Group Relative Policy Optimization - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

Group Relative Policy Optimization (GRPO) is a policy-gradient method tailored to preference- or rule-based post-training of sequence models. It shares PPO's proximal update structure but *replaces* the critic-based advantage with a *group-normalized* reward computed over multiple samples (responses) per input prompt. This eliminates the need to learn a value function while retaining variance reduction via per-prompt baselines and whitening, and it admits both on-policy and off-policy training with importance weighting and a reference-policy KL regularizer.

1.2 Data and Notation

Let $\mathcal{X}$ be the input space (e.g., prompts) and $\mathcal{Y}$ the response space (token sequences of variable length). We consider a stochastic policy $\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$ over $\mathbf{y} \in \mathcal{Y}$ given $\mathbf{x} \in \mathcal{X}$, with parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$.

We are given either:

- **On-policy sampling:** draw $\mathbf{x}_i \sim \mu$ (a prompt distribution), then draw m i.i.d. responses $\{\mathbf{y}_{i,j}\}_{j=1}^m$ from the *current* policy $\pi_{\boldsymbol{\theta}_{\text{old}}}(\cdot \mid \mathbf{x}_i)$;
- **Off-policy sampling:** same, but sample $\{\mathbf{y}_{i,j}\}_{j=1}^m$ from a *behavior* policy $\pi_{\boldsymbol{\phi}}$ (e.g., a frozen or lagged policy).

Each response receives a scalar reward $R(\mathbf{x}_i, \mathbf{y}_{i,j}) \in \mathbb{R}$ (from a preference model, a verifiable checker, etc.). For convenience we define the *group* for prompt i as

$$\mathcal{G}_i \;=\; \big\{R_{i,1}, \ldots, R_{i,m}\big\}, \qquad R_{i,j} \;\equiv\; R(\mathbf{x}_i, \mathbf{y}_{i,j}).$$

Dimensions: n = number of prompts in a batch, m = group size (responses per prompt), p = number of parameters.

We also use a (possibly distinct) *reference* policy $\pi_{\boldsymbol{\theta}_{\text{ref}}}(\cdot \mid \mathbf{x})$ for KL regularization.

1.3 Model Formulation: Whitened (Group-Relative) Advantages

For each prompt i, define the group *sample mean* and *sample standard deviation* of rewards

$$\widehat{\mu}_i \;=\; \frac{1}{m}\sum_{j=1}^{m} R_{i,j}, \qquad \widehat{\sigma}_i \;=\; \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(R_{i,j}-\widehat{\mu}_i\right)^2+\varepsilon^2}, \tag{1}$$

with a small $\varepsilon > 0$ for numerical stability. The *group-relative (whitened) advantage* is then

$$\widehat{A}_{i,j}^{\text{grp}} \;=\; \frac{R_{i,j}-\widehat{\mu}_i}{\widehat{\sigma}_i}. \tag{2}$$

Two simple but important properties follow.

Zero-mean (baseline) property. For fixed group statistics $(\widehat{\mu}_i, \widehat{\sigma}_i)$, $\frac{1}{m}\sum_{j=1}^{m} \widehat{A}_{i,j}^{\text{grp}} = 0$. Thus $\widehat{\mu}_i$ acts as a per-prompt *baseline*, removing first-order reward location effects.

Scale invariance. For any $a > 0$ and $b \in \mathbb{R}$, replacing R by $aR+b$ leaves $\widehat{A}_{i,j}^{\text{grp}}$ unchanged (up to ε), hence improves robustness to reward scale and offset.

1.4 From Policy Improvement to a Constrained Objective

Let $J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\mu,\ \mathbf{y}\sim\pi_{\boldsymbol{\theta}}(\cdot\mid\mathbf{x})}[R(\mathbf{x},\mathbf{y})]$ be the expected reward. Direct maximization of J can be unstable; as in trust-region methods we constrain updates to stay close to a reference (or to the behavior distribution in off-policy training). A GRPO update is obtained by maximizing the *expected whitened advantage* subject to a KL trust region:

$$\max_{\boldsymbol{\theta}} \quad \mathbb{E}_{\mathbf{x}\sim\mu}\Big[\ \underbrace{\mathbb{E}_{\mathbf{y}\sim\pi_{\boldsymbol{\phi}}(\cdot\mid\mathbf{x})}\big[\rho_{\boldsymbol{\theta}/\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})\,A_{\boldsymbol{\phi}}^{\text{grp}}(\mathbf{x},\mathbf{y})\big]}_{\text{off-policy: IS correction; on-policy: } \boldsymbol{\phi}=\boldsymbol{\theta}_{\text{old}}}\ \Big] \quad \text{s.t.} \quad \mathbb{E}_{\mathbf{x}\sim\mu}\Big[D_{\text{KL}}\big(\pi_{\boldsymbol{\theta}}(\cdot\mid\mathbf{x})\,\big\|\,\pi_{\boldsymbol{\theta}_{\text{ref}}}(\cdot\mid\mathbf{x})\big)\Big] \;\le\; \delta, \tag{3}$$

