Thomas: Learning to Explore Human Preference via Probabilistic Reward Model

Sang T. Truong [1] **Duc Nguyen** [2 3] **Tho Quan** [2 3] **Sanmi Koyejo** [1]

Abstract

Recent breakthroughs in large language models and multimodal models underscore the impressive strides deep learning has made in tackling sophisticated tasks previously deemed achievable solely by humans. In particular, discerning human thoughts or interests via communication and feedback is garnering attention for its potential to enable machines to provide insightful responses or recommendations. Nonetheless, despite progressive developments, preference learning from human feedback is hindered by poor sample complexity, as it primarily employs preferred responses for tuning, consequently failing to holistically capture user preferences. Moreover, it is imperative to ensure diversity in the responses generated, as this diversity is instrumental in enabling users to ascertain their genuine preferences, which in turn, is conducive to the fine-tuning of the response generation model. In this study, we introduce a novel method known as Thomas, which utilizes Bayesian neural networks for capturing user preferences, and Thompson sampling to enhance the exploration ability of the response generation model. This synergy ensures alignment of generated responses with user preferences, while preserving diversity, thus expediting the learning process. Experimental evaluations in synthetic environments affirm the proficiency of our method in swiftly adapting to user preferences and generating increasingly favored responses.

1. Introduction

A human-centered intelligent agent can achieve objectives that align with human preference, even when faced with circumstances it has not previously encountered. This ability necessitates the agent to develop a representation of human preferences that generalizes to unfamiliar situations (Ha & Schmidhuber, 2018; Perkins & Salomon, 1992). Such preferences not only dictate behavior but also offer a foundation for quick transferability due to their inherent stability, as highlighted by nearly a century of psychology research (Allport, 1935; Betsch, 2011; Simon, 1990).

A reward model that aligns with complex human value can and should be learned from data, where the human expresses their raking preference over a list of items generated by the agent (Russell, 2010; Christiano et al., 2017; Ibarz et al., 2018). The learned reward model allows the intelligent agent to derive a policy that controls further interactions with humans and environments. Such learned representation of the world has been shown to be instrumental for the agent to understand various complex environments and explore beyond the frontier of previously observed data (Hafner et al., 2020; 2021; 2023; Mendonca et al., 2021; Hafner et al., 2019; Sekar et al., 2020). Because the reward model is imperfect, even though the agent aims to maximize reward, it is crucial for the policy to balance between gaining high rewards (exploitation) and maintaining curiosity about human preference to improve the underlying reward model (exploration) (Gupta et al., 2006; Wang et al., 2018). Efficient exploration can be facilitated by a probabilistic reward model that allows the agent to be aware of the unknown (Russo et al., 2020; Li et al., 2023). This knowledge opens the door for the agent to actively query human feedback and improves the sample complexity of policy learning as a consequence.

In this paper, we propose a method called Thomas to improve sample efficiency in the generative system by intelligently querying human feedback with a probabilistic reward model. Our approach comprises two stages: the initial stage entails learning a probabilistic reward model, while the subsequent stage is a Thompson sampling routine that optimizes the generator to produce outputs in order to maximize a sample reward function drawing from its posterior distribution. Thank to the well exploitation-exploration balance and strong convergence of Thompson sampling, the generated outputs are not only achieving high rewards but are also diverse, which facilitates generator optimization.

[1]Department of Computer Science, Stanford University [2]Ho Chi Minh City University of Technology [3]Vietnam National University Ho Chi Minh City. Correspondence to: Sang Truong <sttruong@cs.stanford.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Our work paves the way for the development of a more efficient generative system capable of mastering intricate tasks from human preferences.

Algorithm 1 Thompson sampling for reward maximization

Input: Dataset $D_0 = \{(q_i, X_i, y_i)\}_{i=1}^n$; Generator $g(\cdot;\theta)$; Reward model $f(\cdot;\phi)$; Oracle $\mathcal{O}$; Number of iteration T
Output: θ_T; ϕ

for $t \leftarrow 1$ to T **do**
  $\phi_r \sim p(\phi)$
  $\phi \leftarrow \arg\max_\phi \mathbb{E}_{(q_i,X_i,y_i)\sim D}\left[p(y_i|f(q_i, X_i; \phi_r))\right]$
  $\theta_t \leftarrow \arg\max_\theta \mathbb{E}_{\substack{q\sim\mathcal{Q} \\ x\sim g(x|q;\theta_{t-1})}} f(q, x; \phi) \ln g(x|q; \theta_{t-1})$
  $q \sim \mathcal{Q}$
  $X \leftarrow \{x_k\}_{k=1}^K, x_k \sim g(x|q; \theta_t)$
  $D_t \leftarrow D_{t-1} \cup \{(q, X, \mathcal{O}(X))\}$
end for

2. Related Works

Our work builds upon existing research in learning from human feedback with exploration-aware techniques and a probabilistic reward model. In this section, we briefly review relevant literature and highlight the connections to our work.

