A Single Stepsize Suffices for Unprojected Linear TD(0)

Simultaneous Robust and Fast Rates via Polyak–Ruppert Averaging

Wei-Cheng Lee Francesco Orabona

King Abdullah University of Science and Technology (KAUST)

LTSS Workshop 2026, Copenhagen

Setting I: from RL successes to policy evaluation

Reinforcement learning learns by interaction.

- **State s_t:** what the agent sees now.
- **Action a_t:** what the agent does next.
- **Policy μ:** a rule mapping states to actions.
- **Reward r_t:** the immediate feedback received after taking an action.

AlphaGo's success

- Once a policy is fixed, we want to know quickly: *how good is this policy from each state?*

Policy evaluation

For a fixed policy μ, estimate

$$V^{\mu}(s) = \mathbb{E}\left[\sum_{k\geq 0} \gamma^k r_k \mid s_0 = s\right]$$

the expected discounted return from state s with discount factor $\gamma \in (0, 1]$.

Setting: Linear TD(0) with Markovian sampling

In large state spaces, exact tabular values are too expensive. Use linear approximation:

$$V_{\boldsymbol{\theta}}(s) = \phi(s)^\top \boldsymbol{\theta}, \qquad \phi(s) \in \mathbb{R}^d, \qquad \|\phi(s)\| \le \phi_\infty.$$

Unprojected linear TD(0) updates by the TD direction

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \boldsymbol{g}(\boldsymbol{\theta}_t, Z_t),$$

$$\boldsymbol{g}(\boldsymbol{\theta}, Z_t) = \big(r_t + \gamma \phi(s_{t+1})^\top \boldsymbol{\theta} - \phi(s_t)^\top \boldsymbol{\theta}\big)\phi(s_t).$$

The stationary TD direction and TD fixed point are

$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}) := \mathbb{E}_{Z \sim \pi_Z}[\boldsymbol{g}(\boldsymbol{\theta}, Z)], \qquad \boldsymbol{\theta}^* : \quad \bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) = \mathbf{0}.$$

Performance measure: the TD potential

For $v \in \mathbb{R}^{|\mathcal{S}|}$, define the stationary value norm and Dirichlet seminorm

$$\|v\|_{\boldsymbol{D}}^2 := \sum_s \pi(s) v(s)^2, \qquad \|v\|_{\mathrm{Dir}}^2 := \frac{1}{2} \sum_{s,s'} \pi(s) P^{\mu}(s,s') \big(v(s) - v(s')\big)^2.$$

We measure error by the TD potential

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*) = (1-\gamma) \, \|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2 + \gamma \, \|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\mathrm{Dir}}^2 .$$

Intuition

The potential rewards two things at once:

- $\|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}$: value accuracy under the stationary distribution
- $\|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\mathrm{Dir}}$: consistency along Markov transitions.

Performance measure: two finite-time regimes

The potential has curvature around the TD fixed point $\boldsymbol{\theta}^*$:

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*) \geq \omega \left\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\right\|^2, \qquad \omega := (1-\gamma)\lambda_{\min}(\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi}) > 0.$$

Two regimes of interest

Robust rate	**Fast rate**
works even when curvature is tiny	exploits curvature when it helps
$\widetilde{\mathcal{O}}(1/\sqrt{T})$	$\widetilde{\mathcal{O}}(1/(\omega T))$

Robust when ω is small; fast when ω is large.

Why stability is hard I: Markovian sampling

The sample $Z_t = (s_t, s_{t+1})$ is generated by the same trajectory that produced $\boldsymbol{\theta}_t$. Thus, in general,

$$\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_t, Z_t) \mid \mathcal{F}_{t-1}] \neq \bar{\boldsymbol{g}}(\boldsymbol{\theta}_t).$$

Why this matters

- For i.i.d. data, the update noise is centered after conditioning on the past.
- For Markov data, the current sample still remembers the previous state, so the noise has a **bias** that must be controlled.

