A Single Stepsize Suffices for Unprojected Linear TD(0)

Wei-Cheng Lee Francesco Orabona

King Abdullah University of Science and Technology

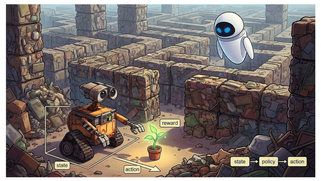
AdaML Adaptive Machine Learning Lab

Abstract

- **A single stepsize.** We run plain, unprojected TD(0) with
$$\eta_t = \frac{1}{c\,\tau_{\text{mix}}\,\phi_\infty^2\sqrt{t+1}\log(t+3)},$$
which is independent of the curvature parameter ω and is used for both convergence regimes.
- **Simultaneous high-probability rates.** For the Polyak–Ruppert average $\bar{\boldsymbol{\theta}}_T$, the TD potential satisfies
$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) = \widetilde{\mathcal{O}}\left(R_{\max}^2 \min\left\{\frac{\tau_{\text{mix}}^2\phi_\infty^4}{\omega T}, \frac{\tau_{\text{mix}}\phi_\infty^2}{\sqrt{T}}\right\}\right).$$
Thus the same run attains both the fast curvature-dependent rate and the curvature-free rate.
- **Implicit stability without projection.** With probability at least $1-\delta$,
$$\sup_{t\geq 0}\|\boldsymbol{\theta}_t\|_2 \leq R_{\max} = \rho(c,\delta)\max\left\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2, \|\boldsymbol{\theta}^*\|_2, \frac{r_\infty}{\phi_\infty}\right\},$$
where ρ is independent of ω. Hence bounded gradients are proved from the TD recursion rather than imposed by a projection radius.

Reinforcement Learning

- state space $\mathcal{S}$,
- action space $\mathcal{A}$,
- transition probability $\boldsymbol{P}(s' \mid s, a)$,
- reward function $r(s, s')$,
- discount factor $\gamma \in [0, 1)$.



A **policy** $\boldsymbol{\mu}$ specifies an action distribution $\boldsymbol{\mu}(a \mid s)$. Starting from an initial state s_0, the system evolves according to
$$a_t \sim \boldsymbol{\mu}(\cdot \mid s_t), \qquad s_{t+1} \sim \boldsymbol{P}(\cdot \mid s_t, a_t),$$
and collects $r_t := r(s_t, s_{t+1})$ along one trajectory. Under irreducibility and aperiodicity, the state chain has transition matrix $\boldsymbol{P}^{\boldsymbol{\mu}}$, where $\boldsymbol{P}^{\boldsymbol{\mu}}(s, s') := \sum_{a\in\mathcal{A}} \boldsymbol{\mu}(a \mid s)\boldsymbol{P}(s' \mid s, a)$, and a unique stationary distribution $\boldsymbol{d}^{\boldsymbol{\mu}}$. Its transition chain $\boldsymbol{Z}_t = (s_t, s_{t+1})$, with transition kernel P_Z and stationary law π_Z, satisfies, for some $\tau_{\text{mix}} \geq 1$, the **geometric mixing property**
$$\sup_z \left\|P_Z^k(z,\cdot) - \pi_Z\right\|_{\text{TV}} \leq 4\cdot 2^{-k/\tau_{\text{mix}}}, \qquad k \geq 0.$$

