Last-Iterate Convergence of Randomized Kaczmarz and SGD with Greedy Step Size

Michał Dereziński DEREZIN@UMICH.EDU
University of Michigan

Xiaoyu Dong XDONG@NUS.EDU.SG
National University of Singapore

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study last-iterate convergence of SGD with greedy step size over smooth quadratics in the interpolation regime, a setting which captures the classical Randomized Kaczmarz algorithm as well as other popular iterative linear system solvers. For these methods, we show that the t-th iterate attains an $O(1/t^{3/4})$ convergence rate, addressing a question posed by Attia, Schliserman, Sherman, and Koren, who gave an $O(1/t^{1/2})$ guarantee for this setting. In the proof, we introduce the family of *stochastic contraction processes*, whose behavior can be described by the evolution of a certain deterministic eigenvalue equation, which we analyze via a careful discrete-to-continuous reduction.

Keywords: Stochastic Gradient Descent, Randomized Kaczmarz, Interpolation regime

1. Introduction

Stochastic Gradient Descent (SGD, Robbins and Monro, 1951) is one of the most extensively studied optimization algorithms. This has led to an in-depth understanding of the convergence properties of many variants of SGD across different optimization settings. It is therefore perhaps surprising that, with such extensive literature, the worst-case convergence remains unresolved for one of the oldest SGD algorithms in one of the most classical settings: the Kaczmarz algorithm for solving consistent systems of linear equations (Kaczmarz, 1937). Given a system of m equations with n unknowns, the Kaczmarz algorithm starts with an initial n-dimensional estimate vector, and then iteratively selects one equation to solve, computing the solution that is the closest to its previous estimate. This method has seen renewed interest in the numerical analysis literature since Strohmer and Vershynin (2009) proposed a randomized sampling scheme for selecting the equations. Yet, the worst-case convergence rate (i.e., independent of any condition numbers) is still unknown for this method, regardless of how the equations are sampled, despite recent efforts in this direction (Steinerberger, 2023; Evron et al., 2025; Attia et al., 2025).

In the broader context, the above question is directly tied to the study of last-iterate convergence of SGD with fixed step size in the smooth interpolation regime, which has seen significant interest thanks to the effectiveness of this type of algorithms for training highly over-parameterized deep learning models (Ma et al., 2018). Here, the Kaczmarz algorithm can be viewed as an instance of SGD minimizing an average of convex β-smooth functions, using step size $1/\beta$. This choice of step size is notable, as this is also the canonical choice for full gradient descent (GD) on a β-smooth function. While SGD is not expected in general to converge under fixed step size, it will do so if all of the averaged functions admit a common minimizer (the *interpolation regime*). In particular, it is known that the average of t such SGD iterates converges at the rate of $O(1/t)$ (Bach and Moulines,

2013; Zou et al., 2021), and after sufficiently shrinking the step size below $1/\beta$, nearly matching rates can be obtained for the last iterate (Varre et al., 2021). Yet, these guarantees do not cover last-iterate SGD with the canonical GD step size $1/\beta$ (called the greedy step size), empirically the most effective choice.

Recent efforts toward the understanding of SGD with greedy step size are motivated not only by the Kaczmarz method (along with its many extensions such as Block Kaczmarz, Coordinate Descent, Sketch-and-Project, etc.), but also by its application to the analysis of catastrophic forgetting in a class of realizable continual learning problems (Evron et al., 2022). In this context, Evron et al. (2025) provided an analysis of SGD with greedy step size (including Kaczmarz), showing that its last iterate converges at an $O(1/t^{1/4})$ rate. Later, Attia et al. (2025) improved this result to $O(1/t^{1/2})$, and asked whether this rate is optimal. (Further related work is provided in Appendix A.)

Our contributions. In this work, we provide a new framework for analyzing the last-iterate convergence of SGD algorithms, and we use it to obtain the following main result:

The last iterate of SGD over β-smooth quadratics in the interpolation regime with step size $1/\beta$, including Randomized Kaczmarz (Strohmer and Vershynin, 2009) and Randomized Coordinate Descent (Leventhal and Lewis, 2010), attains the $O(1/t^{3/4})$ convergence rate.

Curiously, we are able to show that the exponent $3/4$ in the rate is *not* optimal, as our analysis can be pushed further to recover the exponent $3/4 + 0.001$, however we encounter a fundamental barrier around $3/4 + 0.003$. Furthermore, our results apply more generally than the canonical methods mentioned above, for example including all linear system solvers based on the so-called Sketch-and-Project framework (Gower and Richtárik, 2015). In particular, we use our techniques to show that a certain variant of Block Kaczmarz (Elfving, 1980) attains a stronger worst-case last-iterate convergence guarantee than the classical Kaczmarz method.

Overview of our techniques. To attain our results, we characterize the convergence of SGD through what we call a *stochastic contraction process* (Definition 1): a sequence of independent random positive semidefinite (psd) contraction operators applied to a high-dimensional vector. We observe that capturing SGD algorithms with greedy step size involves analyzing such a stochastic process in full generality, without imposing any restrictions (such as upper/lower bounds) on the contraction operators (Theorem 2).

We analyze the stochastic contraction process by characterizing it via a deterministic matrix recursion (Lemma 10). Unfolding this recursion reveals that its spectrum exhibits two regimes: one where the eigenvalues oscillate wildly, and one where they follow a smooth trajectory. We carefully unify these two regimes, and reduce them to a single summation bound (Lemma 11). Establishing this bound proves remarkably delicate (Section 4): We achieve this by performing a discrete-to-continuous reduction and analyzing the resulting ordinary differential equation (ODE).

2. Main Result and Its Implications

In this section, we present our main result, and then describe its implications for Randomized Kaczmarz and other SGD-type algorithms. To highlight the general nature of the claim, we frame it as a characterization of the behavior of a high-dimensional stochastic process defined by a sequence of independent random psd contraction operators with a common mean ($\preceq$ denotes the Loewner order).

Definition 1 *Random sequence $\boldsymbol{\Delta}_0, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, ... \in \mathbb{R}^n$ is called a **stochastic contraction process** with average rate $\bar{\mathbf{M}}$ if it satisfies $\boldsymbol{\Delta}_{t+1} = (\mathbf{I} - \mathbf{M}_t)\boldsymbol{\Delta}_t$ for a sequence of independent random $n \times n$ psd matrices $\mathbf{M}_0, \mathbf{M}_1, \mathbf{M}_2, ...$ such that $\mathbf{0} \preceq \mathbf{M}_t \preceq \mathbf{I}$ and $\mathbb{E}\,\mathbf{M}_t = \bar{\mathbf{M}}$ for all $t \geq 0$.*

Many stochastic algorithms can be cast as instances of such a process, and many existing convergence arguments can be viewed as analyzing this process under additional restrictions on the contractions (such as bounding them away from zero or from identity). Crucially, our result does not impose any such restrictions. Below and throughout, we use the notation $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{\mathbf{x}^\top \mathbf{M}\mathbf{x}}$.

Theorem 2 *There are absolute constants $C > 0$ and $\theta \geq 0.001$ such that any stochastic contraction process $\{\boldsymbol{\Delta}_t\}_{t\geq 0}$ with average rate $\bar{\mathbf{M}}$ satisfies*

$$\mathbb{E}\,\|\boldsymbol{\Delta}_t\|^2_{\bar{\mathbf{M}}} \leq C \cdot \frac{\mathbb{E}\,\|\boldsymbol{\Delta}_0\|^2}{t^{3/4+\theta}}.$$

Remark 3 *The $O(1/t^{3/4+\theta})$ convergence rate appears to be the best rate attainable via our analysis framework (without introducing further restrictions on $\mathbf{M}_t$) up to $\theta \approx 0.003$ (see Section 3.2).*

Remark 4 *A key feature of Theorem 2 is that it allows both $\mathbf{M}_t$ and $\bar{\mathbf{M}}$ to vary in the full range between zero and identity, which enables it to capture the canonical versions of Randomized Kaczmarz and Randomized Coordinate Descent on worst-case inputs. Restricting $\mathbf{M}_t$ (or its expectation) to a smaller range, e.g., $c_1\mathbf{I} \preceq \mathbf{M}_t \preceq c_2\mathbf{I}$ where either $c_1 > 0$ or $c_2 < 1$, leads to simpler results (and potentially faster rates) that are well-known in the literature.*

2.1. Implications for SGD with Greedy Step Size

Theorem 2 can be interpreted as a convergence guarantee for a stochastic gradient algorithm running on a quadratic function in the interpolation regime. To see this, consider minimizing $f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m \psi_i(\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^n$, where each ψ_i is a β-smooth quadratic, i.e., $\|\nabla\psi_i(\mathbf{x}) - \nabla\psi_i(\mathbf{y})\| \leq \beta\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Moreover, suppose that there exists $\mathbf{x}^* \in \mathbb{R}^d$ that simultaneously minimizes all ψ_i (this is referred to as the *interpolation regime*). As a concrete example, we can think of a regression problem defined by m vectors $\mathbf{a}_1, ..., \mathbf{a}_m \in \mathbb{R}^n$ and response values $b_1, ..., b_m \in \mathbb{R}$, and let $\psi_i(\mathbf{x}) = \frac{1}{2}(\mathbf{a}_i^\top\mathbf{x} - b_i)^2$ for each i. Here, $\beta = \max_i \|\mathbf{a}_i\|^2$, and the interpolation regime occurs when there is a linear model $\mathbf{a}_i \to \mathbf{a}_i^\top\mathbf{x}^*$ that perfectly fits all of the response values b_i, in which case $\psi_i(\mathbf{x}) = \frac{1}{2}(\mathbf{a}_i^\top(\mathbf{x}-\mathbf{x}^*))^2$ and $f(\mathbf{x}) = \frac{1}{2m}\|\mathbf{x}-\mathbf{x}^*\|^2_{\mathbf{A}^\top\mathbf{A}}$.

