TABULAR TD DESCENDS THE ABSOLUTE BELLMAN ERROR

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal-difference (TD) learning is the standard method for policy evaluation in reinforcement learning, but its expected update is not, in general, the negative gradient of any global objective, even in the tabular setting. This has limited the use of optimization ideas to interpret TD and to streamline its analysis. Classical contraction arguments establish convergence without requiring a strict decrease in a global error measure after each one-state update. In this paper, we show that expected tabular TD descends the stationary-weighted absolute Bellman error. As our main contribution, we prove that an expected update of any single state strictly decreases this measure whenever the updated state's Bellman error is nonzero, for any step size in $(0, 1]$. This monotonicity yields convergence under standard cumulative-step-size conditions. We then use the same descent property to derive a finite-sample bound on the expected absolute value error of sampled TD under independent sampling. We also give a local-optimization interpretation of this descent property: each expected tabular TD update can be written as the solution of a trust-region subproblem for a linear model of the absolute Bellman error. Although we focus on the tabular case, this perspective may offer a starting point for new algorithms, including ones that use function approximation.

1 INTRODUCTION

Value estimation is a central problem in reinforcement learning. For a fixed policy π, the goal is to estimate the expected discounted sum of future rewards from each state. Temporal difference (TD) learning (Sutton 1988) is the standard algorithm for this purpose (e.g., see Sutton & Barto 2018). By updating estimates based on other learned estimates (a process known as bootstrapping) TD can learn online and efficiently. However, despite its widespread use and empirical effectiveness, the behavior of TD remains difficult to interpret from a classical optimization perspective.

It is well-established that TD learning is not, in general, a gradient-descent algorithm. Even in the tabular setting, its expected update need not correspond to the negative gradient of any global objective function (Barnard 1993). Consequently, while gradient-descent methods are easily understood as moving downhill on an error surface, TD has traditionally been viewed as a root-finding procedure for the Bellman equation. Classical analyses of discounted tabular TD and value iteration use the contraction of the Bellman operator in the ℓ_∞ norm (Bertsekas & Tsitsiklis 1996). For expected one-state updates, however, the maximum value error need not decrease strictly after every nontrivial update.

In this paper, we show that expected tabular TD descends the stationary-weighted absolute Bellman error. In the tabular setting, let $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ denote a vector of state-value estimates, and let $\delta_{\boldsymbol{v}}(s)$ denote the Bellman error (a.k.a. expected TD error) at state s. We study the weighted ℓ_1 Bellman error

$$BE_1(\boldsymbol{v}) \stackrel{\text{def}}{=} \sum_s d_\pi(s)\ |\delta_{\boldsymbol{v}}(s)|\,, \tag{1}$$

where d_π is the strictly positive stationary distribution of the Markov chain induced by π. This choice fixes the objective to the policy being evaluated; it does not depend on how updates are sampled.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

As our main contribution, we establish a strict descent property. We prove that an expected asynchronous update of any single state strictly decreases this global $BE_1(\boldsymbol{v})$ measure for any $\alpha \in (0, 1]$, provided the updated state has a nonzero Bellman error. The result holds regardless of which state is selected. This makes BE_1 a strict Lyapunov function for expected asynchronous tabular TD, and it yields a direct convergence proof under the standard condition that each state receives unbounded total step size. It is worth noting that this monotonicity does not hold for the squared Bellman error (ℓ_2), which an expected one-state TD update can increase; and the strict monotonicity does not hold for the maximum value error (ℓ_∞) which need not decrease after a single update.

We also give a local optimization interpretation of this descent property. The expected TD update is the unique minimizer of a supporting affine model of $BE_1(v)$ over a state-dependent polyhedral trust region (Corollary 2). This complements the descent analysis by connecting TD to optimization methods that minimize local models of a fixed objective, without requiring the update to follow its gradient.

The descent property is also a tool for analysis. In Section 5, we study sampled TD, in which states are selected independently from an arbitrary full-support distribution and updated using independent reward and transition samples under the target policy. We show that the same decrease in ℓ_1 Bellman error that proves convergence also bounds the expected stationary-weighted absolute error of the value estimates after any number of updates. With uniform state sampling and bounded TD-error variance, the resulting sample complexity has the same dependence on the number of states, the effective horizon, and the accuracy as the best known guarantees for tabular TD, although those guarantees control the stronger criterion of maximum error.

The contribution of this paper is therefore theoretical, and its focus is entirely on the tabular prediction setting. Our aim is to show that tabular TD is more closely tied to optimization than its lack of a gradient structure suggests, and that this tie runs through the ℓ_1 norm of Bellman error: Every nontrivial expected one-state update decreases the absolute Bellman error, whereas the squared Bellman error can increase and the maximum value error can remain unchanged. Moreover, the trust-region view may offer a perspective on how this geometry might guide the design of new TD methods in future works, including methods with function approximation.

2 Background and Notation

We consider the tabular prediction problem for a fixed policy π in the finite Markov reward process induced by π. Let $\mathcal{S}$ denote the state space. Under π, the agent moves from state s to a successor state s' with probability $p(s'|s)$ and receives a random reward R_{t+1} with conditional mean

$$r(s) \stackrel{\text{def}}{=} \mathbb{E}[R_{t+1} \mid S_t = s, \pi]. \tag{2}$$

We assume a discounted setting with discount factor $\gamma \in [0, 1)$.

The value of a state is the expected discounted return. The value function v_π is the unique solution to the Bellman equation

$$v_\pi(s) = r(s) + \gamma \sum_{s'} p(s'|s)\, v_\pi(s'), \quad \forall s \in \mathcal{S}. \tag{3}$$

Throughout, $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ denotes an arbitrary vector of state-value estimates, with components $v(s)$ for $s \in \mathcal{S}$. For any $\boldsymbol{v}$, the Bellman error (a.k.a. expected TD error) at state s is

$$\delta_{\boldsymbol{v}}(s) \stackrel{\text{def}}{=} r(s) + \gamma \sum_{s'} p(s'|s)\, v(s') - v(s), \quad \forall s \in \mathcal{S}. \tag{4}$$

In particular, $\delta_{\boldsymbol{v}}(s) = 0$ for all $s \in \mathcal{S}$ if and only if $\boldsymbol{v} = \boldsymbol{v}_\pi$.

We measure Bellman error using the stationary distribution of the policy being evaluated. We assume the Markov chain induced by π is ergodic so that it has a unique stationary distribution d_π that is strictly positive on $\mathcal{S}$ and satisfies

$$d_\pi(s) = \sum_{s'} d_\pi(s')\, p(s|s'), \quad \forall s \in \mathcal{S}. \tag{5}$$

We first study expected tabular TD updates, before turning to sampled updates in Section 5. We distinguish a one-state update from its average over a state-sampling distribution.

Asynchronous expected TD. Given a selected state $s \in \mathcal{S}$, the asynchronous expected TD update with step size $\alpha > 0$ changes only that state's estimate:

$$v'(s) = v(s) + \alpha\, \delta_{\boldsymbol{v}}(s), \quad \text{and} \quad v'(s') = v(s') \text{ for all } s' \neq s. \tag{6}$$

This is the update used in our monotonicity and convergence arguments; its sampled version is analyzed in Section 5. Here, "expected" refers to averaging over the reward and successor state conditional on the selected state, with $\boldsymbol{v}$ held fixed; and asynchronous means that one state is updated at a time using current value estimates.

Mean TD update. Let μ be an arbitrary state-sampling distribution. Averaging the one-state expected update over the selected state gives

$$v'(s) = v(s) + \alpha\, \mu(s)\, \delta_{\boldsymbol{v}}(s), \quad \forall s \in \mathcal{S}. \tag{7}$$

The distribution μ need not equal d_π.

For $x \in \mathbb{R}$, let $\operatorname{sign}(x)$ denote the usual sign when $x \neq 0$, and let $\operatorname{sign}(0)$ be any value in $[-1, 1]$, corresponding to a subgradient of $|x|$ at 0.

3 Descent under Asynchronous Updates

In this section we show that BE_1 decreases under expected asynchronous TD updates and that this monotonicity is enough to prove convergence. An update reduces the absolute Bellman error at the selected state but may increase it at predecessor states. Our first theorem shows that stationary weighting makes the decrease exceed these possible increases whenever the update is nonzero. The result holds for every selected state and every step size $\alpha \in (0, 1]$.

Theorem 1 (Strict Descent for Expected Asynchronous TD)**.** *Consider a state $s \in \mathcal{S}$ and the asynchronous expected TD update* (6) *applied to s with step size $\alpha \in (0, 1]$. Then*

$$BE_1(\boldsymbol{v}') \leq BE_1(\boldsymbol{v}) - \alpha(1-\gamma) d_\pi(s)\, |\delta_{\boldsymbol{v}}(s)|. \tag{8}$$

The decrease is strict whenever $\delta_{\boldsymbol{v}}(s) \neq 0$.

Proof. Only the component $v(s)$ changes, by the amount $\alpha\delta_{\boldsymbol{v}}(s)$. This change affects the Bellman error at state s and Bellman error at other states that have a nonzero transition probability into s. A direct expansion of (4) gives, for any $s' \in \mathcal{S}$,

$$\delta_{\boldsymbol{v}'}(s') = \delta_{\boldsymbol{v}}(s') + \alpha\gamma\, p(s|s')\, \delta_{\boldsymbol{v}}(s) - \alpha\delta_{\boldsymbol{v}}(s)\, \mathbb{1}(s' = s), \tag{9}$$

where $\mathbb{1}(s' = s)$ is 1 if $s' = s$ and 0 otherwise.

