A Unified Paradigm: SFT, RFT, DPO, Online RFT, PPO, GRPO

Main Reference: DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Key Concepts:

Supervised Fine-tuning (SFT) Rejection Sampling Fine-tuning (RFT) Direct Preference Optimization (DPO) Online Rejection Sampling Fine-tuning (Online RFT) Proximal Policy Optimization (PPO) Group Relative Policy Optimization (GRPO)

Table of Contents:

🗄 **A Unified Paradigm: SFT, RFT, DPO, Online RFT, PPO, GRPO**
🧩 **Supervised Fine-tuning (SFT)**
🎯 **Rejection Sampling Fine-tuning (RFT)**
⏳ **Online Rejection Sampling Fine-tuning (Online RFT)**
⚖ **Direct Preference Optimization (DPO)**
🚀 **Proximal Policy Optimization (PPO)**
🎼 **Group Relative Policy Optimization (GRPO)**

A Unified Paradigm: SFT, RFT, DPO, Online RFT, PPO, GRPO

- **Supervised Fine-tuning (SFT):** Fine-tunes a pretrained model on human-curated SFT data.
- **Rejection Sampling Fine-tuning (RFT):** Further fine-tunes the SFT model on filtered outputs sampled from it, keeping only answers deemed correct.
- **Direct Preference Optimization (DPO):** Refines the SFT model using augmented outputs sampled from it, optimized with pairwise DPO loss.
- **Online Rejection Sampling Fine-tuning (Online RFT):** Similar to RFT, but samples outputs from the real-time policy model (initialized from the SFT model).
- **PPO / GRPO:** Initialize the policy model from the SFT model and reinforce it with outputs sampled from the real-time policy model.

General Gradient Formulation

In general, the gradient of a training method $\mathcal{A}$ with respect to the parameter θ can be written as:

$$\nabla_\theta J_{\mathcal{A}}(\theta) = \underbrace{\mathbb{E}[(q,o) \sim \mathcal{D}]}_{\text{Data Source}} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} \underbrace{GC_{\mathcal{A}}(q,o,t,\pi_{rf})}_{\text{Gradient Coefficient}} \nabla_\theta \log \pi_\theta(o_t \mid q, o_{<t}) \right]$$

Key components:

- **Data Source** $\mathcal{D}$: defines the training data.
- **Reward Function** π_{rf}: provides the reward signal during training.
- **Algorithm** $\mathcal{A}$: processes data and reward to compute the gradient coefficient GC, which determines the strength of reinforcement or penalty.

Table 1 | Data Source and Gradient Coefficient of Different Methods

Methods	Data Source	Objective	Reward Function	Gradient Coefficient
SFT	$q, o \sim P_{\text{sft}}(Q, O)$	Eq. (1)	–	1
RFT	$q \sim P_{\text{sft}}(Q), o \sim \pi_{\text{sft}}(O \mid q)$	Eq. (3)	Rule	Eq. (5)
DPO	$q \sim P_{\text{sft}}(Q), o^+, o^- \sim \pi_{\text{sft}}(O \mid q)$	Eq. (6)	Rule	Eq. (5)
Online RFT	$q \sim P_{\text{sft}}(Q), o \sim \pi_\theta(O \mid q)$	Eq. (8)	Rule	Eq. (10)
PPO	$q \sim P_{\text{sft}}(Q), o \sim \pi_\theta(O \mid q)$	Eq. (11)	Model	Eq. (14)
GRPO	$q \sim P_{\text{sft}}(Q), \{o_i\}_{i=1}^G \sim \pi_\theta(O \mid q)$	Eq. (15)	Model	Eq. (17)

📌 **Notes**:

- P_{sft}:= supervised fine-tuning dataset distribution.
- π_{sft}:= supervised fine-tuned model.
- π_θ:= real-time policy model during online training.
- o := a sampled output sequence (e.g., a generated answer).
- o^+:= the preferred (or higher-quality) output in human-labeled preference pairs.
- o^-:= the less-preferred (or lower-quality) output in preference pairs.

