GSPO Gradient Derivation

Yigit Okar

July 2025

1 Introduction

Consider the Group Sequence Policy Optimization (GSPO) objective for reinforcement learning with large language models. For a query x, let $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ denote G sampled responses from an old policy. The GSPO objective is defined as

$$J_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim D, \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(s_i(\theta)\hat{A}_i, \text{clip}(s_i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i \right) \right],$$

where the importance ratio $s_i(\theta)$ is computed at the sequence level using:

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)} \right)^{1/|y_i|} = \exp \left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})} \right),$$

and the normalized advantage is defined as

$$\hat{A}_i = \frac{r(x, y_i) - \text{mean}(\{r(x, y_j)\}_{j=1}^{G})}{\text{std}(\{r(x, y_j)\}_{j=1}^{G})}.$$

The final expression for a valid subgradient of the GSPO objective is:

$$\nabla_\theta J_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim D, \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \nabla_\theta L_i(\theta) \right],$$

where $\nabla_\theta L_i(\theta)$ is a valid subgradient of the loss for a single response, given by:

$$\nabla_\theta L_i(\theta) = C_i(\theta) \cdot \hat{A}_i \cdot s_i(\theta) \cdot \left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) \right),$$

and $C_i(\theta)$ is an indicator function that captures the clipping effect:

$$C_i(\theta) = \mathbf{1}_{\hat{A}_i > 0, s_i(\theta) \le 1+\varepsilon} + \mathbf{1}_{\hat{A}_i < 0, s_i(\theta) \ge 1-\varepsilon}.$$

This indicator is 1 if the update is not clipped and 0 otherwise.

2 Proof Sketch

The derivation of the gradient of the GSPO objective function proceeds through the following logical steps:

1. **Interchange of Subgradient and Expectation:** The GSPO objective $J_{\text{GSPO}}(\theta)$ is an expectation over data sampled from a distribution independent of the policy parameters θ. The inner loss function is non-differentiable due to the 'min' and 'clip' operations. We use subgradient calculus and justify interchanging the subgradient and expectation operators by invoking the Dominated Convergence Theorem for subgradients.

2. **Rigorous Justification:** To rigorously apply the theorem, we first prove a key lemma bounding the subgradient of the scalar loss function.

   - **Lemma (Subgradient Bound):** Let $F(s) = \min(s\hat{A}_i, \text{clip}(s, 1-\varepsilon, 1+\varepsilon)\hat{A}_i)$. For any $s \in \mathbb{R}$ and any subgradient element $k \in \partial_s F(s)$, the bound $|k| \leq |\hat{A}_i|$ holds. We provide a full proof of this lemma via case analysis on the sign of $\hat{A}_i$.
   - **Dominating Function:** Using this lemma and the chain rule for subgradients, we construct a dominating function $g(\omega)$ for the norm of any element in the subgradient set of the full loss term.
   - **Integrability Proof:** We prove that $g(\omega)$ is integrable under a set of standard, formal assumptions, thus validating the interchange of subgradient and expectation.

3. **Derivation of a Specific Subgradient:** With the interchange justified, we compute a valid subgradient of the inner loss term, $L_i(\theta) = \min\big(s_i(\theta)\hat{A}_i, \text{clip}(s_i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i\big)$.

   - **Complete Case Analysis:** We perform a complete case analysis based on the sign of the advantage $\hat{A}_i$: $\hat{A}_i > 0$, $\hat{A}_i < 0$, and $\hat{A}_i = 0$.
   - **Subgradient at Boundaries:** At points of non-differentiability, we select a specific, valid element from the subgradient set (a one-sided derivative), which is a standard and theoretically sound choice for optimization algorithms.
   - **Indicator Function:** The outcome of the case analysis is concisely expressed using an indicator function $C_i(\theta)$, which is 1 when the gradient is passed through and 0 when it is clipped or when $\hat{A}_i = 0$.

4. **Final Assembly:** We derive the gradient of the importance ratio, $\nabla_\theta s_i(\theta)$, and substitute all components back into the main expression to obtain the final form of $\nabla_\theta J_{\text{GSPO}}(\theta)$.

3 Detailed Proof

The GSPO objective function is defined as

$$J_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim D,\ \{y_i\}\sim\pi_{\theta_{\text{old}}}}\Big[\tfrac{1}{G}\sum_{i=1}^{G}\min\big(s_i(\theta)\hat{A}_i,\ \text{clip}(s_i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i\big)\Big].$$

Let $\omega = (x, \{y_i\}_{i=1}^G)$ denote a sample from the data-generating distribution $P(\omega)$, independent of θ. Let $L(\theta;\omega) = \frac{1}{G}\sum_{i=1}^G L_i(\theta;\omega)$, where $L_i(\theta;\omega) = \min\bigl(s_i(\theta)\hat{A}_i, \operatorname{clip}(s_i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i\bigr)$. The objective is $J_{\text{GSPO}}(\theta) = \mathbb{E}_{\omega\sim P}[L(\theta;\omega)]$.

