Reinforcement Learning for LLMs

Karl Stratos

Last updated: November, 2026

Contents

1 Generative Story **3**

2 REINFORCE **3**
2.1 Surrogate Objective . . . 3
2.2 Relation to Importance Sampling . . . 4
2.3 Relation to Rejection Sampling . . . 4

3 Trust-Region Methods **4**
3.1 Surrogate Objective . . . 4
3.2 First-Order Exactness . . . 5

4 Trust Region Policy Optimization (TRPO) **5**
4.1 Natural Policy Gradient . . . 5
4.2 Algorithm . . . 6

5 Proximal Policy Optimization (PPO) **6**
5.1 Asymmetric Clipping Effect . . . 6
5.2 Relative Clipping Effect . . . 7
5.3 Length Normalization . . . 7
5.4 KL Variant . . . 7

6 Group Relative Policy Optimization (GRPO) **8**
6.1 Dr. GRPO . . . 8
6.2 Dynamic Sampling (DAPO) . . . 9
6.3 Group Sequence Policy Optimization (GSPO) . . . 9

7 Mirror Descent Policy Optimization (MDPO) **9**
7.1 Kimi k1.5 Variant . . . 10

8 Divergence Proximal Policy Optimization (DPPO) **11**
8.1 TRPO Motivation . . . 11
8.2 Average Bound . . . 11

9 Training Inference Mismatch (TIM) **12**
9.1 Importance Sampling . . . 12
9.2 Expert Replay . . . 12
9.3 Causes of TIM . . . 12

10 Score Centering (SC) **13**
10.1 Top-k Approximation . . . 13
10.2 Importance Sampling Hybrid . . . 14

11 Advantage Estimation **15**
11.1 Critic . . . 15
11.2 Generalized advantage estimator (GAE) . . . 15

A Markov Decision Process (MDP) **17**
A.1 Training 18

B KL Estimation **18**

C Lemmas **18**

1 Generative Story

An environment issues an initial state $s_1 \sim \mathbf{env}$ consisting of system/user prompts. For $t = 1 \dots T$ where $T \le T_{\max}$ is a bounded random termination step, do

$$\begin{aligned} x_t &\sim p_\theta(\cdot|s_t) && \text{(token } x_t \in \mathcal{V}\text{)} \\ (s_t, x_t) &\mapsto s_{t+1} && \text{(state transition)} \end{aligned} \tag{1}$$

The state s_{t+1} contains all generated tokens $x_1 \dots x_t$ as well as all information present in the past states $s_1 \dots s_t$ so that (1) remains autoregressive. The environment issues a final reward for the terminal state $R = R(s_{T+1}) \in \mathbb{R}$. See Appendix A for how this is implemented in practice. The goal is to maximize

$$J(\theta) = \mathop{\mathbf{E}}_{\substack{s_1 \sim \mathbf{env} \\ \tau \sim p_\theta(\cdot|s_1)}} [R(s_{T+1})] \tag{2}$$

where $\tau = (x_1 \dots x_T)$ is the token trajectory (rollout) generated from s_1 yielding a state trajectory $s_2 \dots s_{T+1}$.[1] We sometimes write the reward $R(\tau) = R(s_{T+1})$ in rollouts. Most methods are **policy-gradient** methods, i.e., maximize (2) by following its gradient.

Endpoints. We distinguish the rollout endpoint $K \le T$ from the termination step, where $K = T$ if the rollout is terminated and $K < T$ if truncated (e.g., due to time/length limit). We assume the reward is always zero for nonterminal states, thus $R(s_{K+1}) = 0$ if truncated. For steps beyond genuine termination, we assume that the state is frozen: $s_t = s_{T+1}$ for $T < t \le T_{\max}$.

Denoting random variables. Failing to specify what is random leads to poor readability, but marking every random variable overwhelms notation. As a middle ground, we will only mark what is helpful in reminding us of the generative story at hand. For instance, if we write $\tau \sim p_\theta$ it is understood that the initial state is drawn as usual (i.e., $p_\theta(\tau) = \mathbf{E}_{s_1 \sim \mathbf{env}}[p_\theta(\tau|s_1)]$). If we write $\mathbf{E}_\theta[R|s_t]$, it means the expected final reward R if the model θ starts the session from s_t.

2 REINFORCE

REINFORCE (Williams, 1992) gives a direct gradient estimator for (2). By the log-derivative trick and a property of the score function, for any "baseline" function $b(s_t)$ independent of x_t we can write (Lemma C.1)

$$\nabla J(\theta) = \mathop{\mathbf{E}}_{\tau \sim p_\theta} \left[\sum_{t=1}^{T} (R(s_{T+1}) - b(s_t)) \nabla \log p_\theta(x_t|s_t) \right] \tag{3}$$

The gradient is now samplable and weighted at a token level. We should make $R(s_{T+1}) - b(s_t)$ large if the eventual outcome is unexpectedly good, so it is natural to define

$$b(s_t) = \mathbf{E}_\theta[R|s_t] \tag{4}$$

(the **value function**). It is also approximately the minimizer of per-step total variance (Lemma C.2).

2.1 Surrogate Objective

To use the gradient estimator (3), we can construct a surrogate objective at each θ as

$$J_{\text{rein}}(\theta) = \mathop{\mathbf{E}}_{\tau \sim p_{\text{sg}(\theta)}} \left[\sum_{t=1}^{T} \text{sg}\left(R(s_{T+1}) - b(s_t)\right) \log p_\theta(x_t|s_t) \right] \tag{5}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. It is evident that $\nabla J_{\text{rein}}(\theta) = \nabla J(\theta)$. If we choose (4) as the baseline, by per-step conditioning we can rewrite (5) as

$$J_{\text{rein}}(\theta) = \mathop{\mathbf{E}}_{\tau \sim p_{\text{sg}(\theta)}} \left[\sum_{t=1}^{T} \text{sg}(A_\theta(s_t, x_t)) \log p_\theta(x_t|s_t) \right] \tag{6}$$

where $A_\theta(s_t, x_t) = \mathbf{E}_\theta[R|s_t, x_t] - \mathbf{E}_\theta[R|s_t]$ is the **advantage function** (Section 11).

[1] We assume deterministic state transition for simplicity, but the stochastic case can be immediately handled by taking an outer expectation.

2.2 Relation to Importance Sampling

To remove the model from the expectation in (2), we can simply use importance sampling. Then the derivation of the gradient is more straightforward:

$$J(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)}R(s_{T+1})\right] \quad \Rightarrow \quad \nabla J(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\sum_{t=1}^{T}\frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)}(R(s_{T+1})-b(s_t))\nabla\log p_\theta(x_t|s_t)\right] \tag{7}$$

We see that the gradient reduces to REINFORCE (3) at $\psi = \mathrm{sg}(\theta)$.

2.3 Relation to Rejection Sampling

If $R \in \{0,1\}$ and $b(s_t) = 0$, the surrogate objective (5) becomes

$$J_{\text{rein}}(\theta) = \Pr_{\tau\sim p_{\mathrm{sg}(\theta)}}(R=1) \times \mathop{\mathbf{E}}_{\tau\sim p_{\mathrm{sg}(\theta)}:\, R=1}[\log p_\theta(\tau|s_1)]$$

This is rejection sampling (i.e., supervised fine-tuning on successful trajectories).

3 Trust-Region Methods

REINFORCE (6) is exact but fully on-policy. In reality, the rollouts come from a slightly lagging version of the model ("policy staleness"):

1. An inference engine (**sampler**) uses the current checkpoint $\psi \leftarrow \theta$ to generate a large batch $\mathcal{B}$ of rollouts.
2. A training engine (**trainer**) updates θ on the frozen $\mathcal{B}$ in minibatches, possibly making many epochs. Thus later updates use stale rollouts.

In principle, there is nothing that prevents us from going off-policy. Importance sampling (7) allows us to use arbitrary ψ without losing consistency. The problem is *variance*.[2]

3.1 Surrogate Objective

Trust-region (TR) methods rely on approximation to trade off bias against variance. The target objective is the "improvement" of the current policy $J(\theta) - J(\psi)$. We write it in a few different forms for convenience:

$$J(\theta)-J(\psi) = \mathop{\mathbf{E}}_{\tau\sim p_\theta}\left[\sum_{t=1}^{T}A_\psi(s_t,x_t)\right] = \sum_{t=1}^{T_{\max}}\mathop{\mathbf{E}}_{\substack{s\sim d_{\theta,t}\\ x\sim p_\psi(\cdot|s)}}\left[\frac{p_\theta(x|s)}{p_\psi(x|s)}A_\psi(s,x)\right] = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\sum_{t=1}^{T}\left(\prod_{i=1}^{t}\frac{p_\theta(x_i|s_i)}{p_\psi(x_i|s_i)}\right)A_\psi(s_t,x_t)\right] \tag{8}$$

where $d_{\theta,t}(s) = \Pr_\theta(s_t = s)$. The first equality is known as the **performance difference lemma** (Kakade and Langford, 2002).[3] Note that (8) is just $J(\theta)$ shifted by a constant, so it has the same true gradient $\nabla J(\theta)$ as REINFORCE. But the estimator's variance blows up if ψ is very different from θ. To reduce the variance, TR assumes

$$d_{\theta,t} \approx d_{\psi,t} \tag{9}$$

at every t. Plugging this assumption in the forms above, we have the surrogate objective:

$$J_\psi(\theta) = \sum_{t=1}^{T_{\max}}\mathop{\mathbf{E}}_{\substack{s\sim d_{\psi,t}\\ x\sim p_\psi(\cdot|s)}}\left[\frac{p_\theta(x|s)}{p_\psi(x|s)}A_\psi(s,x)\right] = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\sum_{t=1}^{T}\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}A_\psi(s_t,x_t)\right] \tag{10}$$

Intuitively, we approximate the probability ratio of the whole history by the current token, banking on $\theta \approx \psi$.

[2] This is mainly because it involves unbounded probability ratios. REINFORCE kills them by assuming $\psi = \mathrm{sg}(\theta)$.

[3] Verify easily by noting that the sum of advantages telescopes.

3.2 First-Order Exactness

A property of the TR approximation is that it is first-order exact at $\theta = \psi$. Define

$$\texttt{Bias}_\psi(\theta) = J_\psi(\theta) - \big(J(\theta) - J(\psi)\big) \tag{11}$$

We can easily check $\texttt{Bias}_\psi(\psi) = 0$ and $\nabla\texttt{Bias}_\psi(\psi) = 0_d$. This implies (Taylor-expand around ψ)

$$\texttt{Bias}_\psi(\psi + \Delta) = O(||\Delta||_2^2) \tag{12}$$

This gives (10) a favorable bias-variance tradeoff (i.e., variance reduction at mild inconsistency) since the bias shrinks quadratically rather than linearly as $\Delta \to 0_d$. The bound's quadratic dependence on policy change cannot be improved in general.