First consider $s' = s$. From (9),

$$\delta_{\boldsymbol{v}'}(s) = \delta_{\boldsymbol{v}}(s)\big(1 - \alpha(1 - \gamma\, p(s|s))\big). \tag{10}$$

Because $\alpha \in (0, 1]$ and $\gamma\, p(s|s) \leq \gamma < 1$, the factor $1 - \alpha(1 - \gamma\, p(s|s))$ is nonnegative. Therefore the update does not reverse the sign of $\delta_{\boldsymbol{v}}(s)$, and

$$|\delta_{\boldsymbol{v}'}(s)| - |\delta_{\boldsymbol{v}}(s)| = -\alpha\big(1 - \gamma\, p(s|s)\big)\, |\delta_{\boldsymbol{v}}(s)|. \tag{11}$$

Next, consider any $s' \neq s$. Then (9) becomes

$$\delta_{\boldsymbol{v}'}(s') = \delta_{\boldsymbol{v}}(s') + \alpha\gamma\, p(s|s')\, \delta_{\boldsymbol{v}}(s), \tag{12}$$

and by the triangle inequality,

$$|\delta_{\boldsymbol{v}'}(s')| - |\delta_{\boldsymbol{v}}(s')| \leq \alpha\gamma\, p(s|s')\, |\delta_{\boldsymbol{v}}(s)|. \tag{13}$$

We now use (11) and (13) in the definition of BE_1:

$$\begin{aligned} BE_1(\boldsymbol{v}') - BE_1(\boldsymbol{v}) &= \sum_{s'} d_\pi(s')\big(|\delta_{\boldsymbol{v}'}(s')| - |\delta_{\boldsymbol{v}}(s')|\big) \\ &\leq -d_\pi(s)\alpha\big(1 - \gamma\, p(s|s)\big)|\delta_{\boldsymbol{v}}(s)| + \sum_{s' \neq s} d_\pi(s')\, \alpha\gamma\, p(s|s')\, |\delta_{\boldsymbol{v}}(s)|. \end{aligned} \tag{14}$$

Factoring out $\alpha|\delta_{\boldsymbol{v}}(s)|$ and using $\sum_{s'} d_\pi(s')p(s|s') = d_\pi(s)$ gives

$$\begin{aligned} BE_1(\boldsymbol{v}') - BE_1(\boldsymbol{v}) &\leq \alpha|\delta_{\boldsymbol{v}}(s)| \left(-d_\pi(s) + \gamma \sum_{s'} d_\pi(s')p(s|s') \right) \\ &= -\alpha(1-\gamma) d_\pi(s)\, |\delta_{\boldsymbol{v}}(s)|. \end{aligned} \tag{15}$$

The right-hand side is strictly negative when $\delta_{\boldsymbol{v}}(s) \neq 0$, completing the proof. □

Theorem 1 shows that every nontrivial expected one-state update produces definite progress measured by BE_1. This descent is enough to prove convergence for expected asynchronous TD when each state receives unbounded total step size.

Corollary 1 (Convergence of Expected Asynchronous Tabular TD)**.** *Consider an expected asynchronous tabular TD process where, at each time step t, an arbitrary state $S_t \in \mathcal{S}$ is selected and updated via $v_{t+1}(S_t) = v_t(S_t) + \alpha_t \delta_{\boldsymbol{v}_t}(S_t)$, with $v_{t+1}(s') = v_t(s')$ for all $s' \neq S_t$. Assume that $0 < \alpha_t \leq 1$ and every state is updated with unbounded total step size:*

$$\sum_t \alpha_t \, \mathbb{1}(S_t = s) = \infty, \quad \textit{for all } s \in \mathcal{S}. \tag{16}$$

Then, the sequence of value estimates $\{\boldsymbol{v}_t\}$ converges to the true value function $\boldsymbol{v}_\pi$.

Proof. Let $\delta_t(s) \stackrel{\text{def}}{=} \delta_{\boldsymbol{v}_t}(s)$. By Theorem 1,

$$BE_1(\boldsymbol{v}_{t+1}) \leq BE_1(\boldsymbol{v}_t) - \alpha_t(1-\gamma) d_\pi(S_t) |\delta_t(S_t)|.$$

Because $BE_1(\boldsymbol{v}_t) \geq 0$ and $\min_s d_\pi(s) > 0$, summing over t implies

$$\sum_{t=0}^{\infty} \alpha_t |\delta_t(S_t)| < \infty. \tag{17}$$

The one-state update (6) and the summability bound (17) give

$$\sum_{t=0}^{\infty} \|\boldsymbol{v}_{t+1} - \boldsymbol{v}_t\|_1 = \sum_{t=0}^{\infty} \alpha_t |\delta_t(S_t)| < \infty.$$

Hence $\{\boldsymbol{v}_t\}$ is a Cauchy sequence and converges to some $\bar{\boldsymbol{v}}$. By (4), the Bellman error is continuous in $\boldsymbol{v}$, so $\delta_t(s) \to \bar{\delta}(s) := \delta_{\bar{\boldsymbol{v}}}(s)$ for every $s \in \mathcal{S}$.

If $\bar{\delta}(s) \neq 0$ for some s, then for some $\epsilon > 0$ and all large t, $|\delta_t(s)| \geq \epsilon$. In this case,

$$\sum_t \alpha_t |\delta_t(S_t)| \geq \sum_t \alpha_t \mathbb{1}(S_t = s) |\delta_t(s)| \geq \epsilon \sum_t \alpha_t \mathbb{1}(S_t = s) = \infty,$$

which contradicts (17). So $\bar{\delta} = \mathbf{0}$, and $\delta_{\boldsymbol{v}_t} \to \mathbf{0}$. Finally, $\delta_{\boldsymbol{v}} = \mathbf{0}$ is the Bellman equation for policy π, whose unique solution is $\boldsymbol{v}_\pi$ (e.g., see Chapter 3 of Sutton & Barto 2018). Therefore $\boldsymbol{v}_t \to \boldsymbol{v}_\pi$. □

The strict descent property in Theorem 1 is specific to the ℓ_1 Bellman-error measure, and need not hold for the squared Bellman error,

$$BE_2(\boldsymbol{v}) \stackrel{\text{def}}{=} \sum_s d_\pi(s)\, \delta_{\boldsymbol{v}}(s)^2, \tag{18}$$

which can increase after an expected one-state TD update. Appendix A gives a two-state example. It also shows an increase under the mean update (7), even when $\mu = d_\pi$. Thus these updates decrease the absolute Bellman error while increasing the squared Bellman error.

4 A Local Optimization Interpretation of Expected TD

The strict descent established in the previous section also admits a local optimization interpretation. In this section, we show that the mean expected TD update (7) can be represented exactly as the solution to a local optimization problem for the BE_1 objective.

To construct a local linear approximation around the current estimate vector $\boldsymbol{v}$, consider a candidate state-value vector $\boldsymbol{v}'$. Using the sign function $\sigma(s) = \operatorname{sign}(\delta_{\boldsymbol{v}}(s))$ introduced in Section 2, we define

$$L_{\boldsymbol{v}}(\boldsymbol{v}') \stackrel{\text{def}}{=} \sum_s d_\pi(s)\, \sigma(s)\, \delta_{\boldsymbol{v}'}(s). \tag{19}$$

Because $\sigma(s)$ is a subgradient of $|\delta_{\boldsymbol{v}}(s)|$, this is a supporting affine model of BE_1:

$$L_{\boldsymbol{v}}(\boldsymbol{v}) = BE_1(\boldsymbol{v}), \qquad L_{\boldsymbol{v}}(\boldsymbol{v}') \le BE_1(\boldsymbol{v}') \quad \text{for all } \boldsymbol{v}'. \tag{20}$$

Since $L_{\boldsymbol{v}}$ is affine, minimizing it is meaningful only over a bounded region. Exactness of the model does not by itself supply one: $L_{\boldsymbol{v}}$ agrees with BE_1 on the polyhedral cell of value functions whose Bellman errors carry the signs σ, and this cell contains the true values $\boldsymbol{v}_\pi$; so the region on which the model is exact returns the solution itself rather than a local step. What the region should control is instead the movement of a single step. Appendix B shows that BE_1 decreases if $\boldsymbol{v}'$ moves each state in the direction of its own Bellman error and $|v'(s) - v(s)| \le |\delta_{\boldsymbol{v}}(s)|$ for every s. Writing the movement allowed at s as a fraction $\alpha\mu(s) \le 1$ of that cap gives the polyhedral trust region

$$\Omega_{\alpha,\mu}(\boldsymbol{v}) \stackrel{\text{def}}{=} \left\{ \boldsymbol{v}' \in \mathbb{R}^{|\mathcal{S}|} : |v'(s) - v(s)| \le \alpha\mu(s)|\delta_{\boldsymbol{v}}(s)| \quad \text{for all } s \in \mathcal{S} \right\}. \tag{21}$$

The following corollary gives the resulting interpretation of the expected TD update.

Corollary 2 (Trust-Region Interpretation of Expected Tabular TD)**.** *Fix a state-sampling distribution μ and a step size $\alpha > 0$. For any current state-value vector $\boldsymbol{v}$, the unique minimizer of the supporting affine model $L_{\boldsymbol{v}}(\boldsymbol{v}')$ over $\Omega_{\alpha,\mu}(\boldsymbol{v})$ is*

$$v^*(s) = v(s) + \alpha\mu(s)\delta_{\boldsymbol{v}}(s), \qquad \forall s \in \mathcal{S}, \tag{22}$$

which is exactly the mean expected TD update (7)*.*

Proof. Since $\delta_{\boldsymbol{v}'}$ is affine in $\boldsymbol{v}'$, so is $L_{\boldsymbol{v}}$, and we write $L_{\boldsymbol{v}}(\boldsymbol{v}') = c + \sum_s g(s)v'(s)$.