The standard SGD algorithm with fixed step size η initialized at $\mathbf{x}_0 \in \mathbb{R}^n$ proceeds by randomly sampling one component function at a time and taking a corresponding gradient descent step:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla\psi_{i_t}(\mathbf{x}_t), \qquad i_t \sim \{1, ..., m\}.$$

This can be mapped to a stochastic contraction process via $\boldsymbol{\Delta}_t = \mathbf{x}_t - \mathbf{x}^*$. Indeed, since ψ_{i_t} is a quadratic minimized by $\mathbf{x}^*$, using the Taylor expansion we have $\nabla\psi_{i_t}(\mathbf{x}_t) = \nabla^2\psi_{i_t}(\mathbf{x}^*)(\mathbf{x}_t - \mathbf{x}^*)$, so by choosing $\mathbf{M}_t = \eta\nabla^2\psi_{i_t}(\mathbf{x}^*)$ we get $\boldsymbol{\Delta}_{t+1} = (\mathbf{I} - \mathbf{M}_t)\boldsymbol{\Delta}_t$. Using β-smoothness, we have $\mathbf{M}_t \preceq \eta\beta\mathbf{I}$ so the condition $\mathbf{0} \preceq \mathbf{M}_t \preceq \mathbf{I}$ is satisfied for any $0 < \eta \leq 1/\beta$. Since $f(\mathbf{x}_t) - f(\mathbf{x}^*) = \frac{1}{2\eta}\|\mathbf{x}_t - \mathbf{x}^*\|^2_{\bar{\mathbf{M}}}$, where $\bar{\mathbf{M}} = \mathbb{E}\,\mathbf{M}_t = \eta\nabla^2 f(\mathbf{x}^*)$, we obtain the following corollary of Theorem 2.

Corollary 5 *Minimizing an average of β-smooth quadratic functions in the interpolation regime, SGD with step size $1/\beta$ satisfies:*

$$\mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^*)\big] = O\left(\frac{\beta\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{t^{3/4+\theta}}\right).$$

Here, the fact that our result allows the choice of step size $\eta = 1/\beta$ is crucial. For standard GD on a β-smooth function, $1/\beta$ is the canonical choice of step size (Bertsekas, 2016). However, when dealing with stochastic gradients, it is common to use either a much smaller fixed step size or a decaying step size schedule in order to compensate for the noise in the convergence analysis (Varre et al., 2021; Liu and Zhou, 2023). Yet, in the interpolation regime, the canonical choice of $\eta = 1/\beta$ (i.e., the greedy step size) is often empirically the most effective one. Corollary 5 continues a recent line of works aiming to close the theory-practice gap in our understanding of SGD with greedy step sizes (Evron et al., 2025; Attia et al., 2025), improving the rate from $O(1/t^{1/2})$ to $O(1/t^{3/4+\theta})$.

2.2. Key Example: Randomized Kaczmarz

Perhaps the most important application of Theorem 2 is in the analysis of randomized iterative methods such as the Kaczmarz algorithm for solving consistent systems of linear equations. Here, we are given an $m \times n$ matrix $\mathbf{A}$ and an m-dimensional vector $\mathbf{b}$ such that there exists $\mathbf{x}^*$ satisfying $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Given an iterate $\mathbf{x}_t$, Kaczmarz chooses one of the m linear equations, $\mathbf{a}_{i_t}^\top \mathbf{x} = b_{i_t}$ (where $\mathbf{a}_i^\top$ denotes the ith row of $\mathbf{A}$), and computes $\mathbf{x}_{t+1}$ as the projection of $\mathbf{x}_t$ onto the subspace of the solutions of that equation:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\mathbf{a}_{i_t}^\top \mathbf{x}_t - b_{i_t}}{\|\mathbf{a}_{i_t}\|^2}\mathbf{a}_{i_t}.$$

The Kaczmarz algorithm can be viewed as a type of weighted SGD minimizing the $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ objective (Needell et al., 2014), and thus it analogously maps to Definition 1 by setting:

$$\boldsymbol{\Delta}_t = \mathbf{x}_t - \mathbf{x}^* \qquad \text{and} \qquad \mathbf{M}_t = \frac{\mathbf{a}_{i_t}\mathbf{a}_{i_t}^\top}{\|\mathbf{a}_{i_t}\|^2}, \tag{1}$$

where note that $\mathbf{M}_t$ is simply the rank-1 projection onto the span of $\mathbf{a}_{i_t}$. Naturally, how we select the equation indices i_t has a great impact on the convergence rate of the Kaczmarz algorithm, and Strohmer and Vershynin (2009) showed that if we sample i_t with probability proportional to the squared row norm, $\Pr[i_t = i] \propto \|\mathbf{a}_i\|^2$, then this Randomized Kaczmarz algorithm will converge to the optimum $\mathbf{x}^*$ at the rate that depends only on the condition number of $\mathbf{A}$, and it requires fewer passes over the matrix than full gradient descent. However, they provide no convergence guarantee that is free of condition number dependence, and despite extensive literature on this subject, the last-iterate convergence rate of Randomized Kaczmarz on worst-case inputs remains unresolved.

Mapping Randomized Kaczmarz to Definition 1, we observe that $\mathbb{E}[\mathbf{M}_t] = \mathbf{A}^\top\mathbf{A}/\|\mathbf{A}\|_F^2$ and $\|\boldsymbol{\Delta}_t\|_{\mathbf{M}}^2 = \|\mathbf{A}\mathbf{x}_t - \mathbf{b}\|^2/\|\mathbf{A}\|_F^2$, where $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})}$ denotes the Frobenius norm of $\mathbf{A}$. This yields the following corollary.

Corollary 6 *For a linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ with solution $\mathbf{x}^*$, Randomized Kaczmarz satisfies:*

$$\mathbb{E}\,\|\mathbf{A}\mathbf{x}_t - \mathbf{b}\|^2 = O\left(\frac{\|\mathbf{A}\|_F^2\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{t^{3/4+\theta}}\right).$$

We note that, just like in existing guarantees for weighted SGD (Needell et al., 2014), the use of importance sampling as opposed to uniform sampling allows us to replace the smoothness parameter $\beta = \max_i \|\mathbf{a}_i\|^2$ with an average smoothness $\bar{\beta} = \frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|^2 = \frac{1}{m}\|\mathbf{A}\|_F^2 \le \beta$, which is why the above bound has $\|\mathbf{A}\|_F^2$ instead of $m \cdot \max\|\mathbf{a}_i\|^2$ in the numerator. Here, again, our result provides an improvement in the last-iterate convergence rate of Randomized Kaczmarz from the previous $O(1/t^{1/2})$ attained by Attia et al. (2025) to $O(1/t^{3/4+\theta})$.

2.3. Further Implications for Sketch-and-Project Algorithms

Thanks to its generality, Theorem 2 covers a number of other randomized iterative methods for linear systems, including all of those that fall under the framework of Sketch-and-Project, developed by Gower and Richtárik (2015), which in addition to Randomized Kaczmarz also includes Block Kaczmarz (Elfving, 1980) and Randomized Coordinate Descent (Leventhal and Lewis, 2010), among many others. Here, the update is defined via a random $b \times m$ matrix $\mathbf{S}_t$ (the sketching operator) and an $n \times n$ positive definite matrix $\mathbf{B}$ (which determines the projection norm):

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x} - \mathbf{x}_t\|_{\mathbf{B}} \quad \text{subject to} \quad \mathbf{S}_t\mathbf{A}\mathbf{x} = \mathbf{S}_t\mathbf{b}. \tag{2}$$

Under this framework, Randomized Kaczmarz is recovered by letting $\mathbf{S}_t = \mathbf{e}_{i_t}^\top \in \mathbb{R}^{1\times m}$ be a random standard basis row-vector and setting $\mathbf{B} = \mathbf{I}$. Stacking $b > 1$ random standard basis row-vectors to produce $\mathbf{S}_t \in \mathbb{R}^{b\times m}$, we recover the Block Kaczmarz method. When $\mathbf{A}$ is a positive definite matrix, then we can consider a different scheme by letting $\mathbf{B} = \mathbf{A}$, which yields the Randomized Coordinate Descent method, proposed by Leventhal and Lewis (2010).

To analyze the last-iterate convergence of all of these methods together, we map the general sketch-and-project update (2) to a stochastic contraction process as follows:

$$\boldsymbol{\Delta}_t = \mathbf{B}^{1/2}(\mathbf{x}_t - \mathbf{x}^*) \qquad \text{and} \qquad \mathbf{M}_t = (\mathbf{S}_t\mathbf{A}\mathbf{B}^{-1/2})^\dagger\mathbf{S}_t\mathbf{A}\mathbf{B}^{-1/2}.$$

Note that here again the matrices $\mathbf{M}_t$ are orthogonal projections, which means that they are between zero and identity, but cannot be bounded away from either, thus requiring the careful convergence analysis framework from Theorem 2. To complete the analysis, observe that $\|\mathbf{A}\mathbf{x}_t - \mathbf{b}\| = \|\boldsymbol{\Delta}_t\|_{\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{A}\mathbf{B}^{-1/2}}$, so in order to analyze convergence in the residual norm it suffices to bound the matrix $\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{A}\mathbf{B}^{-1/2}$ in terms of $\bar{\mathbf{M}} = \mathbb{E}\,\mathbf{M}_t$ in the Loewner ordering.

For Randomized Coordinate Descent, this turns out to be straightforward, given the right choice of sampling probabilities. The distribution proposed by Leventhal and Lewis (2010) samples proportionally to the diagonal entries of $\mathbf{A}$, namely $\Pr[\mathbf{S}_t = \mathbf{e}_{i_t}^\top] \propto \mathbf{A}_{i_t,i_t}$. After a simple calculation, this yields $\mathbb{E}\,\mathbf{M}_t = \mathbf{A}/\mathrm{tr}(\mathbf{A})$. Since $\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{A}\mathbf{B}^{-1/2} = \mathbf{A}$, we obtain the following corollary.[1]

Corollary 7 *For a psd system $\mathbf{A}\mathbf{x} = \mathbf{b}$ with solution $\mathbf{x}^*$, Randomized Coordinate Descent satisfies:*

$$\mathbb{E}\,\|\mathbf{A}\mathbf{x}_t - \mathbf{b}\|^2 = O\left(\frac{\mathrm{tr}(\mathbf{A})\|\mathbf{x}_0 - \mathbf{x}^*\|_{\mathbf{A}}^2}{t^{3/4+\theta}}\right).$$

For Block Kaczmarz, computing the expectation $\mathbb{E}\,\mathbf{M}_t$ explicitly is less straightforward. However, since the projection $\mathbf{M}_t$ only gets larger (in Loewner ordering) when introducing multiple equations, it follows that Block Kaczmarz inherits the guarantee of the corresponding single-row Kaczmarz, such as the one in Corollary 6 (and same is true for block versions of coordinate descent). Nevertheless, this feels somewhat unsatisfying, since we would hope that the convergence guarantee improves as we increase the block size. In fact, Dereziński and Yang (2024) showed such an improved convergence guarantee when the problem is parameterized by an appropriate notion of condition number and as long as we preprocess the linear system with the Randomized Hadamard Transform (RHT, Ailon and Chazelle, 2009). Here, RHT refers to a random orthogonal transformation[2] $\mathbf{Q} \in \mathbb{R}^{m\times m}$ which can be applied to $\mathbf{A}$ and $\mathbf{b}$ in $O(mn\log m)$ time and has the property that

1. Even though the above discussion assumes that $\mathbf{B}$ is invertible, Corollary 7 easily extends to any psd linear system.
2. $\mathbf{Q} = \frac{1}{\sqrt{n}}\mathbf{H}\mathbf{D}$, where $\mathbf{H}$ is the $m \times m$ Hadamard matrix and $\mathbf{D}$ is diagonal with i.i.d. Rademacher entries.

the transformed system $\mathbf{QAx} = \mathbf{Qb}$ is equivalent to the original one, but its equations have roughly equal importance. This allows us to use uniform sampling in the Block Kaczmarz algorithm. In the following corollary, we show that when the block size is proportional to the stable rank of $\mathbf{A}$, such RHT-preprocessed Block Kaczmarz satisfies a stronger convergence bound than single-row Kaczmarz, replacing the Frobenius norm with the spectral norm, thus matching full gradient descent up to the convergence exponent.

