Fast and robust Bayesian inference for modular combinations of dynamic learning and decision models

Krishn Bera (krishn_bera@brown.edu)
Carney Institute for Brain Science
Dept. of Cognitive and Psychological Sciences
Brown University

Alexander Fengler (alexander_fengler@brown.edu)
Carney Institute for Brain Science
Dept. of Cognitive and Psychological Sciences
Brown University

Michael J. Frank (michael_frank@brown.edu)
Carney Institute for Brain Science
Dept. of Cognitive and Psychological Sciences
Brown University

Abstract

In cognitive neuroscience, there has been growing interest in adopting sequential sampling models (SSM) as the generative choice function for reinforcement learning (RLSSM) to jointly account for decision dynamics within and across trials. However, such approaches have been limited by computational tractability due to lack of closed-form likelihoods for the decision process or expensive trial-by-trial evaluation of complex reinforcement learning (RL) processes. We enable hierarchical Bayesian estimation for a broad class of RLSSM models, using Likelihood Approximation Networks (LANs) in conjunction with differentiable RL likelihoods to leverage fast gradient-based inference methods including Hamiltonian Monte Carlo or Variational Inference (VI). To showcase the scalability and faster convergence with our approach, we consider the Reinforcement Learning - Working Memory (RLWM) task and model with multiple interacting generative learning processes. We show that our method enables accurate recovery of the posterior parameter distributions in arbitrarily complex RLSSM paradigms, and moreover, that in comparison, fitting data with the equivalent choice-only model yields a biased estimator of the true generative process. Moreover, leveraging the SSM with efficient inference allows us to uncover a heretofore undescribed cognitive process within the RLWM task, whereby participants proactively adjust the decision threshold as a function of WM load.

Keywords: Hierarchical Bayesian Inference; Variational Inference; Reinforcement Learning; Sequential Sampling Models; Likelihood-Free Inference

Introduction

Reinforcement Learning – Sequential Sampling Models (RLSSM) are a powerful and expressive class of models that are naturally suited for computational modeling of cognitive tasks where the learning process informs the decision-making process. However, to date empirical data analysis has been mostly limited to basic instances of RLSSM that employ Drift Diffusion Models or simple race models and n-armed bandits (Pedersen, Frank, & Biele, 2017; Fontanesi, Gluth, Spektor, & Rieskamp, 2019; Miletić et al., 2021). This is notwithstanding great theoretical interest in more elaborate models, but for which Bayesian parameter inference and quantitative model comparison are traditionally hampered by expensive likelihood evaluations. The computational cost may arise from a lack of closed-form, likelihood functions for the decision process, a more complex reinforcement learning (RL) process, or a combination of both (Fengler, Bera, Pedersen, & Frank, 2022).

Here, we show how Variational Inference (VI) methods (Blei, Kucukelbir, & McAuliffe, 2017; Jordan, Ghahramani, Jaakkola, & Saul, 1998; Liu & Wang, 2016) in combination with LANs and differentiable complex RL dynamics, can be leveraged for computationally efficient and scalable Bayesian inference treatment of a broad class of RLSSM models. While we focus on a specific example, and link to empirical findings of independent interest, we stress that the methods explored are much more generally applicable. Using empirical and synthetic datasets of a widely used cognitive task, the Reinforcement Learning - Working Memory (RLWM) task (Collins & Frank, 2012), we show that this approach can yield fast, tractable inference with a high-dimensional model ($D = 10$ for each participant, with 900 free model parameters) on a large dataset ($> 31k$ trials) in hierarchical settings to recover the posterior distribution over parameters accurately.

Our proposed VI method via differentiable surrogate likelihoods and RL processes allows hierarchical Bayesian inference with arbitrarily complex RLSSM models. We leverage this methodological advance to investigate an outstanding question about how set size (the number of stimulus-response pairs to be learned within a block) during instrumental learning modulates choice RT effects. Previous computational accounts have explained choice RT effects based on Hick's Law (Hick, 1952), value-based differences (Pedersen et al., 2017) or learning and set size interactions (McDougle & Collins, 2021). Here, we build on these models to additionally hypothesize a speed-accuracy tradeoff such that participants might proactively adjust the boundary of the choice process in anticipation of blocks with higher WM load. We provide evidence for this account while also showing the fitting the dynamics of the choice process via SSMs can also more accurately recover RL parameters.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