Fix a state s with $\delta_{\boldsymbol{v}}(s) \neq 0$, and let $\boldsymbol{v}^{(s)}$ be the expected one-state update at s with step size one, which differs from $\boldsymbol{v}$ only in its sth coordinate, by $\delta_{\boldsymbol{v}}(s)$. Then $g(s)\delta_{\boldsymbol{v}}(s) = L_{\boldsymbol{v}}(\boldsymbol{v}^{(s)}) - L_{\boldsymbol{v}}(\boldsymbol{v})$, and

$$g(s)\delta_{\boldsymbol{v}}(s) \;=\; L_{\boldsymbol{v}}(\boldsymbol{v}^{(s)}) - L_{\boldsymbol{v}}(\boldsymbol{v}) \;\le\; BE_1(\boldsymbol{v}^{(s)}) - BE_1(\boldsymbol{v}) \;\le\; -(1-\gamma)d_\pi(s)|\delta_{\boldsymbol{v}}(s)|,$$

where the first inequality is from (20), and the second inequality is Theorem 1 with $\alpha = 1$. The right-hand side is negative, so $\operatorname{sign}(g(s)) = -\sigma(s)$.

Because $\Omega_{\alpha,\mu}(\boldsymbol{v})$ is an axis-parallel hyperrectangle, the minimization of $L_{\boldsymbol{v}}(\boldsymbol{v}')$ decouples completely across dimensions (i.e., states). To minimize the independent linear term $v'(s)\, g(s)$, we must push $v'(s)$ to the boundary of its constraint interval in the direction opposite to the sign of $g(s)$:

$$\begin{aligned} v^*(s) &= v(s) - \alpha\mu(s)|\delta_{\boldsymbol{v}}(s)| \operatorname{sign}(g(s)) \\ &= v(s) + \alpha\mu(s)|\delta_{\boldsymbol{v}}(s)|\sigma(s) \\ &= v(s) + \alpha\mu(s)\delta_{\boldsymbol{v}}(s). \end{aligned}$$

If $\delta_{\boldsymbol{v}}(s) = 0$ or $\mu(s) = 0$, the corresponding interval collapses to the single point $v'(s) = v(s)$, which gives the same expression. Therefore $\boldsymbol{v}^*$ is exactly the mean expected TD update (7), and the minimizer is unique. □

Corollary 2 gives a precise local optimization interpretation of expected tabular TD. Although TD does not, in general, follow the negative gradient of a global objective, its expected update exactly minimizes a supporting affine model of the absolute Bellman error over a residual-scaled polyhedral trust region. The ℓ_1 geometry determines the direction of the update through the signs of the Bellman errors, while the trust-region widths recover the magnitude of the TD step.

We view this result primarily as an interpretation of the descent geometry, rather than as an alternative derivation of TD. Nevertheless, the trust-region perspective may be useful for designing new algorithmic variants of TD with function approximation, in future works.

5 Finite-Sample Guarantees from the Descent Inequality

We now return to the descent property of Section 3 and show that the same inequality can be used to derive finite-sample bounds for sampled TD.

In this section, we use a state-sampling distribution μ that need not equal d_π. At each step t, we draw $s_t \sim \mu$, with $\mu_{\min} = \min_s \mu(s) > 0$. Given s_t, we draw $m \geq 1$ independent reward and successor pairs $(r_{t,i}, s'_{t,i})$ under the target policy π. The batches at different steps are independent. Starting from initial values $\boldsymbol{v}_0$, we use a constant step size $0 < \alpha \leq 1$ and update the value at state s_t by

$$v_{t+1}(s_t) = v_t(s_t) + \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \big(r_{t,i} + \gamma v_t(s'_{t,i})\big) - v_t(s_t) \right], \tag{23}$$

without changing values at other states (i.e., $v_{t+1}(s) = v_t(s)$ for $s \neq s_t$). This is the sampled version of (6), using independent draws from a distribution rather than a single Markov trajectory.

We measure sampling noise by the variance of the TD error at the true values:

$$\omega^2(s) = \mathbb{E}\Big[\big(r_{t,1} + \gamma v_\pi(s'_{t,1}) - v_\pi(s)\big)^2 \mid s_t = s \Big], \qquad \bar{\omega}^{\,2} = \sum_s d_\pi(s)\omega^2(s). \tag{24}$$

By the Bellman equation (3), the TD error inside the square has conditional mean zero. We assume $\omega^2(s) < \infty$ for every state. This assumption concerns the process at the true values, not the value estimates. For a vector $\boldsymbol{x}$, we write $\|\boldsymbol{x}\|_{1,d_\pi} = \sum_s d_\pi(s)|x(s)|$.

Theorem 2 (Finite-sample bound)**.** *Under the sampling and variance assumptions above, for every integer $t \geq 0$,*

$$\mathbb{E}\|\boldsymbol{v}_t - \boldsymbol{v}_\pi\|_{1,d_\pi} \leq \big(1 - \alpha(1-\gamma)\mu_{\min}\big)^t \, \frac{BE_1(\boldsymbol{v}_0)}{1-\gamma} + \bar{\omega}\sqrt{\frac{\alpha}{(1-\gamma)m}}. \tag{25}$$

The proof is given in Appendix C. The two terms in (25) separate the effect of initialization from that of sampling noise. Choosing the step size and number of updates to make each term at most $\varepsilon/2$ gives the following corollary, also proved in Appendix C.

Corollary 3 (Sample complexity)**.** *For any $\varepsilon > 0$ and integer $m \geq 1$, choose*

$$\alpha = \min\left\{1, \frac{m(1-\gamma)\varepsilon^2}{4\,\bar{\omega}^{\,2}}\right\}, \tag{26}$$

where we take $\alpha = 1$ when $\bar{\omega} = 0$. Then $\mathbb{E}\|\boldsymbol{v}_T - \boldsymbol{v}_\pi\|_{1,d_\pi} \leq \varepsilon$, for

$$T = \left\lceil \frac{1}{(1-\gamma)\mu_{\min}} \, \max\left(1, \frac{4\,\bar{\omega}^{\,2}}{(1-\gamma)m\varepsilon^2}\right) \, \log\max\left(1, \frac{2BE_1(\boldsymbol{v}_0)}{(1-\gamma)\varepsilon}\right) \right\rceil. \tag{27}$$

Also, $\mathbb{E}\big[BE_1(\boldsymbol{v}_T)\big] \leq (1+\gamma)\varepsilon$.

The proof separates the final value error into a propagated initialization error and an accumulated sampling error. The descent inequality (8) controls both terms. Repeated expected updates contract the contribution of the initialization. For the sampling contribution, independence across batches and centering of the TD noise allow us to sum the propagated noise variances; (8) then bounds their total contribution in terms of $\bar{\omega}^{\,2}$. Thus the same BE_1 descent inequality used to establish convergence also controls the two sources of error in the finite-sample analysis.

The aim here is to demonstrate the use of L_1 Bellman-error descent in finite-sample analysis, rather than to improve existing optimal sample-complexity bounds. For uniform μ and rewards in $[0,1]$, a law-of-total-variance argument gives $\bar{\omega}^{\,2} \leq (1+\gamma)/(4(1-\gamma)) = O((1-\gamma)^{-1})$ (e.g., see Azar et al., 2013); hence the variance-dependent term of the total sample count mT from (27) has order $\widetilde{O}(|\mathcal{S}|/((1-\gamma)^3\varepsilon^2))$. This term has the same dependence on $|\mathcal{S}|$, $(1-\gamma)^{-1}$, and ε^{-1} as the minimax-optimal sample complexity established for synchronous tabular TD in (Li et al. 2024), although their guarantee controls the stronger criterion of maximum value error with high probability. Bhandari et al. (2021) instead obtain mean-squared error bounds for linear TD through a gradient-like inequality. Our result bounds expected absolute error directly from the BE_1 descent already used to prove convergence. Because these results use different error criteria, existing maximum-error lower bounds (Pananjady & Wainwright 2021) do not directly establish tightness for our stationary-weighted absolute-error guarantee. Section 6 gives a more in-depth comparison.

6 Related Works

The relationship between temporal-difference learning and optimization has been a subject of study since the algorithm's inception. It was recognized early that the expected tabular TD update is not, in general, the negative gradient of the mean squared Bellman error, nor of any other global objective function (Barnard 1993). This is fundamentally due to the asymmetry of the TD-update Jacobian: the Jacobian of the expected update is typically asymmetric, whereas a gradient field must have a symmetric Jacobian. The absence of a gradient representation, however, does not rule out descent on a fixed objective.

Convergence proofs generally follow two main conceptual paths. The first relies on the Bellman operator being a contraction in the ℓ_∞ norm, a property used to establish convergence of asynchronous updates through stochastic approximation (e.g., see Jaakkola, Jordan & Singh 1994; Tsitsiklis 1994; Bertsekas & Tsitsiklis 1996). The second studies stability of the mean ODE associated with TD (e.g., see Tsitsiklis & Van Roy 1997; Borkar & Meyn 2000). These approaches establish asymptotic convergence under their respective assumptions, but they do not single out a simple scalar that strictly decreases after each nontrivial expected one-state update.

One line of work obtains an optimization interpretation by changing the update. Residual-gradient algorithms perform true gradient descent on the squared Bellman error (Baird 1995), but often learn more slowly than TD in tabular and linear settings (e.g., see Sutton et al. 2009). Gradient-TD methods such as GTD2 and TDC target the mean squared *projected* Bellman error (MSPBE), using auxiliary weights to avoid double sampling (Sutton et al. 2009; Maei 2011). These methods guarantee convergence under standard assumptions for off-policy training with linear function approximation, and proximal-gradient variants have also been developed (Liu et al. 2018). In the tabular setting, MSPBE reduces to mean squared Bellman error. Absolute Bellman-error objectives have also been studied: Patterson et al. (2023) derive gradient methods for absolute and Huber Bellman errors through saddlepoint formulations. Our goal here is different: we do not propose a new objective to obtain a new algorithm, but instead interpret the existing expected tabular TD update as a local optimization method.

The geometry of value prediction has also been explored through other optimization frameworks. Linear programming characterizes the value function through a linear objective and Bellman constraints (e.g., see Manne 1960; Puterman 1994). Approximate formulations use state-relevance weights and yield ℓ_1-type performance guarantees (e.g., see Schweitzer & Seidmann 1985; De Farias & Van Roy 2003). These formulations do not directly explain the step-by-step behavior of standard TD.