4 Trust Region Policy Optimization (TRPO)

TRPO (Schulman *et al.*, 2015b) gives an explicit bound on the bias (11) based on total variation (TV):

$$\mathrm{TV}(P, Q) = \frac{1}{2}\sum_x |P(x) - Q(x)| = 1 - \sum_x \min(P(x), Q(x)) = \min_{(X,Y):\, X\sim P,\, Y\sim Q} \Pr(X \neq Y)$$

The second form follows from the identity $|a - b| = a + b - 2\min(a, b)$. The last form is the classical "coupling" characterization of TV (Lemma C.3). TV is useful for bounding the expectation of a function g under different distributions (exactly what a bias is). Using the first form, by triangle inequality

$$|\mathbf{E}_P[g] - \mathbf{E}_Q[g]| \leq 2\,||g||_\infty\,\mathrm{TV}(P, Q) \tag{13}$$

We can use (13) to easily show, for some constants $C, C' > 0$ (Lemma C.5)

$$\begin{aligned} |\texttt{Bias}_\psi(\theta)| &\leq C \times \Big(\max_s \mathrm{TV}(p_\psi(\cdot|s), p_\theta(\cdot|s))\Big)^2 \\ &\leq C' \times \Big(\max_s \mathrm{KL}(p_\psi(\cdot|s), p_\theta(\cdot|s))\Big) \end{aligned} \tag{14}$$

(the latter is Pinsker's inequality $\mathrm{TV}(P, Q)^2 \leq \frac{1}{2}\mathrm{KL}(P, Q)$). Thus to justify $J_\psi(\theta)$ as a valid lower bound (i.e., $J(\theta) - J(\psi) \geq J_\psi(\theta) - \text{const}$), it is sufficient to have the state-wise bounds

$$\Big(\mathrm{TV}(p_\psi(\cdot|s), p_\theta(\cdot|s)) \leq \epsilon \quad \forall s\Big) \qquad \bigvee \qquad \Big(\mathrm{KL}(p_\psi(\cdot|s), p_\theta(\cdot|s)) \leq \epsilon \quad \forall s\Big) \tag{15}$$

TRPO uses the KL-based constraint but approximates the max with mean for computational reasons. Define[4]

$$\mathrm{KL}_\psi(\theta) = \mathop{\mathbf{E}}_{s_1\sim\mathbf{env}}[\mathrm{KL}(p_\psi(\cdot|s_1), p_\theta(\cdot|s_1))] = \mathop{\mathbf{E}}_{\substack{s_1\sim\mathbf{env} \\ \tau\sim p_\psi(\cdot|s_1)}}\left[\sum_{t=1}^{T} \mathrm{KL}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t))\right] \tag{16}$$

The KL-based TRPO objective is

$$\max_{\theta:\, \mathrm{KL}_\psi(\theta)\leq\epsilon} J_\psi(\theta) \tag{17}$$

4.1 Natural Policy Gradient

(17) can be beautifully simplified by local approximations $\theta = \psi + \Delta$:

$$J_\psi(\psi + \Delta) \approx \nabla J(\psi)^\top \Delta \tag{18}$$

$$\mathrm{KL}_\psi(\psi + \Delta) \approx \frac{1}{2}\Delta^\top I(\psi)\Delta \tag{19}$$

[4]This is different from defining it as $\mathrm{KL}(p_\psi, p_\theta)$ between the marginals. But it is the right divergence since we care about the policies' behavioral similarities given an prompt, not the similarities between their marginalized versions. It is also more computationally convenient since we can use the linearity of expectation. Qi *et al.* (2026) retrospectively justifies constraining this KL term in the average bound (40).

(18) uses the first-order exactness of $J_\psi(\theta)$ at $\theta = \psi$. (19) uses $\nabla \mathrm{KL}_\psi(\psi) = 0_d$ and $\nabla^2 \mathrm{KL}_\psi(\psi) = I(\psi)$ where $I(\psi) = \mathbf{E}_{s_1 \sim \mathbf{env}}[\mathrm{Cov}_{\tau \sim p_\psi(\cdot|s_1)}(\nabla \log p_\psi(\tau|s_1))]$ is the expected conditional Fisher matrix of ψ (easy exercise). This is now steepest ascent with a Fisher-ball constraint, and we get the usual exact update

$$\Delta^\star = \underset{\Delta:\ \frac{1}{2}||\Delta||^2_{I(\psi)} \le \epsilon}{\arg\max} \nabla J(\psi)^\top \Delta = \eta I(\psi)^{-1} \nabla J(\psi) \qquad \left(\eta = \sqrt{\frac{2\epsilon}{||\nabla J(\psi)||^2_{I(\psi)^{-1}}}}\right) \tag{20}$$

The direction $v = I(\psi)^{-1}\nabla J(\psi)$ coincides with the **natural policy gradient** of Kakade (2001) which replaces the parameter space with coordinate-free information geometry.

4.2 Algorithm

We can solve for v in $I(\psi)v = \nabla J(\psi)$ without inverting $I(\psi)$ by conjugate gradient, which only requires an efficient mapping $v \mapsto I(\psi)v$. Since $I(\psi) = \nabla G_\psi(\psi)$ is the Jacobian of the gradient function $G_\psi(\theta) = \nabla \mathrm{KL}_\psi(\theta)$ at $\theta = \psi$,

$$I(\psi)v = \mathbf{JVP}(G_\psi, \psi, v)$$

can be computed by the Jacobian-vector product (JVP).[5] Thus a basic implementation may resemble

1. Estimate $g_\psi = \nabla J(\psi)$, e.g., by REINFORCE (6) with samples $\tau \sim p_\psi$.
2. Estimate $v = I(\psi)^{-1} g_\psi$ by conjugate gradient with the mapping $v \mapsto I(\psi)v = \mathbf{JVP}(G_\psi, \psi, v)$.
3. Compute $\eta = \sqrt{2\epsilon / g_\psi^\top v}$ and update $\theta \leftarrow \psi + \eta v$.

In practice, TRPO additionally performs backtracking line search along v that ensures the actual surrogate objective $J_\psi(\theta)$ improves and the KL constraint is satisfied. Unfortunately, conjugate gradient still requires iteratively computing $v \mapsto I(\psi)v$ per update, so the algorithm remains computationally expensive compared to first-order methods.

5 Proximal Policy Optimization (PPO)

PPO (Schulman *et al.*, 2017) directly maximizes $J_\psi(\theta)$ without hard constraints, but promoting a trust region implicitly. The objective uses per-step clipping:

$$J^{\mathrm{ppo}}_\psi(\theta) = \mathop{\mathbf{E}}_{\tau \sim p_\psi}\left[\sum_{t=1}^{T} \min\left(\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)} A_\psi(s_t, x_t), \mathrm{clip}^{1+\epsilon}_{1-\epsilon}\left(\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}\right) A_\psi(s_t, x_t)\right)\right] \tag{21}$$

It is easy to check that (21) is a lower bound on $J_\psi(\theta)$ and remains first-order exact (Section 3.2). Crucially, the operation $\mathrm{clip}^{1+\epsilon}_{1-\epsilon}(r_t) \in [1-\epsilon, 1+\epsilon]$ passes no gradient through r_t if $r_t = \frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}$ falls out of the trust region. The objective combines two separate clipping effects: (i) asymmetric, and (ii) relative.

5.1 Asymmetric Clipping Effect

Because the advantage can be negative, the contribution of the t-th token to training is dropped iff r_t moves excessively in the *right* direction

$$\Big(A_\psi(s_t, x_t) > 0 \quad \bigwedge \quad r_t > 1+\epsilon\Big) \qquad \bigvee \qquad \Big(A_\psi(s_t, x_t) < 0 \quad \bigwedge \quad r_t < 1-\epsilon\Big) \tag{22}$$

In constrast, t is *not* dropped if r_t moves excessively in the wrong direction. This asymmetric design has the effect of pulling θ back into the trust region only when it is mistaken (Figure 1). An intuition is that we should not penalize the right move even when it is outside the trust region, but we should not keep amplifying it either.

[5] Recall that it avoids explicitly computing the Jacobian $\nabla G_\psi(\psi) \in \mathbb{R}^{d \times d}$ by computing the directional derivatives of $G_\psi : \mathbb{R}^d \to \mathbb{R}^d$ at $\theta = \psi$ in $v \in \mathbb{R}^d$ (at the cost of doubling the computation graph).



Figure 1: PPO clipping (21) updates $p_\theta(x_t|s_t)$ only if $r_t = \frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}$ is (i) wrong or (ii) still inside the trust region. When $A_\psi(s_t, x_t) > 0$, it is increased iff $r_t \leq 1 + \epsilon$. When $A_\psi(s_t, x_t) < 0$, it is decreased iff $r_t \geq 1 - \epsilon$.

5.2 Relative Clipping Effect

PPO clipping is based on relative probability change. As a result, if the advantage at step t is positive, we drop the token iff

$$p_\theta(x_t|s_t) > (1+\epsilon)p_\psi(x_t|s_t) \qquad \Leftrightarrow \qquad p_\theta(x_t|s_t) - p_\psi(x_t|s_t) > \epsilon p_\psi(x_t|s_t) \tag{23}$$

We see that high-confidence positive tokens are disproportionately spared ($\rightarrow$ exploitation).[6] For instance, if $p_\psi(x_t|s_t) = 0.9$ and $\epsilon = 0.2$, $x_t|s_t$ never gets clipped and thus continues to be reinforced even when $p_\theta(x_t|s_t) \approx 1$. In contrast, if the advantage is negative, we drop the token iff

$$p_\theta(x_t|s_t) < (1-\epsilon)p_\psi(x_t|s_t) \qquad \Leftrightarrow \qquad p_\psi(x_t|s_t) - p_\theta(x_t|s_t) > \epsilon p_\psi(x_t|s_t) \tag{24}$$

Here low-confidence negative tokens are disproportionately spared ($\rightarrow$ exploration). Researchers have proposed modifications to address PPO's disproportionate clipping effects, for example

- DAPO (Yu *et al.*, 2026) uses a higher upper threshold $\epsilon' > \epsilon$ in $\text{clip}_{1-\epsilon}^{1+\epsilon'}(r_t)$ to promote reinforcing more positive tokens (i.e., low-confidence). Note that it still allows for exploiting high-confidence positive tokens.
- DPPO (Qi *et al.*, 2026) uses an adaptive threshold $\epsilon_t = \frac{\epsilon}{p_\psi(x_t|s_t)}$ corresponding to absolute clipping, leading to *both* stronger reinforcement of low-confidence positive tokens *and* stronger suppression of low-confidence negative tokens (Section 8).

5.3 Length Normalization

Figure 2 summarizes various length normalization schemes developed for PPO, motivated along two orthogonal axes:

1. **Inter-batch scaling.** Without normalization, a batch with longer rollouts makes a larger update. The original PPO paper normalizes by the total number of tokens in each batch. Dr. GRPO avoids modifying the objective by assuming a constant rollout length T_{const}.
2. **Intra-batch weighting.** Since the objective weights all tokens in the batch equally, a longer rollout makes a larger update. GRPO weights all rollouts equally within a prompt. DAPO weights all tokens equally within a prompt (thus a longer rollout makes a larger update—within a prompt).

5.4 KL Variant

PPO also considers a more classical soft-constraint variant that maximizes the Lagrangian $L(\theta, \lambda) = J_\psi(\theta) - \lambda(\text{KL}_\psi(\theta) - \epsilon)$ of (17) over θ, without letting the opponent minimize it over $\lambda > 0$. Equivalently, we maximize the TR objective with a KL penalty

$$J_\psi^{\text{ppo-kl}}(\theta) = J_\psi(\theta) - \lambda \text{KL}_\psi(\theta) = \mathop{\mathbf{E}}_{\tau \sim p_\psi}\left[\sum_{t=1}^{T} \frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)} A_\psi(s_t, x_t) - \lambda \text{KL}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t))\right] \tag{25}$$

So this is just the raw TR objective with a per-step KL regularization. We can use any variance-stabilized KL estimator (Appendix B). Since the penalty is always active, the original paper proposes an adaptive scheme where λ is doubled when the KL is $> 1.5\epsilon$, halved when $< \frac{\epsilon}{1.5}$. However, the empirical observation is that the KL variant is less competitive than the clipping variant, whose per-token regularization is sample-dependent.