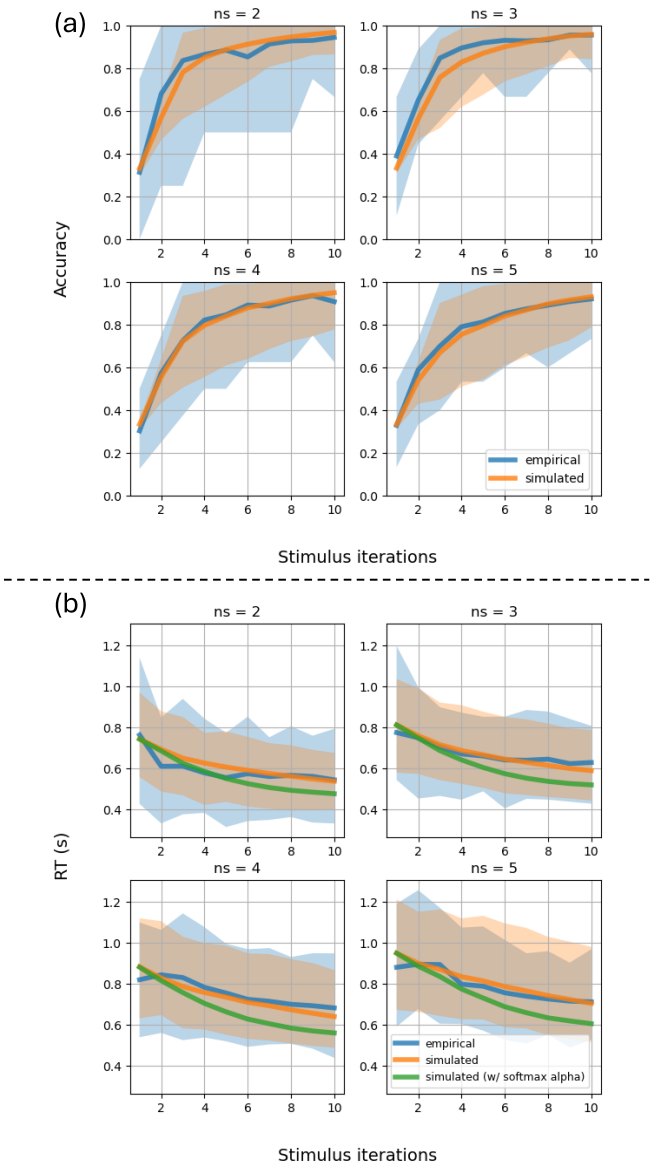

Figure 1: Posterior predictive checks for (a) accuracy and (b) reaction times grouped by set size (*ns*). The solid lines indicate mean across participants and simulations. The shaded region corresponds to 94% HDI ranges.

Method

Participants

All participants were recruited from five different US locations as part of the Cognitive Neuroscience Test Reliability and Computational Applications for Schizophrenia Consortium (CNTRaCS). The experimental procedures were approved by Institution Review Board at Washington University. A total of 255 participants without mental health diagnoses (n=87) or with schizophrenia (n=67), major depressive disorder (n=54) or bipolar disorder (n=47) performed the RLWM task during EEG recording. For this study, we model the behavioral data from the control group (n=87) to experiment with and refine the methodological advances proposed in this work.

Reinforcement Learning Working Memory (RLWM) task

The RLWM task (Collins & Frank, 2012; Collins, Brown, Gold, Waltz, & Frank, 2014) is designed to disentangle the contributions of reinforcement learning and working memory (WM) in stimulus-response learning. Participants learn stimulus-response associations through trial-and-error feedback in a 3-alternative forced-choice paradigm. In addition to incremental RL, the task systematically varies WM demands by manipulating the set size (the number of unique stimuli, ranging from 2 to 5 per block) and delay (the number of trials before re-encountering a stimulus). The participants completed 10 training blocks for 360 trials total. By adjusting WM load and analyzing reward-based accuracy improvements, the RLWM task provides a structured approach to evaluating the interplay and distinct contributions of RL and WM to learning processes.