Temporal Difference Learning for Policy Evaluation

The goal of **policy evaluation** is to estimate
$$\boldsymbol{V}^{\boldsymbol{\mu}}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t \,\middle|\, s_0 = s,\ a_t \sim \boldsymbol{\mu}(\cdot \mid s_t)\right].$$
For a large state space, use the linear approximation
$$\boldsymbol{V}_{\boldsymbol{\theta}}(s) := \boldsymbol{\phi}(s)^\top\boldsymbol{\theta}, \qquad \boldsymbol{\phi}(s) \in \mathbb{R}^d.$$
Let $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S}|\times d}$ be the feature matrix, whose row for state s is $\boldsymbol{\phi}(s)^\top$, so that $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}$. The TD target $\boldsymbol{\theta}^*$ is the unique solution of
$$\boldsymbol{\Phi}\boldsymbol{\theta}^* = \Pi_{\boldsymbol{D}}\,T^{\boldsymbol{\mu}}(\boldsymbol{\Phi}\boldsymbol{\theta}^*), \qquad T^{\boldsymbol{\mu}}\boldsymbol{v} := \boldsymbol{r}^{\boldsymbol{\mu}} + \gamma\boldsymbol{P}^{\boldsymbol{\mu}}\boldsymbol{v}, \qquad \boldsymbol{D} := \operatorname{diag}(\boldsymbol{d}^{\boldsymbol{\mu}}),$$
where $\boldsymbol{r}^{\boldsymbol{\mu}}(s) := \sum_{s'}\boldsymbol{P}^{\boldsymbol{\mu}}(s,s')r(s,s')$, and $\Pi_{\boldsymbol{D}}$ is the $\boldsymbol{D}$-weighted projection onto $\operatorname{span}(\boldsymbol{\Phi})$.
Given $\boldsymbol{Z}_t = (s_t, s_{t+1})$, the unprojected linear TD(0) update is
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{Z}_t) = \boldsymbol{\theta}_t + \eta_t\delta_t\boldsymbol{\phi}(s_t),$$
where $\delta_t = r_t + \gamma\boldsymbol{\phi}(s_{t+1})^\top\boldsymbol{\theta}_t - \boldsymbol{\phi}(s_t)^\top\boldsymbol{\theta}_t$ is the TD error.

Mean TD Dynamics and the Stability Bottleneck

Stationary mean dynamics. For $\boldsymbol{Z} = (s, s') \sim \pi_Z$, define
$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{Z}\sim\pi_Z}[\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{Z})].$$
Then $\boldsymbol{\theta}^*$ is the TD fixed point:
$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) = \boldsymbol{0}, \qquad \langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}), \boldsymbol{\theta}^* - \boldsymbol{\theta}\rangle \geq \omega\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2,$$
where
$$\omega := (1-\gamma)\lambda_{\min}(\boldsymbol{\Phi}^\top\boldsymbol{D}\boldsymbol{\Phi}) > 0.$$
Hence ω quantifies the conditioning of the mean TD dynamics and controls the fast rate.
Why the sample-path analysis is harder than i.i.d. SGD. The iterate $\boldsymbol{\theta}_t$ is a function of the entire past $\boldsymbol{Z}_0, \ldots, \boldsymbol{Z}_{t-1}$, while $\boldsymbol{Z}_t$ is generated from the same Markov chain. Consequently, in general, the current noise $\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{Z}_t) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_t)$ is not conditionally mean-zero with respect to $\mathcal{F}_{t-1} := \sigma(\boldsymbol{Z}_0, \ldots, \boldsymbol{Z}_{t-1})$:
$$\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{Z}_t) \mid \mathcal{F}_{t-1}] \neq \bar{\boldsymbol{g}}(\boldsymbol{\theta}_t).$$
The stability feedback loop. For every $\boldsymbol{\theta}$ and state pair $\boldsymbol{Z}$,
$$\|\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{Z})\|_2 \leq r_\infty\phi_\infty + 2\phi_\infty^2\|\boldsymbol{\theta}\|_2.$$
A loose bound on $\|\boldsymbol{\theta}_t\|_2$ makes the update bound loose, which feeds into $\boldsymbol{\theta}_{t+1}$ and can recursively explode. Projection-based proofs break this loop by imposing a radius.

Main Results

Theorem: Single-stepsize unprojected TD(0)