Why stability is hard II: self-amplification and prior fixes

Self-amplification.

$$\|\boldsymbol{g}(\boldsymbol{\theta}_t, Z_t)\| \leq r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}_t\| .$$

A loose bound on $\|\boldsymbol{\theta}_t\|$ makes the bound on the update $\|\boldsymbol{g}(\boldsymbol{\theta}_t, Z_t)\|$ loose, which feeds back into $\boldsymbol{\theta}_{t+1}$.

What prior work uses

- **Projection:** (Bhandari et al. 2018; Liu and Olshevsky 2021). Stable by design, but changes TD and requires a projection radius.
- **Curvature / contractive stability:** (Srikant and Ying 2019; Patil et al. 2023; Samsonov et al. 2024; Mitra 2025; Chandak and Borkar 2025; Durmus et al. 2025; Li et al. 2026). Projection-free or lightly modified, but rate, burn-in, or tuning can depend on curvature.

Question: one plain TD stepsize for both regimes?

Question

Can unprojected TD(0) achieve high-probability stability and simultaneous robust and fast rates without knowledge of ω?

Best of both worlds target

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \leq \widetilde{\mathcal{O}}\left(\min\left\{\frac{1}{\sqrt{T}}, \frac{1}{\omega T}\right\}\right)$$

- If ω is small: still keep a $1/\sqrt{T}$ guarantee.
- If ω is large: automatically obtain the fast $1/(\omega T)$ rate.

Main result I: one ω-agnostic stepsize gives stability

Use a single decaying stepsize

$$\eta_t = \frac{1}{c\,\tau_{\text{mix}}\,\phi_\infty^2}\,\frac{1}{\sqrt{t+1}\log(t+3)}.$$

Let

$$R_{\text{base}} = \max\left\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, \frac{r_\infty}{\phi_\infty}\right\}.$$

Mixing assumption

Assume the Markov chain $Z_t = (s_t, s_{t+1})$ is irreducible and aperiodic with stationary law π_Z. Then,

$$\max_{z\in\mathcal{Z}} \left\|P_Z^k(z,\cdot) - \pi_Z\right\|_{\text{TV}} \le 4\cdot 2^{-k/\tau_{\text{mix}}}, \qquad k \ge 0.$$

Implicit boundedness

For sufficiently large c, with probability at least $1-\delta$,

$$\sup_{t\ge 0}\|\boldsymbol{\theta}_t\| \le R_{\max}.$$

No projection is applied to the iterates.

Main result II: simultaneous robust and fast rates

Define the weighted Polyak–Ruppert average

$$S_T = \sum_{t=1}^{T} \eta_{t-1}, \qquad \bar{\boldsymbol{\theta}}_T = \frac{1}{S_T} \sum_{t=1}^{T} \eta_{t-1} \boldsymbol{\theta}_{t-1}.$$

Main theorem

For a sufficiently large c depending on δ, set $R_{\max} = \rho(c, \delta) R_{\text{base}}$. With probability at least $1 - \delta$, for all $T \geq 1$,

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \leq \widetilde{\mathcal{O}}\left(R_{\max}^2 \min\left\{ \frac{\tau_{\text{mix}}^2 \phi_\infty^4}{\omega T}, \frac{\tau_{\text{mix}} \phi_\infty^2}{\sqrt{T}} \right\} \right).$$

Stability proof I: make the loop self-bounding

Induct on the past event

$$\mathsf{IH}_{t-1} := \left\{ \max_{0 \le i < t} \|\boldsymbol{\theta}_i\| \le R_{\max} \right\}.$$

Since $\|\boldsymbol{\theta}_t\| \le \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|$, it is enough to control $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|$.