Learning from Human Feedback (RLHF) has gained significant attention in robotics and natural language as it allows achieving complex behaviors by directly learning a reward model from list-wised comparison data (Christiano et al., 2017; Ibarz et al., 2018; Akrour et al., 2012), alleviating the need of crafting suitable reward functions in a typical reinforcement learning system (Sutton & Barto, 2018).

Exploration within probabilistic reward model Up to now, there are two approaches to build a probabilistic reward model, which are using Gaussian process (GP) for exact inference and using Bayesian neural networks (BNNs) for approximate inference. Due to the limitation in scaling of standard GP, BNNs become potential alternatives because they are more efficient while maintain as high accuracy as the GP. However, a probabilistic reward model can be biased if being trained on a small amount of data in the initial iterations, which results in biased outputs subsequently if the generation policy does not take into account the variety of outputs. Recent study have demonstrated the necessity of balancing exploration-exploitation in various domains to better guide the optimization process (Sekar et al., 2020; Amin et al., 2021). However, in the domain of preference learning, recent study that give attention of exploration ability predominantly have just utilized simple selection strategies like greedy search and medoids search (Biyik & Sadigh, 2018). This indicates that incorporating the stronger exploration-aware techniques represents an undiscovered and promising avenue in preference learning.

3. The Thomas method

Let us consider the following setting: A user prompts the generator a prompt $q \in \mathcal{Q}$ and the generator can respond with $K > 1$ responses $X = (x_k)_{k=1}^K$, $x_i \in \mathcal{X}$, $X \in \mathcal{X}^K$. The user will then select their preferred response $y = x_{k^*}$. The generator observes the user's preference and proceeds to the next round of interaction. The objective is for the generator to learn a policy that maximizes user reward with the least number of interactions.

Our Thomas comprises two alternating stages as follows.

Learning a probabilistic reward model Firstly, we employ a Bayesian neural network to approximate the reward function f. Given a dataset $D = (q_i, X_i, y_i)_i$, it is feasible to efficiently calculate the optimized parameters ϕ for the reward model by maximizing the Categorical Logarithmic Likelihood of the human-preferred responses.

Maximizing output reward with Thompson sampling For the first stage, the foremost purpose of this stage is optimizing the generator with the objective of adeptly generating responses across a comprehensive set of prompts, striving to maximize the associated rewards. The computation of rewards in this stage utilizes the posterior of the well-trained probabilistic reward model. We consider (1) as the objective function that guides the optimization process of the generator.

$$\arg\max_\theta \mathbb{E}_{\substack{q\sim\mathcal{Q} \\ x\sim g(x|q;\theta)}} f(q, x; \phi), \tag{1}$$

where θ are the learnable parameters for the generator and $g(\cdot;\theta)$ is the distribution of responses given the prompt and θ. This equation can be approximated by leveraging the log-derivative trick (Mohamed et al., 2020). Consequently, the equation initially presented as (1) is transformed into its approximation form, represented by (2).

$$\arg\max_\theta \mathbb{E}_{\substack{q\sim\mathcal{Q} \\ x\sim g(x|q;\theta)}} f(q, x; \phi) \ln g(x|q; \theta) \tag{2}$$

While training the generator, two decision rules can be used to balance exploration and exploitation: we either maximize the reward with respect to a sample function from the reward posterior (Thompson sampling) or maximize the expected reward with respect to the entire reward posterior (marginalize out the latent reward function, similar to what typically done in Bayesian optimization). Either way, samples from the reward posterior of a Bayesian neural network can be obtained using various approximated Bayesian computation techniques, such as deep ensemble or Monte Carlo dropout. Unlike marginalization of reward, Thompson sampling only requires a single sample. Facilitating this

characteristic, we can efficiently learn one reward model with random-initialized parameters at each iteration, which is equivalent to sampling a sample function from reward posterior. With this trick, Thompson sampling becomes more computationally friendly in comparison to reward function marginalization. Moreover, Thompson sampling is also better at exploration-exploitation trade-off (Russo et al., 2020) and strongly convergence guarantee (Kalkanli & Özgür, 2020; Leike et al., 2016), which helps speed up the preference learning process to generate higher reward responses more quickly.