Corollary 8 *Given a system $\mathbf{Ax} = \mathbf{b}$ with solution $\mathbf{x}^*$ and stable rank $r = \|\mathbf{A}\|_F^2/\|\mathbf{A}\|^2$, after preprocessing with the RHT, Block Kaczmarz with block size $b \geq O(r \log r + \log mt)$ satisfies:*

$$\mathbb{E}\,\|\mathbf{Ax}_t - \mathbf{b}\|^2 = O\left(\frac{\|\mathbf{A}\|^2\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{t^{3/4+\theta}}\right).$$

Proof Lemma 14 of Dereziński et al. (2025a) shows that after RHT preprocessing, $\mathbf{A} \leftarrow \mathbf{QA}$, the projection matrix $\mathbf{M}_t = (\mathbf{S}_t\mathbf{A})^\dagger\mathbf{S}_t\mathbf{A}$ for Block Kaczmarz with $b \geq O(\log(m/\delta))$ satisfies:

$$\bar{\mathbf{M}} = \mathbb{E}\,\mathbf{M}_t \succeq \mathbf{A}^\top(\mathbf{AA}^\top + \lambda\mathbf{I})^{-1}\mathbf{A} - \delta\mathbf{I},$$

where $\lambda = O\big(\frac{\log b}{b}\|\mathbf{A}\|_F^2\big)$. Choosing $\delta = 1/t$ and $b \geq O(r \log r + \log mt)$ so that $\lambda \leq \|\mathbf{A}\|^2$,

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Ax}_t - \mathbf{b}\|^2 = \mathbb{E}\,\|\mathbf{x}_t - \mathbf{x}^*\|^2_{\mathbf{A}^\top\mathbf{A}} &\leq \|\mathbf{AA}^\top + \lambda\mathbf{I}\| \cdot \mathbb{E}\,\|\mathbf{x}_t - \mathbf{x}^*\|^2_{\bar{\mathbf{M}}+\delta\mathbf{I}} \\ &\leq 2\|\mathbf{A}\|^2 \cdot \Big(\mathbb{E}\,\|\mathbf{x}_t - \mathbf{x}^*\|^2_{\bar{\mathbf{M}}} + \delta \cdot \mathbb{E}\,\|\mathbf{x}_t - \mathbf{x}^*\|^2\Big) = O\left(\frac{\|\mathbf{A}\|^2\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{t^{3/4+\theta}}\right),\end{aligned}$$

where we used that $\mathbb{E}\,\|\mathbf{x}_t - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|^2$. Noting that RHT preprocessing preserves all of the above norms concludes the proof. ■

2.4. Connections to Averaged Iterate Analysis

We remark that the main challenge in our analysis of stochastic contraction processes, and their special cases such as SGD with greedy step size and Randomized Kaczmarz, stems from the fact that we seek to bound the last iterate, as opposed to for instance the averaged iterate or a random iterate. To highlight this fact, as an auxiliary result, we provide a simple $O(1/t)$ convergence guarantee for the averaged and random iterates of a stochastic contraction process.

Theorem 9 *Given a stochastic contraction process $\{\boldsymbol{\Delta}_t\}_{t\geq 0}$ with average rate $\bar{\mathbf{M}}$, let τ be a random variable uniformly sampled from $\{0, 1, ..., t\}$, and define $\bar{\boldsymbol{\Delta}}_t = \frac{1}{t+1}\sum_{i=0}^t \boldsymbol{\Delta}_i$. Then:*

$$\mathbb{E}\,\|\bar{\boldsymbol{\Delta}}_t\|^2_{\bar{\mathbf{M}}} \leq \mathbb{E}\,\|\boldsymbol{\Delta}_\tau\|^2_{\bar{\mathbf{M}}} \leq \frac{\mathbb{E}\,\|\boldsymbol{\Delta}_0\|^2}{t+1}.$$

Proof The proof follows a standard argument from averaged iterate analysis of SGD. Observe that:

$$\mathbb{E}\,\|\boldsymbol{\Delta}_{i+1}\|^2 = \mathbb{E}\,\boldsymbol{\Delta}_i^\top(\mathbf{I} - \mathbf{M}_i)^2\boldsymbol{\Delta}_i \leq \mathbb{E}\,\|\boldsymbol{\Delta}_i\|^2 - \mathbb{E}\,\|\boldsymbol{\Delta}_i\|^2_{\bar{\mathbf{M}}}.$$

Summing and canceling out both sides from 0 to t, we get

$$\mathbb{E}\,\|\boldsymbol{\Delta}_{t+1}\|^2 + \sum_{i=0}^t \mathbb{E}\,\|\boldsymbol{\Delta}_i\|^2_{\bar{\mathbf{M}}} \leq \mathbb{E}\,\|\boldsymbol{\Delta}_0\|^2.$$

From this, we immediately obtain that

$$\mathbb{E}\left\|\frac{1}{t+1}\sum_{i=0}^{t}\boldsymbol{\Delta}_i\right\|_{\bar{\mathbf{M}}}^2 \le \frac{1}{t+1}\sum_{i=0}^{t}\mathbb{E}\left\|\boldsymbol{\Delta}_i\right\|_{\bar{\mathbf{M}}}^2 \le \frac{\mathbb{E}\left\|\boldsymbol{\Delta}_0\right\|^2}{t+1},$$

which concludes the proof. ■

Therefore, replacing the last iterate with the averaged/random iterate in each of our examples yields the optimal $O(1/t)$ convergence rate. In particular, using Theorem 9 to analyze the random iterate $\mathbf{x}_\tau$ in Randomized Kaczmarz yields a positive answer to a question posed by Steinerberger (2023) about whether there always exists a sequence of $t = O(\|\mathbf{A}\|_F^2/\epsilon^2)$ Kaczmarz updates that attains the guarantee $\|\mathbf{A}\mathbf{x}_t - \mathbf{b}\| \le \epsilon\|\mathbf{x}_0 - \mathbf{x}^*\|$.

3. Convergence Analysis via Matrix Recursion

In this section, we characterize the convergence behavior of a stochastic contraction process. We start by upper bounding the expected norm of the random vectors by defining a matrix recursion that captures how the process evolves along the eigendirections of the average rate matrix $\bar{\mathbf{M}}$.

Lemma 10 *Given an $n \times n$ psd matrix $\mathbf{0} \preceq \bar{\mathbf{M}} \preceq \mathbf{I}$, define the following matrix recursion:*

$$\mathbf{N}_0 = \bar{\mathbf{M}}, \qquad \mathbf{N}_{t+1} = \mathbf{N}_t(\mathbf{I} - 2\bar{\mathbf{M}}) + \|\mathbf{N}_t\| \cdot \bar{\mathbf{M}}. \tag{3}$$

Then, any stochastic contraction process $\{\boldsymbol{\Delta}_t\}_{t\ge 0}$ with average rate $\bar{\mathbf{M}}$ satisfies:

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t\right\|_{\bar{\mathbf{M}}}^2 \le \mathbb{E}\left\|\boldsymbol{\Delta}_0\right\|_{\mathbf{N}_t}^2.$$

Proof Let $\{\mathbf{M}_t\}_{t\ge 0}$ be the sequence of contractions that define $\{\boldsymbol{\Delta}_t\}_{t\ge 0}$. Recall that $\mathbf{0} \preceq \mathbf{M}_t \preceq \mathbf{I}$ and $\mathbb{E}\,\mathbf{M}_t = \bar{\mathbf{M}}$ for all t. Fix some $t \ge 0$ and define the following sequence for $i = 0, ..., t$:

$$\bar{\mathbf{N}}_{t,0} = \bar{\mathbf{M}}, \qquad \bar{\mathbf{N}}_{t,i} = \mathbb{E}\big[(\mathbf{I} - \mathbf{M}_{t-i})\bar{\mathbf{N}}_{t,i-1}(\mathbf{I} - \mathbf{M}_{t-i})\big].$$

Without loss of generality, assume that $\boldsymbol{\Delta}_0$ is deterministic. Then, $\mathbb{E}\left\|\boldsymbol{\Delta}_0\right\|_{\bar{\mathbf{M}}}^2 = \left\|\boldsymbol{\Delta}_0\right\|_{\bar{\mathbf{N}}_{t,0}}^2$ and

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{\Delta}_t\right\|_{\bar{\mathbf{M}}}^2 &= \mathbb{E}\Big[\boldsymbol{\Delta}_{t-1}^\top \mathbb{E}\big[(\mathbf{I} - \mathbf{M}_{t-1})\bar{\mathbf{M}}(\mathbf{I} - \mathbf{M}_{t-1})\big]\boldsymbol{\Delta}_{t-1}\Big] \\ &= \mathbb{E}\big[\boldsymbol{\Delta}_{t-1}^\top \bar{\mathbf{N}}_{t,1}\boldsymbol{\Delta}_{t-1}\big] = ... = \boldsymbol{\Delta}_0^\top \bar{\mathbf{N}}_{t,t}\boldsymbol{\Delta}_0.\end{aligned}$$

