the magnitudes of the observations. At locations with multiple observations, there are co-centric circles. Red squares indicate the locations of the predicted global maximum. (c) An analogous Bayesian model fit to the questionnaire score, which indicates the number of items with an affirmative response. Higher scores indicate greater engagement.

References

[1] BECK, J. E. Engagement tracing: Using response times to model student disengagement. In *Proceedings of the 2005 Conference on AI in Education* (Amsterdam, 2005), IOS Press, pp. 88–95.

[2] BROCKMYER, J. H., FOX, C. M., CURTISS, K. A., MCBROOM, E., BURKHART, K. M., AND PIDRUZNY, J. N. The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology 45*, 4 (2009), 624–634.

[3] CHAPELLE, O., AND LI, L. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems* (2011), pp. 2249–2257.

[4] COOPER, S., KHATIB, F., TREUILLE, A., BARBERO, J., LEE, J., BEENEN, M., LEAVER-FAY, A., BAKER, D., POPOVIĆ, Z., ET AL. Predicting protein structures with a multiplayer online game. *Nature 466*, 7307 (2010), 756–760.

[5] DE SOUSA BORGES, S., DURELLI, V. H. S., REIS, H. M., AND ISOTANI, S. A systematic mapping on gamification applied to education. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing* (New York, 2014), ACM, pp. 216–222.

[6] FITOCRACY. Fitocracy, 2015.

[7] FREDRICKS, J. A., BLUMENFELD, P. C., AND PARIS, A. H. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research 74* (2004), 59–109.

[8] JURGENS, D., MCCORRISTON, J., AND RUTHS, D. An analysis of exercising behavior in online populations. In *Ninth International AAAI Conference on Web and Social Media* (2015).

[9] KHATIB, F., DIMAIO, F., COOPER, S., KAZMIERCZYK, M., GILSKI, M., KRZYWDA, S., ZABRANSKA, H., PICHOVA, I., THOMPSON, J., POPOVIĆ, Z., ET AL. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology 18*, 10 (2011), 1175–1177.

[10] LIU, Y.-E., MANDEL, T., BRUNSKILL, E., AND POPOVIC, Z. Towards automatic experimentation of educational knowledge. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, 2014), ACM, pp. 3349–3358.

[11] LOMAS, J. D. *Optimizing motivation and learning with large-scale game design experiments*. Unpublished doctoral dissertation, HCI Institute, Carnegie Mellon University, November 2014.

[12] LOMAS, J. D., PATEL, K., FORLIZZI, J. L., AND KOEDINGER, K. R. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, 2013), ACM, pp. 89–98.

[13] WHITEHILL, J., SERPELL, Z., LIN, Y.-C., FOSTER, A., AND MOVELLAN, J. R. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing 5* (2014), 86–98.