➢ A general version of RL policy update forms

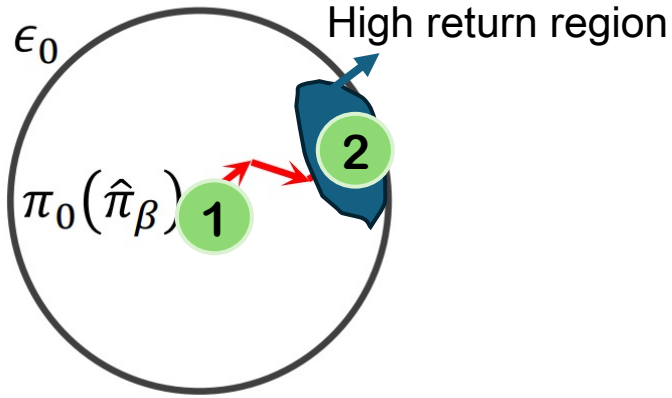
Policy evaluation

$$\hat{Q}^{k+1} = \arg\min_{Q} \mathbb{E}_{(s,a)\sim w}\Big[\Big(\underbrace{T^{\pi^k}\hat{Q}}_{r+\gamma\,\mathbb{E}_{s'|s,a}\,\mathbb{E}_{a'\sim\pi^k(\cdot|s')}[\hat{Q}(s',a')]} - Q(s,a)\Big)^2\Big]. \quad \cdots \; 1$$

Policy improvement

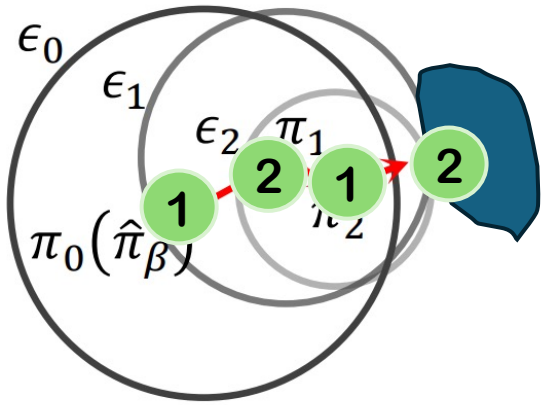
$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{s\sim\rho,\,a\sim\pi}\big[\hat{Q}^{k+1}(s,a)\big] - \beta\,\mathbb{E}_{s\sim\rho}\big[D_{\mathrm{KL}}(\pi \,\|\, \pi_{\mathrm{ref}})\big]. \quad \cdots \; 2$$

➢ One-step RL



- One-step RL/IQL/IDQL…

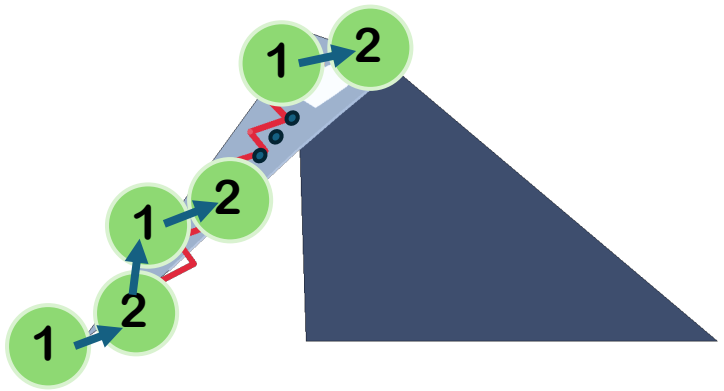
➢ Multi-step RL



- BPPO → Uni-O4/RL-100…

More conservative, but stable

➢ Iterative RL



- PPO/SAC/…

More conservative, but stable