STOCHASTIC TRACKING OF DRIFTING AFFINE FIXED POINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Policy evaluation in a changing environment requires tracking a value function that moves during learning. We study this problem through affine contractions whose fixed points change over time. For a constant step size, we bound the total expected squared tracking error in terms of the initial error, observation noise, and target movement. The result extends known deterministic bounds to observations that can be correlated across time. In a scalar Gaussian model, matching lower bounds identify when a suitably chosen step size achieves the best possible tracking rate and quantify the cost of temporal correlation. We also show how to combine trackers with different step sizes without knowing the noise level or target variation, using additional bounded, unbiased observations of the target. Finally, we apply the analysis to synchronous tabular temporal-difference learning with simulator access and a common invariant distribution, and quantify the extra sampling cost of adaptation.

1 INTRODUCTION

Reinforcement learning algorithms often operate in environments whose conditions change over time. A navigation system may encounter changing traffic conditions, while a recommendation system may face shifts in user preferences. In such settings, the rewards and consequences of an action can change, making earlier observations less representative of the decisions the learner currently faces. These challenges have motivated a growing literature on reinforcement learning in nonstationary environments (Cheung et al., 2020; Wei & Luo, 2021).

A basic part of this problem is keeping value estimates up to date. Policy evaluation estimates the expected future reward obtained by following a given policy, and these estimates support policy improvement in many reinforcement learning algorithms. Finite-time analyses of temporal-difference (TD) learning typically study a fixed policy in a fixed environment (Bhandari et al., 2021; Khamaru et al., 2021). When rewards or transition probabilities change, however, the value function changes as well. Even when the environment remains fixed, changing the policy can change its value function and a critic that evaluates the current policy must keep its estimates up to date as the policy changes (Walton, 2020). The learner must therefore maintain accurate estimates throughout learning, while the quantity being estimated continues to move.

Tracking a changing value function requires balancing responsiveness and noise reduction. A small step size reduces the influence of individual noisy observations but can leave the estimate slow to respond to change. A large step size responds more quickly but produces larger fluctuations. Correlation between observations adds another difficulty: when errors persist across successive samples, averaging them may reduce noise more slowly than it would with independent samples. A useful analysis must account for both the movement of the target and the information available in the observations.

This tradeoff is well established in stochastic tracking and time-varying optimization (Bastianello et al., 2024; Cutler et al., 2023; Sahu et al., 2026). For exact contractive updates, Hadiji et al. (2026) bound cumulative squared tracking error using the sum of squared changes in the target. This measure captures the difference between gradual movement and abrupt changes, even when the overall displacement is the same. We build on this deterministic result to study how temporally correlated observation noise affects the tracking guarantee.

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.

We consider a deterministic sequence of affine contractions $F_t(x) = M_t x + b_t$, each with its own fixed point. At each round, the learner observes a noisy evaluation of the current map and updates its estimate using a constant step size. This formulation includes discounted Bellman updates under the conditions of our TD application. The affine structure allows us to separate the error caused by initialization and target movement from the error caused by noise. We use this separation to obtain a cumulative mean-square bound that retains the deterministic tracking terms and makes the contribution of temporal correlation explicit.

Our contributions.

1. **A bound on tracking error.** We bound the total expected squared distance between the learner's estimates and the changing fixed points. The bound separates the contributions of the initial error, observation noise, and target movement. It shows how the step size and correlation between observations affect tracking accuracy, extending an existing guarantee for exact updates to noisy observations (Theorem 3).
2. **Limits on tracking accuracy.** For a scalar model with Gaussian noise, we establish a worst-case lower bound on the error of any estimator, even one that can use future observations. Over specified ranges of target movement, this lower bound matches, up to constant factors, the guarantee achieved by a suitably chosen step size. We also establish how temporal correlation increases this unavoidable error (Theorems 5 and 7).
3. **Combining trackers with different step sizes.** The appropriate step size depends on the noise level and target movement, which may be unknown. We therefore run several trackers with different step sizes and combine their predictions. To compare their accuracy, the method uses additional bounded observations that are unbiased estimates of the current target. The combined prediction has total expected squared error at most that of the best tracker in the collection plus an explicit penalty that grows logarithmically with the number of trackers (Theorem 9).
4. **Application to policy evaluation.** We apply the tracking bound to synchronous tabular TD with changing rewards and transition probabilities, assuming simulator access and a common invariant distribution. Appendix E gives the tracking guarantees, constructs the additional target observations needed for adaptation, and quantifies their sampling cost. We also extend the analysis to coupled linear updates under explicit contraction assumptions (Appendix F).

Section 2 discusses prior work, Section 3 gives the problem formulation, and Sections 4–5 present the results and their scope. Complete proofs appear in the appendix.

2 Related work

Tracking and dynamic regret. Besbes et al. (2015) establish variation-dependent minimax guarantees for nonstationary stochastic optimization, using a budget on changes in the objective. Yang et al. (2016) study moving minimizers under exact and noisy gradient feedback, while Zhang et al. (2017) introduce squared path length into dynamic-regret bounds with multiple gradient queries. For contractive updates, Bernstein & Dall'Anese (2018) analyze asynchronous tracking. Madden et al. (2021) establish asymptotic tracking bounds and lower bounds for first-order methods in time-varying optimization, while Hadiji et al. (2026) bound cumulative squared error by quadratic path variation. With noisy feedback, Bastianello et al. (2024) treat general time-varying operators, and Cutler et al. (2023) separate optimization, noise, and minimizer drift. Sahu et al. (2026) study the same stepsize tradeoff and the limitations of momentum in nonstationary optimization. Section 5 compares the assumptions and guarantees with our affine result.

Sequence estimation and adaptation. The scalar class $\sum_{t<T}(\theta_{t+1} - \theta_t)^2 \le W$ is a discrete Sobolev class. Sadhanala et al. (2016) give its minimax estimation bounds. Baby & Wang (2019, Section 3.5) obtain online forecasting guarantees that adapt to unknown variation, with the same leading rate up to logarithmic factors and additional terms. Their method assumes a known horizon and sub-Gaussian noise parameter, uses noisy scalar target observations, and requires no separate validation stream. Baby & Wang (2021) develop variation-adaptive guarantees for exp-concave online learning, including squared loss. Correlated errors in nonparametric regression are also well studied (Opsomer et al., 2001). Our Gaussian results specialize this setting to affine queries and keep the finite-horizon dependence on the covariance factor explicit. For general affine maps, we assume a

target-validation oracle and apply standard exponential weighting (Cesa-Bianchi & Lugosi, 2006) to tracking recursions.

Stochastic approximation. Konda & Tsitsiklis (2003) study slowly changing Markov statistics with decreasing stepsizes, and Kumar et al. (2019) analyze constant-stepsize tracking. For fixed targets, finite-sample analyses cover general norms, multiplicative noise, and high-probability guarantees (Chen et al., 2020; Chandak, 2026; Diakonikolas, 2026). Related two-timescale results are discussed alongside the extension in Appendix F.

Policy evaluation and nonstationary RL. Stationary TD has finite-time bounds (Bhandari et al., 2021) and instance-dependent optimality results (Khamaru et al., 2021); Totaro & Jonsson (2021) propose a Kalman-style approach to changing value parameters. Most nonstationary RL theory concerns control and exploration (Cheung et al., 2020; Mao et al., 2021; Wei & Luo, 2021; Zhou et al., 2022; Feng et al., 2023; Jali et al., 2026). Closer to our application, Walton (2020) analyzes tabular TD and Q-learning along a changing Markov trajectory. Our TD guarantee concerns cumulative estimation error under synchronous generative access.

3 Problem formulation

We study the problem of tracking a changing fixed point using noisy observations. Fix a horizon $T \geq 2$. At each time $t = 1, \ldots, T$, consider an affine map

$$F_t(x) = M_t x + b_t, \qquad M_t \in \mathbb{R}^{d\times d}, \quad b_t \in \mathbb{R}^d.$$

The sequence of maps is deterministic and may change over time. We equip $\mathbb{R}^d$ with an inner-product norm $\|\cdot\|$ and use $\|\cdot\|_{\mathrm{op}}$ for the corresponding operator norm.

Assumption 1 (Uniform contraction)**.** *There is a constant $\kappa \in (0, 1]$ such that $\|M_t\|_{\mathrm{op}} \leq 1 - \kappa$ for every $t = 1, \ldots, T$.*

Under this assumption, each map has a unique fixed point,

$$x_t^* = F_t(x_t^*) = (I - M_t)^{-1} b_t.$$

The learner's task is to estimate x_t^* at each time t using the observations available before that time.

Observations and updates. The learner starts from a deterministic initial estimate $x_1 \in \mathbb{R}^d$. Let $\mathcal{F}_t$ denote the information available before the observation at time t, including past observations and any internal randomness used by the learner. The estimate x_t is $\mathcal{F}_t$-measurable. After forming this estimate, the learner receives a noisy evaluation of the current map:

$$\widehat{F}_t(x_t) = F_t(x_t) + \xi_{t+1}, \qquad t = 1, \ldots, T-1,$$

where the observation error ξ_{t+1} is $\mathcal{F}_{t+1}$-measurable.

We analyze the update

$$x_{t+1} = (1 - \eta)x_t + \eta \widehat{F}_t(x_t), \qquad t = 1, \ldots, T-1, \tag{1}$$

with a constant stepsize $\eta \in (0, 1]$. Thus, each new estimate combines the previous estimate with the latest noisy evaluation.

Noise assumptions. We allow observation errors to be correlated across time. In addition to requiring each error to have mean zero and bounded second moment, we control how much noise can accumulate in a linear combination of observations.

Assumption 2 (Noise and temporal correlation)**.** *There are constants $\sigma^2 \geq 0$ and $\tau \geq 1$ such that, for every $t = 1, \ldots, T-1$,*

$$\mathbb{E}[\xi_{t+1}] = 0, \qquad \mathbb{E}\,\|\xi_{t+1}\|^2 \leq \sigma^2.$$

Moreover, for every integer $1 \leq n \leq T-1$ and every collection of deterministic matrices $C_1, \ldots, C_n \in \mathbb{R}^{d\times d}$,

$$\mathbb{E}\left\|\sum_{s=1}^{n} C_s \xi_{s+1}\right\|^2 \leq \tau\sigma^2 \sum_{s=1}^{n} \|C_s\|_{\mathrm{op}}^2 . \tag{2}$$

Here, σ^2 bounds the size of an individual observation error, while τ controls the effect of temporal correlation on accumulated noise. A standard special case is martingale-difference noise:

$$\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\xi_{t+1}\|^2 \mid \mathcal{F}_t] \le \sigma^2.$$

These conditions imply Assumption 2 with $\tau = 1$. The general assumption requires only unconditional mean-zero errors and therefore also permits correlated noise. For example, centered scalar noise with

$$\mathbb{E}[\xi_{s+1}\xi_{t+1}] = \sigma^2 \chi^{|s-t|}, \qquad 0 \le \chi < 1,$$

satisfies the assumption with $\tau = (1+\chi)/(1-\chi)$. Proposition 14 gives the corresponding matrix result.

Tracking objective. We measure performance by the cumulative expected squared distance between the learner's estimates and the current fixed points:

$$\mathcal{E}_T := \sum_{t=1}^{T} \mathbb{E}\,\|x_t - x_t^*\|^2 .$$

To describe how much the target changes, we use its squared variation,

$$W_T := \sum_{t=1}^{T-1} \left\|x_{t+1}^* - x_t^*\right\|^2 .$$

This quantity captures the magnitudes of successive changes and distinguishes movement concentrated in a few rounds from movement spread more evenly over time. For example, a displacement $v \in \mathbb{R}^d$ made in one jump contributes $\|v\|^2$, whereas the same displacement spread evenly over $T-1$ steps contributes $\|v\|^2/(T-1)$.

