A BANACH-SPACE THEORY OF MARKOVIAN HALPERN ITERATION FOR NON-EXPANSIVE MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study stochastic approximation of fixed points of a non-expansive operator T when the oracle samples originate from a continuing Markovian trajectory. A direct block-minibatch implementation of Halpern iteration attains an expected last-iterate residual of order $O(\log N/N)$, but accrues a substantive complexity of $\tilde{O}(\epsilon^{-5})$ Markovian samples. We therefore introduce a variance-reduced Markovian PAGE-Halpern method whose refresh and same-state difference blocks are analyzed through the Poisson equation. In Hilbert spaces, the cocoercivity of $I - T$ results in an $O(\epsilon^{-3})$ sample complexity. Our main result extends this construction to a general finite-dimensional Banach space. A displacement-level Halpern bound replaces the Hilbert-space potential and yields $\tilde{O}(\epsilon^{-3})$ sample complexity. We also establish a high-probability guarantee with the same leading accuracy dependence by measuring the estimator in an auxiliary smooth norm.

1 INTRODUCTION

Many stochastic approximation (SA) algorithms can be expressed as fixed-point iterations for an operator that can be only observed through noisy samples (Robbins & Monro, 1951; Kushner & Yin, 2003; Borkar, 2008). Contractive operators, in particular, already admit a well-developed finite-iteration theory because every exact update is guaranteed to decrease the distance to the unique fixed point (Srikant & Ying, 2019; Qu & Wierman, 2020; Chen et al., 2020). Non-expansive operators present different challenges. Namely, the fixed point is not necessarily unique, there is generally no strict contraction drift, and a sampling error can persist without being geometrically attenuated.

Halpern iteration (Halpern, 1967) provides a natural deterministic method in this regime. Given an anchor x_0 and a non-expansive operator T, the canonical Halpern recursion is given by

$$x_n = \frac{1}{n+1}x_0 + \frac{n}{n+1}Tx_{n-1}. \tag{1}$$

The method has an $O(1/N)$ last-iterate fixed-point residual in Hilbert spaces, and the initial choice of the anchor can be used to select a specific fixed point (Wittmann, 1992; Lieder, 2020). These properties make it attractive for non-expansive stochastic fixed-point problems, including average-reward reinforcement learning (RL) (Tsitsiklis & Van Roy, 1999; Abounadi et al., 2001).

The present paper considers the setting where the samples are generated by one continuing Markov chain. The oracle error need not have zero conditional mean at the start of a Markovian block, and successive blocks remain dependent through the state at their common boundary. We are therefore in need of an analysis that preserves the online trajectory and accounts explicitly for both the martingale component of the Markovian error and its conditional bias at each block boundary. Furthermore, ordinary minibatching estimates every operator value independently, resulting in an overall complexity of $\tilde{O}(\epsilon^{-5})$. To improve this rate, variance reduction methods such as PAGE (Li et al., 2021) can be used when a single Markov state can be used to evaluate the oracle at two query points. The estimator can then track the change in the operator rather than reconstructing its value at every iteration.

In Hilbert spaces, this idea can be coupled with the cocoercivity of $F = I - T$ (Cai et al., 2022). The same argument, however, is unavailable in the sup and block-sup norms that arise frequently in RL (Abounadi et al., 2001; Bellemare et al., 2023). Our approach, in this case, is to move the recursive estimator back from F to T and control the increments of the Halpern displacement directly. This

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.

gives a Banach-space argument in the norm in which the operator is known to be non-expansive, extending the theory to geometries where an inner-product structure is not available.

We summarize our contributions as follows.

- We establish perturbation- and displacement-level bounds for inexact Halpern iteration. The former yields a Markovian block baseline with $\tilde{O}(\epsilon^{-5})$ sample complexity, while the latter enables recursive variance reduction.
- We develop PAGE-Halpern along one uninterrupted Markov trajectory. Using a Poisson-equation decomposition, we obtain $O(\epsilon^{-3})$ sample complexity in Hilbert space and $\tilde{O}(\epsilon^{-3})$ in finite-dimensional Banach spaces, including sup and block-sup geometries.
- We prove high-probability Banach-space guarantees using an auxiliary smooth norm. Smooth approximations of the sup norm incur only logarithmic dimension dependence (Juditsky & Nemirovski, 2023; Chen et al., 2020).

Related work. The closest results study stochastic Halpern iteration under i.i.d. sampling (Cai et al., 2022; Bravo & Contreras, 2026), Markovian SA for non-expansive maps through Krasnosel'skiĭ–Mann iteration (Blaser & Zhang, 2026), and variance reduction along a single Markov trajectory (Sun & Wei, 2026). Our results combine Halpern anchoring, recursive same-state differences, and Poisson-equation control of one continuing trajectory, and extend the analysis to general finite-dimensional Banach norms. Appendix A provides a fuller comparison with these and related lines of work.

2 Setup and inexact Halpern bounds

Let $(\mathbb{R}^d, \|\cdot\|)$ be a finite-dimensional normed space, write $\|\cdot\|_2$ for the Euclidean norm and let

$$\mu_{\|\cdot\|} := \sup_{z \neq 0} \frac{\|z\|}{\|z\|_2} < \infty. \tag{2}$$

Let $(Y_t)_{t\geq 0}$ be a Markov chain on a finite state space $\mathcal{Y}$, with transition matrix P and stationary distribution π. Given a measurable map $H : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$, define the mean operator

$$Tx := \sum_{y\in\mathcal{Y}} \pi(y) H(x, y). \tag{3}$$

We seek an approximate fixed point of T, under the following assumptions.

Assumption 2.1 (Non-expansive operator with fixed points)**.** The operator T is non-expansive, meaning $\|Tx - Tz\| \leq \|x - z\|$ for all $x, z \in \mathbb{R}^d$, and $\mathrm{Fix}(T) \neq \emptyset$.

Assumption 2.2 (Finite ergodic Markov chain)**.** The finite chain $(Y_t)_{t\geq 0}$ is irreducible and aperiodic. Thus, there are constants $C_{\mathrm{mix}} \geq 1$ and $\rho \in (0, 1)$ such that $\sup_{y\in\mathcal{Y}} \|P^t(y, \cdot) - \pi\|_{\mathrm{TV}} \leq C_{\mathrm{mix}}\rho^t$ for every $t \geq 0$.

Assumption 2.3 (Finite oracle variance)**.** For every $x \in \mathbb{R}^d$, $\sum_{y\in\mathcal{Y}} \pi(y)\|H(x, y) - Tx\|_2^2 < \infty$.

Assumption 2.4 (Iterate stability)**.** For $q \in \{1, 2\}$, there exist constants $R_q < \infty$ such that $\sup_{n\geq 0} \mathbb{E}\big[\|x_n - x_0\|^q\big] \leq R_q^q$ for the generated iterates of the SA algorithm under consideration.

The first assumption specifies the fixed-point geometry, and the second describes the sampling process. The remaining two control the random region visited by the iterates and ensure that the oracle has a finite second moment at every fixed query point. All four are standard SA conditions. Their relation to the assumptions used in stochastic Halpern and Markovian SA literature is discussed in Appendix A.

We present two ways to control an inexact Halpern recursion. The first is a standard Markovian block method with adaptive batch size scheduling, which builds a fresh operator estimate at each step and bounds the magnitude of its error. Appendix B.3 analyzes this naive method and obtains an $O(\log N/N)$ expected last-iterate residual with $\tilde{O}(\epsilon^{-5})$ samples. The high cost of repeatedly forming accurate block estimates motivates the use of a more sophisticated, variance-reduced estimator, which updates earlier estimates using the same Markovian states at two successive query points. The analysis of this second method needs a residual bound in terms of changes in the estimation error. For an arbitrary error sequence $(U_n)_{n\geq 1}$, consider the inexact Halpern update

$$x_n = (1 - \beta_n)x_0 + \beta_n(Tx_{n-1} + U_n). \tag{4}$$

The next theorem bounds the last-iterate residual using the increments of the scaled errors $\beta_n U_n$, together with the error at the final step.

Theorem 2.5 (Displacement-level bound)**.** *Under Assumptions 2.1 and 2.4, let* $D_0 := \operatorname{dist}(x_0, \operatorname{Fix}(T))$ *and* $\kappa_1 := R_1 + 2D_0$. *Let* (x_n) *be generated by (4) with* $\beta_n = n/(n+1)$. *Set* $U_0 := 0$ *and* $\Delta_n := \beta_n U_n - \beta_{n-1} U_{n-1}$. *If* $\mathbb{E}[\|\Delta_n\|^2] \le q_n$ *and* $\mathbb{E}[\|U_N\|] \le r$, *then*

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \frac{\kappa_1(1+\log(N+1))}{N+1} + \frac{1}{N+1}\sum_{n=1}^{N} n\sqrt{q_n} + r. \tag{5}$$

The bound is suited to recursive estimation, where changes in the error are easier to control than the error at every iteration. We use it for the Banach-space PAGE analysis in Section 4. The proof is provided in Appendix B.2, and a pathwise version is given in Appendix E.

3 Markovian PAGE-Halpern in Hilbert spaces

The naive block baseline method of Appendix B.3 requires $\tilde{O}(\epsilon^{-5})$ samples. In the i.i.d. setting, Bravo & Contreras (2026) obtain an $O(\epsilon^{-5})$ upper bound and an $\Omega(\epsilon^{-3})$ lower bound for linear-span algorithms with bounded-variance single-point oracle access. The gap between these exponents motivates the recursive construction in this section. Since i.i.d. sampling is a special case of Markovian sampling, the lower bound remains a useful benchmark, although our PAGE method uses same-state evaluations at two query points. The apparent reason for the inefficency of the baseline method is that it treats the oracle errors at different query points as unrelated. Under only pointwise finite variance and affine variance growth, this is essentially all we can use, since a sample at x gives no information about the oracle error at a nearby point z. Furthermore, the resulting residual bound depends on the weighted sum of perturbations $\sum_{n=1}^{N} n\mathbb{E}[\|U_n\|]$. Keeping this sum $O(\log N)$ requires the nth block error to be of order n^{-2}. Since averaging reduces standard deviation only as $k_n^{-1/2}$, the block sizes grow as $k_n = n^4$, and hence to the $\tilde{O}(\epsilon^{-5})$ sample complexity.

Variance reduction can make more efficient use of observed samples, albeit with the prerequisite of a stronger oracle model. The relevant additional regularity is that, using the same Markov state, two evaluations at nearby query points have a small mean-square difference after subtracting their stationary means. Under this condition, a recursive estimator can be updated from same-sample differences instead of being rebuilt from scratch at every iteration. The stochastic cost is then governed by the displacement of the Halpern iterates rather than the full variance of a fresh estimate. Concretely, the estimator alternates between two kinds of blocks. A refresh block averages $\widehat{F}(x, Y)$ at a single query point to estimate $F(x)$. A recursive block uses the same Markov states at two query points and averages $\widehat{F}(x, Y) - \widehat{F}(z, Y)$ to estimate the difference $F(x) - F(z)$.

In a Hilbert space, there is a direct way to combine a PAGE-like variance-reduced estimator with a residual bound. If T is non-expansive, then $F := I - T$ is monotone and $1/2$-cocoercive, since

$$\|Fx - Fz\|_2^2 = \|(x-z) - (Tx - Tz)\|_2^2 \le 2\langle Fx - Fz, x - z\rangle. \tag{6}$$

Thus, controlling the residual $\|x - Tx\|_2$ is equivalent to controlling $\|F(x)\|_2$ (Cai et al., 2022).

3.1 Markovian PAGE oracle

The analysis of this section uses the stochastic residual oracle $\widehat{F}(x, y) := x - H(x, y)$, so that

$$F(x) = \mathbb{E}_\pi\big[\widehat{F}(x, Y)\big]. \tag{7}$$

The additional structure needed for recursive variance reduction is stated in the following condition.

Condition 3.1 (PAGE-compatible Markovian oracle)**.** The stochastic residual oracle permits multi-point access, i.e., after observing a state Y, the algorithm may evaluate both $\widehat{F}(x, Y)$ and $\widehat{F}(z, Y)$. Furthermore, the oracle has stationary second moments that satisfy

$$\sum_{y \in \mathcal{Y}} \pi(y)\|\widehat{F}(x, y) - F(x)\|_2^2 \le \sigma_F^2, \tag{8}$$

and same-sample difference second moments that satisfy

$$\sum_{y \in \mathcal{Y}} \pi(y)\|\widehat{F}(x, y) - \widehat{F}(z, y) - (F(x) - F(z))\|_2^2 \le L_F^2\|x - z\|_2^2. \tag{9}$$

3.2 THE MARKOVIAN PAGE-HALPERN SCHEME

We now describe the Markovian, Hilbert-space PAGE-Halpern recursion. The scheme observes one continuing trajectory of the Markov chain $(Y_t)_{t\geq 0}$ with natural filtration $(\mathcal{F}_t)$ enlarged by past randomization. Each block starts at a stopping time τ, and its length S is an integer-valued $\mathcal{F}_\tau$-measurable random variable, chosen before any state within the block is observed. Query points are also $\mathcal{F}_\tau$-measurable. For a block starting after τ, we define the refresh and difference block averages

$$R(x;\tau,S) := \frac{1}{S}\sum_{i=1}^{S}\widehat{F}(x,Y_{\tau+i}) \quad \text{and} \quad D(x,z;\tau,S) = \frac{1}{S}\sum_{i=1}^{S}\big(\widehat{F}(x,Y_{\tau+i}) - \widehat{F}(z,Y_{\tau+i})\big). \tag{10}$$

Letting $x_0 \in \mathbb{R}^d$, we choose $L^* \geq \max\{2, L_F\}$, and set $p_n = \min\{1, 2/(n+1)\}$ and $\beta_n = n/(n+1)$. The choice of L^* upper bounds both the cocoercivity scale of $F = I - T$ and the same-sample Lipschitz scale in Condition 3.1. We initialize the chain time at $\tau = 0$ and set $v_0 = R(x_0;\tau,S_{1,0})$. Given v_{n-1}, we compute

$$x_n = (1-\beta_n)x_0 + \beta_n\left(x_{n-1} - \frac{1}{L^*}v_{n-1}\right), \tag{11}$$

and then update

$$v_n = \begin{cases} R(x_n;\tau,S_{1,n}), & \text{with probability } p_n, \\ v_{n-1} + D(x_n,x_{n-1};\tau,S_{2,n}), & \text{with probability } 1-p_n. \end{cases} \tag{12}$$

Thus, v_n is a recursive PAGE estimator of $F(x_n)$, updated either by a refresh block or by a difference block. The PAGE coin at iteration n is tossed before choosing the corresponding refresh or difference block length. The random quantities $S_{1,n}$ and $S_{2,n}$ may depend on the past, including the PAGE coin toss and the already computed displacement $\|x_n - x_{n-1}\|$, but not on states within that block. Each estimator used in the update is therefore measurable before the start of the next Markovian block.

3.3 PAGE ERROR AND SAMPLE COMPLEXITY

The chain state at the beginning of each block is inherited from all preceding PAGE updates. Consequently, the samples within the block are not independent of the past and need not begin in stationarity. We control this dependence through the stopping-time Poisson decomposition proved in Appendix C. The main result states that for a function g with zero stationary mean, if the length S is chosen at a stopping time τ, then the block average has conditional bias at most $C_{\mathrm{B}}\|g\|_{2,\pi}/S$ and conditional second moment at most $C_{\mathrm{V}}\|g\|_{2,\pi}^2/S$. The proof uses the Poisson equation to decompose the block sum into a martingale sum and a boundary term involving only the first and last states. The martingale sum gives the $1/S$ second-moment decay, while the boundary term gives the $1/S$ bias. Lemma C.2 applies this result to the two block types used by PAGE, giving a conditional bias bound and a conditional second-moment bound for both the refresh and same-state difference blocks. The following lemma uses these estimates to choose the block sizes for a prescribed estimator accuracy.