A closer line of work gives optimization interpretations of standard TD without changing its update. Ollivier (2018) shows that the expected on-policy TD update is a gradient-descent direction under reversibility, even with nonlinear function approximation. Liu & Olshevsky (2021) relate the mean update in on-policy linear TD to a fixed quadratic through gradient splitting, without requiring reversibility. Asadi et al. (2023) view TD as an iterative optimization procedure in which the function being minimized changes with the iterate, and use this perspective to explain convergence and divergence in linear TD.

Our results focus on the progress made by each individual expected one-state update. The stationary-weighted ℓ_1 Bellman error strictly decreases after an expected update of any state with nonzero Bellman error, for every step size in $(0, 1]$. The guarantee holds before averaging over the selected state and does not require reversibility. The objective is fixed, and the corresponding squared Bellman error can increase under the same update. This descent yields a convergence proof and a local optimization interpretation: the mean TD update exactly minimizes a supporting affine model of the absolute Bellman error over a residual-scaled trust region.

The same descent property also leads to finite-sample guarantees. Bhandari et al. (2021) give finite-sample bounds for TD with linear function approximation. Their independent-sampling analysis includes the unprojected final iterate and controls squared parameter error through a gradient-like inequality. With tabular features and $\mu = d_\pi$, their stated constant-step bound (Theorem 2(b)) gives a variance-dependent sample count of order $\widetilde{O}(\bar{\omega}^{\,2}/((1-\gamma)^2\mu_{\min}^2\varepsilon^2))$ for expected squared Euclidean value error at most ε^2. Corollary 3 has one rather than two inverse factors of $\mu_{\min}$, but controls

the weaker stationary-weighted first-moment criterion. Our proof uses cumulative Bellman-error descent and is restricted to the tabular case, whereas their analysis covers general linear features.

Li et al. (2024) show that synchronous tabular TD, which samples every state at each iteration, achieves maximum value error at most ε with high probability using $\widetilde{O}(|\mathcal{S}|/((1-\gamma)^3\varepsilon^2))$ transitions for rewards in $[0,1]$. Their analysis attains the minimax benchmark established by Pananjady & Wainwright (2021). Under uniform state sampling and the same reward bound, the variance-dependent term in our transition count has the same order, but controls the weaker expected stationary-weighted absolute error. We sample one state per update and allow $\mu \neq d_\pi$. Our proof uses cumulative BE_1 descent to control both initialization and sampling error, connecting the progress of individual expected updates to the finite-sample behavior of sampled TD.

7 Discussion

At first encounter, TD may not seem like a proper optimization method. Even in the tabular setting, its expected update is not, in general, the negative gradient of any global objective function (Barnard 1993). Thus TD can appear to sit somewhat outside the usual optimization picture of moving downhill on a fixed error surface. In this paper we showed that this impression is not accurate. In particular, expected tabular TD moves downhill on a fixed global measure of absolute Bellman error under every nontrivial single-state update, and has a local optimization interpretation.

We identified a simple global progress measure, BE_1, that strictly decreases after an expected update of any state whose Bellman error is nonzero. The result holds for every update schedule and every step size in $(0,1]$, and yields a direct convergence argument under the standard cumulative-step-size condition. This provides a view of asynchronous TD that is complementary to classical contraction arguments: rather than controlling the maximum value error, the analysis assigns a definite decrease in absolute Bellman error to every nontrivial expected one-state update.

This descent property highlights an important difference between ℓ_1 and ℓ_2 Bellman-error measures. With a squared loss (BE_2), the changes induced at other states interact with the magnitudes of their existing Bellman errors, and these terms can outweigh the improvement at the updated state, and make the global objective increase. Moreover, the same descent inequality is useful beyond the deterministic setting. When accumulated over time, it controls both the propagated initialization error and the sampling error in our finite-sample analysis of sampled TD.

The trust-region result provides a complementary local optimization interpretation of tabular TD. The expected TD update is exactly the minimizer of a supporting affine model of BE_1 over a residual-scaled polyhedral trust region. We view this result primarily as an interpretation rather than as the source of the descent theorem. This trust-region view suggests potential design of modified TD updates in more general settings, to be pursued in future works.

Our analysis is restricted to tabular prediction for a fixed policy. The descent and trust-region results concern expected updates, while the finite-sample guarantees cover independent generative-model sampling. We do not address sampling along a single Markov trajectory, control, average-reward formulations, or function approximation. In the linear or nonlinear approximation setting, the Bellman error at one state cannot generally be adjusted independently of others. The statewise argument used here therefore does not extend directly. Understanding whether an analogous absolute-residual descent property, or a useful relaxation of the associated trust-region interpretation, survives under function approximation is an interesting direction for future work.

The main message is that the lack of a global gradient representation does not mean that TD lacks a global notion of progress. In the tabular setting, the stationary-weighted absolute Bellman error provides such a measure: every nontrivial expected single-state update decreases it, its cumulative decrease yields convergence and finite-sample control, and its local geometry admits an optimization interpretation. We hope that this viewpoint will be useful both for understanding TD and for future design of algorithms beyond the tabular setting.

AI USE STATEMENT

In this work, we used generative AI tools for writing improvement and editing, including proofreading and editing mathematical proofs, and literature review. Beyond writing-improvement and editing, generative AI was used to provide critical ingredients for proving the finite sample bound, in particular the formulation and proof of Lemma 2 in Appendix C and interactive development of the proof of the finite sample bound in Appendix C.4. We have not used generative AI tools to propose or refine hypotheses, mathematical claims or theorems, critical proof ingredients of proofs other than those mentioned above, design or provide feedback on research methodology or experiments, help develop theoretical models or conceptual frameworks, or interpret results. Other potential uses of generative AI tools are not applicable to this work; examples include cleaning and reformatting datasets, coding assistance, supporting qualitative and thematic data analysis, generating synthetic data sets, and assisting with translation. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Asadi, K., Sabach, S., Liu, Y., Gottesman, O., & Fakoor, R. (2023). TD convergence: An optimization perspective. *Advances in Neural Information Processing Systems*, *36*, pp. 49169–49186.

Azar, M. G., Munos, R., & Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, *91*(3), 325–349.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 30–37.

Barnard, E. (1993). Temporal-difference methods and Markov models. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(2), 357–365.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Bhandari, J., Russo, D., & Singal, R. (2021). A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, *69*(3), 950–973.

Borkar, V. S., & Meyn, S. P. (2000). The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, *38*(2), 447–469.

De Farias, D. P., & Van Roy, B. (2003). The linear programming approach to approximate dynamic programming. *Operations Research*, *51*(6), 850–865.

Jaakkola, T., Jordan, M. I., & Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, *6*(6), 1185–1201.

Li, G., Cai, C., Chen, Y., Wei, Y., & Chi, Y. (2024). Is Q-learning minimax optimal? A tight sample complexity analysis. *Operations Research*, *72*(1), 222–236.

Liu, B., Gemp, I., Ghavamzadeh, M., Liu, J., Mahadevan, S., & Petrik, M. (2018). Proximal gradient temporal difference learning: Stable reinforcement learning with polynomial sample complexity. *Journal of Artificial Intelligence Research*, *63*, 461–494.

Liu, R., & Olshevsky, A. (2021). Temporal difference learning as gradient splitting. In *Proceedings of the 38th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, *139*, pp. 6905–6913.

Maei, H. R. (2011). *Gradient temporal-difference learning algorithms* (Ph.D. dissertation). University of Alberta.

Manne, A. S. (1960). Linear programming and sequential decisions. *Management Science*, *6*(3), 259–267.

Ollivier, Y. (2018). Approximate temporal difference learning is a gradient descent for reversible policies. *arXiv preprint arXiv:1805.00869*.

Pananjady, A., & Wainwright, M. J. (2021). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, *67*(1), 566–585.

Patterson, A., Liao, V., & White, M. (2023). Robust losses for learning value functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 6157–6167.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Schweitzer, P. J., & Seidmann, A. (1985). Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, *110*(2), 568–582.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., & Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 993–1000.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, *16*(3), 185–202.

Tsitsiklis, J. N., & Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, *42*(5), 674–690.

APPENDIX

A INCREASE OF SQUARED BELLMAN ERROR UNDER TD UPDATES

In Section 3, we established that expected tabular TD strictly and monotonically decreases the weighted ℓ_1 Bellman error, $BE_1(\boldsymbol{v})$. Here, we demonstrate that the corresponding mean squared Bellman error (18) can strictly increase under the exact same updates.

Consider a reward-free environment with two states, $\mathcal{S} = \{s_1, s_2\}$. The transition probabilities are $p(s_2|s_1) = 1$ and $p(s_2|s_2) = 0.9$ (which implies $p(s_1|s_2) = 0.1$). The rewards are identically zero, $r(s_1) = r(s_2) = 0$, so the true value function is $\boldsymbol{v}_\pi = [0, 0]^\top$. We set the discount factor to $\gamma = 0.9$.

The on-policy stationary distribution d_π satisfies $d_\pi(s_1) = 0.1 d_\pi(s_2)$, giving $d_\pi(s_1) = 1/11$ and $d_\pi(s_2) = 10/11$.

Suppose our current value estimates are $v(s_1) = 4$ and $v(s_2) = 3$. The Bellman errors are:

$$\delta_{\boldsymbol{v}}(s_1) = r(s_1) + \gamma v(s_2) - v(s_1) = 0 + 0.9(3) - 4 = -1.3,$$
$$\delta_{\boldsymbol{v}}(s_2) = r(s_2) + \gamma(0.1 v(s_1) + 0.9 v(s_2)) - v(s_2) = 0 + 0.9(0.4 + 2.7) - 3 = -0.21.$$

The initial squared objective is exactly:

$$BE_2(\boldsymbol{v}) = \frac{1}{11}(-1.3)^2 + \frac{10}{11}(-0.21)^2 = \frac{1.69 + 0.441}{11} = \frac{2.131}{11} \approx 0.1937.$$

Increase under Asynchronous Updates. Consider an asynchronous expected TD update applied only to state s_2, using a step size of $\alpha = 1$. The new value for s_2 becomes $v'(s_2) = 3 + (-0.21) = 2.79$, while $v'(s_1)$ remains 4.

