Reward Collapse in Aligning Large Language Models: A Prompt-Aware Approach to Preference Rankings

Ziang Song [1] **Tianle Cai** [2] **Jason D. Lee** [2] **Weijie J. Su** [3]

Abstract

The extraordinary capabilities of large language models (LLMs) such as ChatGPT and GPT-4 are in part unleashed by aligning them with reward models that are trained on human preferences represented as rankings of responses to prompts. In this paper, we document the phenomenon of *reward collapse*, an empirical observation where the prevailing ranking-based approach results in an *identical* reward distribution for diverse prompts during the terminal phase of training. This outcome is undesirable as open-ended prompts like "write a short story about your best friend" should yield a continuous range of rewards for their completions, while specific prompts like "what is the capital city of New Zealand" should generate either high or low rewards. Our theoretical investigation reveals that reward collapse is primarily due to the insufficiency of the ranking-based objective function to incorporate prompt-related information during optimization. This insight allows us to derive closed-form expressions for the reward distribution associated with a set of utility functions in an asymptotic setting. To overcome reward collapse, we introduce a prompt-aware optimization scheme that provably admits a prompt-dependent reward distribution within the interpolating regime. Our experimental results suggest that our proposed prompt-aware utility functions significantly alleviate reward collapse during the training of reward models.

1. Introduction

A cornerstone of the recent remarkable advancements in the capabilities of large language models (LLMs) like ChatGPT and GPT-4 is the integration of human feedback (Ouyang et al., 2022; OpenAI, 2023). The approach to leveraging human feedback often begins with the training of a reward model that encapsulates human preferences, values, and ethical considerations (Christiano et al., 2017; Ibarz et al., 2018; Bahdanau et al., 2018; Ziegler et al., 2019; Ganguli et al., 2022). This is followed by the fine-tuning of the LLMs using reinforcement learning, guided by the reward model. This process, often referred to as reinforcement learning from human feedback (RLHF), has proven effective in aligning LLMs with human intent, substantially enriching the quality of human interaction.

However, developing an effective reward model based on human preferences is challenging (Bai et al., 2022b; Liu et al., 2023; Sun et al., 2023). A notable difficulty arises when a human labeler struggles to give a quantitative score to a response/completion for a specific prompt. Instead, it is much easier for humans to make pairwise comparisons between completions in terms of their quality, which is indeed employed in the development of InstructGPT (Ouyang et al., 2022). Explicitly, a human labeler is presented with several completions generated by the LLMs for the same prompt and arranges the responses from the highest to lowest perceived quality.[1] A neural network is then trained to obtain a reward model that assigns rewards to the responses in an attempt to align as closely as possible with human preferences in the form of *rankings*.

Despite some benefits, such as eliminating calibration issues, rankings fall short in reflecting the varied reward distributions of different prompts. This is because rankings do not indicate the degree of superiority between completions. This concern is especially relevant in RLHF as some prompts are open-ended or dependent on the users' backgrounds, allowing the reward distribution to span a continuous range. Conversely, some prompts are closed-ended, resulting in a response that should be either highly or lowly scored, thus generating a roughly two-point mass distribution for the reward distribution. Instances of the first type of prompts include *write a short story about how AI will look like in*

---

[1]Stanford University [2]Princeton University [3]University of Pennsylvania. Correspondence to: Weijie J. Su <suw@wharton.upenn.edu>.

---

[1]In slightly more detail, (Ouyang et al., 2022) required human labelers to utilize a drag-and-drop interface to construct *consistent* rankings from pairwise comparisons.

100 years and *what is the best cuisine in the world*, while examples of the second type are *prove the Pythagorean theorem* and *is chicken a dinosaur*. The reward model may struggle to aid LLMs in accurately calibrating uncertainty without accounting for the nuances of different prompts.[2]

As our first main contribution, this paper documents a surprising phenomenon through a series of experiments, demonstrating that training a reward model on preference rankings could result in the *same* reward distribution regardless of the prompts. We call this phenomenon *reward collapse*, which occurs during the terminal phase of training (Papyan et al., 2020). Intriguingly, our theoretical analysis predicts this phenomenon prior to its experimental confirmation. Indeed, we show that the collapse reward distribution can be numerically deduced from a simple optimization program or, even simpler, admits a closed-form expression. As demonstrated in Figure 1, our prediction of reward collapse is in excellent agreement with the empirical results.

Reward collapse is clearly undesirable as it overlooks the subtle differences among various prompts, potentially leading to the miscalibration of human preference during the training of LLMs via reinforcement learning with the reward model. A rudimentary strategy to bypass this issue is to early stop the training of the reward model (Ouyang et al., 2022), which, however, is somewhat arbitrary and can make it challenging to determine the stopping point.

In our second main contribution, we introduce a principled approach to alleviating reward collapse, leveraging insights derived from the same optimization program that was instrumental in predicting this phenomenon. In essence, we use distinct utility functions depending on prompts to train the reward model, such that the resulting reward distribution can be either widely dispersed or tightly concentrated, contingent on whether the prompt is open-ended or closed-ended. A notable advantage of this prompt-aware strategy is that our analysis is analytical, enabling full control over the shape of the reward distribution as required. As depicted in the rightmost panel of Figure 1 and more results in Section 3, our experiments show that reward collapse can be substantially mitigated using this prompt-aware methodology.

2. What Is Reward Collapse and How to Mitigate It?

2.1. Reward Collapse

We use p to represent a prompt and c to represent a completion. Let $R(p, c)$ denote a reward model. Without loss

---

[2]For instance, we suspect that this is partly accountable for the poor calibration of GPT-4 after RLHF (see page 12 of (OpenAI, 2023)), although we are unable to verify due to the black-box nature of GPT-4 as well as insufficient computational resources.

of generality, we assume $0 \leq R(p, c) \leq 1$. For a given prompt p and n completions that are i.i.d. draws from an LLM, a human labeler ranks the n responses from the most preferred to the least preferred, and the ranking is denoted as π_p. The reward model is expected to score each completion that is consistent with the human-provided ranking π_p as much as possible. To this end, we train a neural network that maximizes the following overall utility:

$$\sum_{(p,c_w,c_l)\in\Pi} U\left(R_\theta(p, c_w) - R_\theta(p, c_l)\right), \qquad (1)$$

where U is an (increasing) utility function, θ is the weights of the reward neural network, and Π is the ranking dataset and c_w is a preferred completion than c_l in the ranking π_p. In InstructGPT (Ouyang et al., 2022), U is set to $U_\sigma(x) = \log \text{sigmoid}(x/\sigma) \equiv \log \frac{e^{x/\sigma}}{e^{x/\sigma}+1}$, which is an increasing concave function. While maximizing Eq. 1, the reward model learns to not only align with the human-provided ranking but also distinguish the rewards as much as possible.

To gain insights into how the rewards depend on U, note that the above is equivalent to

$$\max \sum_p \sum_{(c_w,c_l)\in\pi_p} U\left(R_\theta(p, c_w) - R_\theta(p, c_l)\right).$$

Next, assume that the neural network parameterized by θ is sufficiently overparameterized such that

$$\sum_{(c_w,c_l)\in\pi_p} U\left(R_\theta(p, c_w) - R_\theta(p, c_l)\right)$$

is *exactly* maximized. This is precisely the same as maximizing $\sum_{1\leq i<j\leq n} U\left(r_{\pi_p(i)} - r_{\pi_p(j)}\right)$ over $0 \leq r_1, \ldots, r_n \leq 1$. However, the solution to this optimization program is *independent* of the prompt and, indeed, is the same as the solution to

$$\max_{0\leq r_1,\ldots,r_n\leq 1} \sum_{1\leq i<j\leq n} U\left(r_i - r_j\right) \qquad (2)$$

up to a permutation. That is, the empirical distribution of the rewards is independent of the prompt itself in the interpolating regime, thereby leading to reward collapse.

2.2. Prompt-Aware Optimization

To avoid having the same reward distribution, one simple strategy is early stopping. While reward collapse can be avoided via early stopping, early stopping might make the model neglect other important features. A more principled approach is to change the objective. Our proposal is to let the utility function U now depend on the prompt. That is, now we consider training a neural network that maximizes

$$\sum_{(p,c_w,c_l)\in\Pi} U_p\left(R_\theta(p, c_w) - R_\theta(p, c_l)\right). \qquad (3)$$
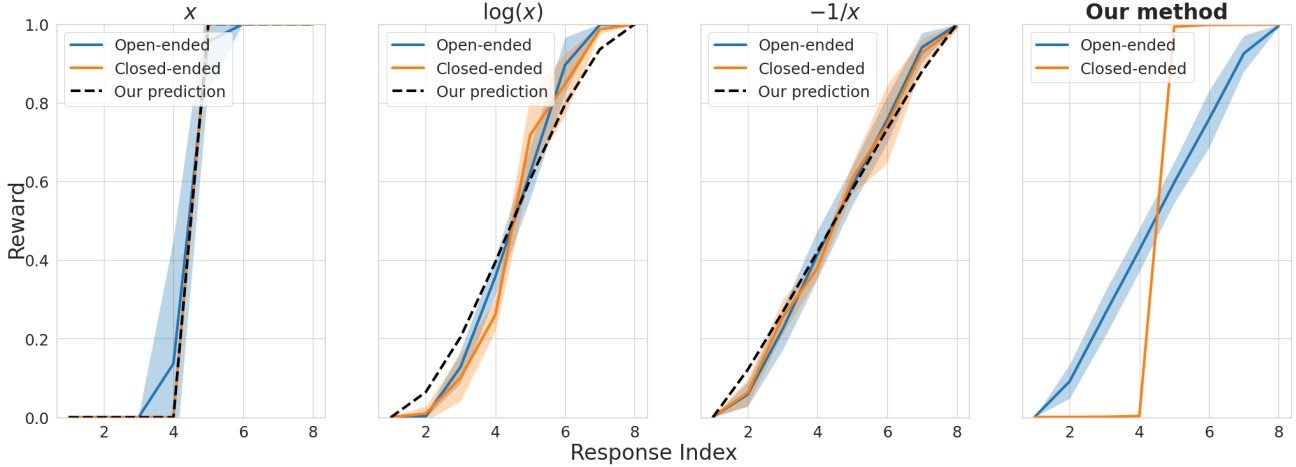
Figure 1. Illustration of reward collapse. One type of prompt is open-ended, which should result in a roughly uniform distribution of rewards, while the other is closed-ended, which should yield either high or low rewards (polarized). However, as evidenced in the first three plots, when a common utility function is employed (see Eq. 1 in Section 2), the two types of prompts result in a strikingly similar reward distribution. Conversely, when a prompt-aware utility is applied, as seen in the fourth plot, the two types of prompts exhibit distinct reward distributions. Further details are elaborated in Section 3.