Lemma 3.2 (Hilbert PAGE schedule). *Fix a target estimator accuracy $\delta > 0$ and choose*

$$S_{1,0} \geq \frac{C_{\mathrm{V}}\sigma_F^2}{\delta^2},\ S_{1,n} \geq \frac{C_{\mathrm{V}}\sigma_F^2}{p_n\delta^2},\ S_{2,n} \geq \max\left\{\frac{C_{\mathrm{V}}L_F^2\|x_n - x_{n-1}\|_2^2}{p_n^2\delta^2}, \frac{C_{\mathrm{B}}L_F\|x_n - x_{n-1}\|_2}{p_n^{3/2}\delta}, 1\right\}. \tag{13}$$

Then the PAGE-Halpern recursion (11)–(12) satisfies

$$\mathbb{E}\big[\|v_n - F(x_n)\|_2^2\big] \leq \frac{100\delta^2}{n+1}, \qquad n \geq 0. \tag{14}$$

The Hilbert-space residual potential absorbs this error while simultaneously controlling the displacements that determine the cost of the difference blocks. The required cocoercive estimate is due to Cai et al. (2022). We state it in our notation below and give the supporting details in Appendix C.

Proposition 3.3 (PAGE-Halpern residual in Hilbert spaces). *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and $1/L^*$-cocoercive with $F^{-1}(0) \neq \emptyset$. Let $R := \mathrm{dist}(x_0, F^{-1}(0))$, and generate (x_n) by (11)–(12). If*

$$\mathbb{E}\big[\|v_n - F(x_n)\|_2^2\big] \leq \frac{\delta^2}{n+1}, \qquad n \geq 0, \tag{15}$$

then there exist universal constants C_0, C_1 such that

$$\mathbb{E}\|F(x_N)\|_2 \le \frac{C_0 L^* R}{N} + C_1\delta. \tag{16}$$

Moreover, there exist universal constants $C_2, c_2 > 0$ such that, if $p_n = 2/(n+1)$ and $\delta N \le c_2 L^* R$,

$$\sum_{n=1}^{N} \frac{\mathbb{E}\big[\|x_n - x_{n-1}\|_2^2\big]}{p_n^2} \le C_2 R^2 N. \tag{17}$$

Combining Proposition 3.3 with the Markovian PAGE schedule provides us with a complexity result. The proposition controls the residual and the displacements that determine the cost of recursive blocks and the Poisson decomposition makes these blocks valid along one continuing Markov trajectory.

Theorem 3.4 (Markovian PAGE-Halpern complexity in Hilbert spaces)**.** *Suppose Assumption 2.1 holds on a Hilbert space, and let $F := I - T$. Suppose further that Assumption 2.2 and Condition 3.1 hold. Consider the PAGE-Halpern recursion (11)–(12), and let $R := \mathrm{dist}(x_0, \mathrm{Fix}(T))$. Choose $L^* \ge \max\{2, L_F\}$ and use the block-size schedules of Lemma 3.2. Let C_0, C_1 be the constants in Proposition 3.3. If $R = 0$ or $\epsilon \ge 2R$, the anchor x_0 already has residual at most ϵ. Otherwise,*

$$\delta = \frac{\epsilon}{20C_1}, \qquad N = \left\lceil \frac{2C_0 L^* R}{\epsilon} \right\rceil \tag{18}$$

ensures

$$\mathbb{E}\big[\|x_N - Tx_N\|_2\big] = \mathbb{E}\big[\|F(x_N)\|_2\big] \le \epsilon. \tag{19}$$

Moreover, the expected number of Markovian samples used by the continuing trajectory is

$$O\left(\frac{C_{\mathrm{V}}\big(\sigma_F^2 L^* R + (L^* R)^3\big)}{\epsilon^3} + \frac{C_{\mathrm{B}}(L^* R)^{5/2}}{\epsilon^{5/2}}\right). \tag{20}$$

In particular, the leading dependence on the target accuracy ϵ is $O(\epsilon^{-3})$.

The theorem improves the $\tilde{O}(\epsilon^{-5})$ block-baseline complexity to $O(\epsilon^{-3})$, with its leading accuracy exponent agrees with the i.i.d. lower bound, while the Markovian boundary contribution is the lower-order $O(\epsilon^{-5/2})$ term. In particular, temporal dependence changes the chain-dependent constants but introduces no additional logarithmic factor in the leading Hilbert-space rate.

4 Markovian PAGE-Halpern in Banach spaces

The Hilbert-space result of the previous section depends on the fortunate geometric identity that if T is non-expansive in a Hilbert norm, then $F := I - T$ is cocoercive. This identity lets us convert residual control into a monotone-inclusion estimate, and provides PAGE with a quadratic potential into which its estimator error can be inserted. However, norms that arise in many common applications, such as the sup norms, span seminorms and block-sup norms of RL and distributional RL, do not have an inner-product geometry. The aim of this section is to establish similar results in general Banach spaces. Rather than estimating the residual operator $F = I - T$, we estimate the fixed-point operator T itself, and use the resulting estimator along with the displacement-level Halpern bound of Theorem 2.5. The anchoring in the Halpern iteration plays the role of the cocoercivity in the Hilbert analysis in converting control of the estimator increments into control of the fixed-point residual.

4.1 Markovian PAGE oracle

We will now use the recursive estimator to estimate Tx rather than $F(x)$. Therefore, we need an analogue of Condition 3.1 which is imposed instead on the original stochastic operator H.

Condition 4.1 (PAGE-compatible Markovian oracle in Banach spaces)**.** The stochastic operator permits multi-point access, i.e., after observing a state Y, the algorithm may evaluate both $H(x, Y)$ and $H(z, Y)$. Furthermore, every query point x generated up to the target horizon satisfies

$$\sum_{y\in\mathcal{Y}} \pi(y)\|H(x,y) - Tx\|_2^2 \le \sigma_T^2, \tag{21}$$

and every query-point pair x, z generated up to the target horizon satisfies

$$\sum_{y\in\mathcal{Y}} \pi(y)\|H(x,y) - H(z,y) - (Tx - Tz)\|_2^2 \le L_T^2\|x - z\|^2. \tag{22}$$

The input displacement in the second bound is measured in $\|\cdot\|$, the non-expansiveness norm of T.

4.2 The Markovian PAGE-Halpern scheme

The scheme observes the same continuing Markovian trajectory as in the previous section. For a block starting after the current chain time τ, let

$$R_T(x;\tau,S) := \frac{1}{S}\sum_{i=1}^{S} H(x, Y_{\tau+i}) \quad \text{and} \quad D_T(x,z;\tau,S) := \frac{1}{S}\sum_{i=1}^{S}\big(H(x,Y_{\tau+i}) - H(z,Y_{\tau+i})\big). \tag{23}$$

Letting $x_0 \in \mathbb{R}^d$, we set $p_n = \min\{1, 2/(n+1)\}$ and $\beta_n = n/(n+1)$. We initialize the chain time at $\tau = 0$ and set $v_0 = R_T(x_0;\tau,S_{1,0})$. Given v_{n-1}, we compute

$$x_n = (1-\beta_n)x_0 + \beta_n v_{n-1}, \tag{24}$$

and then update

$$v_n = \begin{cases} R_T(x_n;\tau,S_{1,n}), & \text{with probability } p_n, \\ v_{n-1} + D_T(x_n, x_{n-1};\tau,S_{2,n}), & \text{with probability } 1-p_n. \end{cases} \tag{25}$$

Evidently, v_n is a recursive PAGE estimator directly of Tx_n. The same Poisson decomposition argument as in the Hilbert analysis controls its refresh and difference blocks. The main difference is how estimator accuracy is converted into residual control. In the absence of cocoercivity, we use Theorem 2.5, which bounds the residual through the increments of the scaled estimator error. The schedule in the next lemma makes these increments decay at the required $1/n$ rate. Since the difference-block lengths depend on $\|x_n - x_{n-1}\|$, their total sample cost can then be controlled by the Halpern displacement bounds. Together, these estimates replace the Hilbert-space cocoercive potential and yield a variance-reduced guarantee in the original Banach norm.

Lemma 4.2 (Banach PAGE schedule)**.** *Fix a target estimator accuracy $\delta > 0$. Choose*

$$S_{1,0} \ge \frac{C_{\mathrm{V}}\sigma_T^2}{\delta^2},\ S_{1,n} \ge \frac{C_{\mathrm{V}}\sigma_T^2}{p_n\delta^2},\ S_{2,n} \ge \max\left\{\frac{C_{\mathrm{V}}L_T^2\|x_n - x_{n-1}\|^2}{p_n^2\delta^2}, \frac{C_{\mathrm{B}}L_T\|x_n - x_{n-1}\|}{p_n^{3/2}\delta}, 1\right\}. \tag{26}$$

Then the PAGE-Halpern recursion (24)–(25) satisfies

$$\mathbb{E}\big[\|v_n - Tx_n\|_2^2\big] \le \frac{100\delta^2}{n+1}, \qquad n \ge 0. \tag{27}$$

The displacement-level reduction of Theorem 2.5 converts the estimator control into a residual guarantee in the non-expansiveness norm. The next theorem is the central geometric extension of the Hilbert result, retaining the sample efficiency of recursive estimation without using cocoercivity.

Theorem 4.3 (Markovian PAGE-Halpern in Banach spaces)**.** *Suppose Assumptions 2.1, 2.2 and 2.4 hold in the non-expansiveness norm $\|\cdot\|$. Suppose further that Condition 4.1 holds. Let $\kappa := R_2 + 2\,\mathrm{dist}(x_0, \mathrm{Fix}(T))$. Consider the PAGE-Halpern recursion (24)–(25) and use the block-size schedules of Lemma 4.2. There exists a universal constant C such that for every $N \ge 1$,*

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \frac{\kappa(1+\log(N+1))}{N+1} + C\mu_{\|\cdot\|}\delta. \tag{28}$$

Consequently, choosing

$$\delta = \frac{\epsilon}{2C\mu_{\|\cdot\|}}, \qquad N = \Theta\left(\frac{\kappa}{\epsilon}\log\frac{\kappa}{\epsilon}\right) \tag{29}$$

ensures

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \epsilon. \tag{30}$$

Moreover, the expected number of Markovian samples used by the continuing trajectory is

$$\tilde{O}\left(\frac{C_V\left(\mu_{\|\cdot\|}^2\sigma_T^2\kappa + (1+\mu_{\|\cdot\|}^2)L_T^2\kappa^3\right)}{\epsilon^3} + \frac{\mu_{\|\cdot\|}C_B L_T \kappa^{5/2}}{\epsilon^{5/2}}\right). \tag{31}$$

In particular, the leading dependence on the target accuracy is $\tilde{O}(\epsilon^{-3})$.

The first term in (28) is the deterministic Halpern contribution. Its harmonic factor introduces a logarithmic overhead in N, so the Banach-space sample complexity is $\tilde{O}(\epsilon^{-3})$ rather than $O(\epsilon^{-3})$. The leading polynomial dependence on the target accuracy remains ϵ^{-3}, as in the Hilbert result. The factor $\mu_{\|\cdot\|}$ is incurred only when Euclidean estimator bounds are converted to the working norm. In particular, $\mu_{\|\cdot\|} \le 1$ for sup and block-sup norms, so these geometries incur no dimension-dependent norm-comparison loss. Further details are provided in Appendix D.

5 High-probability guarantees

The results of the previous sections concern expected residual bounds. In this section we provide the corresponding high-probability statement in Banach spaces. The general structure of the analysis remains similar to Section 4, the new ingredient being estimator concentration. Since martingale concentration in a general norm requires smoothness of said norm, our approach will be to measure the recursive estimator in an auxiliary smooth norm $\|\cdot\|_s$, and to compare it back to $\|\cdot\|$ at the end.

5.1 Smooth high-probability oracle

Condition 5.1 (Smooth high-probability Markovian oracle)**.** There is a norm $\|\cdot\|_{\mathrm{s}}$ on $\mathbb{R}^d$ and constants $\alpha_{\mathrm{s}}, \kappa_{\mathrm{s}} \ge 1$ such that

$$\|z\| \le \alpha_{\mathrm{s}}\|z\|_{\mathrm{s}}, \qquad z \in \mathbb{R}^d, \tag{32}$$

and $(\mathbb{R}^d, \|\cdot\|_{\mathrm{s}})$ is κ_{s}-smooth, meaning $\|z+h\|_{\mathrm{s}}^2 \le \|z\|_{\mathrm{s}}^2 + \langle \nabla\|z\|_{\mathrm{s}}^2, h\rangle + \kappa_{\mathrm{s}}\|h\|_{\mathrm{s}}^2$ for all $z, h \in \mathbb{R}^d$. The oracle permits Markovian multi-point access. Moreover, every generated query point x satisfies

$$\sum_{y\in\mathcal{Y}} \pi(y)\|H(x,y) - Tx\|_{\mathrm{s}}^2 \le \sigma_{\mathrm{s}}^2, \tag{33}$$

and every generated query-point pair x, z satisfies

$$\sum_{y\in\mathcal{Y}} \pi(y)\|H(x,y) - H(z,y) - (Tx - Tz)\|_{\mathrm{s}}^2 \le L_{\mathrm{s}}^2\|x - z\|^2. \tag{34}$$

When the working norm is Euclidean, one may take $\|\cdot\|_{\mathrm{s}} = \|\cdot\|_2$, $\alpha_{\mathrm{s}} = 1$, and $\kappa_{\mathrm{s}} = 1$. When the working norm is non-smooth, the estimator norm can be smoothed without changing the fixed-point geometry. For example, if $\|\cdot\| = \|\cdot\|_\infty$, take $\|\cdot\|_{\mathrm{s}} = \|\cdot\|_q$ with $q = \max\{2, \lceil \log(\max\{e, d\})\rceil\}$. Then

$$\|z\|_\infty \le \|z\|_q \le d^{1/q}\|z\|_\infty \le e\|z\|_\infty, \tag{35}$$

and $\kappa_{\mathrm{s}} = q - 1 = O(\log d)$. The same construction extends easily to block-sup norms.

The concentration certificate must be combined with a pathwise analogue of the stability bound used in expectation. We therefore impose the following finite-horizon condition.

Condition 5.2 (Pathwise stability)**.** For the generated iterates up to the target horizon N, there exists a deterministic constant $K < \infty$ such that

$$\sup_{0\le n\le N} \|Tx_n - x_0\| \le K \quad \text{a.s.} \tag{36}$$

We now revisit the Banach-space PAGE-Halpern recursion of (24)–(25). Let C_{hp} denote the chain-dependent constant in the high-probability block estimate of Lemma E.2 and define

$$\ell_N := 1 + \log(N+1), \qquad \Lambda_{N,\eta} := \kappa_s + \log(32(N+1)/\eta), \tag{37}$$

and set $p_0 := 1$.

Lemma 5.3 (High-probability PAGE schedule). *Let $N \geq 1$, $\eta \in (0,1)$, and $\delta > 0$. Choose*

$$S_{1,0} \geq C_{\mathrm{hp}} \left(\frac{\Lambda_{N,\eta} \sigma_{\mathrm{s}}^2}{\delta^2} + \frac{\sigma_{\mathrm{s}}}{\delta} \right), \quad S_{1,n} \geq C_{\mathrm{hp}} \left(\frac{\Lambda_{N,\eta} \sigma_{\mathrm{s}}^2}{p_n \delta^2} + \frac{\sigma_{\mathrm{s}}}{\sqrt{p_n}\delta} \right), \tag{38}$$

and

$$S_{2,n} \geq C_{\mathrm{hp}} \left(\frac{\Lambda_{N,\eta} L_{\mathrm{s}}^2 \|x_n - x_{n-1}\|^2 \ell_N^2}{p_n^2 \delta^2} + \frac{L_{\mathrm{s}} \|x_n - x_{n-1}\| \ell_N}{p_n \delta} + 1 \right). \tag{39}$$

Let $e_n := v_n - T x_n$, with the convention $e_{-1} := 0$, and let $\mathcal{R}_n$ denote the event that iteration n uses a refresh block. Then, with probability at least $1 - \eta$,

$$\max_{0 \leq n \leq N} \|e_n\|_{\mathrm{s}} \leq \delta, \qquad \sum_{n=1}^{N} n \|\beta_n e_{n-1} - \beta_{n-1} e_{n-2}\|_{\mathrm{s}} \leq C \delta N (1 + \log(1/\eta)), \tag{40}$$

and

$$\sum_{n=1}^{N} \frac{\mathbf{1}_{\mathcal{R}_n}}{p_n} \leq C N (1 + \log(1/\eta)). \tag{41}$$

The schedule controls the estimator error and its scaled increments simultaneously over all iterations up to N. Combining these bounds with pathwise stability and the displacement-level Halpern argument gives the following high-probability guarantee for the last iterate.

