Revisiting Regularized Policy Optimization for Stable and Efficient Reinforcement Learning in Two-Player Games

Kazuki Ota[1,2] Takayuki Osa[2] Motoki Omura[1] Tatsuya Harada[1,2]

[1]The University of Tokyo [2]RIKEN AIP

ICML International Conference On Machine Learning

Key Question

Search-based approaches are highly successful in two-player games, but expensive to train.

Can we develop an efficient model-free RL algorithm for two-player games?

Our Approach

We introduce a model-free RL algorithm *KLENT* with KL and entropy regularizations.

$$\underset{\pi'}{\text{maximize}}\ \underbrace{\mathbb{E}_{A\sim\pi'(\cdot|s)}[Q_\theta(s,A)]}_{\text{expected return}} - \underbrace{\beta D_{\mathrm{KL}}(\pi'(\cdot|s)\|\pi_\theta(\cdot|s))}_{\text{proximity to previous policy}} + \underbrace{\alpha H(\pi'(\cdot|s))}_{\text{entropy bonus}}.$$

Algorithm 1 KLENT Algorithm

Initialize the policy network $\pi_\theta(a|s)$.
Initialize the action-value network $Q_\theta(s,a)$.
repeat
  $\mathcal{D} \leftarrow \{\}$
  repeat
    Initialize the state S_0.
    for $t = 0, \ldots, T$ **do**
      $\pi'(a|S_t) \propto \exp\left(\frac{Q_\theta(S_t,a)+\beta \log \pi_\theta(a|S_t)}{\alpha+\beta}\right)$
      $\hat{v}_t \leftarrow \mathbb{E}_{A\sim\pi'(\cdot|S_t)}\left[Q_\theta(S_t,A)\right]$
      Sample $A_t \sim \pi'(\cdot|S_t)$.
      Execute A_t and observe (S_{t+1}, R_t).
    end for
    Compute λ-returns $\{G_t^\lambda\}_{t=0}^T$.
    $\mathcal{D} \leftarrow \mathcal{D} \cup \left\{(S_t, A_t, (\pi'(a|S_t))_{a\in\mathcal{A}}, G_t^\lambda)\right\}_{t=0}^T$
  until $\mathcal{D}$ reaches a predefined capacity.
  Update θ by minimizing $L(\theta)$.
until convergence.

Prepare policy and action-value networks.

Use the analytical solution of above for action sampling.

Use λ-returns for action value learning.

Fit the networks with: $L(\theta) = \mathbb{E}_{\mathcal{D}}\left[-\sum_{a\in\mathcal{A}} \pi'(a|S)\log\pi_\theta(a|S) + (Q_\theta(S,A) - G^\lambda)^2\right]$.

Theoretical Analysis

We proved the convergence of KLENT on normal-form games and finite-length games.

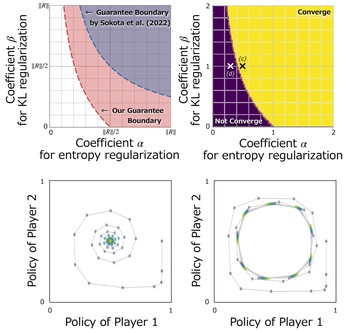
Normal-Form Games

Convergence Guarantee

Theorem 1

KLENT is linearly and locally convergent to the entropy regularized Nash equilibrium in normal-form games, if the payoff matrix R and coefficients α, β satisfies $\alpha(\alpha+2\beta) \geq 4\|R\|$.

Numerical Verification



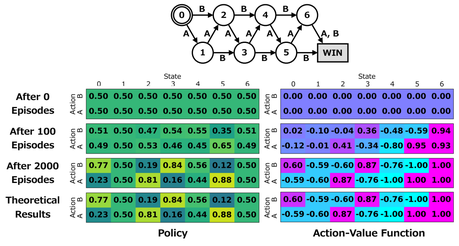
Finite-Length Games

Convergence Guarantee

Theorem 2

KLENT is convergent to the entropy regularized Nash equilibrium in finite-length games.

Numerical Verification

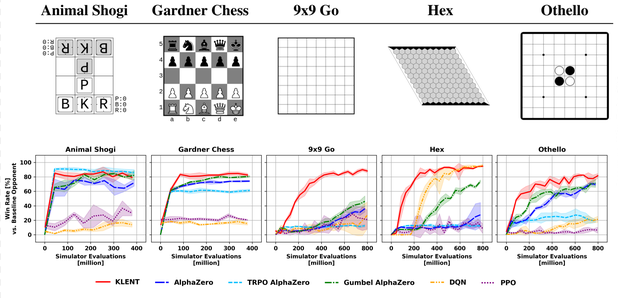


We verified our theoretical results with numerical experiments on synthetic games for both settings.

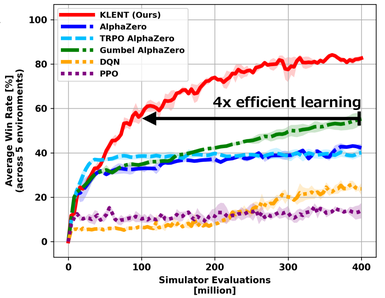
Experiments

Our agent achieves 4x higher training efficiency than existing methods across five board games.

Environments & Performance Comparison

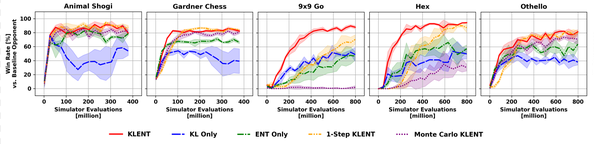


Average Performances

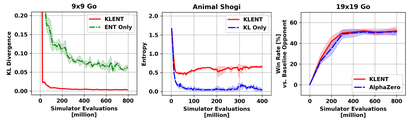


Ablation study shows the importance of combining KL regularization, entropy regularization, and λ-returns.

Performances of Ablated Approaches



Further Results



Conclusion

Model-free RL can effectively learn in two-player games by properly revisiting regularized policy optimization.