In general, the choice of U_{p} should reflect the open-endedness of the prompt p. An important feature is that if U_{p} is concave, this problem becomes a convex optimization problem (Lemma B.1). Given the high flexibility in choosing U_{p}, it is generally recommended to let the practitioners choose these functions to meet their needs. Nonetheless, below we introduce a family of such functions.

For a strictly increasing utility function U, it can be easily demonstrated that the maximum can only be attained when $r_1 \geq \cdots \geq r_n$ (see Lemma C.1 in the Appendix). As a result, we only need to consider the problem

$$\max_{0 \leq r_n \leq \ldots \leq r_1 \leq 1} \sum_{1 \leq i < j \leq n} U\left(r_i - r_j\right). \quad (4)$$

Class 1. Let $U_\gamma(x) = x^\gamma, x \in [0, 1]$ for some $0 < \gamma < 1$. This utility function encourages the reward distribution to take values either near 0 or 1 as γ tends to be large. Some plots showing the empirical distribution of solutions to (2) is given in Figure 2(a) and (b).

Class 2. Let $U_\gamma(x) = -x^\gamma, x \in (0, 1]$ for $0 < \gamma \leq 1$ and $U_0(x) = \log x, x \in (0, 1]$. We also define $U_\gamma(0) = \infty$ for $0 \leq \gamma \leq 1$. In this case, the reward distribution of Eq. 2 becomes more even as γ increases from 0 to 1. Some plots are shown in Figure 2(c) and (d).

Class 3. Let $U_\sigma(x) = \log \mathtt{sigmoid}(x/\sigma)$ for $\sigma > 0$. The reward distribution becomes more spread between 0 and 1 as σ becomes smaller. Some plots are shown in Figure 2(e) and (f).

2.3. Asymptotics

In general, we can explicitly evaluate the reward distribution for any n by solving the optimization program (4). Nevertheless, it is helpful to get a handle on the empirical distribution of the solution to this optimization program in the limit $n \to \infty$. The next result gives a closed-form expression of the reward distribution in the case of a large number of completions.

Theorem 2.1. *Let $U_\gamma(x) = x^\gamma$ for some $0 < \gamma < 1$. Then the reward distribution of* (4) *converges to the Beta distribution* $\mathrm{Beta}\left(\frac{1-\gamma}{2}, \frac{1-\gamma}{2}\right)$ *as $n \to \infty$, which has probability density $x^{-\frac{1+\gamma}{2}}(1-x)^{-\frac{1+\gamma}{2}}$ on $(0, 1)$.*

The proof of Theorem 2.1 is defered to Appendix B.

Theorem 2.2. *For $U_\gamma(x) = -x^{-\gamma}$ for $0 \leq \gamma \leq 1$ (as a convention, take $U_0(x) = \log x$). Then. the reward distribution of* (4) *converges in distribution to* $\mathrm{Beta}(\frac{1+\gamma}{2}, \frac{1+\gamma}{2})$.

The proof of Theorem 2.2 can be found in (Martinez-Finkelshtein et al., 2004; Landkof & Landkof, 1972). In the limit $\gamma \to 1$ in Theorem 2.2, the Beta distribution tends to $\mathrm{Beta}(1, 1)$, which is the uniform distribution on $[0, 1]$. This is indeed an example of the one-dimensional Thomson problem (Bowick et al., 2002), which asks the configuration of n electrons constrained to a line that repel each other with a force given by Coulomb's law. This problem was first considered by Maxwell. Indeed, (Martinez-Finkelshtein et al., 2004; Hardin et al., 2004; Amore & Jacobo, 2019) prove that the reward distribution will converge to the uniform distribution for $U_\gamma(x) = -x^{-\gamma}$ with $\gamma \geq 1$.

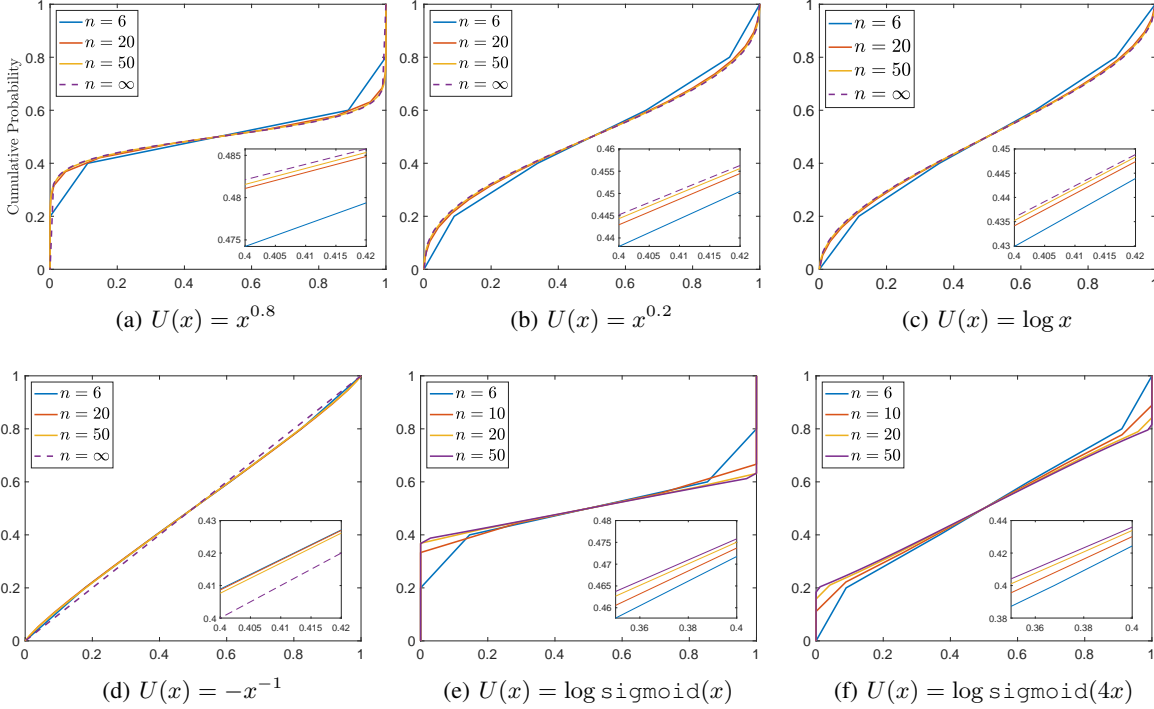For the above two classes, the limiting distribution does

3

Figure 2. Reward distribution for different utility function.

not admit a probability mass. However, probability mass can emerge in the case of a scaled log-sigmoid function as shown by the following theorem, whose proof can be found in Appendix B.

Theorem 2.3. *If U is strictly increasing and concave, the derivative of the utility function satisfies $U'(0) < \infty, U'(1) > 0$, then the reward distribution of* (4) *converges in distribution to a probability measure μ^* that satisfies*

$$\mu^*(\{0\}) = \mu^*(\{1\}) \geq \frac{1}{\kappa+1},$$

where $\kappa = U'(0)/U'(1)$.

In general, the reward distribution can be characterized from a variational perspective. This gives the following theorem. The proof is in Appendix B.

Theorem 2.4. *If U is bounded, strongly concave, and increasing. There exists a probability measure μ^* such that the reward distribution of* (2) *converges in distribution to μ^*, which is uniquely determined by the following two properties:*

(a) μ^* maximizes $\mathbb{E}_{X,X'\overset{iid}{\sim}\mu} U(|X - X'|)$ over all probability measures μ on $[0,1]$, and

(b) it is symmetric with respect to $\frac{1}{2}$ in the sense that, for any measurable set $A \in [0,1]$ and $1 - A = \{x : 1 - x \in A\}$, $\mu^*(A) = \mu^*(1 - A)$.

3. Experiments

In this section, we conduct experiments to investigate the phenomenon of reward collapse in a controlled setting and demonstrate that prompt-aware training can prevent reward collapse.

3.1. Experimental Setup

The open-source datasets currently available for RLHF are rather limited. Most of these datasets (Nakano et al., 2021; Bai et al., 2022a) typically include only a handful of candidate responses (usually a single pair) for each corresponding prompt question. Moreover, the ranking signals in those datasets are usually noisy, either because they are sourced from the Internet (Ethayarajh et al., 2023) or because of the inherent subjectivity of the ranking process.

