Reinforcement Learning from Verifiable Rewards

RL from Verifiable Rewards (RLVR)

Human Feedback versus Verifiable Feedback

- Human annotation is expensive.
- Some tasks admit **automatic correctness checks**.

Examples of verifiable domains

- mathematics and symbolic reasoning
- code generation with unit tests
- theorem proving
- structured tasks with exact-match or checker-based evaluation

This motivates RLVR

Use **reinforcement learning with verifiable rewards** when the task itself can supply the feedback.

RL from Verifiable Rewards

Consider an episodic MDP $(\mathcal{S}, \mathcal{A}, P, H)$. The policy $\pi_\theta(a \mid s)$.

- Given a trajectory $\tau = (s_1, a_1, \ldots, s_H, a_H)$, we assume the trajectory of τ can be verified:

$$r_{\text{ver}}(\tau) = \text{Verifier}(\tau) \in \{0, 1\}.$$

- For example, reasoning tasks where final answers and proof steps can be checked.

RLVR Objective

$$\max_{\pi} \mathbb{E}_{\tau \sim \pi}[r_{\text{ver}}(\tau)] - \beta \, \mathbb{E}_{s \sim d^\pi} \left[\mathrm{KL}\big(\pi(\cdot|s) \| \pi_{\text{ref}}(\cdot|s)\big) \right].$$

Same regularized RL form, but the reward source is now external and verifiable.

RL from Verifiable Rewards

PPO-Clip

Consider a modified objective

$$\max_{\tilde{\theta}} \mathbb{E}_{\rho_\theta, a\sim\pi_\theta}\left[\min\left\{\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)}A_w(s,a), \left(\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)}\right)_{1-\epsilon}^{1+\epsilon} A_w(s,a)\right\}\right]$$

RL from Verifiable Rewards

PPO-Clip

Consider a modified objective

$$\max_{\tilde{\theta}} \mathbb{E}_{\rho_\theta, a\sim\pi_\theta}\left[\min\left\{\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)}A_w(s,a), \left(\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)}\right)_{1-\epsilon}^{1+\epsilon} A_w(s,a)\right\}\right]$$

In RLVR, there is only a verifiable reward $r_{\text{ver}}(\tau)$ obtained after the whole trajectory is produced, so performing TD to compute $A_w(s,a)$ is nontrivial.

From PPO to GRPO

GRPO = Group Relative Policy Optimization

- For each initial state, sample a **group** of outputs from the current policy.
- Obtain the verifiable reward from a verifier for each trajectory.
- Convert absolute rewards into **relative within-group advantages**.

Main idea

Instead of estimating a separate value function baseline (Critic), compare candidate outputs *against each other inside the same group*.

Group-relative advantage construction

Suppose the initial state s_1 produces N trajectories $\mathcal{D}_{s_1} = \{\tau_1, \ldots, \tau_N\}$ with verifiable rewards $r_1, \ldots, r_N$. A GRPO normalized advantage is

$$A_i = \frac{r_i - \mathrm{mean}(r_1, \ldots, r_N)}{\mathrm{std}(r_1, \ldots, r_N)}.$$

Group-relative advantage construction

Suppose the initial state s_1 produces N trajectories $\mathcal{D}_{s_1} = \{\tau_1, \ldots, \tau_N\}$ with verifiable rewards $r_1, \ldots, r_N$. A GRPO normalized advantage is

$$A_i = \frac{r_i - \mathrm{mean}(r_1, \ldots, r_N)}{\mathrm{std}(r_1, \ldots, r_N)}.$$

GRPO learns from **relative quality**, not just from raw absolute reward.

- Positive advantage for above-average outputs.
- Negative advantage for below-average outputs.
- Same initial state acts as a local control variate to reduce variance.

GRPO objective

GRPO

$$\max_{\tilde{\theta}} \mathbb{E}_{\substack{s_1 \sim \rho_0 \\ a \sim \pi_\theta}} \left[\frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} \left(\min \left\{ \frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)} A_i, \left(\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)} \right)_{1-\epsilon}^{1+\epsilon} A_i \right\} - \mathrm{KL}\big(\pi(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s)\big) \right) \right]$$

- Structurally close to PPO
- Advantage comes from group-relative comparison
- Often avoids a learned critic/value model

Zeroth-Order Policy Optimization for RLHF

Preference and Zeroth-Order Optimization

Optimization

$$\min_{x \in \mathbb{R}^n} f(x).$$

for some function $f : \mathbb{R}^n \to \mathbb{R}$.