[6] This is suspected to be related to the empirical phenomenon of "entropy collapse" (θ becomes almost deterministic) which kills exploration in the next round of PPO.

	Unnorm.	PPO	GRPO	DAPO	Dr. GRPO
population objective	$\mathbf{E}_\psi[J]$	$\frac{\mathbf{E}_\psi[J]}{\mathbf{E}_\psi[T]}$	$\mathbf{E}_\psi[\frac{J}{T}]$	$\mathbf{E}_{s_1}[\frac{\mathbf{E}_\psi[J\|s_1]}{\mathbf{E}_\psi[T\|s_1]}]$	$\frac{\mathbf{E}_\psi[J]}{T_{\text{const}}}$
empirical objective	$\frac{1}{N}\sum_i J^{(i)}$	$\frac{1}{\sum_i T^{(i)}}\sum_i J^{(i)}$	$\frac{1}{N}\sum_{p,j}\frac{J^{(p,j)}}{T^{(p,j)}}$	$\frac{1}{P}\sum_p \frac{\sum_j J^{(p,j)}}{\sum_j T^{(p,j)}}$	$\frac{1}{NT_{\text{const}}}\sum_i J^{(i)}$
equal weights for	toks/batch	toks/batch	rollouts/prompt	toks/prompt	toks/batch
longer rollout ↑ weight	◯	◯	✗	◯ (within prompt)	◯
change in grad. dir.	✗	✗	△	△	✗

Figure 2: Summary of length normalization schemes. J denotes the total contribution of a rollout with length T. In a batch of $N = GP$ rollouts, we index $i = 1 \dots N$ (all rollouts), $j = 1 \dots G$ (rollouts within a group), and $p = 1 \dots P$ (all prompts).

6 Group Relative Policy Optimization (GRPO)

GRPO (Shao *et al.*, 2024) uses PPO (21) with length normalization, but uses a *sequence*-level advantage term that can be estimated without a critic (Section 11). Assuming a "group" of $G > 1$ iid rollouts $\tau^{(1)} \dots \tau^{(G)} \sim p_\psi(\cdot|s_1)$ with rewards $R^{(i)}$, the "advantage" of the i-th rollout for the prompt s_1 is measured as (for some $\delta > 0$)

$$\widehat{A}^{\text{grpo}}_{\psi,G}(\tau^{(i)}) = \frac{R^{(i)} - \widehat{\mathbf{E}}_\psi[R|s_1]}{\widehat{\text{stdev}}_\psi(R|s_1) + \delta} \tag{26}$$

where the reward mean and standard deviation can be estimated from the G samples. This is clearly not a valid estimator of $A_\psi(s_t, x_t)$ and does not capture token-level advantages, but it is often sufficiently effective in practice. Thus the empirical GRPO objective per prompt is

$$\widehat{J}^{\text{grpo}}_{\psi,G}(\theta; s_1) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T^{(i)}}\sum_{t=1}^{T^{(i)}}\min\left(\frac{p_\theta(x_t^{(i)}|s_t^{(i)})}{p_\psi(x_t^{(i)}|s_t^{(i)})}\widehat{A}^{\text{grpo}}_{\psi,G}(\tau^{(i)}), \text{clip}^{1+\epsilon}_{1-\epsilon}\left(\frac{p_\theta(x_t^{(i)}|s_t^{(i)})}{p_\psi(x_t^{(i)}|s_t^{(i)})}\right)\widehat{A}^{\text{grpo}}_{\psi,G}(\tau^{(i)})\right) \tag{27}$$

(we omit the reference-policy KL in the paper). The corresponding population version is

$$J^{\text{grpo}}_{\psi,G}(\theta) = \mathop{\mathbf{E}}_{\substack{s_1\sim\mathbf{env}\\ \tau^{(1)}\dots\tau^{(G)}\sim p_\psi(\cdot|s_1)}}\left[\frac{1}{T^{(1)}}\sum_{t=1}^{T^{(1)}}\min\left(\frac{p_\theta(x_t^{(1)}|s_t^{(1)})}{p_\psi(x_t^{(1)}|s_t^{(1)})}\widehat{A}^{\text{grpo}}_{\psi,G}(\tau^{(1)}), \text{clip}^{1+\epsilon}_{1-\epsilon}\left(\frac{p_\theta(x_t^{(1)}|s_t^{(1)})}{p_\psi(x_t^{(1)}|s_t^{(1)})}\right)\widehat{A}^{\text{grpo}}_{\psi,G}(\tau^{(1)})\right)\right]$$

using exchangeability. We can further simplify it in the limit $G \to \infty$ since $\widehat{A}^{\text{grpo}}_{\psi,G}(\tau)$ for any τ in the group converges to a group-independent quantity $A^{\text{grpo}}_{\psi,\infty}(\tau) = (R(\tau) - \mathbf{E}_\psi[R|s_1])/(\text{stdev}_\psi(R|s_1) + \delta)$, yielding

$$J^{\text{grpo}}_{\psi,\infty}(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\frac{1}{T}\sum_{t=1}^{T}\min\left(\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}A^{\text{grpo}}_{\psi,\infty}(\tau), \text{clip}^{1+\epsilon}_{1-\epsilon}\left(\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}\right)A^{\text{grpo}}_{\psi,\infty}(\tau)\right)\right] \tag{28}$$

The popularity of GRPO motivated several refinements, a few of which we cover below.

6.1 Dr. GRPO

Dr. GRPO (Liu *et al.*, 2025) skips the whitening step in (26).[7] The advantage of the i-th rollout is now

$$\widehat{A}^{\text{drgrpo}}_{\psi,G}(\tau^{(i)}) = R^{(i)} - \widehat{\mathbf{E}}_\psi[R|s_1] = \frac{G-1}{G}\Bigg(R^{(i)} - \underbrace{\frac{1}{G-1}\sum_{j\neq i}R^{(j)}}_{\text{(leave-one-out baseline)}}\Bigg) \tag{29}$$

Dr. GRPO also replaces per-rollout normalization with constant $T_{\max}$ (so longer rollouts can make larger updates). With these two changes, GRPO's empirical objective per prompt becomes

$$\widehat{J}^{\text{drgrpo}}_{\psi,G}(\theta; s_1) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_{\max}}\sum_{t=1}^{T^{(i)}}\min\left(\frac{p_\theta(x_t^{(i)}|s_t^{(i)})}{p_\psi(x_t^{(i)}|s_t^{(i)})}\widehat{A}^{\text{drgrpo}}_{\psi,G}(\tau^{(i)}), \text{clip}^{1+\epsilon}_{1-\epsilon}\left(\frac{p_\theta(x_t^{(i)}|s_t^{(i)})}{p_\psi(x_t^{(i)}|s_t^{(i)})}\right)\widehat{A}^{\text{drgrpo}}_{\psi,G}(\tau^{(i)})\right) \tag{30}$$

[7] GRPO's advantage whitening upweights a skewed prompt (i.e., mostly successful or mostly failed rollouts) compared to a balanced prompt. This is a questionable design choice to begin with, presumably adopted to follow standard normalization practices.

We can easily check that $J^{\text{drgrpo}}_{\psi,G}(\theta) = \mathbf{E}_\psi[\widehat{J}^{\text{drgrpo}}_{\psi,G}(\theta)]$ now has the correct gradient at $\theta = \psi$ (up to scaling), since

$$\nabla J^{\text{drgrpo}}_{\psi,G}(\psi) = \frac{1}{GT_{\max}} \sum_{i=1}^{G} \mathbf{E}_\psi[\widehat{A}^{\text{drgrpo}}_{\psi,G}(\tau^{(i)}) \nabla \log p_\psi(\tau^{(i)}|s_1)] = \frac{G-1}{GT_{\max}} \nabla J(\psi)$$

where the last equality follows from $\mathbf{E}_\psi[\widehat{A}^{\text{drgrpo}}_{\psi,G}(\tau^{(i)})|s_1, \tau^{(i)}] = \frac{G-1}{G}(R^{(i)} - \mathbf{E}_\psi[R|s_1])$ and REINFORCE (3).

6.2 Dynamic Sampling (DAPO)

GRPO's group-relative advantage (26) is zero if all rollouts have the same reward $R^{(1)} = \cdots = R^{(G)}$, yielding zero gradient in (27). DAPO (Yu *et al.*, 2026) proposes to oversample rollouts and ensure the group satisfies

$$0 < \sum_{i=1}^{G} R^{(i)} < G$$

(assuming binary reward). This way, every rollout in the group has nonzero advantage. While oversampling until the batch is filled with such groups adds sampling cost, it does not necessarily increase wall-clock time proportionally (since batched generation is often bottlenecked by a small number of exceptionally long rollouts). This empirically results in improved training efficiency (i.e., higher reward faster).

6.3 Group Sequence Policy Optimization (GSPO)

Since GRPO replaces the token-level advantage to sequence-level, a natural direction is to extend existing token-level TR methods to sequence-level. GSPO (Zheng *et al.*, 2025) takes this direction, observing that with the GRPO approximation $A_\psi(s_t, x_t) \approx A_\psi(\tau)$ for every t (infinite-group regime for simplicity) the policy improvement objective (8) can be written as

$$J(\theta) - J(\psi) = \mathop{\mathbf{E}}_{\tau\sim p_\theta}\left[\sum_{t=1}^{T} A_\psi(s_t, x_t)\right] \approx \mathop{\mathbf{E}}_{\tau\sim p_\theta}[T A_\psi(\tau)] = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[T \frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)} A_\psi(\tau)\right]$$

Following GRPO's sequence-level normalization, this motivates a PPO-style objective:

$$J^{\text{gspo}}_\psi(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\min\left(\left(\frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)}\right)^{1/T} A_\psi(\tau), \text{clip}^{1+\epsilon}_{1-\epsilon}\left(\left(\frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)}\right)^{1/T}\right) A_\psi(\tau)\right)\right] \tag{31}$$

As in the token-level case, it updates θ only for rollouts τ whose scaled importance ratio $r^{1/T} = (\frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)})^{1/T}$ is in the wrong direction or inside the trust region (Figure 1). The exponential normalization by $1/T$ (i.e., taking a geometric mean of token-level ratios $r_t = \frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}$) is not justified for importance sampling, but it makes the clipping threshold less sensitive to rollout length and matches GRPO's length normalization in the gradient. Ignoring clipping,

$$J^{\text{grpo}}_{\psi,\infty}(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\frac{1}{T}\sum_{t=1}^{T} r_t A_\psi(\tau)\right] \quad \Rightarrow \quad \nabla J^{\text{grpo}}_{\psi,\infty}(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\frac{1}{T}\sum_{t=1}^{T} r_t A_\psi(\tau) \nabla \log p_\theta(x_t|s_t)\right]$$

$$J^{\text{gspo}}_\psi(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[r^{1/T} A_\psi(\tau)\right] \quad \Rightarrow \quad \nabla J^{\text{gspo}}_\psi(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\frac{1}{T}\sum_{t=1}^{T} r^{1/T} A_\psi(\tau) \nabla \log p_\theta(x_t|s_t)\right]$$

(using $\nabla_\theta(r^{1/T}) = r^{1/T}\nabla_\theta(\frac{1}{T}\sum_t \log r_t)$). Thus GSPO replaces the importance ratio in the GRPO gradient from r_t to $r^{1/T}$ at every step, in particular preserving the initial gradient $\nabla J^{\text{grpo}}_{\psi,\infty}(\psi) = \nabla J^{\text{gspo}}_\psi(\psi)$.

