Adaptive Model Design for Markov Decision Process

Siyu Chen*[1], Donglin Yang*[1], Jiayang Li[2], Senmiao Wang[2], Zhuoran Yang[3] and Zhaoran Wang[2]

July, 2021

[1]Tsinghua University [2]Northwestern University [3]Yale University

Designing Markov Decision Process

- MDP: A powerful tool for modeling various dynamic planning problems
  - financial investment, repair and maintenance, resource management, robotic control, …
- Externality
  - Uncooperative agent with self interest.
  - Detrimental to other individuals in the system or the system's overall performance.

Regularized Markov Decision Process

- Optimal policy π_ϵ^* with policy entropy regularization

$$\pi_\epsilon^* = \underset{\pi}{\operatorname{argmax}} \langle \pi(\cdot\,|s), Q_\epsilon^*(s,\cdot) \rangle_{\mathcal{A}} - \epsilon^{-1} \sum_{\text{a}} \Omega(\pi(a|s))$$

$$Q_\epsilon^*(s,a) = r(s,a) + \gamma \cdot \mathbb{E}_{P(\cdot|s,a)}[V_\epsilon^*(\cdot)]$$

$$V_\epsilon^*(s) = \max_{\pi} \langle \pi(\cdot\,|s), Q_\epsilon^*(s,\cdot) \rangle_{\mathcal{A}} - \epsilon^{-1} \sum_{\text{a}} \Omega(\pi(a|s))$$

  - Ω is a strictly convex and doubly differentiable function
- Example: KL divergence $\sum_{\text{a}} \Omega(\pi(a|s)) = \langle \pi, \log \pi \rangle_{\mathcal{A}}$

  - $\pi_\epsilon^*(a|s) = \exp Q_\epsilon^*(s,a) / \sum_a \exp Q_\epsilon^*(s,a)$

Bounded rationality

Question

- How to adaptively design the reward function/transition kernel in an MDP to induce a desirable outcome that fulfills the designer's objective?

Problem Formulation

$$\pi_\epsilon^* = \underset{\pi}{\operatorname{argmax}} \langle \pi(\cdot\,|s), Q_\epsilon^*(s,\cdot)\rangle_{\mathcal{A}} - \epsilon^{-1} \sum_{a} \Omega(\pi(a|s))$$

- The original MDP design (OMD) problem

$$\text{OMD}: \quad \max_{\theta \in \mathcal{X}} F(\theta, \pi^*),$$

$$\text{s.t.} \quad \pi^* \in \Pi^*(\mathcal{S}, \mathcal{A}, \gamma, P(\theta), r(\theta)),$$

- OMD is non-singleton and ill-defined when Π^* has more than one element
- The optimal policy can be dis-continuous concerning θ

- The regularized MDP design (RMD) problem

$$\text{RMD}: \quad \max_{\theta \in \mathcal{X}} F(\theta, \pi),$$

$$\text{s.t.} \quad \pi = \pi_\epsilon^*(\mathcal{S}, \mathcal{A}, \gamma, P(\theta), r(\theta)),$$

- Assume bounded rationality in the MDP agent by introducing entropy regularization in the agent's policy

Sub-optimality of the RMD

Given Δ_π, Δ_r s.t., $\Delta_r \geq \epsilon^{-1}\left(\gamma U_\Omega + (1+\gamma)\log\left(\frac{2|\mathcal{A}|}{\Delta_\pi}\right)\right)$

We have,

Optimistic OMD

$$\max_\theta F(\theta, \pi_\epsilon^*(r_\theta)) \leq \max_\theta \max_{\substack{\pi \in \Pi^*(P(\theta), \hat{r}(\theta)), \\ \hat{r}(\cdot) \in \hat{R}(\Delta_r)}} F(\theta, \pi) + \Delta_\pi L_{F,\pi,0},$$

Pessimistic OMD

$$\max_\theta F(\theta, \pi_\epsilon^*(r_\theta)) \geq \max_\theta \min_{\substack{\pi \in \Pi^*(P(\theta), \hat{r}(\theta)), \\ \hat{r}(\cdot) \in \hat{R}(\Delta_r)}} F(\theta, \pi) - \Delta_\pi L_{F,\pi,0}.$$

General Framework for Solving RMD-Gradients

- Use KL divergence as the entropy regularization, we can obtain

$$\nabla_\theta \pi_\epsilon^*(a|s) = \epsilon \cdot \pi_\epsilon^*(a|s) \cdot \nabla_\theta A_\epsilon^*(s,a)$$

$$\nabla_\theta V_\epsilon^*(s) = \mathbb{E}_{\pi_\epsilon^*(\cdot|s)}[\nabla_\theta Q_\epsilon^*(s,\cdot)]$$

$$\nabla_\theta Q_\epsilon^* = \mathcal{T}^\theta_{\nabla_\vartheta r,\gamma}(\nabla_\theta V_\epsilon^* + V_\epsilon^* \nabla_\theta \ln P)$$

$$\nabla_\vartheta A_\epsilon^*(s,a) = \nabla_\theta Q_\epsilon^*(s,a) - \nabla_\theta V_\epsilon^*(s)$$

  - $\mathcal{T}$ is a Bellman operator defined as follows,

$$\mathcal{T}^\theta_{r,\gamma}(V)(s,a) = r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a;\vartheta)}[V(\cdot)]$$

- The gradient of the designer's objective function F

$$\nabla_\theta F = \frac{\partial F}{\partial \theta} + \epsilon \mathbb{E}_{\rho^{\pi_\epsilon^*}}\left[\rho^{-1} \cdot \frac{\partial F}{\partial \vartheta} \cdot \nabla_\theta A_\epsilon^*\right]$$

  - ρ is a reference distribution for sampling across the state space.