Theorem 5.4 (High-probability Markovian PAGE-Halpern). *Suppose Assumptions 2.1 and 2.2 and Conditions 5.1 and 5.2 hold. Run the Banach PAGE-Halpern recursion (24)–(25) with the block-size schedules of Lemma 5.3. Then, with probability at least $1 - \eta$,*

$$\|x_N - T x_N\| \leq \frac{K(1 + \log(N+1))}{N+1} + C \alpha_{\mathrm{s}} \delta (1 + \log(1/\eta)). \tag{42}$$

Consequently, the choices

$$N = \Theta \left(\frac{K}{\epsilon} \log \frac{K}{\epsilon} \right), \qquad \delta = \Theta \left(\frac{\epsilon}{\alpha_{\mathrm{s}}(1 + \log(1/\eta))} \right) \tag{43}$$

ensure $\|x_N - T x_N\| \leq \epsilon$ with probability at least $1 - \eta$. On the same event, the number of Markovian samples used through time N is

$$\tilde{O} \left(\frac{\alpha_{\mathrm{s}}^2 (1 + \log(1/\eta))^3 (\kappa_{\mathrm{s}} + \log(1/\eta)) \left(\sigma_{\mathrm{s}}^2 K + L_{\mathrm{s}}^2 K^3 \right)}{\epsilon^3} \right). \tag{44}$$

The theorem shows that requiring confidence $1 - \eta$ does not change the leading ϵ^{-3} dependence. Confidence and smoothing enter only through logarithmic factors and the smoothness and norm-comparison constants of the auxiliary norm. The proof follows the same structure as the expected Banach result, but controls all estimator errors and increments simultaneously on an event of probability at least $1 - \eta$, rather than only in second moment. The smooth-norm Poisson decomposition establishes these bounds for every refresh and difference block. On this event, the deterministic displacement-level inequality converts the increment bounds into the displayed residual estimate. The block tail bound and complete proofs are given in Appendix E.

6 Experiments

We use two eight-state average-reward problems driven by the same continuing cycle chain. In the first, policy evaluation has the linear Bellman map $Bh = \bar{r} + Ph$, which is non-expansive in the Euclidean norm. In the second, a two-action control problem has the nonlinear optimality map $\mathcal{T}Q = \bar{r} + P \max_a Q(\cdot, a)$, which is non-expansive in the sup norm but not in any inner-product norm. Both transition oracles can be evaluated at two query points after observing a single transition. The constructions, fixed points, and unbiasedness calculations are given in Appendix F.

We compare block Halpern with the corresponding Hilbert or Banach PAGE-Halpern method on fast- and slow-mixing versions of the chain. Each pair receives the same realized trajectory, beginning

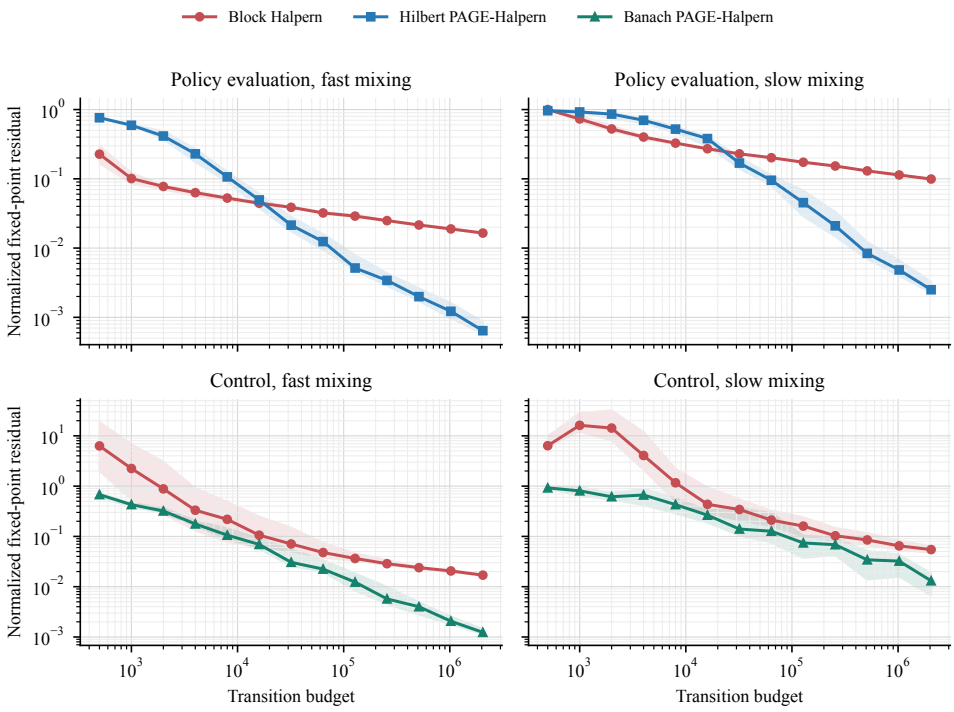


Figure 1: Normalized fixed-point residual at common transition budgets. The top row evaluates the Hilbert method on policy evaluation, and the bottom row evaluates the Banach method on average-reward control. Each PAGE method is compared with the block method on the same realized trajectories. Lines show medians over twenty trajectories and shaded regions interquartile ranges.

at state zero, and every observed transition counts toward the budget. We report the normalized last-iterate residual at common budgets $500 \cdot 2^j$, $j = 0, \ldots, 12$, over twenty repetitions. Lines in Figure 1 show medians and the shaded regions show interquartile ranges, which summarize the central half of the trajectories without allowing a few large residuals to dominate. Hyperparameters and target-hitting transition counts are reported in Appendix F.

On the fast evaluation chain, block Halpern benefits initially from its short blocks, but PAGE is more accurate once its initialization cost is amortized. At $128{,}000$ transitions, its median residual is 0.00515, compared with 0.0290 for block Halpern. The advantage is clearer on the slow chain, where PAGE reaches residual 0.05 after a median of $107{,}374$ transitions and the block method does not reach that target within the maximum budget. The Banach control experiment gives the same qualitative separation. On the slow chain, Banach PAGE reaches residual 0.05 after a median of $80{,}000$ transitions in all repetitions, whereas block Halpern reaches it in only 8/20. Thus, same-state differences become increasingly valuable as the target tightens or the trajectory mixes more slowly.

7 Conclusion

We developed a finite-sample theory for stochastic Halpern iteration along one continuing Markovian trajectory. A direct block method gives an expected $O(\log N/N)$ last-iterate residual with $\tilde{O}(\epsilon^{-5})$ samples, while same-state recursive variance estimation improves the complexity to $O(\epsilon^{-3})$ in Hilbert space and $\tilde{O}(\epsilon^{-3})$ in finite-dimensional Banach spaces. The Banach analysis works in the norm where T is non-expansive, and smooth auxiliary norms give high-probability guarantees for sup and block-sup geometries. The theory makes the required information structure explicit. Stability controls the visited region, the Poisson equation handles Markovian dependence and bias, and PAGE requires each sampled state to be reused at two query points. The average-reward experiments implement this model on continuing trajectories without stationary-oracle access. Thus, Halpern anchoring and same-state variance reduction provide last-iterate guarantees in the Banach geometries natural to RL.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, designing and implementing the synthetic experiments, and language and formatting revisions. The authors independently checked all citations. The authors made all final scientific and editorial decisions and take responsibility for the complete contents of the submission.

ETHICS STATEMENT

This work is theoretical and uses only synthetic finite-state reinforcement learning problems. It involves no human subjects, personal or private data, external datasets, or deployed decision systems, and raises no specific ethical concerns.

REPRODUCIBILITY STATEMENT

The main text states all assumptions, oracle conditions, algorithms, and block-size schedules. Complete proofs are provided in the appendix. In particular, Appendix B gives the stability and baseline analysis, Appendix C develops the stopping-time Poisson decomposition and Hilbert-space block estimates, and Appendices D and E contain the Banach-space and high-probability arguments. The experiments use no external data. Appendix F specifies both finite-state problems and their transition oracles, all initialization and hyperparameter choices, the sampling protocol, transition budgets, residual metrics, number of repetitions, and reporting procedure. The supplementary material includes the code used to run the experiments and generate the reported figure and transition-count data.

REFERENCES

Jinane Abounadi, Dimitrib Bertsekas, and Vivek S Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pp. 778–816. PMLR, 2022.

Heinz H Bauschke and Patrick L Combettes. Correction to: convex analysis and monotone operator theory in Hilbert spaces. In *Convex analysis and monotone operator theory in Hilbert spaces*, pp. C1–C4. Springer, 2020.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. Pmlr, 2017.

Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Springer Science & Business Media, 2012.

Ethan Blaser and Shangtong Zhang. Asymptotic and finite sample analysis of nonexpansive stochastic approximations with Markovian noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19764–19772, 2026.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.

Vivek S Borkar and Sean P Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Mario Bravo and Roberto Cominetti. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 62(1):191–219, 2024.

Mario Bravo and Juan Pablo Contreras. Stochastic Halpern iteration in normed spaces and applications to reinforcement learning: M. bravo, jp contreras. *Mathematical Programming*, pp. 1–39, 2026.

Xufeng Cai, Chaobing Song, Cristóbal Guzmán, and Jelena Diakonikolas. Stochastic Halpern iteration with variance reduction for stochastic monotone inclusions. *Advances in Neural Information Processing Systems*, 35:24766–24779, 2022.

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

Zaiwei Chen, Siva T Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A Lyapunov theory for finite-sample guarantees of Markovian stochastic approximation. *Operations Research*, 72(4): 1352–1367, 2024.

Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pp. 1428–1451. PMLR, 2020.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with Markovian noise: Application to linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pp. 1711–1752. PMLR, 2021.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Benjamin Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

Shaan Ul Haque and Siva Theja Maguluri. Stochastic approximation with unbounded Markovian noise: A general-purpose theorem. In *International Conference on Artificial Intelligence and Statistics*, pp. 3718–3726. PMLR, 2025.

Hideaki Iiduka. Mini-batch stochastic Halpern algorithm for nonexpansive fixed point problems. *arXiv preprint arXiv:2604.21443*, 2026.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Anatoli Juditsky and Arkadii S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces, 2023. URL https://arxiv.org/abs/0809.0813.

Ege C Kaya and Abolfazl Hashemi. A finite-iteration theory for asynchronous categorical distributional temporal-difference learning. *arXiv preprint arXiv:2605.06866*, 2026.

Ege Can Kaya, Aliasghar Pourghani, Vijay Gupta, and Abolfazl Hashemi. Quotient-categorical representations for Bellman-compatible average-reward distributional reinforcement learning. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=qVHRXJt8X2.

M. A. Krasnoselskii. Two comments on the method of successive approximations. *Usp. Math. Nauk*, 10:123–127, 1955.

Harold J Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

Jongmin Lee, Ernest K Ryu, and Vaneet Aggarwal. Learning policy from a single trajectory in average-reward Markov decision process. *arXiv preprint arXiv:2606.16729*, 2026.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pp. 6286–6295. PMLR, 2021.

Felix Lieder. On the convergence rate of the Halpern-iteration. *Optimization letters*, 15(2):405–418, 2020.

Zhankun Luo, Antesh Upadhyay, M Berk Sahin, Sang Bin Moon, Anuran Makur, and Abolfazl Hashemi. Unified high-probability analysis of stochastic variance-reduced estimation. *arXiv preprint arXiv:2605.15388*, 2026.

W Robert Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Nicholas Pischke and Thomas Powell. Asymptotic regularity of a generalised stochastic Halpern scheme. *Journal of Optimization Theory and Applications*, 210(1):3, 2026.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Conference on learning theory*, pp. 3185–3205. PMLR, 2020.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Mark Rowland, Marc Bellemare, Will Dabney, Remi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on learning theory*, pp. 2803–2830. PMLR, 2019.

Zhenyu Sun and Ermin Wei. Learning from a single Markovian trajectory: Optimality and variance reduction. *Advances in Neural Information Processing Systems*, 38:20004–20033, 2026.

John N Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35 (11):1799–1808, 1999.

Yi Wan, Abhishek Naik, and Richard S Sutton. Learning and planning in average-reward Markov decision processes. In *International Conference on Machine Learning*, pp. 10653–10662. PMLR, 2021.

Rainer Wittmann. Approximation of fixed points of nonexpansive mappings. *Archiv der mathematik*, 58(5):486–491, 1992.

Tengyu Xu, Zhe Wang, Yi Zhou, and Yingbin Liang. Reanalysis of variance reduced temporal difference learning. In *International Conference on Learning Representations*, 2020.

APPENDIX TABLE OF CONTENTS

A Additional related work **15**

B Perturbation-level analysis and the Markovian block baseline **16**

B.1 Discussion of assumptions 16

B.2 Perturbation-level Halpern bounds 16

B.3 Markovian block method 19

C Supporting results for Hilbert PAGE-Halpern **21**

C.1 Poisson decomposition of Markovian error 21

D Supporting results for Banach PAGE-Halpern **23**

E Supporting results for high-probability guarantees **23**

E.1 High-probability Markovian blocks 23

E.2 Proof of Lemma 5.3 24

E.3 High-probability residual bound 25

F Experimental details **26**

F.1 Hilbert-space policy evaluation 26

F.2 Banach-space average-reward control 27

F.3 Methods and evaluation 27

G Proof of the displacement-level bound **28**

G.1 Proof of Theorem 2.5 28

H Proofs of the main Hilbert-space results **29**

H.1 Proof of Lemma 3.2 29

H.2 Proof of Proposition 3.3 30

H.3 Proof of Theorem 3.4 30

I Proofs of the main Banach-space results **31**

I.1 Proof of Lemma 4.2 31

I.2 Proof of Theorem 4.3 32

J Proof of the main high-probability result **34**

J.1 Proof of Theorem 5.4 34

A Additional related work

The classical averaged iterations of Mann (1953) and Krasnoselskii (1955) form the basis of Krasnosel'skiĭ–Mann methods for non-expansive fixed-point problems. A modern account of their relation to monotone operators and splitting methods is given by Bauschke & Combettes (2020). Halpern iteration introduces an anchor into this recursion (Halpern, 1967). The strong convergence of this iteration to the fixed point selected by the anchor was established under standard weight conditions by Wittmann (1992). For the canonical weights in Hilbert space, Lieder (2020) proved a tight $O(1/N)$ residual bound. Diakonikolas (2020) used this rate and a residual potential to obtain near-optimal methods for monotone inclusions and variational inequalities. We use the same anchored last-iterate structure, but develop perturbation estimates that remain valid for stochastic operator evaluations.

For stochastic non-expansive maps, Bravo & Cominetti (2024) analyze Krasnosel'skiĭ–Mann iterations under martingale noise. Bravo & Contreras (2026) study stochastic Halpern iteration in normed spaces under i.i.d. oracle access. Their explicit minibatch construction has $O(\epsilon^{-5})$ complexity, while their lower bound has order $\Omega(\epsilon^{-3})$ for a broad single-point linear-span oracle model. However, the same-sample multi-point access used by PAGE is stronger than their lower-bound oracle. The lower bound is therefore an important benchmark, but does not constitute a matched minimax statement for our information model. Recent stochastic Halpern analyses also study independent minibatches and abstract asymptotic regularity (Iiduka, 2026; Pischke & Powell, 2026). Cai et al. (2022) combine stochastic Halpern iteration with recursive variance reduction for monotone inclusions under independent sampling.

The analysis of stochastic approximation with Markovian observations has a long history. The ODE and Poisson-equation approaches are developed systematically by Kushner & Yin (2003), Borkar (2008), and Benveniste et al. (2012). Modern finite-time results cover linear stochastic approximation and temporal-difference learning (Srikant & Ying, 2019; Durmus et al., 2021), contractive operators in arbitrary norms (Chen et al., 2020; 2024; Qu & Wierman, 2020), and unbounded Markovian noise under Lyapunov growth conditions (Haque & Maguluri, 2025). In the non-expansive setting, Blaser & Zhang (2026) develop asymptotic and finite-sample guarantees for stochastic Krasnosel'skiĭ–Mann recursion with Markovian noise. Our blocks also preserve one continuing trajectory. Their boundaries are predictable stopping times, and the Poisson equation controls both the conditional bias and the second moment of each estimator.