Our goal is to bound $\mathcal{E}_T$ in terms of the initial error, the contraction strength, the noise, and the target variation, and to understand how the stepsize affects tracking accuracy.

4 Main results

4.1 Finite-time tracking bound

We first bound the cumulative tracking error for any constant stepsize. The result quantifies the effects of initialization, observation noise, and movement of the fixed points.

Theorem 3 (Cumulative tracking error)**.** *Suppose Assumptions 1–2 hold. For every constant stepsize* $\eta \in (0, 1]$*,*

$$\boxed{\mathcal{E}_T \le \frac{\|x_1 - x_1^*\|^2}{\eta\kappa} + \frac{\eta\tau\sigma^2 T}{\kappa(2-\eta\kappa)} + \frac{W_T}{\eta^2\kappa^2}.} \tag{3}$$

The first term bounds the contribution of the initial estimation error. The second accounts for accumulated observation noise, while the third accounts for the lag behind the moving target. Increasing the stepsize reduces the initialization and movement terms but increases the noise term. The covariance factor τ enters only through the noise term.

To obtain an explicit stepsize choice, we next assume exact initialization, $x_1 = x_1^*$. This removes the first term and isolates the tradeoff between noise and target movement. Using $2 - \eta\kappa \ge 1$, we obtain the simpler bound

$$\mathcal{E}_T \le \frac{\tau\sigma^2 T}{\kappa}\eta + \frac{W_T}{\kappa^2}\eta^{-2}.$$

Minimizing this expression over positive stepsizes gives the following corollary, provided the minimizing stepsize does not exceed one.

Corollary 4 (Tracking bound with a tuned stepsize)**.** *Under the assumptions of Theorem 3, suppose* $x_1 = x_1^*$*,* $W_T > 0$*, and* $\tau\sigma^2 > 0$*. Define* $\eta_{\mathrm{bal}} := \left(\frac{2W_T}{\kappa\tau\sigma^2 T}\right)^{1/3}$*. If* $\eta_{\mathrm{bal}} \le 1$*, running the iteration with* $\eta = \eta_{\mathrm{bal}}$ *gives*

$$\mathcal{E}_T < 2\frac{(\tau\sigma^2 T)^{2/3} W_T^{1/3}}{\kappa^{4/3}}.$$

The tuned stepsize increases with target variation and decreases with the noise level and covariance factor. This choice requires knowledge of those quantities and minimizes the simplified upper bound above. Theorem 3 itself requires neither exact initialization nor positive noise or variation, and remains valid for every $\eta \in (0, 1]$.

4.2 A SCALAR GAUSSIAN BENCHMARK

We next examine whether the tracking rate in Corollary 4 can be improved when the observation noise is independent. We consider a scalar model in which noisy operator evaluations can be converted into direct observations of the changing target.

Consider the affine maps

$$F_t^\theta(x) = (1 - \kappa)x + \kappa\theta_t,$$

whose fixed points are θ_t. These are also the Bellman maps of a one-state reward process with discount factor $1 - \kappa$ and mean reward $\kappa\theta_t$.

At each round, the learner may make at most q queries, where $q \geq 1$ is an integer. Query locations may depend on previous observations. The response to the ith query at location x is

$$\mathsf{O}_{t,i}(x) = F_t^\theta(x) + \varepsilon_{t,i},$$

where the errors are independent across queries and rounds, with $\varepsilon_{t,i} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 > 0$.

For a variation budget $W > 0$, define the target class

$$\Theta_T(W) := \left\{ \theta \in \mathbb{R}^T : \theta_1 = 0, \quad \sum_{t=1}^{T-1} (\theta_{t+1} - \theta_t)^2 \leq W \right\}.$$

Thus, W bounds the variation of every sequence in the class, whereas W_T denotes the variation of an individual sequence. The initial target $\theta_1 = 0$ is known to the learner.

When κ is known, each response can be transformed as

$$\frac{\mathsf{O}_{t,i}(x) - (1 - \kappa)x}{\kappa} = \theta_t + \frac{\varepsilon_{t,i}}{\kappa}.$$

The query location disappears from this expression. Averaging q transformed responses therefore gives a Gaussian observation of θ_t with variance

$$\nu^2 := \frac{\sigma^2}{q\kappa^2}.$$

The problem reduces to estimating a sequence from noisy observations under a bound on its squared increments. Writing $(D\theta)_t = \theta_{t+1} - \theta_t$, this constraint is $\|D\theta\|_2^2 \leq W$, commonly called a discrete Sobolev constraint. Its estimation rates are classical (Sadhanala et al., 2016). The following theorem specializes the lower bound to our query model, with an explicit variation range and a fixed initial target.

Theorem 5 (Gaussian lower bound for moderate variation)**.** *In the scalar model above, suppose the learner is told κ, σ^2, q, W, and $\theta_1 = 0$. Let $\nu^2 = \sigma^2/(q\kappa^2)$. There is a universal constant $c > 0$ such that, for $T \geq 16$ and*

$$\frac{16\nu^2}{T^2} \leq W \leq \frac{\nu^2 T}{32},$$

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_T(W)} \mathbb{E}_\theta \sum_{t=1}^{T} (\widehat{\theta}_t - \theta_t)^2 \geq c \frac{(\sigma^2 T/q)^{2/3} W^{1/3}}{\kappa^{4/3}}.$$

The infimum allows estimators to use all qT observations before producing any estimate. The bound therefore also applies to online algorithms, which have access only to past observations.

The appendix gives a self-contained proof with $c = 1/4096$. We now compare this lower bound with the constant-stepsize tracker. Averaging q independent responses at each update reduces the oracle noise variance to σ^2/q, allowing us to apply Theorem 3 with this smaller variance.

Corollary 6 (Matching tracking guarantee)**.** *In the same scalar model, start from $x_1 = 0$. At each update time $t = 1, \dots, T-1$, make q queries at the current iterate and average the responses. Define*

$$\eta_{q,\mathrm{bal}} := \left(\frac{2qW}{\kappa \sigma^2 T} \right)^{1/3}.$$

If $\eta_{q,\mathrm{bal}} \le 1$, running the iteration in equation 1 with $\eta = \eta_{q,\mathrm{bal}}$ gives

$$\sup_{\theta \in \Theta_T(W)} \mathbb{E}_\theta \sum_{t=1}^{T} (x_t - \theta_t)^2 < 2 \frac{(\sigma^2 T/q)^{2/3} W^{1/3}}{\kappa^{4/3}}.$$

Consequently, for $T \ge 16$, the upper and lower bounds agree up to universal constants whenever

$$\frac{16\nu^2}{T^2} \le W \le \min\left\{ \frac{\nu^2 T}{32}, \frac{\kappa^3 \nu^2 T}{2} \right\}.$$

Within this range, the constant-stepsize tracker attains the smallest possible worst-case error up to constants, despite using only past observations. The two upper restrictions on W come from the lower-bound theorem and the requirement that the balancing stepsize be at most one.

4.3 The cost of temporal correlation

The previous subsection considered independent observation errors. We now allow errors to be correlated across rounds and show that the dependence on τ in the tuned tracking bound is necessary over a specified range of target variation.

We retain the scalar maps F_t^θ and target class $\Theta_T(W)$ from Section 4.2. The ith query at round t uses error $\varepsilon_{t,i}$. For each $i = 1, \dots, q$, the sequence $(\varepsilon_{t,i})_{t=1}^T$ is jointly Gaussian, with

$$\mathbb{E}[\varepsilon_{t,i}] = 0, \qquad \mathbb{E}[\varepsilon_{s,i}\varepsilon_{t,i}] = \sigma^2 \chi^{|s-t|}, \qquad \chi = \frac{\tau - 1}{\tau + 1}.$$

Sequences corresponding to different query indices are independent. Thus, $\tau = 1$ gives independent errors, while larger values of τ give more persistent temporal correlation. Averaging the q responses at each round produces noise satisfying Assumption 2 with variance parameter σ^2/q and the same covariance factor τ.

Theorem 7 (Gaussian lower bound with temporal correlation)**.** *Let $T \ge 16$, $\sigma^2 > 0$, $q \ge 1$ be an integer, $1 \le \tau \le T/8$, and $\nu^2 = \sigma^2/(q\kappa^2)$. For the correlated observations above, if*

$$\frac{4\tau\nu^2}{T^2} \le W \le \frac{\nu^2 T}{128\tau^2},$$

then

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_T(W)} \mathbb{E}_\theta \sum_{t=1}^{T} (\widehat{\theta}_t - \theta_t)^2 \ge \frac{1}{8192} \frac{(\tau\sigma^2 T/q)^{2/3} W^{1/3}}{\kappa^{4/3}}.$$

The bound holds even if the learner knows $\kappa, \sigma^2, q, W, \tau$, and $\theta_1 = 0$, and receives all qT observations before estimating the targets.

To compare this lower bound with the tracker, start from $x_1 = 0$, average q responses at each of the first $T-1$ update times, and choose

$$\eta = \left(\frac{2qW}{\kappa\tau\sigma^2 T} \right)^{1/3}.$$

If $W \le \kappa^3 \tau \nu^2 T/2$, this stepsize is at most one. Theorem 3 then bounds the worst-case cumulative error over $\Theta_T(W)$ by a quantity smaller than $2(\tau\nu^2 T)^{2/3} W^{1/3}$. Consequently, the upper and lower bounds match up to universal constants whenever

$$\frac{4\tau\nu^2}{T^2} \le W \le \min\left\{ \frac{\nu^2 T}{128\tau^2}, \frac{\kappa^3 \tau \nu^2 T}{2} \right\}.$$

Under the theorem's conditions, this interval is nonempty when $\kappa T \ge 2$. Within this range, the factor $\tau^{2/3}$ is unavoidable in the Gaussian model, even when the noise covariance is known. Appendix C.4 proves the lower bound using target sequences that change slowly relative to the decay of noise correlations.

A constant-target example. To isolate the effect of temporal correlation, consider a target that remains fixed at zero: $\theta_t = 0$ for every $t \geq 1$. Initialize the tracker at $x_1 = 0$, so there is no initial error or target movement, and all subsequent estimation error comes from observation noise. For this calculation, extend the Gaussian noise model to all $t \geq 1$ and allow any $\tau \geq 1$. With q averaged responses per update and a fixed stepsize $\eta \in (0, 1]$, write $r = 1 - \eta\kappa$. The limiting mean-square error is

$$\lim_{t\to\infty} \mathbb{E}x_t^2 = \frac{\eta\sigma^2}{q\kappa(2-\eta\kappa)} \frac{1+r\chi}{1-r\chi}. \tag{4}$$

Indeed, expanding the recursion gives $\mathbb{E}x_t^2 = (\eta^2\sigma^2/q) \sum_{i,j=0}^{t-2} r^{i+j}\chi^{|i-j|}$; summing the resulting geometric series yields equation 4.

The first factor in equation 4 is the limiting error under independent noise. Temporal correlation multiplies it by

$$\frac{1+r\chi}{1-r\chi} = \frac{2\tau - \eta\kappa(\tau-1)}{2+\eta\kappa(\tau-1)}.$$

When $\eta\kappa\tau$ is small, this multiplier is approximately τ. In this regime, the tracker averages observations over a period longer than the noise remains strongly correlated. The upper restriction on W in Theorem 7 ensures that $\eta\kappa\tau \leq 1/4$ for the balancing stepsize, placing the matching result in this regime.

For a fixed positive stepsize, the limiting error instead approaches $\sigma^2/(q\kappa^2)$ as $\tau \to \infty$. Thus, the error does not continue to grow proportionally to τ when correlations persist much longer than the tracker's averaging period. The general upper bound can overestimate the correlation cost in that setting.