where $\rho_{\boldsymbol{\theta}/\boldsymbol{\phi}}(\mathbf{x},\mathbf{y}) = \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}\mid\mathbf{x})}{\pi_{\boldsymbol{\phi}}(\mathbf{y}\mid\mathbf{x})}$, and $A_{\boldsymbol{\phi}}^{\text{grp}}$ denotes the whitened reward using group statistics computed under the sampling policy $\pi_{\boldsymbol{\phi}}$.[1]

Using the Lagrangian with multiplier $\beta > 0$ yields the penalized form

$$\mathcal{L}_{\text{pen}}(\boldsymbol{\theta}) \;=\; \mathbb{E}_{\mathbf{x}\sim\mu,\ \mathbf{y}\sim\pi_{\boldsymbol{\phi}}}\big[\rho_{\boldsymbol{\theta}/\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})\,A_{\boldsymbol{\phi}}^{\text{grp}}(\mathbf{x},\mathbf{y})\big] \;-\; \beta\,\mathbb{E}_{\mathbf{x}\sim\mu}\Big[D_{\text{KL}}\big(\pi_{\boldsymbol{\theta}}(\cdot\mid\mathbf{x})\,\big\|\,\pi_{\boldsymbol{\theta}_{\text{ref}}}(\cdot\mid\mathbf{x})\big)\Big]. \tag{4}$$

With Pinskers inequality, the KL-penalized problem is equivalent (for suitable β, δ) to the trust-region form Eqn. (3).

1.5 Likelihood-Ratio Gradient and Unbiasedness

Fix $\mathbf{x}$. For off-policy sampling $\mathbf{y} \sim \pi_{\boldsymbol{\phi}}(\cdot \mid \mathbf{x})$ and treating $A_{\boldsymbol{\phi}}^{\text{grp}}$ as a constant w.r.t. $\boldsymbol{\theta}$,

$$\nabla_{\boldsymbol{\theta}}\ \mathbb{E}_{\mathbf{y}\sim\pi_{\boldsymbol{\phi}}}\big[\rho_{\boldsymbol{\theta}/\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})\,A_{\boldsymbol{\phi}}^{\text{grp}}\big] \;=\; \mathbb{E}_{\mathbf{y}\sim\pi_{\boldsymbol{\phi}}}\big[\nabla_{\boldsymbol{\theta}}\rho_{\boldsymbol{\theta}/\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})\,A_{\boldsymbol{\phi}}^{\text{grp}}\big] = \mathbb{E}_{\mathbf{y}\sim\pi_{\boldsymbol{\theta}}}\big[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{y}\mid\mathbf{x})\,A_{\boldsymbol{\phi}}^{\text{grp}}\big]. \tag{5}$$

Thus the Monte-Carlo estimator $\widehat{g} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{y}_{i,j}\mid\mathbf{x}_i)\,\widehat{A}_{i,j}^{\text{grp}}$ is *unbiased* for the policy-gradient term and inherits variance reduction from the per-prompt baseline.

[1] In practice, the group statistics (mean/std) are treated as *stop-gradient* quantities when differentiating w.r.t. $\boldsymbol{\theta}$.

1.6 Clipped Surrogate Objective (On- and Off-Policy)

As in PPO, we replace Eqn. (4) by a *clipped* surrogate to enforce a soft trust region. For samples $\{(\mathbf{x}_i, \mathbf{y}_{i,j})\}$ drawn from $\pi_{\boldsymbol{\phi}}$, define importance ratios $r_{i,j}(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}_{i,j}|\mathbf{x}_i)}{\pi_{\boldsymbol{\phi}}(\mathbf{y}_{i,j}|\mathbf{x}_i)}$ and group advantages $\widehat{A}_{i,j}^{\text{grp}}$ from Eqn. (2). Given a clipping parameter $\epsilon \in (0,1)$, the GRPO clipped surrogate is

$$\mathcal{L}_{\text{clip}}^{\text{GRPO}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\min\Big(r_{i,j}(\boldsymbol{\theta})\,\widehat{A}_{i,j}^{\text{grp}},\ \text{clip}\big(r_{i,j}(\boldsymbol{\theta}),\,1-\epsilon,\,1+\epsilon\big)\,\widehat{A}_{i,j}^{\text{grp}}\Big) \\ -\ \beta\cdot\frac{1}{n}\sum_{i=1}^{n}D_{\text{KL}}\big(\pi_{\boldsymbol{\theta}}(\cdot\mid\mathbf{x}_i)\,\big\|\,\pi_{\boldsymbol{\theta}_{\text{ref}}}(\cdot\mid\mathbf{x}_i)\big). \tag{6}$$