Recursive variance reduction was developed for stochastic optimization through methods including SVRG, SARAH, SPIDER, and PAGE (Johnson & Zhang, 2013; Nguyen et al., 2017; Fang et al., 2018; Li et al., 2021). These methods exploit same-sample differences whose second moments decrease as successive query points become close. Related constructions have been developed for monotone variational inequalities (Alacaoglu & Malitsky, 2022) and for stochastic Halpern iteration under independent sampling (Cai et al., 2022). Variance reduction from a single Markov trajectory has also been studied for smooth nonconvex optimization and temporal-difference learning (Xu et al., 2020; Sun & Wei, 2026). Our estimator uses the same recursive principle, while its refresh and difference samples are consecutive segments of one Markov chain.

The Banach-space and high-probability parts require a separate geometric argument. Smooth Lyapunov functions based on generalized Moreau envelopes give finite-sample stochastic approximation bounds in arbitrary contraction norms (Chen et al., 2020; 2024). Concentration for vector-valued martingales in smooth Banach spaces is developed by Pinelis (1994) and Juditsky & Nemirovski (2023). Our expected Banach result does not smooth the fixed-point residual, but instead replaces the Hilbert cocoercive potential with a displacement estimate in the original non-expansiveness norm. Smoothing enters only when the recursive estimator must be controlled with high probability, where we use the unified estimator analysis of Luo et al. (2026).

Average-reward policy evaluation is a natural continuing-trajectory application. Classical temporal-difference methods estimate differential values from one endless Markov trajectory (Tsitsiklis & Van Roy, 1999), while average-reward control leads to relative-value and Q-learning recursions whose Bellman maps are naturally studied modulo additive constants (Abounadi et al., 2001; Wan et al., 2021). Concurrently, Lee et al. (2026) combine anchored value iteration with recursive Bellman-difference estimation for finite weakly communicating average-reward MDPs and obtain a high-probability $\tilde{O}(\epsilon^{-2})$ transition complexity for finding an ϵ-optimal policy. Their construction uses return times to extract state-conditional i.i.d. transition samples from one trajectory and exploits the

exact Bellman decomposition. Our analysis instead controls consecutive Markovian blocks through the Poisson equation and applies to general non-expansive operators in finite-dimensional Banach spaces. Distributional RL replaces scalar values by return distributions and uses supremum-over-state metrics for its Bellman operators (Bellemare et al., 2017; Rowland et al., 2018; Bellemare et al., 2023). Recent finite-iteration results study asynchronous categorical TD under Markovian sampling in statewise supremum geometries (Kaya & Hashemi, 2026), while quotient-categorical representations yield non-expansive average-reward distributional operators in a coordinate Cramér geometry (Kaya et al., 2026). These settings motivate a last-iterate theory that does not replace the application norm by an inner-product norm.

B Perturbation-level analysis and the Markovian block baseline

B.1 Discussion of assumptions

Assumption 2.3 is the Markovian analogue of the pointwise finite-variance oracle model used for stochastic Halpern iteration in Bravo & Contreras (2026, Assumption 1). Related finite-sample stochastic approximation analyses impose the same kind of second-moment control, often in a growth form. See, for instance, Chen et al. (2020, Assumption 2.2) and Chen et al. (2024, Assumption 4). The explicit complexity theorem below states the additional growth bound needed to turn this oracle model into exact numerical constants.

Assumption 2.4 is the standard stochastic-approximation stability condition. See, for instance, the stability theory of Borkar (2008) and Borkar & Meyn (2000). In the Markovian non-expansive setting, Blaser & Zhang (2026) identify stability through the boundedness of the iterates. The non-expansive fixed-point residual analyses of Bravo & Cominetti (2024, Condition (H4)) use the same type of stability condition for stochastic Krasnosel'skiĭ–Mann iterations, while Bravo & Contreras (2026, Lemma 1) use the closely related boundedness of the operator images $Tx_n - x_0$. Lemma B.2 below shows that this operator-image bound is a consequence of Assumptions 2.1 and 2.4.

Remark B.1 (A sufficient condition for stability). Suppose that, for $q \in \{1, 2\}$, there exist constants $K_q < \infty$ such that

$$\sup_{x,y} \|H(x, y) - x_0\|^q \le K_q^q. \tag{45}$$

Then Assumption 2.4 holds for any update

$$x_n = (1 - \beta_n)x_0 + \beta_n G_n \tag{46}$$

with $\beta_n \in [0, 1]$ in which G_n is a convex combination of oracle values $H(x_{n-1}, y)$. Indeed, if $G_n = \sum_j w_j H(x_{n-1}, y_j)$ with $w_j \ge 0$ and $\sum_j w_j = 1$, then Jensen's inequality gives

$$\|x_n - x_0\|^q = \beta_n^q \|G_n - x_0\|^q \le \sum_j w_j \|H(x_{n-1}, y_j) - x_0\|^q \le K_q^q. \tag{47}$$

Thus, Assumption 2.4 holds with $R_q = K_q$. In particular, this condition applies to the Markovian block recursion in Appendix B.3.

B.2 Perturbation-level Halpern bounds

For $1 \le i \le n$, let

$$B_i^n := \prod_{j=i}^{n} \beta_j, \tag{48}$$

with the empty-product convention $B_i^n := 1$ when $0 \le n < i$, and define $\beta_0 := \sigma_0 := 0$, where $\sigma_n := \mathbb{E}\big[\|U_n\|\big]$.

Lemma B.2 (Stability constants). *If Assumption 2.4 holds, then*

$$\sup_{n\ge 0} \mathbb{E}\big[\|Tx_n - x_0\|\big] \le \kappa_1, \qquad \sup_{n\ge 0} \mathbb{E}\big[\|Tx_n - x_0\|^2\big] \le \kappa_2^2. \tag{49}$$

More precisely, if $D_0 := \operatorname{dist}(x_0, \operatorname{Fix}(T))$, then we may take

$$\kappa_q := R_q + 2D_0, \qquad q \in \{1, 2\}. \tag{50}$$

Proof of Lemma B.2. Choose $x^* \in \mathrm{Fix}(T)$ such that $\|x^* - x_0\| = D_0$. The choice exists because $\mathrm{Fix}(T)$ is closed in finite dimension. By non-expansiveness and $Tx^* = x^*$,

$$\|Tx_n - x_0\| \le \|Tx_n - Tx^*\| + \|x^* - x_0\| \le \|x_n - x^*\| + D_0 \le \|x_n - x_0\| + 2D_0. \tag{51}$$

Taking expectations gives the $q = 1$ claim with $\kappa_1 = R_1 + 2D_0$. Taking second moments in the same inequality and then applying Minkowski's inequality gives the $q = 2$ claim with $\kappa_2 = R_2 + 2D_0$. □

Proposition B.3 (Perturbation-level inexact Halpern residual)**.** *Suppose Assumptions 2.1 and 2.4 hold, and let (β_n) be nondecreasing. Let (x_n) be generated by (4). Then, for all $n \ge 1$,*

$$\mathbb{E}\big[\|x_n - Tx_n\|\big] \le \kappa_1(1-\beta_n) + \sum_{i=1}^{n} B_i^n\big(\kappa_1(\beta_i - \beta_{i-1}) + \beta_i\sigma_i + \beta_{i-1}\sigma_{i-1}\big) + \beta_n\sigma_n. \tag{52}$$

Proof of Proposition B.3. Let $r_n := \|Tx_n - x_0\|$, $P_0 := 0$, and, for $n \ge 1$,

$$P_n := r_{n-1}(\beta_n - \beta_{n-1}) + \beta_{n-1}P_{n-1} + \beta_n\|U_n\| + \beta_{n-1}\|U_{n-1}\|. \tag{53}$$

We have

$$x_n - x_{n-1} = (\beta_n - \beta_{n-1})(Tx_{n-1} - x_0) + \beta_{n-1}(Tx_{n-1} - Tx_{n-2}) + \beta_n U_n - \beta_{n-1}U_{n-1}. \tag{54}$$

Hence, using the non-expansiveness in Assumption 2.1,

$$\begin{aligned}\|x_n - x_{n-1}\| &\le (\beta_n - \beta_{n-1})r_{n-1} + \beta_{n-1}\|x_{n-1} - x_{n-2}\| + \beta_n\|U_n\| + \beta_{n-1}\|U_{n-1}\|. \\ &= P_n.\end{aligned} \tag{55}$$

Unrolling the recursion for P_n, we obtain

$$P_n = \sum_{i=1}^{n} B_i^{n-1}\big(r_{i-1}(\beta_i - \beta_{i-1}) + \beta_i\|U_i\| + \beta_{i-1}\|U_{i-1}\|\big). \tag{56}$$

Now, we have

$$\begin{aligned}\|x_n - Tx_n\| &\le (1-\beta_n)\|x_0 - Tx_n\| + \beta_n\|Tx_{n-1} - Tx_n\| + \beta_n\|U_n\| \\ &\le (1-\beta_n)r_n + \beta_n\|Tx_{n-1} - Tx_n\| + \beta_n\|U_n\| \\ &\le (1-\beta_n)r_n + \beta_n\|x_{n-1} - x_n\| + \beta_n\|U_n\| \\ &\le (1-\beta_n)r_n + \beta_n P_n + \beta_n\|U_n\|,\end{aligned} \tag{57}$$

using the definition of r_n, non-expansiveness of T, and (55). Plugging in the definition of P_n and using the fact that $\beta_n B_i^{n-1} = B_i^n$, and taking expectations,

$$\mathbb{E}\big[\|x_n - Tx_n\|\big] \le (1-\beta_n)\mathbb{E}[r_n] + \sum_{i=1}^{n} B_i^n\big(\mathbb{E}[r_{i-1}](\beta_i - \beta_{i-1}) + \beta_i\mathbb{E}\big[\|U_i\|\big] + \beta_{i-1}\mathbb{E}\big[\|U_{i-1}\|\big]\big) + \beta_n\mathbb{E}\big[\|U_n\|\big]. \tag{58}$$

By Assumption 2.4, $\sup_{n\ge0}\mathbb{E}[r_n] = \sup_{n\ge0}\mathbb{E}\big[\|Tx_n - x_0\|\big] \le \kappa_1$. Then,

$$\mathbb{E}\big[\|x_n - Tx_n\|\big] \le \kappa_1(1-\beta_n) + \sum_{i=1}^{n} B_i^n\big(\kappa_1(\beta_i - \beta_{i-1}) + \beta_i\sigma_i + \beta_{i-1}\sigma_{i-1}\big) + \beta_n\sigma_n. \tag{59}$$

□

Lemma B.4 (Perturbation-level bound)**.** *Under the conditions of Proposition B.3, let $\beta_n = n/(n+1)$ and $N \in \mathbb{N}$. Then,*

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \frac{\kappa_1(1 + \log(N+1))}{N+1} + \frac{2}{N+1}\sum_{n=1}^{N} n\sigma_n. \tag{60}$$

Consequently, if $\sum_{n=1}^{N} n\sigma_n \le C_\sigma(1 + \log(N+1))$, then

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \frac{(\kappa_1 + 2C_\sigma)(1 + \log(N+1))}{N+1}. \tag{61}$$

Proof of Lemma B.4. Proposition B.3 directly gives

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \kappa_1(1-\beta_N) + \sum_{i=1}^{N} B_i^N\big(\kappa_1(\beta_i-\beta_{i-1}) + \beta_i\sigma_i + \beta_{i-1}\sigma_{i-1}\big) + \beta_N\sigma_N. \quad (62)$$

With the choice of the step size sequence, we get

$$\kappa_1(1-\beta_N) = \kappa_1\left(1 - \frac{N}{N+1}\right) = \frac{\kappa_1}{N+1}. \quad (63)$$

Also,

$$B_i^N = \prod_{j=i}^{N} \frac{j}{j+1} = \frac{i}{N+1} \quad (64)$$

by means of a telescoping product. Finally,

$$\beta_i - \beta_{i-1} = \frac{i}{i+1} - \frac{i-1}{i} = \frac{1}{i(i+1)}. \quad (65)$$

Therefore, the exact-Halpern part of the bound becomes

$$\frac{\kappa_1}{N+1} + \sum_{i=1}^{N} B_i^N \kappa_1(\beta_i-\beta_{i-1}) = \frac{\kappa_1}{N+1} + \frac{\kappa_1}{N+1}\sum_{i=1}^{N}\frac{1}{i+1} = \frac{\kappa_1}{N+1}\sum_{i=0}^{N}\frac{1}{i+1}. \quad (66)$$

Since $\sum_{i=0}^{N}\frac{1}{i+1} \le 1+\log(N+1)$, the exact-Halpern terms are bounded by

$$\frac{\kappa_1(1+\log(N+1))}{N+1}. \quad (67)$$

We now need to handle the perturbation terms, i.e.,

$$\sum_{i=1}^{N} B_i^N(\beta_i\sigma_i + \beta_{i-1}\sigma_{i-1}) + \beta_N\sigma_N. \quad (68)$$

We separate the two sums and shift the index of the second:

$$\begin{aligned}\sum_{i=1}^{N} B_i^N(\beta_i\sigma_i + \beta_{i-1}\sigma_{i-1}) + \beta_N\sigma_N &= \sum_{i=1}^{N} B_i^N\beta_i\sigma_i + \sum_{i=1}^{N} B_i^N\beta_{i-1}\sigma_{i-1} + \beta_N\sigma_N \\ &= \sum_{i=1}^{N-1}\beta_i(B_i^N + B_{i+1}^N)\sigma_i + \beta_N(B_N^N+1)\sigma_N.\end{aligned} \quad (69)$$

Using $B_i^N = i/(N+1)$, each $1 \le i < N$ satisfies

$$\beta_i(B_i^N + B_{i+1}^N) = \frac{i}{i+1}\frac{2i+1}{N+1} \le \frac{2i}{N+1}, \quad (70)$$

and

$$\beta_N(B_N^N+1) = \frac{N}{N+1}\left(\frac{N}{N+1}+1\right) \le \frac{2N}{N+1}. \quad (71)$$

Substituting these into Proposition B.3,

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \frac{\kappa_1(1+\log(N+1))}{N+1} + \frac{2}{N+1}\sum_{n=1}^{N} n\sigma_n. \quad (72)$$

□

B.3 MARKOVIAN BLOCK METHOD

Fix an anchor $x_0 \in \mathbb{R}^d$, a step size sequence $(\beta_n) \subset (0,1)$, block lengths $(k_n) \subset \mathbb{N}$, and optional burn-in lengths $(b_n) \subset \mathbb{N} \cup \{0\}$. At iteration n, starting from the current Markov state Y_{τ_n}, run the chain for b_n unused steps, and then collect the next k_n states. Let

$$\tau_1 := 0, \qquad \tau_n := \sum_{i=1}^{n-1} (b_i + k_i) \quad \text{for } n \geq 2, \tag{73}$$

so that τ_n indicates the time at which the nth block begins. Further, let

$$G_n := \frac{1}{k_n} \sum_{j=1}^{k_n} H(x_{n-1}, Y_{\tau_n + b_n + j}). \tag{74}$$

The Markovian-block Halpern update is

$$x_n = (1 - \beta_n) x_0 + \beta_n G_n, \qquad n \geq 1. \tag{75}$$

Equivalently,

$$x_n = (1 - \beta_n) x_0 + \beta_n (T x_{n-1} + U_n), \qquad U_n := G_n - T x_{n-1}. \tag{76}$$

We can interpret U_n as the block estimation error, which includes both the error due to sampling and the Markovian bias from using a dependent, nonstationary segment of the chain.

A mixing-time argument would first discard a burn-in segment to make the conditional bias small. Instead, we use the Poisson-equation decomposition of Markovian noise (Haque & Maguluri, 2025; Blaser & Zhang, 2026). The conditional bias is then a boundary term of order k^{-1}, while the conditional second moment is of order k^{-1}. Thus, a block can begin immediately at the current stopping time.