4.4 Adaptation using target validation

The tuned stepsizes above depend on the target variation and noise parameters, which may be unknown. We run several copies of the tracking iteration, each with a different constant stepsize, and form a weighted average of their predictions. After predicting at round t, we receive an additional observation of x_t^* whose conditional mean, given the learner's history, is x_t^*. We score each tracker by its squared distance to this observation and assign larger weights to trackers with smaller cumulative scores.

Additional target observations. Let $\mathcal{X}$ be a known nonempty closed convex set containing every fixed point x_t^* and contained in a ball of known radius $B > 0$. Let $\mathcal{G}_t$ contain all information available before the validation observation at round t, including the history, the internal iterates of all trackers, and their current weights.

Assumption 8 (Target validation)**.** *At each round $t = 1, \ldots, T$, after forming its prediction, the learner receives a validation vector $Y_t \in \mathcal{X}$ satisfying*

$$\mathbb{E}[Y_t \mid \mathcal{G}_t] = x_t^*. \tag{5}$$

For $t < T$, conditional on $\mathcal{G}_t$, Y_t is independent of the collection of operator samples used to update the trackers at round t.

Combining the trackers. Set $J = \lceil \log_2 T \rceil$ and run $J + 1$ trackers with stepsizes $\eta_j = 2^{-j}$, $j = 0, \ldots, J$. We refer to these trackers as experts. Expert j maintains an internal iterate $x_{t,j}$ and predicts $\widetilde{x}_{t,j} = \Pi_{\mathcal{X}}(x_{t,j})$, where $\Pi_{\mathcal{X}}$ is the nearest-point projection onto $\mathcal{X}$ with respect to the norm $\|\cdot\|$ introduced in Section 3. The internal iterates follow the original affine update; projection is used to bound the predictions that are combined and scored.

At each update, expert j receives $\widehat{F}_{t,j}(x_{t,j}) = F_t(x_{t,j}) + \xi_{t+1,j}$. For each fixed j, assume that its noise sequence satisfies Assumption 2 with the same parameters (σ^2, τ). Different experts may share update samples, and their noise sequences need not be independent. Algorithm 1 combines the projected predictions by exponential weighting. Experts whose past predictions were closer to the validation observations receive larger weights.

Algorithm 1 META-TRACK: combining trackers with target validation

Set $J = \lceil \log_2 T \rceil$ and $\lambda = 1/(8B^2)$.
For each $j = 0, \dots, J$, set $\eta_j = 2^{-j}$, $p_{1,j} = 1/(J+1)$, and $x_{1,j} = x_1$.
for $t = 1, \dots, T$ **do**
  Set $\widetilde{x}_{t,j} = \Pi_{\mathcal{X}}(x_{t,j})$ for every j.
  Predict $\bar{x}_t = \sum_{j=0}^{J} p_{t,j} \widetilde{x}_{t,j}$.
  Observe the validation vector Y_t.
  if $t < T$ **then**
    Set $p_{t+1,j} \propto p_{t,j} \exp\{-\lambda \left\| \widetilde{x}_{t,j} - Y_t \right\|^2\}$ and normalize the weights to sum to one.
    Obtain the operator observations $\widehat{F}_{t,j}(x_{t,j})$ for all experts.
    Update $x_{t+1,j} = (1-\eta_j) x_{t,j} + \eta_j \widehat{F}_{t,j}(x_{t,j})$ for every j.
  end if
end for

Theorem 9 (Adaptation with target validation)**.** *Suppose Assumptions 1 and 8 hold, and each expert's noise sequence satisfies Assumption 2 with the same parameters (σ^2, τ). Then Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E} \left\| \bar{x}_t - x_t^* \right\|^2 \le \min_{0 \le j \le J} \sum_{t=1}^{T} \mathbb{E} \left\| x_{t,j} - x_t^* \right\|^2 + 8B^2 \log(J+1), \tag{6a}$$

$$\le 4 \inf_{\eta \in [1/T, 1]} \left\{ \frac{\left\| x_1 - x_1^* \right\|^2}{\eta \kappa} + \frac{\eta \tau \sigma^2 T}{\kappa(2 - \eta\kappa)} + \frac{W_T}{\eta^2 \kappa^2} \right\} + 8B^2 \log(J+1). \tag{6b}$$

If $x_1 = x_1^*$, $W_T > 0$, $\tau\sigma^2 > 0$, and the balancing stepsize η_{bal} from Corollary 4 lies in $[1/T, 1]$, then

$$\sum_{t=1}^{T} \mathbb{E} \left\| \bar{x}_t - x_t^* \right\|^2 < 8 \frac{(\tau\sigma^2 T)^{2/3} W_T^{1/3}}{\kappa^{4/3}} + 8B^2 \log(1 + \lceil \log_2 T \rceil).$$

Equation 6a bounds the combined predictor's cumulative expected squared error by the smallest such error among the $J+1$ fixed-stepsize trackers, plus $8B^2 \log(J+1)$. Equation 6b gives an explicit guarantee in terms of initialization, noise, and target variation. It is obtained by applying Theorem 3 to the trackers and comparing the stepsize grid with all $\eta \in [1/T, 1]$, which introduces the factor four. Appendix D gives the proof.

The algorithm uses $J + 1 = O(\log T)$ trackers and does not require knowledge of W_T, σ^2, or τ. With separate pointwise queries, it uses $J+1$ operator calls per update. It also requires one validation observation per prediction round. Excluding operator evaluations and projections, computing the tracker updates, combining their predictions, and updating their weights requires $O(T(J+1)d)$ arithmetic operations, assuming that squared norms can be computed in $O(d)$ operations.

Direct smoothing of validation observations. The additional target observations also support a simpler tracker that uses no operator evaluations. For a fixed $a \in (0, 1]$, initialize $u_1 = x_1$ and update

$$u_{t+1} = (1-a) u_t + a Y_t, \qquad t = 1, \dots, T-1.$$

Conditional unbiasedness and boundedness imply $\mathbb{E}[\left\| Y_t - x_t^* \right\|^2 \mid \mathcal{G}_t] \le B^2$. Applying Theorem 3 to this recursion gives

$$\sum_{t=1}^{T} \mathbb{E} \left\| u_t - x_t^* \right\|^2 \le \frac{\left\| x_1 - x_1^* \right\|^2}{a} + \frac{aB^2 T}{2-a} + \frac{W_T}{a^2}.$$

Appendix D gives the details.

At $a = \eta\kappa$, this bound has the same initialization and movement terms as the operator-tracking bound in Theorem 3. The respective noise coefficients are B^2 for direct smoothing and $\tau\sigma^2/\kappa^2$ for operator tracking. Thus, the operator-tracking bound has a smaller noise term when $\tau\sigma^2/\kappa^2 < B^2$. This comparison concerns the upper bounds; it does not establish that either method always has smaller actual error.

4.5 Applications and extensions

Appendix E applies our results to synchronous tabular TD when rewards and transition probabilities change. This application requires a simulator that can sample from every state and assumes that all transition kernels share the same invariant distribution. When this distribution is known, we also show how additional simulator rollouts provide the target observations needed for stepsize adaptation and quantify their sampling cost. Appendix F extends the analysis to two coupled linear updates, each with its own stepsize, and gives tracking bounds under additional contraction assumptions.

5 Discussion and comparison with prior work

Our analysis builds on the deterministic tracking guarantee of Hadiji et al. (2026). For the affine updates considered here, Theorem 3 retains their initialization and target-movement terms and adds a noise term that accounts for temporal correlation. Earlier stochastic analyses already describe the tradeoff between noise reduction and responsiveness to change, including in nonlinear settings (Bastianello et al., 2024; Cutler et al., 2023). The deterministic affine structure allows us to separate the effects of initialization and target movement from accumulated observation noise. We use this separation to obtain a bound in terms of squared target variation and the covariance condition in Assumption 2. Table 1 in Appendix G summarizes the assumptions and guarantees of selected prior results.

The Gaussian benchmarks clarify when this bound is sharp. For independent observations, the dependence on the horizon and squared target variation agrees with classical results for estimating sequences with bounded squared increments (Sadhanala et al., 2016). Our formulation makes the dependence on the contraction strength and number of operator queries explicit. It also identifies variation ranges in which a suitably chosen constant stepsize attains the smallest possible worst-case error up to constant factors. The correlated Gaussian benchmark further shows that the $\tau^{2/3}$ dependence in the tuned bound is unavoidable in the matching range described in Section 4.3. These optimality statements concern the specified scalar Gaussian models and variation ranges.

The adaptive result addresses how to choose among stepsizes when the noise level and target variation are unknown. Adaptation to unknown variation is already possible in scalar forecasting: Baby & Wang (2019) obtain guarantees directly from noisy target observations, without a separate validation stream. Our method combines affine trackers using standard exponential weighting (Cesa-Bianchi & Lugosi, 2006) and requires additional bounded, unbiased observations of the current fixed point. Theorem 9 compares its cumulative expected error with that of the best fixed-stepsize tracker in the collection. The additional observations also support direct smoothing, as discussed in Section 4.4.

For policy evaluation, the sampling model is an essential distinction. Walton (2020) studies asynchronous TD and Q-learning using observations from a changing Markov trajectory. Our TD application assumes simulator access from every state and transition kernels that share a common invariant distribution. Under these conditions, we obtain cumulative value-estimation bounds and quantify the additional sampling cost of adaptation. Extending the analysis to observations from a single changing trajectory, and adapting the stepsize using operator observations alone, remain directions for further work.

AI use statement

Generative AI tools were used to help draft portions of the text, identify related literature, and prepare LaTeX. The authors remain responsible for independently checking every cited source, theorem statement, algebraic derivation, and proof, and for the final content of the manuscript. No experimental data were generated by AI.

Reproducibility statement

Sections 3–4 state the assumptions, algorithms, and constants for every main-text result and specify which observations are available when each prediction is formed. The appendix gives complete proofs and states the additional localization condition used for the nested fast–slow extension. The paper contains no experiments or datasets.

REFERENCES

Dheeraj Baby and Yu-Xiang Wang. Online forecasting of total-variation-bounded sequences. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.03364.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 359–409. PMLR, 2021. URL https://proceedings.mlr.press/v134/baby21a.html.

Nicola Bastianello, Liam Madden, Ruggero Carli, and Emiliano Dall'Anese. A stochastic operator framework for optimization and learning with sub-Weibull errors. *IEEE Transactions on Automatic Control*, 69(12):8722–8737, 2024.

Andrey Bernstein and Emiliano Dall'Anese. Asynchronous and distributed tracking of time-varying fixed points. In *2018 IEEE Conference on Decision and Control*, pp. 3236–3243. IEEE, 2018.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015. doi: 10.1287/opre.2015.1408.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021.

Guido Carnevale, Nicola Bastianello, Luca Schenato, Giuseppe Notarstefano, and Ruggero Carli. Timescale separation through the lens of operator theory. *arXiv preprint arXiv:2608.02443*, 2026.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

Siddharth Chandak. Concentration and mean-square bounds for contractive stochastic approximation: A unified elementary approach. *arXiv preprint arXiv:2607.17595*, 2026.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8223–8234, 2020.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1843–1854. PMLR, 2020.

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under distributional drift. *Journal of Machine Learning Research*, 24(147):1–56, 2023.

Jelena Diakonikolas. Solving stochastic fixed-point equations with high probability. *arXiv preprint arXiv:2607.09097*, 2026.

Songtao Feng, Ming Yin, Ruiquan Huang, Yu-Xiang Wang, Jing Yang, and Yingbin Liang. Non-stationary reinforcement learning under general function approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9976–10007. PMLR, 2023.