Gradient descent (GD)

$$x_{k+1} = x_k - h \nabla f(x_k).$$

with stepsize (learning rate) $h > 0$.

Preference and Zeroth-Order Optimization

Zeroth-Order Method: Gradient Free

- Sample $u_k \sim \mathcal{N}(0, \mathbf{I})$ and form the estimator

$$g_\mu(x_k; u_k) := \frac{f(x_k + \mu u_k) - f(x_k)}{\mu} u_k.$$

- Update

$$x_{k+1} = x_k - h\, g_\mu(x_k; u_k).$$

Preference and Zeroth-Order Optimization

Zeroth-Order Method: Gradient Free

- Sample $u_k \sim \mathcal{N}(0, \mathbf{I})$ and form the estimator

$$g_\mu(x_k; u_k) := \frac{f(x_k + \mu u_k) - f(x_k)}{\mu} u_k.$$

- Update

$$x_{k+1} = x_k - h\, g_\mu(x_k; u_k).$$

Connection to RLHF

Recall the Bradley-Terry Model.

$$\mathbb{P}((x_k + \mu u_k) \succ x_k) = \sigma(f(x_k + \mu u_k) - f(x_k)).$$

Preference and Zeroth-Order Optimization

Why does a random perturbation work?

Connecting to the gradient method, we expect $g_\mu(\cdot)$ to be the gradient of something. Define a Gaussian-smoothed function

$$f_\mu(x) := \mathbb{E}_u\big[f(x + \mu u)\big].$$

We will show that

$$\nabla f_\mu(x) = \mathbb{E}_u\big[g_\mu(x; u)\big].$$

Proof

Since $u \sim \mathcal{N}(0, \mathbf{I})$, the pdf of u_k is

$$\frac{1}{(2\pi)^{n/2}} \exp(-\frac{||u||^2}{2}).$$

Therefore, we have

$$\begin{aligned}\nabla f_\mu(x) &= \nabla \mathbb{E}_u[f(x+\mu u)] \\ &= \nabla \int f(x+\mu u) \frac{1}{(2\pi)^{n/2}} \exp(-\frac{||u||^2}{2}) du\end{aligned}$$

Proof

Define $y = x + \mu u$. We have

$$
\begin{aligned}
&\nabla f_\mu(x) \\
=&\nabla_x \int f(y) \frac{1}{(2\pi)^{n/2}} \exp(-\frac{||y-x||^2}{2\mu^2}) d(\frac{y-x}{\mu}) \\
=&\nabla_x \int f(y) \frac{1}{(2\pi)^{n/2}} \exp(-\frac{||y-x||^2}{2\mu^2}) \frac{1}{\mu^n} dy \\
=&\int f(y) \frac{1}{(2\pi)^{n/2}} \exp(-\frac{||y-x||^2}{2\mu^2})(-\frac{1}{2\mu^2})(2(y-x))(-1)\frac{1}{\mu^n} dy \\
=&\int f(y) \frac{1}{(2\pi)^{n/2}} \exp(-\frac{||y-x||^2}{2\mu^2}) \frac{y-x}{\mu^2} \frac{1}{\mu^n} dy \\
=&\int f(x+\mu u) \frac{1}{(2\pi)^{n/2}} \exp(-\frac{||u||^2}{2}) \frac{u}{\mu} du
\end{aligned}
$$

Proof

Note that $\frac{1}{\mu}\mathbb{E}[f(x)u] = 0$. Therefore,

$$\begin{aligned}
&\nabla f_\mu(x) \\
=&\nabla f_\mu(x) - \frac{1}{\mu}\mathbb{E}[f(x)u] \\
=&\int \frac{f(x+\mu u) - f(x)}{\mu} u \frac{1}{(2\pi)^{n/2}} \exp(-\frac{||u||^2}{2}) du \\
=&\mathbb{E}_u[g_\mu(x;u)]
\end{aligned}$$

Stochastic Zeroth-Order Policy Optimization (SZPO) for RLHF

Without reward signals, it is difficult to compute the gradient $\nabla_\theta V(\pi_\theta)$. But we can try a small perturbation:

$$\theta' = \theta + \mu v,$$

where $v \sim \mathcal{N}(0, \mathbb{I})$ is a random direction.

RLHF

Preference feedback can be turned into a policy improvement direction.