Lemma B.5 (Markovian block error)**.** *Suppose Assumption 2.2 holds. There exists a constant C_{blk}, depending only on the finite-state chain and $\mu_{\|\cdot\|}$, such that the following holds. Let τ be a stopping time and let x be an $\mathcal{F}_\tau$-measurable query point. Define*

$$V_H(x) := \sum_{y \in \mathcal{Y}} \pi(y) \|H(x,y) - Tx\|_2^2. \tag{77}$$

Then, for every integer-valued $\mathcal{F}_\tau$-measurable block length $k \geq 1$,

$$\mathbb{E}\Big[\Big\| \frac{1}{k} \sum_{j=1}^{k} H(x, Y_{\tau+j}) - Tx \Big\| \mid \mathcal{F}_\tau \Big] \leq \frac{C_{\mathrm{blk}} V_H(x)^{1/2}}{\sqrt{k}}. \tag{78}$$

Proof of Lemma B.5. Condition on $\mathcal{F}_\tau$. The query point x and the block length k are then fixed. Let

$$f_x(y) := H(x,y) - Tx. \tag{79}$$

By definition, $Tx = \sum_{y \in \mathcal{Y}} \pi(y) H(x,y)$, so f_x has stationary mean zero and

$$\|f_x\|_{2,\pi}^2 = \sum_{y \in \mathcal{Y}} \pi(y) \|f_x(y)\|_2^2 = V_H(x). \tag{80}$$

By the Poisson block estimate in Lemma C.1,

$$\mathbb{E}\Big[\Big\| \frac{1}{k} \sum_{j=1}^{k} f_x(Y_{\tau+j}) \Big\|_2^2 \mid \mathcal{F}_\tau \Big] \leq \frac{C_{\mathrm{V}} V_H(x)}{k}. \tag{81}$$

Conditional Jensen's inequality and $\|z\| \leq \mu_{\|\cdot\|} \|z\|_2$ give the result after absorbing $\mu_{\|\cdot\|} \sqrt{C_{\mathrm{V}}}$ into C_{blk}. □

Although Assumption 2.3 gives us a qualitative property of the oracle model, in that it requires finite variance at each fixed query point, establishing a finite-sample rate further requires a quantitative way to control the variances at the random query points generated by the algorithm. The affine growth condition that follows, together with trajectory stability, provides such a bound.

Condition B.6 (Affine variance growth)**.** Every query point x generated up to the target horizon satisfies

$$\sum_{y\in\mathcal{Y}} \pi(y)\|H(x,y) - Tx\|_2^2 \le \sigma_0^2 + \sigma_1^2\|Tx - x_0\|^2. \tag{82}$$

Uniformly bounded variance is a special case that is recovered when $\sigma_1 = 0$. This distinction is the same one present in the i.i.d. stochastic Halpern analysis of Bravo & Contreras (2026). Their standing oracle assumption is pointwise finite variance, but their concrete non-expansive minibatch complexity results require uniformly bounded variance in order to turn the abstract perturbation condition into an explicit schedule for the batch sizes. Condition B.6 relaxes that uniformly bounded variance specialization while still yielding an explicit finite-sample constant through Assumption 2.4.

The affine-growth condition is also natural in applications such as RL. In TD or distributional TD updates, the random target contains the current value or return-distribution estimate at a sampled next state. Consequently, the conditional second moment of the update noise typically scales with the size of the current estimate. This is the reason why standard stochastic approximation analyses of RL use affine-growth variance conditions. See, for instance, the discussion by Bellemare et al. (2023, Section 6.8, pp. 180-181) and the stochastic approximation stability theory of Borkar (2008) and Borkar & Meyn (2000).

The next theorem quantifies the best guarantee obtained by rebuilding the operator estimate from a fresh Markovian block at every Halpern iteration. It serves as the non-variance-reduced reference point for the PAGE results in the main body.

Theorem B.7 (Inexact Halpern residual with Markovian block minibatches)**.** *Suppose Assumptions 2.1 to 2.4 and Condition B.6 hold. Run (75) with*

$$\beta_n = \frac{n}{n+1}, \qquad k_n = n^4, \tag{83}$$

and let (b_n) be any nonnegative burn-in schedule satisfying $\sum_{n=1}^{N} b_n = O(N\log N)$. In particular, the no burn-in scenario with $b_n = 0$ is also admissible. Let $D_0 := \mathrm{dist}(x_0, \mathrm{Fix}(T))$. Then, there exists a constant C, depending only on $R_1, R_2, D_0, C_{\mathrm{blk}}, \sigma_0, \sigma_1$ and ρ such that

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le C\frac{1+\log(N+1)}{N+1}. \tag{84}$$

Consequently, to obtain $\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \epsilon$, it suffices to take

$$N = O\left(\frac{C}{\epsilon}\log\frac{C}{\epsilon}\right), \tag{85}$$

and the total number of Markov samples needed is

$$\sum_{n=1}^{N}(b_n + k_n) = \tilde{O}(\epsilon^{-5}). \tag{86}$$

The theorem gives the desired $O(\log N/N)$ last-iterate residual, but its sample cost is substantially larger. The perturbation bound requires the nth block error to be of order n^{-2}, which in turn requires $k_n = n^4$ samples. Summing these costs over N iterations produces the fifth-order dependence on the target accuracy. This is the bottleneck removed by recursive variance estimation.

Proof of Theorem B.7. Let $\mathcal{F}_{\tau_n}$ denote the sigma-algebra at the start of the nth block. By construction, x_{n-1}, b_n and k_n are $\mathcal{F}_{\tau_n}$-measurable, while the states used in G_n are sampled after step τ_n. Therefore, Lemma B.5 applies after conditioning on $\mathcal{F}_{\tau_n}$, with query point $x = x_{n-1}$, burn-in length $b = b_n$, and retained block length $k = k_n$. By Condition B.6 and Assumption 2.4,

$$\mathbb{E}\big[V_H(x_{n-1})\big] \le \sigma_0^2 + \sigma_1^2\mathbb{E}\big[\|Tx_{n-1} - x_0\|^2\big] \le \bar{\sigma}_H^2, \tag{87}$$

where $\bar{\sigma}_H^2 := \sigma_0^2 + \sigma_1^2\kappa_2^2$. By Lemma B.5,

$$\mathbb{E}\big[\|U_n\| \mid \mathcal{F}_{\tau_n}\big] \le C_{\mathrm{blk}} V_H(x_{n-1})^{1/2} k_n^{-1/2}. \tag{88}$$

Taking expectations using Jensen's inequality, and the choice $k_n = n^4$, we get

$$\sigma_n = \mathbb{E}\big[\|U_n\|\big] \le C_{\mathrm{blk}} \bar{\sigma}_H n^{-2} \le C_0 n^{-2} \tag{89}$$

for some constant C_0. Therefore, the weighted perturbation sum required by Lemma B.4 satisfies

$$\sum_{n=1}^{N} n\sigma_n \le C_0 \sum_{n=1}^{N} \frac{1}{n} \le C_0(1 + \log N). \tag{90}$$

For the sample complexity, each iteration uses exactly $b_n + k_n$ Markovian samples. By the assumed bound on the burn-in schedule,

$$\sum_{n=1}^{N} (b_n + k_n) = O(N^5) + O(N \log N) = O(N^5). \tag{91}$$

Substituting $N = O\big((C/\epsilon)\log(C/\epsilon)\big)$ gives a sample complexity of $\tilde{O}(\epsilon^{-5})$. □

C Supporting results for Hilbert PAGE-Halpern

C.1 Poisson decomposition of Markovian error

The i.i.d. analyses of stochastic approximation methods leverage the martingale-difference property of the sequence of perturbations. This property is evidently not available when the samples are Markovian. However, an analogue is recovered through the Poisson-equation decomposition (Blaser & Zhang, 2026). For any stationary-mean zero function $g : \mathcal{Y} \to \mathbb{R}^d$, finiteness and ergodicity of the chain guarantee a solution ν_g to

$$g(y) = \nu_g(y) - (P\nu_g)(y). \tag{92}$$

Along a block beginning at τ, this identity can be written as

$$g(Y_{\tau+i}) = M_i + (P\nu_g)(Y_{\tau+i-1}) - (P\nu_g)(Y_{\tau+i}), \tag{93}$$

where

$$M_i = \nu_g(Y_{\tau+i}) - (P\nu_g)(Y_{\tau+i-1}) \tag{94}$$

is a martingale difference after conditioning on $\mathcal{F}_\tau$. Summing over a block leaves a martingale sum and two boundary terms. This gives the usual S^{-1} second-moment scaling for the random part of the block average, together with an additional S^{-1} conditional bias from the boundary terms. Lemma B.5 used the resulting first-moment estimate for the baseline method. We now state the full conditional bounds and specialize them to the PAGE refresh and difference estimators.

Lemma C.1. *Suppose Assumption 2.2 holds. Let $g : \mathcal{Y} \to \mathbb{R}^d$ satisfy $\sum_y \pi(y)g(y) = 0$. There exist constants $C_{\mathrm{B}}, C_{\mathrm{V}}$, depending only on the finite-state Markov chain, such that for any stopping time τ and any integer-valued $\mathcal{F}_\tau$-measurable block length $S \ge 1$,*

$$\Big\| \mathbb{E}\Big[\frac{1}{S} \sum_{i=1}^{S} g(Y_{\tau+i}) \mid \mathcal{F}_\tau \Big] \Big\|_2 \le \frac{C_{\mathrm{B}} \|g\|_{2,\pi}}{S}, \tag{95}$$

and

$$\mathbb{E}\Big[\Big\| \frac{1}{S} \sum_{i=1}^{S} g(Y_{\tau+i}) \Big\|_2^2 \mid \mathcal{F}_\tau \Big] \le \frac{C_{\mathrm{V}} \|g\|_{2,\pi}^2}{S}. \tag{96}$$

Proof of Lemma C.1. Condition on $\mathcal{F}_\tau$. Since S is $\mathcal{F}_\tau$-measurable, it is fixed before the block is sampled. It is therefore enough to prove the displayed inequalities for deterministic S, with constants independent of τ and S. Since g has stationary mean zero and the chain is finite and geometrically mixing, the Poisson equation

$$g(y) = \nu_g(y) - (P\nu_g)(y) \tag{97}$$

has the solution

$$\nu_g(y) = \sum_{t=0}^{\infty} (P^t g)(y). \tag{98}$$

Finiteness and ergodicity give a uniform Poisson-solution bound: using the displayed geometric bound from Assumption 2.2 and $\|g\|_{\infty,2} \le \pi_{\min}^{-1/2}\|g\|_{2,\pi}$, we have, for some constant C_ν

$$\|\nu_g\|_{\infty,2} \le C_\nu \|g\|_{2,\pi}. \tag{99}$$

The constant C_ν is uniform over all centered functions g, it depends only on the transition matrix and the stationary distribution. Define

$$M_i = \nu_g(Y_{\tau+i}) - (P\nu_g)(Y_{\tau+i-1}). \tag{100}$$

By the strong Markov property at τ, $(M_i)_{i=1}^S$ is a martingale difference sequence when conditioned on $\mathcal{F}_\tau$, since

$$\mathbb{E}[\nu_g(Y_{\tau+i}) \mid \mathcal{F}_{\tau+i-1}] = (P\nu_g)(Y_{\tau+i-1}). \tag{101}$$

Moreover,

$$g(Y_{\tau+i}) = M_i + (P\nu_g)(Y_{\tau+i-1}) - (P\nu_g)(Y_{\tau+i}). \tag{102}$$

Summing over i telescopes the two boundary terms and yields

$$\sum_{i=1}^S g(Y_{\tau+i}) = \sum_{i=1}^S M_i + (P\nu_g)(Y_\tau) - (P\nu_g)(Y_{\tau+S}). \tag{103}$$

Taking conditional expectation removes the martingale sum. The remaining boundary terms combined have norm at most $2C_\nu\|g\|_{2,\pi}$. After division by S this gives the bias bound with $C_{\mathrm{B}} = 2C_\nu$.

For the second moment, using the fact that the martingale difference terms are conditionally orthogonal in Euclidean norm,

$$\mathbb{E}\Big[\Big\|\sum_{i=1}^S M_i\Big\|_2^2 \mid \mathcal{F}_\tau\Big] = \sum_{i=1}^S \mathbb{E}[\|M_i\|_2^2 \mid \mathcal{F}_\tau] \le 4SC_\nu^2\|g\|_{2,\pi}^2. \tag{104}$$

The boundary terms contribute at most $4C_\nu^2\|g\|_{2,\pi}^2$. Combining the martingale and boundary contributions with $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ and dividing by S^2, we have

$$\mathbb{E}\Big[\Big\|\frac{1}{S}\sum_{i=1}^S g(Y_{\tau+i})\Big\|_2^2 \mid \mathcal{F}_\tau\Big] \le \frac{C_{\mathrm{V}}\|g\|_{2,\pi}^2}{S} \tag{105}$$

with $C_{\mathrm{V}} = 16C_\nu^2$. □

Lemma C.2. *Suppose Assumption 2.2 and Condition 3.1 hold, and consider the PAGE-Halpern recursion (11)–(12). Let τ be the current chain time, $x, z \in \mathbb{R}^d$ be $\mathcal{F}_\tau$-measurable query points, and let $S \ge 1$ be an integer-valued $\mathcal{F}_\tau$-measurable block length. Then, conditioned on $\mathcal{F}_\tau$,*

$$\begin{aligned}
\big\|\mathbb{E}\big[R(x;\tau,S) - F(x) \mid \mathcal{F}_\tau\big]\big\|_2 &\le \frac{C_{\mathrm{B}}\sigma_F}{S}, \\
\mathbb{E}\big[\big\|R(x;\tau,S) - F(x)\big\|_2^2 \mid \mathcal{F}_\tau\big] &\le \frac{C_{\mathrm{V}}\sigma_F^2}{S}, \\
\big\|\mathbb{E}\big[D(x,z;\tau,S) - (F(x)-F(z)) \mid \mathcal{F}_\tau\big]\big\|_2 &\le \frac{C_{\mathrm{B}}L_F\|x-z\|_2}{S}, \\
\mathbb{E}\big[\big\|D(x,z;\tau,S) - (F(x)-F(z))\big\|_2^2 \mid \mathcal{F}_\tau\big] &\le \frac{C_{\mathrm{V}}L_F^2\|x-z\|_2^2}{S}.
\end{aligned} \tag{106}$$

Proof of Lemma C.2. All query points and the block length are $\mathcal{F}_\tau$-measurable. Hence, after conditioning on $\mathcal{F}_\tau$, the functions below and the block length are fixed and Lemma C.1 applies directly. For the refresh estimator, apply Lemma C.1 to

$$g_x(y) := \widehat{F}(x,y) - F(x), \tag{107}$$

whose stationary mean is zero and whose $L^2(\pi)$ norm is at most σ_F by Condition 3.1. For the difference estimator, apply Lemma C.1 to

$$g_{x,z}(y) := \widehat{F}(x,y) - \widehat{F}(z,y) - (F(x) - F(z)). \tag{108}$$

This function also has stationary mean zero, and by Condition 3.1 its $L^2(\pi)$ norm is at most $L_F\|x-z\|_2$. Substituting these two bounds into Lemma C.1 proves the statement. □

D SUPPORTING RESULTS FOR BANACH PAGE-HALPERN

Lemma D.1. *Suppose Assumption 2.2 and Condition 4.1 hold, and consider the PAGE-Halpern recursion (24)–(25). Let τ be the current chain time, $x, z \in \mathbb{R}^d$ be $\mathcal{F}_\tau$-measurable query points, and let $S \geq 1$ be an integer-valued $\mathcal{F}_\tau$-measurable block length. Then, conditioned on $\mathcal{F}_\tau$,*

$$\begin{aligned}
\big\|\mathbb{E}\big[R_T(x;\tau,S) - Tx \mid \mathcal{F}_\tau\big]\big\|_2 &\leq \frac{C_{\mathrm{B}}\sigma_T}{S}, \\
\mathbb{E}\big[\big\|R_T(x;\tau,S) - Tx\big\|_2^2 \mid \mathcal{F}_\tau\big] &\leq \frac{C_{\mathrm{V}}\sigma_T^2}{S}, \\
\big\|\mathbb{E}\big[D_T(x,z;\tau,S) - (Tx - Tz) \mid \mathcal{F}_\tau\big]\big\|_2 &\leq \frac{C_{\mathrm{B}}L_T\|x-z\|}{S}, \\
\mathbb{E}\big[\big\|D_T(x,z;\tau,S) - (Tx - Tz)\big\|_2^2 \mid \mathcal{F}_\tau\big] &\leq \frac{C_{\mathrm{V}}L_T^2\|x-z\|^2}{S}.
\end{aligned} \tag{109}$$

Proof of Lemma D.1. After conditioning on $\mathcal{F}_\tau$, the query points and block length are deterministic. For the refresh estimator R_T, we apply Lemma C.1 to $g_x(y) := H(x,y) - Tx$. For the difference estimator D_T, we apply the same lemma to

$$g_{x,z}(y) := H(x,y) - H(z,y) - (Tx - Tz). \tag{110}$$

Both functions have stationary-mean zero. Condition 4.1 bounds their $L^2(\pi)$ norms by σ_T and $L_T\|x-z\|$, respectively. Substituting these into Lemma C.1 gives the stated inequalities. □

Remark D.2 (Norm-comparison constants). The norm-comparison cost factor $\mu_{\|\cdot\|}$ appears only when Euclidean estimator estimates are converted into the working norm. In the proof of Theorem 4.3, it appears in the terminal perturbation bound (197) and squared in the increment estimate (207). For many geometries, this comparison factor is mild. For instance, for ℓ_∞ and block-sup norms, $\mu_{\|\cdot\|} \leq 1$. For the span seminorm, $\|z\|_{\mathrm{sp}} \leq 2\|z\|_\infty \leq 2\|z\|_2$, so $\mu_{\|\cdot\|} \leq 2$ before quotient normalization.