Hédi Hadiji, Sarah Sachs, and Cristóbal Guzmán. Tracking solutions of time-varying variational inequalities. *Mathematics of Operations Research*, 2026. doi: 10.1287/moor.2024.0571. Articles in Advance.

Shaan Ul Haque, Sajad Khodadadian, and Siva Theja Maguluri. Tight finite time bounds of two-time-scale linear stochastic approximation with Markovian noise. *arXiv preprint arXiv:2401.00364*, 2024.

Neharika Jali, Eshika Pathak, Pranay Sharma, Guannan Qu, and Gauri Joshi. Natural policy gradient for average reward non-stationary reinforcement learning. *Transactions on Machine Learning Research*, 2026.

Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2144–2203. PMLR, 2020.

Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J. Wainwright, and Michael I. Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040, 2021. doi: 10.1137/20M1331524.

Vijay R. Konda and John N. Tsitsiklis. Linear stochastic approximation driven by slowly varying Markov chains. *Systems & Control Letters*, 50(2):95–102, 2003.

Bhumesh Kumar, Vivek S. Borkar, and Akhil Shetty. Non-asymptotic error bounds for constant stepsize stochastic approximation for tracking mobile agents. *Mathematics of Control, Signals, and Systems*, 31(4):589–614, 2019.

Jeongyeol Kwon, Luke Dotson, Yudong Chen, and Qiaomin Xie. Two-timescale linear stochastic approximation: Constant stepsizes go a long way. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3781–3789. PMLR, 2025.

Liam Madden, Stephen Becker, and Emiliano Dall'Anese. Bounds for the tracking error of first-order online optimization methods. *Journal of Optimization Theory and Applications*, 189(2):437–457, 2021. doi: 10.1007/s10957-021-01836-9.

Vagul Mahadevan, Claire Chen, Shuze Daniel Liu, and Shangtong Zhang. Convergence of two-timescale Markovian stochastic approximations with applications in reinforcement learning. *arXiv preprint arXiv:2605.31172*, 2026. Accepted at ICML 2026.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Başar. Near-optimal model-free reinforcement learning in non-stationary episodic MDPs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7447–7458. PMLR, 2021.

Jean Opsomer, Yuedong Wang, and Yuhong Yang. Nonparametric regression with correlated errors. *Statistical Science*, 16(2):134–153, 2001. doi: 10.1214/ss/1009213287.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J. Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL `https://arxiv.org/abs/1605.08400`.

Sharan Sahu, Cameron J. Hogan, and Martin T. Wells. On the provable suboptimality of momentum SGD in nonstationary stochastic optimization. *arXiv preprint arXiv:2601.12238*, 2026. Accepted at ICML 2026.

Dhruv Sarkar and Vaneet Aggarwal. Non-expansive two-time-scale stochastic approximation: A fixed-schedule one-quarter barrier and bias-corrected acceleration. *arXiv preprint arXiv:2607.13414*, 2026a.

Dhruv Sarkar and Vaneet Aggarwal. Nonlinear two-time-scale stochastic approximation: A sharp phase transition and how to beat it. *arXiv preprint arXiv:2606.14488*, 2026b.

Simone Totaro and Anders Jonsson. Fast stochastic Kalman gradient descent for reinforcement learning. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pp. 1118–1129. PMLR, 2021.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Neil Walton. An adiabatic theorem for policy tracking with TD-learning. *arXiv preprint arXiv:2010.12848*, 2020.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4300–4354. PMLR, 2021.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 449–457. PMLR, 2016. URL https://proceedings.mlr.press/v48/yangb16.html.

Lijun Zhang, Tianbao Yang, Jinfeng Yi, Rong Jin, and Zhi-Hua Zhou. Improved dynamic regret for non-degenerate functions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1608.03933.

Huozhi Zhou, Jinglin Chen, Lav R. Varshney, and Ashish Jagmohan. Nonstationary reinforcement learning with linear function approximation. *Transactions on Machine Learning Research*, 2022.

APPENDIX ROADMAP

The appendix begins with preliminary lemmas and the proof of the affine tracking upper bound, including the tuned-stepsize guarantee. It then proves the Gaussian lower bounds for independent and temporally correlated noise and establishes the ranges in which they match the tracking upper bounds. Appendix D proves the adaptive guarantee and the bound for direct smoothing of validation observations. Appendix E presents the synchronous tabular TD application, constructs unbiased target-validation observations, and gives the associated proofs and sampling costs. Appendix F develops tracking bounds for linear recursions with two stepsizes, covering both a stacked contraction and a nested fast–slow system. Finally, Appendix G provides a table comparing our assumptions and guarantees with selected prior results.

A PRELIMINARY LEMMAS

We first collect the deterministic filter bounds, the covariance bound, and the exponential-weighting inequality used later.

Lemma 10 (Stable products)**.** *Let $K_1, \dots, K_{T-1}$ be deterministic and satisfy $\|K_t\|_{\mathrm{op}} \le \rho < 1$. For $t > s$, define $\Phi(t,s) = K_{t-1} \cdots K_s$, and let $\Phi(s,s) = I$. Then $\|\Phi(t,s)\|_{\mathrm{op}} \le \rho^{t-s}$ and*

$$\sum_{t=1}^{T} \|\Phi(t,1)u\|^2 \le \frac{\|u\|^2}{1-\rho^2}.$$

Proof. Submultiplicativity gives $\|\Phi(t,s)\|_{\mathrm{op}} \le \rho^{t-s}$. Hence

$$\sum_{t=1}^{T} \|\Phi(t,1)u\|^2 \le \|u\|^2 \sum_{k=0}^{T-1} \rho^{2k} \le \frac{\|u\|^2}{1-\rho^2}.$$

□

Lemma 11 (Deterministic contraction tracking)**.** *Suppose $d_{t+1} = K_t d_t - u_t$ and $\|K_t\|_{\mathrm{op}} \le \rho < 1$ for $t = 1, \dots, T-1$. Then*

$$\sum_{t=1}^{T} \|d_t\|^2 \le \frac{\|d_1\|^2}{1-\rho} + \frac{1}{(1-\rho)^2} \sum_{t=1}^{T-1} \|u_t\|^2.$$

Proof. For nonnegative a, b,

$$(\rho a + b)^2 \le \rho a^2 + \frac{b^2}{1-\rho},$$

because the right-hand side minus the left-hand side equals $\rho(\sqrt{1-\rho}\, a - b/\sqrt{1-\rho})^2$. Therefore

$$\|d_{t+1}\|^2 \le \rho \|d_t\|^2 + \frac{\|u_t\|^2}{1-\rho}.$$

Let $S_T = \sum_{t=1}^{T} \|d_t\|^2$ and $U_T = \sum_{t=1}^{T-1} \|u_t\|^2$. Summing the preceding inequality for $t = 1, \dots, T-1$ gives

$$S_T - \|d_1\|^2 \le \rho\big(S_T - \|d_T\|^2\big) + \frac{U_T}{1-\rho}.$$

Dropping the nonpositive term $-\rho \|d_T\|^2$, rearranging, and dividing by $1-\rho$ prove the claim. □

Lemma 12 (Stable convolution)**.** *For vectors $d_1, \dots, d_{T-1}$ and $\rho \in [0,1)$,*

$$\sum_{t=2}^{T} \left(\sum_{s=1}^{t-1} \rho^{t-1-s} \|d_s\| \right)^2 \le \frac{1}{(1-\rho)^2} \sum_{s=1}^{T-1} \|d_s\|^2 .$$

Consequently, under the hypotheses of Lemma 10,

$$\sum_{t=2}^{T} \left\| \sum_{s=1}^{t-1} \Phi(t, s+1) d_s \right\|^2 \le \frac{1}{(1-\rho)^2} \sum_{s=1}^{T-1} \|d_s\|^2 .$$

Proof. Extend $a_s = \|d_s\|$ by zero outside $\{1, \dots, T-1\}$ and let $k_j = \rho^j \mathbf{1}\{j \ge 0\}$. The first left-hand side is at most $\|k * a\|_{\ell_2}^2$. Young's convolution inequality gives $\|k * a\|_{\ell_2} \le \|k\|_{\ell_1} \|a\|_{\ell_2} = (1-\rho)^{-1} \|a\|_{\ell_2}$. For the matrix statement, use $\|\Phi(t, s+1) d_s\| \le \rho^{t-1-s} \|d_s\|$ and apply the scalar bound. □

Lemma 13 (Accumulated oracle error)**.** *Under Assumption 2 and the hypotheses of Lemma 10,*

$$\sum_{t=2}^{T} \mathbb{E} \left\| \sum_{s=1}^{t-1} \Phi(t, s+1) \xi_{s+1} \right\|^2 \le \frac{\tau \sigma^2 (T-1)}{1-\rho^2}.$$

Proof. Fix t. Apply equation 2 with $C_s = \Phi(t, s+1)$ for $s < t$ and $C_s = 0$ for $s \ge t$. Lemma 10 gives

$$\mathbb{E} \left\| \sum_{s=1}^{t-1} \Phi(t, s+1) \xi_{s+1} \right\|^2 \le \tau \sigma^2 \sum_{s=1}^{t-1} \rho^{2(t-1-s)} \le \frac{\tau \sigma^2}{1-\rho^2}.$$

There are $T-1$ nonzero prediction-time contributions, so summing over $t = 2, \dots, T$ proves the result. □

Proposition 14 (Geometric centered correlation implies integrated covariance)**.** *Suppose $\mathbb{E}[\xi_{t+1}] = 0$ for every $t = 1, \dots, T-1$ and, for all deterministic linear maps $C_s, C_r : \mathbb{R}^d \to \mathbb{R}^d$,*

$$|\mathbb{E} \langle C_s \xi_{s+1}, C_r \xi_{r+1} \rangle| \le \sigma^2 \chi_{\mathrm{corr}}^{|s-r|} \|C_s\|_{\mathrm{op}} \|C_r\|_{\mathrm{op}}, \qquad \chi_{\mathrm{corr}} \in [0,1).$$

Then Assumption 2 holds with $\tau = (1+\chi_{\mathrm{corr}})/(1-\chi_{\mathrm{corr}})$.

Proof. Taking $s = r$ and $C_s = I$ gives $\mathbb{E} \|\xi_{s+1}\|^2 \le \sigma^2$. For arbitrary deterministic linear maps, let $a_s = \|C_s\|_{\mathrm{op}}$. Expanding the square gives

$$\mathbb{E} \left\| \sum_{s} C_s \xi_{s+1} \right\|^2 \le \sigma^2 \sum_{s,r} a_s a_r \chi_{\mathrm{corr}}^{|s-r|}.$$

The convolution kernel $h_k = \chi_{\mathrm{corr}}^{|k|}$ has ℓ_1 norm $(1+\chi_{\mathrm{corr}})/(1-\chi_{\mathrm{corr}})$, which upper bounds its ℓ_2 operator norm. The last display is therefore at most $\tau \sigma^2 \sum_s a_s^2$. □

Lemma 15 (Exp-concavity of squared loss)**.** *Let $B > 0$, let $\mathcal{X}$ be convex and contained in a ball of radius B, and let $y \in \mathcal{X}$. For every $0 < \lambda \le 1/(8B^2)$, the map $x \mapsto \exp(-\lambda \|x - y\|^2)$ is concave on $\mathcal{X}$. Consequently, for $x_j \in \mathcal{X}$ and a probability vector p,*

$$\left\| \sum_j p_j x_j - y \right\|^2 \le -\lambda^{-1} \log \sum_j p_j \exp(-\lambda \|x_j - y\|^2).$$

Proof. Because the norm is induced by an inner product, choose coordinates in which it is Euclidean. For $f(x) = \exp(-\lambda \|x - y\|^2)$,

$$\nabla^2 f(x) = f(x) \left(4\lambda^2 (x-y)(x-y)^\top - 2\lambda I\right).$$

Since $\|x - y\| \le 2B$, the largest eigenvalue of the rank-one term is at most $16\lambda^2 B^2$. Hence $\nabla^2 f(x) \preceq 0$ whenever $0 < \lambda \le 1/(8B^2)$. Jensen's inequality for f, followed by $-\lambda^{-1}\log(\cdot)$, gives the result. □

B Proof of the affine tracking upper bound

We combine the deterministic tracking lemma with the covariance bound for the propagated noise.