7 Mirror Descent Policy Optimization (MDPO)

MDPO (Tomar *et al.*, 2020) revisits the KL variant of PPO:

$$J^{\text{mdpo}}_\psi(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\sum_{t=1}^{T} \frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)} A_\psi(s_t, x_t) - \lambda_{\text{iter}} \text{KL}(p_\theta(\cdot|s_t), p_\psi(\cdot|s_t))\right] \tag{32}$$

The only practical differences from (25) are: the per-step KL direction is reversed, and the regularization strength is scheduled over policy iterations.[8] But (32) is motivated more specifically from a classical tabular setting. The key observation is that the TR surrogate (10) is already linear in the model's token distribution $q_s(x) := p_\theta(x|s)$ for a fixed s. Shown again without importance sampling,

$$J_\psi(\theta) = \sum_{t=1}^{T_{\max}} \mathop{\mathbf{E}}_{s\sim d_{\psi,t}} \left[\sum_x q_s(x) A_\psi(s,x) \right] \tag{33}$$

So the *per-state* objective is $q_s^\top A_\psi(s,\cdot)$ where the gradient is literally the advantages $A_\psi(s,\cdot) \in \mathbb{R}^V$. In a tabular setting, at each s we use the optimal token distribution

$$q_s^\star = \mathop{\arg\max}_{q_s \in \Delta^{V-1}} \sum_x q_s(x) A_\psi(s,x) - \lambda \mathrm{KL}(q_s, p_\psi(\cdot|s)) \tag{34}$$

(34) is equivalent to exponentiated gradient descent (i.e., mirror descent with probability simplex + KL) on a trivial (linear) loss. The fact that $\sum_x q_s^\star(x) A_\psi(s,x) \geq 0$ implies $J(q^\star) - J(\psi) = \sum_t \mathbf{E}_{q^\star}[A_\psi(s,x)] \geq 0$. MDPO (32) is motivated as taking gradient steps toward $q_s^\star$ instead of computing the well-known "tilting" solution $q_s^\star(x) \propto e^{A_\psi(s,x)/\lambda} p_\psi(x|s)$. With a bit of analysis, we can show the exact relationship:

$$J_\psi^{\mathrm{mdpo}}(\theta) = \sum_{t=1}^{T_{\max}} \mathop{\mathbf{E}}_{s\sim d_{\psi,t}} [C_s^\star - \lambda \mathrm{KL}(p_\theta(\cdot|s), q_s^\star)] \tag{35}$$

where $C_s^\star$ is the maximum of (34).[9]

7.1 Kimi k1.5 Variant

Kimi k1.5 (Team *et al.*, 2025a) applies the mirror descent approach to a sequence level and minimizes a regression loss toward the tilting solution. It observes that the original objective $J(\theta) = \mathbf{E}_{\tau\sim p_\theta}[R(\tau)]$ (2) is also linear in the model's *rollout* distribution $r_{s_1}(\tau) = p_\theta(\tau|s_1)$. So by the same line of tabular logic, at each s_1 we use the optimal rollout distribution

$$r_{s_1}^\star = \mathop{\arg\max}_{r_{s_1} \in (\text{all policies})} \sum_\tau r_{s_1}(\tau) R(\tau) - \lambda \mathrm{KL}(r_{s_1}, p_\psi(\cdot|s_1)) \qquad \Rightarrow \qquad r_{s_1}^\star(\tau) \propto e^{R(\tau)/\lambda} p_\psi(\tau|s_1) \tag{36}$$

This gives the characterization of an ideal policy. Taking the log and rearranging, it can be stated as $\lambda \log \frac{r_{s_1}^\star(\tau)}{p_\psi(\tau|s_1)} = R(\tau) - \lambda \log Z_\psi(s_1)$ where $Z_\psi(s_1) = \mathbf{E}_{\tau\sim p_\psi(\cdot|s_1)}[e^{R(\tau)/\lambda}]$. The paper proposes to promote this property by minimizing the regression loss

$$L_\psi^{\mathrm{k1.5}}(\theta) = \mathop{\mathbf{E}}_{\substack{s_1\sim\mathbf{env} \\ \tau\sim p_\psi(\cdot|s_1)}} \left[\left(R(\tau) - \mathbf{E}_\psi[R|s_1] - \lambda \log \frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)} \right)^2 \right] \tag{37}$$

which samples from ψ for convenience (any full-support rollout distribution is valid) and uses the approximation $\lambda \log \mathbf{E}[e^{R/\lambda}] = \mathbf{E}[R] + \frac{\mathrm{Var}(R)}{2\lambda} + O(\frac{1}{\lambda^2}) \approx \mathbf{E}[R]$ (justified for large λ).[10] To understand what it learns, we can write the gradient as

$$-\frac{1}{2\lambda} \nabla L_\psi^{\mathrm{k1.5}}(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi} \left[\Big((R(\tau) - \mathbf{E}_\psi[R|s_1]) \nabla \log p_\theta(\tau|s_1) \Big) - \frac{\lambda}{2} \nabla \left(\log \frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)} \right)^2 \right]$$

The first part reduces to the REINFORCE gradient (3) (with constant baseline) when $\psi = \mathrm{sg}(\theta)$. The second part is KL-like regularization, discouraging divergence.

[8] The paper uses $\lambda_{\mathrm{iter}} = \frac{1}{t_{\mathrm{iter}}}$ where $t_{\mathrm{iter}} = 1 - \frac{\mathrm{iter}}{\mathrm{iter}_{\max}}$, so the strength increases over policy iterations each of which corresponds to a mirror-descent step (34) and involves sampling batches from a particular rollout policy.

[9] Expanding $\mathrm{KL}(q_s, q_s^\star)$ with $q_s^\star(x) = e^{A_\psi(s,x)/\lambda} p_\psi(x|s)/Z_\psi(s)$ and rearranging, we have $\sum_x q_s(x) A_\psi(s,x) - \lambda \mathrm{KL}(q_s, p_\psi(\cdot|s)) = \lambda \log Z_\psi(s) - \lambda \mathrm{KL}(q_s, q_s^\star)$. The RHS is maximized exactly at $q_s = q_s^\star$ to $C_s^\star = \lambda \log Z_\psi(s)$.

[10] The expression can be verified using $e^x = 1 + x + \frac{x^2}{2} + O(x^3)$ and $\log(1+x) = x - \frac{x^2}{2} + O(x^3)$. But note that the general estimation problem of form $\log \mathbf{E}[e^{f(X)}]$ can be extremely sample-inefficient (McAllester and Stratos, 2020). Thus this approximation is more of a bias-variance tradeoff.

8 Divergence Proximal Policy Optimization (DPPO)

DPPO (Qi *et al.*, 2026) introduces an adaptive clipping threshold $\epsilon_t = \frac{\epsilon}{p_\psi(x_t|s_t)}$ in the PPO objective (21):

$$J_\psi^{\text{dppo}}(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\sum_{t=1}^{T} \min\left(\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}A_\psi(s_t,x_t), \text{clip}_{1-\epsilon_t}^{1+\epsilon_t}\left(\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}\right)A_\psi(s_t,x_t)\right)\right] \tag{38}$$

Per-token clipping is still asymmetric (i.e., drop only in the correct direction), but now based on absolute probability change (contrast with (23–24)):

$$\Big(A_\psi(s_t,x_t) > 0 \;\bigwedge\; p_\theta(x_t|s_t) - p_\psi(x_t|s_t) > \epsilon\Big) \;\bigvee\; \Big(A_\psi(s_t,x_t) < 0 \;\bigwedge\; p_\psi(x_t|s_t) - p_\theta(x_t|s_t) > \epsilon\Big) \tag{39}$$

As a result, more low-confidence positive tokens are reinforced (exploration ↑ in the positive case) *and* more low-confidence negative tokens are suppressed (exploitation ↑ in the negative case).

8.1 TRPO Motivation

DPPO motivates absolute clipping $|P(x_t) - Q(x_t)| > \epsilon$ as a sample-dependent approximation of TRPO's original constraint $\text{TV}(P,Q) = \frac{1}{2}\sum_x |P(x) - Q(x)| \le \epsilon$. When P, Q are concentrated on x_t, the fake distributions P', Q' over $\{x_t, \varnothing\}$ with $P'(x_t) = P(x_t)$ and $Q'(x_t) = Q(x_t)$ are reasonable proxies. Then

$$\text{TV}(P',Q') = |P(x_t) - Q(x_t)| \le \text{TV}(P,Q)$$

While DPPO treats the LHS as an actual approximation of the RHS, note that it is technically a lower bound (by traingle inequality) so enforcing $\text{TV}(P',Q') \le \epsilon$ does not guarantee $\text{TV}(P,Q) \le \epsilon$.[11] Using this approximation to replace the state-wise TV constraint in TRPO (15) gives the step-wise surrogate constraint

$$\begin{aligned}\max_{\theta:\ \text{TV}(p_\psi(\cdot|s),p_\theta(\cdot|s))\le\epsilon\ \forall s} \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\sum_{t=1}^{T}\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}A_\psi(s_t,x_t)\right] &\approx \max_{\theta:\ \text{TV}(p_\psi(\cdot|s_t),p_\theta(\cdot|s_t))\le\epsilon\ \forall t} \sum_{t=1}^{T}\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}A_\psi(s_t,x_t) \\ &\approx \max_{\theta:\ |p_\theta(x_t|s_t) - p_\psi(x_t|s_t)|\le\epsilon\ \forall t} \sum_{t=1}^{T}\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}A_\psi(s_t,x_t)\end{aligned}$$

Promoting this constraint in the asymmetric style of PPO motivates the masked objective of form

$$\mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\sum_{t=1}^{T} M_t \times r_t A_t\right] \qquad M_t = \begin{cases} 0 & \text{if } (A_t > 0 \wedge p_{\theta,t} - p_{\psi,t} > \epsilon) \vee (A_t < 0 \wedge p_{\psi,t} - p_{\theta,t} > \epsilon) \\ 1 & \text{otherwise}\end{cases}$$

which is gradient-equivalent to the ratio-clipped objective (38).

8.2 Average Bound

The DPPO paper also gives an average bound on the TR surrogate objective's bias (Lemma C.6, Corollary C.7):

$$|\texttt{Bias}_\psi(\theta)| \le C \times \mathop{\mathbf{E}}_{\tau\sim p_\psi}\left[\sum_{t=1}^{T}\text{TV}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t))\right] \le C' \times \sqrt{\text{KL}_\psi(\theta)} \tag{40}$$

for some constants $C, C' > 0$. It retrospectively justifies the KL constraint in TRPO (17). It is not needed to motivate the per-step TV constraint in DPPO, and it is a weaker bound than TRPO's max bound (14) locally in the size of policy change (first-order, rather than second-order).[12]

[11] It is easy to generalize this to include top-k tokens (i.e., support size in $\{k+2, k+1\}$) with a tighter bound, but the paper does not find it improve over just using x_t.