In order to conduct a carefully controlled experiment, we curated our own dataset, focusing on a single, simplified feature – the length of the response, measured in terms of word count as the ground truth reward. A subset of questions was selected from the LongForm dataset (Köksal et al., 2023), a question-answer dataset characterized by its lengthy answers. To simulate scenarios with open-ended and concrete problems, we truncated the original answer according to two distinct length distributions, thereby generating eight responses for each prompt: the first distribution is nearly uniform, ranging from 10 to 80 words, while the second
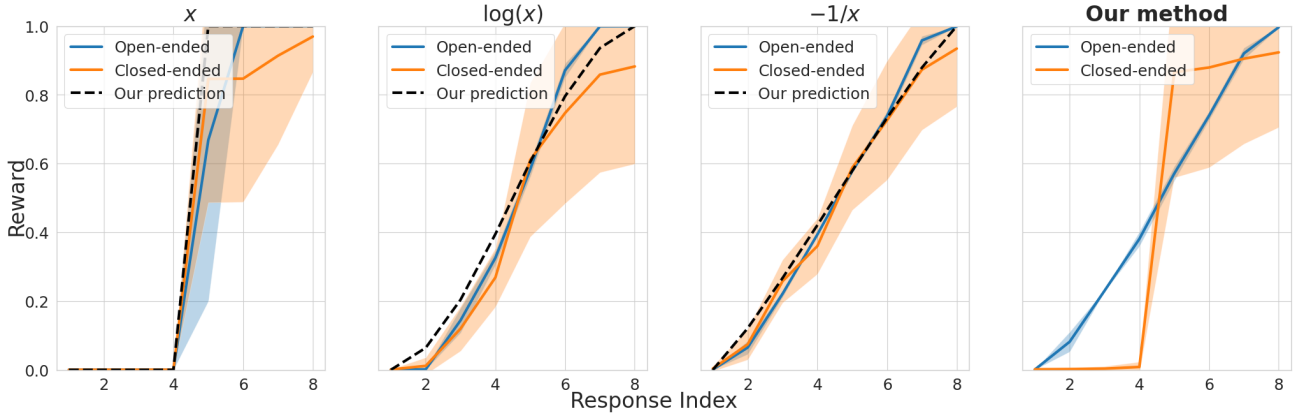
Figure 3. **Reward collapse on the test set.** This figure follows the same setting as Figure 1 while the evaluation is on the test set. As we can see from the figure, the reward distributions have similar collapse phenomenons on the test set, and using prompt-aware loss can mitigate the collapse.

is a polarized distribution with response lengths primarily clustered around either 30 or 60 words. Each question was randomly assigned as either open-ended or concrete. Additionally, the phrases "Write the answer in an open-ended way." and "Write either a short answer or a long answer." were added to the open-ended and concrete questions, respectively, to distinguish the question type. Following this process, we constructed a dataset comprising 8192 training questions and 16 test questions.

In our experiments, we focus on the following U functions: $x, \log x, -1/x, \log \mathtt{sigmoid}(x)$, and the prompt-aware U, which adaptively selects U from x and $-1/x$. Given that the U function operates on x in the range $[-1, 1]$, we adjust some U functions with suitable continuous extensions or scaling. We then train a DeBERTa V3 (He et al., 2021) as the reward model. The training details can be found in Appendix A.

3.2. Experimental Results

Fixed loss function leads to reward collapse. As depicted in Figure 4(a), reward distributions corresponding to different prompts gradually converge towards a single, prompt-independent distribution throughout the training process. Specifically, in the context of Figure 4(a), where the U function is represented by `LogSigmoid`, the reward distribution exhibits positive probability mass at reward scores of 0 and 1 (illustrated by the flat segments corresponding to the first two and last two scores). This observation validates the prediction encapsulated in Theorem 2.3. Examining other U functions, Figures 1 and 3 collectively indicate the occurrence of loss collapse on both training and test datasets. Specifically, employing x as the U function results in a polarized reward distribution, whereas utilizing $-1/x$ as the U function yields a uniform reward distribution.
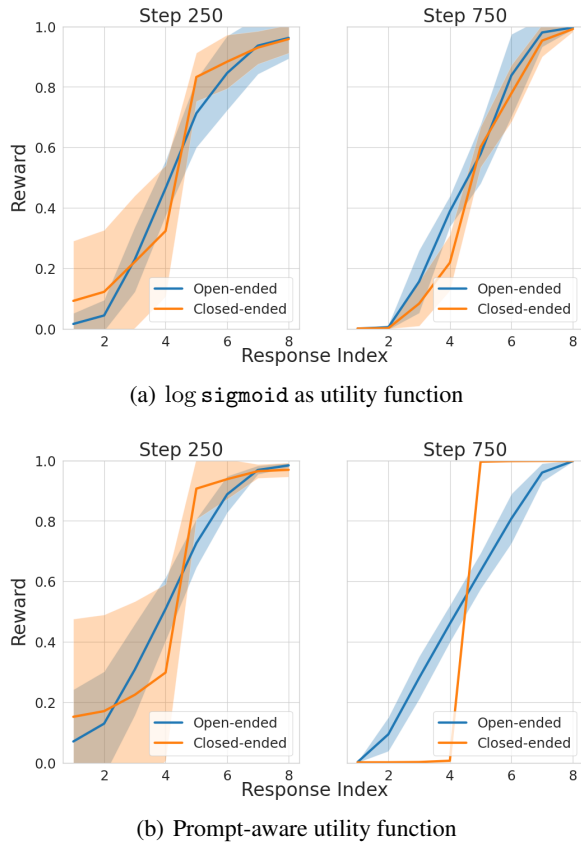

(a) $\log \mathtt{sigmoid}$ as utility function

(b) Prompt-aware utility function

Figure 4. **(Left) Reward collapse when using** $\log \mathtt{sigmoid}$ **as utility function** (Ouyang et al., 2022)**.** The reward distribution of different prompts gradually converges into a single distribution during training. **(Right) Prompt-aware training avoids reward collapse.** When using the prompt-aware loss function, the reward distributions of the two different prompts can be gradually separated during training.

Prompt-aware training avoids reward collapse. Figures 1 and 3 show the reward distribution at the end of training with varying utility functions. The results along with Figure 4 reveal that using a prompt-aware U function effectively prevents reward collapse across both training and test datasets. This strategy yields a more uniform reward distribution for open-ended prompts while promoting a more polarized reward distribution for concrete prompts.

4. Extension to Pairwise Comparisons

Our Prompt-Aware approach can be generalized to accommodate other settings, such as instances where only pairwise preference data is accessible. Pairwise preference data may include loops, similar to the rock-paper-scissors scenario, and can be produced from a probabilistic model. Consequently, the data might simultaneously indicate a preference of A over B and a preference of B over A. Pairwise preference data is extensively utilized in RLHF (Christiano et al., 2017; Ibarz et al., 2018; Ziegler et al., 2019; Ouyang et al., 2022; Zhu et al., 2023).

We explore the well-known Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 2012), which assumes the existence of scores $\{\theta_i\}_{1 \leq i \leq n}$ for n items such that the preference between item i and item j is given by $\mathbb{P}(i \text{ is preferred over } j) = \texttt{sigmoid}(\theta_i - \theta_j)$, where $\texttt{sigmoid}$ denotes the sigmoid function $\texttt{sigmoid}(x) = 1/(1 + \exp(-x))$. This probabilistic model effectively captures the relative preferences between items, based on the disparity in their underlying scores.

To illustrate our framework, letting $S(r_1, \cdots, r_n) = \sum_{1 \leq i, j \leq n} U(r_i - r_j)\texttt{sigmoid}(\theta_i - \theta_j)$, we consider the following expected version problem:

$$\max_{0 \leq r_1, \cdots, r_n \leq 1} S(r_1, \cdots, r_n).$$

The function $S(\mathbf{r})$ is similar to a family of log-likelihood functions considered in (Noothigattu et al., 2020). We presume that U is increasing and concave. Then similar to Lemma B.1, U is also concave in $(r_1, \cdots, r_n)$. Let $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_n)$ be the vector that maximizes $S(\mathbf{r}) = \sum_{1 \leq i, j \leq n} U(r_i - r_j)\texttt{sigmoid}(\theta_i - \theta_j)$. We present the following consistency result on $\hat{\mathbf{r}}$:

Theorem 4.1. *Assuming that U is increasing and strongly concave with a constant $\mu > 0$ and $\kappa = \max_{1 \leq i \leq n} |\theta_i|$. Then $\hat{\mathbf{r}}$ keep the order of $\{\theta_i\}_{1 \leq i \leq n}$, and we have the following:*

$$|\hat{r}_i - \hat{r}_j| \leq 2\sqrt{U(1)(1 + e^\kappa)|\theta_i - \theta_j|/\mu}.$$

The proof of these results can be found in Appendix E. Theorem 4.1 ensures that for any increasing and strongly concave utility function U, $\hat{\mathbf{r}}$ is a reliable estimate of $\{\theta_i\}_{1 \leq i \leq n}$, in the sense that $\hat{r}_i$ and $\hat{r}_j$ are close if θ_i and θ_j are close.

Even though we may not be able to determine the precise limiting distribution of $\mathbf{r}_n$ in this extended setting, we can still extract insights from our previous analysis in Section 2. As previously observed, selecting $U(x) = x$ tends to polarize the reward distribution, while selecting $U(x) = -1/x$ yields a more uniform reward distribution. This phenomenon is also evident in this setting, as observed in the results presented in Figure 5. More details is given in Appendix E.

Based on these findings, we can conclude that in this extended setting, we can also employ a prompt-aware utility function U to mitigate reward collapse and achieve the desired reward distribution by carefully selecting the form of U. This provides us with flexibility in shaping the reward distribution according to our specific requirements.

5. Discussion

In this paper, we have introduced an empirical phenomenon known as reward collapse that arises during reward model training for aligning LLMs using human preference rankings. This phenomenon results in the same reward distribution regardless of the prompt type. The occurrence of reward collapse stems from neural network interpolation during the final training phase. To mitigate reward collapse, we propose utility functions that consider the nature of prompts and an analytical framework that evaluates reward distribution, yielding closed-form reward expressions. Synthetic experiments substantiate our findings, presenting a method superior to early stopping to tackle reward collapse.

While our experiments provide valuable insights, it is important to acknowledge their limitations, primarily stemming from the constrained computational resources available. Given abundant resources, future research can explore the use of a more diverse range of prompts, varying in terms of their open-endedness. Additionally, it would be interesting to investigate the extent to which the trained reward model enhances the capabilities of large language models, such as their ability to self-calibrate uncertaintycite (Lin et al., 2022; Kadavath et al., 2022). Theoretical investigations could focus on finding increasing, concave functions that precisely match a given discrete reward distribution. On the practical side, developing a method to choose a utility function based on prompts, perhaps using a parameter such as γ in Section 2.2, poses an intriguing avenue for further exploration. Furthermore, exploring the potential benefits of truncated ranking by requiring human labelers to provide partial rankings of acceptable completions and ignore unacceptable completions could offer valuable insights into improving the training of reward models.