On-policy GRPO corresponds to $\boldsymbol{\phi} = \boldsymbol{\theta}_{\text{old}}$ (ratios vs. the most recent policy and group stats computed on-policy). **Off-policy** GRPO sets $\boldsymbol{\phi}$ to a fixed behavior policy: the same clipped form Eqn. (6) applies, with group statistics computed under $\pi_{\boldsymbol{\phi}}$ and fresh batches sampled from $\pi_{\boldsymbol{\phi}}$.

1.7 Reward-Improvement Lower Bound (Sketch)

Let $J(\boldsymbol{\theta})$ be the expected reward and let $A_{\boldsymbol{\phi}}^{\text{grp}}$ be the whitened reward under a policy $\pi_{\boldsymbol{\phi}}$. Under bounded, non-degenerate rewards and a small update staying near $\pi_{\boldsymbol{\phi}}$ (or $\pi_{\boldsymbol{\theta}_{\text{ref}}}$) in total variation (hence in KL), one shows

$$J(\boldsymbol{\theta})\ -\ J(\boldsymbol{\theta}_{\text{old}})\ \gtrsim\ C\,\mathbb{E}_{\mathbf{x}\sim\mu}\Big[\ \mathbb{E}_{\mathbf{y}\sim\pi_{\boldsymbol{\phi}}}\big[r_{\boldsymbol{\theta}/\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})\,A_{\boldsymbol{\phi}}^{\text{grp}}(\mathbf{x},\mathbf{y})\big]\ \Big]\ -\ \widetilde{C}\cdot\mathbb{E}_{\mathbf{x}\sim\mu}\big[\text{TV}(\pi_{\boldsymbol{\theta}},\pi_{\boldsymbol{\phi}})\big], \tag{7}$$

for positive constants $C, \widetilde{C}$ depending on reward dispersion terms. Bounding TV by $\sqrt{\frac{1}{2}D_{\text{KL}}}$ (Pinsker) motivates the KL penalty and the clipped ratio in Eqn. (6) as practical surrogates ensuring monotone improvement under small steps. (Formal statements and proofs for on- and off-policy GRPO follow this template; see Theorem/Corollary analogues in recent analyses.)

1.8 Token-Level Factorization (Sequence Models)

For autoregressive models with tokens $y_{1:T}$, $\pi_{\boldsymbol{\theta}}(\mathbf{y}\mid\mathbf{x}) = \prod_{t=1}^{T}\pi_{\boldsymbol{\theta}}(y_t \mid y_{<t},\mathbf{x})$. Then

$$\log r_{i,j}(\boldsymbol{\theta}) = \sum_{t=1}^{T_{i,j}}\log\frac{\pi_{\boldsymbol{\theta}}(y_{i,j,t}\mid y_{i,j,<t},\mathbf{x}_i)}{\pi_{\boldsymbol{\phi}}(y_{i,j,t}\mid y_{i,j,<t},\mathbf{x}_i)},$$

and the policy-gradient estimator decomposes over tokens. In practice, the group-relative advantage $\widehat{A}_{i,j}^{\text{grp}}$ is a *sequence-level* scalar broadcast across tokens of that response.

1.9 Algorithm (GRPO, On- or Off-Policy)

1. **Input:** batch size n, group size m, clip ϵ, KL weight β, reference $\pi_{\boldsymbol{\theta}_{\text{ref}}}$, sampling policy $\pi_{\boldsymbol{\phi}}$ (on-policy: $\boldsymbol{\phi} = \boldsymbol{\theta}_{\text{old}}$).
2. **Collect groups:** For $i = 1,\ldots,n$, sample $\mathbf{x}_i \sim \mu$; sample m responses $\mathbf{y}_{i,1:m} \sim \pi_{\boldsymbol{\phi}}(\cdot\mid\mathbf{x}_i)$; compute rewards $R_{i,1:m}$.
3. **Compute group advantages:** For each i, compute $(\widehat{\mu}_i, \widehat{\sigma}_i)$ and $\widehat{A}_{i,j}^{\text{grp}}$ via Eqn. (2).
4. **Optimize:** Update $\boldsymbol{\theta}$ to *maximize* $\mathcal{L}_{\text{clip}}^{\text{GRPO}}(\boldsymbol{\theta})$ in Eqn. (6) (stop-grad through group stats).
5. **Iterate / stage:** Optionally refresh $\boldsymbol{\phi}$ (on-policy: set $\boldsymbol{\phi} \leftarrow \boldsymbol{\theta}$), and repeat.