[12] But the max bound is $O(T_{\max}^2\alpha^2)$ (Lemma C.5) whereas this bound is $O(T_{\max}\alpha)$ (Lemma C.6) where $\alpha = \max_s \text{TV}(p_\theta(\cdot|s), p_\psi(\cdot|s))$, thus it may be numerically tighter if we account for the rollout length.

9 Training Inference Mismatch (TIM)

Infra issues can cause a so-called "TIM":

$$p_\psi^{\text{sampler}}(x|s) \neq p_\psi^{\text{trainer}}(x|s) \tag{41}$$

where the superscript marks if the rollout policy's token probability is computed by the sampler or the trainer. This mismatch can create an extra layer of error. Notably, the basic score function identity breaks:

$$\mathop{\mathbf{E}}_{x\sim p_\psi^{\text{sampler}}(\cdot|s)}\left[\nabla \log p_\psi^{\text{trainer}}(x|s)\right] \neq 0 \tag{42}$$

invalidating the justification for the baseline function in REINFORCE.

9.1 Importance Sampling

We can fix (42) by importance sampling:

$$\mathop{\mathbf{E}}_{x\sim p_\psi^{\text{sampler}}(\cdot|s)}\left[\frac{p_\psi^{\text{trainer}}(x|s)}{p_\psi^{\text{sampler}}(x|s)}\nabla \log p_\psi^{\text{trainer}}(x|s)\right] = \mathop{\mathbf{E}}_{x\sim p_\psi^{\text{trainer}}(\cdot|s)}\left[\nabla \log p_\psi^{\text{trainer}}(x|s)\right] = 0 \tag{43}$$

We can compute p_ψ^{trainer} efficiently by batched forward pass using a copy of ψ on the trainer's side (aka. behavior policy). While not always justified, this motivates an empirical strategy of importance weighting every per-token objective as

$$\mathbf{E}_\psi\left[\sum_{t=1}^{T} J_t\right] \longrightarrow \mathbf{E}_\psi\left[\sum_{t=1}^{T} \mathbf{G}\left(\frac{p_\psi^{\text{trainer}}(x_t|s_t)}{p_\psi^{\text{sampler}}(x_t|s_t)}\right) J_t\right]$$

where $\mathbf{G}$ is a guard function to prevent extreme values from breaking optimization. Examples include

$$\mathbf{G}(r) = \begin{cases} \min(r, C) & \text{(truncated importance sampling (TIS), Yao } et\ al.\text{ (2025))} \\ [[C \leq r \leq C']]r & \text{(masked importance sampling (MIS), Team } et\ al.\text{ (2025b))} \end{cases}$$

Note that TIS caps importance weights, but MIS completely drops the token contribution outside a specified range.

9.2 Expert Replay

Expert mismatch can make TIM particularly severe (Section 9.3). Expert replay forces the sampler's expert assignments on the trainer. For instance, R3 (Rollout Routing Replay, Ma *et al.* (2025)) changes both the student and behavior policies in the trainer as

$$p_\theta^{\text{trainer}}(x|s) \longrightarrow p_\theta^{\text{trainer}}(x|s, \mathcal{E}^{\text{sampler}}) \tag{44}$$

$$p_\psi^{\text{trainer}}(x|s) \longrightarrow p_\psi^{\text{trainer}}(x|s, \mathcal{E}^{\text{sampler}}) \tag{45}$$

where $\mathcal{E}^{\text{sampler}}$ is the expert assignment under p_ψ^{sampler}. Note that only (45) fixes expert mismatch. (44) fixes neither TIM nor policy staleness (i.e., $\mathcal{E}^{\text{sampler}}$ is no longer a correct assignment when $\theta \neq \psi$). However, since it completely prevents expert reallocation when θ is updated (which can cause huge changes in probabilities), it can be argued as indirectly reducing policy staleness.

9.3 Causes of TIM

We list some known causes of TIM. Recall that addition and multiplication lose associativity in finite precision due to rounding errors, so any change in operation order can change the end result.

- **Kernel mismatch**: The sampler and trainer use different kernels (generation vs batched forward pass). Different kernels may differ in reduction order, choices of numerical approximation, or intermediate precision, even if they are implementing the same function.

- **Execution nondeterminism**: Some parallel operations (e.g., reduction with atomic updates) may execute in nondeterministic order and produce different results even with identical kernels, inputs, and hardware.
- **Batch dependency**: Some operations may depend on the given batch. This dependency can be mathematical (e.g., batchnorm, capacity-constrained expert routing) or numerical (e.g., different batch sizes may yield different matmul kernels).

Small numerical differences can also spike into more drastic changes. For instance, **expert mismatch** can happen from any of these issues, making the result very different for a mixture of experts.

10 Score Centering (SC)

SC (Marek and Ryabinin, 2026) presents an alternative to importance sampling for addressing policy staleness. The derivation assumes a practically motivated scenario where

1. We just sample from a wrong policy $\psi \neq \theta$ in the true policy gradient (3), and
2. We use a general advantage estimator $\widehat{A}_\psi(s_t, x_t) = R(\tau) - b(s_t)$ for $\tau \sim p_\psi$.

then tries to fix the resulting mismatch. That is,

$$\nabla J(\theta) = \sum_{t=1}^{T_{\max}} \mathop{\mathbf{E}}_{s\sim d_{\theta,t}} \left[\mathop{\mathbf{E}}_{x\sim p_\theta(\cdot|s)} \left[\widehat{A}_\theta(s,x) \nabla \log p_\theta(x|s) \right] \right] \approx \sum_{t=1}^{T_{\max}} \mathop{\mathbf{E}}_{s\sim d_{\psi,t}} \left[\mathop{\mathbf{E}}_{x\sim p_\psi(\cdot|s)} \left[\widehat{A}_\psi(s,x) \nabla \log p_\theta(x|s) \right] \right] \tag{46}$$

The two sources of mismatch are: (i) wrong state distribution, (ii) wrong token distribution. The former is exactly the approximation in the TR surrogate; (46) reduces to $\nabla J_\psi(\theta)$ if we multiply each term by $\frac{p_\theta(x|s)}{p_\psi(x|s)}$. Instead, the paper proposes a different correction by comparing the true and false gradient side by side:

$$\mathop{\mathbf{E}}_{x\sim p_\theta(\cdot|s)} \left[\widehat{A}_\theta(s,x) \nabla \log p_\theta(x|s) \right] = \mathop{\mathrm{Cov}}_{x\sim p_\theta(\cdot|s)} \left(\widehat{A}_\theta(s,x), \nabla \log p_\theta(x|s) \right) \tag{47}$$

$$\mathop{\mathbf{E}}_{x\sim p_\psi(\cdot|s)} \left[\widehat{A}_\psi(s,x) \nabla \log p_\theta(x|s) \right] = \mathop{\mathrm{Cov}}_{x\sim p_\psi(\cdot|s)} \left(\widehat{A}_\psi(s,x), \nabla \log p_\theta(x|s) \right) + \underbrace{\mathop{\mathbf{E}}_{x\sim p_\psi(\cdot|s)} \left[\widehat{A}_\psi(s,x) \right] \mathop{\mathbf{E}}_{x\sim p_\psi(\cdot|s)} \left[\nabla \log p_\theta(x|s) \right]}_{\text{“drift”}}$$

In the true gradient, the drift term vanishes since $\mathbf{E}_\psi[\nabla \log p_\theta(x|s)] = -\nabla \mathrm{KL}(p_\psi(\cdot|s), p_\theta(\cdot|s))$ becomes zero. But it *already* vanishes if $\widehat{A}_\psi(s,x) = A_\psi(s,x)$, thus the motivation only applies to non-exact advantage estimate (assumption 2). We can remove the drift term even when $\psi \neq \theta$ by centering the score function:[13]

$$\mathop{\mathbf{E}}_{x\sim p_\psi(\cdot|s)} \left[\widehat{A}_\psi(s,x) \left(\nabla \log p_\theta(x|s) - \mathop{\mathbf{E}}_{x\sim p_\psi(\cdot|s)} \left[\nabla \log p_\theta(x|s) \right] \right) \right] = \mathop{\mathrm{Cov}}_{x\sim p_\psi(\cdot|s)} \left(\widehat{A}_\psi(s,x), \nabla \log p_\theta(x|s) \right) \tag{48}$$

Note that this is still not equal to (47); SC only removes the drift and does not correct the remaining distribution mismatch. The gradient (48) can be computed with the surrogate objective:

$$J^{\mathrm{SC}}_\psi(\theta) = \mathop{\mathbf{E}}_{\tau\sim p_\psi} \left[\sum_{t=1}^{T} \widehat{A}_\psi(s_t, x_t) \left(\log p_\theta(x_t|s_t) - \mathop{\mathbf{E}}_{x\sim p_\psi(\cdot|s)} \left[\log p_\theta(x|s) \right] \right) \right] \tag{49}$$

10.1 Top-k Approximation

Given a rollout $x_t \sim p_\psi(\cdot|s_t)$ from the sampler, SC requires estimating the center (negative cross entropy)

$$C_{s_t}(\theta) = \sum_{x\in\mathcal{V}} p_\psi(x|s_t) \log p_\theta(x|s_t)$$

[13] To motivate this, the paper considers the case $\widehat{A}_\psi(s,x) = C \neq 0$ is constant. Then the true gradient is zero but the new estimator is nonzero (since $C\mathbf{E}_\psi[\nabla \log p_\theta(x|s)] \neq 0$). SC can be seen as exercising an inductive bias “there should be nothing to learn here”.

for each $t = 1 \dots T$. Naively this requires the sampler to send full-vocab logits for every token in the entire rollout batch. The paper proposes a top-k (e.g., 128) approximation that uses p_θ (with stop-gradient) to sketch the tail distribution since the latter is already available in a given minibatch. Letting $\mathcal{V} = \mathcal{H} \cup \mathcal{T}$ denote the partition,

$$\widehat{C}_{s_t}(\theta) = \sum_{x \in \mathcal{V}} q(x|s_t) \log p_\theta(x|s_t) \qquad q(x|s_t) = \begin{cases} p_\psi(x|s_t) & \text{if } x \in \mathcal{H} \\ \rho \text{sg}(p_\theta(x|s_t)) & \text{if } x \in \mathcal{T} \end{cases} \tag{50}$$

Solving for ρ to satisfy $\sum_{x \in \mathcal{V}} q(x|s_t) = 1$, we have $\rho = \frac{1 - \sum_{x \in \mathcal{H}} p_\psi(x|s_t)}{1 - \sum_{x \in \mathcal{H}} \text{sg}(p_\theta(x|s_t))}$ which can be easily computed per minibatch assuming the availability of the sampler's top-k probabilities. The paper goes a step further and note that the *gradient* of $\widehat{C}_{s_t}(\theta)$ only involves a top-k sum:

$$\begin{aligned} \nabla \widehat{C}_{s_t}(\theta) &= \sum_{x \in \mathcal{H}} p_\psi(x|s_t) \nabla \log p_\theta(x|s_t) + \rho \sum_{x \in \mathcal{T}} p_\theta(x|s_t) \nabla \log p_\theta(x|s_t) \\ &= \sum_{x \in \mathcal{H}} p_\psi(x|s_t) \nabla \log p_\theta(x|s_t) - \rho \sum_{x \in \mathcal{H}} p_\theta(x|s_t) \nabla \log p_\theta(x|s_t) \qquad (51) \\ &= \sum_{x \in \mathcal{H}} (p_\psi(x|s_t) - \rho p_\theta(x|s_t)) \nabla \log p_\theta(x|s_t) \end{aligned}$$

where (51) uses the score function identity $\mathbf{E}_{x \sim p_\theta}[\nabla \log p_\theta(x)] = 0$. This yields the gradient-equivalent surrogate center

$$\widetilde{C}_{s_t}(\theta) = \sum_{x \in \mathcal{H}} \text{sg}\left(p_\psi(x|s_t) - \rho p_\theta(x|s_t)\right) \log p_\theta(x)$$

which only involves a sum over $\mathcal{H}$ and can be plugged into (49) to approximate $\mathbf{E}_{x \sim p_\psi(\cdot|s)}[\log p_\theta(x|s)]$.