Because state s_2 transitions back to s_1, the update to s_2 shifts the bootstrap target for s_1:

$$\delta_{\boldsymbol{v}'}(s_1) = 0 + 0.9(2.79) - 4 = 2.511 - 4 = -1.489,$$
$$\delta_{\boldsymbol{v}'}(s_2) = 0 + 0.9(0.1(4) + 0.9(2.79)) - 2.79 = 0.9(2.911) - 2.79 = -0.1701.$$

The new squared objective is exactly:

$$BE_2(\boldsymbol{v}') = \frac{1}{11}(-1.489)^2 + \frac{10}{11}(-0.1701)^2 = \frac{2.217121 + 0.2893401}{11} = \frac{2.5064611}{11} \approx 0.2279.$$

The BE_2 objective has strictly increased (it grew from 0.1937 to 0.2279). Note that, as Theorem 1 guarantees, the ℓ_1 objective strictly decreased for this exact same update.

Increase under Synchronous On-Policy Updates. The BE_2 objective can also increase under synchronous expected-updates, even when strictly sampling from the on-policy distribution ($\mu = d_\pi$). Using Equation 7 with $\alpha = 1$, both values are updated proportionally to their stationary probabilities:

$$v'(s_1) = 4 + \frac{1}{11}(-1.3) = \frac{44 - 1.3}{11} = \frac{42.7}{11},$$
$$v'(s_2) = 3 + \frac{10}{11}(-0.21) = \frac{33 - 2.1}{11} = \frac{30.9}{11}.$$

Computing the new errors using these exact fractions yields:

$$\delta_{\boldsymbol{v}'}(s_1) = 0.9\left(\frac{30.9}{11}\right) - \frac{42.7}{11} = \frac{27.81 - 42.7}{11} = -\frac{14.89}{11},$$
$$\delta_{\boldsymbol{v}'}(s_2) = 0.9\left(0.1\left(\frac{42.7}{11}\right) + 0.9\left(\frac{30.9}{11}\right)\right) - \frac{30.9}{11} = 0.9\left(\frac{32.08}{11}\right) - \frac{30.9}{11} = -\frac{2.028}{11}.$$

The new squared objective is exactly:

$$BE_2(\boldsymbol{v}') = \frac{1}{11}\left(-\frac{14.89}{11}\right)^2 + \frac{10}{11}\left(-\frac{2.028}{11}\right)^2 = \frac{221.7121 + 10(4.112784)}{11^3} \approx 0.1975.$$

Therefore, the squared error has strictly increased from 0.1937 to 0.1975. This shows that squared Bellman error is not a Lyapunov function for the expected TD update, affecting both asynchronous and strictly on-policy synchronous updates.

B The Statewise Cap of the Trust Region

This appendix supports the choice of the trust region (21) in Section 4. It shows that a step that moves each state in the direction of its own Bellman error decreases BE_1 as long as no state moves by more than its own Bellman error, and that this cap cannot be relaxed. The expected TD updates of Corollary 2 are the steps of this form with change magnitude $w(s) = \alpha\mu(s)|\delta_{\boldsymbol{v}}(s)|$.

Notation. We use bold lower-case symbols for column vectors, upper-case symbols for matrices, and lower-case symbols for scalars. For a vector $\boldsymbol{x}$, $x(s)$ denotes its component at state s and $|\boldsymbol{x}|$ denotes componentwise absolute value. Vector and matrix inequalities are componentwise. The sign function is also applied componentwise. We use $\mathbf{1}_s$ for the vector with a one at state s and zeros elsewhere, and I for the identity matrix.

Let $\boldsymbol{r}$ and $\boldsymbol{d}_\pi$ collect the mean rewards $r(s)$ and stationary probabilities $d_\pi(s)$, respectively. We define the transition matrix P by

$$P(s, s') = p(s' \mid s), \qquad s, s' \in \mathcal{S}.$$

For every value vector $\boldsymbol{v}$, we use $\boldsymbol{\delta}_{\boldsymbol{v}}$ for the vector of expected TD errors $\delta_{\boldsymbol{v}}(s)$ defined in (4). Writing that definition in vector form gives

$$\boldsymbol{\delta}_{\boldsymbol{v}} = \boldsymbol{r} + \gamma P\boldsymbol{v} - \boldsymbol{v}. \tag{B.1}$$

In this notation, the stationary-weighted absolute Bellman error defined in (1) is $BE_1(\boldsymbol{v}) = \boldsymbol{d}_\pi^\top|\boldsymbol{\delta}_{\boldsymbol{v}}|$.

The following proposition provides the main result of this appendix.

Proposition 1. *Let $\boldsymbol{w}$ satisfy $0 \le w(s) \le |\delta_{\boldsymbol{v}}(s)|$ for all s, and let $\boldsymbol{v}' = \boldsymbol{v} + \boldsymbol{w}\,\mathrm{sign}(\boldsymbol{\delta}_{\boldsymbol{v}})$. Then*

$$BE_1(\boldsymbol{v}') \;\le\; BE_1(\boldsymbol{v}) - (1-\gamma)\sum_s d_\pi(s)\, w(s).$$

Proof. The proof is conceptually similar to the proof of Theorem 1. Here we provide a more compact proof. Let D be the diagonal matrix with entries $w(s)/|\delta_{\boldsymbol{v}}(s)|$ at the states with $\delta_{\boldsymbol{v}}(s) \neq 0$ and zero elsewhere; its entries lie in $[0, 1]$. Then, for any $s \in \mathcal{S}$,

$$v'(s) = v(s) + w(s)\,\mathrm{sign}(\delta_{\boldsymbol{v}}(s)) = v(s) + \big(w(s)/|\delta_{\boldsymbol{v}}(s)|\big)\,\delta_{\boldsymbol{v}}(s) = v(s) + D_{ss}\delta_{\boldsymbol{v}}(s).$$

Therefore, $\boldsymbol{v}' = \boldsymbol{v} + D\boldsymbol{\delta}_{\boldsymbol{v}}$. Subtracting (B.1) at $\boldsymbol{v}$ from the same identity at $\boldsymbol{v}'$ gives

$$\boldsymbol{\delta}_{\boldsymbol{v}'} = \boldsymbol{\delta}_{\boldsymbol{v}} - (I - \gamma P)(\boldsymbol{v}' - \boldsymbol{v}) = \boldsymbol{\delta}_{\boldsymbol{v}} - (I - \gamma P)D\boldsymbol{\delta}_{\boldsymbol{v}} = \big(I - (I - \gamma P)D\big)\boldsymbol{\delta}_{\boldsymbol{v}}.$$

The sth column of $I - (I - \gamma P)D$ is a nonnegative vector, being $(1 - D_{ss})\mathbf{1}_s + \gamma D_{ss} P\mathbf{1}_s$ with $D_{ss} \in [0, 1]$, $P \geq 0$. Taking absolute values componentwise and using $\boldsymbol{d}_\pi \geq 0$,

$$BE_1(\boldsymbol{v}') \;=\; \boldsymbol{d}_\pi^\top |\boldsymbol{\delta}_{\boldsymbol{v}'}| \;=\; \boldsymbol{d}_\pi^\top \big|\big(I - (I - \gamma P)D\big)\boldsymbol{\delta}_{\boldsymbol{v}}\big| \;\leq\; \boldsymbol{d}_\pi^\top \big(I - (I - \gamma P)D\big)|\boldsymbol{\delta}_{\boldsymbol{v}}|.$$

Since $d_\pi^\top (I - \gamma P) = (1 - \gamma) d_\pi^\top$, the right-hand side equals $BE_1(\boldsymbol{v}) - (1 - \gamma)\sum_s d_\pi(s) D_{ss} |\delta_{\boldsymbol{v}}(s)|$. Proposition 1 then follows by $D_{ss}|\delta_{\boldsymbol{v}}(s)| = w(s)$. □

The cap in Proposition 1 is sharp for the stated decrease bound. Even an arbitrarily small multiplicative enlargement can invalidate that bound. To see this, fix $\gamma \in [0, 1)$ and consider the two-state chain

$$P = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}, \qquad \boldsymbol{d}_\pi = \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}, \qquad \boldsymbol{r} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The chain is irreducible and aperiodic, with stationary distribution $\boldsymbol{d}_\pi$. Start from $\boldsymbol{v} = (0, 0)^\top$. Then $\boldsymbol{\delta}_{\boldsymbol{v}} = (1, 0)^\top$ and $BE_1(\boldsymbol{v}) = 1/3$.

Let $\varepsilon' > 0$ and update only the first state, using $\boldsymbol{w} = (1 + \varepsilon', 0)^\top$. This exceeds the cap at that state by ε' and gives $\boldsymbol{v}' = (1 + \varepsilon', 0)^\top$. By (B.1),

$$\boldsymbol{\delta}_{\boldsymbol{v}'} = \begin{pmatrix} -\varepsilon' \\ \gamma(1 + \varepsilon')/2 \end{pmatrix}.$$

The Bellman error at the updated state has crossed zero. The excess movement therefore increases its absolute error rather than reducing it further. Taking the stationary-weighted sum of absolute errors gives

$$BE_1(\boldsymbol{v}') = \frac{\varepsilon' + \gamma(1 + \varepsilon')}{3}.$$

In contrast, the right-hand side of the bound in Proposition 1 is

$$BE_1(\boldsymbol{v}) - (1 - \gamma)\sum_s d_\pi(s) w(s) = \frac{\gamma(1 + \varepsilon') - \varepsilon'}{3}.$$

The actual Bellman error exceeds this bound by $2\varepsilon'/3$. Thus the stated bound fails for every $\varepsilon' > 0$, however small.