Proof of Theorem 3. Let $e_t = x_t - x_t^*$ and $\Delta_t = x_{t+1}^* - x_t^*$. Since $F_t(x_t^*) = x_t^*$,

$$\begin{aligned} e_{t+1} &= (1-\eta)(x_t - x_t^*) + \eta\big(F_t(x_t) - F_t(x_t^*)\big) + \eta\xi_{t+1} - \Delta_t \\ &= K_t e_t + \eta\xi_{t+1} - \Delta_t, \qquad K_t = (1-\eta)I + \eta M_t. \end{aligned}$$

Assumption 1 gives $\|K_t\|_{\mathrm{op}} \le (1-\eta) + \eta(1-\kappa) = 1 - \eta\kappa =: \rho < 1$.

Split $e_t = d_t + n_t$, where

$$d_1 = e_1, \quad n_1 = 0, \qquad d_{t+1} = K_t d_t - \Delta_t, \qquad n_{t+1} = K_t n_t + \eta\xi_{t+1}.$$

All K_t, d_t, and Δ_t are deterministic. Marginal centering in Assumption 2 gives $\mathbb{E}[n_t] = 0$, and therefore

$$\mathbb{E}\|e_t\|^2 = \|d_t\|^2 + \mathbb{E}\|n_t\|^2.$$

Lemma 11 yields

$$\sum_{t=1}^{T} \|d_t\|^2 \le \frac{\|e_1\|^2}{1-\rho} + \frac{W_T}{(1-\rho)^2}.$$

Moreover,

$$n_t = \eta \sum_{s=1}^{t-1} \Phi(t, s+1)\xi_{s+1}.$$

Lemma 13 gives

$$\sum_{t=1}^{T} \mathbb{E}\|n_t\|^2 \le \frac{\eta^2\tau\sigma^2(T-1)}{1-\rho^2} \le \frac{\eta^2\tau\sigma^2 T}{1-\rho^2}.$$

Combining the deterministic and stochastic bounds and using $1 - \rho = \eta\kappa$ and $1 - \rho^2 = \eta\kappa(2 - \eta\kappa)$ proves equation 3. □

Proof of Corollary 4. When $x_1 = x_1^*$, Theorem 3 and $2 - \eta\kappa \ge 1$ give

$$\mathcal{E}_T \le a\eta + b\eta^{-2}, \qquad a = \frac{\tau\sigma^2 T}{\kappa}, \quad b = \frac{W_T}{\kappa^2}.$$

The unconstrained minimizer of this simpler bound is $(2b/a)^{1/3} = \eta_{\mathrm{bal}}$. Substituting it gives

$$a\eta_{\mathrm{bal}} + b\eta_{\mathrm{bal}}^{-2} = (2^{1/3} + 2^{-2/3}) a^{2/3} b^{1/3},$$

which is the claimed bound. □

C Proof of the Gaussian lower bound

Proof of Theorem 5. We give the full testing argument to enforce $\theta_1 = 0$ and track the finite-horizon range explicitly. We proceed in three steps. First, the query locations are removed and the q observations at each round are reduced to a sufficient statistic. Second, we construct signed triangular signals that satisfy the path budget while keeping neighboring hypotheses difficult to distinguish. Third, a testing reduction converts sign errors into squared estimation error.

C.1 Reduction to a Gaussian sequence experiment

Let $\mathsf{O}_{t,i}(x)$ denote the ith response to a query x at round t. A possibly adaptive query $x_{t,i}$ produces

$$Z_{t,i} := \frac{\mathsf{O}_{t,i}(x_{t,i}) - (1-\kappa)x_{t,i}}{\kappa} = \theta_t + \frac{\varepsilon_{t,i}}{\kappa}.$$

Thus the query location disappears after transformation. Conditional on the learner's internal randomization, the transformed observations have the same product Gaussian law for every adaptive query strategy; the query record is a function of past transformed observations and carries no additional information about θ. Granting exactly q queries at every round can only help the learner.

Lemma 16 (Sufficiency of the per-round sample mean)**.** *For each t, let $Z_{t,1}, \dots, Z_{t,q}$ be independent $\mathcal{N}(\theta_t, \sigma^2/\kappa^2)$ observations. For squared-error estimation of $\theta_{1:T}$, the full experiment is statistically equivalent to observing only*

$$\bar{Z}_t = q^{-1}\sum_{i=1}^{q} Z_{t,i} \sim \mathcal{N}(\theta_t, \nu^2), \qquad \nu^2 = \frac{\sigma^2}{q\kappa^2}.$$

Every estimator based on the full samples can therefore be replaced by one based on $\bar{Z}_{1:T}$ with no larger risk.

Proof. For fixed t, apply an orthogonal transformation whose first coordinate is $\sqrt{q}\,\bar{Z}_t$. The remaining $q-1$ coordinates are centered residual contrasts. Joint Gaussianity and orthogonality imply that these residuals are independent of $\bar{Z}_t$, and their law does not depend on θ_t. The same decomposition applies independently across rounds. The residuals therefore contain no information about θ. Conditioning any estimator on $\bar{Z}_{1:T}$ and using Jensen's inequality gives an estimator based only on the sample means with no larger squared risk. □

It remains to lower bound the offline Gaussian sequence model $\bar{Z}_t \sim \mathcal{N}(\theta_t, \nu^2)$.

C.2 Triangular packing

Let $m = 2r+1$ be odd with $r \geq 2$, and define $\psi \in \mathbb{R}^m$ by

$$\psi_i = \frac{\min\{i-1, m-i\}}{r}, \qquad i = 1, \dots, m.$$

Thus $\psi_1 = \psi_m = 0$ and the center equals one.

Lemma 17 (Triangle geometry)**.** *For $m = 2r+1$ with $r \geq 2$,*

$$\frac{m}{6} \leq \|\psi\|^2 \leq m, \qquad \sum_{i=1}^{m-1}(\psi_{i+1} - \psi_i)^2 = \frac{2}{r} \leq \frac{8}{m}.$$

Proof. The nonzero values are $1/r, 2/r, \dots, (r-1)/r, 1, (r-1)/r, \dots, 1/r$, so

$$\|\psi\|^2 = 1 + \frac{2}{r^2}\sum_{j=1}^{r-1} j^2 = 1 + \frac{(r-1)(2r-1)}{3r}.$$

For $r \geq 2$, this lies between $(2r+1)/6 = m/6$ and m. Each of the $2r$ nonzero increments has magnitude $1/r$, so the derivative energy is $2/r$. Finally, $m = 2r+1 \leq 4r$ implies $2/r \leq 8/m$. □

Partition the first Km indices into $K = \lfloor T/m \rfloor$ consecutive blocks $\mathcal{I}_j = \{(j-1)m+1, \dots, jm\}$. Let $\psi^{(j)} \in \mathbb{R}^T$ place a copy of ψ on $\mathcal{I}_j$ and zero elsewhere. For $u \in \{-1, +1\}^K$, set

$$\theta^u = a\sum_{j=1}^{K} u_j \psi^{(j)}, \qquad a^2 = \frac{\nu^2}{32m}.$$

Every packing element has $\theta^u_1 = 0$.

Lemma 18 (Budget and neighboring information)**.** *The packing satisfies*

$$\sum_{t=1}^{T-1}(\theta_{t+1}^{u}-\theta_{t}^{u})^{2}\le\frac{\nu^{2}T}{4m^{3}}.$$

If u and u' differ in one coordinate, then the Kullback–Leibler divergence between the corresponding Gaussian observation laws is at most $1/16$.

Proof. The blocks start and end at zero, so no cross-block increment appears. Lemma 17 gives

$$\sum_{t=1}^{T-1}(\theta_{t+1}^{u}-\theta_{t}^{u})^{2}=Ka^{2}\sum_{i=1}^{m-1}(\psi_{i+1}-\psi_{i})^{2}\le\frac{T}{m}\frac{\nu^{2}}{32m}\frac{8}{m}=\frac{\nu^{2}T}{4m^{3}}.$$

Neighboring mean vectors differ by $2a\psi^{(j)}$, and therefore

$$D_{\mathrm{KL}}(\mathbb{P}_{u}\|\mathbb{P}_{u'})=\frac{\left\|2a\psi^{(j)}\right\|^{2}}{2\nu^{2}}=\frac{2a^{2}\left\|\psi\right\|^{2}}{\nu^{2}}\le\frac{1}{16}.$$

□

C.3 Testing reduction and completion

For an estimator $\widehat{\theta}$, infer the sign on block j by $\widehat{u}_j=\operatorname{sign}\left\langle\widehat{\theta}_{\mathcal{I}_j},\psi\right\rangle$, with an arbitrary tie rule. If $\widehat{u}_j\neq u_j$, then

$$\left\|\widehat{\theta}_{\mathcal{I}_j}-u_j a\psi\right\|^{2}\ge a^{2}\left\|\psi\right\|^{2}.$$

Indeed, after expanding the square, the sign error makes the cross term nonnegative when moved to the right-hand side.

Place the uniform prior on the sign vector. The next step is a standard binary-testing argument (Tsybakov, 2009). Holding all other signs fixed, the minimum average error probability for the remaining sign is $(1-\mathrm{TV}(\mathbb{P}_{+},\mathbb{P}_{-}))/2$, where TV denotes total-variation distance. Pinsker's inequality and Lemma 18 give $\mathrm{TV}(\mathbb{P}_{+},\mathbb{P}_{-})\le\sqrt{1/32}<1/4$, so the average sign-error probability is at least $1/4$. Averaging over signs and summing over blocks yields

$$2^{-K}\sum_{u}\mathbb{E}_{u}\left\|\widehat{\theta}-\theta^{u}\right\|^{2}\ge\frac{K}{4}a^{2}\left\|\psi\right\|^{2}\ge\frac{K\nu^{2}}{768}.$$

Set $m_0=(\nu^2T/(4W))^{1/3}$ and choose m as the smallest odd integer at least $\max\{5,m_0\}$. Under the range in Theorem 5, $2\le m_0\le T/4$. If $m_0\le 5$, then $m=5\le T/2$; if $m_0>5$, then $m\le m_0+2\le T/2$. In both cases $m\le 3m_0$ and $K=\lfloor T/m\rfloor\ge T/(2m)$. Lemma 18 and $m\ge m_0$ imply $\theta^u\in\Theta_T(W)$. Therefore

$$\begin{aligned}\sup_{\theta\in\Theta_T(W)}\mathbb{E}_{\theta}\left\|\widehat{\theta}-\theta\right\|^{2}&\ge\frac{\nu^{2}T}{1536m}\ge\frac{\nu^{2}T}{4608m_0}\\&=\frac{4^{1/3}}{4608}(\nu^{2}T)^{2/3}W^{1/3}\ge\frac{1}{4096}(\nu^{2}T)^{2/3}W^{1/3}.\end{aligned}$$

Taking the infimum over estimators and substituting $\nu^2=\sigma^2/(q\kappa^2)$ proves Theorem 5 with $c=1/4096$. □

Proof of Corollary 6. Averaging the q oracle responses gives additive noise variance σ^2/q. For every $\theta\in\Theta_T(W)$, Theorem 3, $x_1=\theta_1=0$, and $2-\eta\kappa\ge 1$ give

$$\mathbb{E}_{\theta}\sum_{t=1}^{T}(x_t-\theta_t)^{2}\le\frac{\eta\sigma^{2}T}{q\kappa}+\frac{W_T}{\eta^{2}\kappa^{2}}\le\frac{\eta\sigma^{2}T}{q\kappa}+\frac{W}{\eta^{2}\kappa^{2}}.$$

Take $\eta=\eta_{q,\mathrm{bal}}$. Substitution, as in the proof of Corollary 4, gives the claimed rate with leading constant $2^{1/3}+2^{-2/3}<2$, uniformly over $\Theta_T(W)$. In terms of ν^2, the condition $\eta_{q,\mathrm{bal}}\le 1$ is $W\le\kappa^3\nu^2T/2$. Intersecting it with the lower-bound range gives the displayed overlap. For $T\ge 16$, the lower endpoint does not exceed $\nu^2T/32$; it does not exceed $\kappa^3\nu^2T/2$ exactly when $\kappa T\ge 32^{1/3}$. □

C.4 Lower bound with correlated Gaussian noise

Proof of Theorem 7. We first verify the covariance condition and compute the information between two target sequences. We then reuse the triangular signals from the independent-noise proof, with an amplitude rescaled to account for correlation.