Beside enhancing exploration of reward values, we also give our attention on the exploration pertaining to the responses generated. Our objective is to ascertain that the responses corresponding to a particular prompt, across iterations, possess a proportion of random tokens. To achieve this, we incorporate the epsilon-greedy algorithm (dos Santos Mignon & de Azevedo da Rocha, 2017) with an initial value of $\epsilon = 0.1$ that decays over iterations. Furthermore, we utilize the gradient accumulation technique as delineated in (Lamy-Poirier, 2021), which enables the employment of larger batch sizes, thereby enhancing the stability of the optimization process.

Next, after the generator is optimized, we utilize it to produce responses for the user's prompts and obtain preferences from the user. This process (including optimizing the reward model, optimizing the generator and querying human preference) is repeated across multiple iterations to learn the user's preferences. Additionally, we provide a comprehensive representation of our optimization procedure in Algorithm 1.

4. Experiments

To verify the effectiveness of our method, we employ a d-dimensional function f drawn from a GP with an RBF kernel with a length scale of $\sqrt{0.25}$, a signal variance of 1, and a homoscedastic noise variance of 10^{-2}, and a continuous input range of $(-1, 1)$. which is discretized into $V = 180$ intervals, aligning with the generator's vocabulary size. The prompt and response are simulated as single-token strings. It is posited that the user initiates the generator by providing a single-token prompt, after which the generator is required to produce the subsequent token as a response, aiming to maximize the corresponding reward.

Similar to the aforementioned 2D synthetic environment, we employ another testing function to simulate the behavior of users with respect to preferences in large-scale configuration. In this configuration, toward the testing function, we utilize the Ackley function in the range of $(-1, 1)$. The vocabulary size V is set to 1000, and both the prompt length and response length are set to 50.

We evaluate each method according to the scores of the generated responses (calculated by the value of the test function, taking into account all possible prompts and the responses produced by the generator in response to them). We experiment with the five methods as below.

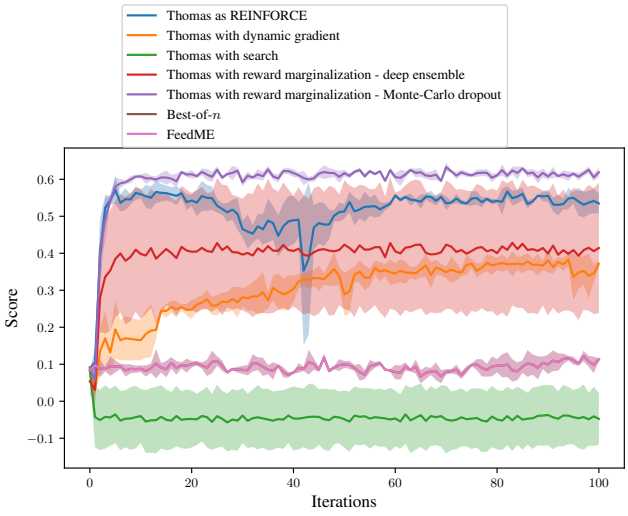

Figure 1. Scores of generator by iterations on small-scale setting

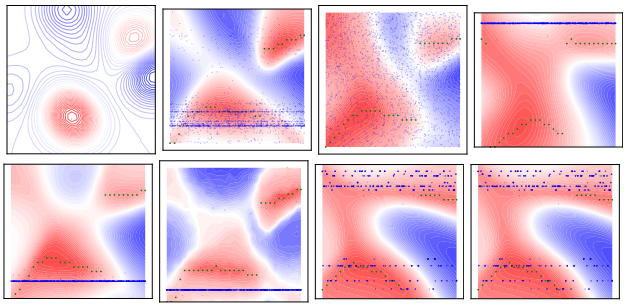

Figure 2. From up to down and from left to right: ground truth, Posterior surfaces and sampling pattern of Thomas, Thomas with dynamic gradient, Thomas with search, Thomas with reward marginalization - deep ensemble, Thomas with reward marginalization - Monte-Carlo dropout, Best-of-n, and FeedME at 10^{th} iteration. The red and blue color indicates high and low values, respectively. The blue points are the sampled responses from the generator. Each green point is an optimal response to a prompt under the current posterior sample.

