Understanding Curriculum Learning in Policy Optimization for Online Combinatorial Optimization

Runlong Zhou[1] Yuandong Tian[2] Yi Wu[3] Simon S. Du[1]

[1]University of Washington [2]Meta AI [3]Tsinghua University

Motivation & Goal

- ML is good at Combinatorial Optimization (CO) problems — Is (any part of) this success explainable?
- Online CO matches the nature of RL — **Sequential decision-making**
- Theoretical understanding of RL techniques — **Curriculum Learning** — on online CO

Example 1: Secretary Problem

Setting:

- Hire one secretary among n candidates, each with different score.
- Arrive sequentially, but the order is unknown.
- Once reject someone, cannot revoke; Once hire someone, ends.
- **Maximize the probability of hiring the candidate with the highest score.**

Abstraction:

- A distribution over all $n!$ permutations (ordering of the candidates).
- The observation of the agent can only be "whether the i-th candidate is the so-far best", so it cannot distinguish between each permutation.

A policy working averagely well on the instance distribution?

Example 2: Online Knapsack

- n items arrive sequentially, value v_i and size s_i **revealed upon arrival**.
- Once reject something, cannot revoke; To pick something, requires total size $\leqslant B$.
- **Maximize total value of picked items.**
- Distribution of instances: each $(v_i, s_i) \sim F_v \times F_s$ i.i.d.

Formulation: Latent MDP

- A distribution $\{w_1, \ldots, w_M\}$ over MDPs $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$.
- Shared state set $\mathcal{S}$, action set $\mathcal{A}$, horizon H.
- Possibly distinct initial state distribution ν_m, transition P_m, reward r_m.
- Value of policy π, V^π, is defined as the w-weighted average on those of individual MDPs.

Result 1: Natural Policy Gradient for LMDP

Log-linear policy: $\pi_\theta(a|s) = \frac{\exp(\theta^\top\phi(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\theta^\top\phi(s,a'))}$, where $\theta \in \mathbb{R}^d$.

Regularized value, Q and advantage functions: $V^{\pi,\lambda}_{m,h}(s)$, $Q^{\pi,\lambda}_{m,h}(s,a)$ w.r.t. reward $r_m(s_t,a_t) + \lambda \ln \frac{1}{\pi(a_t|s_t)}$, and $A^{\pi,\lambda}_{m,h}(s,a) := Q^{\pi,\lambda}_{m,h}(s,a) - V^{\pi,\lambda}_{m,h}(s)$.

Visitation distribution: $d^\pi_{m,h}(s)$, $d^\pi_{m,h}(s,a)$ are the probability that the h-th step is a certain state(-action pair). $\widetilde{d}^\pi_{m,h}(s,a) := d^\pi_{m,h}(s) \circ \mathsf{Unif}_{\mathcal{A}}(a)$.

Function approximation loss: Let v be any visitation distribution,

$$L(g;\theta,v) := \sum_{m=1}^{M} w_m \sum_{h=1}^{H} \mathbb{E}_{s,a\sim v_{m,H-h}}\left[\left(A^{\pi_\theta,\lambda}_{m,h}(s,a) - g^\top \nabla_\theta \ln \pi_\theta(a|s)\right)^2\right].$$

Learning procedure: $\theta_{t+1} \approx \theta_t + \eta \arg\min_{\|g\|_2\leqslant G} L(g;\theta_t, d^{\pi_{\theta_t}})$ because we have no access to the true value of L.

Fisher information matrix: Let v be any visitation distribution,

$$\Sigma^\theta_v := \sum_{m=1}^{M} w_m \sum_{h=1}^{H} \mathbb{E}_{s,a\sim v_{m,H-h}}\left[\nabla_\theta \ln\pi_\theta(a|s)\left(\nabla_\theta \ln \pi_\theta(a|s)\right)^\top\right].$$

Let $g^\star_t \in \arg\min_{\|g\|_2\leqslant G} L(g;\theta_t,d^t)$ denote the true minimizer. Define:

- (Excess risk) $\epsilon_{\text{stat}} := \max_t \mathbb{E}[L(g_t;\theta_t,d^t) - L(g^\star_t;\theta_t,d^t)]$;
- (Transfer error) $\epsilon_{\text{bias}} := \max_t \mathbb{E}[L(g^\star_t;\theta_t,d^\star)]$;
- (Relative condition number) $\kappa := \max_t \mathbb{E}\left[\sup_{x\in\mathbb{R}^d} \frac{x^\top \Sigma^{\theta_t}_{d^\star} x}{x^\top \Sigma_t x}\right]$.