Supervised Fine-tuning (SFT)

Objective: The goal is to maximize

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

Gradient:

$$\nabla_\theta J_{\mathrm{SFT}}(\theta) = \mathbb{E}_{(q,o)\sim P_{\mathrm{sft}}(Q,O)} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} \nabla_\theta \log \pi_\theta(o_t \mid q, o_{<t}) \right] \tag{2}$$

- **Data Source:** SFT dataset
- **Reward Function:** Human selection
- **Gradient Coefficient:** Always 1

Rejection Sampling Fine-tuning (RFT)

Objective: Multiple outputs are first sampled from the SFT model for each question. The model is then trained on the sampled outputs that correspond to the correct answers:

$$J_{\mathrm{RFT}}(\theta) = \mathbb{E}_{q\sim P_{\mathrm{sft}}(Q),\, o\sim \pi_{\mathrm{sft}}(O|q)} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} I(o) \log \pi_\theta(o_t \mid q, o_{<t}) \right] \tag{3}$$

Gradient:

$$\nabla_\theta J_{\mathrm{RFT}}(\theta) = \mathbb{E}_{q\sim P_{\mathrm{sft}}(Q),\, o\sim \pi_{\mathrm{sft}}(O|q)} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} I(o) \nabla_\theta \log \pi_\theta(o_t \mid q, o_{<t}) \right] \tag{4}$$

- **Data Source:** Questions from the SFT dataset with outputs sampled from the SFT model
- **Reward Function:** Rule-based (answer correctness)
- **Gradient Coefficient:**

$$GC_{\mathrm{RFT}}(q,o,t) = I(o) = \begin{cases} 1, & \text{if answer of } o \text{ is correct} \\ 0, & \text{if answer of } o \text{ is incorrect} \end{cases} \tag{5}$$

Online Rejection Sampling Fine-tuning (Online RFT)

The only difference from RFT is that outputs are sampled from the **real-time policy model** π_θ, instead of the SFT model π_{sft}:

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

Gradient:

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

Direct Preference Optimization (DPO)

Objective:

$$J_{\mathrm{DPO}}(\theta) = \mathbb{E}_{q\sim P_{\mathrm{sft}}(Q),\, o^+,o^-\sim \pi_{\mathrm{sft}}(O|q)} \left[\log \sigma\Big(\beta \frac{1}{|o^+|} \sum_{t=1}^{|o^+|} \log \frac{\pi_\theta(o_t^+ \mid q, o_{<t}^+)}{\pi_{\mathrm{ref}}(o_t^+ \mid q, o_{<t}^+)} - \beta \frac{1}{|o^-|} \sum_{t=1}^{|o^-|} \log \frac{\pi_\theta(o_t^- \mid q, o_{<t}^-)}{\pi_{\mathrm{ref}}(o_t^- \mid q, o_{<t}^-)} \Big) \right] \tag{8}$$

Gradient:

$$\begin{aligned} \nabla_\theta J_{\mathrm{DPO}}(\theta) = \mathbb{E}_{q\sim P_{\mathrm{sft}}(Q),\, o^+,o^-\sim \pi_{\mathrm{sft}}(O|q)} \Bigg[& \frac{1}{|o^+|} \sum_{t=1}^{|o^+|} GC_{\mathrm{DPO}}(q,o,t) \nabla_\theta \log \pi_\theta(o_t^+ \mid q, o_{<t}^+) \\ & - \frac{1}{|o^-|} \sum_{t=1}^{|o^-|} GC_{\mathrm{DPO}}(q,o,t) \nabla_\theta \log \pi_\theta(o_t^- \mid q, o_{<t}^-) \Bigg] \end{aligned} \tag{9}$$

- **Data Source:** Questions in the SFT dataset with outputs sampled from the SFT model
- **Reward Function:** Human preference (or rule-based for math tasks)
- **Gradient Coefficient:**

$$GC_{\mathrm{DPO}}(q,o,t) = \sigma\Big(\beta \log \frac{\pi_\theta(o_t^- \mid q, o_{<t}^-)}{\pi_{\mathrm{ref}}(o_t^- \mid q, o_{<t}^-)} - \beta \log \frac{\pi_\theta(o_t^+ \mid q, o_{<t}^+)}{\pi_{\mathrm{ref}}(o_t^+ \mid q, o_{<t}^+)} \Big) \tag{10}$$

Proximal Policy Optimization (PPO)

Objective:

$$J_{\text{PPO}}(\theta) = \mathbb{E}_{q\sim P_{\text{sft}}(Q),\ o\sim\pi_\theta^{old}(O|q)}\left[\frac{1}{|o|}\sum_{t=1}^{|o|}\min\left(\frac{\pi_\theta(o_t\mid q,o_{<t})}{\pi_\theta^{old}(o_t\mid q,o_{<t})}A_t,\ \text{clip}\left(\frac{\pi_\theta(o_t\mid q,o_{<t})}{\pi_\theta^{old}(o_t\mid q,o_{<t})},1-\epsilon,1+\epsilon\right)A_t\right)\right] \tag{11}$$