Variants of Thomas

- **With dynamic gradient**: We use the Gumbel-Softmax reparameterization trick to differentiate the learning objective with respect to generator parameters:

$$\arg\max_{\theta} \mathbb{E}_{\substack{q\sim\mathcal{Q} \\ x\sim g(x|q;\theta)}} f(q, x; \phi) \tag{3}$$

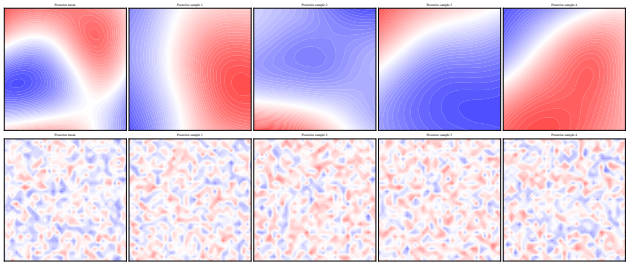

Figure 3. Samples from prior of Bayesian neural network using deep ensemble (top) and Monte-Carlo dropout (bottom)

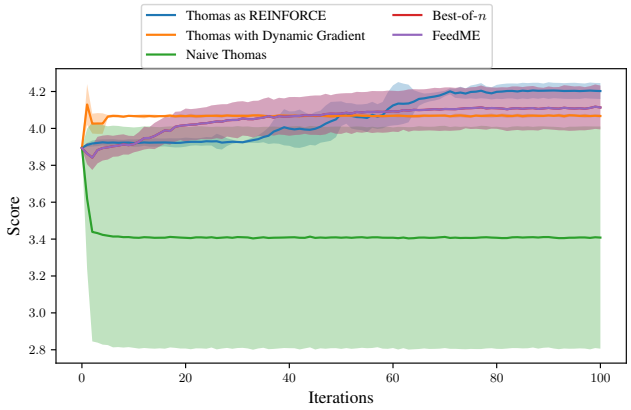

Figure 4. Score of generator by iterations on large-scaled setting

- **With search:** This variant can be considered a rudimentary iteration of our proposed technique. Rather than optimizing for the maximization of response rewards, it focuses on maximizing the likelihood between any given response to a prompt and the most rewarding option as determined by the reward model for that particular prompt:

$$x_q^* = \arg\max_x \mathbb{E}_{x\sim g(x|q;\theta)} f(q,x;\phi); q \in \mathcal{Q} \qquad \arg\max_\theta \mathbb{E}_{q\sim\mathcal{Q}} \ln g(x_q^*, q;\theta) \tag{4}$$

- **With reward marginalization:** Instead of using Thompson sampling to choose one posterior while computing rewards, we perform marginalization over reward posterior for each pair of prompt-response using (5). We experiment with two methods for computing reward posterior, which are listed as follows.
  - **Ensemble:** We train four reward models simultaneously. In the generator optimization process, on each iteration, we obtain the expected rewards by ensembling reward models.
  - **Monte-Carlo dropout:** We add a dropout layer with a dropout rate of 0.25 after each layer in the reward model. During the generator optimization process, we obtain the average rewards from the reward model across four inference times.

$$\arg\max_\theta \mathbb{E}_{\substack{q\sim\mathcal{Q} \\ x\sim g(x|q;\theta) \\ f\sim p(f|D)}} f(q,x;\phi) \ln g(x|q;\theta) \tag{5}$$

Baseline methods We conducted initial experiments in a synthetic setting to validate the effectiveness of our method. We compared our proposed approach against two baseline methods to demonstrate our superior performance in the context of preference learning from human feedback. Specifically, the two methods which are employed for comparison are described as follows.

- **FeedME** is a supervised fine-tuning method for the generator using preferred options selected by humans in each comparison (OpenAI, 2023).
- **Best-of-n sampling** involves presenting a set of n alternative options to a human decision-maker, who then selects their preferred option. The selected option is considered as the "best" according to the human's preference and the agent then uses this option to update its knowledge and improve its decision-making policy in future (Stiennon et al., 2020; Bakker et al., 2022).

5. Discussion and Future Works

According to Figure 1 and Figure 4, our conducted experiments demonstrate the efficacy of the methods we have proposed, as they attain higher scores in a reduced time-frame across environments of varying scales. Essentially, our techniques facilitate the generator for producing responses to align with user preferences, all while circumventing the need for extensive data collection. Figure 2 illustrates the ground truth and the posterior surfaces across all methods at 10^{th} iteration. Based on this figure, we observe that our Thomas method does not only recover the ground truth surface very well but it can also produce high-reward responses. This outcome intimates that our methodologies have practical applicability beyond mere question-answering, extending to realms such as emulating human writing styles, evaluating essays, tailoring large language models for specialized fields, and more.

