# A. Appendix

## A.1. Analysis of Reinforcement Learning

We provide the detailed derivation of the data source and gradient coefficient (algorithm and reward function) across various methods, including SFT, RFT, Online RFT, DPO, PPO, and GRPO.

### *A.1.1. Supervised Fine-tuning*

The objective of Supervised Fine-tuning is maximizing the following objective:

$$\mathcal{J}_{SFT}(\theta) = \mathbb{E}[q, o \sim P_{sft}(Q, O)] \left( \frac{1}{|o|} \sum_{t=1}^{|o|} \log \pi_\theta(o_t|q, o_{<t}) \right). \tag{6}$$

The gradient of $\mathcal{J}_{SFT}(\theta)$ is:

$$\nabla_\theta \mathcal{J}_{SFT} = \mathbb{E}[q, o \sim P_{sft}(Q, O)] \left( \frac{1}{|o|} \sum_{t=1}^{|o|} \nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) \right). \tag{7}$$

Data Source: The dataset employed for SFT. Reward Function: This can be regarded as human selection. Gradient Coefficient: always set to 1.

### *A.1.2. Rejection Sampling Fine-tuning*

Rejection Sampling Fine-tuning first samples multiple outputs from the supervised fine-tuned LLMs for each question, and then trains LLMs on the sampled outputs with the correct answer. Formally, the objective of RFT is to maximize the following objectives:

$$\mathcal{J}_{RFT}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), o \sim \pi_{sft}(O|q)] \left( \frac{1}{|o|} \sum_{t=1}^{|o|} \mathbb{I}(o) \log \pi_\theta(o_t|q, o_{<t}) \right). \tag{8}$$

The gradient of $\mathcal{J}_{RFT}(\theta)$ is:

$$\nabla_\theta \mathcal{J}_{RFT}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), o \sim \pi_{sft}(O|q)] \left( \frac{1}{|o|} \sum_{t=1}^{|o|} \mathbb{I}(o) \nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) \right). \tag{9}$$

Data Source: question in SFT dataset with outputs sampled from SFT model. Reward Function: Rule (whether the answer is correct or not). Gradient Coefficient:

$$GC_{RFT}(q, o, t) = \mathbb{I}(o) = \begin{cases} 1 & \text{the answer of o is correct} \\ 0 & \text{the answer of o is incorrect} \end{cases} \tag{10}$$

### *A.1.3. Online Rejection Sampling Fine-tuning*

The only difference between RFT and Online RFT is that the outputs of Online RFT are sampled from the real-time policy model $\pi_\theta$, rather than from the SFT model $\pi_{\theta_{sft}}$. Therefore, the gradient of online RFT is:

$$\nabla_\theta \mathcal{J}_{OnRFT}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), o \sim \pi_\theta(O|q)] \left( \frac{1}{|o|} \sum_{t=1}^{|o|} \mathbb{I}(o) \nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) \right). \tag{11}$$

### A.1.4. Direct Preference Optimization (DPO)

The objective of DPO is:

$$\mathcal{J}_{DPO}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), o^+, o^- \sim \pi_{sft}(O|q)] \log \sigma\left(\beta \frac{1}{|o^+|} \sum_{t=1}^{|o^+|} \log \frac{\pi_\theta(o_t^+|q, o_{<t}^+)}{\pi_{\text{ref}}(o_t^+|q, o_{<t}^+)} - \beta \frac{1}{|o^-|} \sum_{t=1}^{|o^-|} \log \frac{\pi_\theta(o_{<t}^-|q, o_{<t}^-)}{\pi_{\text{ref}}(o_{<t}^-|q, o_{<t}^-)}\right) \tag{12}$$

The gradient of $\mathcal{J}_{DPO}(\theta)$ is:

$$\begin{aligned}\nabla_\theta \mathcal{J}_{DPO}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), o^+, o^- \sim \pi_{sft}(O|q)] \Bigg( & \frac{1}{|o^+|} \sum_{t=1}^{|o^+|} GC_{DPO}(q, o, t) \nabla_\theta \log \pi_\theta(o_t^+|q, o_{<t}^+) \\ & - \frac{1}{|o^-|} \sum_{t=1}^{|o^-|} GC_{DPO}(q, o, t) \nabla_\theta \log \pi_\theta(o_t^-|q, o_{<t}^-)\Bigg)\end{aligned} \tag{13}$$

Data Source: question in SFT dataset with outputs sampled from SFT model. Reward Function: human preference in the general domain (can be 'Rule' in mathematical tasks). Gradient Coefficient:

$$GC_{DPO}(q, o, t) = \sigma\left(\beta \log \frac{\pi_\theta(o_t^-|q, o_{<t}^-)}{\pi_{\text{ref}}(o_t^-|q, o_{<t}^-)} - \beta \log \frac{\pi_\theta(o_t^+|q, o_{<t}^+)}{\pi_{\text{ref}}(o_t^+|q, o_{<t}^+)}\right) \tag{14}$$

### A.1.5. Proximal Policy Optimization (PPO)

The objective of PPO is:

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min\left[\frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip}\left(\frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1-\varepsilon, 1+\varepsilon\right) A_t\right]. \tag{15}$$

To simplify the analysis, it is assumed that the model only has a single update following each exploration stage, thereby ensuring that $\pi_{\theta_{old}} = \pi_\theta$. In this case, we can remove the min and clip operation:

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t. \tag{16}$$

The gradient of $\mathcal{J}_{PPO}(\theta)$ is:

$$\nabla_\theta \mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} A_t \nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) \tag{17}$$

Data Source: question in SFT dataset with outputs sampled from policy model. Reward Function: reward model. Gradient Coefficient:

$$GC_{PPO}(q, o, t, \pi_{\theta_{rm}}) = A_t, \tag{18}$$

where $A_t$ is the advantage, which is computed by applying Generalized Advantage Estimation (GAE) (Schulman et al., 2015), based on the rewards $\{r_{\geq t}\}$ and a learned value function $V_\psi$.

### A.1.6. Group Relative Policy Optimization (GRPO)

The objective of GRPO is (assume $\pi_{\theta_{old}} = \pi_\theta$ for simplified analysis):

$$\begin{aligned}\mathcal{J}_{GRPO}(\theta) = & \mathbb{E}[q \sim P_{sft}(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \\ & \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left[\frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t} - \beta\left(\frac{\pi_{ref}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - \log \frac{\pi_{ref}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - 1\right)\right].\end{aligned} \tag{19}$$

The gradient of $\mathcal{J}_{GRPO}(\theta)$ is:

$$\begin{aligned}\nabla_\theta \mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P_{sft}(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)] \\ \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left[\hat{A}_{i,t} + \beta\left(\frac{\pi_{ref}(o_{i,t}|o_{i,<t})}{\pi_\theta(o_{i,t}|o_{i,<t})} - 1\right)\right]\nabla_\theta \log \pi_\theta(o_{i,t}|q, o_{i,<t}).\end{aligned} \tag{20}$$

Data Source: question in SFT dataset with outputs sampled from policy model. Reward Function: reward model. Gradient Coefficient:

$$GC_{GRPO}(q, o, t, \pi_{\theta_{rm}}) = \hat{A}_{i,t} + \beta\left(\frac{\pi_{ref}(o_{i,t}|o_{i,<t})}{\pi_\theta(o_{i,t}|o_{i,<t})} - 1\right), \tag{21}$$

where $\hat{A}_{i,t}$ is computed based on the group reward scores.