Computational Modeling

We employ three different RLWM models: a base model and two additional variants to test the hypothesis concerning WM load modulated speed-accuracy trade-offs. For our base model, we employ a version of the RLWM model which decomposes choice behavior into RL and WM processes (Collins & Frank, 2012). It is common for researchers to fit only choice versions of RLWM models, despite known systematic effects of WM load on reaction times, which strongly suggest the inclusion of RTs for inferring the core generative process. The RTs contain valuable information that could constrain and help interpret the generative RL and decision process. Because the optimal action is largely convergent between WM and RL processes, the RL learning rate can sometimes mimic WM processes and vice versa (Yoo & Collins, 2022). We reason that modeling RTs would allow us to further disentangle these processes: for example, RL influences should gradually speed RTs over learning especially over higher set size. In addition, the underlying psychological processes involved in arbitrating WM and RL processes could have potentially rich connections to the vast literature investigating speed-accuracy trade-offs in simple decision mak-

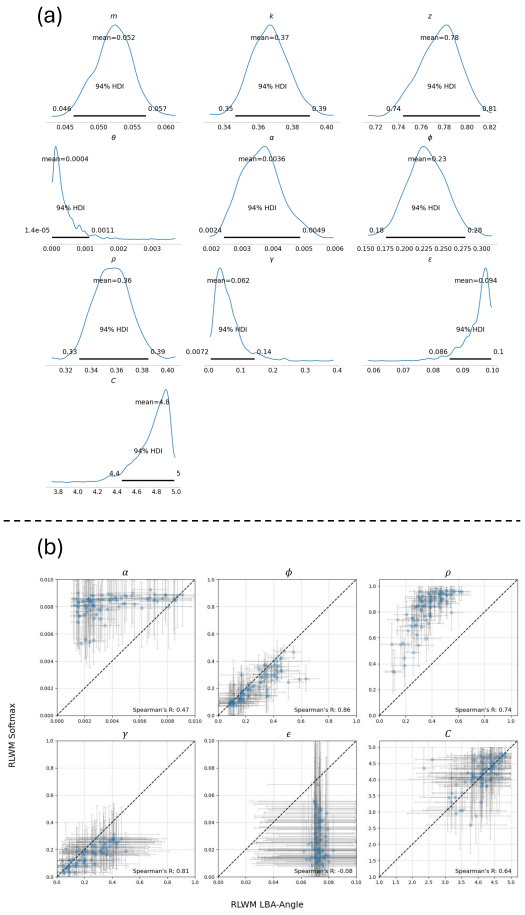

Figure 2: (a) Group-level parameter posterior distributions from RLWM LBA-Angle model fitting. Solid black lines indicate 94% HDI range. (b) Comparison of common participant-level parameters recovered between RLWM LBA-Angle and RLWM Softmax models. The error bars indicate 94% HDI range. Spearman's R is computed for each parameter.

ing. That is, we hypothesized that people engage in a speed-accuracy trade-off that could potentially influence the degree to which WM and RL processes contribute to learning depending on factors such as WM load. Hence, to better characterize these mechanisms using both RT and choice data, we replace the conventional softmax choice rule with a sequential sampling model. The significance of this work is three-fold. First, we demonstrate an efficient Bayesian inference method (via differentiable RL likelihoods) for high-dimensional RLSSM models that can leverage RTs for better characterization of the underlying neuro-cognitive processes (in the case of RLWM task and model, for disentangling RL learning rate from WM). Second, we extend the previous RLSSM accessibility to SSMs that traditionally don't have an analytical likelihoods to showcase the flexibility of our approach. And lastly, we leverage the model to investigate speed-accuracy trade offs to improve RT fits.

Our modeling builds of a previous report (McDougle & Collins, 2021) showed that the softmax rule can be replaced by a Linear Ballistic Accumulator (LBA) to capture RTs. That is, the LBA assumes that each of the three actions has an accumulator whose rate is proportional to the learned value (within RL and WM processes), and that choices are determined when one of the accumulators reaches a decision bound. As the WM and RL processes update over trials, the RTs evolve accordingly; variance across trials in the starting points of the accumulation process can account for further RT variability. The basic LBA has an analytical likelihood, however for proof of concept to expose the generality of our method, we employ an LBA variant that introduces linear collapsing bounds (LBA-Angle) for which the analytical likelihood has not yet been derived, thus requiring surrogate likelihoods (via LANs). (This also allows us to test whether the bound does indeed collapse with time, although in practice we found that it did not and that the resulting angle was zero). The LBA-Angle model is parameterized by six parameters – three drift rates ($v0, v1, v2$ - each corresponding to a choice in the RLWM task), a decision boundary (a), an upper limit of starting-point uniform distribution (z) and a linear collapse parameter (θ).

