1. The Bradley-Terry Model

The Bradley-Terry model is a probabilistic model for paired comparisons. Given two items, say A and B, the model estimates the probability that item A is preferred over item B.

The core assumption is that each item i has a latent "strength" or "skill" parameter, often denoted as π_i. The probability that item i is preferred over item j is given by:

$$P(i \text{ preferred over } j) = \frac{\pi_i}{\pi_i + \pi_j}$$

In the context of reward modeling with chosen (c) and rejected (r) responses, we can think of the model as assigning a score or "strength" to each response. Let $s(response)$ be the score the reward model assigns to a given response. The probability that the chosen response is preferred over the rejected response is:

$$P(\text{chosen preferred over rejected}) = \frac{e^{s(\text{chosen})}}{e^{s(\text{chosen})} + e^{s(\text{rejected})}}$$

The exponential is used to ensure the scores are positive, and the ratio represents the relative strength.

2. Loss Function Used by RewardTrainer

The RewardTrainer in trl typically uses a loss function that aims to maximize the likelihood of the observed preferences. This is often a negative log-likelihood loss based on the pairwise comparisons.

For a given pair of chosen (c) and rejected (r) responses, the loss is designed to penalize the model when the score of the rejected response is higher than or equal to the score of the chosen response.

A common form of this loss, derived from the negative log-likelihood of the Bradley-Terry probability, is:

$$L = -\log\left(\frac{e^{s(\text{chosen})}}{e^{s(\text{chosen})} + e^{s(\text{rejected})}}\right)$$

This can be rewritten as:

$$L = \log(e^{s(\text{chosen})} + e^{s(\text{rejected})}) - s(\text{chosen})$$

Alternatively, some implementations might use a hinge loss or a margin-based loss, such as:

$$L = \max(0, s(\text{rejected}) - s(\text{chosen}) + \text{margin})$$

However, the negative log-likelihood derived from the probabilistic formulation is very common in reward modeling as it directly corresponds to the probability of the observed preference. The RewardTrainer in trl most likely uses a variant of the negative log-likelihood loss, which aligns with the principles of the Bradley-Terry model, even if the exact implementation might have minor differences for numerical stability or efficiency.

In summary, the RewardTrainer uses a loss function that encourages the model to output a higher score for the chosen response compared to the rejected response. This loss is typically based on the probabilistic framework of pairwise comparison models like the Bradley-Terry model.

Derivation of Formula

The formula you provided is $L = -\log(\sigma(s(\text{chosen}) - s(\text{rejected})))$.

Let $\Delta s = s(\text{chosen}) - s(\text{rejected})$. Then the loss is $L = -\log(\sigma(\Delta s))$.

Substitute the definition of the sigmoid function:

$$L = -\log\left(\frac{1}{1+e^{-\Delta s}}\right)$$

Using the logarithm property $\log(\frac{1}{x}) = -\log(x)$:

$$L = -(-\log(1 + e^{-\Delta s})) \; L = \log(1 + e^{-\Delta s})$$

Now, let's relate this back to the Bradley-Terry derived negative log-likelihood:

$$L_{BT} = \log(e^{s(\text{chosen})} + e^{s(\text{rejected})}) - s(\text{chosen})$$

Let's manipulate this expression:

$L_{BT} = \log(e^{s(\text{rejected})}(e^{s(\text{chosen})-s(\text{rejected})} + 1)) - s(\text{chosen})\ L_{BT} = \log(e^{s(\text{rejected})}) + \log(e^{\Delta s} + 1) - s(\text{chosen})\ L_{BT} = s(\text{rejected}) + \log(e^{\Delta s} + 1) - s(\text{chosen})\ L_{BT} = \log(e^{\Delta s} + 1) - (s(\text{chosen}) - s(\text{rejected}))\ L_{BT} = \log(e^{\Delta s} + 1) - \Delta s$

Now, let's go back to our sigmoid-based loss $L = \log(1 + e^{-\Delta s})$.

Using the property $e^{-\Delta s} = \frac{1}{e^{\Delta s}}$:

$L = \log\left(1 + \frac{1}{e^{\Delta s}}\right)\ L = \log\left(\frac{e^{\Delta s}+1}{e^{\Delta s}}\right)\ L = \log(e^{\Delta s} + 1) - \log(e^{\Delta s})\ L = \log(e^{\Delta s} + 1) - \Delta s$

This is the same formula as L_{BT}!

Conclusion:

The formula $L = -\log(\sigma(s(\text{chosen}) - s(\text{rejected})))$ is indeed a form of the negative log-likelihood loss for pairwise comparisons, and it is mathematically equivalent to the negative log-likelihood derived directly from the Bradley-Terry probability formulation.

So, while the RewardTrainer might use this specific formulation with the sigmoid function, it is fundamentally implementing a loss that aligns with the principles of the Bradley-Terry model, aiming to maximize the likelihood of the chosen response having a higher score.