Characterizing Output Distribution Shift under Preference Tuning under the Bradley-Terry Model

Brian R.Y. Huang

Abstract

Playing with the math of DPO leads to a neat theorem. TL;DR: by casting an LLM output distribution as a mixture of an 'aligned' distribution and 'unaligned' distribution with certain human preference constraints, we can write the exact form of the LLM output distribution after preference tuning in terms of these mixture components; we can also show that output distribution mass always shifts from the unaligned component to the aligned component.

1 Preference tuning shrinks, but not truncates, unaligned outputs

1.1 Background: preference tuning

In preference tuning, we finetune a model, usually a base pretrained model or SFT-treated model, using a dataset of human preference samples $(x_i, y_i^1, ..., y_i^k) \sim \mathcal{D}$. To create a preference dataset, for each prompt x_i, we independently sample k completions from our model, and have human raters supply k labels; for example, human raters can say which one of $k = 2$ completions they prefer, listwise rank $k > 2$ completions, or give a thumbs up/down for $k = 1$ completion. In any case, we model our sampled human preferences as generated from a latent reward function $r(y; x)$, and our goal is to find preference-tuned model parameters θ which maximize the expected reward of continuations sampled from the model, while regularizing for KL distance between the continuation output distributions of θ and the pre-tuning model checkpoint, henceforth called the "reference model" ref. Our preference-tuning objective function is

$$\mathcal{L}(\theta) = \max_{\theta} E_{x;y\in p_\theta(y|x)}[r(y;x)] - \beta\, \mathbb{D}_{KL}(\, p_\theta(y|x) \,||\, p_{ref}(y|x)) \tag{1}$$

To approximate the latent reward, we may take one of several approaches from the literature. For example, RLHF [LJX+22] trains a reward model (often, a linear probe on the last layer of our language model) to estimate rewards, following the Bradley-Terry model [RM52] for preference probabilities based on latent reward. Another method, DPO [RAE+23], optimizes the language model directly with a classification loss, which corresponds to an implicit reward function also following the Bradley-Terry model and has the same optimal solution as the original preference-tuning objective. Other preference-tuning methods in the literature, like LiPO, GRPO, new one comes out every week these days..., KTO [KWD+23] and SLiC-HF [YRT+23], are also possible. In our subsequent analysis of preference tuning, we only assume the preference tuning objective in (1), so our treatment is general to any preference tuning method. (For example, we don't even assume Bradley-Terry preferences, so our analysis will apply to KTO and SLiC-HF.) Add formal statement of Bradley-Terry?

1.2 Distribution shift in preference-tuned model outputs

The typical end goal of preference tuning is to "align" the language model. In other words, if we assume the latent reward model driving human raters' preferences corresponds to human values such as helpfulness, honesty, morality, and more, preference tuning procedures should push the model's outputs on any prompt to embody these human values more effectively on average.

Preprint.

One interpretation of this goal is that, for any given prompt x, we can construct any aligned distribution of outputs $\mathcal{A}_x$ from which sampled outputs $y \sim p_{\mathcal{A}_x}(\cdot|x)$ have a higher expected reward—i.e., the output is more "aligned"—compared to sampling from the reference model $p_{ref}(\cdot|x)$. On the flipside, the remainder of the probability mass of $p_{ref}(\cdot|x)$ not comprised of $\mathcal{A}$ would be represented by the "unaligned" distribution $\mathcal{U}_x$, for which expected rewards of sampled outputs are lower than those of ref. In this interpretation, we may write $p_{ref}(\cdot|x)$ as a mixture of the aligned and unaligned distributions

$$p_{ref}(y|x) = \alpha_x p_{\mathcal{A}_x}(y|x) + (1-\alpha_x) p_{\mathcal{U}_x}(y|x). \tag{2}$$

In this view, our hope for preference tuning is that the resulting model's output distribution conforms as closely as possible to $\mathcal{A}_x$ and, in turn, is influenced as little as possible by $\mathcal{U}_x$. We find that preference-tuning shifts model outputs towards the aligned distribution in a certain fashion: crucially, the unaligned component $\mathcal{U}_x$ of mixture (2) is shrunk but not truncated, and even outputs y that exist only in the unaligned distribution ($p_{\mathcal{U}_x}(y|x) > 0$ and $p_{\mathcal{A}_x}(y|x) = 0$) are not fully expunged from the model ($p_\theta(y|x) > 0$). We present a theorem which characterizes the distribution shift of a preference-tuned model with respect to the aligned and unaligned component distributions.