10.2 Importance Sampling Hybrid

We first point out that applying IS in (46) recovers $\nabla J_\psi(\theta)$ (for $\widehat{A}_\psi \approx A_\psi$) and removes the need for score centering:

$$\mathop{\mathbf{E}}_{x \sim p_\psi(\cdot|s)} \left[\frac{p_\theta(x|s)}{p_\psi(x|s)} \widehat{A}_\psi(s,x) \nabla \log p_\theta(x|s) \right] = \mathop{\mathbf{E}}_{x \sim p_\theta(\cdot|s)} \left[\widehat{A}_\psi(s,x) \nabla \log p_\theta(x|s) \right] = \mathop{\text{Cov}}_{x \sim p_\theta(\cdot|s)} \left(\widehat{A}_\psi(s,x), \nabla \log p_\theta(x|s) \right)$$

But it also creates the usual variance problem. The paper proposes truncating/masking the ratio $r \mapsto \mathbf{G}(r)$ to prevent extreme values (at the cost of bias), which *then* re-introduces the need for score centering because

$$\begin{aligned} &\mathop{\mathbf{E}}_{x \sim p_\psi(\cdot|s)} \left[\mathbf{G}\left(\frac{p_\theta(x|s)}{p_\psi(x|s)} \right) \widehat{A}_\psi(s,x) \nabla \log p_\theta(x|s) \right] \qquad (52) \\ &= \mathop{\text{Cov}}_{x \sim p_\psi(\cdot|s)} \left(\widehat{A}_\psi(s,x), \mathbf{G}\left(\frac{p_\theta(x|s)}{p_\psi(x|s)} \right) \nabla \log p_\theta(x|s) \right) + \mathop{\mathbf{E}}_{x \sim p_\psi(\cdot|s)} \left[\widehat{A}_\psi(s,x) \right] \mathop{\mathbf{E}}_{x \sim p_\psi(\cdot|s)} \left[\mathbf{G}\left(\frac{p_\theta(x|s)}{p_\psi(x|s)} \right) \nabla \log p_\theta(x|s) \right] \end{aligned}$$

Removing the drift from (52) yields the SC-IS gradient, which can be computed with the surrogate objective:

$$J_\psi^{\text{SC-IS}}(\theta) = \mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[\sum_{t=1}^{T} \widehat{A}_\psi(s_t, x_t) \left(\text{sg}\left(\mathbf{G}\left(\frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)} \right) \right) \log p_\theta(x_t|s_t) - \mathop{\mathbf{E}}_{x \sim p_\psi(\cdot|s_t)} \left[\text{sg}\left(\mathbf{G}\left(\frac{p_\theta(x|s_t)}{p_\psi(x|s_t)} \right) \right) \log p_\theta(x|s_t) \right] \right) \right] \tag{53}$$

The paper experiments with $\mathbf{G}(r) = \min(r, C)$ and $\mathbf{G}(r) = [[C \le r \le C']]r$ which are originally designed for TIM (TIS and MIS, Section 9). Experiments show that SC-IS improves over SC when the rollout policy is substantially stale (i.e., we make many updates within a rollout batch).

Top-k approximation. Again, an efficient center estimator can be obtained from the gradient. Using the top-k approximation in (50) (i.e., $q(x|s_t) = p_\psi(x|s_t)$ for $x \in \mathcal{H}$ and $q(x|s_t) = \rho \text{sg}(p_\theta(x|s_t))$ for $x \in \mathcal{T}$), the gradient of the

center term in (53) is

$$\begin{aligned}
&\sum_{x\in\mathcal{V}} q(x|s_t)\mathbf{G}\left(\frac{p_\theta(x|s_t)}{q(x|s_t)}\right)\nabla\log p_\theta(x|s_t)\\
&=\sum_{x\in\mathcal{H}} p_\psi(x|s_t)\mathbf{G}\left(\frac{p_\theta(x|s_t)}{p_\psi(x|s_t)}\right)\nabla\log p_\theta(x|s_t)+\sum_{x\in\mathcal{T}}\rho p_\theta(x|s_t)\mathbf{G}\left(\frac{p_\theta(x|s_t)}{\rho p_\theta(x|s_t)}\right)\nabla\log p_\theta(x|s_t)\\
&=\sum_{x\in\mathcal{H}} p_\psi(x|s_t)\mathbf{G}\left(\frac{p_\theta(x|s_t)}{p_\psi(x|s_t)}\right)\nabla\log p_\theta(x|s_t)-\rho\mathbf{G}\left(\frac{1}{\rho}\right)\sum_{x\in\mathcal{H}} p_\theta(x|s_t)\nabla\log p_\theta(x|s_t)\\
&=\sum_{x\in\mathcal{H}}\left(p_\psi(x|s_t)\mathbf{G}\left(\frac{p_\theta(x|s_t)}{p_\psi(x|s_t)}\right)-\rho\mathbf{G}\left(\frac{1}{\rho}\right)p_\theta(x|s_t)\right)\nabla\log p_\theta(x|s_t)
\end{aligned}$$

where we additionally use the fact that the guarded ratio becomes constant for the tail distribution. Thus the corresponding gradient-equivalent surrogate center is

$$\widetilde{C}^{\mathrm{IS}}_{s_t}(\theta)=\sum_{x\in\mathcal{H}}\mathrm{sg}\left(p_\psi(x|s_t)\mathbf{G}\left(\frac{p_\theta(x|s_t)}{p_\psi(x|s_t)}\right)-\rho\mathbf{G}\left(\frac{1}{\rho}\right)p_\theta(x|s_t)\right)\log p_\theta(x|s_t)$$

11 Advantage Estimation

We often need to estimate the advantage term

$$A_\theta(s_t,x_t)=\mathbf{E}_\theta[R|s_t,x_t]-\mathbf{E}_\theta[R|s_t] \tag{54}$$

for per-token weights in training objectives (e.g., (6), (21)). This typically must be estimated from a single rollout (possibly truncated) with endpoint reward $R\in\mathbb{R}$ since it would be difficult to get another sample conditioned on these particular s_t and x_t.

11.1 Critic

To amortize over samples, we can train a **critic** ϕ that parameterizes the value function of θ. Given a sample rollout $\tau_{1:K}\sim p_\theta$ ending in step $K\le T$ with endpoint reward $R(s_{K+1})$ (zero if truncated), the simplest training objective is to minimize the squared error for all states

$$\min_\phi\sum_{t=1}^{K}\left(V_\phi(s_t)-\widehat{\mathbf{E}}_\theta[R|s_t]\right)^2 \tag{55}$$

where $\widehat{\mathbf{E}}_\theta[R|s_t]\approx\mathbf{E}_\theta[R|s_t]$ is any estimator of the expected reward from s_t under p_θ computable from this rollout. A straightforward choice is

$$\widehat{\mathbf{E}}_\theta[R|s_t]=R(s_{K+1})+\nu(s_{K+1})V_{\mathrm{sg}(\phi)}(s_{K+1}) \tag{56}$$

where $\nu(s_{K+1})=[[s_{K+1}\text{ is nonterminal}]]$ (i.e., we bootstrap the value at s_{K+1} if needed). Note that every $t=1\dots K$ gets the same estimate from the shared rollout endpoint. The critic ϕ is trained by (55) simultaneously with the main model θ (optionally with a warmup), during which it can provide a running estimate of (54) as

$$\widehat{A}_{\theta,t}=R(s_{K+1})+\nu(s_{K+1})V_\phi(s_{K+1})-V_\phi(s_t) \tag{57}$$

We view (56) and (57) as the default critic target and advantage estimator. This is equal to using GAE with $\lambda=1$.

11.2 Generalized advantage estimator (GAE)

Upon inspection, we can see that (57) is the lowest-bias and highest-variance instance of a series of estimators

$$\widehat{A}^{(k)}_{\theta,t}=R(s_{t+k})+\nu(s_{t+k})V_\phi(s_{t+k})-V_\phi(s_t)$$

More explicitly, at each t, we have $K-t+1$ estimators with decreasing bias (less dependent on ϕ) and increasing variance (further out):

$$\begin{aligned} \widehat{A}^{(1)}_{\theta,t} &= V_\phi(s_{t+1}) - V_\phi(s_t) & \cdots & & \widehat{A}^{(K-t)}_{\theta,t} &= V_\phi(s_K) - V_\phi(s_t) \\ \widehat{A}^{(2)}_{\theta,t} &= V_\phi(s_{t+2}) - V_\phi(s_t) & \cdots & & \widehat{A}^{(K-t+1)}_{\theta,t} &= R(s_{K+1}) + \nu(s_{K+1})V_\phi(s_{K+1}) - V_\phi(s_t) = (57) \end{aligned}$$

We could trade off bias and variance by picking different k. Instead, GAE (Schulman *et al.*, 2015a) proposes to introduce a knob $\lambda \in [0,1]$ and take the exponential average of all estimators:[14]

$$\widehat{A}^{\mathrm{GAE}}_{\theta,t}(\lambda) = (1-\lambda)\widehat{A}^{(1)}_{\theta,t} + (1-\lambda)\lambda\widehat{A}^{(2)}_{\theta,t} + \cdots + (1-\lambda)\lambda^{K-t-1}\widehat{A}^{(K-t)}_{\theta,t} + \lambda^{K-t}\widehat{A}^{(K-t+1)}_{\theta,t} \tag{58}$$

(check the extremes at $\lambda = 0$ and $\lambda = 1$). A naive implementation would take $O(K^2)$ time to obtain all estimates for the trajectory (i.e., for each t, compute $\widehat{A}^{(k)}_{\theta,t}$ for all k). GAE uses $O(K)$-time dynamic programming that exploits shared parts between consecutive estimators. Define the value deltas

$$\begin{aligned} d_t &= V_\phi(s_{t+1}) - V_\phi(s_t) & \cdots & & d_{K-1} &= V_\phi(s_K) - V_\phi(s_{K-1}) \\ d_{t+1} &= V_\phi(s_{t+2}) - V_\phi(s_{t+1}) & \cdots & & d_K &= R(s_{K+1}) + \nu(s_{K+1})V_\phi(s_{K+1}) - V_\phi(s_K) \end{aligned}$$

All deltas $d_1 \dots d_K$ can be precomputed in $O(K)$ time. We can verify that

$$\begin{aligned} \widehat{A}^{\mathrm{GAE}}_{\theta,t}(\lambda) &= d_t + \lambda d_{t+1} + \cdots + \lambda^{K-t-1}d_{K-1} + \lambda^{K-t}d_K \\ &= d_t + \lambda\widehat{A}^{\mathrm{GAE}}_{\theta,t+1}(\lambda) \end{aligned}$$

The last expression gives a natural $O(K)$ backward pass to compute all estimates starting from $\widehat{A}^{\mathrm{GAE}}_{\theta,K}(\lambda) = d_K$. GAE also suggests an alternative value estimator for the critic's training objective (55)

$$\widehat{\mathbf{E}}_\theta[R|s_t] = \widehat{A}^{\mathrm{GAE}}_{\theta,t}(\lambda) + V_{\mathrm{sg}(\phi)}(s_t) \tag{59}$$

where $\widehat{A}^{\mathrm{GAE}}_{\theta,t}(\lambda)$ is computed by (58) using the frozen $\mathrm{sg}(\phi)$. If $\lambda = 1$, it degenerates to (56) and therefore unbiased (assuming perfect bootstrapping). For $\lambda < 1$, it is generally biased but yields step-dependent estimates unlike (56), thus potentially lowering variance.