Next, we show by induction that $\bar{\mathbf{N}}_{t,i} \preceq \mathbf{N}_i$ for each $i = 0, ..., t$. Clearly, $\bar{\mathbf{N}}_{t,0} = \mathbf{N}_0$, so suppose that $\bar{\mathbf{N}}_{t,i-1} \preceq \mathbf{N}_{i-1}$ for some $1 \le i \le t$. Then:

$$\begin{aligned}\bar{\mathbf{N}}_{t,i} &\preceq \mathbb{E}\big[(\mathbf{I} - \mathbf{M}_{t-i})\mathbf{N}_{i-1}(\mathbf{I} - \mathbf{M}_{t-i})\big] \\ &= \mathbf{N}_{i-1} - \mathbb{E}\,\mathbf{M}_{t-i}\mathbf{N}_{i-1} - \mathbf{N}_{i-1}\mathbb{E}\,\mathbf{M}_{t-i} + \mathbb{E}\,\mathbf{M}_{t-i}\mathbf{N}_{i-1}\mathbf{M}_{t-i} \\ &= \mathbf{N}_{i-1} - \bar{\mathbf{M}}\mathbf{N}_{i-1} - \mathbf{N}_{i-1}\bar{\mathbf{M}} + \mathbb{E}\,\mathbf{M}_{t-i}\mathbf{N}_{i-1}\mathbf{M}_{t-i} \\ &\preceq \mathbf{N}_{i-1}(\mathbf{I} - 2\bar{\mathbf{M}}) + \|\mathbf{N}_{i-1}\| \cdot \mathbb{E}\,\mathbf{M}_{t-i}^2, \\ &\preceq \mathbf{N}_{i-1}(\mathbf{I} - 2\bar{\mathbf{M}}) + \|\mathbf{N}_{i-1}\| \cdot \bar{\mathbf{M}},\end{aligned}$$

where we used that by definition $\mathbf{N}_{i-1}$ commutes with $\bar{\mathbf{M}}$, and then that $\mathbb{E}\,\mathbf{M}_{t-i}^2 \preceq \mathbb{E}\,\mathbf{M}_{t-i}$. Thus, we have shown that $\bar{\mathbf{N}}_{t,i} \preceq \mathbf{N}_i$ for each $i = 0, ..., t$ and each $t \ge 0$. In particular, this implies that $\bar{\mathbf{N}}_{t,t} \preceq \mathbf{N}_t$. The claim now follows since $\mathbb{E}\left\|\boldsymbol{\Delta}_t\right\|_{\bar{\mathbf{M}}}^2 = \left\|\boldsymbol{\Delta}_0\right\|_{\bar{\mathbf{N}}_{t,t}}^2 \le \left\|\boldsymbol{\Delta}_0\right\|_{\mathbf{N}_t}^2$. ■

We next prove Theorem 2 by analyzing the evolution of the eigenvalues in recursion (3).

3.1. Proof of Theorem 2

Lemma 10 implies that $\mathbb{E}\,\|\boldsymbol{\Delta}_t\|_{\bar{\mathbf{M}}}^2 \le \|\mathbf{N}_t\| \cdot \mathbb{E}\,\|\boldsymbol{\Delta}_0\|^2$, so it now suffices to bound $\|\mathbf{N}_t\|$ for matrices $\mathbf{N}_t$ defined by the recursion (3). Since these matrices all commute with $\bar{\mathbf{M}}$ (and therefore, with each other), we can rewrite the recursion purely as a transformation of their eigenvalues in the common basis. Let $\bar{\mathbf{M}} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ be the eigendecomposition of $\bar{\mathbf{M}}$, where $\mathbf{D} = \mathrm{diag}(\rho_1, ..., \rho_n)$ and $\rho_1 \ge \rho_2 \ge ...$ denote the eigenvalues of $\bar{\mathbf{M}}$. Also, let $\lambda_{k,t}$ be the eigenvalue of $\mathbf{N}_t$ associated with the kth eigenvector in $\mathbf{U}$. Then, these eigenvalues are governed by the following set of recursions:

$$\lambda_{k,t+1} = \lambda_{k,t} \cdot (1 - 2\rho_k) + \rho_k \cdot \max_i \lambda_{i,t}, \qquad k = 1, ..., n. \tag{4}$$

It now suffices to show that $\|\mathbf{N}_t\| = \lambda_{\max,t} \le \frac{C}{t^\alpha}$ for each t, where we define the shorthand $\lambda_{\max,t} := \max_k \lambda_{k,t}$, whereas the constant $C \ge 2$ will be chosen later. We do this via induction, using the following slightly stronger inductive hypothesis, distinguishing between even and odd t:

$$\textbf{Inductive hypothesis:} \qquad \lambda_{\max,t} \le \begin{cases} \frac{C}{(t+1)^\alpha} & \text{if } t \text{ is even,} \\ \frac{C}{(t+2)^\alpha} & \text{if } t \text{ is odd.} \end{cases}$$

As the base case, consider all $0 \le t \le C - 2$. Then, for all k we have

$$\lambda_{k,t} = \lambda_{k,t-1}(1 - \rho_k) + (\lambda_{\max,t-1} - \lambda_{k,t-1})\rho_k \le \lambda_{\max,t-1} \le ... \le \lambda_{\max,0} \le 1 \le \frac{C}{(t+2)^\alpha},$$

where we used that $\lambda_{k,t-1} \ge 0$ since $\mathbf{N}_{t-1}$ is positive semidefinite.

Next, suppose that the induction hypothesis holds for for all $0, 1, ..., t-1$. We will show the hypothesis for t. We break the analysis down into two subsets of the indices k, depending on whether $\rho_k \le 1/2$ or not, to account for the possibility of changing signs in the first term of (3).

Case 1: $\rho_k \le 1/2$**.** We start by focusing only on k such that $\rho_k \le 1/2$. Here, we can treat the even and odd t cases together by showing that $\lambda_{k,t} \le \frac{C}{(t+2)^\alpha}$. We start by expanding the recursion:

$$\begin{aligned} \lambda_{k,t} &= \lambda_{k,t-1}(1 - 2\rho_k) + \rho_k \lambda_{\max,t-1} \\ &= \rho_k(1 - 2\rho_k)^t + \rho_k \sum_{i=1}^{t} (1 - 2\rho_k)^{t-i} \lambda_{\max,i-1} \\ &\le \rho_k(1 - 2\rho_k)^t + C\rho_k \sum_{i=1}^{t} \frac{(1 - 2\rho_k)^{t-i}}{i^\alpha}. \end{aligned}$$

In the last inequality we use the assumption that $\rho_k \le 1/2$ to ensure that $1 - 2\rho_k \ge 0$ and all of the terms in the expression are non-negative. It now suffices to show that the above summation formula is bounded by $\frac{C}{(t+2)^\alpha}$, which is obtained in the following key technical lemma (see Section 4).

Lemma 11 *For all $K, t \ge 1000$, $\rho \in (0, 1/2]$, and $\alpha = 3/4 + 0.001$, we have*

$$\rho(1 - 2\rho)^t + K\rho \sum_{i=1}^{t} \frac{(1 - 2\rho)^{t-i}}{i^\alpha} \le \frac{K}{(t+2)^\alpha}.$$

Despite being entirely elementary, Lemma 11 turns out to require a quite involved and technical analysis. In Section 4, we give a proof sketch of a slightly weaker claim with $\alpha = 3/4$ (which is still highly technical, but much more readable), and the complete argument is given in the Appendix.

Case 2: $\rho_k > 1/2$**.** Next, we consider only k such that $\rho_k > 1/2$. In this case, $1-2\rho_k$ is negative, which leads the recursion to oscillate up and down between even and odd t's (see Figure 1). We thus consider these two sub-cases.

Case 2a: Even t**.** Note that from the inductive hypothesis we have $\lambda_{\max,t-1} \le \frac{C}{(t+1)^\alpha}$, since $t-1$ is odd. Furthermore, as shown before, we have $\lambda_{k,t} \le \lambda_{\max,t-1}$, so the claim follows.

Case 2b: Odd t**.** We again expand the recursion, but in a slightly different way:

$$\begin{aligned}
\lambda_{k,t} &= \lambda_{k,t-1}(1-2\rho_k) + \rho_k\lambda_{\max,t-1} \\
&= \big(\lambda_{k,t-2}(1-2\rho_k) + \rho_k\lambda_{\max,t-2}\big)(1-2\rho_k) + \rho_k\lambda_{\max,t-1} \\
&= \lambda_{k,t-2}(1-2\rho_k)^2 + 2\rho_k(1-\rho_k)\lambda_{\max,t-1} + \rho_k(2\rho_k-1)(\lambda_{\max,t-1} - \lambda_{\max,t-2}) \\
&\le \lambda_{k,t-2}(1-2\rho_k)^2 + 2\rho_k(1-\rho_k)\lambda_{\max,t-2},
\end{aligned}$$

where in the last inequality we used again that $\lambda_{\max,t-1} \le \lambda_{\max,t-2}$. In order to recover an analog of Case 1, we substitute $\beta_k = 2\rho_k(1-\rho_k) \in [0,1/2]$, observing that $(1-2\rho_k)^2 = 1-2\beta_k$, and then we re-index the recursion to go over only odd indices t. Defining $\gamma_{k,i} = \lambda_{k,2i+1}$ and $\gamma_{\max,i} = \lambda_{\max,2i+1}$ and letting $t = 2s+1$, we obtain:

$$\lambda_{k,t} = \gamma_{k,s} \le \gamma_{k,s-1}(1-2\beta_k) + \beta_k\gamma_{\max,s-1} \le \gamma_{k,0}(1-2\beta_k)^s + \beta_k\sum_{i=1}^{s}(1-2\beta_k)^{s-i}\gamma_{\max,i-1}.$$

Note that we can bound $\gamma_{k,0} = \lambda_{k,1}$ as follows:

$$\lambda_{k,1} = \lambda_{k,0}(1-2\rho_k) + \rho_k\lambda_{\max,0} \le \rho_k(1-2\rho_k) + \rho_k = 2\rho_k(1-\rho_k) = \beta_k.$$

Using the inductive hypothesis, $\gamma_{\max,i-1} = \lambda_{\max,2i-1} \le \frac{C}{(2i)^\alpha}$, so we conclude:

$$\lambda_{k,t} \le \beta_k(1-2\beta_k)^s + (C/2^\alpha)\beta_k\sum_{i=1}^{s}\frac{(1-2\beta_k)^{s-i}}{i^\alpha} \le \frac{C/2^\alpha}{(s+2)^\alpha} \le \frac{C}{(t+2)^\alpha},$$

where in the second-to-last step we again used Lemma 11, choosing $C = 2^\alpha \cdot 1000$ and $\rho = \beta_k$. Thus, we have shown the inductive hypothesis, which concludes the proof of Theorem 2.

3.2. (Near-)Optimality of the Recursion Analysis

Given the peculiarity of the exponent $3/4+0.001$ in the convergence rate shown above, it is natural to ask what is the minimax optimal rate achieved by the matrix recursion defined in Lemma 10. To examine this question, we first give a nearly matching lower bound construction which shows that the exponent cannot be improved beyond $3/4+0.003$.

Theorem 12 *For any $T \ge 1$, there is a psd matrix $\mathbf{0} \preceq \bar{\mathbf{M}} \preceq \mathbf{I}$ such that the sequence*

$$\mathbf{N}_0 = \bar{\mathbf{M}}, \qquad \mathbf{N}_{t+1} = \mathbf{N}_t(\mathbf{I} - 2\bar{\mathbf{M}}) + \|\mathbf{N}_t\| \cdot \bar{\mathbf{M}}$$

satisfies $\|\mathbf{N}_t\| \ge c/t^{3/4+0.003}$ for every $t = 1, 2, ..., T$, where $c > 0$ is an absolute constant.