First, given any prompt x, decompose the output distribution of the reference model as $p_{ref}(y|x) = \alpha(x) p_{\mathcal{A}}(y|x) + (1-\alpha_x) p_{\mathcal{U}}(y|x)$ for some "aligned" and "unaligned" distributions $\mathcal{A}$ and $\mathcal{U}$, respectively. (For brevity, we've dropped the subscripts from $\mathcal{A}_x$ and $\mathcal{U}_x$.) Let θ be the optimal model parameters found via preference tuning.

Lemma 1 *Assume the preference-tuning objective function used is eq. (1). With no additional assumptions on $\mathcal{A}$ and $\mathcal{U}$, the output distribution of θ can be written as*

$$p_\theta(y|x) = \alpha_{\mathcal{A}}(y;x) p_{\mathcal{A}}(y|x) + \alpha_{\mathcal{U}}(y;x) p_{\mathcal{U}}(y|x)$$

$$\alpha_{\mathcal{A}}(y;x) = \frac{1}{Z(x)} \alpha(x) e^{r(y;x)/\beta}, \quad \alpha_{\mathcal{U}}(y;x) = \frac{1}{Z(x)} (1-\alpha(x)) e^{r(y;x)/\beta}$$

where $Z(x) = \sum_{y \in \operatorname{supp} p_\theta(\cdot|x)} p_{ref}(y|x) e^{r(y;x)/\beta}$ is the partition function of $p_\theta(y;x)$.

Theorem 2 *Write the "total probability mass" of $\mathcal{A}$ in $p_\theta(\cdot|x)$ as $\mathrm{TPM}(\mathcal{A};x) = \sum_{y \in \operatorname{supp} p_\theta(\cdot|x)} \alpha_{\mathcal{A}}(y;x) p_{\mathcal{A}}(y|x)$ (and define this respectively for $\mathcal{U}$). If the reward function $r(y;x)$ satisfies the following additional assumptions:*

- *$r(y;x)$ and $p_{\mathcal{A}}(y|x)$ are positively correlated*
- *$r(y;x)$ and $p_{\mathcal{U}}(y|x)$ are negatively correlated*
- *r is finite*

then the total probability masses of $\mathcal{A}$ and $\mathcal{U}$ in $p_\theta(\cdot|x)$ satisfy

$$1 > \mathrm{TPM}(\mathcal{A};x) \geq \alpha(x), \quad 0 < \mathrm{TPM}(\mathcal{U};x) \leq (1-\alpha(x)).$$

The total probability masses of $\mathcal{A}$ and $\mathcal{U}$ can be viewed as a measure of the aligned (resp. unaligned) distribution's influence on the tuned language model's output distribution. Intuitively, we can interpret Theorem 2 as the following idea: if we select aligned and unaligned distributions that indeed correspond to a level of "alignment" or utility measured by our rewards, then the preference-tuned model draws outputs more heavily from the aligned distribution, while the influence of the unaligned distribution is diminished but not expunged. Crucially, the preference-tuned model still has nonzero probability of producing any harmful or otherwise undesirable output that was possible in the base model.

1.3 Proofs

Our proofs of Lemma 1 and Theorem 2 will illuminate further details for how the probability of eliciting a given output from our model is amplified or diminished as a function of the reward.

First, the closed-form solution for the preference-tuning objective (1) is well-known [cite something?], but we reproduce it here. First, rewrite the objective as

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_y p_\theta(y|x) r(y;x) - \beta p_\theta(y|x) \log\left(\frac{p_\theta(y|x)}{p_{ref}(y|x)}\right) \\ &= \beta \sum_y p_\theta(y|x) \Big[\log e^{r(y;x)/\beta} - \log\left(\frac{p_\theta(y|x)}{p_{ref}(y|x)}\right)\Big] \\ &= -\beta \sum_y p_\theta(y|x) \log\left(\frac{p_\theta(y|x)}{\frac{1}{Z(x)} p_{ref}(y|x) e^{r(y;x)/\beta}}\right) + \beta \log Z(x)\end{aligned}$$

where $Z(x) = \sum_y p_{ref}(y|x) e^{r(y;x)/\beta}$ is the partition function. Notice that $\frac{1}{Z(x)} p_{ref}(y|x) e^{r(y;x)/\beta}$ is a valid probability distribution, so by Gibbs' inequality, the objective $\mathcal{L}(\theta)$ is maximized when