References

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274.

Kakade, S. M. (2001). A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, pages 1531–1538.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. (2025). Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Ma, W., Zhang, H., Zhao, L., Song, Y., Wang, Y., Sui, Z., and Luo, F. (2025). Stabilizing moe reinforcement learning by aligning training and inference routers. *arXiv preprint arXiv:2510.11370*.

Marek, M. and Ryabinin, M. (2026). Score centering stabilizes off-policy reinforcement learning. *arXiv preprint arXiv:2609.20807*.

McAllester, D. and Stratos, K. (2020). Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR.

Qi, P., Zhou, X., Liu, Z., Pang, T., Du, C., Lin, M., and Lee, W. S. (2026). Rethinking the trust region in llm reinforcement learning. *arXiv preprint arXiv:2602.04879*.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. (2015a). High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

[14] The generative process is "continue with probability λ". The mass sums to $(1-\lambda)\sum_{k=0}^{K-t-1}\lambda^k + \lambda^{K-t} = 1$.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015b). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., *et al.* (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Team, K., Du, A., Gao, B., Xing, B., Jiang, C., Chen, C., Li, C., Xiao, C., Du, C., Liao, C., *et al.* (2025a). Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Team, L., Shen, A., Li, B., Hu, B., Jing, B., Chen, C., Huang, C., Zhang, C., Yang, C., Lin, C., *et al.* (2025b). Every step evolves: Scaling reinforcement learning for trillion-scale thinking model. *arXiv preprint arXiv:2510.18855*.

Tomar, M., Shani, L., Efroni, Y., and Ghavamzadeh, M. (2020). Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, **8**(3-4), 229–256.

Yao, F., Liu, L., Zhang, D., Dong, C., Shang, J., and Gao, J. (2025). On the rollout-training mismatch in modern rl systems. In *OPT 2025: Optimization for Machine Learning*.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., *et al.* (2026). Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, **38**, 113222–113244.

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., Zhou, J., and Lin, J. (2025). Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

A Markov Decision Process (MDP)

Let p_θ denote an LLM that generates a token $x \in \mathcal{V}$ with arbitrary conditioning. The model navigates an **environment** by state transition $(s, x) \mapsto s'$ (possibly stochastic) where a state $s \in \mathcal{S}$ is a full history of the interactions. To illustrate, a Unix environment could start with an initial state like

$$s_1 = \boxed{\begin{array}{l}\texttt{<system>} \text{ You may use these tools: terminal(command: string): Execute a shell command. } \texttt{</system>} \\ \texttt{<user>} \text{ How many files have write permissions in the current directory? } \texttt{</user>} \\ \texttt{<assistant>}\end{array}}$$

The model may decide to use the terminal tool and generate, token by token,

$$(x_1 \dots x_t) = \boxed{\text{Let me look. } \texttt{<tool>} \text{ \{“name”: “terminal”, “args”: \{“command”: “ls -l”\}\} } \texttt{</tool>} \ \textbf{</assistant>}}$$

with state transition $s_{i+1} = (s_i, x_i)$. Upon $x_t = \textbf{</assistant>}$, the controller inspects the completed message to (i) print *Let me look.*, (ii) execute `ls -l`, and (iii) append the observation to the next state, e.g.,

$$s_{t+1} = \boxed{\begin{array}{l}\texttt{<system>} \text{ You may use these tools: terminal(command: string): Execute a shell command. } \texttt{</system>} \\ \texttt{<user>} \text{ How many files have write permissions in the current directory? } \texttt{</user>} \\ \texttt{<assistant>} \text{ Let me look. } \texttt{<tool>} \text{ \{“name”: “terminal”, “args”: \{“command”: “ls -l”\}\} } \texttt{</tool>} \texttt{</assistant>} \\ \textbf{<observation> -rw-r–r– file1\textbackslash n-rwxr-xr-x file2 </observation> <assistant>}\end{array}}$$

The model goes on until it stops tool calling (e.g., just prints “There are 3 files with write permissions”). The user may engage and continue the session (“That’s wrong, check again”).

A.1 Training

For training purposes, the environment additionally supplies the knowledge of

- *Endpoint step* $K \geq 1$: Either s_{K+1} is a terminal state (**terminated**), or the rollout has reached a specified length/time limit (**truncated**).[15] Using T to denote the (random) termination step of the full trajectory, we have $K = T$ if terminated and $K < T$ if truncated.
- *Endpoint reward* $R(s_{K+1}) \in \mathbb{R}$: The reward is naturally binary in many verifiable tasks, but can also be more granular (e.g., rubric-based score by an LLM judge). We assume the reward is zero for nonterminal states, so $R(s_{K+1}) = 0$ if truncated.

B KL Estimation

From (16), it is sufficient to consider estimating the KL between two conditional token distributions

$$\mathrm{KL}(p, q) = \mathop{\mathbf{E}}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right]$$

Define the random variable $r = \frac{q(x)}{p(x)}$ where $x \sim p$. An unbiased estimator is obviously

$$k_1 = -\log r \tag{60}$$

John Schulman popularized an alternative estimator by this blog:

$$k_3 = r - 1 - \log r \tag{61}$$

which remains unbiased simply by the virtue of $\mathbf{E}[r] = 1$. The idea is that $r - 1$ offsets the first-order effect of $\log r$ around $r = 1$. Write $r = 1 + u$ and Taylor expand both around $u = 0$:

$$k_1 = -u + \frac{u^2}{2} + O(u^3) = O(u) \qquad\qquad k_3 = \frac{u^2}{2} + O(u^3) = O(u^2)$$

So k_3 is much flatter than k_1 around $u = 0$, implying lower variance in the regime $p \approx q$. For completeness, Schulman considers one other (biased) estimator:[16]

$$k_2 = \frac{1}{2}(\log r)^2 \tag{62}$$

C Lemmas

Lemma C.1. The gradient of $J(\theta) = \mathbf{E}_\theta[R(s_{T+1})]$ can be written as $\mathbf{E}_\theta[\sum_{t=1}^{T}(R(s_{T+1}) - b(s_t))\nabla \log p_\theta(x_t|s_t)]$.

Proof. Let $J_{s_1}(\theta) = \mathbf{E}_{\tau\sim p_\theta(\cdot|s_1)}[R(s_{T+1})]$ for any fixed s_1. Then

$$\begin{aligned}
\nabla J_{s_1}(\theta) = \sum_\tau R(s_{T+1})\nabla p_\theta(\tau|s_1) &= \sum_\tau R(s_{T+1}) p_\theta(\tau|s_1)\nabla \log p_\theta(\tau|s_1) && \text{(log-derivative trick)} \\
&= \mathop{\mathbf{E}}_{\tau\sim p_\theta(\cdot|s_1)}\left[R(s_{T+1})\nabla \log p_\theta(\tau|s_1)\right] \\
&= \mathop{\mathbf{E}}_{\tau\sim p_\theta(\cdot|s_1)}\left[\sum_{t=1}^{T} R(s_{T+1})\nabla \log p_\theta(x_t|s_t)\right] \\
&= \mathop{\mathbf{E}}_{\tau\sim p_\theta(\cdot|s_1)}\left[\sum_{t=1}^{T} (R(s_{T+1}) - b(s_t))\nabla \log p_\theta(x_t|s_t)\right] && \text{(baseline function)}
\end{aligned} \tag{63}$$

[15] A natural condition for termination is if $x_K =$ </assistant> and the response contains no tool call or simulated user followup. But it can also be domain-specific, e.g., generating an explicit `submit` action for coding.

[16] Although Schulman motivates (62) via f-divergence, there is no need. It can be directly seen as a quadratic approximation of k_3, since $k_3 = e^z - 1 - z \approx \frac{z^2}{2}$ where $z = \log r$. This is why k_2 has lower variance than k_1 around $p \approx q$ but is biased unlike k_3.

The last step uses a well-known property of the score function along with the fact that $b(s_t)$ is independent of x_t:

$$\begin{aligned}\mathop{\mathbf{E}}_{\tau\sim p_\theta(\cdot|s_1)}\left[\sum_{t=1}^{T} b(s_t)\nabla \log p_\theta(x_t|s_t)\right] &= \mathop{\mathbf{E}}_{\tau\sim p_\theta(\cdot|s_1)}\left[\sum_{t=1}^{T} \mathop{\mathbf{E}}_{x_t\sim p_\theta(\cdot|s_t)}\left[b(s_t)\nabla \log p_\theta(x_t|s_t)\right]\right] \\ &= \mathop{\mathbf{E}}_{\tau\sim p_\theta(\cdot|s_1)}\left[\sum_{t=1}^{T} b(s_t)\underbrace{\mathop{\mathbf{E}}_{x_t\sim p_\theta(\cdot|s_t)}\left[\nabla \log p_\theta(x_t|s_t)\right]}_{0_d}\right] = 0_d\end{aligned}$$

Taking $\nabla J(\theta) = \mathbf{E}_{s_1\sim\mathbf{env}}[\nabla J_{s_1}(\theta)]$ gives the result. □

Lemma C.2. Holding s_t fixed, we have

$$b^\star = \mathop{\arg\min}_{b\in\mathbb{R}} \operatorname{tr}\left(\operatorname{Cov}\left(\underbrace{(R(s_{T+1}) - b)\nabla \log p_\theta(x_t|s_t)}_{X}\right)\right) = \frac{\mathbf{E}[R(s_{T+1})\,||\nabla \log p_\theta(x_t|s_t)||_2^2]}{\mathbf{E}[||\nabla \log p_\theta(x_t|s_t)||_2^2]} \tag{64}$$

Proof. Recall the total variance of $X \in \mathbb{R}^d$ is computed as $\operatorname{tr}(\operatorname{Cov}(X)) = \operatorname{tr}\left(\mathbf{E}[XX^\top]\right) - \operatorname{tr}\left(\mathbf{E}[X]\mathbf{E}[X]^\top\right) = \mathbf{E}[\|X\|_2^2] - \|\mathbf{E}[X]\|_2^2$. Here we have the form $X = (R-b)g$ where $\mathbf{E}[X] = \mathbf{E}[Rg] - b\mathbf{E}[g] = \mathbf{E}[Rg]$ is independent of the optimization parameter, so the minimization problem (64) is equivalent to

$$b^\star = \mathop{\arg\min}_{b\in\mathbb{R}} \operatorname{tr}(\operatorname{Cov}((R-b)g)) = \mathop{\arg\min}_{b\in\mathbb{R}} \mathbf{E}[(R-b)^2\|g\|_2^2] = \frac{\mathbf{E}[R\|g\|_2^2]}{\mathbf{E}[\|g\|_2^2]}$$

□

Lemma C.3.

$$1 - \sum_x \min(P(x), Q(x)) = \min_{(X,Y):\ X\sim P,\ Y\sim Q} \Pr(X \neq Y) \tag{65}$$

Proof. Consider any (X, Y) with marginals $X \sim P$ and $Y \sim Q$. We have

$$\Pr(X = Y) = \sum_x \Pr(X = Y = x) \le \sum_x \min(P(x), Q(x)) \quad \Rightarrow \quad 1 - \sum_x \min(P(x), Q(x)) \le \Pr(X \neq Y) \tag{66}$$

Now we construct such (X, Y) that achieves equality, which shows (65). Let $m = \sum_x \min(P(x), Q(x))$.