E SUPPORTING RESULTS FOR HIGH-PROBABILITY GUARANTEES

E.1 HIGH-PROBABILITY MARKOVIAN BLOCKS

We first prove the Markovian concentration estimate that replaces the second-moment block estimate used in Sections 3 and 4.

Lemma E.1 (Smooth-norm Markovian block tail)**.** *Suppose Assumption 2.2 holds. Let $g : \mathcal{Y} \to \mathbb{R}^d$ satisfy*

$$\sum_y \pi(y)g(y) = 0, \qquad \sup_{y\in\mathcal{Y}}\|g(y)\|_{\mathrm{s}} \leq M. \tag{111}$$

There exists a constant C_{hp}, depending only on the finite-state chain, such that for any stopping time τ, any integer-valued $\mathcal{F}_\tau$-measurable block length $S \geq 1$, and any $\eta \in (0,1)$, conditioned on $\mathcal{F}_\tau$,

$$\Big\|\frac{1}{S}\sum_{i=1}^{S} g(Y_{\tau+i})\Big\|_{\mathrm{s}} \leq C_{\mathrm{hp}}M\left(\sqrt{\frac{\kappa_{\mathrm{s}} + \log(1/\eta)}{S}} + \frac{1}{S}\right) \tag{112}$$

with probability at least $1-\eta$.

Proof of Lemma E.1. Condition on $\mathcal{F}_\tau$. Since S is $\mathcal{F}_\tau$-measurable, the block length is fixed before the block is sampled. It is therefore enough to prove the estimate for deterministic S.

Let $\nu_g = \sum_{t\geq 0} P^t g$ be the Poisson solution. The finite-state geometric-ergodicity bound gives

$$\sup_{y\in\mathcal{Y}}\|\nu_g(y)\|_{\mathrm{s}} \leq C_\nu M, \tag{113}$$

where C_ν depends only on the chain. Let

$$M_i := \nu_g(Y_{\tau+i}) - (P\nu_g)(Y_{\tau+i-1}), \qquad 1 \leq i \leq S. \tag{114}$$

Then $(M_i)_{i=1}^S$ is a martingale-difference sequence conditioned on $\mathcal{F}_\tau$, and by the Poisson equation,

$$\sum_{i=1}^{S} g(Y_{\tau+i}) = \sum_{i=1}^{S} M_i + (P\nu_g)(Y_\tau) - (P\nu_g)(Y_{\tau+S}). \tag{115}$$

The boundary term has $\|\cdot\|_{\mathrm{s}}$-norm at most $2C_\nu M$. Moreover, $\|M_i\|_{\mathrm{s}} \le 2C_\nu M$ almost surely. Hence the conditional light-tail hypothesis of the vector-valued Freedman inequality in κ_{s}-smooth spaces is satisfied with variance proxy of order M^2 (Luo et al., 2026, Theorem B.2). Thus, with conditional probability at least $1-\eta$,

$$\Big\|\sum_{i=1}^{S} M_i\Big\|_{\mathrm{s}} \le CM\sqrt{S(\kappa_{\mathrm{s}} + \log(1/\eta))}. \tag{116}$$

Dividing the martingale and boundary bounds by S proves (112). □

Lemma E.2 (High-probability block estimates)**.** *Suppose Assumption 2.2 and Condition 5.1 hold. Let τ be the current chain time, let x, z be $\mathcal{F}_\tau$-measurable query points, and let $S \ge 1$ be an integer-valued $\mathcal{F}_\tau$-measurable block length. Then, conditioned on $\mathcal{F}_\tau$, the refresh block satisfies*

$$\|R_T(x;\tau,S) - Tx\|_{\mathrm{s}} \le C_{\mathrm{hp}}\sigma_{\mathrm{s}}\left(\sqrt{\frac{\kappa_{\mathrm{s}} + \log(1/\eta)}{S}} + \frac{1}{S}\right) \tag{117}$$

with probability at least $1-\eta$, and the difference block satisfies

$$\|D_T(x,z;\tau,S) - (Tx - Tz)\|_{\mathrm{s}} \le C_{\mathrm{hp}}L_{\mathrm{s}}\|x-z\|\left(\sqrt{\frac{\kappa_{\mathrm{s}} + \log(1/\eta)}{S}} + \frac{1}{S}\right) \tag{118}$$

with probability at least $1-\eta$.

Proof of Lemma E.2. We simply apply Lemma E.1 to

$$g_x(y) := H(x,y) - Tx \tag{119}$$

for the refresh block, and to

$$g_{x,z}(y) := H(x,y) - H(z,y) - (Tx - Tz) \tag{120}$$

for the difference block. Both functions have stationary mean zero. Writing $\pi_{\min} := \min_{y\in\mathcal{Y}} \pi(y) > 0$, we have by Condition 5.1,

$$\sup_{y\in\mathcal{Y}}\|g_x(y)\|_{\mathrm{s}} \le \pi_{\min}^{-1/2}\sigma_{\mathrm{s}}, \qquad \sup_{y\in\mathcal{Y}}\|g_{x,z}(y)\|_{\mathrm{s}} \le \pi_{\min}^{-1/2}L_{\mathrm{s}}\|x-z\|. \tag{121}$$

Lemma E.1 therefore provides the stated estimates after absorbing $\pi_{\min}^{-1/2}$ into the chain-dependent constant C_{hp}. □

E.2 Proof of Lemma 5.3

Proof of Lemma 5.3. By Lemma E.2, the block-size schedules and a union bound over the blocks used through time N imply that, with probability at least $1-\eta/2$, every refresh block satisfies

$$\|R_T(x_n;\tau,S_{1,n}) - Tx_n\|_{\mathrm{s}} \le \frac{\sqrt{p_n}\delta}{8}, \tag{122}$$

and every difference block satisfies

$$\|D_T(x_n, x_{n-1};\tau,S_{2,n}) - (Tx_n - Tx_{n-1})\|_{\mathrm{s}} \le \frac{p_n\delta}{8\ell_N}. \tag{123}$$

The probability bounds in Lemma E.2 are conditional on the sigma-algebra at the corresponding block start. We obtain the simultaneous event above by applying the tower property successively over the blocks. On this event, the estimator error remains uniformly bounded. At a refresh iteration, we have $\|e_n\|_{\mathrm{s}} \le \delta/8$. At a difference iteration, we write

$$\zeta_n := D_T(x_n, x_{n-1};\tau,S_{2,n}) - (Tx_n - Tx_{n-1}). \tag{124}$$

The recursive update then implies

$$e_n = e_{n-1} + \zeta_n, \qquad \|\zeta_n\|_{\mathrm{s}} \le \frac{p_n \delta}{8\ell_N}. \tag{125}$$

Since $\sum_{n=1}^{N} p_n \le 2\ell_N$, the total error accumulated during any sequence of difference iterations is at most $\delta/4$. Together with the refresh-block bound, this proves (40).

We next control the scaled increments. For $n \ge 2$, we define

$$\Delta_n := \beta_n e_{n-1} - \beta_{n-1} e_{n-2}. \tag{126}$$

If iteration $n-1$ uses a difference block, then

$$\Delta_n = (\beta_n - \beta_{n-1}) e_{n-2} + \beta_n \zeta_{n-1}. \tag{127}$$

By (40), (123), and $\beta_n - \beta_{n-1} = 1/(n(n+1))$, we have

$$n\|\Delta_n\|_{\mathrm{s}} \le \frac{C\delta}{n} + \frac{C\delta}{\ell_N}. \tag{128}$$

Summing over the recursive iterations gives a contribution of at most $C\delta N$. If iteration $n-1$ instead uses a refresh block, then by (40),

$$n\|\Delta_n\|_{\mathrm{s}} \le Cn\delta. \tag{129}$$

We let τ_m denote the chain time immediately before the PAGE coin is tossed at iteration m. We have

$$\mathbb{E}\big[\mathbf{1}_{\mathcal{R}_m} \mid \mathcal{F}_{\tau_m}\big] = p_m = \frac{2}{m+1}, \qquad m \ge 1. \tag{130}$$

Therefore, by the scalar Bernstein–Freedman inequality applied to the martingale differences $(m+1)(\mathbf{1}_{\mathcal{R}_m} - p_m)$, with probability at least $1 - \eta/2$,

$$\sum_{m=1}^{N} (m+1)\mathbf{1}_{\mathcal{R}_m} \le CN(1 + \log(1/\eta)). \tag{131}$$

Combining the difference and refresh contributions proves (40). Since $p_m = 2/(m+1)$, the same estimate also proves (41). □

E.3 High-probability residual bound

Lemma E.3 (Pathwise displacement-level residual)**.** *Let $\beta_n = n/(n+1)$, and consider the recursion*

$$x_n = (1-\beta_n)x_0 + \beta_n(Tx_{n-1} + U_n), \qquad n \ge 1. \tag{132}$$

Let $U_0 := 0$ and $\Delta_n := \beta_n U_n - \beta_{n-1} U_{n-1}$. Suppose that, up to horizon N,

$$\sup_{0 \le n \le N} \|Tx_n - x_0\| \le K, \qquad \|U_N\| \le r, \qquad \sum_{n=1}^{N} n\|\Delta_n\| \le A. \tag{133}$$

Then,

$$\|x_N - Tx_N\| \le \frac{K(1 + \log(N+1))}{N+1} + \frac{A}{N+1} + r. \tag{134}$$

Proof of Lemma E.3. Let $d_n := x_n - x_{n-1}$. Subtracting two consecutive inexact Halpern updates gives

$$d_n = (\beta_n - \beta_{n-1})(Tx_{n-1} - x_0) + \beta_{n-1}(Tx_{n-1} - Tx_{n-2}) + \Delta_n. \tag{135}$$

By non-expansiveness of T and the assumed pathwise bound,

$$\|d_n\| \le \frac{K}{n(n+1)} + \frac{n-1}{n}\|d_{n-1}\| + \|\Delta_n\|. \tag{136}$$

Multiplying by n and summing from 1 to N yields

$$N\|d_N\| \le K \sum_{n=1}^{N} \frac{1}{n+1} + \sum_{n=1}^{N} n\|\Delta_n\| \le K \log(N+1) + A. \tag{137}$$

The final residual satisfies

$$\|x_N - Tx_N\| \le \frac{K}{N+1} + \frac{N}{N+1}\|d_N\| + r. \tag{138}$$

Substituting the displacement bound proves the result. □

F EXPERIMENTAL DETAILS

F.1 HILBERT-SPACE POLICY EVALUATION

We arrange the states $\{0, \dots, 7\}$ on a cycle. The resulting process is a lazy nearest-neighbor random walk. At each step, the state remains unchanged with probability $1 - 2q$ and moves to either neighboring state with probability q:

$$P(s, s) = 1 - 2q, \qquad P(s, s - 1 \bmod 8) = P(s, s + 1 \bmod 8) = q. \tag{139}$$

Smaller values of q make the trajectory remain near its current state for longer. Since the transition matrix is symmetric, its stationary distribution is uniform. We choose the alternating differential value function $\phi(s) = (-1)^s$. The transition reward is

$$r(s, s') = \phi(s) - \phi(s') + 0.2\big(\mathbf{1}\{s' = s + 1 \bmod 8\} - \mathbf{1}\{s' = s - 1 \bmod 8\}\big). \tag{140}$$

The first term is a potential difference. Along any realized trajectory, these differences telescope according to

$$\sum_{t=0}^{m-1} \big(\phi(S_t) - \phi(S_{t+1})\big) = \phi(S_0) - \phi(S_m). \tag{141}$$

Thus, the relative return associated with the initial state is described by ϕ. The second term does not alter this mean behavior. It is 0.2 for a clockwise transition, -0.2 for a counterclockwise transition, and zero for a self-transition. The clockwise and counterclockwise moves have equal probability, so this term has conditional mean zero. We include it to ensure that individual transition samples remain noisy even at the exact solution. More precisely, the mean reward satisfies

$$\bar{r}(s) := \mathbb{E}\big[r(S_t, S_{t+1}) \mid S_t = s\big] = \phi(s) - (P\phi)(s). \tag{142}$$

Let π denote the uniform stationary distribution. Since $\pi^\top P = \pi^\top$, we have

$$\sum_s \pi(s)\bar{r}(s) = \pi^\top(\phi - P\phi) = \pi^\top\phi - (\pi^\top P)\phi = 0. \tag{143}$$

Equivalently, under stationarity, S_t and S_{t+1} have the same marginal distribution, so the potential difference has mean zero. Thus, the average reward is zero. The differential Bellman equation is (Tsitsiklis & Van Roy, 1999; Wan et al., 2021)

$$h = \bar{r} + Ph. \tag{144}$$

Substitution shows directly that $h^* = \phi$ is a solution, since $\bar{r} + P\phi = \phi$. We define the Bellman operator by $Bh := \bar{r} + Ph$. The matrix P averages the values at the current state and its two neighbors. Since this averaging matrix is symmetric, all of its eigenvalues lie in $[-1, 1]$. Consequently,

$$\|Bh - Bz\|_2 = \|P(h - z)\|_2 \le \|h - z\|_2, \tag{145}$$

so this problem falls within the Hilbert-space theory of Section 3. The algorithm observes the continuing transition process $Y_t = (S_t, S_{t+1})$. For a candidate value vector h, the temporal-difference error associated with one transition is

$$\delta_h(s, s') := r(s, s') + h(s') - h(s). \tag{146}$$

For e_s denoting the sth coordinate vector, we define the single-transition operator evaluation

$$H_{\text{eval}}(h, (s, s')) = h + 8e_s\delta_h(s, s'). \tag{147}$$

This operator samples the Bellman equation only at the observed state s. The multiplier eight is an importance weight rather than an algorithmic stepsize. Each state appears with stationary probability $1/8$, so the multiplier converts the randomly selected coordinate correction into the full Bellman correction in expectation. Indeed,

$$\mathbb{E}_\pi\left[8e_{S_t}\delta_h(S_t, S_{t+1})\right] = \bar{r} + Ph - h, \tag{148}$$

and hence $\mathbb{E}_\pi[H_{\text{eval}}(h, Y)] = Bh$. Once a transition (s, s') has been observed, the expression in (147) can be evaluated at two query points. The same-state difference oracle required by PAGE is therefore available without observing an additional transition.