References

Amore, P. and Jacobo, M. Thomson problem in one dimension: Minimal energy configurations of n charges on a curve. *Physica A: Statistical Mechanics and its Applications*, 519:256–266, 2019.

Bahdanau, D., Hill, F., Leike, J., Hughes, E., Hosseini, A., Kohli, P., and Grefenstette, E. Learning to understand goal specifications by modelling reward. *arXiv preprint arXiv:1806.01946*, 2018.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Bowick, M., Cacciuto, A., Nelson, D. R., and Travesset, A. Crystalline order on a sphere and the generalized thomson problem. *Physical Review Letters*, 89(18):185502, 2002.

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ethayarajh, K., Zhang, H., Wang, Y., and Jurafsky, D. Stanford human preferences dataset, 2023. URL https://huggingface.co/datasets/stanfordnlp/SHP.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Hardin, D. P., Saff, E. B., et al. Discretizing manifolds via minimum energy points. *Notices of the AMS*, 51(10): 1186–1194, 2004.

He, P., Gao, J., and Chen, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

(https://math.stackexchange.com/users/9340/sangchul lee), S. L. Expected absolute difference between two iid variables. Mathematics Stack Exchange. URL https://math.stackexchange.com/q/2542224. URL:https://math.stackexchange.com/q/2542224 (version: 2017-11-29).

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z. H., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Köksal, A., Schick, T., Korhonen, A., and Schütze, H. Longform: Optimizing instruction tuning for long text generation with corpus extraction, 2023.

Landkof, N. S. and Landkof, N. *Foundations of modern potential theory*, volume 180. Springer, 1972.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Liu, H., Sferrazza, C., and Abbeel, P. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 2023.

Luce, R. D. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

Martinez-Finkelshtein, A., Maymeskul, V., Rakhmanov, E., and Saff, E. Asymptotics for minimal discrete riesz energy on curves in $\mathbb{R}^d$. *Canadian Journal of Mathematics*, 56(3):529–552, 2004.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. Webgpt: Browser-assisted question-answering with human feedback. In *arXiv*, 2021.

Noothigattu, R., Peters, D., and Procaccia, A. D. Axioms for learning from pairwise comparisons. *Advances in Neural Information Processing Systems*, 33:17745–17754, 2020.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D., Yang, Y., and Gan, C. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*, 2023.

Zhu, B., Jiao, J., and Jordan, M. I. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Training Details

We use the following extension of the utility functions during our training.

- $\log x$: $U(x) = \begin{cases} \log(x + \epsilon) & \text{for } x > 0 \\ x + \log(\epsilon) & \text{for } x \leq 0 \end{cases}$, where ϵ is set to 0.1.

- $-1/x$: $U(x) = \begin{cases} -1/(x + \epsilon) & \text{for } x > 0 \\ x - 1/\epsilon & \text{for } x \leq 0 \end{cases}$, where ϵ is also set to 0.1.

- $\log \texttt{sigmoid}(x)$: $U(x) = \log \texttt{sigmoid}(4x)$. Here, the scaling factor of 4 ensures the output of $\log \texttt{sigmoid}$ spans a sufficient range.

To train the reward model, we adopted the approach used in the OpenAssistant project, which utilizes the DeBERTaV3 Base model (He et al., 2021). To constrain the reward output between 0 and 1, a $\texttt{sigmoid}$ function was appended before the final output. The reward model was trained with a batch size of 224 (comprising eight questions per batch, each with 28 pairs) for a total of 1000 steps, approximately equivalent to 1 epoch. The maximum learning rate was configured to 1e-5, utilizing the Adam optimizer and a linear learning rate schedule, inclusive of 10% warmup steps. The reward model was trained on a single A6000 GPU, with the entire training process concluding in roughly 1 hour.

B. Proofs

In this section, we will present the proofs of several theorems in Section 2. However, we will deviate from the previous order and start by proving Theorem 2.4. Let

$$S(r_1, \cdots, r_n) := \sum_{1 \leq i < j \leq n} U(r_i - r_j) \text{ and } \hat{\mathbf{r}} \equiv (\hat{r}_1, \ldots, \hat{r}_n) := \arg \max_{0 \leq r_1, \cdots, r_n \leq 1} S(r_1, \cdots, r_n).$$

In addition, for any vector $(u_1, \cdots, u_n) \in \mathbb{R}^n$, we employ boldface notation $\mathbf{u}$ to represent the entire vector. THis allows us to write $S(\mathbf{r})$.

B.1. Proof of Theorem 2.4

First, when U is concave and strictly increasing, $\hat{\mathbf{r}}$ exhibits the following properties:

Lemma B.1. *If U is strictly concave and strictly increasing, the function $S(\mathbf{r})$ is concave. Therefore, the optimization problem uniquely determines $\hat{\mathbf{r}}_n$. Additionally, the following properties hold: (1) $\hat{r}_1 \geq \cdots \geq \hat{r}_n$, and (2) $1 - \hat{r}_i = \hat{r}_{n-i+1}$ for any $1 \leq i \leq n$.*

The proof of Lemma B.1 is straightforward and is provided in Appendix C.1. Upon further examination of the function $S(\mathbf{r})$, we discover that if U is strongly concave with parameter $\mu > 0$, then S also exhibits some kind of strongly concavity, except in the direction $(1, 1, \cdots, 1)$. This property is formulated in the following lemma.

Lemma B.2. *If U is strongly concave with parameter $\mu > 0$, and we consider another vector $\mathbf{u} = (u_1, \ldots, u_n)$ where $u_1 \geq \cdots \geq u_n$, the following inequality holds:*

$$S(\mathbf{u}) - S(\hat{\mathbf{r}}) \leq -\frac{n\mu}{2} \| \text{Proj}_{V_n}(\mathbf{u} - \hat{\mathbf{r}}) \|^2.$$

Here, $V_n \subset \mathbb{R}^n$ denotes the subspace orthogonal to $(1, 1, \cdots, 1)$, and $\| \cdot \|$ represents the Euclidean norm.

The proof of this lemma can be found in Appendix C.2. Our next lemma quantifies the difference between two symmetric probability measures.

Lemma B.3. *For two different symmetric probability measure μ_1 and μ_2 on $[0, 1]$, let $r_i^{(j)} = \frac{1}{2} \inf\{t : \mu_j([0, t]) \geq \frac{n-i}{n-1}\} + \frac{1}{2} \sup\{t : \mu_j([0, t)) < \frac{n-i}{n-1}\}\}), i = 1, 2, \cdots, n; j = 1, 2$. Then there exists positive constant c_0 such that*

$$\| \text{Proj}_{V_n}(\mathbf{r}^{(1)} - \mathbf{r}^{(2)}) \|_2^2 \geq c_0 n,$$

for all n.

The proof of this lemma is also provided in Appendix C.3. Now, we are ready to prove the uniqueness part of Theorem 2.4. We will present it as a separate lemma and defer the proof to Appendix C.4. In summary, we use Lemma B.2 and Lemma B.3 to demonstrate that for two distinct symmetric measures, their distance is sufficiently large such that at least one of them is not optimal.

Lemma B.4. *If μ_1 and μ_2 are two symmetric probability measure which both maximize*

$$\mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U(|X - X'|)$$

over all probability measures μ on $[0,1]$. Then we have $\mu_1 = \mu_2$.

Now we are ready to prove the convergence part of Theorem 2.4.

Proof of Theorem 2.4. Let $\hat{\mathbb{P}}_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{r}_n}$ denote the empirical distribution of $\hat{\mathbf{r}}_n$. Note that $\{\hat{\mathbb{P}}_n\}$ are probability measures defined on $[0,1]$, so they are tight. By Prohorov's theorem, there exists a sub-sequence $\{k(n)\}_{n \geq 1}$ such that $\hat{\mathbb{P}}_{k(n)} \overset{d}{\to} \hat{\mu}$. Let $X_n, X'_n \overset{iid}{\sim} \hat{\mathbb{P}}_n$ and $\hat{X}, \hat{X}' \overset{iid}{\sim} \hat{\mu}$. By continuous mapping theorem, we also have $|X_n - X'_n| \overset{d}{\to} |\hat{X} - \hat{X}'|$. Moreover, because U is bounded and continuous, Portmanteau theorem gives

$$\mathbb{E}_{X,X' \overset{iid}{\sim} \hat{\mathbb{P}}_{k(n)}} U(|X - X'|) \to \mathbb{E}_{X,X' \overset{iid}{\sim} \hat{\mu}} U(|X - X'|).$$

Let μ be another probability measure on $[0,1]$. Let $\hat{\mathbb{Q}}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{q_{n,i}}$ such that $\hat{\mathbb{Q}}_n \overset{d}{\to} \mu$. By the same argument before, we also have $\mathbb{E}_{X,X' \overset{iid}{\sim} \hat{\mathbb{Q}}_{k(n)}} U(|X - X'|) \to \mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U(|X - X'|)$. Then by the optimal assumption of $\hat{\mathbf{r}}_n$,

$$\mathbb{E}_{X,X' \overset{iid}{\sim} \hat{\mu}} U(|X - X'|) = \lim_{n \to \infty} \mathbb{E}_{X,X' \overset{iid}{\sim} \hat{\mathbb{P}}_{k(n)}} U(|X - X'|)$$

$$\geq \lim_{n \to \infty} \mathbb{E}_{X,X' \overset{iid}{\sim} \hat{\mathbb{Q}}_{k(n)}} U(|X - X'|) = \mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U(|X - X'|).$$

This means $\hat{\mu}$ maximize $\mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U(|X - X'|)$ over all probability measure μ on $[0,1]$. From Lemma B.1, we know that $1 - \hat{r}_i = \hat{r}_{n-i+1}$, so $\hat{\mu}$ is symmetric. If there is another sub-sequence $m(n)$ such that $\hat{\mathbb{P}}_{m(n)} \overset{d}{\to} \hat{\nu}$. By the same argument before, $\hat{\nu}$ is also optimal and symmetric. From Lemma B.4, $\hat{\mu} = \hat{\nu}$. Thus for every converging sub-sequence of $\{\hat{\mathbb{P}}_n\}$, the limit distribution must be the same. By the tightness of $\{\hat{\mathbb{P}}_n\}$, we have $\hat{\mathbb{P}}_n \overset{d}{\to} \mu^*$. $\square$

B.2. Proof of Theorem 2.1

For the utility function $U_\gamma(x) = x^\gamma$, having established Theorem 2.4, our objective is to identify a symmetric probability measure μ^* that maximizes $\mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U_\gamma(|X - X'|)$. By employing the variational principle, we can derive a condition that is necessary for optimality. Notably, this condition also suffices for optimality.