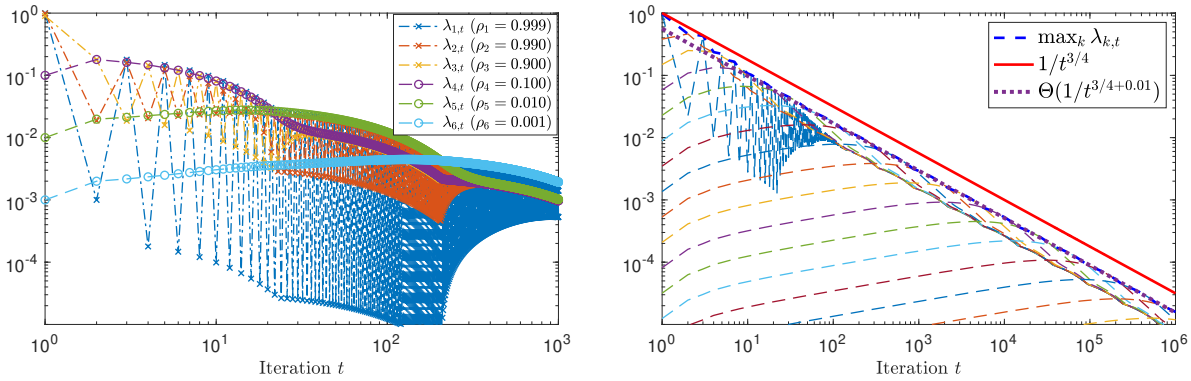

Figure 1: Two simulations of the eigenvalue recursion (4). On the left is a small example ($n = 6$) illustrating the two regimes that distinguish the behavior of eigenvalues below and above the 1/2 threshold. On the right is a larger example consisting of $n = 50$ eigenvalues spread evenly in the log-scale between 1 and 10^{-20}. Here, the dotted line is a best linear fit on the log-log plot for the evolution of $\max_k \lambda_{k,t}$ (the exponent $3/4 + 0.01$ was rounded). To best accommodate the log-log plot we start indexing the recursions at $t = 1$, not at 0.

Proof sketch. The key idea is to first show that the inequality in the summation bound (Lemma 11) does not hold with $\alpha = 3/4 + 0.003$. This follows essentially by showing the opposite inequality for some sufficiently small $\rho > 0$. We do this using a discrete-to-continuous reduction that, while different from the argument described in Section 4, has a similar flavor. Then, we extend this lower bound to the deterministic recursion by constructing a sufficiently large and ill-conditioned matrix $\bar{\mathbf{M}}$ so that it has eigenvalues that densely cover the interval $[\rho, 1]$, forcing the recursion to exhibit the worst-case behavior represented by the summation in Lemma 11. See Appendix G for details.

Numerical simulations. To further illustrate this phenomenon, we performed numerical simulations of the recursion using a number of initial conditions (defined by the eigenvalues ρ_k of a hypothetical matrix $\bar{\mathbf{M}}$). The results of these simulations exhibit a convergence behavior that aligns closely with our theory.

In Figure 1, we show the results of two simulations by plotting the evolution of all of the eigenvalues of the matrices $\mathbf{N}_t$. The left plot is a smaller example that is designed to highlight the two distinct convergence regimes that arise in our analysis, depending on whether the initial eigenvalue is below or above $1/2$. Here, we consider matrices of dimension $n = 6$, with the first three eigenvalues close to 1, and the latter three close to 0. As we can see, the initially large eigenvalues exhibit an oscillating behavior as the convergence progresses, whereas the initially small eigenvalues exhibit a more smooth trajectory. This is because the eigenvalue recursion (4) encounters a sign-change in the term $1 - 2\rho_k$ if and only if the initial eigenvalue ρ_k is larger than $1/2$. This is handled in the proof by separating out these two cases. Notably, in the end both cases rely on the same summation bound, but it is applied in a different way each time.

In order to demonstrate a convergence rate close to the minimax rate, we construct a larger example with $n = 50$ eigenvalues that exhibit fast exponential decay (Figure 1, right). As we can see, the maximum eigenvalue decays at a rate that nearly matches $O(1/t^{3/4})$, but a closer

examination shows that the actual rate is slightly better, at around $\alpha \approx 3/4 + 0.01$. We note that the gap between this rate and the minimax lower bound of $\alpha \approx 3/4 + 0.003$ from Theorem 12 is likely caused by the small size of the matrices we use in our simulation (due to computational constraints) compared to the matrices we use in the lower bound construction, which are much larger.

4. Proof Sketch of the Main Summation Bound

In this section, we give a sketch of the proof of the main technical lemma (Lemma 11) used to establish Theorem 2. For the sake of clarity, we overview the argument needed to obtain a slightly modified claim, with $\alpha = 3/4$ instead of $\alpha = 3/4 + 0.001$, but with a better constant, 300 instead of 1000. The remaining details for this simplified claim are in Appendices C-E, whereas the remaining details of the full proof of Lemma 11 with $\alpha = 3/4 + 0.001$ is in Appendix F.

Declaration of LLM assistance. An LLM-based assistant was used during the development of the proof of Lemma 11. The authors verified all steps and take responsibility for the result.

Modified Lemma 11. We will show that for all $K, t \geq 300$, $\rho \in (0, 1/2]$, and $\alpha = 3/4$, we have

$$\rho(1-2\rho)^t + K\rho \sum_{i=1}^{t} \frac{(1-2\rho)^{t-i}}{i^\alpha} \leq \frac{K}{(t+2)^\alpha}.$$

In our argument, we will treat the two terms on the left hand side separately. To that end, define

$$A_t(\rho) := \rho(1-2\rho)^t \quad \text{and} \quad B_t(\rho, \alpha) := \rho \sum_{i=1}^{t} \frac{(1-2\rho)^{t-i}}{i^\alpha}.$$

Multiplying both terms by $(t+2)^\alpha$, the inequality is now equivalent to

$$(t+2)^\alpha A_t(\rho) + K(t+2)^\alpha B_t(\rho, \alpha) \leq K.$$

We will next bound the suprema of both terms over all admissible values of t and ρ:

$$A_{\sup}(\alpha) := \sup_{t \geq 300} \sup_{\rho \in (0,1/2]} (t+2)^\alpha A_t(\rho), \qquad B_{\sup}(\alpha) := \sup_{t \geq 300} \sup_{\rho \in (0,1/2]} (t+2)^\alpha B_t(\rho, \alpha). \quad (5)$$

The following lemma provides a simple characterization of the constant K that can satisfy our inequality.

Lemma 13 *Fix $\alpha > 0$. Assume $A_{\sup}(\alpha) < \infty$ and $B_{\sup}(\alpha) < 1$. Then for any $K > 0$ satisfying*

$$K \geq \frac{A_{\sup}(\alpha)}{1 - B_{\sup}(\alpha)} \quad (6)$$

we have

$$\rho(1-2\rho)^t + K\rho \sum_{i=1}^{t} \frac{(1-2\rho)^{t-i}}{i^\alpha} \leq \frac{K}{(t+2)^\alpha}$$

for all $t \geq 300$, $\rho \in (0, 1/2]$.

Proof If K satisfies (6), then $A_{\sup}(\alpha) \le K(1 - B_{\sup}(\alpha))$, so

$$A_{\sup}(\alpha) + K\, B_{\sup}(\alpha) \le K(1 - B_{\sup}(\alpha)) + K\, B_{\sup}(\alpha) = K.$$

Therefore, we have

$$\sup_{t \ge 300} \sup_{\rho \in (0, \frac{1}{2}]} \left((t+2)^\alpha A_t(\rho) + K(t+2)^\alpha B_t(\rho)\right) \le K$$

which implies the desired result. ∎

The first term $A_{\sup}(\alpha)$ is easier to bound because we can find the exact maximizers ρ and t that attain the value $A_{\sup}(\alpha) := \sup_{t \ge 300} \sup_{\rho \in (0, 1/2]} (t+2)^\alpha A_t(\rho)$ (see Lemma 20 and Lemma 21 for details), and eventually we obtain

$$A_{\sup}(3/4) \le \frac{363307}{7228832} < 0.051.$$

To bound the second term, which is a discrete sum, we perform a discrete-to-continuous reduction by comparing it with an integral. For simplicity, we will replace the $(t+2)^\alpha$ factor with t^α, which can be easily addressed later for sufficiently large t. Then, letting $\delta := -\log(1 - 2\rho) \ge 0$, we have:

$$(1 - 2\rho)^{t-i} = e^{-\delta(t-i)}.$$

Therefore, the quantity $t^\alpha B_t(\rho, \alpha)$ corresponds to the integral $\rho\, t \int_0^1 e^{-\delta t(1-u)} u^{-\alpha}\, du$, which after a simple change of variable, $\theta := \delta t/2$ (see Remark 23), results in the following function for $\theta > 0$:

$$L_\alpha(\theta) := \theta \int_0^1 e^{-2\theta u} (1-u)^{-\alpha}\, du. \tag{7}$$

This way, the integral comparison reduces the two-parameter family (t, ρ) to the one-parameter family θ. Formally, by Lemma 25 and Lemma 26, for any $0 < \alpha < 1$, $t \ge 1$, and $\rho \in (0, 1/2]$, we have the following comparison:

$$t^\alpha B_t(\rho, \alpha) \le \max\left\{ L_\alpha\Big(\frac{t}{2}\big(-\log(1-2\rho)\big)\Big), 1/2 \right\}. \tag{8}$$

Therefore, it suffices to bound $\sup_{\theta > 0} L_\alpha(\theta)$. To this end, we first observe the following ordinary differential equation (ODE) property of $L_\alpha(\theta)$,

$$L'_\alpha(\theta) = 1 - \Big(2 - \frac{\alpha}{\theta}\Big) L_\alpha(\theta), \qquad \theta > 0. \tag{9}$$

(see Lemma 27 for details). This ODE property implies that we can bound $\sup_{\theta > 0} L_\alpha(\theta)$ by checking a single point. More precisely, we have the following result.

Lemma 14 (One-point criterion) *For any $0 < \alpha < 1$ and $\ell \in (1/2, 1)$, define*

$$\theta_\ell := \frac{\alpha}{2 - \frac{1}{\ell}}. \tag{10}$$

If $L_\alpha(\theta_\ell) < \ell$, then $L_\alpha(\theta) < \ell$ for all $\theta > 0$, and therefore $\sup_{\theta > 0} L_\alpha(\theta) \le \ell$.