Theorem 1: Assume $\|\phi(s,a)\|_2 \leqslant B$. NPG for LMDP satisfies:

$$\mathbb{E}\left[\min_{0\leqslant t\leqslant T} V^{\star,\lambda} - V^{t,\lambda}\right] \leqslant \frac{\lambda(1-\eta\lambda)^{T+1}\Phi(\pi_0)}{1-(1-\eta\lambda)^{T+1}} + \eta\frac{B^2G^2}{2} + \sqrt{H\epsilon_{\text{bias}}} + \sqrt{H\kappa\epsilon_{\text{stat}}},$$

where $\Phi(\pi_0)$ is only relevant to the initialization.

Remark 1:

- First result of sample-based, regularized NPG on LMDP.
- Fix $\lambda > 0 \Rightarrow$ linear convergence, matching discounted infinite horizon MDP; $\lambda \to 0 \Rightarrow O(1/(\eta T) + \eta) \Rightarrow O(1/\sqrt{T})$ convergence.
- ϵ_{stat} can be reduced using a larger batch size N, $\epsilon_{\text{stat}} = \widetilde{O}(1/\sqrt{N})$.
- If some d_t (especially the initialization d_0) is far away from $d^\star$, κ may be extremely large. Initialization with a small κ is of great help.

Result 2: Curriculum Learning for Online CO

Concept of CL: Learn to solve problems from easier ones to harder ones.

Example: For the classical Secretary Problem with 100 candidates, use curricula of classical Secretary Problems with $10, 20, \ldots, 90$ candidates.

We found this multi-step CL unnecessary!

Why? From Remark 1, a pre-trained model reduces κ.

Learning from a different distribution is also possibly helpful!

Theoretical result for the Secretary Problem:

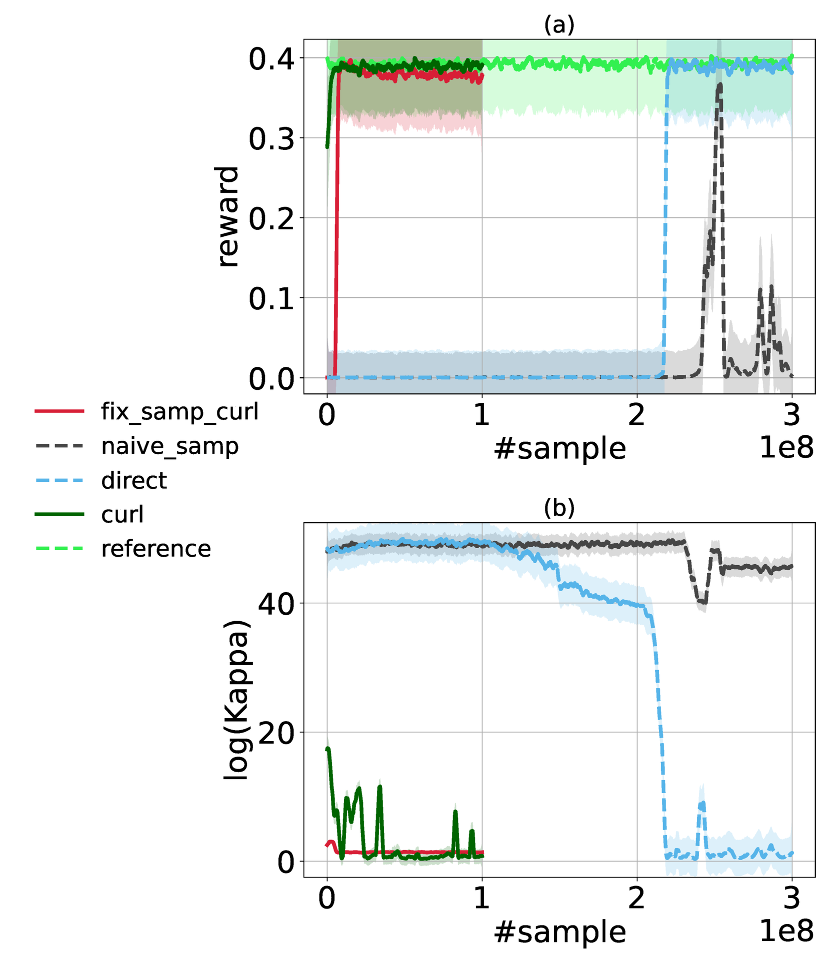
- Suppose for each candidate i, it has probability P_i to be the so-far best. For classical SP, $P_i = 1/i$.
- Optimal policy is always a p-threshold policy: accept if and only if $i/n > p$ and is so-far best. For classical SP, $p = 1/\mathrm{e}$.

