Non-parametric TD(0) for policy evaluation

**Objective:** Given a Markov reward process, compute the value function:

\[ V^*(x) = \mathbb{E}\left[ \sum_{k=0}^{\infty} r(x_k) | x_0 = x \right]. \]

**Algorithm:** Sample \((x_n, r(x_n), x'_{n+1})\) from the Markov chain, and update:

\[ V_n = V_{n-1} + \rho_n [r(x_n) + \gamma V_{n+1}(x_{n+1}) - V_n(x_n)] K(x_n, \cdot), \]

where \(K\) is a positive-definite kernel associated with an RKHS \(\mathcal{H}\).

**Generalization of:**

- Tabular setting with \(V(x) = \sum_{y \in \mathcal{X}} \theta^T \phi(y)\) with \(K(x, y) = \phi(x)^T \phi(y)\).

**Challenge:** Proving convergence to \(V^*\)

**Existing results:**

- In tabular setting, a.s. convergence to \(V^*\) if all states are visited often with linear approximation, convergence to a minimizer of the mean-squared projected Bellman error, in general different from \(V^*\).

**Proposed solution:**

- Use a universal kernel as approximator, i.e., such that:

\[ V_n = (1 - \rho_n \lambda) V_{n-1} + \rho_n [r(x_n) + \gamma V_{n+1}(x_{n+1}) - V_n(x_n)] \Phi(x_n). \]

The ODE method

1. Study the mean-path version of the algorithm, in continuous-time:

\[ \frac{dV_t}{dt} = -M V_t + \mathbb{E} \left[ (r(x_t) + \gamma V_{t+1}(x') - V_t(x_t)) \Phi(x_t) \right]. \]

2. Back to the stochastic, discrete-time version, choose the step size properly for convergence, according to Robbins-Monro conditions.

**Exponential convergence to \(V^*_x\)**

The recursion can be written as:

\[ \frac{dV_t}{dt} = (\gamma \Sigma_1 - \Sigma - \lambda) V_t + \Sigma x, \]

where \(\Sigma_1\) and \(\Sigma\) are the first two autocovariance operators:

\[ \Sigma = \mathbb{E}[\Phi(x) \otimes \Phi(x)] \quad \text{and} \quad \Sigma_1 = \mathbb{E}[\Phi(x) \otimes \Phi(x')]. \]

**Key property:**

\[ \|\Sigma_1^{1/2} \Sigma_1^{1/2} - I\|_F^2 \leq 1 \] (Schur complement)

- the ODE has a unique fixed point \(V^*_x \in \mathcal{H}\) defined by:

\[ (\gamma \Sigma_1 - \Sigma - \lambda) V^*_x + \Sigma x = 0 \]

\[ W(f) = \|V^*_x - V^*_x\|_H^2 \leq \frac{\|V^*_x\|_H^2 e^{-2\lambda t}}{O(1/\lambda^2)} \]

**Regularity assumption on \(V^*_x\)**

**Source condition:**

\[ \|\Sigma^{\theta/2} V^*_x\|_H < +\infty \quad \text{for some} \quad \theta \in [-1, 1]. \]

The parameter \(\theta\) quantifies the regularity of \(V^*_x\) with respect to \(\mathcal{H}\):

- \(\theta = -1\) is equivalent to \(V^*_x \in L^2(\mu)\) (always ok if \(r \in L^1\))
- \(\theta = 0\) is equivalent to \(V^*_x \in H^1(\mu)\) (stronger)
- \(\theta \in (-1, 0)\) and \(\theta \in (0, 1)\) are respectively stronger and weaker conditions than \(V^*_x \in H^1(\mu)\).

**Convergence of \(V^*_x \rightarrow V^*_x\)** is faster for larger values of \(\theta\):

\[ \|V^*_x - V^*_x\|_H^2 = O(1/\lambda^{1+\theta}). \]

**Optimal choice** of the regularization \(\lambda\) is a trade-off depending on \(\theta\):

\[ \|V^*_x - V^*_x\|_H^2 = O\left(\frac{e^{-2\lambda t}}{\lambda^2} + O(1/\lambda^{1+\theta})\right). \]

**Convergence rates of TD learning**

We consider two different settings for the sampling of the \((x_n, r(x_n), x'_{n+1})\):

- **i.i.d. sampling** from the stationary distribution of the Markov chain
- **successive samples** from the Markov chain, with exponential mixing (requires additional boundedness assumption, see details in the paper)

**Main theorem:**

- With \(\lambda = \alpha^{-1} - \frac{1}{2}\), constant step size \(\rho = \frac{\log n}{2\lambda}\) and exponential averaging:

\[ \mathbb{E}\left[\|V_n - V^*_x\|_H^2\right] = O\left(\left(\frac{\log n}{n}\right)^2 \frac{1}{\lambda^2}\right). \]

- recovers existing \(1/\sqrt{n}\) rate for \(\hat{\theta} = 0\)
- the rates are adaptive to the regularity of \(V^*_x\) with respect to \(\mathcal{H}\) and can be slower (\(\hat{\theta} < 0\)) or faster (\(\hat{\theta} > 0\)) than \(1/\sqrt{n}\)
- for \(\hat{\theta} = 1\), we only prove asymptotic convergence to \(V^*_x\)

**Numerical experiment**

We use the Sobolev kernels of regularity \(y\) on the 1d torus.

**The source condition** is equivalent to:

\[ \|V^*_x - \sum_{k \neq 0} \|e^{2\pi ikx} \|_H^2 < \infty \]

(decrease rate of Fourier coefficients)

We consider two different settings for the sampling of the \((x_n, r(x_n), x'_{n+1})\):

- **i.i.d. sampling** from the stationary distribution of the Markov chain
- **successive samples** from the Markov chain, with exponential mixing (requires additional boundedness assumption, see details in the paper)

**Main theorem:**

- With \(\lambda = \alpha^{-1} - \frac{1}{2}\), constant step size \(\rho = \frac{\log n}{2\lambda}\) and exponential averaging:

\[ \mathbb{E}\left[\|V_n - V^*_x\|_H^2\right] = O\left(\left(\frac{\log n}{n}\right)^2 \frac{1}{\lambda^2}\right). \]

- recovers existing \(1/\sqrt{n}\) rate for \(\hat{\theta} = 0\)
- the rates are adaptive to the regularity of \(V^*_x\) with respect to \(\mathcal{H}\) and can be slower (\(\hat{\theta} < 0\)) or faster (\(\hat{\theta} > 0\)) than \(1/\sqrt{n}\)
- for \(\hat{\theta} = 1\), we only prove asymptotic convergence to \(V^*_x\)

**Numerical experiment**

We use the Sobolev kernels of regularity \(y\) on the 1d torus.

**The source condition** is equivalent to:

\[ \|V^*_x - \sum_{k \neq 0} \|e^{2\pi ikx} \|_H^2 < \infty \]

(decrease rate of Fourier coefficients)