Let $R_\chi \in \mathbb{R}^{T\times T}$ have entries $(R_\chi)_{st} = \chi^{|s-t|}$. For any deterministic scalar coefficients c_t, the inequality $2|c_s c_t| \le c_s^2 + c_t^2$ gives

$$\sum_{s,t} c_s c_t \chi^{|s-t|} \le \left(1 + 2\sum_{h=1}^{\infty} \chi^h\right) \sum_t c_t^2 = \tau \sum_t c_t^2.$$

This proves Assumption 2 for each stream and, by independence across streams, for their mean with variance parameter σ^2/q.

The query transformation from Section 4.2 gives q independent Gaussian vectors with mean θ and covariance $(\sigma^2/\kappa^2)R_\chi$. Grant all these vectors to the estimator. If two means differ by v, their joint Kullback–Leibler divergence is

$$D_{\mathrm{KL}} = \frac{v^\top R_\chi^{-1} v}{2\nu^2}.$$

To compute the quadratic form, a Gaussian vector with covariance R_χ can be generated by $Z_1 \sim \mathcal{N}(0,1)$ and $Z_t = \chi Z_{t-1} + \sqrt{1-\chi^2}\, U_t$, with independent standard normal innovations. Its density therefore gives

$$v^\top R_\chi^{-1} v = v_1^2 + \frac{1}{1-\chi^2}\sum_{t=2}^{T} (v_t - \chi v_{t-1})^2.$$

For $v_1 = v_T = 0$, expansion yields the useful identity

$$v^\top R_\chi^{-1} v = \frac{1}{\tau}\|v\|_2^2 + \frac{\tau^2 - 1}{4\tau}\sum_{t=1}^{T-1} (v_{t+1} - v_t)^2.$$

Set $m_0 = (\tau\nu^2 T/(16W))^{1/3}$ and let m be the smallest odd integer at least $\max\{5, m_0\}$. The assumed variation range gives $2\tau \le m_0 \le T/4$. As in the independent-noise proof, $m \le 3m_0$, $m \le T/2$, and $K = \lfloor T/m \rfloor \ge T/(2m)$. In particular, $m \ge \tau$.

Use the same disjoint triangular blocks $\psi^{(j)}$, but set

$$\theta^u = a\sum_{j=1}^{K} u_j \psi^{(j)}, \qquad a^2 = \frac{\tau\nu^2}{128m}, \qquad u \in \{-1,+1\}^K.$$

Their endpoints vanish, and Lemma 17 gives

$$\sum_{t=1}^{T-1} (\theta^u_{t+1} - \theta^u_t)^2 \le \frac{8Ka^2}{m} \le \frac{\tau\nu^2 T}{16m^3} \le W.$$

Thus every signal belongs to $\Theta_T(W)$. Neighboring signals differ by $v = 2a\psi^{(j)}$, which has $v_1 = v_T = 0$. The precision identity and $m \ge \tau$ imply

$$D_{\mathrm{KL}}(\mathbb{P}_u \| \mathbb{P}_{u'}) \le \frac{2a^2}{\nu^2}\left(\frac{m}{\tau} + \frac{2\tau}{m}\right) \le \frac{6a^2 m}{\tau\nu^2} = \frac{3}{64} < \frac{1}{16}.$$

The binary-testing argument used above gives average sign-error probability at least $1/4$. Since $\|\psi\|_2^2 \ge m/6$, the resulting squared-error lower bound is

$$\begin{aligned}\sup_{\theta\in\Theta_T(W)} \mathbb{E}_\theta \left\|\widehat{\theta} - \theta\right\|_2^2 &\ge \frac{K}{4} a^2 \|\psi\|_2^2 \ge \frac{K\tau\nu^2}{3072} \ge \frac{\tau\nu^2 T}{18432 m_0} \\ &= \frac{16^{1/3}}{18432}(\tau\nu^2 T)^{2/3} W^{1/3} \ge \frac{1}{8192}(\tau\nu^2 T)^{2/3} W^{1/3}.\end{aligned}$$

Taking the infimum and substituting $\nu^2 = \sigma^2/(q\kappa^2)$ proves the theorem.

For the matching upper bound, the averaged oracle satisfies Assumption 2 with parameters $(\sigma^2/q, \tau)$. Apply Theorem 3 with $x_1 = 0$ and $W_T \le W$, then balance as in Corollary 4. The prescribed stepsize is at most one exactly when $W \le \kappa^3\tau\nu^2 T/2$. The lower endpoint $4\tau\nu^2/T^2$ is below this cap exactly when $\kappa T \ge 2$, and below $\nu^2 T/(128\tau^2)$ exactly when $T \ge 8\tau$. □

D PROOF OF THE ADAPTIVE RESULT

We first prove a deterministic aggregation inequality for the observable validation losses. Conditional unbiasedness then converts those losses into tracking errors, and comparison over the stepsize grid applies Theorem 3 to the best expert.

Proof of Theorem 9. Let $\ell_{t,j} = \|\widetilde{x}_{t,j} - Y_t\|^2$ and $\bar{\ell}_t = \|\bar{x}_t - Y_t\|^2$. Lemma 15 gives

$$\bar{\ell}_t \le -\lambda^{-1} \log \sum_j p_{t,j} e^{-\lambda \ell_{t,j}}.$$

For the analysis, define

$$Z_t = \sum_j p_{1,j} \exp\left(-\lambda \sum_{s=1}^{t-1} \ell_{s,j}\right), \qquad t = 1, \dots, T+1.$$

For every prediction time $t \le T$,

$$p_{t,j} = \frac{p_{1,j} \exp(-\lambda \sum_{s=1}^{t-1} \ell_{s,j})}{Z_t}.$$

Consequently, $\bar{\ell}_t \le -\lambda^{-1} \log(Z_{t+1}/Z_t)$. Here Z_{T+1} is only an analysis quantity; the final validation observation does not trigger another update. Summing over t and using the uniform prior yields the pathwise inequality

$$\sum_{t=1}^{T} \bar{\ell}_t \le \min_j \sum_{t=1}^{T} \ell_{t,j} + \lambda^{-1} \log(J+1).$$

Write $Y_t = x_t^* + \varepsilon_t^{\mathrm{val}}$. By equation 5, every $\mathcal{G}_t$-measurable prediction x satisfies

$$\mathbb{E}[\|x - Y_t\|^2 \mid \mathcal{G}_t] = \|x - x_t^*\|^2 + \mathbb{E}[\|\varepsilon_t^{\mathrm{val}}\|^2 \mid \mathcal{G}_t].$$

Take expectations in the pathwise inequality and use $\mathbb{E}[\min_j L_j] \le \min_j \mathbb{E}[L_j]$, where $L_j = \sum_{t=1}^{T} \ell_{t,j}$. The common validation-variance terms cancel, giving

$$\sum_{t=1}^{T} \mathbb{E}\,\|\bar{x}_t - x_t^*\|^2 \le \min_j \sum_{t=1}^{T} \mathbb{E}\,\|\widetilde{x}_{t,j} - x_t^*\|^2 + 8B^2 \log(J+1).$$

Since $x_t^* \in \mathcal{X}$, the projection inequality $\|\Pi_{\mathcal{X}}(x) - x_t^*\| \le \|x - x_t^*\|$ allows us to replace each projected prediction by its raw expert state. This proves the first inequality in equation 6.

Now fix $\eta \in [1/T, 1]$. The grid contains η_j such that $\eta/2 \le \eta_j \le \eta$. Since $u \mapsto u/(2 - \kappa u)$ is increasing on $(0, 1]$, Theorem 3 gives

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\,\|x_{t,j} - x_t^*\|^2 &\le \frac{2\,\|x_1 - x_1^*\|^2}{\eta\kappa} + \frac{\eta\tau\sigma^2 T}{\kappa(2 - \eta\kappa)} + \frac{4W_T}{\eta^2\kappa^2} \\ &\le 4\left\{\frac{\|x_1 - x_1^*\|^2}{\eta\kappa} + \frac{\eta\tau\sigma^2 T}{\kappa(2 - \eta\kappa)} + \frac{W_T}{\eta^2\kappa^2}\right\}.\end{aligned}$$

Taking the infimum over η proves the second inequality in equation 6. Under the stated initialization and stepsize conditions, Corollary 4 gives the final bound. □

Proof of the bound for direct smoothing. Let c be the center of a radius-B ball containing $\mathcal{X}$. Conditional unbiasedness gives

$$\mathbb{E}[\|Y_t - x_t^*\|^2 \mid \mathcal{G}_t] = \mathbb{E}[\|Y_t - c\|^2 \mid \mathcal{G}_t] - \|x_t^* - c\|^2 \le B^2.$$

Because $\mathcal{G}_t$ contains the full past, the errors $Y_t - x_t^*$ are martingale differences. The update $u_{t+1} = (1-a)u_t + aY_t$ is the relaxed fixed-point iteration for the deterministic constant map $u \mapsto x_t^*$, whose contraction gap is one. Theorem 3 therefore applies with stepsize a, noise parameters $(B^2, 1)$, and initialization $u_1 = x_1$, proving the displayed bound in Section 4.4. This argument uses only the validation stream. Substituting $a = \eta\kappa$ rewrites its noise term as $aB^2T/(2-a)$ and the operator tracker's noise term as $a(\tau\sigma^2/\kappa^2)T/(2-a)$, giving the stated comparison. □

E Application to tabular policy evaluation

E.1 Model and guarantees

Consider a deterministic sequence of discounted Markov reward processes $(\mathcal{S}, P_t, r_t, \gamma)$ on a finite state space, fixed before learning begins, with $\gamma \in [0, 1)$. Let the initial value vector v_1 be deterministic. At round t, the current MRP is held fixed while the learner queries a generative model. Conditional on all previous calls and algorithmic randomness, a call at state s returns (R, S') with

$$|R| \le R_{\max} \text{ a.s.}, \qquad \mathbb{E}[R] = r_t(s), \qquad \mathrm{Var}(R) \le \sigma_R^2, \qquad S' \sim P_t(\cdot \mid s).$$

All expectations and laws in this display are conditional on the past. Different calls are conditionally independent. The reward and next state from the same call may be dependent.

