A Kernel Loss for Solving Bellman Equation

Yihao Feng  Lihong Li  Qiang Liu

December 2, 2019
Problems involving an agent interacting with an environment, which provides numeric reward signals

**Goal:** Learn how to take actions in order to maximize reward
**Objective:** Balance a pole on top of a movable cart

**State:** angle, angular speed, position, horizontal velocity

**Action:** horizontal force applied on the cart

**Reward:** 1 at each time step if the pole is upright
Markov Decision Process (MDP)

- Mathematical formulation of RL problem: \( M = \langle S, A, P, R, \gamma \rangle \)
  - Set of states \( S \)
  - Set of actions \( A \)
  - Transition probabilities \( P(s' | s, a) \)
  - Immediate expected reward \( R(s, a) \)
  - Discount factor \( \gamma \in (0, 1) \)
- Goal: Find \( \pi^* : S \rightarrow A \) to maximize

\[
\forall, s \in S, V(s) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_t \sim \pi(a_t | s_t), s_{t+1} \sim P(\cdot | a_t, s_t) \right].
\]

- \( V(s) \) is the value function of policy \( \pi \).
- This presentation we focus on learning value function \( V(s) \).
**Bellman Equation**

- **Bellman equation**: $V = B^\pi V$, where $B^\pi$ is the Bellman operator for policy $\pi$, defined by

  $$B^\pi V(s) := \mathbb{E}_{a \sim \pi(\cdot|s), s' \sim P(\cdot|a, s)}[r(s, a) + \gamma V(s')| s].$$

  $V = V^\pi$ is the unique solution of Bellman equation.

- **Bellman Optimal equation**: $V = B^* V$, where $B^*$ is the Bellman Optimal operator, defined by

  $$B^* V(s) := \max_a \mathbb{E}_{s' \sim P(\cdot|s, a)}[r(s, a) + \gamma V(s')| s, a].$$

  The unique solution of $V = B^* V$ is the optimal value function, denoted $V^*$.  

- In this presentation, we focus on $V = B^\pi V$, or how to learn the value function $V^\pi$ given a fixed policy $\pi$.  

We want to estimate $V^\pi$ from a parametric family \( \{V_\theta : \theta \in \Theta\} \) from data $D := \{(s_i, a_i, r_i, s'_i)\}_{1 \leq i \leq n}$.

A basic algorithm (residual gradient) minimizes the squared TD error:

$$
\hat{L}_{RG}(V_\theta) := \frac{1}{n} \sum_{i=1}^{n} \left( \hat{B}_\pi V_\theta(s_i) - V_\theta(s_i) \right)^2,
$$

which is a biased and inconsistent estimate of the squared Bellman error: $L_2(V_\theta) := \|B_\pi V - V\|_\mu^2$.

**Double-sampling** problem!

$$
\mathbb{E}_{s \sim \mu}[\hat{L}_{RG}(V)] = L_2(V_\theta) + \mathbb{E}_{s \sim \mu}[\text{var}(\hat{B}_\pi V(s))|s] \neq L_2(V_\theta).
$$

$$
\mathbb{E}[X^2] = (\mathbb{E}[X])^2 + \text{Var}[X].
$$

Biased estimation unless the MDP is deterministic.
Fitted Value Iteration (FVI).

\[ \theta_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \hat{L}_{FVI}^{t+1}(V_\theta) := \frac{1}{n} \sum_{i=1}^{n} \left( V_\theta(s_i) - \hat{B}_\pi V_\theta(s_i) \right)^2 \right\}. \]

- TD(0) can be viewed as a stochastic version of FVI.
- FVI-style algorithm does not optimize any objective function.
- Convergence is guaranteed in rather restricted case, which is also known as deadly triad.
A Divergent Example

- Init $w \neq 0$
- FVI diverges when $\gamma > 5/6$.
- But we know the optimal $w^* = 0$. 
Alternative View of squared Bellman Error

- Denote $R_\pi V = B_\pi V - V$. Suppose we have an auxiliary function $f(s)$ with $\mathbb{E}_{s \sim \mu}[f(s)^2] \leq 1$, and according to Cauchy–Schwarz inequality, we have

$$L_2(V_\theta) := \mathbb{E}_{s \sim \mu}[(R_\pi V_\theta(s))^2] \geq \mathbb{E}_{s \sim \mu}[(R_\pi V_\theta(s))^2] \cdot \mathbb{E}_{s \sim \mu}[(f(s))^2] \geq (\mathbb{E}_{s \sim \mu}[R_\pi V_\theta(s) \cdot f(s)])^2.$$

- We have a new objective function as:

$$L_2(V_\theta) \geq \max_{f} \left\{ (\mathbb{E}_{s \sim \mu}[R_\pi V_\theta(s) \cdot f(s)])^2 : \mathbb{E}_{s \sim \mu}[f(s)^2] \leq 1 \right\}.$$

- How can we choose $f$?
Choice of $f$

- Suppose $V^\pi \in \{V_\theta : \theta \in \Theta\}$, and $V_\theta$ is smooth, and we can assume $f(s) \in \mathcal{H}_K$, where $\mathcal{H}_K$ is a Reproducing Kernel Hilbert Space (RKHS) associated with a kernel $K(s, \bar{s})$:

$$L_K(V_\theta) := \max_{f \in \mathcal{H}_K} \left\{ \left( \mathbb{E}_{s \sim \mu} [R_\pi V_\theta(s) \cdot f(s)] \right)^2 : \|f\|_{\mathcal{H}_K} \leq 1 \right\}.$$ 

- Solving above optimization problem, we have a new loss for minimizing Bellman residual:

$$L_K(V_\theta) = \|R_\pi V_\theta\|_{K, \mu}^2 := \mathbb{E}_{s, \bar{s} \sim \mu} [K(s, \bar{s}) \cdot R_\pi V(s) \cdot R_\pi V(\bar{s})]$$

where $s, \bar{s}$ are drawn i.i.d from a states distribution $\mu$. 
Advantages of Kernel Loss

\[ L_K(V_\theta) = \| R_\pi V_\theta \|^2_{K,\mu} := \mathbb{E}_{s,\bar{s} \sim \mu} [K(s, \bar{s}) \cdot R_\pi V(s) \cdot R_\pi V(\bar{s})] . \]

- A valid loss! \( L_K(V_\theta) \geq 0 \) for any \( V_\theta \). And \( L_K(V) = 0 \) iff \( V = V^\pi \).
- No \textit{double sampling} problem! (using an ISPD kernel and U-statistics)
- It can be easily estimated and optimized from observed transitions:

\[ \hat{L}_K(V_\theta) := \frac{1}{n(n-1)} \sum_{1 \leq i,j \leq n, i \neq j} k(s_i, s_j) \cdot \hat{R}_\pi V_\theta(s_i) \cdot \hat{R}_\pi V_\theta(s_j) . \]

- We can use standard optimization techniques such as SGD to minimize \( L_K(V_\theta) \).
(a) MDP Example

(b) MSE vs. Iteration

(c) \[ \| w - w^* \| \] vs. Iteration

Figure: Modified example of Tsitsiklis & Van Roy (1997).