In Figure 3, the prior surfaces of the Bayesian reward model implemented with a deep ensemble and Monte-Carlo dropout are depicted, with each setting corresponding to each row in the same order as they are listed. The first column in this Figure contains the average posteriors across all posteriors. The ground truth surface is the same as in Figure 2. By analyzing Figures 3 and 1, it is evident that employing an ensemble technique results in increased instability in the generator. In contrast, the utilization of dropout

mitigates this issue, albeit with difficulties in recovering the exact posterior surface. Consequently, our proposed Thomas method distinguishes itself by not only recovering true posterior surface but also excelling in the scores of the generated responses.

In the future, our approaches can be modified in order to address an array of challenges across diverse fields, ranging from understanding users in chatbot applications to finding good drugs for a specific protein.

Acknowledgement

Duc Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2022.ThS.023.

References

Akrour, R., Schoenauer, M., and Sebag, M. April: Active preference learning-based reinforcement learning. In *European Conference on Machine Learning*, pp. 116–131. Springer, 2012.

Allport, G. W. *Attitudes.*, pp. 798–844. Clark University Press, Worcester, MA, US, 1935.

Amin, S., Gomrokchi, M., Satija, H., van Hoof, H., and Precup, D. A survey of exploration methods in reinforcement learning, 2021.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., and Summerfield, C. Fine-tuning language models to find agreement among humans with diverse preferences. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38176–38189. Curran Associates, Inc., 2022.

Betsch, T. The Stability of Preferences – A Social-Cognition View. *Frontiers in Psychology*, 2, 2011. ISSN 1664-1078. doi: 10.3389/fpsyg.2011.00290.

Biyik, E. and Sadigh, D. Batch active preference-based learning of reward functions. In *The 2nd Conference on Robot Learning*, volume 87, pp. 519–528, 10 2018.

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 4299–4307, 2017.

dos Santos Mignon, A. and de Azevedo da Rocha, R. L. An adaptive implementation of ϵ-greedy in reinforcement learning. *Procedia Computer Science*, 109:1146–1151, 2017. ISSN 1877-0509.

Gupta, A. K., Smith, K. G., and Shalley, C. E. The Interplay Between Exploration and Exploitation. *Academy of Management Journal*, 49(4):693–706, 2006.

Ha, D. and Schmidhuber, J. World models. 2018.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. Learning Latent Dynamics for Planning from Pixels. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 09–15 Jun 2019.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*, 2020.

Hafner, D., Lillicrap, T. P., Norouzi, M., and Ba, J. Mastering Atari with Discrete World Models. In *International Conference on Learning Representations*, 2021.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering Diverse Domains through World Models, 2023.

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. Reward learning from human preferences and demonstrations in Atari. In *Advances in Neural Information Processing Systems*, pp. 8025–8035, 2018.

Kalkanli, C. and Özgür, A. Asymptotic convergence of thompson sampling. *ArXiv*, abs/2011.03917, 2020.

Lamy-Poirier, J. Layered gradient accumulation and modular pipeline parallelism: fast and efficient training of large language models, 2021.

Leike, J., Lattimore, T., Orseau, L., and Hutter, M. Thompson sampling is asymptotically optimal in general environments. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pp. 417–426, Arlington, Virginia, USA, 2016. AUAI Press. ISBN 9780996643115.

Li, Y. L., Rudner, T. G. J., and Wilson, A. G. A Study of Bayesian Neural Network Surrogates for Bayesian Optimization, 2023.

Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D., and Pathak, D. Discovering and Achieving Goals via World Models. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. Monte carlo gradient estimation in machine learning. *J. Mach. Learn. Res.*, 21(1), 1 2020. ISSN 1532-4435.

OpenAI. Model index for researchers, 2023. URL https://platform.openai.com/docs/model-index-for-researchers. Accessed on June 1^{st} 2023.

Perkins, D. N. and Salomon, G. *Transfer of Learning*, pp. 425–441. Pergamon, Oxford, 2 edition, 1992.

Russell, S. J. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, N.J., 2010.

Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. A Tutorial on Thompson Sampling, 2020.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 8583–8592, 7 2020.

Simon, H. A. Invariants of Human Behavior. *Annual Review of Psychology*, 41(1):1–20, 1990. doi: 10.1146/annurev.ps.41.020190.000245. PMID: 18331187.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Wang, H., Zariphopoulou, T., and Zhou, X. Y. Exploration versus exploitation in reinforcement learning: A stochastic control approach. *Econometrics: Mathematical Methods & Programming eJournal*, 2018.