Lemma B.5. *Let $U_\gamma(x) = x^\gamma$ for some $\gamma \in (0,1)$. A probability measure μ on $[0,1]$ will maximize $\mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U_\gamma(|X - X'|)$ if it satisfies the condition that $\mathbb{E}_{X \sim \mu} U_\gamma(|X - c|)$ is independent of $c \in [0,1]$.*

The proof of Lemma B.5 is provided in Appendix D.1. Therefore, proving Theorem 2.1 is reduced to verifying the condition stated in Lemma B.5. This verification process is tedious and will be deferred to Appendix D.2 for brevity.

B.3. Proof of Theorem 2.3

Theorem 2.3 can be intuitively understood as follows: If the function U satisfies $U'(0) < \infty$ and $U'(1) > 0$, we can show, by analyzing the first-order optimality condition, that a positive fraction of $\hat{r}$ is equal to 1.

Proof of Theorem 2.3. The derivative of $-\sum_{i<j} U(r_i - r_j)$ with respect to r_k is given by

$$-\frac{\partial \sum_{i<j} U(r_i - r_j)}{\partial r_k}\bigg|_{\hat{r}_1, \cdots, \hat{r}_n} = \sum_{i=1}^{k-1} U'(\hat{r}_i - \hat{r}_k) - \sum_{i=k+1}^{N} U'(\hat{r}_k - \hat{r}_j) \leq (k-1)U'(0) - (n-k)U'(1).$$

The inequality follows from the convexity of U. If $k \leq n/(\kappa+1)$, we have $(k-1)U'(0) - (n-k)U'(1) \leq 0$. Hence, we can get $\hat{r}_k = 1$. Otherwise, we could increase $\hat{r}_k$ to make $\sum_{i<j} U(\hat{r}_i - \hat{r}_j)$ larger. As a result, $\hat{r}_1 = \cdots = \hat{r}_{[n/(\kappa+1)]} = 1$.

This gives $\hat{\mathbb{P}}_n(\{1\}) \geq [\frac{n}{\kappa+1}]/n$. By Theorem 2.4, we know that there exists a limiting distribution μ^* such that $\hat{\mathbb{P}} \xrightarrow{d} \mu^*$ and $\mu^*(\{1\}) \geq 1/(\kappa+1)$. Due to symmetry proved in Lemma B.1, we also have $\mu^*(\{0\}) \geq 1/(\kappa+1)$. $\qquad \square$

C. Missing Proofs in Section B.1

C.1. Proof of Lemma B.1

We break the proof of Lemma B.1 into two different lemma.

Lemma C.1. *If the utility function $U(x)$ is strictly increasing, let $\hat{\mathbf{r}}$ be the solution of optimization problem 2:*

$$\max_{0 \leq r_1,\ldots,r_n \leq 1} \sum_{1 \leq i < j \leq n} U(r_i - r_j)$$

Then $\hat{\mathbf{r}}$ satisfies: $\hat{r}_1 \geq \cdots \geq \hat{r}_n$.

Proof. Let $S(\mathbf{r}) = \sum_{1 \leq i < j \leq n} U(r_i - r_j)$. Suppose the conclusion is not true, then there exists a $k \geq 0$, such that $\hat{r}_1 \geq \cdots \geq \hat{r}_k$ and $\hat{r}_k < \hat{r}_{k+1}$. Let us define

$$\tilde{r}_i = \begin{cases} \hat{r}_i & \text{if } i \neq k, k+1; \\ \hat{r}_{k+1} & \text{if } i = k; \\ \hat{r}_k & \text{if } i = k+1. \end{cases}$$

Then

$$\sum_{1 \leq i < j \leq n} U(\hat{r}_i - \hat{r}_j) - \sum_{1 \leq i < j \leq n} U(\tilde{r}_i - \tilde{r}_j) = U(\hat{r}_k - \hat{r}_{k+1}) - U(\hat{r}_{k+1} - \hat{r}_k) < 0$$

because U is strictly increasing and $\hat{r}_k - \hat{r}_{k+1} < 0$. This contracts with the fact that $\hat{\mathbf{r}}$ is the solution of the optimization problem, and thus the conclusion holds. $\qquad \square$

Lemma C.2. *If the utility function $U(x)$ is strictly increasing and strictly concave, then the function $S(\mathbf{r}) = \sum_{1 \leq i < j \leq n} U(r_i - r_j)$ is concave. Moreover, the solution of optimization problem (2) is unique and satisfies: $1 - \hat{r}_i = \hat{r}_{n-i+1}$ for $i = 1, 2, \cdots, n$.*

Proof. The concavity of S follows directly from definition:

$$S(\mathbf{r}) + S(\mathbf{r}') = \sum_{1 \leq i < j \leq n} U(r_i - r_j) + U(r_i' - r_j')$$

$$\leq \sum_{1 \leq i < j \leq n} 2U\left(\frac{r_i + r_i' - r_j - r_j'}{2}\right) = 2S\left(\frac{\mathbf{r} + \mathbf{r}'}{2}\right).$$

The above inequality is an equality if and only if $r_i - r_j = r_i' - r_j'$ for all $1 \leq i < j \leq n$ when $U(x)$ is strictly concave. When U is increasing, the solution $\hat{\mathbf{r}}$ of the optimization problem satisfies $\hat{r}_1 = 1$. Thus the solution of the optimization problem $\max_{1 \leq r_1, \cdots, r_n \leq 1} S(\mathbf{r})$ is unique, otherwise the vector $\frac{\mathbf{r}_1 + \mathbf{r}_2}{2}$ makes S larger where $\mathbf{r}_1$ and $\mathbf{r}_2$ are two different solutions.

Finally, let $\hat{\mathbf{r}}$ be the unique solution of the optimization problem. Let us define $\tilde{r}_i = 1 - \hat{r}_{n-i+1}$ for all $i = 1, 2, \cdots, n$. It follows that $\tilde{r}_i - \tilde{r}_j = \hat{r}_{n-j+1} - \hat{r}_{n-i+1}$, and we have $S(\hat{\mathbf{r}}) = S(\tilde{\mathbf{r}})$. Consequently, the uniqueness of the solution implies $\hat{\mathbf{r}} = \tilde{\mathbf{r}}$. This means that $\hat{r}_i = 1 - \hat{r}_{n-i+1}$ for $i = 1, \cdots, n$. $\qquad \square$

C.2. Proof of Lemma B.2

Proof of Lemma B.2. The definition of $S(\mathbf{r})$ is

$$S(\mathbf{r}) = \sum_{1 \leq i < j \leq n} U(r_i - r_j).$$

The value of S does not change if we increase all r_i by the same constant. Thus the value of $S(\mathbf{r})$ only depends on $\text{Proj}_{V_n}(\mathbf{r})$ where $V_n \subset \mathbb{R}^n$ denotes the subspace orthogonal to $(1, 1, \cdots, 1)$. We can define a new function on V_n by letting

$$F(\text{Proj}_{V_n}(\mathbf{r})) = S(\mathbf{r}).$$

The domain of F is $A = \{\mathbf{v} \in V_n | \exists \mathbf{r} \in \mathbb{R}^n \text{ such that } 0 \leq r_i \leq 1 \text{ and } \mathbf{v} = \text{Proj}_{V_n}(\mathbf{r})\}$. First, we can show that F is $n\mu$-strongly concave.

Because U is μ-strongly concave, $U(x) + \frac{\mu}{2}x^2$ is concave. It follows that

$$S(\mathbf{r}) + \frac{\mu}{2} \sum_{1 \leq i < j \leq n} (r_i - r_j)^2$$

is also concave. We can write $\sum_{1 \leq i < j \leq n}(r_i - r_j)^2$ as

$$\sum_{1 \leq i < j \leq n} (r_i - r_j)^2 = n \sum_{i=1}^{n} r_i^2 - (\sum_{i=1}^{n} r_i)^2$$

by Lagrange identity. Then note that V_n is the subspace orthogonal to $(1, 1, \cdots, 1)$. The projection onto V_n is given by

$$\text{Proj}_{V_n}(\mathbf{r}) = (r_1 - \frac{1}{n}\sum_{i=1}^{n} r_i, \cdots, r_n - \frac{1}{n}\sum_{i=1}^{n} r_i).$$

As a result,

$$\| \text{Proj}_{V_n}(\mathbf{r})\|^2 = \sum_{i=1}^{n} \left(r_i - \frac{1}{n}\sum_{j=1}^{n} r_j \right)^2 = \sum_{i=1}^{n} r_i^2 - \frac{1}{n}(\sum_{i=1}^{n} r_i)^2 = \frac{1}{n} \sum_{1 \leq i < j \leq n} (r_i - r_j)^2.$$

From this equation and the concavity of $S(\mathbf{r}) + \frac{\mu}{2}\sum_{1 \leq i < j \leq n}(r_i - r_j)^2$, we know that

$$S(\mathbf{r}) + \frac{n\mu}{2}\| \text{Proj}_{V_n}(\mathbf{r})\|^2$$

is also concave. Consequently, $F(\text{Proj}_{V_n}(\mathbf{r})) + \frac{n\mu}{2}\| \text{Proj}_{V_n}(\mathbf{r})\|^2$ is concave, which lead to the strong concavity of F because

$$F(\mathbf{v}) + \frac{n\mu}{2}\|\mathbf{v}\|^2$$

is concave. Let $\hat{\mathbf{v}}$ be the optimal vector that maximizes $F(\mathbf{v})$, strong concavity implies (See e.g. Section 9.1.2 in (Boyd & Vandenberghe, 2004))

$$F(\mathbf{v}) - F(\hat{\mathbf{v}}) \leq -\frac{n\mu}{2}\|\mathbf{v} - \hat{\mathbf{v}}\|^2.$$