- With probability m, draw $X = Y = x$ proportional to $\min(P(x), Q(x))$.
- With probability $1 - m$
  - Draw $X = x$ proportional to $P(x) - \min(P(x), Q(x))$, and
  - Draw $Y = y$ proportional to $Q(y) - \min(P(y), Q(y))$.

  Crucially, note that $X \neq Y$ in this case.

By construction, this defines a valid joint distribution over (X, Y) with $\Pr(X \neq Y) = 1 - m$. Thus we only need to verify they have the correct marginals:

$$\Pr(X = x) = m \times \frac{\min(P(x), Q(x))}{\sum_x \min(P(x), Q(x))} + (1-m) \times \frac{P(x) - \min(P(x), Q(x))}{\sum_x (P(x) - \min(P(x), Q(x)))} = P(x)$$

Similarly $\Pr(Y = x) = Q(x)$. □

Lemma C.4. $\operatorname{TV}(d_{\theta,t}, d_{\psi,t}) \le (t-1)\alpha$ where $\alpha = \max_s \operatorname{TV}(p_\theta(\cdot|s), p_\psi(\cdot|s))$.

Proof. By (66)

$$\mathrm{TV}(d_{\theta,t}, d_{\psi,t}) \le \Pr(s_{\theta,t} \neq s_{\psi,t})$$

where $s_{\theta,t} \sim d_{\theta,t}$ and $s_{\psi,t} \sim d_{\psi,t}$. By the union bound

$$\Pr(s_{\theta,t+1} \neq s_{\psi,t+1}) \le \Pr(s_{\theta,t} \neq s_{\psi,t}) + \Pr(s_{\theta,t} = s_{\psi,t} \wedge x_{\theta,t} \neq x_{\psi,t})$$

For the last term, WLOG we may characterize $x_{\theta,t}, x_{\psi,t}$ by the maximum coupling in Lemma C.3 whenever $s_{\theta,t} = s_{\psi,t} = s$ since they have the correct conditional marginals $p_\theta(\cdot|s)$ and $p_\psi(\cdot|s)$. Then

$$\begin{aligned}\Pr(s_{\theta,t} = s_{\psi,t} \wedge x_{\theta,t} \neq x_{\psi,t}) &= \sum_s \Pr(s_{\theta,t} = s_{\psi,t} = s) \Pr(x_{\theta,s} \neq x_{\psi,s} | s_{\theta,t} = s_{\psi,t} = s) \\ &= \sum_s \Pr(s_{\theta,t} = s_{\psi,t} = s) \underbrace{\mathrm{TV}(p_\theta(\cdot|s), p_\psi(\cdot|s))}_{\le \alpha} \le \alpha\end{aligned}$$

Since $s_{\theta,1} = s_{\psi,1}$, it follows $\Pr(s_{\theta,t} \neq s_{\psi,t}) \le (t-1)\alpha$. □

Lemma C.5. Let $\alpha = \max_s \mathrm{TV}(p_\theta(\cdot|s), p_\psi(\cdot|s))$ and assume $|R| \le \xi$ for some constant $\xi \ge 0$. Then

$$|(J(\theta) - J(\psi)) - J_\psi(\theta)| \le 4\xi T_{\max}(T_{\max} - 1)\alpha^2$$

Proof. The common factor between $J(\theta) - J(\psi)$ and $J_\psi(\theta)$ is the advantage $f(s) = \mathop{\mathbf{E}}_{x \sim p_\theta(\cdot|s)}[A_\psi(s,x)]$, where by (13)

$$|f(s)| = \left| \mathop{\mathbf{E}}_{x \sim p_\theta(\cdot|s)}[A_\psi(s,x)] - \mathop{\mathbf{E}}_{x \sim p_\psi(\cdot|s)}[A_\psi(s,x)] \right| \le 2 \underbrace{||A_\psi(s,\cdot)||_\infty}_{\le 2\xi} \underbrace{\mathrm{TV}(p_\theta(\cdot|s), p_\psi(\cdot|s))}_{\le \alpha} \le 4\xi\alpha$$

Expressing the difference in the common factor, again by (13)

$$|(J(\theta) - J(\psi)) - J_\psi(\theta)| \le \sum_{t=1}^{T_{\max}} \left| \mathop{\mathbf{E}}_{s \sim d_{\theta,t}}[f(s)] - \mathop{\mathbf{E}}_{s \sim d_{\psi,t}}[f(s)] \right| \le 2 \underbrace{||f||_\infty}_{\le 4\xi\alpha} \sum_{t=1}^{T_{\max}} \underbrace{\mathrm{TV}(d_{\theta,t}, d_{\psi,t})}_{\le (t-1)\alpha} \le 4\xi T_{\max}(T_{\max} - 1)\alpha^2$$

where $\mathrm{TV}(d_{\theta,t}, d_{\psi,t}) \le (t-1)\alpha$ by Lemma C.4. □

Lemma C.6. Let $\alpha = \max_s \mathrm{TV}(p_\theta(\cdot|s), p_\psi(\cdot|s))$ and assume $|R| \le \xi$ for some constant $\xi \ge 0$. Then

$$|J_\psi(\theta) - (J(\theta) - J(\psi))| \le 4\xi \mathop{\mathbf{E}}_{\tau \sim p_\psi}\left[\sum_{t=1}^{T} \mathrm{TV}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t)) \right] \le 4\xi T_{\max}\alpha \tag{67}$$

Proof. Let $r_t = \frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)}$, which is deterministic given s_t and x_t but random given only s_t with the expectation $\mathbf{E}_\psi[r_t|s_t] = 1$. We can write

$$J(\theta) - J(\psi) = \mathop{\mathbf{E}}_{\tau \sim p_\psi}\left[\frac{p_\theta(\tau|s_1)}{p_\psi(\tau|s_1)} R - R \right] = \mathop{\mathbf{E}}_{\tau \sim p_\psi}\left[R\left(\prod_{t=1}^{T} r_t - 1 \right) \right]$$

and also

$$\begin{aligned}J_\psi(\theta) = \mathop{\mathbf{E}}_{\tau \sim p_\psi}\left[\sum_{t=1}^{T} r_t \Big(\mathbf{E}_\psi[R|s_t, x_t] - \mathbf{E}_\psi[R|s_t] \Big) \right] &= \sum_{t=1}^{T_{\max}} \mathbf{E}_\psi[\mathbf{E}_\psi[r_t R|s_t, x_t] - \mathbf{E}_\psi[r_t|s_t]\mathbf{E}_\psi[R|s_t]] \\ &= \sum_{t=1}^{T_{\max}} \mathbf{E}_\psi[\mathbf{E}_\psi[r_t R|s_t, x_t] - \mathbf{E}_\psi[R|s_t]] \\ &= \mathop{\mathbf{E}}_{\tau \sim p_\psi}\left[R \sum_{t=1}^{T} (r_t - 1) \right]\end{aligned}$$

So they differ by prod-vs-sum and can be related through the telescoping product identity:

$$\prod_{t=1}^{T} r_t - 1 = \sum_{t=1}^{T} (r_t - 1) \prod_{j=t+1}^{T} r_j$$

Plugging this in, we have by triangle inequality

$$|J_\psi(\theta) - (J(\theta) - J(\psi))| = \left| \mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[R \sum_{t=1}^{T} (r_t - 1) \left(1 - \prod_{j=t+1}^{T} r_j \right) \right] \right| \leq \xi \mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[\sum_{t=1}^{T} \left| r_t - 1 \right| \left| 1 - \prod_{j=t+1}^{T} r_j \right| \right]$$

Here we can use the law of iterated expectations (LIE) with $|1 - z| \leq 1 + |z|$ to bound[17]

$$\mathbf{E}_\psi \left[\left| 1 - \prod_{j=t+1}^{T} r_j \right| \middle| s_{t+1} \right] \leq \mathbf{E}_\psi \left[1 + \prod_{j=t+1}^{T} r_j \middle| s_{t+1} \right] \leq 2$$

resulting in

$$|J_\psi(\theta) - (J(\theta) - J(\psi))| \leq 2\xi \mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[\sum_{t=1}^{T} |r_t - 1| \right] = 4\xi \mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[\sum_{t=1}^{T} \mathrm{TV}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t)) \right]$$

where the last equality uses another round of LIE

$$\mathbf{E}_\psi[|r_t - 1| \, |s_t] = \sum_{x_t} p_\psi(x_t|s_t) \left| \frac{p_\theta(x_t|s_t)}{p_\psi(x_t|s_t)} - 1 \right| = \sum_{x_t} |p_\theta(x_t|s_t) - p_\psi(x_t|s_t)| = 2\mathrm{TV}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t))$$

□

Corollary C.7. Assume $|R| \leq \xi$ for some constant $\xi \geq 0$. Then

$$|J_\psi(\theta) - (J(\theta) - J(\psi))| \leq 4\xi \sqrt{\frac{T_{\max}}{2} \mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[\sum_{t=1}^{T} \mathrm{KL}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t)) \right]} = 4\xi \sqrt{\frac{T_{\max}}{2} \mathrm{KL}_\psi(\theta)}$$

We may also switch the direction of KL inside the expectation without losing correctness.

Proof. We can bound the expectation in (67) (Lemma C.6) as

$$\begin{aligned}
\mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[\sum_{t=1}^{T} \mathrm{TV}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t)) \right] &\leq \sqrt{\mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[\left(\sum_{t=1}^{T} \mathrm{TV}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t)) \right)^2 \right]} && \text{(Jensen/Cauchy-Schwarz)} \\
&\leq \sqrt{\mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[T \left(\sum_{t=1}^{T} \mathrm{TV}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t))^2 \right) \right]} && \text{(Cauchy-Schwarz)} \\
&\leq \sqrt{\frac{T_{\max}}{2} \mathop{\mathbf{E}}_{\tau \sim p_\psi} \left[\sum_{t=1}^{T} \mathrm{KL}(p_\psi(\cdot|s_t), p_\theta(\cdot|s_t)) \right]} && \text{(Pinsker)}
\end{aligned}$$

Note that Pinsker could choose the other KL direction per token, but the rollout comes from ψ. □

[17] This is the loose part, since we bound this by 2 ignoring the ratios even when they are close to 1.