Expanding the squared distance gives

$$\begin{aligned} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 = \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \sum_{i=0}^{t-1} \eta_i^2 \|\boldsymbol{g}_i\|^2 \\ + 2 \sum_{i=0}^{t-1} \eta_i \langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}^* \rangle + 2 \sum_{i=0}^{t-1} \eta_i h_i(Z_i), \end{aligned}$$

where

$$h_i(z) := \langle \boldsymbol{g}(\boldsymbol{\theta}_i, z) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}^* \rangle .$$

Stability proof I: isolate the only hard term

The mean drift helps:

$$\langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}^* \rangle = -\big[f(\boldsymbol{\theta}_i) - f(\boldsymbol{\theta}^*)\big] \leq 0.$$

Under IH_{t-1}, the quadratic term is controlled because $\sum_i \eta_i^2 < \infty$.

Only hard term: Markov bias

$$B_t := \sum_{i=0}^{t-1} \eta_i h_i(Z_i).$$

This is not a standard martingale sum because Z_i is Markovian and h_i changes with $\boldsymbol{\theta}_i$.

Stability proof II: Poissonize the Markov bias

Fix $\boldsymbol{\theta}_i$. Since $\mathbb{E}_{Z\sim\pi_Z}[h_i(Z)] = 0$, solve the Poisson equation

$$u_i - P_Z u_i = h_i, \qquad u_i(z) = \sum_{\ell=0}^{\infty} (P_Z^\ell h_i)(z), \qquad \|u_i\|_\infty \le 16\tau_{\mathrm{mix}} \|h_i\|_\infty .$$

Under IH_{t-1},

$$\|h_i\|_\infty \lesssim \phi_\infty^2 R_{\max}^2, \qquad \|u_i\|_\infty \lesssim \tau_{\mathrm{mix}} \phi_\infty^2 R_{\max}^2.$$

Each Markov-noise term decomposes as

$$h_i(Z_i) = \underbrace{u_i(Z_{i+1}) - (P_Z u_i)(Z_i)}_{\text{martingale difference}} + \underbrace{u_i(Z_i) - u_i(Z_{i+1})}_{\text{telescoping remainder}}.$$

Two key implications

- The first term is controlled by Pinelis' inequality from bounded differences.
- The second term is rewritten by Abel summation; the moving u_i's are controlled by Lipschitz continuity and $\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1} = \eta_{i-1}\boldsymbol{g}_{i-1}$.

Together, these bounds make B_t small enough to feed back into the bootstrap.

Stability proof III: close the bootstrap

The previous bounds give

$$\|\boldsymbol{\theta}_t\|^2 \le \left(4 + \frac{A_1(\delta)\rho^2}{c} + \frac{A_2\rho^2}{c^2}\right) R_{\text{base}}^2.$$