Assume a finite, irreducible, aperiodic Markov reward process (MRP) with a full-column-rank feature matrix $\boldsymbol{\Phi}$, $\|\boldsymbol{\phi}(s)\|_2 \leq \phi_\infty$, and $|r(s,s')| \leq r_\infty$. For any $\delta \in (0,1)$ and sufficiently large $c = c(\delta)$, run
$$\eta_t = \frac{1}{c\,\tau_{\text{mix}}\,\phi_\infty^2\sqrt{t+1}\log(t+3)}.$$
Let
$$R_{\text{base}} := \max\left\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2, \|\boldsymbol{\theta}^*\|_2, \frac{r_\infty}{\phi_\infty}\right\}, \qquad R_{\max} := \rho R_{\text{base}}, \qquad S_T := \sum_{t=0}^{T-1}\eta_t, \qquad \bar{\boldsymbol{\theta}}_T := \frac{1}{S_T}\sum_{t=0}^{T-1}\eta_t\boldsymbol{\theta}_t.$$
Here $\rho > 2$ is a fixed constant dependent of only c and δ. For the TD potential
$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*) = (1-\gamma)\|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2 + \gamma\|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\text{Dir}}^2,$$
where
$$\|v\|_{\boldsymbol{D}}^2 := v^\top\boldsymbol{D}v, \qquad \|v\|_{\text{Dir}}^2 := \frac{1}{2}\sum_{s,s'}\boldsymbol{d}^{\boldsymbol{\mu}}(s)\boldsymbol{P}^{\boldsymbol{\mu}}(s,s')\big(v(s) - v(s')\big)^2.$$
With probability at least $1-\delta$, simultaneously for every $T \geq 1$,
$$\sup_{t\geq 0}\|\boldsymbol{\theta}_t\|_2 \leq R_{\max},$$
and
$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \leq \widetilde{\mathcal{O}}\left(R_{\max}^2\min\left\{\frac{\tau_{\text{mix}}^2\phi_\infty^4}{\omega T}, \frac{\tau_{\text{mix}}\phi_\infty^2}{\sqrt{T}}\right\}\right).$$

Comparisons

Rate	Paper	Algorithm Design: Extra inputs	Algorithm Design: No projection	Guarantees: Optimal dependence on ω	Guarantees: High probability
$\widetilde{\mathcal{O}}(1/T)$	Bhandari et al. [2018]	ω	✗	✓	✗
	Srikant and Ying [2019]	ω, $\|\boldsymbol{\theta}^*\|_2$	✓	✓	✗
	Patil et al. [2023]	–	✓	✓	✗
	Samsonov et al. [2024]	–	✓	✓	✓
	Mitra [2024]	ω	✓	✓	✗
	Ll et al. [2025]	ω	✓	✓	✗
	Our work	–	✓	✓	✓
$\widetilde{\mathcal{O}}(1/\sqrt{T})$	Bhandari et al. [2018]	–	✗	✓	✗
	Liu and Olshevsky [2021]	–	✗	✓	✗
	Sun et al. [2021]	–	✗	✗	✗
	Our work	–	✓	✓	✓