Simplification: If we assume a single update step such that $\pi_\theta^{old} = \pi_\theta$, the objective reduces to:

$$J_{\text{PPO}}(\theta) = \mathbb{E}_{q\sim P_{\text{sft}}(Q),\ o\sim\pi_\theta^{old}(O|q)}\left[\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t\mid q,o_{<t})}{\pi_\theta^{old}(o_t\mid q,o_{<t})}A_t\right] \tag{12}$$

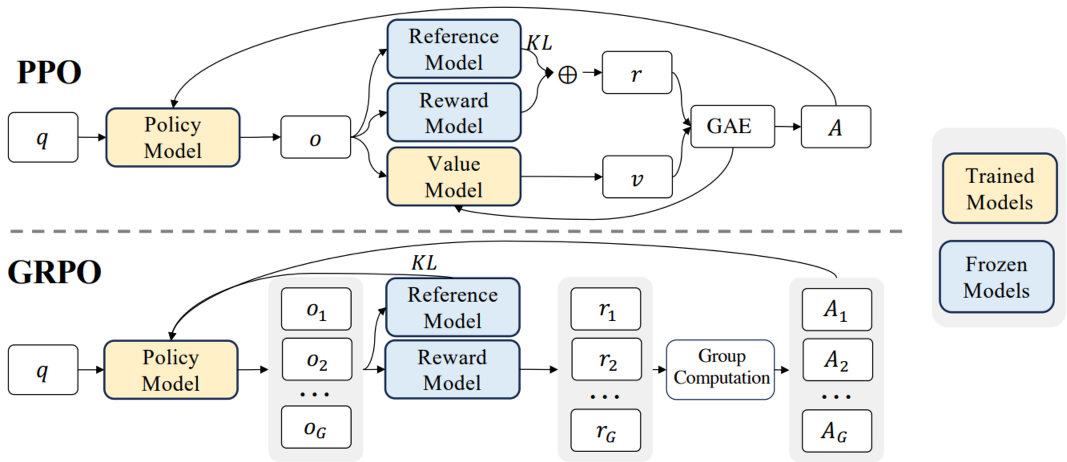
Gradient:

$$\nabla_\theta J_{\text{PPO}}(\theta) = \mathbb{E}_{q\sim P_{\text{sft}}(Q),\ o\sim\pi_\theta^{old}(O|q)}\left[\frac{1}{|o|}\sum_{t=1}^{|o|}A_t\nabla_\theta\log\pi_\theta(o_t\mid q,o_{<t})\right] \tag{13}$$

- **Data Source:** Questions in the SFT dataset with outputs sampled from the policy model
- **Reward Function:** Reward model
- **Gradient Coefficient:**

$$GC_{\text{PPO}}(q,o,t,\pi_\theta^{rm}) = A_t, \tag{14}$$

where A_t denotes the *advantage function*, computed using **Generalized Advantage Estimation (GAE)** based on rewards $r_{\geq t}$ and a learned value function V_ψ. See the below figure for a reference:



Group Relative Policy Optimization (GRPO)

Objective (also assume $\pi_\theta^{old} = \pi_\theta$ for simplified analysis):

$$\begin{aligned} J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P_{\text{sft}}(Q),\ \{o_i\}_{i=1}^G\sim\pi_\theta^{old}(O|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Bigg(&\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_\theta^{old}(o_{i,t}\mid q,o_{i,<t})}\hat{A}_{i,t} \\ &-\beta\left(\frac{\pi_{ref}(o_{i,t}\mid q,o_{i,<t})}{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}-\log\frac{\pi_{ref}(o_{i,t}\mid q,o_{i,<t})}{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}-1\right)\Bigg)\Bigg] \end{aligned} \tag{15}$$

Gradient:

$$\nabla_\theta J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P_{\text{sft}}(Q),\ \{o_i\}_{i=1}^G\sim\pi_\theta^{old}(O|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\hat{A}_{i,t}+\beta\left(\frac{\pi_{ref}(o_{i,t}\mid o_{i,<t})}{\pi_\theta(o_{i,t}\mid o_{i,<t})}-1\right)\right)\nabla_\theta\log\pi_\theta(o_{i,t}\mid q,o_{i,<t})\right] \tag{16}$$

- **Data Source:** Questions in the SFT dataset with outputs sampled from the policy model
- **Reward Function:** Reward model
- **Gradient Coefficient:**

$$GC_{\text{GRPO}}(q,o,t,\pi_\theta^{rm}) = \hat{A}_{i,t}+\beta\left(\frac{\pi_{ref}(o_{i,t}\mid o_{i,<t})}{\pi_\theta(o_{i,t}\mid o_{i,<t})}-1\right), \tag{17}$$

where $\hat{A}_{i,t}$ is the advantage term computed from **group reward scores**.