Therefore, by the definition of $F(\text{Proj}_{V_n}(\mathbf{r})) = S(\mathbf{r})$, we have

$$S(\mathbf{u}) - S(\hat{\mathbf{r}}) \leq -\frac{n\mu}{2}\| \text{Proj}_{V_n}(\mathbf{u} - \hat{\mathbf{r}})\|^2.$$

$\square$

C.3. Proof of Lemma B.3

Proof of Lemma B.3. Because $\mu_j, j = 1, 2$ are symmetric, we have

$$
\begin{aligned}
r_{n,n-i+1}^{(j)} &= \frac{1}{2}\inf\{t : \mu_j([0,t]) \geq \frac{i-1}{n-1}\} + \frac{1}{2}\sup\{t : \mu_j([0,t)) < \frac{i-1}{n-1}\}\}) \\
&= \frac{1}{2}(1 - \sup\{t : \mu_j([t,1]) \geq \frac{i-1}{n-1}\}) + \frac{1}{2}(1 - \inf\{t : \mu_j((t,1]) < \frac{i-1}{n-1}\}) \\
&= \frac{1}{2}(1 - \sup\{t : \mu_j([0,t)) < \frac{n-i}{n-1}\}) + \frac{1}{2}(1 - \inf\{t : \mu_j([0,t]) \geq \frac{n-i}{n-1}\} \\
&= 1 - r_{n,i}^{(j)}.
\end{aligned}
$$

So we have $\sum_{i=1}^{n}(r_{n,i}^{(1)} - r_{n,i}^{(2)}) = 0$. Note that $V_n \subset \mathbb{R}^n$ is the subspace which is orthogonal to $(1, 1, \cdots, 1)$, the projection of $\mathbf{x} = (x_1, \cdots, x_n)$ onto V_n is given by

$$
\mathrm{Proj}_{V_n}(\mathbf{x}) = (x_1 - \frac{1}{n}\sum_{i=1}^{n}x_i, \cdots, x_n - \frac{1}{n}\sum_{i=1}^{n}x_i).
$$

Consequently,

$$
\|\mathrm{Proj}_{V_n}(\mathbf{r}_n^{(1)} - \mathbf{r}_n^{(2)})\|_2^2 = \sum_{i=1}^{n}(r_{n,i}^{(1)} - r_{n,i}^{(2)})^2.
$$

If μ_1 and μ_2 are two different symmetric probability measure on $[0, 1]$, we can assume that there exists $q_1 < q_2 \in [0, 1]$ and $\delta \geq 0$, such that $\mu_1([0, q_2]) < \mu_2([0, q_1]) - \delta$. So when $\frac{i-1}{n-1} \in (\mu_1([0, q_2]), \mu_2([0, q_1]) - \delta)$, we have $r_{n,n-i+1}^{(1)} \geq q_2$ because $\mu_1([0, q_2]) < \frac{i-1}{n-1}$. We also have $r_{n,n-i+1}^{(2)} \leq q_1$ because $\mu_2([0, q_1]) > \frac{i-1}{n-1}$. As a result, $r_{n,n-i+1}^{(1)} - r_{n,n-i+1}^{(2)} \geq q_2 - q_1$ whenever $(i-1)/(n-1) \in (\mu_1([0, q_2]), \mu_2([0, q_1]) - \delta)$. Because the length of the interval is positive, the number of such i is larger than $c_1 n$ where c_1 is a constant independent of n. Then we conclude that

$$
\begin{aligned}
\|\mathrm{Proj}_{V_n}(\mathbf{r}_n^{(1)} - \mathbf{r}_n^{(2)})\|_2^2 &= \sum_{i=1}^{n}(r_{n,i}^{(1)} - r_{n,i}^{(2)})^2 \\
&\geq c_1 n(q_1 - q_2)^2.
\end{aligned}
$$

Choosing $c_0 = c_1(q_1 - q_2)^2$ gives the inequality

$$
\|\mathrm{Proj}_{V_n}(\mathbf{r}_n^{(1)} - \mathbf{r}_n^{(2)})\|_2^2 \geq c_0 n.
$$

$\square$

C.4. Proof of Lemma B.4

Proof of Lemma B.4. Suppose there exist two different symmetric probability measure μ_1 and μ_2, they both maximize $\mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U(|X - X'|)$. Let $M = \mathbb{E}_{X,X' \overset{iid}{\sim} \mu_j} U(|X - X'|), j = 1, 2$. Now let $r_{n,i}^{(j)} = \frac{1}{2}\inf\{t : \mu_j([0,t]) \geq \frac{i-1}{n-1}\} + \frac{1}{2}\sup\{t : \mu_j([0,t)) < \frac{i-1}{n-1}\}\}), i = 1, 2, \cdots, n; j = 1, 2$ as defined in Lemma B.3. Accordingly, let $\mathbb{P}_n^{(j)} = \frac{1}{n}\sum_{i=1}^{n}\delta_{r_{n,i}^{(j)}}$. Then we have

$$
\mathbb{P}_n^{(j)} \overset{d}{\to} \mu_j, j = 1, 2.
$$

This can be proved easily by considering the definition of convergence in distribution. Since G is bounded, this lead to $\mathbb{E}_{X,X' \overset{iid}{\sim} \mathbb{P}_n^{(j)}} U(|X - X'|) \to M, j = 1, 2$ as $n \to \infty$.

The expectation $\mathbb{E}_{X,X' \overset{iid}{\sim} \mathbb{P}_n^{(j)}} U(|X - X'|)$ can be written more precisely as

$$
\mathbb{E}_{X,X' \overset{iid}{\sim} \mathbb{P}_n^{(j)}} U(|X - X'|) = \frac{1}{n^2}\sum_{1 \leq i,i' \leq n} U(|r_{n,i}^{(j)} - r_{n,i'}^{(j)}|).
$$

By Lemma B.2, we can bound the difference

$$\frac{1}{n^2} \sum_{1 \leq i, i' \leq n} U(|r_{n,i}^{(j)} - r_{n,i'}^{(j)}|) - \frac{1}{n^2} \sum_{1 \leq i \leq i' \leq n} U(|\hat{r}_{n,i} - \hat{r}_{n,i'}|)$$

$$= \frac{2}{\binom{n}{2}} \sum_{1 \leq i < i' \leq n} U(r_{n,i}^{(j)} - r_{n,i'}^{(j)}) - \frac{2}{\binom{n}{2}} \sum_{1 \leq i < i' \leq n} U(\hat{r}_{n,i} - \hat{r}_{n,i'})$$

$$\leq - \frac{2\mu}{n-1} \| \operatorname{Proj}_{V_n}(\mathbf{r}_n^{(j)} - \hat{\mathbf{r}}_n) \|_2^2.$$

Then apply Lemma B.3, there exist $c_0 \geq 0$ such that

$$2\| \operatorname{Proj}_{V_n}(\mathbf{r}_n^{(1)} - \hat{\mathbf{r}}_n) \|^2 + 2\| \operatorname{Proj}_{V_n}(\mathbf{r}_n^{(2)} - \hat{\mathbf{r}}_n) \|^2 \geq \| \operatorname{Proj}_{V_n}(\mathbf{r}_n^{(1)} - \mathbf{r}_n^{(2)}) \|^2.$$

Here, we uses $2\|x\|_2^2 + 2\|y\|_2^2 \geq \|x - y\|_2^2$. So

$$\min_{j=1,2} \frac{1}{n^2} \left[\sum_{1 \leq i, i' \leq n} U(|r_{n,i}^{(j)} - r_{n,i'}^{(j)}|) - U(|\hat{r}_{n,i} - \hat{r}_{n,i'}|) \right]$$

$$= - \frac{2\mu}{n-1} \max_{j=1,2} \| \operatorname{Proj}_{V_n}(\mathbf{r}_n^{(j)} - \hat{\mathbf{r}}_n) \|_2^2$$

$$\leq - \frac{2\mu}{n-1} \frac{\| \operatorname{Proj}_{V_n}(\mathbf{r}_n^{(1)} - \mathbf{r}_n^{(2)}) \|^2}{4}$$

$$\leq - \frac{\mu c_0 n}{2n-2} \leq - \frac{c_0 \mu}{2}.$$

Since $M = \max \mathbb{E}_{X, X' \overset{iid}{\sim} \mu} U(|X - X'|)$, we know $\frac{1}{n^2} \sum_{1 \leq i, i' \leq n} U(|\hat{r}_{n,i} - \hat{r}_{n,i'}|) \leq M$. As a result,

$$\min_{j=1,2} \mathbb{E}_{X, X' \overset{iid}{\sim} \mathbb{P}_n^{(j)}} U(|X - X'|) \leq \frac{1}{n^2} \sum_{1 \leq i \leq i' \leq n} U(|\hat{r}_{n,i} - \hat{r}_{n,i'}|) - \mu c_0 / 2 \leq M - \mu c_0 / 2.$$

This contradicts the assumption that $\mathbb{E}_{X, X' \overset{iid}{\sim} \mathbb{P}_n^{(j)}} U(|X - X'|) \to M, j = 1, 2, \ n \to \infty$.