Failure of this bound does not by itself mean that BE_1 increases. In this example,

$$BE_1(\boldsymbol{v}') - BE_1(\boldsymbol{v}) = \frac{(1 + \gamma)\varepsilon' - (1 - \gamma)}{3}.$$

For a fixed discount, sufficiently small excess steps still decrease the objective. However, for any $\varepsilon' > 0$, choosing $\gamma = 1/(1 + \varepsilon')$ gives

$$BE_1(\boldsymbol{v}') = \frac{1 + \varepsilon'}{3} > \frac{1}{3} = BE_1(\boldsymbol{v}).$$

Consequently, no fixed multiplicative enlargement of the cap guarantees descent uniformly over all discounts.

C Proofs of the Finite-Sample Results

We first derive a cumulative form of the descent inequality (8). We then express the sampled value error as a propagated initial error plus centered sampling contributions. This separates the proof into the effect of initialization and the effect of sampling. Repeated descent bounds the initial contribution, and cumulative descent bounds the variance of the sampling contribution. Independence across batches allows us to account for cancellation between sampling contributions.

C.1 Notation and Bellman errors

We use the vector notation introduced in Appendix B. Integer superscripts on matrices denote powers. Let $\boldsymbol{\mu}$ collect the state-sampling probabilities and define

$$D_\mu = \operatorname{diag}(\boldsymbol{\mu}). \tag{C.1}$$

The vector $\boldsymbol{\omega}^2$ has components $\omega^2(s)$ from (24). Recall that $d_\pi(s) > 0$ and $\sum_s d_\pi(s) = 1$.

The notation $\boldsymbol{\delta}_{\boldsymbol{v}}$ always denotes the vector of expected TD errors, including when $\boldsymbol{v}$ is a sampled iterate or an auxiliary value vector. By (B.1) and the Bellman equation (3),

$$\boldsymbol{\delta}_{\boldsymbol{v}} = -(I - \gamma P)(\boldsymbol{v} - \boldsymbol{v}_\pi). \tag{C.2}$$

Since P is a transition matrix and $\gamma < 1$,

$$(I - \gamma P)^{-1} = \sum_{j=0}^{\infty} \gamma^j P^j \geq 0. \tag{C.3}$$

Solving (C.2) for the value error therefore gives

$$|\boldsymbol{v} - \boldsymbol{v}_\pi| = |(I - \gamma P)^{-1}\boldsymbol{\delta}_{\boldsymbol{v}}| \leq (I - \gamma P)^{-1}|\boldsymbol{\delta}_{\boldsymbol{v}}|. \tag{C.4}$$

Stationarity (5) states that $\boldsymbol{d}_\pi^\top P = \boldsymbol{d}_\pi^\top$. It follows that $\boldsymbol{d}_\pi^\top P^j = \boldsymbol{d}_\pi^\top$ for every j. Multiplying the series in (C.3) by this row vector gives

$$\boldsymbol{d}_\pi^\top (I - \gamma P)^{-1} = \frac{\boldsymbol{d}_\pi^\top}{1 - \gamma}. \tag{C.5}$$

Multiplying both sides of (C.4) with $\boldsymbol{d}_\pi^\top$ and applying (C.5) then implies $\|\boldsymbol{v} - \boldsymbol{v}_\pi\|_{1,d_\pi} \leq BE_1(\boldsymbol{v})/(1 - \gamma)$.

C.2 Expected updates and cumulative descent

We begin with the synchronous expected update (7). It averages the one-state updates to which (8) applies.

Lemma 1 (Synchronous descent)**.** *Let $\boldsymbol{v}^+ = \boldsymbol{v} + \alpha D_\mu \boldsymbol{\delta}_{\boldsymbol{v}}$, with $0 < \alpha \leq 1$. Then*

$$BE_1(\boldsymbol{v}^+) \leq BE_1(\boldsymbol{v}) - \alpha(1 - \gamma) \sum_s \mu(s) d_\pi(s) |\delta_{\boldsymbol{v}}(s)|. \tag{C.6}$$

If $\boldsymbol{\delta}_{\boldsymbol{v}} \geq 0$, then $\boldsymbol{\delta}_{\boldsymbol{v}^+} \geq 0$ and equality holds in (C.6)*.*

Proof. For each state s, let $\boldsymbol{v}^{(s)} = \boldsymbol{v} + \alpha\delta_{\boldsymbol{v}}(s)\mathbf{1}_s$ be the one-state expected update (6). Their weighted average is $\boldsymbol{v}^+ = \sum_s \mu(s)\boldsymbol{v}^{(s)}$. The function BE_1 is convex because it is a weighted sum of absolute values of affine functions. Hence

$$BE_1(\boldsymbol{v}^+) \leq \sum_s \mu(s) BE_1(\boldsymbol{v}^{(s)}). \tag{C.7}$$

We now apply (8) separately to each one-state update:

$$BE_1(\boldsymbol{v}^{(s)}) \leq BE_1(\boldsymbol{v}) - \alpha(1 - \gamma) d_\pi(s) |\delta_{\boldsymbol{v}}(s)|. \tag{C.8}$$

This also holds when $\delta_{\boldsymbol{v}}(s) = 0$, since the update is then zero. We multiply by $\mu(s)$ and sum over states. Substituting the result into (C.7) proves (C.6).

To establish equality for nonnegative Bellman errors, we expand (4) after the synchronous update:

$$\delta_{\boldsymbol{v}^+}(s) = (1 - \alpha\mu(s))\delta_{\boldsymbol{v}}(s) + \alpha\gamma \sum_{s'} p(s' \mid s)\mu(s')\delta_{\boldsymbol{v}}(s'). \tag{C.9}$$

All coefficients are nonnegative because $\alpha \leq 1$. Thus nonnegative Bellman errors remain nonnegative, and we can omit the absolute values in $BE_1(\boldsymbol{v}^+)$. Multiplying (C.9) by $d_\pi(s)$ and summing gives

$$BE_1(\boldsymbol{v}^+) = BE_1(\boldsymbol{v}) - \alpha \sum_s d_\pi(s)\mu(s)\delta_{\boldsymbol{v}}(s) + \alpha\gamma \sum_{s'} \mu(s')\delta_{\boldsymbol{v}}(s') \sum_s d_\pi(s)p(s' \mid s). \tag{C.10}$$

By stationarity (5), the inner sum is $d_\pi(s')$. The last term is therefore γ times the term subtracted in the first line. We obtain

$$BE_1(\boldsymbol{v}) - BE_1(\boldsymbol{v}^+) = \alpha(1-\gamma) \sum_s d_\pi(s)\mu(s)\delta_{\boldsymbol{v}}(s), \tag{C.11}$$

which proves the equality case. □

To relate these decreases to value errors, we define the matrix of the expected value-error update:

$$A \stackrel{\text{def}}{=} I - \alpha D_\mu (I - \gamma P). \tag{C.12}$$

The expression $A = I - \alpha D_\mu + \alpha\gamma D_\mu P$ shows that $A \geq 0$. The synchronous expected update can be written as

$$\boldsymbol{v}^+ = A\boldsymbol{v} + \alpha D_\mu \boldsymbol{r}, \qquad \boldsymbol{v}^+ - \boldsymbol{v}_\pi = A(\boldsymbol{v} - \boldsymbol{v}_\pi). \tag{C.13}$$

The next lemma sums the decreases from Lemma 1 along an expected-update sequence with a chosen initial Bellman-error vector. We use this sequence only in the analysis. Let

$$\rho \stackrel{\text{def}}{=} 1 - \alpha(1-\gamma)\mu_{\min}. \tag{C.14}$$

Lemma 2 (Cumulative descent)**.** *For any vector $\boldsymbol{q} \geq 0$, define*

$$\widetilde{\boldsymbol{v}}_0 = \boldsymbol{v}_\pi - (I - \gamma P)^{-1}\boldsymbol{q}, \qquad \widetilde{\boldsymbol{v}}_{k+1} = \widetilde{\boldsymbol{v}}_k + \alpha D_\mu \boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_k}. \tag{C.15}$$

Then, for every integer $k \geq 0$,

$$\boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_0} = \boldsymbol{q}, \qquad \boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_k} \geq 0, \qquad BE_1(\widetilde{\boldsymbol{v}}_k) \leq \rho^k \|\boldsymbol{q}\|_{1,d_\pi}. \tag{C.16}$$

Moreover, for every integer $n \geq 0$,

$$\alpha \sum_{k=0}^{n-1} \boldsymbol{d}_\pi^\top A^k D_\mu \boldsymbol{q} = \frac{BE_1(\widetilde{\boldsymbol{v}}_0) - BE_1(\widetilde{\boldsymbol{v}}_n)}{1-\gamma} \leq \frac{\|\boldsymbol{q}\|_{1,d_\pi}}{1-\gamma}. \tag{C.17}$$

Proof. The choice of $\widetilde{\boldsymbol{v}}_0$ gives $\boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_0} = \boldsymbol{q}$ by (C.2). Lemma 1 shows inductively that all subsequent Bellman-error vectors are nonnegative. For each such vector,

$$\sum_s d_\pi(s)\mu(s)\delta_{\widetilde{\boldsymbol{v}}_k}(s) \geq \mu_{\min} BE_1(\widetilde{\boldsymbol{v}}_k). \tag{C.18}$$

Substituting into the descent inequality (C.6) gives $BE_1(\widetilde{\boldsymbol{v}}_{k+1}) \leq \rho BE_1(\widetilde{\boldsymbol{v}}_k)$. We iterate this inequality and use $BE_1(\widetilde{\boldsymbol{v}}_0) = \|\boldsymbol{q}\|_{1,d_\pi}$ to obtain (C.16).