Proof sketch Here we explain the key intuition and ideas of the proof, while the full proof is in Appendix D. Assume for contradiction that there exists $\theta_0 > 0$ with $L_\alpha(\theta_0) \geq \ell$. By standard limit calculation, we have $0 \leq L_\alpha(\theta) \to 0$, as $\theta \to 0$. Therefore, $L_\alpha(\theta) < \ell$ for all sufficiently small $\theta > 0$. By continuity of L_α , the set $\{\theta > 0 : \ L_\alpha(\theta) = \ell\}$ is nonempty, and the first hitting time

$$\tau := \inf\{\theta > 0 : \ L_\alpha(\theta) = \ell\}$$

is therefore well defined. Then $\tau > 0$, $L_\alpha(\theta) < \ell$ for all $\theta \in (0, \tau)$, and $L_\alpha(\tau) = \ell$. Using a standard argument from calculus (Lemma 18), this implies that $L'_\alpha(\tau) \geq 0$.

Next, using the ODE property, we make the following key observation about the structure of $L'_\alpha(\theta)$ when L_α passes the $y = \ell$ threshold. Define for $\theta > 0$ the scalar function

$$\Psi(\theta) := 1 - \Big(2 - \frac{\alpha}{\theta}\Big)\ell.$$

Since L_α satisfies the ODE (9), at any $\theta > 0$ with $L_\alpha(\theta) = \ell$ we have

$$L'_\alpha(\theta) = \Psi(\theta).$$

Also, a simple calculation shows that

$$\Psi(\theta) > 0 \ \text{ for } \ \theta < \theta_\ell, \qquad \Psi(\theta) = 0 \ \text{ for } \ \theta = \theta_\ell, \qquad \Psi(\theta) < 0 \ \text{ for } \ \theta > \theta_\ell.$$

Note that we cannot conclude that θ_ℓ is global maximizer because $L'_\alpha(\theta) = \Psi(\theta)$ only holds for $L_\alpha(\theta) = \ell$. However, the above observation shows that every time when L_α passes the line $y = \ell$ before θ_ℓ, the function must be increasing, so it will not be able to go down and satisfy our assumption $L_\alpha(\theta_\ell) < \ell$. This is sufficient to attain contradiction. ■

Having shown Lemma 14, we just need to check

$$L_{3/4}\Big(\theta_{\ell_0}\Big) < \ell_0,$$

where $\ell_0 = \frac{497}{500}$ (see Lemma 32 for details) and then conclude that $B_{\sup}(3/4) \leq \frac{201}{200} \cdot \frac{497}{500} = \frac{99897}{100000} < 1$ (see Lemma 33 for details).

Finally, combining $A_{\sup}(3/4) \leq 0.051$ and $B_{\sup}(3/4) \leq \frac{99897}{100000}$, we have

$$\frac{A_{\sup}(3/4)}{1 - B_{\sup}(3/4)} \leq \frac{\frac{51}{1000}}{\frac{103}{100000}} = \frac{51}{1000} \cdot \frac{100000}{103} = \frac{5100}{103} < 50 < 300.$$

Therefore, using Lemma 13, we obtain the desired claim.

5. Conclusions

We gave a new last-iterate convergence analysis for a class of SGD algorithms with greedy step size, which includes Randomized Kaczmarz and Randomized Coordinate Descent, among many others, showing that they attain an $O(1/t^{3/4})$ expected convergence rate. Our results provide a direct improvement over the previously known $O(1/t^{1/2})$ guarantee for this setting and have implications for the analysis of catastrophic forgetting in realizable continual learning problems.

Acknowledgments

MD was supported in part by NSF CAREER Grant CCF-233865 and a Google ML and Systems Junior Faculty Award. This work was done in part while MD was visiting the Simons Institute for the Theory of Computing. The authors thank Shabarish Chenakkod, Yuji Nakatsukasa, Elizaveta Rebrova, and Mark Rudelson for helpful conversations.

References

Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

Amit Attia, Matan Schliserman, Uri Sherman, and Tomer Koren. Fast last-iterate convergence of sgd in the smooth interpolation regime. *arXiv preprint arXiv:2507.11274*, 2025.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3rd edition, 2016. ISBN 978-1-886529-05-2. URL http://www.athenasc.com/nonlinbook.html.

Xufeng Cai and Jelena Diakonikolas. Last iterate convergence of incremental methods and applications in continual learning. *arXiv preprint arXiv:2403.06873*, 2024.

Michał Dereziński and Michael W Mahoney. Recent and upcoming developments in randomized numerical linear algebra for machine learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6470–6479, 2024.

Michał Dereziński and Elizaveta Rebrova. Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition. *SIAM Journal on Mathematics of Data Science*, 6(1):127–153, 2024.

Michał Dereziński and Jiaming Yang. Solving dense linear systems faster than via preconditioning. In *56th Annual ACM Symposium on Theory of Computing*, 2024.

Michał Dereziński, Daniel LeJeune, Deanna Needell, and Elizaveta Rebrova. Fine-grained analysis and faster algorithms for iteratively solving linear systems. *Journal of Machine Learning Research*, 26(144):1–49, 2025a.

Michał Dereziński, Deanna Needell, Elizaveta Rebrova, and Jiaming Yang. Randomized kaczmarz methods with beyond-krylov convergence. *SIAM Journal on Matrix Analysis and Applications*, 46(4):2558–2588, 2025b.

Tommy Elfving. Block-iterative methods for consistent and inconsistent linear equations. *Numerische Mathematik*, 35(1):1–12, 1980.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pages 4028–4079. PMLR, 2022.

Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. From continual learning to sgd and back: Better rates for continual linear models. In *Fourth Conference on Lifelong Learning Agents-Workshop Track*, 2025.

Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in neural information processing systems*, 32, 2019.

Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755. PMLR, 2019.

M. S. Kaczmarz. Angenaherte auflosung von systemen linearer gleichungen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres*, 35:355–357, 1937.

Dennis Leventhal and Adrian S Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. *arXiv preprint arXiv:2506.06501*, 2025.

Chaoyue Liu, Dmitriy Drusvyatskiy, Misha Belkin, Damek Davis, and Yian Ma. Aiming towards the minimizers: fast convergence of sgd for overparametrized problems. *Advances in neural information processing systems*, 36:60748–60767, 2023.

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. *The Twelfth International Conference on Learning Representations*, 2023.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects, 2013.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

Pratik Rathore, Zachary Frangella, Jiaming Yang, Michał Dereziński, and Madeleine Udell. Have askotch: A neat solution for large-scale kernel ridge regression. *arXiv preprint arXiv:2407.10070*, 2025.

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

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013.

Stefan Steinerberger. Approximate solutions of linear systems at a universal rate. *SIAM Journal on Matrix Analysis and Applications*, 44(3):1436–1446, 2023.

Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

William Swartworth, Deanna Needell, Rachel Ward, Mark Kong, and Halyun Jeong. Nearly optimal bounds for cyclic forgetting. *Advances in neural information processing systems*, 36:68197–68206, 2023.

Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 34:21581–21591, 2021.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International conference on machine learning*, pages 24280–24314. PMLR, 2022.

Moslem Zamani and Francois Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35(3):2182–2201, 2025.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on learning theory*, pages 4633–4635. PMLR, 2021.

Appendix A. Additional Related Work

In this section, we overview the most relevant prior work related to the convergence analysis of SGD and Randomized Kaczmarz.

Last-iterate convergence of SGD. Extensive work has been dedicated to last-iterate analysis of SGD, particularly in the general convex and Lipschitz setting (Shamir and Zhang, 2013; Jain et al., 2019; Harvey et al., 2019; Liu and Zhou, 2023; Zamani and Glineur, 2025). These results essentially match the corresponding guarantees attained by the average iterate in the convex Lipschitz setting. The last-iterate convergence of SGD has also been extensively studied in the interpolation regime (Ge et al., 2019; Vaswani et al., 2019; Berthier et al., 2020; Varre et al., 2021; Wu et al., 2022; Liu et al., 2023), which is motivated by modern over-parameterized machine learning models (Ma et al., 2018). However, until the work of Evron et al. (2025); Attia et al. (2025), none of these prior results captured last-iterate SGD with the greedy step size, or the Randomized Kaczmarz method.

Connections to continual learning models. SGD with greedy step size is closely connected with the analysis of continual learning models, and particularly the phenomenon of catastrophic forgetting (Mermillod et al., 2013). A number of theoretical works have shown that catastrophic forgetting can be mitigated by using randomization (Evron et al., 2022; Swartworth et al., 2023; Cai and Diakonikolas, 2024). Most notably, Evron et al. (2022, 2025) showed that convergence bounds for SGD with greedy step size in the smooth interpolation regime can be used to provide bounds for continual linear regression in the realizable setting. Our results, such as Corollary 5, can be directly applied to their framework, improving on those guarantees. We note that an alternative approach has been developed by Levinstein et al. (2025) that attains optimal bounds for continual linear regression without relying on SGD with greedy step size.

Randomized linear system solvers. The last-iterate convergence analysis of randomized iterative methods for solving linear systems, such as Randomized Kaczmarz (Strohmer and Vershynin, 2009), Randomized Coordinate Descent (Leventhal and Lewis, 2010), or Sketch-and-Project (Gower and Richtárik, 2015), has focused primarily on guarantees that depend on some form of a condition number of the problem (Dereziński and Mahoney, 2024), which effectively corresponds to the strongly convex setting in SGD analysis. For example, in the case of Randomized Kaczmarz, Strohmer and Vershynin (2009) established convergence of the form $\mathbb{E}\,\|\mathbf{x}_t - \mathbf{x}^*\|^2 \leq (1-\rho)^t \|\mathbf{x}_0 - \mathbf{x}^*\|^2$ where $\rho = \|\mathbf{A}\|_F^2/\sigma_{\min}^2(\mathbf{A})$. Faster rates are known for Block Kaczmarz (Dereziński and Yang, 2024; Dereziński et al., 2025b) and other instances of sketch-and-project (Dereziński and Rebrova, 2024; Rathore et al., 2025), but all of them become vacuous as $\sigma_{\min}(\mathbf{A})$ tends to zero. Steinerberger (2023) studies conditioning-free convergence of a non-constructive Kaczmarz procedure that attains an $O(\log(t)/t)$ rate but only for the optimal sequence of equation selections. As an auxiliary result, we show that a constructive Kaczmarz procedure attains an $O(1/t)$ rate, by choosing a random iterate (instead of the last iterate) in Randomized Kaczmarz.

Appendix B. Calculus Preliminaries

In this section, we collect auxiliary lemmas that describe standard results from calculus which are useful in our proof of Lemma 11.

Lemma 15 (Bernoulli's inequality for $0 < \alpha \leq 1$) *If $0 < \alpha \leq 1$, then for all $u \geq 0$,*

$$(1+u)^\alpha \leq 1 + \alpha u.$$

Lemma 16 (Tangent and quadratic lower bounds for e^x)

(i) For all $x \in \mathbb{R}$, $e^x \geq 1 + x$.