$\square$

D. Proof of Theorem 2.1

Given Theorem 2.4, we only need to find a symmetric probability measure on $[0, 1]$, which maximizes

$$\mathbb{E}_{X, X' \overset{iid}{\sim} \mu} U(|X - X'|).$$

The following proof in this section is adapted from (, https://math.stackexchange.com/users/9340/sangchul lee). Let $M(\mathcal{B}([0,1]))$ denote the sets of all finite signed measure on the Borel sigma algebra $\mathcal{B}([0,1])$. Apparently, $P(\mathcal{B}([0,1])) \subset M(\mathcal{B}([0,1]))$. Then we define the following "inner product" in $M(\mathcal{B}([0,1]))$:

$$\langle \mu, \nu \rangle = \mathbb{E}_{X \sim \mu, X' \sim \nu, \text{independent}} U(|X - X'|) = \int_{[0,1]^2} U(|x - y|) \mu(dx) \nu(dy).$$

We also define $I(\mu)$ as $I(\mu) := \langle \mu, \mu \rangle$. With these notations, the problem becomes

$$\max_{\mu \in P(\mathcal{B}([0,1]))} I(\mu).$$

Lemma D.1. *For $U(x) = x^\gamma$ with $\gamma \in (0, 1)$. If μ is a signed measure satisfying $\mu([0,1]) = 0$, then we have $I(\mu) \leq 0$. Moreover, $I(\mu) = 0$ if and only if $\mu(E) = 0$ for all $E \subset [0, 1]$.*

Proof. $f(t) = \frac{1-\cos(xt)}{t^{1+\gamma}}$ is integrable on $(0, \infty)$. As a result, using change of variables, we have

$$|x|^\gamma = C \int_0^\infty \frac{1 - \cos(xt)}{t^{1+\gamma}} dt$$

for come constant $C > 0$. Then by Fubini's theorem, we have

$$\langle \mu, \mu \rangle = \int_{[0,1]^2} |x - y|^\gamma \mu(dx)\mu(dy)$$

$$= C \int_{[0,1]^2} \int_0^\infty \frac{1 - \cos((x - y)t)}{t^{1+\gamma}} dt\mu(dx)\mu(dy)$$

$$= C \int_0^\infty \left(\int_{[0,1]^2} \frac{1 - \cos((x - y)t)}{t^{1+\gamma}} \mu(dx)\mu(dy) \right) dt.$$

Note that $\cos((x - y)t) = \Re(e^{ixt-iyt})$, we have

$$\int_{[0,1]^2} \frac{1 - \cos((x - y)t)}{t^{1+\gamma}} \mu(dx)\mu(dy)$$

$$= -\Re \left(\int_{[0,1]^2} \frac{e^{ixt}e^{-iyt}}{t^{1+\gamma}} \mu(dx)\mu(dy) \right)$$

$$= -\Re \left(|\hat{\mu}(t)|^2 \right),$$

where $\hat{\mu}(t) = \int_{[0,1]} e^{itx} \mu(dx)$ is the Fourier transform of μ. Then

$$I(\mu) = -C \int_0^\infty \frac{|\hat{\mu}(t)|^2}{t^{1+\gamma}} dt \le 0.$$

Moreover, $I(\mu) = 0$ if and only if $\hat{\mu}(t) = 0$ for all $t \in [0, \infty)$ if and only if $\mu(E) = 0$ for all $E \in \mathcal{B}([0, 1])$. $\square$

D.1. Proof of Lemma B.5

We first restate the lemma.

Lemma D.2. *Let* $U(x) = x^\gamma$ *for some* $\gamma \in (0, 1)$. *If a probability measure* μ *on* $[0, 1]$ *maximize* $\mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U(|X - X'|)$ *if it satisfies that* $\mathbb{E}_{X \sim \mu} U(|X - c|)$ *does not depend on* $c \in [0, 1]$.

Proof of Lemma B.5. For two probability measure μ and ν on $[0, 1]$, $(\mu - \nu)([0, 1]) = 0$. Suppose μ satisfies $\mathbb{E}_{X \sim \mu} U(|X - c|) = K$ does not depend on $c \in [0, 1]$. Note that

$$\langle \nu - \mu, \mu \rangle = \int_{[0,1]} \left(\int_{[0,1]} |x - y|^\gamma \mu(dx) \right) (\nu - \mu)(dy) = \int_{[0,1]} K(\nu - \mu)(dy) = 0.$$

And by lemma D.1, $\langle \nu - \mu, \nu - \mu \rangle \le 0$. Therefore,

$$\langle \nu, \nu \rangle = \langle \mu, \mu \rangle + 2 \langle \nu - \mu, \mu \rangle + \langle \nu - \mu, \nu - \mu \rangle \le \langle \mu, \mu \rangle.$$

This means that μ maximize $\mathbb{E}_{X,X' \overset{iid}{\sim} \mu} U(|X - X'|)$. $\square$

D.2. Proof of Theorem 2.1

Proof of Theorem 2.1. Let μ be the probability measure induced by $\text{Beta}(\frac{1-\gamma}{2}, \frac{1-\gamma}{2})$. It has probability density function

$$f_\gamma(x) = \frac{1}{B(\frac{1-\gamma}{2}, \frac{1-\gamma}{2})} x^{-\frac{1+\gamma}{2}} (1 - x)^{-\frac{1+\gamma}{2}}.$$

For any $c \in [0,1]$, $\mathbb{E}_{X \sim \mu} U(|X - c|)$ can be expressed as

$$
\begin{aligned}
\mathbb{E}_{X \sim \mu} U(|X - c|) &= \frac{1}{B(\frac{1-\gamma}{2}, \frac{1-\gamma}{2})} \int_0^1 |x - c|^\gamma x^{-\frac{1+\gamma}{2}} (1-x)^{-\frac{1+\gamma}{2}} dx \\
&= \frac{1}{B(\frac{1-\gamma}{2}, \frac{1-\gamma}{2})} \int_0^{\frac{\pi}{2}} |\sin^2\theta - c|^\gamma (\sin\theta)^{-1-\gamma} (\cos\theta)^{-1-\gamma} d\sin^2\theta \\
&= \frac{2}{B(\frac{1-\gamma}{2}, \frac{1-\gamma}{2})} \int_0^{\frac{\pi}{2}} \left| \frac{\sin^2\theta - c}{\sin\theta\cos\theta} \right|^\gamma d\theta \\
&= \frac{2}{B(\frac{1-\gamma}{2}, \frac{1-\gamma}{2})} \int_0^\infty \left(\int_0^{\pi/2} \mathbf{1}\left\{ \left| \frac{\sin^2\theta - c}{\sin\theta\cos\theta} \right|^\gamma \geq t \right\} d\theta \right) dt.
\end{aligned}
$$

Because

$$
\begin{aligned}
\int_0^{\pi/2} \mathbf{1}\left\{ \left| \frac{\sin^2\theta - c}{\sin\theta\cos\theta} \right|^\gamma \geq t \right\} d\theta &= \frac{1}{2} \int_0^\pi \mathbf{1}\left\{ \left| \frac{\cos\theta + 2c - 1}{\sin\theta} \right|^\gamma \geq t \right\} d\theta \\
&= \frac{\pi}{2} - \frac{1}{2} \int_0^\pi \mathbf{1}\left\{ -\cos\theta - t^{1/\gamma}\sin\theta \leq 2c - 1 \leq -\cos\theta + t^{1/\gamma}\sin\theta \right\} d\theta \\
&= \frac{\pi}{2} - \frac{1}{2} \int_0^\pi \mathbf{1}\left\{ -\cos(\theta - \phi) \leq \frac{2c - 1}{\sqrt{1 + t^{2/\gamma}}} \leq -\cos(\theta + \phi) \right\} d\theta \\
&= \frac{\pi}{2} - \phi,
\end{aligned}
$$

where $\tan\phi = t^{1/\gamma}$ and $\phi \in [0, \pi/2]$, and the last equation use the fact that $c \in [0,1]$. As a result, $\mathbb{E}_{X \sim \mu} U(|X - c|)$ does not depend on c.

Note that Beta distribution is also symmetric. It follows from Theorem 2.4 that the reward distribution converges to Beta $\left(\frac{1-\gamma}{2}, \frac{1-\gamma}{2} \right)$ □

E. Proof in Section 4

E.1. Proof of Theorem 4.1

Proof of Theorem 4.1. First, we prove that $\hat{\mathbf{r}}$ keeps the order of $\{\theta_i\}_{1 \leq i \leq n}$. If there exist i and j such that $\theta_i < \theta_j$ and $\hat{r}_i > \hat{r}_j$, we define

$$
\tilde{r}_k = \begin{cases} \hat{r}_k & \text{if } k \neq i, j; \\ \hat{r}_j & \text{if } k = i; \\ \hat{r}_i & \text{if } k = j. \end{cases}
$$

Then

$$
\begin{aligned}
S(\hat{\mathbf{r}}) - S(\tilde{\mathbf{r}}) = &\sum_{k \notin \{i,j\}} U(\hat{r}_i - \hat{r}_k)\texttt{sigmoid}(\theta_i - \theta_k) + \sum_{k \notin \{i,j\}} U(\hat{r}_k - \hat{r}_i)\texttt{sigmoid}(\theta_k - \theta_i) \\
&+ \sum_{k \notin \{i,j\}} U(\hat{r}_j - \hat{r}_k)\texttt{sigmoid}(\theta_j - \theta_k) + \sum_{k \notin \{i,j\}} U(\hat{r}_k - \hat{r}_j)\texttt{sigmoid}(\theta_k - \theta_j) \\
&- \sum_{k \notin \{i,j\}} U(\hat{r}_j - \hat{r}_k)\texttt{sigmoid}(\theta_i - \theta_k) - \sum_{k \notin \{i,j\}} U(\hat{r}_k - \hat{r}_j)\texttt{sigmoid}(\theta_k - \theta_i) \\
&- \sum_{k \notin \{i,j\}} U(\hat{r}_i - \hat{r}_k)\texttt{sigmoid}(\theta_j - \theta_k) - \sum_{k \notin \{i,j\}} U(\hat{r}_k - \hat{r}_i)\texttt{sigmoid}(\theta_k - \theta_j) \\
&+ U(\hat{r}_i - \hat{r}_j)\texttt{sigmoid}(\theta_i - \theta_j) + U(\hat{r}_j - \hat{r}_i)\texttt{sigmoid}(\theta_j - \theta_i) \\
&- U(\hat{r}_j - \hat{r}_i)\texttt{sigmoid}(\theta_i - \theta_j) + U(\hat{r}_i - \hat{r}_j)\texttt{sigmoid}(\theta_j - \theta_i).
\end{aligned}
$$