To identify the cumulative sum, we first examine the increments of the auxiliary sequence. Subtracting two consecutive expected updates in (C.13) gives

$$\widetilde{\boldsymbol{v}}_{k+2} - \widetilde{\boldsymbol{v}}_{k+1} = A(\widetilde{\boldsymbol{v}}_{k+1} - \widetilde{\boldsymbol{v}}_k). \tag{C.19}$$

Induction therefore yields

$$\widetilde{\boldsymbol{v}}_{k+1} - \widetilde{\boldsymbol{v}}_k = A^k(\widetilde{\boldsymbol{v}}_1 - \widetilde{\boldsymbol{v}}_0) = \alpha A^k D_\mu \boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_0} = \alpha A^k D_\mu \boldsymbol{q}. \tag{C.20}$$

On the other hand, the update in (C.15) makes this same increment equal to $\alpha D_\mu \boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_k}$. Consequently, $D_\mu \boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_k} = A^k D_\mu \boldsymbol{q}$. We use this identity in the exact decrease (C.11):

$$BE_1(\widetilde{\boldsymbol{v}}_k) - BE_1(\widetilde{\boldsymbol{v}}_{k+1}) = \alpha(1-\gamma)\boldsymbol{d}_\pi^\top A^k D_\mu \boldsymbol{q}. \tag{C.21}$$

Summing over $k = 0, \ldots, n-1$ cancels consecutive Bellman errors, leaving $BE_1(\widetilde{\boldsymbol{v}}_0) - BE_1(\widetilde{\boldsymbol{v}}_n)$. The final Bellman error is nonnegative, so this difference is at most $\|\boldsymbol{q}\|_{1,d_\pi}$. Division by $1-\gamma$ proves (C.17). □

The left side of (C.17) is a sum of propagated nonnegative vectors. The identity bounds it by the total decrease of BE_1 under expected TD. We will use $\boldsymbol{q}$ to represent either the absolute initial Bellman errors or the sampling variances.

C.3 THE SAMPLED VALUE-ERROR RECURSION

We now separate sampling noise from the value error. Define

$$\xi_t = \frac{1}{m}\sum_{i=1}^{m}\big(r_{t,i} + \gamma v_\pi(s'_{t,i}) - v_\pi(s_t)\big). \tag{C.22}$$

This is the batch TD error evaluated at $\boldsymbol{v}_\pi$, not at $\boldsymbol{v}_t$. It depends only on the current batch. The Bellman equation centers each summand, and conditional independence within the batch gives

$$\mathbb{E}[\xi_t \mid s_t = s] = 0, \qquad \mathbb{E}[\xi_t^2 \mid s_t = s] = \frac{\omega^2(s)}{m}. \tag{C.23}$$

Let $\boldsymbol{e}_t = \boldsymbol{v}_t - \boldsymbol{v}_\pi$ and define the empirical successor distribution

$$\widehat{\boldsymbol{p}}_t = \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{s'_{t,i}}. \tag{C.24}$$

Then, from $e_t(s) = v_t(s) - v_\pi(s)$ and (23), we have

$$\begin{aligned}
e_{t+1}(s_t) &= (1-\alpha)v_t(s_t) + \frac{\alpha}{m}\sum_{i=1}^{m}\big[r_{t,i} + \gamma v_t(s'_{t,i})\big] - v_\pi(s_t) \\
&= (1-\alpha)e_t(s_t) + \frac{\alpha}{m}\sum_{i=1}^{m}\big[r_{t,i} + \gamma v_t(s'_{t,i}) - v_\pi(s_t)\big] \\
&= (1-\alpha)e_t(s_t) + \frac{\alpha}{m}\sum_{i=1}^{m}\big[r_{t,i} + \gamma v_\pi(s'_{t,i}) - v_\pi(s_t)\big] + \frac{\alpha\gamma}{m}\sum_{i=1}^{m} e_t(s'_{t,i}) \\
&= (1-\alpha)e_t(s_t) + \alpha\xi_t + \alpha\gamma\,\widehat{\boldsymbol{p}}_t^{\top}\boldsymbol{e}_t.
\end{aligned} \tag{C.25}$$

The first equality is (23) rearranged, with $v_\pi(s_t)$ subtracted from both sides. The second writes $v_\pi(s_t) = (1-\alpha)v_\pi(s_t) + \alpha v_\pi(s_t)$ and uses $e_t(s_t) = v_t(s_t) - v_\pi(s_t)$. The third uses $v_t(s'_{t,i}) = v_\pi(s'_{t,i}) + e_t(s'_{t,i})$. The fourth uses the definition (C.22) of ξ_t and definition (C.24) of $\widehat{\boldsymbol{p}}_t$. All other components are unchanged. To write this recursion in vector form, we define

$$U_t = I - \alpha\mathbf{1}_{s_t}\big(\mathbf{1}_{s_t} - \gamma\widehat{\boldsymbol{p}}_t\big)^{\top}. \tag{C.26}$$

Then

$$\boldsymbol{e}_{t+1} = U_t\boldsymbol{e}_t + \alpha\xi_t\mathbf{1}_{s_t}. \tag{C.27}$$

The matrix U_t depends on the selected state and sampled successors, not on the current value estimates.

Because $\alpha \le 1$, the s_t row of U_t is nonnegative and sums to $1 - \alpha(1-\gamma)$. Every other row is a row of the identity. Thus

$$U_t(s,s') \ge 0, \qquad \sum_{s'} U_t(s,s') \le 1. \tag{C.28}$$

Averaging over the selected state and its successors gives

$$\mathbb{E}_{s_t\sim\mu}[U_t] = I - \alpha D_\mu + \alpha\gamma D_\mu P = A. \tag{C.29}$$

Although U_t and ξ_t share samples and may be dependent, both are independent of batches at other times. We use only this independence across batches.

C.4 PROOF OF THEOREM 2

Proof. We fix $n \ge 1$. For $0 \le t < n$, define

$$F_t = U_{n-1}U_{n-2}\cdots U_t, \qquad F_n = I. \tag{C.30}$$

Nonnegativity and row sums at most one are preserved under matrix multiplication. Thus, (C.28) implies that every entry of F_t lies in $[0,1]$. Independence across batches lets us take expectations one factor at a time in this product. By (C.29),

$$0 \le F_t(s,s') \le 1, \qquad \mathbb{E}F_t = A^{n-t}. \tag{C.31}$$

Iterating (C.27) separates the two sources of error:

$$\boldsymbol{e}_n = F_0 \boldsymbol{e}_0 + \boldsymbol{z}, \qquad \boldsymbol{z} = \alpha \sum_{t=0}^{n-1} \xi_t F_{t+1} \mathbf{1}_{s_t}. \tag{C.32}$$

We first bound the contribution from initialization. Since $F_0 \geq 0$, $|F_0 \boldsymbol{e}_0| \leq F_0 |\boldsymbol{e}_0|$. Taking expectations and weighting by $\boldsymbol{d}_\pi$ gives

$$\mathbb{E}\big[\|F_0 \boldsymbol{e}_0\|_{1,d_\pi}\big] = \boldsymbol{d}_\pi^\top \mathbb{E}\big[|F_0 \boldsymbol{e}_0|\big] \leq \boldsymbol{d}_\pi^\top \mathbb{E}\big[F_0 |\boldsymbol{e}_0|\big] = \boldsymbol{d}_\pi^\top A^n |\boldsymbol{e}_0|. \tag{C.33}$$

Apply Lemma 2 with $\boldsymbol{q} = |\boldsymbol{\delta}_{\boldsymbol{v}_0}|$, and denote the resulting sequence by $\widetilde{\boldsymbol{v}}_k$. Then

$$\begin{aligned} A^n |\boldsymbol{e}_0| &\leq A^n (I - \gamma P)^{-1} |\boldsymbol{\delta}_{\boldsymbol{v}_0}| \\ &= A^n (\boldsymbol{v}_\pi - \widetilde{\boldsymbol{v}}_0) \\ &= \boldsymbol{v}_\pi - \widetilde{\boldsymbol{v}}_n \\ &= (I - \gamma P)^{-1} \boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_n}. \end{aligned} \tag{C.34}$$

The inequality uses (C.4), which gives $|\boldsymbol{e}_0| \leq (I - \gamma P)^{-1} |\boldsymbol{\delta}_{\boldsymbol{v}_0}|$, together with $A^n \geq 0$, which preserves componentwise inequalities. The second line is the definition (C.15) of $\widetilde{\boldsymbol{v}}_0$ with $\boldsymbol{q} = |\boldsymbol{\delta}_{\boldsymbol{v}_0}|$. The third holds because $\widetilde{\boldsymbol{v}}_{k+1}$ is the synchronous expected update of $\widetilde{\boldsymbol{v}}_k$, so by (C.13), $\widetilde{\boldsymbol{v}}_{k+1} - \boldsymbol{v}_\pi = A(\widetilde{\boldsymbol{v}}_k - \boldsymbol{v}_\pi)$ for every k; iterating n times gives the equality in the third line. The last uses the residual identity (C.2), $\boldsymbol{\delta}_{\boldsymbol{v}} = -(I - \gamma P)(\boldsymbol{v} - \boldsymbol{v}_\pi)$, with $\boldsymbol{v} = \widetilde{\boldsymbol{v}}_n$.

