A Non-asymptotic Analysis of Non-parametric Temporal-Difference Learning

Eloïse Berthier, Ziad Kobeissi, Francis Bach

NEURAL INFORMATION PROCESSING SYSTEMS

Non-parametric TD(0) for policy evaluation

Objective: given a Markov reward process, compute the value function:

$$V^*(x) = \mathbb{E}\Big[\sum_{n=0}^{+\infty} \gamma^n r(x_n) \,\Big|\, x_0 = x\Big].$$

Algorithm: sample $(x_n, r(x_n), x_n')$ from the Markov chain, and update:

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

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

Generalization of:

- tabular setting with $K(x,y) = \mathbf{1}_{x=y}$
- linear approximation $V(x) = \theta^\top \varphi(x)$ with $K(x,y) = \varphi(x)^\top \varphi(y)$.

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 $\overline{\mathcal{H}} = L^2$
- add a regularization λ if $V^* \notin \mathcal{H}$:

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

The ODE method

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

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = -\lambda V_t + \mathbb{E}\Big[\big(r(x) + \gamma V_t(x') - V_t(x)\big)\Phi(x)\Big]$$

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

References

J. N. Tsitsiklis and B. Van Roy (1997). "An analysis of temporal-difference learning with function approximation." IEEE Transactions on Automatic Control.

J. Bhandari, D. Russo and R. Singal (2018). "A finite time analysis of temporal difference learning with linear function approximation." COLT.

A. Dieuleveut and F. Bach (2016). "Nonparametric stochastic approximation with large step-sizes." The Annals of Statistics.

Exponential convergence to V_λ^*

The recursion can be written as:

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = (\gamma \Sigma_1 - \Sigma - \lambda I) V_t + \Sigma r,$$

where Σ and Σ_1 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: $\|\Sigma^{-1/2}\Sigma_1\Sigma^{-1/2}\|_{\mathrm{op}} \leq 1$ (Schur complement)

- the ODE has a unique fixed point $V_\lambda^* \in \mathcal{H}$ defined by:

$$(\gamma\Sigma_1 - \Sigma - \lambda I) V_\lambda^* + \Sigma r = 0$$

- $W(t) = \|V_t - V_\lambda^*\|_{\mathcal{H}}^2$ is a Lyapunov function, so that:

$$\|V_t - V_\lambda^*\|_{\mathcal{H}}^2 \leq \underbrace{\|V_\lambda^*\|_{\mathcal{H}}^2}_{O(1/\lambda^2)} e^{-2\lambda t}$$

Regularity assumption on V^*

Source condition: $\|\Sigma^{-\theta/2} V^*\|_{\mathcal{H}} < +\infty$ for some $\theta \in [-1, 1]$.

The parameter θ quantifies the regularity of V^* with respect to $\mathcal{H}$:

- ★ $\theta = -1$ is equivalent to $V^* \in L^2$ (always ok if $r \in L^2$)
- ★ $\theta = 0$ is equivalent to $V^* \in \mathcal{H}$ (stronger)
- ★ $\theta \in (0, 1]$ and $\theta \in (-1, 0)$ are respectively stronger and weaker conditions than $V^* \in \mathcal{H}$.

Convergence of $V_\lambda^* \xrightarrow[\lambda \to 0]{} V^*$ is faster for larger values of θ :

$$\|V_\lambda^* - V^*\|_{L^2}^2 = O(\lambda^{1+\theta}).$$

$$\begin{array}{ccc} V_t & \xrightarrow[t \to +\infty]{} & V_\lambda^* \\ & & \Big\downarrow {\scriptstyle \lambda \to 0} \\ & & V^* \end{array}$$

Optimal choice of the regularization λ is a trade-off depending on θ:

$$\|V_t - V^*\|_{L^2}^2 = O\Big(\frac{e^{-2\lambda t}}{\lambda^2}\Big) + O(\lambda^{1+\theta})\ .$$

Convergence rates of TD learning

We consider two different settings for the sampling of the $(x_n, r(x_n), x_n')$:

- ***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 $\lambda = n^{-\frac{1+\theta}{2+\theta}}$, constant step size $\rho = \frac{\log n}{2\lambda n}$ and exponential averaging:

$$\mathbb{E}\Big[\|\overline{V}_n - V^*\|_{L^2}^2\Big] = O\Big((\log n)^2 n^{-\frac{1+\theta}{2+\theta}}\Big)\ .$$

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

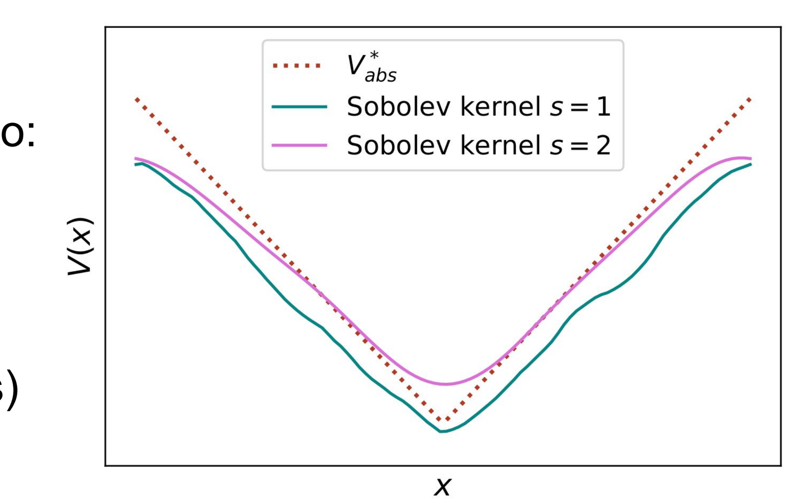
Numerical experiment

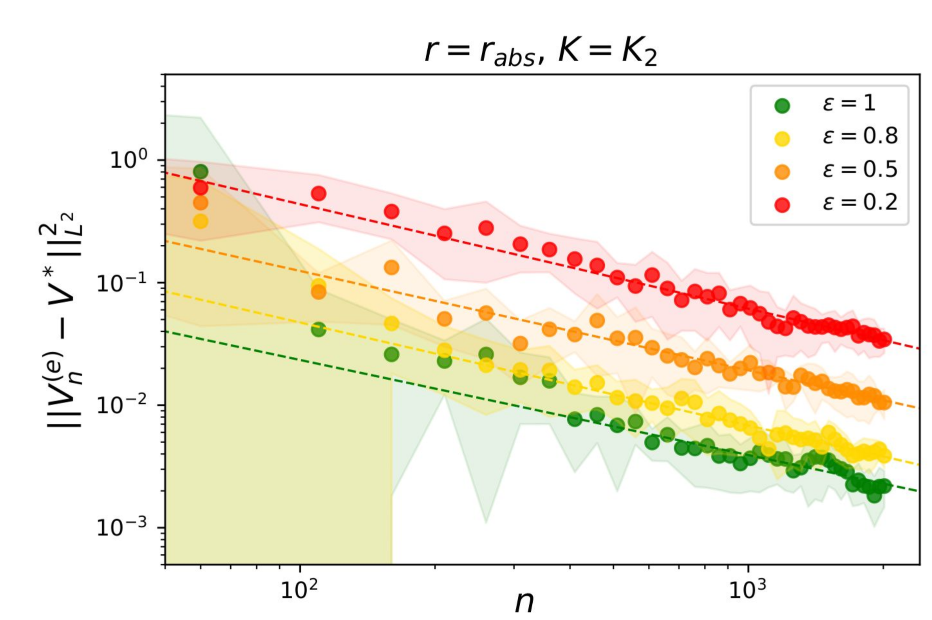
We use the Sobolev kernels of regularity s on the 1d torus.

The **source condition** is equivalent to:

$$|\hat{V}_0^*|^2 + \sum_{\omega \neq 0} |\omega|^{2s(1+\theta)} |\hat{V}_\omega^*|^2 < \infty$$

(decrease rate of Fourier coefficients)





The effect of the mixing parameter $1 - \varepsilon$ is in the constants, not the rate.

Predicted slope: -0.43
Observed slope: -0.58