The trial-wise learning updates for the RL (eq. 1) and WM (eq. 2) processes are given by the following learning equations, where s is the stimulus / state of the current trial, a is the chosen action, r_{t-1} is the reward obtained in the current trial, α is the basic learning rate, and γ is a perseverance parameter.

$$Q_{RL}(s,a)_t = \begin{cases} Q_{RL}(s,a)_{t-1} + \alpha \cdot (1 - Q_{RL}(s,a)_{t-1}) & \text{if } r_{t-1} = 1 \\ Q_{RL}(s,a)_{t-1} + \alpha \cdot \gamma \cdot (-Q_{RL}(s,a)_{t-1}) & \text{if } r_{t-1} = 0 \end{cases} \quad (1)$$

$$Q_{WM}(s,a)_t = \begin{cases} 1 & \text{if } r_{t-1} = 1 \\ Q_{WM}(s,a)_{t-1} + \gamma \cdot (-Q_{WM}(s,a)_{t-1}) & \text{if } r_{t-1} = 0 \end{cases} \quad (2)$$

The model includes a forgetting mechanism by allowing the WM Q-values to decay toward their initial values of $1/n_a$, where n_a represents the number of possible actions. This WM decay process is governed by the parameter ϕ (eq. 3), which reflects the degree of WM interference from intervening trials before a stimulus is re-encountered.

$$Q_{WM}(s,a)_t = Q_{WM}(s,a)_{t-1} + \phi \cdot \left(\frac{1}{n_a} - Q_{WM}(s,a)_{t-1}\right) \quad (3)$$

In each trial, the Q-values generated by the RL and WM systems are converted into a softmax policy. The variability in participants' choices due to attention lapses is captured by the ε parameter (described below). The resulting probabilities, represented as the RL policy pol_{RL} and the WM policy pol_{WM}, are then combined to produce a weighted mixture policy, pol_{mix}. This mixture policy reflects the relative influence of WM reliance, denoted by ω, and is defined mathematically by equations 4 and 5. This formulation allows the model to dynamically balance the contributions of the RL and WM systems in driving behavior. The ρ parameter represents a participant's baseline tendency to rely on WM across contexts, C represents the participant's WM capacity and ns is the block set size.

$$\omega = \rho \cdot min\left(1, \frac{C}{ns}\right) \quad (4)$$

$$pol_{mix} = (1-\omega) \cdot pol_{RL} + \omega \cdot pol_{WM} \quad (5)$$

$$pol_{eps} = (1-\varepsilon) \cdot pol_{mix} + \varepsilon \cdot \left(\frac{1}{n_a}\right) \quad (6)$$

The final evaluation policy is obtained by multiplying the epsilon policy with a scalar parameter η, which in turn determines the trial-by-trial drift-rate of the decision process (McDougle & Collins, 2021).

$$pol_{final} = \eta \cdot pol_{eps} \quad (7)$$

We refer to the above described RLWM-LBA-Angle model as the "Base" model. The following parameters were estimated for the Base model at the participant-level and the group-level - $\alpha, \phi, \rho, \gamma, \varepsilon, C, \eta, a, z, \theta$.

In addition to the Base model, we tested two additional RLWM-LBA-Angle variants ("A-reg" and "A-reg-eta") to account for speed accuracy tradeoff as a function of WM load. These variants introduced a linear regression on decision threshold as a function of set size (in the RLWM task, participants were shown the set size of the block of trials they would be learning in advance of each block and could thus proactively prepare). The decision threshold a in these models is computed as a function of slope parameter m, intercept parameter k and set size ns. All the other aspects of the model remained the same as in the Base model. With the introduction of this regression, the A-reg-eta model is parameterized by the following parameters - $\alpha, \phi, \rho, \gamma, \varepsilon, C, \eta, m, k, z, \theta$.

The A-reg model is similar to the A-reg-eta model in all aspects except that it drops the policy scaling η parameter in order to constrain the model further and maintain parity of the number of free parameters with respect to the Base model. The A-reg is parameterized by the following parameters - $\alpha, \phi, \rho, \gamma, \varepsilon, C, m, k, z, \theta$.

$$a = m \cdot (ns) + k \quad (8)$$

Note that we originally tested a logistic regression function on set size to account for set size-dependent decision threshold (results not shown here). Our decision to incorporate only the a-regression, linear regression version here, reflects our finding that the recovered parameters of the logistic function suggested a simple linear increase in decision threshold.