1.10 Proofs: Baseline Invariance and Whitening

Lemma (baseline invariance). Fix prompt i. For any constants $a > 0, b \in \mathbb{R}$, define $\widetilde{R}_{i,j} = aR_{i,j} + b$. Then the group-relative advantages computed from $\widetilde{R}_{i,j}$ equal those from $R_{i,j}$: $\widetilde{A}^{\text{grp}}_{i,j} = A^{\text{grp}}_{i,j}$.

Proof. The group mean transforms as $\widetilde{\mu}_i = a\mu_i + b$, the (population) std as $\widetilde{\sigma}_i = a\sigma_i$. Hence $\frac{\widetilde{R}_{i,j}-\widetilde{\mu}_i}{\widetilde{\sigma}_i} = \frac{a(R_{i,j}-\mu_i)}{a\sigma_i} = \frac{R_{i,j}-\mu_i}{\sigma_i}$, and the same holds for empirical statistics up to ε. ∎

Lemma (zero-mean whitened reward). For fixed $(\widehat{\mu}_i, \widehat{\sigma}_i)$, $\frac{1}{m}\sum_j \widehat{A}^{\text{grp}}_{i,j} = 0$ and $\frac{1}{m}\sum_j (\widehat{A}^{\text{grp}}_{i,j})^2 = 1$ (up to ε).

Proof. Immediate from centering and scaling by the empirical mean and std. ∎

Proposition (unbiased policy-gradient estimator). Assume off-policy sampling $\mathbf{y} \sim \pi_{\boldsymbol{\phi}}(\cdot \mid \mathbf{x})$ and treat $A^{\text{grp}}_{\boldsymbol{\phi}}$ as constant w.r.t. $\boldsymbol{\theta}$. Then

$$\nabla_{\boldsymbol{\theta}}\, \mathbb{E}_{\mathbf{y}\sim\pi_{\boldsymbol{\phi}}}\big[\rho_{\boldsymbol{\theta}/\boldsymbol{\phi}}\, A^{\text{grp}}_{\boldsymbol{\phi}}\big] = \mathbb{E}_{\mathbf{y}\sim\pi_{\boldsymbol{\theta}}}\big[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}\, A^{\text{grp}}_{\boldsymbol{\phi}}\big],$$

hence the empirical estimator using Eqn. (2) is unbiased.

Proof. Use $\nabla_{\boldsymbol{\theta}}\rho = \rho\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}$ and change of measure. ∎

1.11 Summary of Variables and Their Dimensions

- $\mathbf{x}_i \in \mathcal{X}$: prompt; $i = 1, \ldots, n$ (batch size n).
- $\mathbf{y}_{i,j} \in \mathcal{Y}$: jth response for prompt i; $j = 1, \ldots, m$ (group size m).
- $R_{i,j} \in \mathbb{R}$: scalar reward for $(\mathbf{x}_i, \mathbf{y}_{i,j})$.
- $\widehat{\mu}_i, \widehat{\sigma}_i \in \mathbb{R}$: per-prompt sample mean/std of $R_{i,1:m}$.
- $\widehat{A}^{\text{grp}}_{i,j} \in \mathbb{R}$: group-relative (whitened) advantage Eqn. (2).
- $\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$: target policy; $\boldsymbol{\theta} \in \mathbb{R}^p$.
- $\pi_{\boldsymbol{\phi}}$: sampling/behavior policy (on-policy: $\boldsymbol{\phi} = \boldsymbol{\theta}_{\text{old}}$).
- $\pi_{\boldsymbol{\theta}_{\text{ref}}}$: reference policy for KL regularization.
- $r_{i,j}(\boldsymbol{\theta})$: importance ratio $\frac{\pi_{\boldsymbol{\theta}}}{\pi_{\boldsymbol{\phi}}}$ for $(\mathbf{x}_i, \mathbf{y}_{i,j})$.
- $\epsilon \in (0, 1)$: clipping parameter in Eqn. (6); $\beta \geq 0$: KL weight; $\varepsilon > 0$: numerical stability in std.

1.12 Summary

GRPO replaces value-based advantages by per-prompt, group-normalized rewards, preserving the baseline benefits (zero-mean) and adding scale invariance. Starting from a KL-constrained objective and applying standard PPO machinery yields the clipped surrogate Eqn. (6) for both on- and off-policy training. Under mild assumptions, maximizing the whitened-advantage surrogate with small KL steps ensures reward improvement, while avoiding the cost and instability of learning a critic.