Benefits of regularization

- Well-defined problem
- Smoother landscape, Improved stability
- Improved exploration and robustness
- Easy gradient

RMD-Algorithm

Total Reward as Design Objective

for $t = 0$ **to** $T-1$ **do**
  for $k = 0$ **to** $K-1$ **do**
    $\pi_\epsilon^k(\cdot|s) \propto \exp\left(\epsilon Q_\epsilon^k(s,\cdot)\right)$
    $V_\epsilon^k(s) = \epsilon^{-1}\ln\left(\sum_a \exp\left(\epsilon Q_\epsilon^k(s,a)\right)\right)$
    $\nabla_{\theta_t} V_\epsilon^k(s) = \mathbb{E}_{\pi_\epsilon^K}\left[\nabla_{\theta_t} Q_\epsilon^K(s,a)\right]$
    $Q_\epsilon^{k+1} = \mathcal{T}_{r,\gamma}^{\theta}(V_\epsilon^k)$
    $\nabla_{\theta_t} Q_\epsilon^{k+1} = \mathcal{T}_{\nabla_{\theta_t} r,\gamma}^{\theta}(\nabla_{\theta_t} V_\epsilon^k + V_\epsilon^k \nabla_{\theta_t} \ln P)$
  end for
  $\nabla_{\theta_t} A_\epsilon^K(s,a) = \nabla_{\theta_t} Q_\epsilon^K(s,a) - \nabla_{\theta_t} V_\epsilon^K(s)$
  $\nabla_{\theta_t} F = \frac{\partial F}{\partial \theta_t} + \epsilon \mathbb{E}_{\rho^{\pi_\epsilon^K}}\left[\rho^{-1} \cdot \frac{\partial F}{\partial \pi_\epsilon^K} \cdot \nabla_{\theta_t} A_\epsilon^K\right]$
  $\theta_{t+1} = \theta_t + \eta \nabla_{\theta_t} F$
  Reinitialize $Q_\epsilon^0 = Q_\epsilon^K$ and $\nabla_{\theta_{t+1}} Q_\epsilon^0 = \nabla_{\theta_t} Q_\epsilon^K$
end for

for $t = 0$ **to** $T-1$ **do**
  for $k = 0$ **to** $K-1$ **do**
    $\pi_\epsilon^k(\cdot|s) \propto \exp\left(\epsilon Q_\epsilon^k(s,\cdot)\right)$
    Calculate $V_\epsilon^k, \nabla_{\theta_t} V_\epsilon^k, V_u^k, \nabla_{\theta_t} A_\epsilon^k, A_u^k, \tilde{V}^k$
    $Q_\epsilon^{k+1} = \mathcal{T}_{r,\gamma}(V_\epsilon^k)$
    $\nabla_{\theta_t} Q_\epsilon^{k+1} = \mathcal{T}_{\nabla_{\theta_t} r,\gamma}(\nabla_{\theta_t} V_\epsilon^k + V_\epsilon^k \nabla_{\theta_t} \ln P)$
    $Q_u^{k+1} = \mathcal{T}_{r_u,\gamma_u}(V_u^k)$
    $\tilde{Q}^{k+1} = \mathcal{T}_{\nabla_{\theta_t} r_u + \epsilon A_u \nabla_{\theta_t} A_\epsilon, \gamma_u}(\tilde{V}^k + V_u^k \nabla_{\theta_t} \ln P)$
  end for
  $\nabla_{\theta_t} F = \mathbb{E}_{\mathcal{D}_0}[\tilde{V}^K]$
  $\theta_{t+1} = \theta_t + \eta \nabla_{\theta_t} F$
  Reinitialize $Q_\epsilon^0 = Q_\epsilon^K$, $\nabla_{\theta_{t+1}} Q_\epsilon^0 = \nabla_{\theta_t} Q_\epsilon^K$, and $\nabla_{\theta_{t+1}} Q_\epsilon^0 = \nabla_{\theta_t} Q_\epsilon^K$.
end for