Variational Inference

VI approximates the posterior by optimizing a parameterized family of distributions to minimize the KL (or other) divergence from the true posterior. The combination of differentiable surrogate likelihoods (via LANs: see below) and differentiable RL implementation allows computation of ELBO gradients with respect to variational parameters, enabling efficient updates via gradient-based methods for VI. We employed VI instead of MCMC for multiple reasons. Firstly, VI is faster than MCMC and sometimes shows better convergence. Secondly, it affords utilization of LANs in high-dimensional settings which allows a novel approach for tractable Bayesian inference. In this work, Automatic Differentiation Variational Inference (ADVI; Kucukelbir, Tran, Ranganath, Gelman, and Blei (2017)), as implemented in PyMC (Abril-Pla et al., 2023), was used for all the parameter fitting experiments. All experiments were performed by recovering the parameters hierarchically with non-informative priors. The number of VI iterations were set to 20000 with a learning rate of 0.1 and Adagrad (window) optimizer. After the fit, 1000 samples were drawn from the variational posterior for each parameter. To guard against reporting results based on local minima, we repeated each optimization run 20 times. What is reported for further analysis is respectively the best-fit run as determined by the log probability of fit. The ELBO loss was monitored for convergence and stability. The same VI settings were used for all the models compared in this work.

The likelihood was implemented using JAX (Bradbury et al., 2018) and wrapped in PyTensor (Developers, 2024) to interface with PyMC for all autodiff and numerical computations. We trained a multilayer perceptron with five layers (hidden size = 120) to approximate the likelihoods of the joint choices and decision dynamics (RT distributions for each of the three choices), as per the procedures outlined in the Likelihood Approximation Network (LAN) reference paper (Fengler, Govindarajan, Chen, & Frank, 2021). Although the LBA has an analytical likelihood, we used the version with a linear collapsing decision bound (with angle parameter θ) which is one example among many variants that

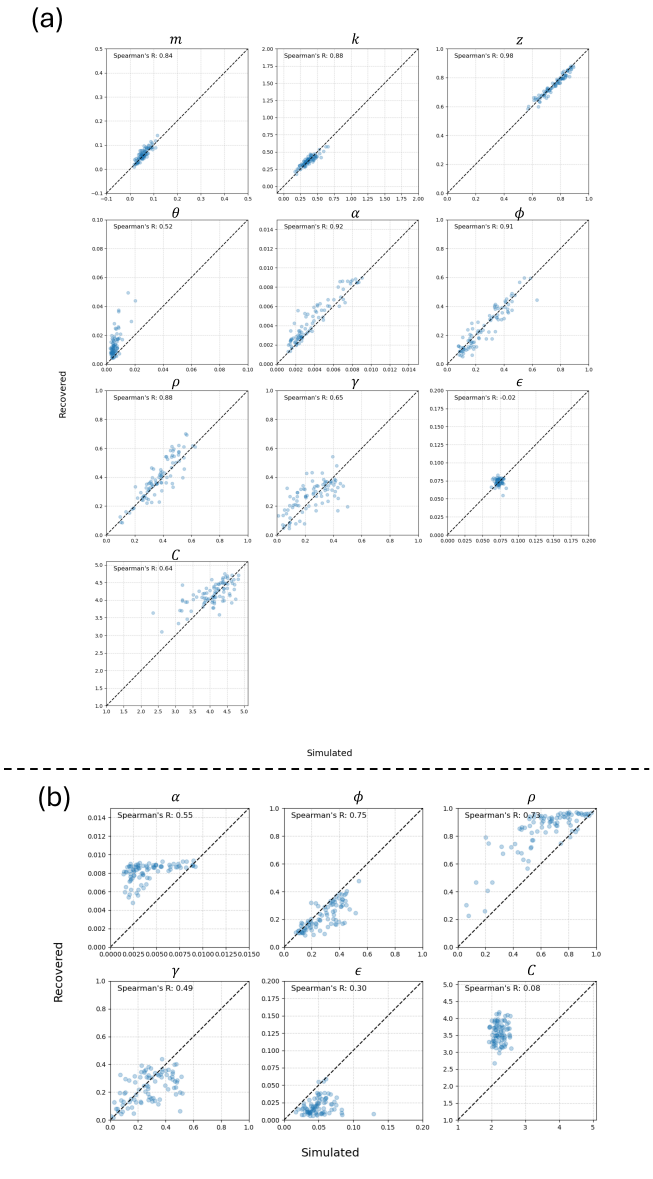

Figure 3: Parameter recovery results with (a) RLWM LBA-Angle and (b) RLWM Softmax model. The simulated dataset was generated from the best-fit parameters from the RLWM LBA-Angle model fitting.