(ii) For all $x \geq 0$, $e^x \geq 1 + x + \frac{x^2}{2}$.

Lemma 17 *For every real number x with $x > 0$ or $x < -1$,*

$$\log\Big(1 + \frac{1}{x}\Big) \geq \frac{1}{x+1}.$$

Lemma 18 (Derivative at a first hitting time from below) *Let $f : (0,\infty) \to \mathbb{R}$ be continuous and differentiable at $\tau > 0$. Fix $\ell \in \mathbb{R}$ and assume*

$$\tau = \inf\{\theta > 0 : \ f(\theta) = \ell\}, \qquad \textit{and} \qquad f(\theta) < \ell \ \textit{for all } \theta \in (0,\tau).$$

Then $f'(\tau) \geq 0$.

Proof For $h > 0$ small, $f(\tau - h) < \ell = f(\tau)$, so

$$\frac{f(\tau) - f(\tau - h)}{h} \geq 0.$$

Since f is differentiable at τ, the desired result holds by definition of derivative. ■

Lemma 19 (Derivative at a first return time from above) *Let $f : (0,\infty) \to \mathbb{R}$ be continuous and differentiable at $\sigma > 0$. Fix $\ell \in \mathbb{R}$ and assume*

$$\sigma = \inf\{\theta > \tau : \ f(\theta) = \ell\}$$

for some $\tau \geq 0$, and $f(\theta) > \ell$ for all $\theta \in (\tau, \sigma)$. Then $f'(\sigma) \leq 0$.

Proof For $h > 0$ small, $f(\sigma - h) > \ell = f(\sigma)$, so

$$\frac{f(\sigma) - f(\sigma - h)}{h} \leq 0.$$

Since f is differentiable at σ, the desired result holds by definition of derivative. ■

Appendix C. Bounds for $A_{\sup}(\alpha)$

In this section, we provide the intermediate results for bounding the $A_t(\rho)$, which is the easier of the two terms analyzed in the proof of Lemma 11.

Lemma 20 (Exact maximizer of A_t) *Fix $t \geq 1$. The function $\rho \mapsto A_t(\rho) = \rho(1-2\rho)^t$ on $(0, \frac{1}{2}]$ attains its maximum at $\rho^* = \frac{1}{2(t+1)}$, with*

$$\sup_{\rho \in (0,1/2]} A_t(\rho) = \frac{1}{2(t+1)}\Big(\frac{t}{t+1}\Big)^t.$$

Proof For $\rho \in (0, \frac{1}{2})$, direct differentiation shows that

$$A_t'(\rho) = (1-2\rho)^t + \rho \cdot t(1-2\rho)^{t-1}(-2) = (1-2\rho)^{t-1}\big(1 - 2(t+1)\rho\big).$$

Since $(1-2\rho)^{t-1} > 0$ on $(0, \frac{1}{2})$, the unique critical point is $\rho^* = \frac{1}{2(t+1)}$. In addition, $A_t(\rho) \to 0$ as $\rho \to 0$ and $A_t(\frac{1}{2}) = 0$, and therefore the unique critical point is the global maximizer. Evaluating A_t at ρ^* gives the stated formula. ■

In the following lemma, we provide a bound for $A_{\sup}(\alpha)$ that is used in the analysis of the $\alpha = 3/4$ version of Lemma 11 that is described in Section 4.

Lemma 21 (Bound for $A_{\sup}(\alpha)$ for $t \geq 300$) *Let $\alpha_0 = \frac{3}{4}$. Then*

$$A_{\sup}(\alpha_0) \leq \frac{363307}{7228832} < 0.051.$$

Proof Fix an integer $t \geq 300$. By Lemma 20, we have

$$\sup_{\rho \in (0, \frac{1}{2}]} (t+2)^{\alpha_0} A_t(\rho) = \frac{(t+2)^{3/4}}{2(t+1)} \Big(\frac{t}{t+1}\Big)^t. \tag{11}$$

Define, for real $x > 0$,

$$f(x) := \frac{(x+2)^{3/4}}{2(x+1)} \Big(\frac{x}{x+1}\Big)^x. \tag{12}$$

Then (11) reads $\sup_\rho (t+2)^{\alpha_0} A_t(\rho) = f(t)$.

Since $f(x) > 0$ for $x > 0$, we define

$$g(x) := \log f(x) \qquad (x > 0).$$

We will compute $g'(x)$ and show $g'(x) < 0$ for all $x > 0$. Direct differentiation show that for every $x > 0$,

$$\begin{aligned} g'(x) &= \frac{3}{4(x+2)} - \frac{1}{x+1} + \log\Big(\frac{x}{x+1}\Big) + \frac{1}{x+1} \\ &= \frac{3}{4(x+2)} + \log\Big(\frac{x}{x+1}\Big) = \frac{3}{4(x+2)} - \log\Big(1 + \frac{1}{x}\Big). \end{aligned} \tag{13}$$

Using the inequality $\log\Big(1 + \frac{1}{x}\Big) \geq \frac{1}{x+1}$ (see Lemma 17 for the formal proof), we have

$$g'(x) \leq \frac{3}{4(x+2)} - \frac{1}{x+1} = \frac{3(x+1) - 4(x+2)}{4(x+1)(x+2)} = -\frac{x+5}{4(x+1)(x+2)} < 0.$$

for all $x > 0$.

Therefore, the function f is strictly decreasing on $(0, +\infty)$, so we have

$$A_{\sup}(\alpha_0) \leq f(300). \tag{14}$$

Direct calculation (see Lemma 22 for details) shows that

$$f(300) = \frac{302^{3/4}}{2 \cdot 301} \Big(\frac{300}{301}\Big)^{300} < 0.051 \tag{15}$$

■

Lemma 22 *Let*

$$f(x) := \frac{(x+2)^{3/4}}{2(x+1)} \Big(\frac{x}{x+1}\Big)^x. \tag{16}$$

defined on $x \in (0, +\infty)$. Then we have

$$f(300) = \frac{302^{3/4}}{2 \cdot 301} \Big(\frac{300}{301}\Big)^{300} < 0.051 \tag{17}$$

Proof From (16), we have

$$f(300) = \frac{302^{3/4}}{2 \cdot 301} \Big(\frac{300}{301}\Big)^{300}. \tag{18}$$

We bound the two factors in (15).

(i) Bound $\left(\frac{300}{301}\right)^{300}$. By Lemma 16 applied at $x = -\frac{1}{301}$,

$$e^{-\frac{1}{301}} \geq 1 - \frac{1}{301} = \frac{300}{301}.$$

Raising both sides to the power 300 (which preserves the inequality since both sides are positive) yields

$$\Big(\frac{300}{301}\Big)^{300} \leq e^{-\frac{300}{301}}.$$

By Lemma 16 (second inequality) applied at $x = \frac{300}{301} \geq 0$,

$$e^{-\frac{300}{301}} \leq \frac{1}{1 + \frac{300}{301} + \frac{1}{2}\left(\frac{300}{301}\right)^2} = \frac{90601}{225901}.$$

Therefore,

$$\Big(\frac{300}{301}\Big)^{300} \leq \frac{90601}{225901}. \tag{19}$$

(ii) Bound $\frac{302^{3/4}}{301}$. Write $302 = 301\left(1 + \frac{1}{301}\right)$, hence

$$\frac{302^{3/4}}{301} = 301^{-1/4}\Big(1 + \frac{1}{301}\Big)^{3/4}.$$

Since $301 \geq 256 = 4^4$, we have $301^{-1/4} \leq \frac{1}{4}$. By Lemma 15 with $\alpha = \frac{3}{4}$ and $u = \frac{1}{301}$,

$$\Big(1 + \frac{1}{301}\Big)^{3/4} \leq 1 + \frac{3}{4 \cdot 301} = \frac{1207}{1204}.$$

Therefore

$$\frac{302^{3/4}}{301} \leq \frac{1}{4} \cdot \frac{1207}{1204} = \frac{1207}{4816}. \tag{20}$$

(iii) Combine (15), (19), and (20). Using (19) and (20) in (15) gives

$$f(300) \leq \frac{1}{2} \cdot \frac{1207}{4816} \cdot \frac{90601}{225901} = \frac{1207 \cdot 90601}{2 \cdot 4816 \cdot 225901}.$$

Compute the numerator and denominator:

$$1207 \cdot 90601 = 109355407, \qquad 2 \cdot 4816 \cdot 225901 = 2175878432.$$

Since $109355407 = 301 \cdot 363307$ and $2175878432 = 301 \cdot 7228832$, we simplify to

$$f(300) \leq \frac{363307}{7228832}. \tag{21}$$

Finally, to verify $\frac{363307}{7228832} < 0.051 = \frac{51}{1000}$, cross-multiply:

$$363307 \cdot 1000 = 363307000 < 368670432 = 7228832 \cdot 51.$$

Thus $\frac{363307}{7228832} < 0.051$.

Therefore, we have

$$f(300) \le \frac{363307}{7228832} < 0.051,$$

as claimed. ■

Appendix D. Properties of $L_\alpha(\theta)$ and connection with $B_{\sup}(\alpha)$

In this section, we describe the discrete-to-continuous reduction, and the resulting ODE, that is used later to bound $B_t(\alpha)$, the second of the two terms in the analysis of Lemma 11.

Remark 23 (Motivation of the definition of $L_\alpha(\theta)$) *Write $r := 1-2\rho \in [0,1)$ and $\delta := -\log r = -\log(1-2\rho) \ge 0$, so that $r = e^{-\delta}$. Then for $0 < \rho < \frac{1}{2}$ we have $\delta > 0$ and*

$$(1-2\rho)^{t-i} = r^{\,t-i} = e^{-\delta(t-i)}.$$

In particular, when ρ is small one has $\delta \sim 2\rho$, so $\delta(t-i) \approx 2\rho(t-i)$. Now rewrite

$$B_t(\rho,\alpha) = \rho \sum_{i=1}^{t} \frac{(1-2\rho)^{t-i}}{i^\alpha} = \rho \sum_{i=1}^{t} \frac{e^{-\delta(t-i)}}{i^\alpha}.$$

Therefore the quantity $t^\alpha B_t(\rho,\alpha)$

$$\rho\, t \int_0^1 e^{-\delta t(1-u)} u^{-\alpha}\, du,$$

which depends on t and ρ only through the combined parameter δt. Therefore, by change of variable

$$\theta := \frac{\delta t}{2} = \frac{t}{2}\big(-\log(1-2\rho)\big),$$

the resulting integral for comparison takes the form

$$L_\alpha(\theta) = \theta \int_0^1 e^{-2\theta u}(1-u)^{-\alpha}\, du.$$

In this way the integral comparison reduces the two-parameter family (t,ρ) to the one-parameter family θ.