Note that for $a < b$ and $c < d$, we have inequality $ad + bc - ac - bd = (a - b)(d - c) < 0$. It follows that from the monotonicity of U and $\texttt{sigmoid}$ function, and our assumption $\theta_i < \theta_j$ and $\hat{r}_i > \hat{r}_j$, we have

$$
\begin{aligned}
& U(\hat{r}_i - \hat{r}_k)\texttt{sigmoid}(\theta_i - \theta_k) + U(\hat{r}_j - \hat{r}_k)\texttt{sigmoid}(\theta_j - \theta_k) \\
& - U(\hat{r}_j - \hat{r}_k)\texttt{sigmoid}(\theta_i - \theta_k) - U(\hat{r}_i - \hat{r}_k)\texttt{sigmoid}(\theta_j - \theta_k) < 0, \\
& U(\hat{r}_k - \hat{r}_i)\texttt{sigmoid}(\theta_k - \theta_i) + U(\hat{r}_k - \hat{r}_j)\texttt{sigmoid}(\theta_k - \theta_j) \\
& - U(\hat{r}_k - \hat{r}_j)\texttt{sigmoid}(\theta_k - \theta_i) - U(\hat{r}_k - \hat{r}_i)\texttt{sigmoid}(\theta_k - \theta_j) < 0, \\
& U(\hat{r}_i - \hat{r}_j)\texttt{sigmoid}(\theta_i - \theta_j) + U(\hat{r}_j - \hat{r}_i)\texttt{sigmoid}(\theta_j - \theta_i) \\
& - U(\hat{r}_j - \hat{r}_i)\texttt{sigmoid}(\theta_i - \theta_j) - U(\hat{r}_i - \hat{r}_j)\texttt{sigmoid}(\theta_j - \theta_i) < 0.
\end{aligned}
$$

As a result, $S(\hat{\mathbf{r}}) < S(\tilde{\mathbf{r}})$, which contradicts the optimality of $\hat{\mathbf{r}}$. This gives that $\hat{\mathbf{r}}$ keep the order of $\{\theta_i\}_{1 \leq i \leq n}$.

To prove the inequality in Theorem 4.1, let us consider the case where $i = 1$ and $j = 2$ without loss of generality. We also assume $\theta_1 \geq \theta_2$. Then it follows from previous argument that $\hat{r}_1 \geq \hat{r}_2$. From the optimality of $\hat{r}$, we have

$$
S(\hat{\mathbf{r}}) \geq S(\frac{\hat{r}_1 + \hat{r}_2}{2}, \frac{\hat{r}_1 + \hat{r}_2}{2}, \hat{r}_3, \cdots, \hat{r}_n).
$$

The difference can be written as

$$
\begin{aligned}
& S(\frac{\hat{r}_1 + \hat{r}_2}{2}, \frac{\hat{r}_1 + \hat{r}_2}{2}, \hat{r}_3, \cdots, \hat{r}_n) - S(\hat{\mathbf{r}}) \\
= & \sum_{i>2} U(\frac{\hat{r}_1 + \hat{r}_2}{2} - \hat{r}_i)(\texttt{sigmoid}(\theta_1 - \theta_i) + \texttt{sigmoid}(\theta_2 - \theta_i)) \\
& - \sum_{i>2} U(\hat{r}_1 - \hat{r}_i)\texttt{sigmoid}(\theta_1 - \theta_i) - \sum_{i>2} U(\hat{r}_2 - \hat{r}_i)\texttt{sigmoid}(\theta_2 - \theta_i) \\
& + \sum_{i>2} U(\hat{r}_i - \frac{\hat{r}_1 + \hat{r}_2}{2})(\texttt{sigmoid}(\theta_i - \theta_1) + \texttt{sigmoid}(\theta_i - \theta_2)) \\
& - \sum_{i>2} U(\hat{r}_i - \hat{r}_1)\texttt{sigmoid}(\theta_i - \theta_1) - \sum_{i>2} U(\hat{r}_i - \hat{r}_2)\texttt{sigmoid}(\theta_i - \theta_2) \\
& + U(0)\texttt{sigmoid}(\theta_1 - \theta_2) + U(0)\texttt{sigmoid}(\theta_2 - \theta_1) \\
& - U(\hat{r}_1 - \hat{r}_2)\texttt{sigmoid}(\theta_1 - \theta_2) - U(\hat{r}_2 - \hat{r}_1)\texttt{sigmoid}(\theta_2 - \theta_1).
\end{aligned}
$$

From strong concavity of U, we have

$$
\frac{U(x) + U(y)}{2} \leq U(\frac{x + y}{2}) - \frac{\mu(x - y)^2}{8}.
$$

As a result,

$$
\begin{aligned}
& U(\frac{\hat{r}_1 + \hat{r}_2}{2} - \hat{r}_i)(\texttt{sigmoid}(\theta_1 - \theta_i) + \texttt{sigmoid}(\theta_2 - \theta_i)) \\
\geq & 2U(\frac{\hat{r}_1 + \hat{r}_2}{2} - \hat{r}_i)\texttt{sigmoid}(\theta_2 - \theta_i) \\
\geq & \texttt{sigmoid}(\theta_2 - \theta_i)(U(\hat{r}_1 - \hat{r}_i) + U(\hat{r}_2 - \hat{r}_i) + \frac{\mu(\hat{r}_1 - \hat{r}_2)^2}{4}).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& U(\hat{r}_i - \frac{\hat{r}_1 + \hat{r}_2}{2})(\texttt{sigmoid}(\theta_i - \theta_1) + \texttt{sigmoid}(\theta_i - \theta_2)) \\
\geq & \texttt{sigmoid}(\theta_i - \theta_1)(U(\hat{r}_i - \hat{r}_1) + U(\hat{r}_i - \hat{r}_2) + \frac{\mu(\hat{r}_1 - \hat{r}_2)^2}{4}), \\
& U(0)\texttt{sigmoid}(\theta_1 - \theta_2) + U(0)\texttt{sigmoid}(\theta_2 - \theta_1) \\
\geq & \texttt{sigmoid}(\theta_2 - \theta_1)(U(\hat{r}_1 - \hat{r}_2) + U(\hat{r}_2 - \hat{r}_1) + \mu(\hat{r}_1 - \hat{r}_2)^2).
\end{aligned}
$$

As a consequence, letting $m = \min_{|x|<\kappa} \texttt{sigmoid}(x) = \frac{1}{1+e^\kappa}$, $M = \max_{x\in[-1,1]} U(x) = U(1)$, and L be the Lipschitz constant of $\texttt{sigmoid}(\cdot)$, which is bounded by 1, we have

$$
\begin{aligned}
&S(\frac{\hat{r}_1 + \hat{r}_2}{2}, \frac{\hat{r}_1 + \hat{r}_2}{2}, \hat{r}_3, \cdots, \hat{r}_n) - S(\hat{\mathbf{r}}) \\
\geq\ & \sum_{i>2} (\texttt{sigmoid}(\theta_2 - \theta_i) - \texttt{sigmoid}(\theta_1 - \theta_i)) U(\hat{r}_1 - \hat{r}_i) \\
& + \sum_{i>2} (\texttt{sigmoid}(\theta_i - \theta_1) - \texttt{sigmoid}(\theta_i - \theta_2)) U(\hat{r}_i - \hat{r}_2) \\
& + (\texttt{sigmoid}(\theta_2 - \theta_1) - \texttt{sigmoid}(\theta_1 - \theta_2)) U(\hat{r}_1 - \hat{r}_2) + \frac{m(2(n-2)+4)\mu(\hat{r}_1 - \hat{r}_2)^2}{4} \\
\geq\ & -(2n-2)LM|\theta_1 - \theta_2| + \frac{mn\mu(\hat{r}_1 - \hat{r}_2)^2}{2}.
\end{aligned}
$$

So the optimality of $\hat{r}$ gives

$$
\frac{mn\mu(\hat{r}_1 - \hat{r}_2)^2}{2} \leq 2nLM|\theta_1 - \theta_2|.
$$

This yields the inequality

$$
|\hat{r}_1 - \hat{r}_2| \leq 2\sqrt{LM|\theta_1 - \theta_2|/(m\mu)} \leq 2\sqrt{U(1)(1+e^\kappa)|\theta_1 - \theta_2|/\mu}.
$$

$\square$

E.2. Reward distribution in Section 4

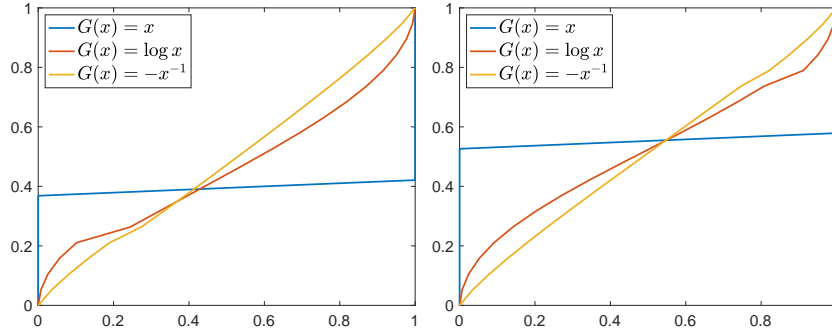*Figure 5.* Reward distribution with different choice of $\{\theta\}_{1\leq i\leq n}$ when $n = 20$.

The choice of $\{\theta_i\}_{1\leq i\leq n}$ in the left is

$$
\theta_i = \begin{cases} i/20 & \text{if } i \leq 15, \\ (i+10)/6 & \text{if } i > 15. \end{cases}
$$

The choice of $\{\theta_i\}_{1\leq i\leq n}$ in the left is

$$
\theta_i = \begin{cases} i/10 & \text{if } i \leq 5, \\ (i+10)/6 & \text{if } i > 5. \end{cases}
$$