Proof Idea: A Poissonized Self-Bounding Bootstrap

1. Induct on the past; use $\boldsymbol{\theta}_t - \boldsymbol{\theta}^*$ as the surrogate. Let
$$\boldsymbol{e}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}^*,$$
and assume only
$$\mathsf{IH}_{t-1}: \qquad \max_{0\leq i\leq t-1}\|\boldsymbol{\theta}_i\|_2 \leq R_{\max}.$$
Expanding around the fixed point yields
$$\|\boldsymbol{e}_t\|_2^2 = \|\boldsymbol{e}_0\|_2^2 + \sum_{i=0}^{t-1}\eta_i^2\|\boldsymbol{g}(\boldsymbol{\theta}_i, \boldsymbol{Z}_i)\|_2^2 + 2\sum_{i=0}^{t-1}\eta_i\big\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\big\rangle + 2\sum_{i=0}^{t-1}\eta_i h_i(\boldsymbol{Z}_i)$$
$$\leq \|\boldsymbol{e}_0\|_2^2 + Q_t + 2B_t,$$
$$Q_t := \sum_{i=0}^{t-1}\eta_i^2\|\boldsymbol{g}(\boldsymbol{\theta}_i, \boldsymbol{Z}_i)\|_2^2, \qquad B_t := \sum_{i=0}^{t-1}\eta_i h_i(\boldsymbol{Z}_i),$$
where
$$h_i(z) := \big\langle\boldsymbol{g}(\boldsymbol{\theta}_i, z) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_i),\ \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\big\rangle,$$
and
$$\big\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}^*\big\rangle \leq 0.$$
Every iterate-dependent term on the right uses only $\boldsymbol{\theta}_i$ with $i \leq t-1$. Under IH_{t-1},
$$\|\boldsymbol{g}(\boldsymbol{\theta}_i, z)\|_2 \leq 3\phi_\infty^2 R_{\max}, \qquad \|\boldsymbol{e}_i\|_2 \leq 2R_{\max}, \qquad \sum_{i\geq 0}\eta_i^2 < \infty.$$
Thus, if B_t is uniformly controlled, the bound on the new iterate follows from $\|\boldsymbol{\theta}_t\|_2 \leq \|\boldsymbol{e}_t\|_2 + \|\boldsymbol{\theta}^*\|_2$.
2. Poissonize the only hard term B_t. Freeze the past iterate $\boldsymbol{\theta}_i$. Since $\mathbb{E}_{\boldsymbol{Z}\sim\pi_Z}[h_i(\boldsymbol{Z})] = 0$, define
$$(P_Z u)(z) := \mathbb{E}[u(\boldsymbol{Z}_1) \mid \boldsymbol{Z}_0 = z], \qquad u_i - P_Z u_i = h_i, \qquad u_i(z) = \sum_{\ell=0}^{\infty}(P_Z^\ell h_i)(z).$$
Geometric mixing gives
$$\|u_i\|_\infty \leq 16\tau_{\text{mix}}\|h_i\|_\infty.$$
The Poisson identity decomposes one Markov-noise term as
$$h_i(\boldsymbol{Z}_i) = \underbrace{u_i(\boldsymbol{Z}_{i+1}) - (P_Z u_i)(\boldsymbol{Z}_i)}_{\Delta M_{i+1}} + \underbrace{u_i(\boldsymbol{Z}_i) - u_i(\boldsymbol{Z}_{i+1})}_{\text{boundary remainder}}.$$
Moreover, IH_{t-1} implies
$$\|h_i\|_\infty \lesssim \phi_\infty^2 R_{\max}^2 \quad\Longrightarrow\quad \|u_i\|_\infty \lesssim \tau_{\text{mix}}\phi_\infty^2 R_{\max}^2.$$
Hence
$$|\eta_i\Delta M_{i+1}| \lesssim \frac{R_{\max}^2}{c\sqrt{i+1}\log(i+3)}, \qquad \sum_{i\geq 0}\|\eta_i\Delta M_{i+1}\|_\infty^2 < \infty.$$
With $M_t := \sum_{i=0}^{t-1}\eta_i\Delta M_{i+1}$, Pinelis' inequality therefore controls $\sup_t|M_t|$ with high probability.
3. Summation by parts makes the remainder past-only. Write $B_t = M_t + R_t$. An equivalent Abel decomposition is
$$R_t = \eta_0 u_0(\boldsymbol{Z}_0) - \eta_{t-1}u_{t-1}(\boldsymbol{Z}_t) - \sum_{i=1}^{t-1}(\eta_{i-1} - \eta_i)u_{i-1}(\boldsymbol{Z}_i)$$
$$+\sum_{i=1}^{t-1}\eta_i\big(u_i - u_{i-1}\big)(\boldsymbol{Z}_i).$$
The first three terms use bounded u_i, while the last term uses local Lipschitz continuity together with
$$\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1} = \eta_{i-1}\boldsymbol{g}(\boldsymbol{\theta}_{i-1}, \boldsymbol{Z}_{i-1}),$$
which gives
$$\|u_i - u_{i-1}\|_\infty \lesssim \tau_{\text{mix}}\phi_\infty^4 R_{\max}^2\eta_{i-1}.$$
Thus, $R_t = \mathcal{O}(\tau_{\text{mix}}R_{\max}^2/c + R_{\max}^2/c^2)$, which is small for large c.
4. Close the self-bounding bootstrap. Combining the quadratic, martingale, and remainder bounds gives, with high probability, for some finite constants $A_1(\delta) > 0$ and $A_2 > 0$,
$$\|\boldsymbol{\theta}_t\|_2^2 \leq \left(4 + \frac{A_1(\delta)\rho^2}{c} + \frac{A_2\rho^2}{c^2}\right)R_{\text{base}}^2 \leq \rho^2 R_{\text{base}}^2 = R_{\max}^2,$$
for a sufficiently large stepsize multiplier c. Hence $\mathsf{IH}_{t-1} \Rightarrow \mathsf{IH}_t$.

weicheng.lee@kaust.edu.sa