To close the induction, it is enough that

$$4 + \frac{A_1(\delta)\rho^2}{c} + \frac{A_2\rho^2}{c^2} \le \rho^2.$$

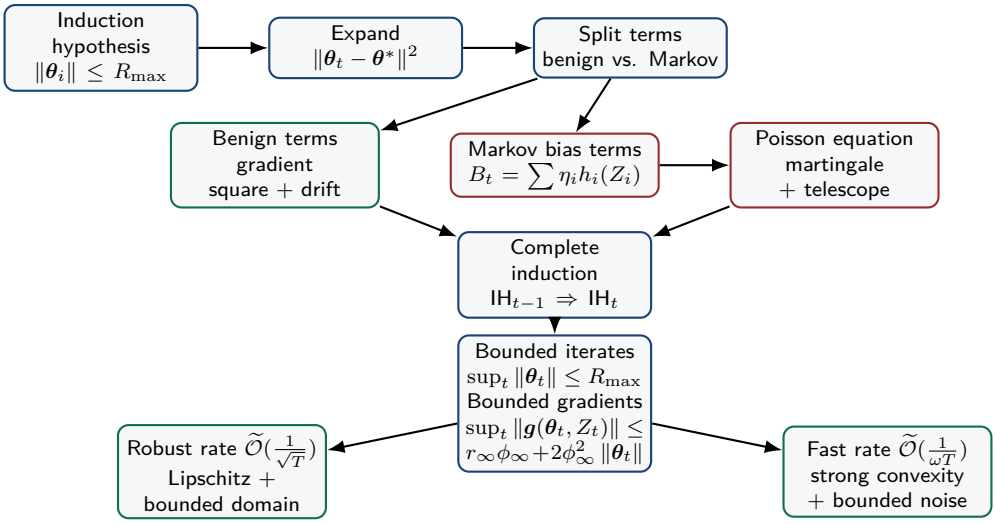
Many pairs (c, ρ) close the induction

- As $c \to \infty$, the admissible radius factor satisfies $\rho \downarrow 2$.
- There is a minimum lower bound $c > c_{\min}(\delta)$.
- Larger c: smaller stepsize, smaller radius bound.
- Smaller c above the threshold: larger stepsize, larger radius bound.

Bootstrap conclusion

$$\mathsf{IH}_{t-1} \Longrightarrow \mathsf{IH}_t, \qquad \sup_{t \ge 0} \|\boldsymbol{\theta}_t\| \le \rho R_{\text{base}}.$$

Proof road map: stability first, rates second



Conclusion

Main message

A single ω-agnostic stepsize suffices for projection-free linear TD(0) under Markov sampling.

- **Algorithmic takeaway:** no projection radius and no tuning by the curvature/eigenvalue parameter ω.
- **Statistical takeaway:** the averaged iterate gets simultaneous high-probability rates

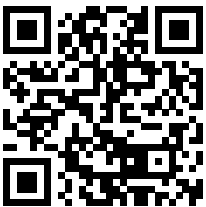
$$\min\left\{\widetilde{\mathcal{O}}\left(\frac{1}{\omega T}\right),\ \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right)\right\}.$$

- **Proof takeaway:** self-bounding induction gives implicit boundedness; a Poisson-equation decomposition controls Markov bias.

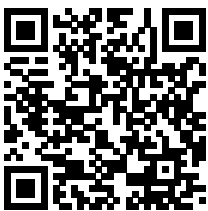
Open problem: Can we still get a reasonable best-of-both-worlds rate using a τ_{mix}-agnostic stepsize? (Currently, $\eta_t = \frac{1}{c\,\tau_{\mathrm{mix}}\,\phi_\infty^2}\,\frac{1}{\sqrt{t+1}\log(t+3)}$, but τ_{mix} is usuallyunknown in practice.)

Thank you! Any questions?

Slides



Paper



Personal website

References

Bhandari, J., D. Russo, and R. Singal (2018). "A finite time analysis of temporal difference learning with linear function approximation". In: *Conference on Learning Theory*. PMLR, pp. 1691–1692.

Chandak, S. and V. S. Borkar (2025). "A Concentration Bound for TD(0) with Function Approximation". In: *Stochastic Systems*.

Durmus, A. et al. (2025). "Finite-Time High-Probability Bounds for Polyak–Ruppert Averaged Iterates of Linear Stochastic Approximation". In: *Mathematics of Operations Research* 50.2, pp. 935–964.

Li, Y. et al. (2026). "Towards Parameter-Free Temporal Difference Learning". In: *arXiv preprint arXiv:2603.02577*.

Liu, R. and A. Olshevsky (2021). "Temporal difference learning as gradient splitting". In: *International Conference on Machine Learning*. PMLR, pp. 6905–6913.

Mitra, A. (2025). "A Simple Finite-Time Analysis of TD Learning With Linear Function Approximation". In: *IEEE Transactions on Automatic Control* 70.2, pp. 1388–1394.

Patil, G. et al. (2023). "Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 5438–5448.

Samsonov, S. et al. (2024). "Improved high-probability bounds for the temporal difference learning algorithm via exponential stability". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR, pp. 4511–4547.

Srikant, R. and L. Ying (2019). "Finite-time error bounds for linear stochastic approximation and TD learning". In: *Conference on Learning Theory*. PMLR, pp. 2803–2830.