$$p_\theta(y|x) = \frac{1}{Z(x)} p_{ref}(y|x) e^{r(y;x)/\beta}.$$

Substituting the mixture formulation for $p_{ref}(y|x)$ and performing a little more algebra gives us

$$p_\theta(y|x) = \alpha_{\mathcal{A}}(y;x) p_{\mathcal{A}}(y|x) + \alpha_{\mathcal{U}}(y;x) p_{\mathcal{U}}(y|x)$$

$$\alpha_{\mathcal{A}}(y;x) = \frac{1}{Z(x)} \alpha(x) e^{r(y;x)/\beta}, \;\; \alpha_{\mathcal{U}}(y;x) = \frac{1}{Z(x)} (1-\alpha(x)) e^{r(y;x)/\beta}$$

as Lemma 1 states.

For Theorem 2, we assumed $r(y;x)$ was positively correlated with $p_{\mathcal{A}}(y|x)$ and negatively correlated with $p_{\mathcal{U}}(y|x)$. Since $Z(x)$ and $\alpha(x)$ are constants with respect to y, $\alpha_{\mathcal{A}}(y;x)$ and $\alpha_{\mathcal{U}}(y;x)$ are clearly monotonic functions of $r(y;x)$, and as a result, the correlation properties are preserved: $\alpha_{\mathcal{A}}(y;x)$ is positively correlated with $p_{\mathcal{A}}(y|x)$, and $\alpha_{\mathcal{U}}(y;x)$ is negatively correlated with $p_{\mathcal{U}}(y|x)$.

Notice that finiteness of r implies $e^{r(y;x)/\beta} > 0$ everywhere, so the output distribution of θ must have nonzero probability for any output in $\operatorname{supp} p_{ref}$. Therefore, $\operatorname{supp} p_\theta(\cdot|x) = \operatorname{supp} p_{ref}(\cdot|x)$. Consider the expectations of $\alpha_{\mathcal{A}}(y;x)$ and $\alpha_{\mathcal{U}}(y;x)$ over a uniform distribution over (WLOG) $\operatorname{supp} p_\theta(\cdot|x)$, and use the correlation properties:

$$\begin{aligned}E_{y \in \text{unif}(\operatorname{supp} p_\theta(\cdot|x))}[\alpha_{\mathcal{A}}(y;x) p_{\mathcal{A}}(y|x)] &> E_y[\alpha_{\mathcal{A}}(y;x)] E_y[p_{\mathcal{A}}(y|x)] \\ &= \frac{1}{Z(x)} \alpha(x) E_y[e^{r(y;x)/\beta}] E_y[p_{\mathcal{A}}(y|x)].\end{aligned}$$

$$\begin{aligned}E_y[\alpha_{\mathcal{U}}(y;x) p_{\mathcal{U}}(y|x)] &< E_y[\alpha_{\mathcal{U}}(y;x)] E_y[p_{\mathcal{U}}(y|x)] \\ &= \frac{1}{Z(x)} (1-\alpha(x)) E_y[e^{r(y;x)/\beta}] E_y[p_{\mathcal{U}}(y|x)].\end{aligned}$$

(For the sake of brevity, we mostly omit writing the distribution over which the expectation is taken; for the rest of the proof, $E_y[\cdot]$ is shorthand for $E_{y \in \text{unif}(\operatorname{supp} p_\theta(\cdot|x))}[\cdot]$.) The $E_y[e^{r(y;x)/\beta}]$ and partition function terms vanish upon dividing the inequalities:

$$\frac{E_y[\alpha_{\mathcal{A}}(y;x) p_{\mathcal{A}}(y|x)]}{E_y[\alpha_{\mathcal{U}}(y;x) p_{\mathcal{U}}(y|x)]} > \frac{E_y[\alpha(x) p_{\mathcal{A}}(y|x)]}{E_y[(1-\alpha(x)) p_{\mathcal{U}}(y|x)]}$$

Reciprocating the inequality, adding 1 to both sides, and reciprocating again yields:

$$\frac{E_y[\alpha_{\mathcal{A}}(y;x) p_{\mathcal{A}}(y|x)]}{E_y[p_\theta(y|x)]} > \frac{E_y[\alpha(x) p_{\mathcal{A}}(y|x)]}{E_y[p_{ref}(y|x)]}$$