F.2 Banach-space average-reward control

The second example retains the state process and transition noise from the policy-evaluation problem. We introduce two actions $\mathcal{A} = \{0, 1\}$, but the selected action does not affect the transition matrix. The behavior policy selects each action with probability $1/2$, so the stationary state-action distribution is uniform. Action one incurs a penalty $\Delta = 0.5$, and the reward is

$$r(s, a, s') = \phi(s) - \phi(s') - \Delta a + 0.2\big(\mathbf{1}\{s' = s + 1 \bmod 8\} - \mathbf{1}\{s' = s - 1 \bmod 8\}\big). \quad (149)$$

Action zero is therefore optimal. The control decision is deliberately simple. The purpose of introducing the actions is to place the maximization operator inside the Bellman map while keeping the Markovian sampling problem unchanged. The mean reward is $\bar{r}(s, a) = \phi(s) - (P\phi)(s) - \Delta a$, the average-reward optimality operator is

$$(\mathcal{T}Q)(s, a) = \bar{r}(s, a) + \sum_{s'} P(s, s') \max_{b \in \mathcal{A}} Q(s', b), \quad (150)$$

and the optimal average reward is zero. We have the fixed point

$$Q^*(s, a) = \phi(s) - \Delta a, \quad (151)$$

since $\max_b Q^*(s', b) = Q^*(s', 0) = \phi(s')$. Therefore,

$$(\mathcal{T}Q^*)(s, a) = \phi(s) - (P\phi)(s) - \Delta a + (P\phi)(s) = Q^*(s, a). \quad (152)$$

Although the optimal action is simple, the operator $\mathcal{T}$ is globally nonlinear because the maximizing action changes with the input Q. For any $Q, Z \in \mathbb{R}^{16}$, we have

$$\begin{aligned} \left| \max_b Q(s, b) - \max_b Z(s, b) \right| &\le \max_b |Q(s, b) - Z(s, b)| \le \|Q - Z\|_\infty, \\ \|\mathcal{T}Q - \mathcal{T}Z\|_\infty &\le \max_s \sum_{s'} P(s, s') \|Q - Z\|_\infty = \|Q - Z\|_\infty. \end{aligned} \quad (153)$$

Thus, $\mathcal{T}$ is non-expansive in the sup norm, whereas no norm induced by an inner product makes $\mathcal{T}$ non-expansive. The observed process is now $Y_t = (S_t, A_t, S_{t+1})$. For a candidate action-value vector Q, we define the sampled optimality error by

$$\delta_Q(s, a, s') := r(s, a, s') + \max_{b \in \mathcal{A}} Q(s', b) - Q(s, a). \quad (154)$$

For $e_{(s,a)}$ denoting the corresponding coordinate vector in $\mathbb{R}^{16}$, the single-transition evaluation is

$$H_{\mathrm{ctl}}(Q, (s, a, s')) = Q + 16 e_{(s,a)} \delta_Q(s, a, s'). \quad (155)$$

The multiplier sixteen is again an importance weight. Each of the sixteen state-action pairs has stationary probability $1/16$, and hence $\mathbb{E}_\pi[H_{\mathrm{ctl}}(Q, Y)] = \mathcal{T}Q$. Once (s, a, s') has been observed, the sampled optimality error can be evaluated at two different vectors Q and Z without another environment transition. The PAGE difference evaluation is therefore available in the same online trajectory.

F.3 Methods and evaluation

We use $q = 0.24$ for the fast-mixing chain and $q = 0.04$ for the slow-mixing chain. The corresponding second-largest eigenvalue moduli are 0.859 and 0.977. We initialize the policy-evaluation iterate at $h_0 = 0$ and the control iterate at $Q_0 = 0$. We generate twenty independent trajectories for each problem and each chain. Within every repetition, the block and PAGE methods use the same realized trajectory. Each Markov chain is initialized at $S_0 = 0$ rather than from its stationary distribution.

For the block method, we use $k_n = \lceil 0.01 n^4 \rceil$ and the burn-in schedule $b_n = \lceil 2 \log(n+1) / \log(1/\rho) \rceil$ with the exact eigenvalue modulus in place of ρ. This optional burn-in reduces finite-horizon sensitivity to the initial state and block boundaries. For both PAGE methods, we use $p_n = \min\{1, 2/(n+1)\}$ and the fixed estimator scale $\delta = 0.02$. The problem-dependent constants in the PAGE block schedules are replaced by a common multiplier $c = 0.15$. Thus,

$$S_{1,0} = \left\lceil \frac{c}{\delta^2} \right\rceil, \qquad S_{1,n} = \left\lceil \frac{c}{p_n \delta^2} \right\rceil, \quad (156)$$

Table 1: Median number of observed transitions required to reach a common residual target. The counts in parentheses give the number of repetitions that reach the target when fewer than twenty do so. The notation $> 2{,}048{,}000$ means that the median transition count exceeds the available budget.

Problem	Mixing	Method	$\epsilon = 0.20$	$\epsilon = 0.10$	$\epsilon = 0.05$
Evaluation	Fast	Block Halpern	585	1,238	8,839
		Hilbert PAGE-Halpern	4,567	9,001	14,684
	Slow	Block Halpern	68,885	1,923,959	$> 2{,}048{,}000$ (0/20)
		Hilbert PAGE-Halpern	25,705	62,611	107,374
Control	Fast	Block Halpern	6,205	21,021	63,081
		Banach PAGE-Halpern	3,302	8,606	20,749
	Slow	Block Halpern	51,481 (19/20)	289,018 (17/20)	$> 2{,}048{,}000$ (8/20)
		Banach PAGE-Halpern	13,430	46,415	80,000

and

$$S_{2,n} = \left\lceil \max\left\{ \frac{cd_n^2}{p_n^2\delta^2}, \frac{cd_n}{p_n^{3/2}\delta}, 1 \right\} \right\rceil. \tag{157}$$

For policy evaluation, $d_n = \|h_n - h_{n-1}\|_2$, and the Hilbert method uses $L^* = 2$. For control, $d_n = \|Q_n - Q_{n-1}\|_\infty$, and the Banach method applies the estimator of $\mathcal{T}Q_n$ directly. The same parameters are fixed for both chains and all transition budgets.

We evaluate all methods at the common budgets

$$B_j = 500 \cdot 2^j, \qquad j = 0, \dots, 12. \tag{158}$$

At budget B_j, we report the last iterate available after at most B_j transitions. If the method is partway through a block at that time, the incomplete block is not used and the preceding iterate is reported. Each method consumes one uninterrupted trajectory. Every transition is included in the budget, including the transitions discarded during the optional burn-in periods of the baseline block method. A PAGE difference block evaluates the oracle at two query points for each observed transition. We use the number of environment transitions as the primary cost. For policy evaluation, we report

$$\frac{\|h - \bar{r} - Ph\|_2}{\|\bar{r}\|_2}. \tag{159}$$

For control, we report

$$\frac{\|Q - \mathcal{T}Q\|_\infty}{\|\bar{r}\|_\infty}. \tag{160}$$

Because a small number of trajectories can produce unusually large residuals, Figure 1 reports the median residual at each common budget and shades the interquartile range from the 25th to the 75th percentile. We also record the first completed iterate whose residual is at most ϵ for $\epsilon \in \{0.20, 0.10, 0.05\}$. These transition counts are reported in Table 1. If a run does not reach a target within $B_{12} = 2{,}048{,}000$ transitions, we record it as unsuccessful at that target.

G Proof of the displacement-level bound

G.1 Proof of Theorem 2.5

Proof of Theorem 2.5. Let $d_n := x_n - x_{n-1}$. By the Halpern recursion,

$$d_n = (\beta_n - \beta_{n-1})(Tx_{n-1} - x_0) + \beta_{n-1}(Tx_{n-1} - Tx_{n-2}) + \Delta_n, \tag{161}$$

(with the second term omitted when $n = 1$.) By the step size schedule, we also have

$$\beta_n - \beta_{n-1} = \frac{1}{n(n+1)}, \qquad \beta_{n-1} = \frac{n-1}{n}. \tag{162}$$

By Assumption 2.4 and using non-expansiveness,

$$\mathbb{E}\big[\|d_n\|\big] \le \frac{\kappa_1}{n(n+1)} + \frac{n-1}{n}\mathbb{E}\big[\|d_{n-1}\|\big] + a_n. \tag{163}$$

Multiplying both sides by n results in the telescoping recursion

$$n\mathbb{E}\big[\|d_n\|\big] \le \frac{\kappa_1}{(n+1)} + (n-1)\mathbb{E}\big[\|d_{n-1}\|\big] + na_n. \tag{164}$$

Summing both sides from 1 to N:

$$N\mathbb{E}\big[\|d_N\|\big] \le \kappa_1 \sum_{n=1}^{N} \frac{1}{n+1} + \sum_{n=1}^{N} na_n \le \kappa_1 \log(N+1) + \sum_{n=1}^{N} na_n. \tag{165}$$

Now, the final iterate residual is

$$x_N - Tx_N = (1-\beta_N)(x_0 - Tx_N) + \beta_N(Tx_{N-1} - Tx_N) + \beta_N U_N. \tag{166}$$

By non-expansiveness and Assumption 2.4,

$$\mathbb{E}\big[\|x_N - Tx_N\|\big] \le \frac{\kappa_1}{N+1} + \frac{N}{N+1}\mathbb{E}\big[\|d_N\|\big] + r. \tag{167}$$

Using (165) and increasing $1+\log(N+1)$ to absorb the extra $\kappa_1/(N+1)$ term yields the first claim. The remaining statements follow by Jensen's inequality. □

H Proofs of the main Hilbert-space results

H.1 Proof of Lemma 3.2

Proof of Lemma 3.2. Let $e_n := v_n - F(x_n)$ and $a_n := \mathbb{E}\big[\|e_n\|_2^2\big]$. The proof is an induction on the estimator error. The initialization uses a refresh block whose size satisfies the first inequality in (13), so Lemma C.2 directly gives

$$a_0 = \mathbb{E}\big[\|v_0 - F(x_0)\|_2^2\big] \le \delta^2. \tag{168}$$

Now, fix $n \ge 1$ and assume the induction hypothesis

$$a_{n-1} \le \frac{100\delta^2}{n}. \tag{169}$$

At the moment of the PAGE coin toss, x_n, x_{n-1}, v_{n-1}, and therefore e_{n-1}, are all measurable with respect to the past filtration. The next block is sampled only after the branch and block length have been chosen. By the refresh-block bound of Lemma C.2 and the second inequality in (13),

$$\mathbb{E}\big[\|R(x_n;\tau,S_{1,n}) - F(x_n)\|_2^2\big] \le p_n\delta^2. \tag{170}$$

This is the contribution to a_n in case that a refresh block is used at iteration n, the probability of which is p_n. In the case a difference block is used, we write

$$\zeta_n := D(x_n, x_{n-1};\tau,S_{2,n}) - (F(x_n) - F(x_{n-1})). \tag{171}$$

By Lemma C.2 and the third inequality in (13),

$$\mathbb{E}\big[\|\zeta_n\|_2^2 \mid \mathcal{F}_\tau\big] \le p_n^2\delta^2, \qquad \big\|\mathbb{E}\big[\zeta_n \mid \mathcal{F}_\tau\big]\big\|_2 \le p_n^{3/2}\delta. \tag{172}$$

In this case, we have by the update rule in (11)–(12) that

$$v_n = v_{n-1} + D(x_n, x_{n-1};\tau,S_{2,n}), \tag{173}$$

so the estimator error evolves as

$$e_n = v_n - F(x_n) = \big(v_{n-1} + D(x_n,x_{n-1};\tau,S_{2,n})\big) - F(x_n) + \big(F(x_{n-1}) - F(x_{n-1})\big) = e_{n-1} + \zeta_n. \tag{174}$$

Since e_{n-1} is $\mathcal{F}_\tau$-measurable, we have

$$\begin{aligned}\mathbb{E}\big[\big\|e_{n-1}+\zeta_n\big\|_2^2\big] &= a_{n-1} + 2\mathbb{E}\big[\langle e_{n-1}, \mathbb{E}[\zeta_n \mid \mathcal{F}_\tau]\rangle\big] + \mathbb{E}\big[\|\zeta_n\|_2^2\big] \\ &\le a_{n-1} + 2\sqrt{a_{n-1}}\,p_n^{3/2}\delta + p_n^2\delta^2 \\ &\le a_{n-1} + 21p_n^2\delta^2,\end{aligned} \tag{175}$$

where the last inequality uses the induction hypothesis and $p_n^{3/2}/\sqrt{n} \le p_n^2$, which follows from $p_n = 2/(n+1)$ for $n \ge 1$. Thus the difference branch has conditional second moment at most $a_{n-1} + 21p_n^2\delta^2$. Combining the bounds for the refresh and difference branches gives

$$a_n \le p_n^2\delta^2 + (1-p_n)\left(\frac{100\delta^2}{n} + 21p_n^2\delta^2\right). \tag{176}$$

For $n \ge 1$, $p_n = 2/(n+1)$, hence

$$(1-p_n)\frac{100\delta^2}{n} = \frac{100(n-1)\delta^2}{n(n+1)} = \frac{100\delta^2}{n+1} - \frac{100\delta^2}{n(n+1)}. \tag{177}$$

The remaining terms satisfy $p_n^2\delta^2 + (1-p_n)21p_n^2\delta^2 \le 22p_n^2\delta^2 \le 88\delta^2/(n(n+1))$, and this is absorbed by the negative term $-100\delta^2/(n(n+1))$ in (177). Therefore

$$a_n \le \frac{100\delta^2}{n+1}. \tag{178}$$

This proves the induction. □

H.2 Proof of Proposition 3.3

Proof of Proposition 3.3. We state a known result on cocoercive monotone inclusions in the notation of the present paper for completeness, see Cai et al. (2022). Under the stated estimator accuracy condition, the PAGE–Halpern iterates satisfy (Cai et al., 2022, Theorem 3.1)

$$\mathbb{E}\big[\|F(x_N)\|_2\big] \le \frac{C_0 L^* R}{N} + C_1\delta \tag{179}$$

for universal constants C_0, C_1. Although the cited theorem states unbiasedness at initialization, its proof uses this property only to remove the initial mixed error term. Bounding that term by Cauchy's inequality under the displayed second-moment condition changes only the universal constant C_1. Furthermore, with $p_n = 2/(n+1)$ and $\delta N \le c_2 L^* R$, the weighted displacement accumulated by the PAGE recursion satisfies (Cai et al., 2022, Lemma 3.3)

$$\sum_{n=1}^{N} \frac{\mathbb{E}\big[\|x_n - x_{n-1}\|_2^2\big]}{p_n^2} \le C_2 R^2 N \tag{180}$$

for a universal constant C_2. □

H.3 Proof of Theorem 3.4

Proof of Theorem 3.4. The claim is immediate in the two trivial cases identified in the theorem, so we assume that $R > 0$ and $\epsilon < 2R$. The operator $F = I - T$ is monotone and $1/2$-cocoercive, as shown at the start of this section. Since $L^* \ge 2$, it is also $1/L^*$-cocoercive. Thus, the conditions required by Proposition 3.3 are satisfied. Lemma 3.2 guarantees $\mathbb{E}\big[\|v_n - F(x_n)\|_2^2\big] \le 100\delta^2/(n+1)$, so the residual conclusion of Proposition 3.3 applies with estimator scale 10δ. Hence

$$\mathbb{E}\big[\|F(x_N)\|_2\big] \le \frac{C_0 L^* R}{N} + 10C_1\delta \le \epsilon, \tag{181}$$

since the choices of N and δ make each of the two terms at most $\epsilon/2$. Because $F(x_N) = x_N - Tx_N$, this is exactly the fixed-point residual bound. For the sample-complexity calculation, we use the displacement conclusion of Proposition 3.3. We may enlarge the universal constant C_1 so that $C_1 \ge (2C_0+1)/(2c_2)$. Since $L^* \ge 2$ and $\epsilon < 2R$, the choices of N and δ then give

$$10\delta N \le \frac{2C_0 L^* R + \epsilon}{2C_1} \le c_2 L^* R. \tag{182}$$

Thus, the displacement bound in Proposition 3.3 applies with estimator scale 10δ. It remains to bound the sample complexity. We count expected Markov transitions along the single continuing chain. At time n, a refresh block is used with probability p_n, while its prescribed length is

$$S_{1,n} = O\left(\frac{C_V \sigma_F^2}{p_n \delta^2}\right). \tag{183}$$

Therefore the refresh cost satisfies

$$\sum_{n=1}^{N} p_n S_{1,n} = O\left(\frac{C_{\mathrm{V}}\sigma_F^2 N}{\delta^2}\right). \tag{184}$$

The unit term in the third inequality of (13) contributes at most $O(N)$, which is lower order compared with the other two accuracy-dependent terms. For the first term in the maximum, the variance-control part of difference batches, the probability of occurrence is at most one, so it is enough to sum the prescribed variance-control lengths. Using (17),

$$\sum_{n=1}^{N} \frac{C_{\mathrm{V}}L_F^2}{\delta^2}\frac{\mathbb{E}\big[\|x_n - x_{n-1}\|_2^2\big]}{p_n^2} = O\left(\frac{C_{\mathrm{V}}L_F^2 R^2 N}{\delta^2}\right). \tag{185}$$

The second term in the maximum in the third inequality of (13), which controls the bias, contributes

$$O\left(\frac{C_{\mathrm{B}}L_F}{\delta}\sum_{n=1}^{N}\frac{\mathbb{E}\big[\|x_n - x_{n-1}\|_2\big]}{p_n^{3/2}}\right). \tag{186}$$

To bound the remaining sum, apply Cauchy's inequality:

$$\begin{aligned}\sum_{n=1}^{N}\frac{\mathbb{E}\big[\|x_n - x_{n-1}\|_2\big]}{p_n^{3/2}} &\le \left(\sum_{n=1}^{N}\frac{\mathbb{E}\big[\|x_n - x_{n-1}\|_2^2\big]}{p_n^2}\right)^{1/2}\left(\sum_{n=1}^{N}\frac{1}{p_n}\right)^{1/2} \\ &= O(RN^{3/2}).\end{aligned} \tag{187}$$