SZPO

For each iteration $t = 1, \ldots, T$:

1. Sample a random direction v_t and form a perturbed policy $\theta'_t = \theta_t + \mu_t v_t$.
2. For each comparison $n = 1, \ldots, N$:
   - sample a batch of trajectories from π_{θ_t};
   - sample a batch of trajectories from $\pi_{\theta'_t}$;
   - query M human evaluators which batch is better and record $o_{t,n,m} \in \{0, 1\}$;
3. Aggregate to policy preference to find the improvement direction

$$\hat{g}_t = s_t v_t.$$

4. Update the policy:

$$\theta_{t+1} = \theta_t + \alpha_t \hat{g}_t.$$

SZPO with a Known Preference Model

Recall

$$x_{k+1} = x_k - h\, g_\mu(x_k; u_k).$$

with

$$g_\mu(x_k; u_k) := \frac{f(x_k + \mu u_k) - f(x_k)}{\mu}\, u_k.$$

So

$$s_t = \frac{V_{\theta'_t} - V_{\theta_t}}{\mu_t} = \frac{\mathbb{E}[r(\tau_{\theta'_t})] - \mathbb{E}[r(\tau_{\theta_t})]}{\mu_t}$$

SZPO with a Known Preference Model

Zeroth-Order

$$s_t = \frac{\mathbb{E}[r(\tau_{\theta'_t}) - r(\tau_{\theta_t})]}{\mu_t}$$

Bradley-Terry Model

$$\mathbb{P}(\tau_{\theta'_t} \succ \tau_{\theta_t}) = \sigma(r(\tau_{\theta'_t}) - r(\tau_{\theta_t})).$$

SZPO with a Known Preference Model

Zeroth-Order

$$s_t = \frac{\mathbb{E}[r(\tau_{\theta'_t}) - r(\tau_{\theta_t})]}{\mu_t}$$

Bradley-Terry Model

$$\mathbb{P}(\tau_{\theta'_t} \succ \tau_{\theta_t}) = \sigma(r(\tau_{\theta'_t}) - r(\tau_{\theta_t})).$$

- Use M human evaluations to estimate $\hat{\mathbb{P}}(\tau_{\theta'_t} \succ \tau_{\theta_t})$
- Use the link function to estimate the value difference $r(\tau_{\theta'_t}) - r(\tau_{\theta_t}) \approx \sigma^{-1}(\hat{\mathbb{P}}(\tau_{\theta'_t} \succ \tau_{\theta_t}))$.
- Use N pairs of batches to estimate $s_t \approx \frac{1}{\mu_t N} \sum_n \sigma^{-1}(\hat{\mathbb{P}}_n(\tau_{\theta'_t} \succ \tau_{\theta_t}))$

SZPO with an Unknown Preference Model

Preference Model ???

Zeroth-Order

s_t =???

SZPO with an Unknown Preference Model

Preference Model ???

Zeroth-Order

s_t =???

Sign-SZPO

$$s_t = \text{sign}\left(\mathbb{E}[r(\tau_{\theta'_t}) - r(\tau_{\theta_t})]\right)$$

- For each pair of batches, obtain preference o_n, and use majority vote to estimate the sign.

SZPO

For each iteration $t = 1, \ldots, T$:

1. Sample a random direction v_t and form a perturbed policy $\theta'_t = \theta_t + \mu_t v_t$.
2. For each comparison $n = 1, \ldots, N$:
   - sample a batch of trajectories from π_{θ_t};
   - sample a batch of trajectories from $\pi_{\theta'_t}$;
   - query M human evaluators which batch is better and record $o_{t,n,m} \in \{0, 1\}$;
3. Aggregate to policy preference to find the improvement direction

$$\hat{g}_t = s_t v_t.$$

4. Update the policy:

$$\theta_{t+1} = \theta_t + \alpha_t \hat{g}_t.$$

Both SZPO and Sign-SZPO converge.

Reference

- Nesterov, Yurii, and Vladimir Spokoiny. Random gradient-free minimization of convex functions. Foundations of Computational Mathematics 17, no. 2 (2017): 527-566.
- Z. Shao, *et al.* Deepseekmath: pushing the limits of mathematical reasoning in open language models, 2024.
- Q. Zhang, L. Ying. Zeroth-order policy gradient for reinforcement learning from human feedback without reward inference. ICLR, 2025.
- Q. Zhang, L. Ying. Provable Reinforcement Learning from Human Feedback with an Unknown Link Function. arXiv:2506.03066, 2025.