Assume every P_t preserves the same full-support probability distribution μ. This gives a single inner-product norm in which all mean Bellman maps contract; it restricts the allowed changes in the transition kernels. Define $\|v\|_\mu^2 = \sum_s \mu(s)v(s)^2$, $\|A\|_{\mu\to\mu} = \sup_{v\neq 0} \|Av\|_\mu / \|v\|_\mu$, and $v_t^* = (I - \gamma P_t)^{-1} r_t$. One call from each state defines a sampled Bellman operator

$$(\widehat{\mathcal{T}}_t v)(s) = R_{t,s} + \gamma v(S'_{t,s}),$$

and synchronous TD uses $v_{t+1} = (1-\eta)v_t + \eta\widehat{\mathcal{T}}_t v_t$.

Corollary 19 (Synchronous tabular TD)**.** *Let $V_{\max} = R_{\max}/(1-\gamma)$ and assume $\|v_1\|_\infty \le V_{\max}$. Define*

$$\sigma_{\mathrm{TD}}^2 := 2\sigma_R^2 + 2\gamma^2 V_{\max}^2, \qquad W_T^V := \sum_{t=1}^{T-1} \left\|v_{t+1}^* - v_t^*\right\|_\mu^2.$$

Then, for every $\eta \in (0, 1]$,

$$\sum_{t=1}^{T} \mathbb{E}\,\|v_t - v_t^*\|_\mu^2 \le \frac{\|v_1 - v_1^*\|_\mu^2}{\eta(1-\gamma)} + \frac{\eta\sigma_{\mathrm{TD}}^2 T}{(1-\gamma)(2-\eta(1-\gamma))} + \frac{W_T^V}{\eta^2(1-\gamma)^2}.$$

Moreover, with $W_T^r = \sum_{t=1}^{T-1} \|r_{t+1} - r_t\|_\mu^2$ and $W_T^P = \sum_{t=1}^{T-1} \|P_{t+1} - P_t\|_{\mu\to\mu}^2$,

$$W_T^V \le \frac{2}{(1-\gamma)^2}\left(W_T^r + \gamma^2 V_{\max}^2 W_T^P\right).$$

The common invariant distribution ensures $\|P_t v\|_\mu \le \|v\|_\mu$. Standard synchronous TD does not need to know μ; the adaptive mixture does because it scores predictions in $L_2(\mu)$. To obtain an unbiased target observation, we sample a reward at a random geometric horizon: its sampling probability supplies the discount factor, and rescaling removes the geometric normalization.

Proposition 20 (An unbiased validation vector)**.** *Fix a round and keep its MRP unchanged. For each starting state s, draw an independent copy of L with $\mathbb{P}(L = \ell) = (1-\gamma)\gamma^\ell$, follow the current transition kernel for L steps, and make one final generative-model call at the terminal state. If R^{val} is the reward from that call, set $Y_t(s) = R^{\mathrm{val}}/(1-\gamma)$. Using independent rollouts for different states and for the update samples gives $\mathbb{E}[Y_t \mid \mathcal{G}_t] = v_t^*$ and $\|Y_t\|_\mu \le V_{\max}$ almost surely. The expected cost is $|\mathcal{S}|/(1-\gamma)$ calls per round.*

Corollary 21 (Adaptive synchronous TD)**.** *Under the conditions of Corollary 19, assume μ is known and $R_{\max} > 0$. Run Algorithm 1 in $L_2(\mu)$ with $B = V_{\max}$ and $\mathcal{X} = [-V_{\max}, V_{\max}]^{|\mathcal{S}|}$, use one sampled Bellman operator for all experts at each update, and use Proposition 20 for validation. Then*

$$\sum_{t=1}^{T} \mathbb{E}\,\|\bar{v}_t - v_t^*\|_\mu^2 \le 4 \inf_{\eta\in[1/T,1]} \left\{ \frac{\|v_1 - v_1^*\|_\mu^2}{\eta(1-\gamma)} + \frac{\eta\sigma_{\mathrm{TD}}^2 T}{(1-\gamma)(2-\eta(1-\gamma))} + \frac{W_T^V}{\eta^2(1-\gamma)^2} \right\} + 8V_{\max}^2 \log(1 + \lceil \log_2 T \rceil).$$

The update samples require $(T-1)|\mathcal{S}|$ calls. Validation adds an expected $T|\mathcal{S}|/(1-\gamma)$ calls, a factor $1/(1-\gamma)$ more per prediction round than a synchronous update. This is the sampling cost of adaptation under the proposed validation construction.

Direct smoothing uses the same validation rollouts, with $B = V_{\max}$ and $W_T = W_T^V$, and saves the $(T-1)|\mathcal{S}|$ TD update calls. It is therefore a natural baseline under this feedback model. Corollary 21 establishes comparison with TD experts; the present results do not establish a sampling advantage over direct smoothing.

E.2 Proofs

Proof of Corollary 19. Because $\mu P_t = \mu$, Jensen's inequality gives, for every v,

$$\begin{aligned}\|P_t v\|_\mu^2 &= \sum_s \mu(s) \left(\sum_{s'} P_t(s' \mid s) v(s') \right)^2 \\ &\le \sum_s \mu(s) \sum_{s'} P_t(s' \mid s) v(s')^2 = \sum_{s'} \mu(s') v(s')^2 = \|v\|_\mu^2 .\end{aligned}$$

Thus the mean Bellman map $\mathcal{T}_t v = r_t + \gamma P_t v$ is a γ-contraction in $L_2(\mu)$.

The box $[-V_{\max}, V_{\max}]^{|\mathcal{S}|}$ is invariant. Indeed, if $\|v\|_\infty \le V_{\max}$, then $|R_{t,s} + \gamma v(S'_{t,s})| \le R_{\max} + \gamma V_{\max} = V_{\max}$, and the relaxed update is a convex combination of two points in the box. Since $|r_t(s)| \le R_{\max}$, the fixed point also satisfies $\|v_t^*\|_\infty \le V_{\max}$.

Let $\xi_{t+1} = \widehat{\mathcal{T}}_t v_t - \mathcal{T}_t v_t$. Conditional on the pre-update history, each coordinate has mean zero. Moreover,

$$\begin{aligned}\mathbb{E}[\|\xi_{t+1}\|_\mu^2 \mid \mathcal{F}_t] &\le 2 \sum_s \mu(s) \mathbb{E}[(R_{t,s} - r_t(s))^2 \mid \mathcal{F}_t] \\ &\quad + 2\gamma^2 \sum_s \mu(s) \mathbb{E}[(v_t(S'_{t,s}) - (P_t v_t)(s))^2 \mid \mathcal{F}_t] \\ &\le 2\sigma_R^2 + 2\gamma^2 \sum_s \mu(s) \mathbb{E}[v_t(S'_{t,s})^2 \mid \mathcal{F}_t] \\ &= 2\sigma_R^2 + 2\gamma^2 \|v_t\|_\mu^2 \le \sigma_{\mathrm{TD}}^2.\end{aligned}$$

Theorem 3 with $\kappa = 1-\gamma$ and $\tau = 1$ proves the tracking bound.

To relate value-function movement to changes in rewards and transitions, subtract the two Bellman equations to obtain

$$(I - \gamma P_{t+1})(v_{t+1}^* - v_t^*) = (r_{t+1} - r_t) + \gamma(P_{t+1} - P_t) v_t^*.$$

Since $\|P_{t+1}\|_{\mu\to\mu} \le 1$, the Neumann series gives $\left\|(I - \gamma P_{t+1})^{-1}\right\|_{\mu\to\mu} \le (1-\gamma)^{-1}$. Therefore

$$\left\|v_{t+1}^* - v_t^*\right\|_\mu \le \frac{\|r_{t+1} - r_t\|_\mu + \gamma \|P_{t+1} - P_t\|_{\mu\to\mu} \|v_t^*\|_\mu}{1-\gamma}.$$

Use $\|v_t^*\|_\mu \le V_{\max}$, square with $(a+b)^2 \le 2a^2 + 2b^2$, and sum over $t = 1, \dots, T-1$. □

Proof of Proposition 20. Fix $s \in \mathcal{S}$. Conditional on $\mathcal{G}_t$, the round-t MRP is deterministic and remains frozen throughout the rollout. The horizon $L_{t,s}$ is independent of all rollout samples. Conditional on

$L_{t,s} = \ell$, the final call is made at $S_\ell^{t,s}$, and its reward has conditional mean $r_t(S_\ell^{t,s})$. Hence

$$\mathbb{E}[Y_t(s) \mid \mathcal{G}_t] = \frac{1}{1-\gamma} \sum_{\ell=0}^{\infty} (1-\gamma)\gamma^\ell \mathbb{E}[R_{t,s}^{\mathrm{val}} \mid L_{t,s} = \ell, S_0^{t,s} = s, \mathcal{G}_t]$$
$$= \sum_{\ell=0}^{\infty} \gamma^\ell \mathbb{E}[r_t(S_\ell^{t,s}) \mid S_0^{t,s} = s] = v_t^*(s).$$

The reward bound gives $|Y_t(s)| \le R_{\max}/(1-\gamma) = V_{\max}$ almost surely, so $\|Y_t\|_\mu \le V_{\max}$. Finally, $\mathbb{E}[L_{t,s}+1] = 1/(1-\gamma)$. Summing over states gives the per-round expected call count, and summing over the T validation rounds gives the total expected count. Conditional independence from the update-map calls verifies the independence requirement in Assumption 8. □

Proof of Corollary 21. Every raw TD expert remains in $\mathcal{X}$, so projection is redundant but harmless. One synchronous collection of statewise samples defines a random affine Bellman map that can be reused at all expert value vectors. For each fixed expert, Corollary 19 supplies the tracking bound with $\kappa = 1-\gamma$, $\tau = 1$, and noise parameter σ_{TD}^2. Proposition 20 supplies a validation vector in the radius-$V_{\max}$ ball of $L_2(\mu)$. Substitution into Theorem 9 proves the displayed inequality. The sample and arithmetic costs follow directly from the synchronous-map and validation constructions. □

F Linear recursions with two stepsizes

Finite-time analyses of linear two-timescale stochastic approximation include Kaledin et al. (2020); Haque et al. (2024); Kwon et al. (2025). Recent work treats nonlinear recursions, Markovian noise, nonexpansive maps, and operator-based timescale separation (Mahadevan et al., 2026; Sarkar & Aggarwal, 2026b;a; Carnevale et al., 2026). The results below address a narrower question: how target movement enters cumulative tracking error once contraction is available. The first result assumes a common quadratic contraction for the stacked update. The second uses an autonomous slow recursion and separately bounds the movement it induces in the fast target.

F.1 A stacked contraction bound

Let $x_t \in \mathbb{R}^{d_x}$ and $y_t \in \mathbb{R}^{d_y}$ have deterministic initial values. We write $\|\cdot\|_2$ for the Euclidean component norms and measure stacked tracking error in the quadratic norm defined below. For constant stepsizes $0 < \beta \le \alpha$, consider

$$x_{t+1} = x_t + \alpha(A_{11,t}x_t + A_{12,t}y_t + c_{1,t} + \xi_{t+1}),$$
$$y_{t+1} = y_t + \beta(A_{21,t}x_t + A_{22,t}y_t + c_{2,t} + \zeta_{t+1}),$$

for $t = 1, \dots, T-1$, where all coefficient paths are deterministic. Define

$$z_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix}, \quad A_t = \begin{bmatrix} A_{11,t} & A_{12,t} \\ A_{21,t} & A_{22,t} \end{bmatrix}, \quad c_t = \begin{bmatrix} c_{1,t} \\ c_{2,t} \end{bmatrix}, \quad D = \mathrm{diag}(\alpha I_{d_x}, \beta I_{d_y}).$$

Assume that there are $Q = Q^\top \succ 0$ and $\delta \in (0,1]$ such that

$$K_t^\top Q K_t \preceq (1-\delta)^2 Q, \qquad K_t = I + DA_t. \tag{7}$$

This condition implies that A_t is invertible, so $z_t^* = -A_t^{-1}c_t$ is well defined. Write $\|z\|_Q^2 = z^\top Q z$.

