Goal: Understand the need of stop gradient for GSPO-token.

Observation 1. Let f be a differentiable function, x a scalar or vector input, and $\mathrm{sg}[\cdot]$ the stop-gradient operator. The stop-gradient operator returns the same numerical value as its argument but is treated as a constant during backpropagation. Formally,

$$\frac{d}{dx}\,\mathrm{sg}\big(f(x)\big) = 0.$$

Observation 2. Assume we had no stop gradient. Then we have

$$s_{i,t}(\theta) = s_i(\theta)\,\frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}.$$

Taking the derivative with respect to the parameters:

$$\frac{d}{d\theta}s_{i,t}(\theta) = \frac{d}{d\theta}s_i(\theta)\,\frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}.$$

Note that the fraction on the right is constant (equal to 1), so its derivative is 0. This results in us pushing gradient through the importance weight $s_i(\theta)$, which is exactly what we want to prevent:

$$\frac{d}{d\theta}s_{i,t}(\theta) = \frac{d}{d\theta}s_i(\theta).$$

In GSPO-token, $s_i(\theta)$ is meant to be a stable sequence-level importance ratio that corrects for off-policy sampling. If gradients are allowed to update $s_i(\theta)$ here, we are no longer treating it as a fixed correction factor. Instead, we are optimizing it directly. I believe the authors claim this can destabilize training and undermine the intended off-policy correction.

Observation 3: Consider the version with stop gradient. In GSPO-token, they define

$$s_{i,t}(\theta) = \mathrm{sg}[s_i(\theta)] \cdot \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\mathrm{sg}[\pi_\theta(y_{i,t} \mid x, y_{i,<t})]}.$$

Applying Observation 1, both $\mathrm{sg}[s_i(\theta)]$ and $\mathrm{sg}[\pi_\theta(\cdot)]$ are constants with respect to θ during backpropagation. Differentiating gives

$$\frac{d}{d\theta}s_{i,t}(\theta) = \frac{\mathrm{sg}[s_i(\theta)]}{\mathrm{sg}[\pi_\theta(y_{i,t} \mid x, y_{i,<t})]} \cdot \frac{d}{d\theta}\pi_\theta(y_{i,t} \mid x, y_{i,<t}).$$

Recognizing that $\frac{1}{\mathrm{sg}[\pi_\theta]}\,\frac{d}{d\theta}\pi_\theta = \frac{d}{d\theta}\log\pi_\theta$, we can write

$$\frac{d}{d\theta}s_{i,t}(\theta) = \mathrm{sg}[s_i(\theta)] \cdot \frac{d}{d\theta}\log\pi_\theta(y_{i,t} \mid x, y_{i,<t}).$$

This shows that the sequence-level importance ratio $\mathrm{sg}[s_i(\theta)]$ acts purely as a fixed multiplier on the token-level log-probability gradient.