Convergence Analysis

- Convergence of the Inner Loop
  - After K inner iterations

$$\|\nabla_\theta Q_\epsilon^K - \nabla_\theta Q_\epsilon^*\|_{\theta\sim 2,(s,a)\sim\infty}$$
$$\leq \gamma^K K \|Q_\epsilon^0 - Q_\epsilon^*\|_\infty \cdot \left(4\epsilon\|\nabla_\theta Q_\epsilon^*\|_{\theta\sim 2,(s,a)\sim\infty} + \|\nabla_\theta P\|_{\theta\sim 2,s'\sim 1,(s,a)\sim\infty}\right)$$
$$+ \gamma^K \|\nabla_\theta Q_\epsilon^0 - \nabla_\theta Q_\epsilon^*\|_{\theta\sim 2,(s,a)\sim\infty}$$

- Convergence of the Outer Loop
  - Under proper regularization conditions, by appropriately setting the inner iteration number K and the learning rate η, it holds that

$$l_\epsilon(\theta_T) - l_\epsilon(\theta^*) \leq O(T^{-1/2})$$

where $l_\epsilon(\theta) = -F\left(\theta, \pi_\epsilon^*(\theta)\right)$.

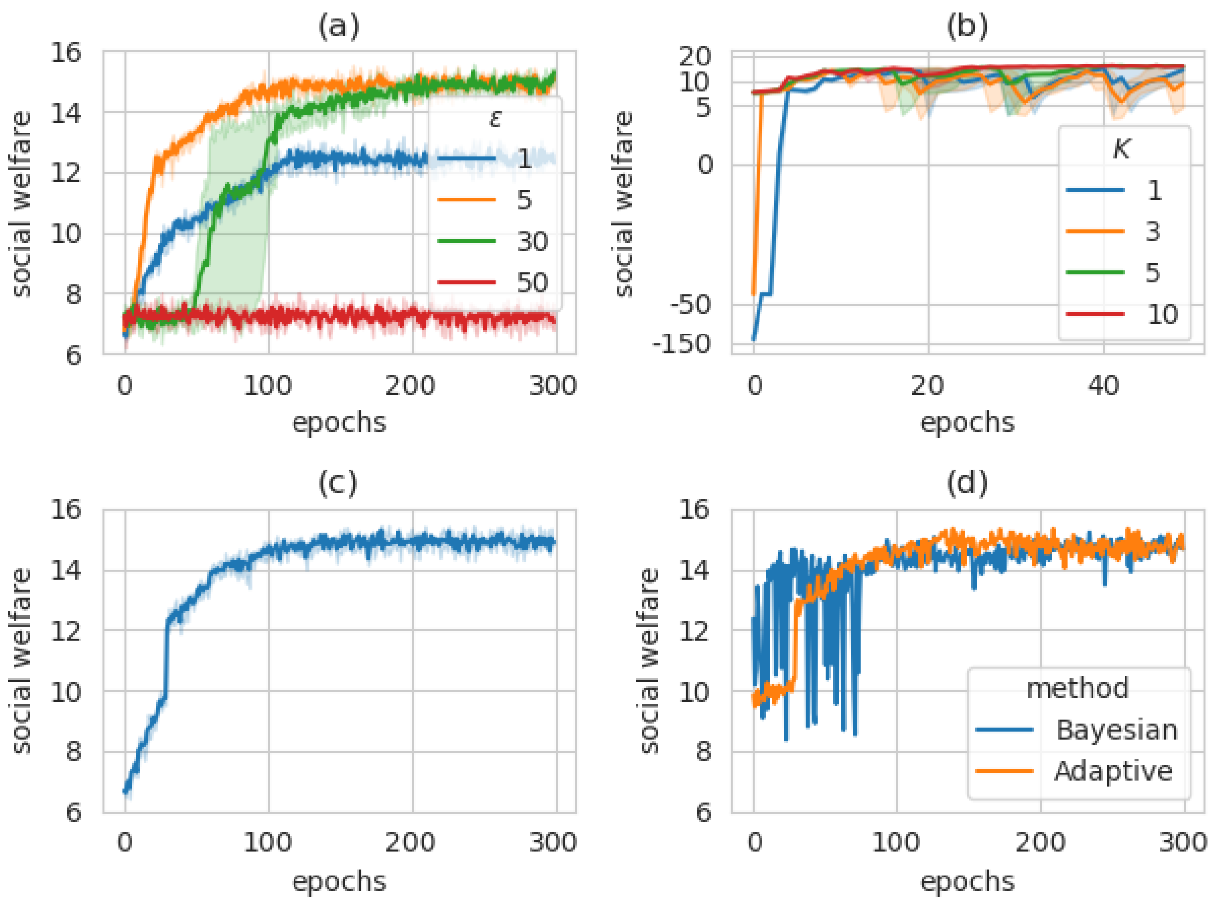
Extentions

- ϵ-Adaptive Strategy
  - a smaller ϵ: smoother optimization landscape, improved stability, fewer inner iterations
  - a larger ϵ: more accurate in the design objective function
  - We adjust ϵ during the update (from small to large).

Experiments

Tax Design for Macroeconomic Model

- (a) different ϵ
- (b) different inner loop K
- (c) adaptive strategy ϵ
- (e) comparison with Bayesian optimization



Thank you!