Theorem 22 (Tracking for a stacked contraction)**.** *Let $\omega_{t+1} = (\xi_{t+1}, \zeta_{t+1})$ be a martingale difference relative to the joint filtration, and suppose $\mathbb{E}[\|D\omega_{t+1}\|_Q^2 \mid \mathcal{F}_t] \le \bar{\sigma}^2$. Under equation 7, with $W_T^z = \sum_{t=1}^{T-1} \|z_{t+1}^* - z_t^*\|_Q^2$,*

$$\sum_{t=1}^{T} \mathbb{E}\,\|z_t - z_t^*\|_Q^2 \le \frac{\|z_1 - z_1^*\|_Q^2}{\delta} + \frac{\bar{\sigma}^2 T}{\delta(2-\delta)} + \frac{W_T^z}{\delta^2}. \tag{8}$$

If $\mathbb{E}[\|\xi_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \sigma_x^2$ and $\mathbb{E}[\|\zeta_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \sigma_y^2$, then one may take $\bar{\sigma}^2 = \lambda_{\max}(Q)(\alpha^2\sigma_x^2 + \beta^2\sigma_y^2)$.

Proof. If $A_t u = 0$, then $K_t u = u$, while equation 7 gives $\|K_t u\|_Q \le (1-\delta)\|u\|_Q$. Thus $u = 0$, and A_t is invertible.

Stacking the two recursions and setting $e_t = z_t - z_t^*$ and $\Delta_t^z = z_{t+1}^* - z_t^*$ gives

$$e_{t+1} = K_t e_t + D\omega_{t+1} - \Delta_t^z.$$

The common quadratic inequality means that $\|K_t u\|_Q \le (1-\delta)\|u\|_Q$. The deterministic–stochastic decomposition used in Theorem 3, now in the Q norm, yields

$$\sum_{t=1}^{T} \mathbb{E}\,\|e_t\|_Q^2 \le \frac{\|e_1\|_Q^2}{\delta} + \frac{\bar\sigma^2 T}{1-(1-\delta)^2} + \frac{W_T^z}{\delta^2}.$$

Since $1-(1-\delta)^2 = \delta(2-\delta)$, this proves equation 8. Finally,

$$\mathbb{E}[\|D\omega_{t+1}\|_Q^2 \mid \mathcal{F}_t] \le \lambda_{\max}(Q)\left(\alpha^2\mathbb{E}[\|\xi_{t+1}\|_2^2 \mid \mathcal{F}_t] + \beta^2\mathbb{E}[\|\zeta_{t+1}\|_2^2 \mid \mathcal{F}_t]\right),$$

which gives the stated choice of $\bar\sigma^2$. □

F.2 Separate slow and fast bounds for a nested system

Assume throughout this subsection that all coefficient sequences and initial iterates are deterministic. Consider

$$y_{t+1} = (1-\beta)y_t + \beta(N_t y_t + g_t + \zeta_{t+1}),$$
$$x_{t+1} = (1-\alpha)x_t + \alpha(M_t x_t + H_t y_t + f_t + \xi_{t+1}),$$

where $\kappa_x, \kappa_y \in (0,1]$ and

$$\|N_t\|_{\mathrm{op}} \le 1-\kappa_y, \qquad \|M_t\|_{\mathrm{op}} \le 1-\kappa_x, \qquad 0 < \beta \le \alpha \le 1.$$

Define

$$y_t^* = (I-N_t)^{-1} g_t, \qquad \Lambda_t(y) = (I-M_t)^{-1}(H_t y + f_t),$$

and errors $e_t^y = y_t - y_t^*$ and $e_t^x = x_t - \Lambda_t(y_t)$. Let

$$W_T^y = \sum_{t=1}^{T-1} \left\|y_{t+1}^* - y_t^*\right\|^2, \qquad W_T^\Lambda = \sum_{t=1}^{T-1} \mathbb{E}\,\|\Lambda_{t+1}(y_{t+1}) - \Lambda_t(y_t)\|^2.$$

The quantity W_T^y is a deterministic target-variation budget. In contrast, W_T^Λ includes the random movement caused by the slow iterate and depends on the algorithm. Proposition 24 bounds it using coefficient changes, slow tracking error, and slow noise.

Theorem 23 (Separate slow and fast tracking bounds)**.** *Suppose ξ_{t+1} and ζ_{t+1} are martingale differences relative to the joint filtration, with conditional second moments at most σ_x^2 and σ_y^2. Then*

$$\sum_{t=1}^{T} \mathbb{E}\,\|e_t^y\|^2 \le \frac{\|e_1^y\|^2}{\beta\kappa_y} + \frac{\beta\sigma_y^2 T}{\kappa_y(2-\beta\kappa_y)} + \frac{W_T^y}{\beta^2\kappa_y^2},$$

and

$$\sum_{t=1}^{T} \mathbb{E}\,\|e_t^x\|^2 \le \frac{3\,\|e_1^x\|^2}{\alpha\kappa_x} + \frac{3\alpha\sigma_x^2 T}{\kappa_x(2-\alpha\kappa_x)} + \frac{3W_T^\Lambda}{\alpha^2\kappa_x^2}.$$

Proof. The slow fixed-point identity gives

$$e_{t+1}^y = ((1-\beta)I + \beta N_t)e_t^y + \beta\zeta_{t+1} - (y_{t+1}^* - y_t^*).$$

The coefficient norm is at most $1-\beta\kappa_y$, and the target increments are deterministic. The decomposition used in Theorem 3 gives the slow bound.

For the fast error,

$$e_{t+1}^x = ((1-\alpha)I + \alpha M_t)e_t^x + \alpha\xi_{t+1} - (\Lambda_{t+1}(y_{t+1}) - \Lambda_t(y_t)).$$

The last term is random and may be correlated with ξ_{t+1}, so the expected cross term need not vanish. Unrolling the recursion, applying $\|a+b+c\|^2 \le 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$, and then using Lemmas 10 and 12 together with the martingale case of Lemma 13 prove the fast bound. No independence between the two recursions is needed. □

The next results assume that the unprojected slow trajectory remains in a set $\mathcal{Y}$ almost surely. This can follow from a forward-invariant set or a separate boundedness argument. A second-moment noise bound alone does not imply this condition.

Proposition 24 (Bounding movement of the fast fixed-point map)**.** *Under the assumptions of Theorem 23, suppose additionally that each Λ_t is L_Λ-Lipschitz on a set $\mathcal{Y}$ containing $y_1, \dots, y_T$ almost surely. Define*

$$d_t^\Lambda = \sup_{y \in \mathcal{Y}} \|\Lambda_{t+1}(y) - \Lambda_t(y)\|, \qquad \Omega_T^\Lambda = \sum_{t=1}^{T-1} (d_t^\Lambda)^2, \qquad \rho_y = \sup_t \|N_t\|_{\mathrm{op}} \le 1 - \kappa_y.$$

Then

$$W_T^\Lambda \le 2\Omega_T^\Lambda + 2L_\Lambda^2 \beta^2 (1+\rho_y)^2 \sum_{t=1}^{T-1} \mathbb{E}\,\|e_t^y\|^2 + 2L_\Lambda^2 \beta^2 \sigma_y^2 T.$$

Proof. For every t,

$$\|\Lambda_{t+1}(y_{t+1}) - \Lambda_t(y_t)\|^2 \le 2L_\Lambda^2 \|y_{t+1} - y_t\|^2 + 2(d_t^\Lambda)^2.$$

The slow update gives $y_{t+1} - y_t = \beta((N_t - I)e_t^y + \zeta_{t+1})$. Since e_t^y is measurable before ζ_{t+1} is drawn,

$$\mathbb{E}\,\|y_{t+1} - y_t\|^2 \le \beta^2 (1+\rho_y)^2 \mathbb{E}\,\|e_t^y\|^2 + \beta^2 \sigma_y^2.$$

Substitution and summation prove the claim. □

Proposition 25 (Coefficient-based bound)**.** *Assume* $\sup_{y \in \mathcal{Y}} \|y\| \le R_y$, $\bar{H} = \sup_t \|H_t\|_{\mathrm{op}} < \infty$, *and* $\bar{f} = \sup_t \|f_t\| < \infty$. *Define*

$$C_\Lambda = \frac{\bar{H} R_y + \bar{f}}{\kappa_x}, \quad W_T^M = \sum_{t=1}^{T-1} \|M_{t+1} - M_t\|_{\mathrm{op}}^2,$$

$$W_T^H = \sum_{t=1}^{T-1} \|H_{t+1} - H_t\|_{\mathrm{op}}^2, \qquad W_T^f = \sum_{t=1}^{T-1} \|f_{t+1} - f_t\|^2.$$

One may take $L_\Lambda = \bar{H}/\kappa_x$, and

$$\Omega_T^\Lambda \le \frac{3}{\kappa_x^2} \left(R_y^2 W_T^H + W_T^f + C_\Lambda^2 W_T^M \right).$$

Together with Proposition 24 and Theorem 23, this expresses the fast-error bound in terms of coefficient changes and the slow-error bound.

Proof. The resolvent bound $\left\|(I - M_t)^{-1}\right\|_{\mathrm{op}} \le 1/\kappa_x$ gives

$$\|\Lambda_t(y) - \Lambda_t(y')\| \le \frac{\|H_t\|_{\mathrm{op}}}{\kappa_x} \|y - y'\|, \qquad \sup_{y \in \mathcal{Y}} \|\Lambda_t(y)\| \le C_\Lambda.$$

Moreover,

$$\begin{aligned}(I - M_{t+1})(\Lambda_{t+1}(y) - \Lambda_t(y)) = (H_{t+1} - H_t)y + (f_{t+1} - f_t) \\ + (M_{t+1} - M_t)\Lambda_t(y).\end{aligned}$$

Therefore

$$d_t^\Lambda \le \frac{1}{\kappa_x} \left(R_y \|H_{t+1} - H_t\|_{\mathrm{op}} + \|f_{t+1} - f_t\| + C_\Lambda \|M_{t+1} - M_t\|_{\mathrm{op}} \right).$$

Squaring and summing prove the claim. □

G Comparison with prior work

Table 1: Selected guarantees, with numerical constants suppressed. In the optimization row, m is the strong-convexity parameter and Δ^2 bounds each expected squared minimizer increment. In the forecasting row, ν^2 is the noise proxy and $|\theta_1| \le U$; $\widetilde{O}$ suppresses logarithmic factors.

Work	Model and feedback	Variation	Selected guarantee
Hadiji et al. (2026), Theorem 2.1	Contractive updates; exact operators	Sum of squared target increments	Cumulative squared error; deterministic initialization and movement terms used here
Bastianello et al. (2024), Propositions 1 and 3	General contractive operators; possibly biased errors	Bound on per-round target movement	Expected distance under second-moment bounds; high-probability tracking with additional tail and update assumptions
Cutler et al. (2023), Theorem 3	Strongly convex optimization; unbiased noisy gradients	Uniform second-moment bound on minimizer increments	Expected squared distance; initialization plus $\eta\sigma^2/m + \Delta^2/(m^2\eta^2)$
Baby & Wang (2019), Corollary 2	Scalar targets; independent sub-Gaussian observations; known horizon and noise parameter	$\|D\theta\|_2^2 \le W$	High-probability adaptive squared-error bound $\widetilde{O}(T^{2/3}W^{1/3}\nu^{4/3} + U^2 + W + \nu^2)$
This paper, Theorems 3 and 9	Deterministic affine maps; centered noise satisfying equation 2	Sum of squared fixed-point increments, W_T	Explicit cumulative mean-square bound; stepsize adaptation with an additional bounded, unbiased target observation