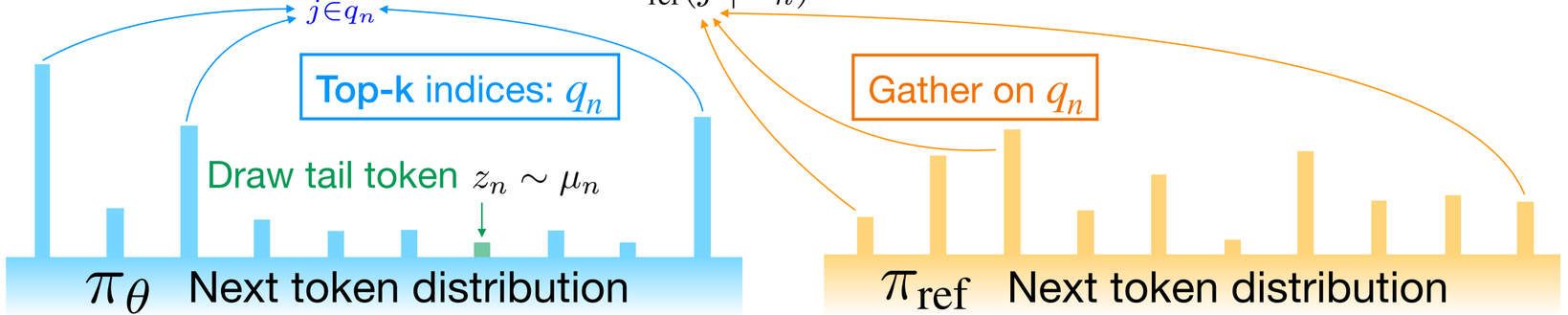
$$(\mathrm{TopkKL})_n = \sum_{j \in q_n} \pi_\theta(j \mid h_n) \log \frac{\pi_\theta(j \mid h_n)}{\pi_{\mathrm{ref}}(j \mid h_n)} - \Delta_n + s_n \, (\widehat{\mathrm{KL}}_{\mathrm{sampled}})_n$$



Sampled KL estimator **K4**

$$(\widehat{\mathrm{KL}}_{\mathrm{sampled}})_n = \frac{\pi_\theta(z_n \mid h_n)}{\mathrm{sg}(\pi_\theta(z_n|h_n))} \, \mathrm{sg}\left(\log \frac{\pi_\theta(z_n \mid h_n)}{\pi_{\mathrm{ref}}(z_n \mid h_n)}\right)$$

$z_n \sim \mu_n, \quad \mu_n(j) \propto \pi_\theta(j \mid h_n) \, \mathbb{1}(j \notin q_n)$

$s_n = \mathrm{sg}(\pi_\theta(z_n \mid h_n) \,/\, \mu_n(z_n))$ importance weight

$\Delta_n = \sum_{j \in q_n} [\pi_\theta(j \mid h_n) - \mathrm{sg}(\pi_\theta(j \mid h_n))]$

h_n: past context $(x, y_{0:n-1})$ sg: stop gradient