3.1 Justification for Interchanging Subgradient and Expectation

The function $L(\theta;\omega)$ is non-differentiable due to the 'min' and 'clip' operations. We therefore work with subgradients. To compute the gradient of the objective we must justify interchanging the subgradient and expectation operators:

$$\nabla_\theta \mathbb{E}[L(\theta;\omega)] \; = \; \mathbb{E}\bigl[\partial_\theta L(\theta;\omega)\bigr],$$

where $\partial_\theta L$ denotes the subgradient set. The Dominated Convergence Theorem for subgradients allows this interchange if two conditions hold:

1. For each sample ω, the function $L(\theta;\omega)$ is locally Lipschitz in θ. 2. There exists an integrable function $g(\omega)$ with $\mathbb{E}[g(\omega)] < \infty$ that dominates every subgradient: $\|\zeta\| \le g(\omega)$ for all $\zeta \in \partial_\theta L(\theta;\omega)$ and all θ.

3.1.1 Condition 1: Local Lipschitz Continuity

Under standard assumptions of policy smoothness, $s_i(\theta)$ is a composition of differentiable functions, hence locally Lipschitz. The 'clip' and 'min' functions are globally Lipschitz (constant 1). Their composition is therefore locally Lipschitz, so $L(\theta;\omega)$ satisfies Condition 1.

3.1.2 Condition 2: Dominating Integrable Function

We first establish a key lemma:

Lemma: Bounding the Subgradient of the Scalar Loss

Let $F(s) = \min(s\hat{A}_i, \operatorname{clip}(s, 1-\varepsilon, 1+\varepsilon)\hat{A}_i)$. For any $s \in \mathbb{R}$ and any $k \in \partial_s F(s)$ we have $|k| \le |\hat{A}_i|$.

Proof: (sketch).* We split on the sign of $\hat{A}_i$.

Case 1: $\hat{A}_i > 0$ $F(s) = \hat{A}_i \min\bigl(s, \operatorname{clip}(s, 1-\varepsilon, 1+\varepsilon)\bigr)$. Write $G(s) = \min(s, \operatorname{clip}(s, 1-\varepsilon, 1+\varepsilon))$. Then $\partial_s F(s) = \hat{A}_i\, \partial_s G(s)$ and $\partial_s G(s) \subseteq [0,1]$, so $|k| \le \hat{A}_i$.

Case 2: $\hat{A}_i < 0$ $F(s) = \hat{A}_i \max\bigl(s, \operatorname{clip}(s, 1-\varepsilon, 1+\varepsilon)\bigr)$. A symmetric argument gives $|k| \le |\hat{A}_i|$.

Case 3: $\hat{A}_i = 0$ $F(s) \equiv 0$, so $k = 0$.

□

Using the lemma and standard bounds (e.g. $\|\nabla_\theta \log \pi_\theta\| \le G_{\max}$, $|\hat{A}_i| \le A_{\max}$) we construct

$$g(\omega) \; = \; \frac{A_{\max} G_{\max}}{G} \sum_{i=1}^{G} \sup_{\theta' \in \Theta} s_i(\theta'),$$

which dominates $\|\partial_\theta L(\theta;\omega)\|$ and is integrable under the usual finite-importance-ratio assumption, establishing Condition 2.

3.2 Derivation of a One-Sample Subgradient

We compute a valid selection from $\partial_\theta L_i(\theta)$ by a case analysis on $\hat{A}_i$.

Case $\hat{A}_i > 0$

$$\nabla_\theta L_i(\theta) = \begin{cases} \hat{A}_i \, \nabla_\theta s_i(\theta) & s_i(\theta) \le 1 + \varepsilon, \\ \mathbf{0} & s_i(\theta) > 1 + \varepsilon. \end{cases}$$

Case $\hat{A}_i < 0$

$$\nabla_\theta L_i(\theta) = \begin{cases} \mathbf{0} & s_i(\theta) < 1 - \varepsilon, \\ \hat{A}_i \, \nabla_\theta s_i(\theta) & s_i(\theta) \ge 1 - \varepsilon. \end{cases}$$

Case $\hat{A}_i = 0$ Gradient is zero.

These three branches combine into

$$\nabla_\theta L_i(\theta) = C_i(\theta) \, \hat{A}_i \, \nabla_\theta s_i(\theta),$$

with

$$C_i(\theta) = \mathbf{1}[\hat{A}_i > 0 \wedge s_i(\theta) \le 1 + \varepsilon] + \mathbf{1}[\hat{A}_i < 0 \wedge s_i(\theta) \ge 1 - \varepsilon].$$

3.3 Gradient of the Importance Ratio

$$\nabla_\theta s_i(\theta) = s_i(\theta) \, \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}).$$

3.4 Final Gradient

Substituting $\nabla_\theta s_i(\theta)$ into $C_i(\theta)\hat{A}_i\nabla_\theta s_i(\theta)$ and averaging over i gives

$$\nabla_\theta J_{\text{GSPO}}(\theta) = \mathbb{E}_{x,\{y_i\}} \Big[\tfrac{1}{G} \sum_{i=1}^{G} C_i(\theta) \, \hat{A}_i \, s_i(\theta) \, \big(\tfrac{1}{|y_i|} \sum_{t=1}^{|y_i|} \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) \big) \Big].$$