CB Regret

$$\max_{\pi^*} \sum_t \mathbb{E}_{a \sim \pi^*(o_t)} r(a) - \sum_t r(a_t)$$

RL Regret:

$R_t = \sum_h r(a_h)$ in t^{th} episode

$$\text{Regret} = \max_{\pi^*} \sum_t \mathbb{E}_{\tau \sim \pi^*} \left[\sum_h r_h \right]$$

τ ↑ $\langle o_1, a_1, r_1, o_2, \ldots, a_{H-1}, r_H \rangle$

$$- \sum_t R_t$$

$$\mathbb{E}_{o_1 \sim \text{world}} \mathbb{E}_{a_1 \sim \pi(o_1)} \mathbb{E}_{r_1, o_2 \sim \text{world}} \mathbb{E}_{a_2 \sim \pi(o_2)}$$

$$\mathbb{E} \ldots \ldots \mathbb{E}_{r_H} \left[\sum_h r_h \right]$$

Policy gradient (for CB)

$$\pi(a|o) = \frac{\exp[g(o,a)]}{\sum_{a'} \exp[g(o,a')]}$$

Goal: $\max_{\pi} \underbrace{\mathbb{E}_{o,r} \mathbb{E}_{a\sim\pi(o)} r(a)}_{R(\pi)}$

attempt: $\nabla_\pi R(\pi)$

$$R(\pi) = \mathbb{E}_{o,r} \sum_a \pi(a|o)\, r(a)$$

∇_π

$$\frac{\partial \log f(x)}{\partial x} = \frac{\partial f}{\partial x} \cdot \frac{1}{f(x)}$$

$$\Leftrightarrow \frac{\partial f}{\partial x} = f(x) \frac{\partial \log f(x)}{\partial x}$$

$$\nabla_\pi R(\pi) = \mathbb{E}_{o,r} \sum_a \left[\nabla_\pi \pi(a|o) \right] r(a)$$

$$= \mathbb{E}_{o,r} \sum_a \left[\pi(a|o) \nabla_\pi \log \pi(a|o) \right] \times r(a)$$

$$= \mathbb{E}_{o,r} \mathbb{E}_{a \sim \pi} r(a) \nabla_\pi \log \pi(a|o)$$

Monte-Carlo estimate

[redacted]

$$\approx \sum_{(o,r)} r(\pi(o)) \nabla_\pi \log \pi(\pi(o)|o)$$

sample from $\pi(a)$