Finally, we multiply all expectations by $\operatorname{card}(\operatorname{supp} p_\theta(\cdot|x))$. Using the property that

$$\operatorname{card}(\operatorname{supp} p) \cdot E_{y \in \text{unif}(\operatorname{supp} p)}[p(x)] = 1$$

for any discrete probability distribution p, along with our earlier observation that $\operatorname{supp} p_\theta(\cdot|x) = \operatorname{supp} p_{ref}(\cdot|x)$, we see that the expectations of $p_\theta(y|x)$ and $p_{ref}(y|x)$ vanish. On the other hand, multiplying the expectation of $\alpha_{\mathcal{A}}(y;x) p_{\mathcal{A}}(y|x)$ by $\operatorname{card}(\operatorname{supp} p_\theta(\cdot|x))$ gives the total probability mass. We recover

$$\text{TPM}(\mathcal{A};x) > \alpha(x) \qquad \text{and likewise, } \text{TPM}(\mathcal{U};x) < (1-\alpha(x)).$$

1.4 More Notes

This section is informal, I'm jotting down more observations about the post-tuning output distribution and what intuition that might build.

An important point we raised midway through the previous proof is that, with finite r, as long as some continuation y of a given prompt can be naturally outputted from the model pre-preference tuning, the tuned model θ must have nonzero probability of outputting that same continuation for the prompt. More specifically, the output probability for y given x is shrunk by a factor of $e^{r(y;x)/\beta}$.

In practice, great pains are taken to make sure the reward model does not degenerate into negative infinite outputs [should cite something for this], so the finiteness assumption is virtually guaranteed. As such, it may be concerning that most, if not all, preference-tuned frontier models still retain some base model-level capabilities for harmful outputs; with certain prompts, these preference-tuned models may be able to produce harmful outputs at a rate much closer to that of their respective base models. The (rough) intuition is that, if we can discover prompts where the unaligned mixture component greatly dominates over the aligned mixture component for continuations, then, even after preference-tuning amplifies the aligned component and shrinks the unaligned component, we still have significant risk of sampling from the unaligned component using these prompts on the preference-tuned model.

This idea came to mind when I realized a link between the training data extraction paper [MNJ+23] and "What's In My Big Data?" paper [YAI+23]. The training data extraction paper finds that using long single-token repeats as inputs (for example, "aaaaaaaaaaaaaaaaaaaa" or "////////////////////") causes OpenAI chat models to stop behaving in the user-facing chat format and instead give printouts of boilerplate documents/information, mostly hallucinated but occasionally regurgitations of genuine private data. The "WIMBD?" paper analyzes web-scale datasets with a mapreduce-based system and finds out the most common 10-grams in all explored datasets are single-token repeats (i.e. for each dataset, across the 10 most common 10-grams, majority if not all are single-token repeats). The paper looks at the context of these 10-grams and finds that they usually come from delimiters in common documents and files. The realization here is that these single-token repeats are virtually never found as outputs of preference-tuned chat models, but they're very common in the pretraining data, and would actually be highly weighted in the output distribution of a base model. By having very higher mass in a base model's output distribution vs. very low mass in a preference-tuned model's output distribution, do long single-token repeats steer a tuned model to behave like a base model?

$$logits_{tkbk} = \alpha * logits_{\theta} - \beta * logits_{ref}$$

References

[KWD+23] Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Kiela Douwe. "Human-Centered Loss Functions (HALOs)". In: (2023). URL: https://github.com/ContextualAI/HALOs/blob/main/assets/report.pdf.

[LJX+22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. "Training language models to follow instructions with human feedback". In: *NeurIPS* (2022).

[MNJ+23] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. "Scalable Extraction of Training Data from (Production) Language Models". In: *arXiv preprint arXiv:2311.17035* (2023).

[RAE+23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model". In: *NeurIPS* (2023).

[RM52] Ralph Allan Bradley and Milton E. Terry. "Rank analysis of incomplete block designs: I. the method of paired comparisons." In: *Biometrika* 39.3/4 (1952), pp. 324–345.

[YAI+23] Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. "What's In My Big Data?" In: *arXiv preprint arXiv:2310.20707* (2023).

[YRT+23] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. "SLiC-HF: Sequence Likelihood Calibration with Human Feedback". In: *arXiv preprint arXiv:2305.10425* (2023).