could benefit from LANs. The resulting network takes the data and model parameters (features during training) as input and outputs trial-wise (approximate) log-likelihood values (labels, based on kernel density estimates of empirical likelihood functions). The specific network used in this article was trained on 5×10^4 parameter combinations (with 2×10^4 simulations of each run) of the LBA-Angle model using Pytorch (Paszke et al., 2019).

Results

VI shows strongly favorable runtime when compared to NUTS (Hamiltonian MCMC). Posterior inference on the full hierarchical model reduces to a runtime of less than 2h, which represents a speedup of at least one order of magnitude. The A-reg model is able to capture the summary statistics of the data better than the other candidate models (see details below in sec. Model Comparison) and hence, we utilize this model for all subsequent analysis.

RLSSM model captures both choice and RT distributions

Figure 2(a) shows the posterior distribution of the group-level parameters. Figure 2(b) shows a comparison of common participant-level parameters between the RLWM LBA-Angle (A-reg) model and RLWM Softmax models. The RLWM Softmax overestimates the WM reliance parameter (ρ) as well as RL learning rate (α) relative to the RLWM LBA-Angle model. Indeed, if one simulates RTs from the LBA-Angle model using the αs estimated by the Softmax model, the RTs become progressively faster at a higher rate than observed empirically (Figure 1). In contrast, the RLSSM jointly models the choice and RT distributions, and hence, imposes further constraints on identifying the generative RL parameters separable from WM. Posterior predictive checks confirmed this interpretation. We generated data using estimated parameters from the empirical dataset (sub-sampled by a factor of 50 from the posterior) and compared the observed and simulated results. The simulated dataset was obtained by repeating the simulation process 10 times for each subject. To evaluate the accuracy and RTs across learning for observed and simulated data, we bin trials by set size and stimulus iterations and plot 94% highest density intervals of the mean responses (Figure 1). As predicted, the RL-SSM model more appropriately captures the RT dynamics specifically at higher set sizes when RL is more dominant. It is important to note here that the large variability in HDI is largely due to between-subject variability in overall RTs. Another intriguing finding is that the θ parameter of LBA-Angle was estimated close to 0 on empirical data suggesting that vanilla LBA without collapsing bound better captures the empirical patterns.

Model comparison

To perform model comparison, we employed the WAIC criterion (Watanabe, 2013). The Base model fit the data better than the A-reg and A-reg-eta models but not by a significant margin especially over the A-reg model ($\Delta_{Base}elpd =$

267.99). The A-reg-eta model was the least favored model with $\Delta_{Base}elpd = 663.78$. We note that the WAIC-based model comparison ranking is not necessarily reliable for two reasons. First, as noted in Vehtari, Gelman, and Gabry (2017), the variance of log predictives exceeded 0.4 for a fraction of samples. Second, the difference in WAIC between the Base and A-reg model ($\Delta_{Base}elpd = 267.99$) is smaller than the standard error of the elpd for either model ($SE_{Base} = 272.28$ and $SE_{A-reg} = 279.75$). This suggests the difference may not be statistically significant. Consequently, the computed weights (interpreted as the probability of each model among the compared models given the data) for the Base and A-reg models are 0.522 and 0.478 - at about a random coin flip suggesting very low confidence in ranking both the models. In other words, both the Base and A-reg model are able to account for the data identically.

Since the Base model is strictly nested under the A-reg-eta model, we examined the m and k estimates to check if the additional a-regression was better able to account for the choice RT distributions. With the A-reg-eta model, the group-level m parameter was recovered with mean = 0.051 and 94% HDI = [0.046, 0.056]. With the A-reg model, the group-level m parameter was recovered with mean = 0.052 and 94% HDI = [0.046, 0.057]. The 94% HDI does not contain 0 for both the a-regression models suggesting a significant effect of set-size-dependency on the decision threshold. The observed data provides evidence that set-size modulated decision threshold plays a meaningful role in the data generating process. We further examine absolute model fits using posterior predictive checks with the two best-fitting models - Base and A-reg. We observed that the A-reg model is better able to account for the choice RT distributions especially during the initial stimulus encounters across set sizes. This confirms the intuition that participants may proactively engage cognitive control to increase decision caution for higher WM loads. The difference in mean absolute error in RT during the initial encounters supported this observation ($MAE_{RT}(Base) = 0.1413$ and $MAE_{RT}(A-reg) = 0.1252$).