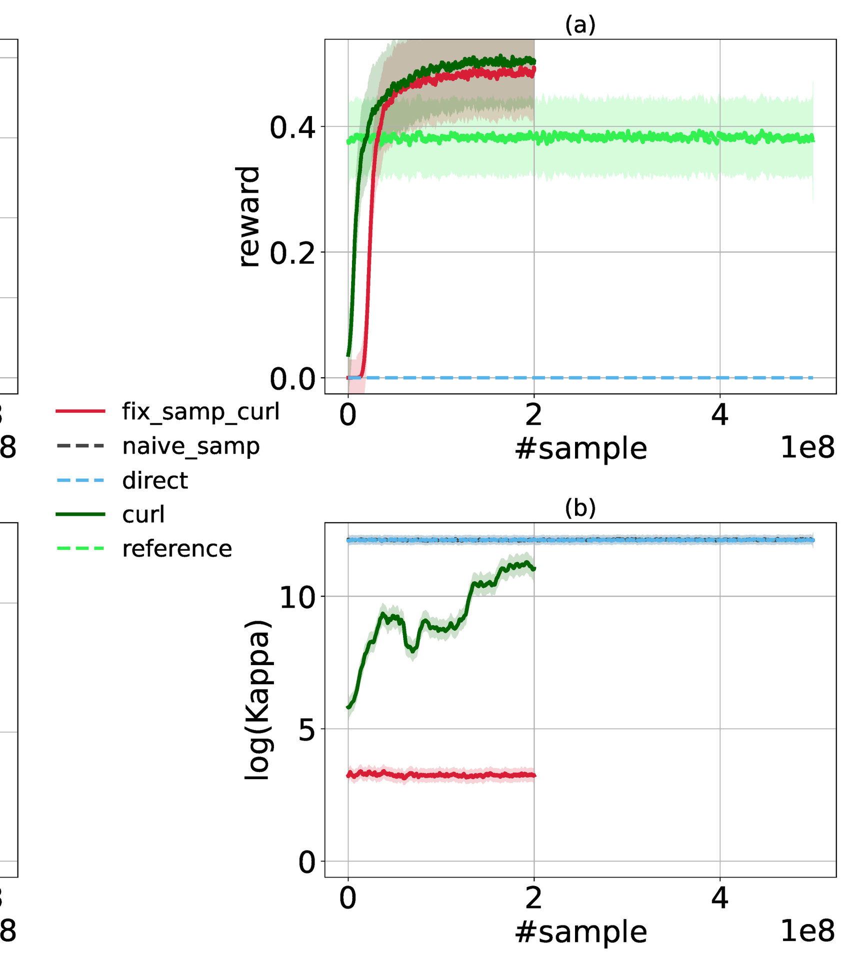
Theorem 2: For the Secretary Problem, assume the optimal policy is a p-threshold policy and CL returns a q-threshold policy as initialization:

$$\kappa_{\text{CL}} = \Theta\left(\begin{cases}\prod_{j=\lfloor nq\rfloor+1}^{\lfloor np\rfloor}\frac{1}{1-P_j}, & q \leqslant p,\\ 1, & q > p,\end{cases}\right),$$

$$\kappa_{\text{naïve}} = \Theta\left(2^{\lfloor np\rfloor}\max\left\{1, \max_{i\geqslant\lfloor np\rfloor+2}\prod_{j=\lfloor np\rfloor+1}^{i-1} 2(1-P_j)\right\}\right).$$

- **Classical case:** $q \geqslant 1/n \Rightarrow \kappa_{\text{CL}} = O(n)$ while $\kappa_{\text{naïve}} = \Omega(2^n)$.
- $P_i \leqslant 1/2$ **case:** $\kappa_{\text{CL}} \leqslant 2^{\lfloor np\rfloor - \lfloor nq\rfloor}$ while $\kappa_{\text{naïve}} \geqslant 2^{\lfloor np\rfloor}$.
- **Failure case:** If $q < 1 - 1/n$ and $P_j > 1 - 2^{-\frac{n}{n-\lfloor nq\rfloor-1}}$ for any $\lfloor nq\rfloor + 1 \leqslant j \leqslant n-1$, then $\kappa_{\text{CL}} > 2^n > \kappa_{\text{naïve}}$. (See paper.)





vectorzh@cs.washington.edu