The initialization $S_{1,0} = O(C_{\mathrm{V}}\sigma_F^2/\delta^2)$ is dominated by the refresh sum when $N \ge 1$. Substituting $N = O(L^*R/\epsilon)$, $\delta = \Theta(\epsilon)$, and $L_F \le L^*$, the refresh term contributes

$$O\left(\frac{C_{\mathrm{V}}\sigma_F^2 L^* R}{\epsilon^3}\right), \tag{188}$$

the variance-control difference term contributes

$$O\left(\frac{C_{\mathrm{V}}(L^*R)^3}{\epsilon^3}\right), \tag{189}$$

and the bias-control difference term contributes

$$O\left(\frac{C_{\mathrm{B}}(L^*R)^{5/2}}{\epsilon^{5/2}}\right). \tag{190}$$

Combining these contributions gives (20). □

I Proofs of the main Banach-space results

I.1 Proof of Lemma 4.2

Proof of Lemma 4.2. Let $e_n := v_n - Tx_n$ and $a_n := \mathbb{E}\big[\|e_n\|_2^2\big]$. The proof is identical in structure to Lemma 3.2, with T replacing F. $a_0 \le \delta^2$ follows from the refresh block bound in Lemma D.1 and the first inequality in (26). Now, assume $a_{n-1} \le 100\delta^2/n$. In the case of using a refresh block at iteration n, the second inequality in (26) gives

$$\mathbb{E}\big[\|R_T(x_n;\tau,S_{1,n}) - Tx_n\|_2^2\big] \le p_n\delta^2. \tag{191}$$

In the case of using a difference block, write

$$\zeta_n := D_T(x_n, x_{n-1};\tau,S_{2,n}) - (Tx_n - Tx_{n-1}). \tag{192}$$

By Lemma D.1 and the third inequality in (26),

$$\mathbb{E}\big[\|\zeta_n\|_2^2 \mid \mathcal{F}_\tau\big] \le p_n^2\delta^2, \qquad \big\|\mathbb{E}[\zeta_n \mid \mathcal{F}_\tau]\big\|_2 \le p_n^{3/2}\delta \tag{193}$$

In this branch, $e_n = e_{n-1} + \zeta_n$. Since e_{n-1} is $\mathcal{F}_\tau$-measurable,

$$\begin{aligned}\mathbb{E}\big[\|e_{n-1}+\zeta_n\|_2^2\big] &= a_{n-1} + 2\mathbb{E}\big[\langle e_{n-1}, \mathbb{E}[\zeta_n \mid \mathcal{F}_\tau]\rangle\big] + \mathbb{E}\big[\|\zeta_n\|_2^2\big] \\ &\le a_{n-1} + 2\sqrt{a_{n-1}}p_n^{3/2}\delta + p_n^2\delta^2 \\ &\le a_{n-1} + 21p_n^2\delta^2.\end{aligned} \tag{194}$$

Combining the bounds for the two blocks,

$$a_n \le p_n^2\delta^2 + (1-p_n)\left(\frac{100\delta^2}{n} + 21p_n^2\delta^2\right). \tag{195}$$

Finally, the same algebra as in the proof of Lemma 3.2 yields $a_n \le 100\delta^2/(n+1)$. □

I.2 Proof of Theorem 4.3

Proof of Theorem 4.3. The proof will follow through establishing bounds on the final perturbation $\mathbb{E}\big[\|U_N\|\big]$ and on the scaled perturbation increments Δ_n in Theorem 2.5. Let $e_n := v_n - Tx_n$. We start by writing

$$x_n = (1-\beta_n)x_0 + \beta_n v_{n-1} = (1-\beta_n)x_0 + \beta_n(Tx_{n-1} + e_{n-1}). \tag{196}$$

Thus (24) is exactly (4) with $U_n = e_{n-1}$. Lemma 4.2 provides the Euclidean estimator bound (27). Converting this estimate to the non-expansiveness norm gives us a bound on the final perturbation norm:

$$\mathbb{E}\big[\|U_N\|\big] = \mathbb{E}\big[\|e_{N-1}\|\big] \le \mu_{\|\cdot\|}\mathbb{E}\big[\|e_{N-1}\|_2\big] \le \mu_{\|\cdot\|}\big(\mathbb{E}\big[\|e_{N-1}\|_2^2\big]\big)^{1/2} \le C\mu_{\|\cdot\|}\delta. \tag{197}$$

It remains to control the scaled perturbation increments

$$\Delta_n = \beta_n U_n - \beta_{n-1}U_{n-1} = \beta_n e_{n-1} - \beta_{n-1}e_{n-2}, \tag{198}$$

with the convention $e_{-1} = 0$. For $n = 1$, we have $\Delta_1 = \beta_1 e_0$, and (27) implies $\mathbb{E}\big[\|\Delta_1\|_2^2\big] \le C\delta^2$. Now, fix $n \ge 2$, and let τ be the current chain time immediately before the PAGE coin is tossed at iteration $n-1$. The variables x_{n-1}, x_{n-2}, v_{n-2}, and hence e_{n-2}, are $\mathcal{F}_\tau$-measurable. Let $\mathcal{R}_{n-1}$ and $\mathcal{D}_{n-1}$ denote the events that the coin selects a refresh block and a difference block, respectively. Since the PAGE coin is independent of $\mathcal{F}_\tau$,

$$\mathbb{P}(\mathcal{R}_{n-1} \mid \mathcal{F}_\tau) = p_{n-1}, \qquad \mathbb{P}(\mathcal{D}_{n-1} \mid \mathcal{F}_\tau) = 1 - p_{n-1}. \tag{199}$$

On the event $\mathcal{D}_{n-1}$, we define

$$\zeta_{n-1} := D_T(x_{n-1}, x_{n-2}; \tau, S_{2,n-1}) - (Tx_{n-1} - Tx_{n-2}), \tag{200}$$

so that $e_{n-1} = e_{n-2} + \zeta_{n-1}$. Therefore,

$$\Delta_n = (\beta_n - \beta_{n-1})e_{n-2} + \beta_n\zeta_{n-1}. \tag{201}$$

Using $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$, the estimator bound (27), and the difference-block bound in Lemma 4.2, the contribution of $\mathcal{D}_{n-1}$ to the second moment of Δ_n satisfies

$$\begin{aligned}\mathbb{E}[\|\Delta_n\|_2^2\mathbf{1}_{\mathcal{D}_{n-1}}] &\le 2(\beta_n-\beta_{n-1})^2\mathbb{E}[\|e_{n-2}\|_2^2\mathbf{1}_{\mathcal{D}_{n-1}}] + 2\beta_n^2\mathbb{E}[\|\zeta_{n-1}\|_2^2\mathbf{1}_{\mathcal{D}_{n-1}}] \\ &\le 2(\beta_n-\beta_{n-1})^2\mathbb{E}[\|e_{n-2}\|_2^2] + 2\beta_n^2p_{n-1}^2\delta^2 \\ &\le C\left((\beta_n-\beta_{n-1})^2\frac{\delta^2}{n} + p_{n-1}^2\delta^2\right) \le \frac{C\delta^2}{n^2}.\end{aligned} \tag{202}$$

Here we use $\beta_n - \beta_{n-1} = 1/(n(n+1))$, $\beta_n \le 1$, and $p_{n-1} = O(1/n)$. On the event $\mathcal{R}_{n-1}$, by the refresh block schedule

$$\mathbb{E}\big[\|e_{n-1}\|_2^2 \mid \mathcal{F}_\tau, \mathcal{R}_{n-1}\big] \le p_{n-1}\delta^2, \tag{203}$$

while by (27),

$$\mathbb{E}\big[\|e_{n-2}\|_2^2\big] \le \frac{C\delta^2}{n}. \tag{204}$$

Using $\mathbb{E}\big[\mathbf{1}_{\mathcal{R}_{n-1}} \mid \mathcal{F}_\tau\big] = p_{n-1}$ and $\|a-b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$,

$$\begin{aligned}\mathbb{E}\big[\|\Delta_n\|_2^2 \mathbf{1}_{\mathcal{R}_{n-1}}\big] &\le C\mathbb{E}\big[\|e_{n-1}\|_2^2 \mathbf{1}_{\mathcal{R}_{n-1}}\big] + C\mathbb{E}\big[\|e_{n-2}\|_2^2 \mathbf{1}_{\mathcal{R}_{n-1}}\big] \\ &\le Cp_{n-1}\left(p_{n-1}\delta^2 + \frac{\delta^2}{n}\right) \\ &\le \frac{C\delta^2}{n^2},\end{aligned} \tag{205}$$

where we use $p_{n-1} = O(1/n)$ in the last inequality. Since $\mathcal{R}_{n-1}$ and $\mathcal{D}_{n-1}$ constitute a partition of the sample space, combining their contributions yields

$$\mathbb{E}\big[\|\Delta_n\|_2^2\big] \le \frac{C\delta^2}{n^2}, \qquad n \ge 1. \tag{206}$$

By the norm comparison in Section 2,

$$\mathbb{E}\big[\|\Delta_n\|^2\big] \le \mu_{\|\cdot\|}^2 \mathbb{E}\big[\|\Delta_n\|_2^2\big] \le \frac{C\mu_{\|\cdot\|}^2\delta^2}{n^2}. \tag{207}$$

Applying Theorem 2.5 with (197) and (207) proves the residual bound statement (28). We now establish the sample complexity result. A refresh block is used with probability p_n and has length

$$S_{1,n} = O\left(\frac{C_{\mathrm{V}}\sigma_T^2}{p_n\delta^2}\right), \tag{208}$$

so the expected refresh cost through time N is

$$O\left(\frac{C_{\mathrm{V}}\sigma_T^2 N}{\delta^2}\right). \tag{209}$$

For the difference blocks, we need a displacement estimate in $\|\cdot\|$. Let

$$d_n := x_n - x_{n-1}, \qquad D_n := \big(\mathbb{E}\big[\|d_n\|^2\big]\big)^{1/2}. \tag{210}$$

From the inexact Halpern recursion,

$$d_n = (\beta_n - \beta_{n-1})(Tx_{n-1} - x_0) + \beta_{n-1}(Tx_{n-1} - Tx_{n-2}) + \Delta_n. \tag{211}$$

Taking L^2 norms, using Minkowski's inequality, non-expansiveness of T, Lemma B.2, and (207), we obtain

$$D_n \le \frac{\kappa}{n(n+1)} + \frac{n-1}{n}D_{n-1} + \frac{C\mu_{\|\cdot\|}\delta}{n}. \tag{212}$$

Multiplying by n,

$$nD_n \le \frac{\kappa}{n+1} + (n-1)D_{n-1} + C\mu_{\|\cdot\|}\delta. \tag{213}$$

Summing this telescoping recursion from 1 through n, we get

$$D_n \le \frac{C\kappa(1+\log n)}{n} + C\mu_{\|\cdot\|}\delta. \tag{214}$$

Since $p_n^{-2} = O(n^2)$, by (214),

$$\sum_{n=1}^{N} \frac{\mathbb{E}\big[\|x_n - x_{n-1}\|^2\big]}{p_n^2} \le C\sum_{n=1}^{N} n^2\left(\frac{\kappa^2(1+\log n)^2}{n^2} + \mu_{\|\cdot\|}^2\delta^2\right) = \tilde{O}\big(\kappa^2 N + \mu_{\|\cdot\|}^2\delta^2 N^3\big). \tag{215}$$

The variance-control part of the difference-block cost is therefore

$$\tilde{O}\left(\frac{C_{\mathrm{V}}L_T^2}{\delta^2}\big(\kappa^2 N + \mu_{\|\cdot\|}^2\delta^2 N^3\big)\right). \tag{216}$$

Similarly, using (214) and $p_n^{-3/2} = O(n^{3/2})$,

$$\sum_{n=1}^{N} \frac{\mathbb{E}\big[\|x_n - x_{n-1}\|\big]}{p_n^{3/2}} = \tilde{O}\big(\kappa N^{3/2} + \mu_{\|\cdot\|}\delta N^{5/2}\big). \tag{217}$$

Thus the bias-control part of the difference-block cost is

$$\tilde{O}\left(\frac{C_{\mathrm{B}}L_T}{\delta}\left(\kappa N^{3/2}+\mu_{\|\cdot\|}\delta N^{5/2}\right)\right). \tag{218}$$

Finally, substituting

$$N=\tilde{\Theta}(\kappa/\epsilon), \qquad \delta=\Theta(\epsilon/\mu_{\|\cdot\|}), \tag{219}$$

the refresh cost (209) becomes

$$\tilde{O}\left(\frac{C_{\mathrm{V}}\mu_{\|\cdot\|}^2\sigma_T^2\kappa}{\epsilon^3}\right), \tag{220}$$

the variance-control cost (216) becomes

$$\tilde{O}\left(\frac{C_{\mathrm{V}}\mu_{\|\cdot\|}^2L_T^2\kappa^3}{\epsilon^3}+\frac{C_{\mathrm{V}}L_T^2\kappa^3}{\epsilon^3}\right), \tag{221}$$

the first term in the bias-control cost (218) becomes

$$\tilde{O}\left(\frac{\mu_{\|\cdot\|}C_{\mathrm{B}}L_T\kappa^{5/2}}{\epsilon^{5/2}}\right), \tag{222}$$

and the second term becomes

$$\tilde{O}\left(\frac{\mu_{\|\cdot\|}C_{\mathrm{B}}L_T\kappa^{5/2}}{\epsilon^{5/2}}\right). \tag{223}$$

Combining all of the contributions yields the stated cost (31). □

J Proof of the main high-probability result

J.1 Proof of Theorem 5.4

Proof of Theorem 5.4. The recursion is an inexact Halpern recursion with

$$U_n=e_{n-1}=v_{n-1}-Tx_{n-1}. \tag{224}$$

By Lemma 5.3, with probability at least $1-\eta$, we have

$$\|U_N\|_{\mathrm{s}}\le\delta \tag{225}$$

and

$$\sum_{n=1}^{N}n\|\beta_nU_n-\beta_{n-1}U_{n-1}\|_{\mathrm{s}}\le C\delta N(1+\log(1/\eta)). \tag{226}$$

By Condition 5.1, we may convert both bounds to the working norm with the factor α_{s}. Applying Lemma E.3 and Condition 5.2 proves (42). It remains to count the samples used on the same event. We let $d_n:=x_n-x_{n-1}$. The proof of Lemma E.3 also shows that, for every $n\le N$,

$$n\|d_n\|\le K(1+\log(N+1))+C\alpha_{\mathrm{s}}\delta N(1+\log(1/\eta)). \tag{227}$$

Under the stated choices of δ and N, the right-hand side is $\tilde{O}(K)$. Therefore,

$$\sum_{n=1}^{N}\frac{\|x_n-x_{n-1}\|^2}{p_n^2}=\tilde{O}(K^2N), \qquad \sum_{n=1}^{N}\frac{\|x_n-x_{n-1}\|}{p_n}=\tilde{O}(KN). \tag{228}$$

By (41), the refresh cost satisfies

$$\sum_{n=1}^{N}\mathbf{1}_{\mathcal{R}_n}S_{1,n}\le\tilde{O}\left(\frac{\Lambda_{N,\eta}\sigma_{\mathrm{s}}^2N(1+\log(1/\eta))}{\delta^2}\right). \tag{229}$$

The term proportional to δ^{-1} in the second inequality of (38) contributes at most

$$\tilde{O}\left(\frac{\sigma_{\mathrm{s}}N(1+\log(1/\eta))}{\delta}\right). \tag{230}$$

The corresponding initialization term in the first inequality of (38) is bounded by the same expression when $N \geq 1$. For the difference blocks, (39) and (228) give

$$\sum_{n=1}^{N} S_{2,n} \leq \tilde{O}\left(\frac{\Lambda_{N,\eta} L_{\mathrm{s}}^2 K^2 N}{\delta^2} + \frac{L_{\mathrm{s}} K N}{\delta} + N\right). \tag{231}$$

Substituting $N = \tilde{O}(K/\epsilon)$ and $\delta = \Theta\big(\epsilon/(\alpha_{\mathrm{s}}(1 + \log(1/\eta)))\big)$ proves (44). The terms proportional to δ^{-1} and N are lower order in ϵ. □