By multiplying both sides of (C.34) by $\boldsymbol{d}_\pi^\top$, we obtain

$$\boldsymbol{d}_\pi^\top A^n |\boldsymbol{e}_0| \leq \boldsymbol{d}_\pi^\top (I - \gamma P)^{-1} \boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_n} = \frac{1}{1-\gamma} \boldsymbol{d}_\pi^\top \boldsymbol{\delta}_{\widetilde{\boldsymbol{v}}_n} = \frac{BE_1(\widetilde{\boldsymbol{v}}_n)}{1-\gamma} \leq \rho^n \frac{BE_1(\boldsymbol{v}_0)}{1-\gamma}, \tag{C.35}$$

where the second relation follows from (C.5), and the last inequality is due to (C.16) in Lemma 2. Plugging this into (C.33), we obtain a bound on the contribution from initialization to the value error in (C.32):

$$\mathbb{E}\big[\|F_0 \boldsymbol{e}_0\|_{1,d_\pi}\big] \leq \rho^n \frac{BE_1(\boldsymbol{v}_0)}{1-\gamma}. \tag{C.36}$$

We next bound the accumulated noise, $\boldsymbol{z}$, in (C.32). Fix a state s. The sth component of $\boldsymbol{z}$ in (C.32) is the sum of the contributions of the individual batches:

$$z(s) = \sum_{t=0}^{n-1} x_t(s), \qquad x_t(s) \stackrel{\text{def}}{=} \alpha F_{t+1}(s, s_t)\, \xi_t. \tag{C.37}$$

Since $0 \leq F_{t+1}(s, s_t) \leq 1$ by (C.31) and ξ_t has a finite second moment by (C.23), each $x_t(s)$ has a finite second moment. We first show that contributions from different steps are uncorrelated. Let $\mathcal{B}_{t+1}$ denote the collection of batches at steps $t+1, \ldots, n-1$ (empty if $t = n-1$). For $t < u$,

$$\begin{aligned} \mathbb{E}[x_t(s) x_u(s)] &= \mathbb{E}\big[x_u(s)\, \mathbb{E}[x_t(s) \mid \mathcal{B}_{t+1}]\big] \\ &= \mathbb{E}\Big[x_u(s)\, \alpha \sum_{s'} F_{t+1}(s, s') \mu(s')\, \mathbb{E}[\xi_t \mid s_t = s']\Big] \\ &= 0. \end{aligned} \tag{C.38}$$

The first equality holds because $x_u(s)$ depends only on the batches at steps $u, \ldots, n-1$, and is therefore determined by $\mathcal{B}_{t+1}$. The second holds because F_{t+1} is determined by $\mathcal{B}_{t+1}$, while the batch at step t is independent of $\mathcal{B}_{t+1}$, with $s_t \sim \mu$. The third follows from $\mathbb{E}[\xi_t \mid s_t = s'] = 0$; see (C.23). Note that we have not assumed that the contributions $x_t(s)$ are independent; only the centering of ξ_t at the true values is used. Expanding the square of the sum in (C.37) and using (C.38), we obtain

$$\mathbb{E}[z(s)^2] = \sum_{t=0}^{n-1} \mathbb{E}[x_t(s)^2]. \tag{C.39}$$

We next bound each term in (C.39):

$$\begin{aligned}\mathbb{E}[x_t(s)^2] &= \alpha^2 \sum_{s'} \mu(s')\, \mathbb{E}\big[F_{t+1}(s,s')^2\big]\, \mathbb{E}[\xi_t^2 \mid s_t = s'] \\ &= \frac{\alpha^2}{m} \sum_{s'} \mu(s')\omega^2(s')\, \mathbb{E}\big[F_{t+1}(s,s')^2\big] \\ &\le \frac{\alpha^2}{m} \sum_{s'} \mu(s')\omega^2(s')\, \mathbb{E}\big[F_{t+1}(s,s')\big] \\ &= \frac{\alpha^2}{m} \big[A^{n-1-t} D_\mu \boldsymbol{\omega}^{\,2}\big](s).\end{aligned} \tag{C.40}$$

The first equality holds because F_{t+1} is independent of the batch at step t. The second uses the conditional second moment in (C.23). The inequality holds because every entry of F_{t+1} lies in $[0,1]$ by (C.31), so its square is at most itself. The last equality follows from $\mathbb{E}F_{t+1} = A^{n-1-t}$; see (C.31).

We now sum over states with weights $d_\pi(s)$ and use cumulative descent. Let $\widetilde{\boldsymbol{v}}_k$ now denote the sequence of Lemma 2 with $\boldsymbol{q} = \boldsymbol{\omega}^{\,2}$, so that $BE_1(\widetilde{\boldsymbol{v}}_0) = \|\boldsymbol{\omega}^{\,2}\|_{1,d_\pi} = \bar{\omega}^{\,2}$. Then

$$\begin{aligned}\sum_s d_\pi(s)\, \mathbb{E}[z(s)^2] &\le \frac{\alpha^2}{m} \sum_{k=0}^{n-1} \boldsymbol{d}_\pi^\top A^k D_\mu \boldsymbol{\omega}^{\,2} \\ &= \frac{\alpha}{(1-\gamma)m} \big[BE_1(\widetilde{\boldsymbol{v}}_0) - BE_1(\widetilde{\boldsymbol{v}}_n)\big] \\ &\le \frac{\alpha\, \bar{\omega}^{\,2}}{(1-\gamma)m}.\end{aligned} \tag{C.41}$$

The first inequality follows from (C.39) and (C.40), after reindexing by $k = n-1-t$. The equality is (C.17) in Lemma 2. The last inequality uses $BE_1(\widetilde{\boldsymbol{v}}_n) \ge 0$ and $BE_1(\widetilde{\boldsymbol{v}}_0) = \bar{\omega}^{\,2}$. Thus, cumulative descent bounds the entire variance-propagation sum at once.

It remains to pass from second moments to the weighted ℓ_1 norm:

$$\mathbb{E}\|\boldsymbol{z}\|_{1,d_\pi} = \sum_s d_\pi(s)\, \mathbb{E}|z(s)| \le \sum_s d_\pi(s) \sqrt{\mathbb{E}[z(s)^2]} \le \sqrt{\sum_s d_\pi(s)\, \mathbb{E}[z(s)^2]} \le \bar{\omega} \sqrt{\frac{\alpha}{(1-\gamma)m}}. \tag{C.42}$$

The first inequality is Jensen's inequality, applied to each state. The second is Jensen's inequality for the concave square root, since d_π is a probability distribution. The last follows from (C.41).

Finally, by (C.32) and the triangle inequality,

$$\mathbb{E}\|\boldsymbol{e}_n\|_{1,d_\pi} \le \mathbb{E}\|F_0 \boldsymbol{e}_0\|_{1,d_\pi} + \mathbb{E}\|\boldsymbol{z}\|_{1,d_\pi} \le \rho^n \frac{BE_1(\boldsymbol{v}_0)}{1-\gamma} + \bar{\omega} \sqrt{\frac{\alpha}{(1-\gamma)m}},$$

where the second inequality follows from (C.36) and (C.42). Since $\rho = 1 - \alpha(1-\gamma)\mu_{\min}$, this is (25). This completes the proof of Theorem 2. □

C.5 Proof of Corollary 3

Proof. We show that each of the two terms in (25), with $n = T$, is at most $\varepsilon/2$. Note that the choice of α in (26) is equivalent to

$$\frac{1}{\alpha} = \max\left\{1, \frac{4\bar{\omega}^{\,2}}{(1-\gamma)m\varepsilon^2}\right\}, \tag{C.43}$$

including the case $\bar{\omega} = 0$. Hence (27) can be written as

$$T = \left\lceil \frac{1}{\alpha(1-\gamma)\mu_{\min}} \log \max\left(1, \frac{2BE_1(\boldsymbol{v}_0)}{(1-\gamma)\varepsilon}\right) \right\rceil. \tag{C.44}$$

Sampling term. If $\bar{\omega} = 0$, this term is zero. Otherwise, (26) gives $\alpha \le (1-\gamma)m\varepsilon^2/(4\bar{\omega}^2)$, and therefore

$$\bar{\omega}\sqrt{\frac{\alpha}{(1-\gamma)m}} \le \bar{\omega}\sqrt{\frac{\varepsilon^2}{4\bar{\omega}^2}} = \frac{\varepsilon}{2}.$$

Initialization term. If $2BE_1(\boldsymbol{v}_0) \le (1-\gamma)\varepsilon$, then $T = 0$, and the initialization term (i.e, the first term on the right hand side of (25)) equals $BE_1(\boldsymbol{v}_0)/(1-\gamma) \le \varepsilon/2$. For the other case where $2BE_1(\boldsymbol{v}_0) > (1-\gamma)\varepsilon$, the expression for T in (C.44) gives $\alpha(1-\gamma)\mu_{\min}T \ge \log\big(2BE_1(\boldsymbol{v}_0)/((1-\gamma)\varepsilon)\big)$, and

$$\rho^T \frac{BE_1(\boldsymbol{v}_0)}{1-\gamma} \le \exp\big(-\alpha(1-\gamma)\mu_{\min}T\big)\frac{BE_1(\boldsymbol{v}_0)}{1-\gamma} \le \frac{(1-\gamma)\varepsilon}{2BE_1(\boldsymbol{v}_0)} \cdot \frac{BE_1(\boldsymbol{v}_0)}{1-\gamma} = \frac{\varepsilon}{2}. \tag{C.45}$$

The first inequality uses $1 - x \le e^{-x}$ in the definition of ρ in (C.14), and the second uses the above lower bound on T. This proves $\mathbb{E}\|\boldsymbol{v}_T - \boldsymbol{v}_\pi\|_{1,d_\pi} \le \varepsilon$.

Bellman error. We first note that P is nonexpansive in $\|\cdot\|_{1,d_\pi}$. For any vector $\boldsymbol{x}$,

$$\|P\boldsymbol{x}\|_{1,d_\pi} = \sum_s d_\pi(s)\Big|\sum_{s'} p(s' \mid s)x(s')\Big| \le \sum_{s'} |x(s')| \sum_s d_\pi(s)p(s' \mid s) = \|\boldsymbol{x}\|_{1,d_\pi}, \tag{C.46}$$

where the inequality is the triangle inequality and the last equality is stationarity (5). Then, for any $\boldsymbol{v}$,

$$BE_1(\boldsymbol{v}) = \|(I-\gamma P)(\boldsymbol{v}-\boldsymbol{v}_\pi)\|_{1,d_\pi} \le \|\boldsymbol{v}-\boldsymbol{v}_\pi\|_{1,d_\pi} + \gamma\|P(\boldsymbol{v}-\boldsymbol{v}_\pi)\|_{1,d_\pi} \le (1+\gamma)\|\boldsymbol{v}-\boldsymbol{v}_\pi\|_{1,d_\pi}.$$

The equality is the residual identity (C.2). The first inequality is the triangle inequality, and the second follows from (C.46). Setting $\boldsymbol{v} = \boldsymbol{v}_T$ and taking expectations gives $\mathbb{E}BE_1(\boldsymbol{v}_T) \le (1+\gamma)\varepsilon$. This completes the proof of Corollary 3. □