Aside from this novel result implicating an additional cognitive process not previously accounted for in RLWM models, we also tested whether incorporating the SSM allowed us to better recover RL and WM parameters common with the choice only model. As noted above, because RL and WM systems both converge on the same optimal discrete policy, they can partially mimic each other (Yoo & Collins, 2022). But because RL process updates more incrementally than the WM process, we reasoned that this would be more identifiable in terms of continuously varying RTs across trials.

RLSSM model recovers the parameters better than the RL-only model

To further assess whether these differences allow RLSSM parameters to be more identifiable than the Softmax version, we performed parameter recovery experiments. We simulate a synthetic dataset based on the parameters that best-fit to our empirical dataset and attempted to recover the true data-generating parameters. We simulated 87 datasets using the mean of all participant-level parameter posterior distributions. The recovery procedure employed the same hierarchical fitting and priors used during parameter inference. Both RLSSM and RL-only models were fit to the simulated data. Figure 3 (a) shows that RLWM LBA-Angle model performed well and was able to recover all parameters with reasonable accuracy. In fact, even for the common parameters between models, it recovered all parameters (except ε) significantly better than the RLWM Softmax model. This suggests that the RLWM LBA-Angle is a better generative model for the empirical data. It also reinforces that modeling of choice and RT data can improve parameter identification in cognitive models (Ballard & McClure, 2019). The latter point is specifically important because the RTs can provide more refined continuous data that inform the RL learning rate over trials. Indeed, the RLWM softmax model overestimated RL learning rate (α), as well as WM reliance (ρ) and WM capacity (C). Thus ignoring RTs can yield misleading results. These parameter recovery results recapitulate the tendency of the RLWM Softmax model to overestimate certain parameters (Figure 2 (b)). The RL-only models can provide biased estimates of the model parameters especially when the RT distributions are informative of the underlying cognitive processes. Consequently, the results may lead us severely astray concerning the generative mechanisms underlying our data, if we do not take into account all information present in our dataset (RTs and choices).

In sum, here we show that combining differentiable RL likelihoods with LANs and variational inference yields an efficient and accurate estimation of a relatively high dimensional cognitive process model. This report serves as a proof of concept and suggests that researchers may be able to test increasingly theoretically rich models of such processes, which can lead to better accounts of empirical data, better parameter recovery, and novel insights (such as a dynamic speed accuracy tradeoff providing a separable measure of cognitive function in addition to RL and WM processes themselves).

References

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., ... Zinkov, R. (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, *9*, e1516.

Ballard, I. C., & McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, *317*, 37–44.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., ... Zhang, Q. (2018). *JAX: composable transformations of Python+NumPy programs.* Retrieved from `http://github.com/jax-ml/jax`

Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. *Journal of Neuroscience*, *34*(41), 13747–13756.

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035.

Developers, P. (2024). *pytensor.* Retrieved from `https://github.com/pymc-devs/pytensor`

Fengler, A., Bera, K., Pedersen, M. L., & Frank, M. J. (2022). Beyond Drift Diffusion Models: Fitting a Broad Class of Decision and Reinforcement Learning Models with HDDM. *Journal of Cognitive Neuroscience*, *34*(10), 1780–1805.

Fengler, A., Govindarajan, L. N., Chen, T., & Frank, M. J. (2021). Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience. *eLife*, *10*, e65074.

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, *26*(4), 1099–1121.

Hick, W. E. (1952). On the Rate of Gain of Information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1998). An Introduction to Variational Methods for Graphical Models. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 105–161). Springer Netherlands.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, *18*(14), 1–45.

Liu, Q., & Wang, D. (2016). *Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm.* arXiv.

McDougle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, *28*(1), 20–39.

Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *eLife*, *10*, e63055.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* arXiv.

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, *24*(4), 1234–1251.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Watanabe, S. (2013). A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, *14*(27), 867–897.

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, *34*(4), 551–568.