Remark 24 (A simplification of $L_{3/4}$) *For every $\theta > 0$, the function $L_{3/4}$ admits the non-singular representation*

$$L_{3/4}(\theta) = 4\theta e^{-2\theta} \int_0^1 e^{2\theta y^4}\, dy = 4\theta \int_0^1 e^{-2\theta(1-y^4)}\, dy. \tag{22}$$

In particular, for each fixed $\theta > 0$ the integrand $y \mapsto e^{-2\theta(1-y^4)}$ is C^∞ on $[0,1]$.

Proof Fix $\theta > 0$. By definition,

$$L_{3/4}(\theta) = \theta \int_0^1 e^{-2\theta u}(1-u)^{-3/4}\,du.$$

Since $\int_0^1 (1-u)^{-3/4}\,du = 4 < \infty$ and $0 \le e^{-2\theta u} \le 1$, the integral is absolutely convergent (and is understood as an improper integral at $u = 1$).

First, for $\varepsilon \in (0,1)$ set

$$I_\varepsilon := \int_0^{1-\varepsilon} e^{-2\theta u}(1-u)^{-3/4}\,du.$$

On $[0, 1-\varepsilon]$ the substitution $w = 1-u$ is legitimate (it is C^1 and strictly monotone), giving

$$I_\varepsilon = \int_\varepsilon^1 e^{-2\theta(1-w)} w^{-3/4}\,dw = e^{-2\theta}\int_\varepsilon^1 e^{2\theta w} w^{-3/4}\,dw.$$

Letting $\varepsilon \to 0^+$ (which is valid since the original integral converges absolutely) yields

$$L_{3/4}(\theta) = \theta e^{-2\theta}\int_0^1 e^{2\theta w} w^{-3/4}\,dw. \tag{23}$$

Second, for $\varepsilon \in (0,1)$ define

$$J_\varepsilon := \int_\varepsilon^1 e^{2\theta w} w^{-3/4}\,dw.$$

On $[\varepsilon, 1]$ the substitution $w = y^4$ is legitimate (again C^1 and strictly increasing), and we obtain

$$J_\varepsilon = \int_{\varepsilon^{1/4}}^1 e^{2\theta y^4}(y^4)^{-3/4}\,4y^3\,dy = 4\int_{\varepsilon^{1/4}}^1 e^{2\theta y^4}\,dy.$$

As $\varepsilon \to 0^+$, the left-hand side satisfies $J_\varepsilon \to \int_0^1 e^{2\theta w} w^{-3/4}\,dw$ by the definition of the improper integral. On the right-hand side, since $0 \le e^{2\theta y^4} \le e^{2\theta}$ on $[0,1]$,

$$0 \le 4\int_0^{\varepsilon^{1/4}} e^{2\theta y^4}\,dy \le 4e^{2\theta}\varepsilon^{1/4} \xrightarrow[\varepsilon\to 0^+]{} 0,$$

so

$$4\int_{\varepsilon^{1/4}}^1 e^{2\theta y^4}\,dy \xrightarrow[\varepsilon\to 0^+]{} 4\int_0^1 e^{2\theta y^4}\,dy.$$

Therefore,

$$\int_0^1 e^{2\theta w} w^{-3/4}\,dw = 4\int_0^1 e^{2\theta y^4}\,dy.$$

Insert this into (23) to obtain

$$L_{3/4}(\theta) = 4\theta e^{-2\theta}\int_0^1 e^{2\theta y^4}\,dy,$$

which is the first equality in (22). The second equality follows from $e^{-2\theta}e^{2\theta y^4} = e^{-2\theta(1-y^4)}$.

Finally, since $y \mapsto 1-y^4$ is a polynomial and $x \mapsto e^x$ is C^∞ on $\mathbb{R}$, the composition $y \mapsto e^{-2\theta(1-y^4)}$ is C^∞ on $[0,1]$ for each fixed $\theta > 0$. ∎

Lemma 25 (Discrete to continuous) *Fix $0 < \alpha < 1$, $t \geq 1$, and $\rho \in (0, \frac{1}{2})$. Then*

$$t^\alpha B_t(\rho, \alpha) \leq L_\alpha\Big(\frac{t}{2}\big(-\log(1-2\rho)\big)\Big). \tag{24}$$

Proof Set $r := 1 - 2\rho \in (0,1)$ and $\delta := -\log r > 0$, so $r = e^{-\delta}$. Reindex with $k = t - i$:

$$B_t(\rho, \alpha) = \rho \sum_{k=0}^{t-1} \frac{r^k}{(t-k)^\alpha}.$$

Using $\rho = \frac{1-r}{2}$, for each $k \geq 0$, we have

$$\rho r^k = \frac{1-r}{2} r^k = \frac{1}{2}(r^k - r^{k+1}).$$

Since $r = e^{-\delta}$, we have

$$r^k - r^{k+1} = e^{-\delta k} - e^{-\delta(k+1)} = \int_k^{k+1} \delta e^{-\delta x}\, dx.$$

Therefore, we have

$$\rho r^k = \frac{1}{2} \int_k^{k+1} \delta e^{-\delta x}\, dx.$$

Substitute the above identity into the sum, we have

$$B_t(\rho, \alpha) = \frac{1}{2} \sum_{k=0}^{t-1} \int_k^{k+1} \delta e^{-\delta x} \frac{dx}{(t-k)^\alpha}.$$

For $x \in [k, k+1]$, we have $x \geq k$ so $t - x \leq t - k$, and since $y \mapsto y^{-\alpha}$ is decreasing on $(0, \infty)$,

$$(t-k)^{-\alpha} \leq (t-x)^{-\alpha}.$$

Therefore

$$B_t(\rho, \alpha) \leq \frac{1}{2} \int_0^t \delta e^{-\delta x} (t-x)^{-\alpha}\, dx.$$

Using change of variable $x = tu$, we have

$$B_t(\rho, \alpha) \leq \frac{1}{2}\, \delta t^{1-\alpha} \int_0^1 e^{-\delta t u} (1-u)^{-\alpha}\, du.$$

Multiplying both sides by t^α and setting $\theta = \delta t/2$ shows the desired inequality (24). ■

Lemma 26 (Endpoint $\rho = \frac{1}{2}$) *Fix $0 < \alpha < 1$ and $t \geq 1$. Then*

$$B_t\Big(\tfrac{1}{2}, \alpha\Big) = \frac{1}{2t^\alpha}, \qquad \textit{equivalently} \qquad t^\alpha B_t\Big(\tfrac{1}{2}, \alpha\Big) = \frac{1}{2}.$$

Proof If $\rho = \frac{1}{2}$, then $r = 1 - 2\rho = 0$. In $B_t(\rho, \alpha)$, the factor r^{t-i} vanishes unless $i = t$, in which case $r^0 = 1$. Thus $B_t(\frac{1}{2}, \alpha) = \frac{1}{2} \cdot t^{-\alpha}$. ∎

Lemma 27 (ODE for L_α) *Fix $0 < \alpha < 1$. The function L_α is continuously differentiable on $(0, \infty)$ and satisfies*

$$L'_\alpha(\theta) = 1 - \Big(2 - \frac{\alpha}{\theta}\Big) L_\alpha(\theta), \qquad \theta > 0. \tag{25}$$

Proof Define

$$I(\theta) := \int_0^1 e^{-2\theta u}(1-u)^{-\alpha}\, du, \qquad \text{so that} \qquad L_\alpha(\theta) = \theta I(\theta).$$

For fixed $u \in [0, 1)$,

$$\frac{\partial}{\partial \theta}\big(e^{-2\theta u}(1-u)^{-\alpha}\big) = (-2u)e^{-2\theta u}(1-u)^{-\alpha}.$$

Since $0 \le u \le 1$ and $e^{-2\theta u} \le 1$,

$$\big|(-2u)e^{-2\theta u}(1-u)^{-\alpha}\big| \le 2(1-u)^{-\alpha}.$$

Because $0 < \alpha < 1$, $(1-u)^{-\alpha} \in L^1(0,1)$, hence by differentiation under integration (justified by dominate convergence)

$$I'(\theta) = \int_0^1 (-2u)e^{-2\theta u}(1-u)^{-\alpha}\, du.$$

Therefore $L'_\alpha(\theta) = I(\theta) + \theta I'(\theta)$ exists.

Now define

$$J(\theta) := \int_0^1 e^{-2\theta u}(1-u)^{1-\alpha}\, du.$$

Since $u(1-u)^{-\alpha} = (1-u)^{-\alpha} - (1-u)^{1-\alpha}$, we have

$$L'_\alpha(\theta) = (1 - 2\theta)I(\theta) + 2\theta J(\theta).$$

Let $w(u) = (1-u)^{1-\alpha}$, so $w(0) = 1$, $w(1) = 0$, and $w'(u) = -(1-\alpha)(1-u)^{-\alpha}$. Integrating by parts on $[0, 1-\varepsilon]$ and letting $\varepsilon \to 0^+$ (justified by Dominated Convergence Theorem since the integrands are dominated by an L^1 function), one obtains

$$I(\theta) = \frac{1}{1-\alpha}\big(1 - 2\theta J(\theta)\big) \qquad \Longleftrightarrow \qquad J(\theta) = \frac{1 - (1-\alpha)I(\theta)}{2\theta}.$$

Substituting into $L'_\alpha(\theta) = (1-2\theta)I(\theta) + 2\theta J(\theta)$ yields

$$L'_\alpha(\theta) = 1 - (2\theta - \alpha)I(\theta) = 1 - \Big(2 - \frac{\alpha}{\theta}\Big) L_\alpha(\theta),$$

which is (25). Continuity of L'_α follows from continuity of I and I' obtained via dominated convergence. ∎

Lemma 28 (Restated Lemma 14) *For any $0 < \alpha < 1$ and $\ell \in (\frac{1}{2}, 1)$, we define*

$$\theta_\ell := \frac{\alpha}{2 - \frac{1}{\ell}}. \tag{26}$$

If $L_\alpha(\theta_\ell) < \ell$, then $L_\alpha(\theta) < \ell$ for all $\theta > 0$, and therefore $\sup_{\theta>0} L_\alpha(\theta) \le \ell$.

Proof Assume for contradiction that there exists $\theta_0 > 0$ with $L_\alpha(\theta_0) \ge \ell$. Since $0 < e^{-2\theta u} \le 1$ and $(1-u)^{-\alpha} \in L^1(0,1)$ for $0 < \alpha < 1$,

$$0 \le L_\alpha(\theta) = \theta \int_0^1 e^{-2\theta u}(1-u)^{-\alpha}\, du \le \theta \int_0^1 (1-u)^{-\alpha}\, du = \frac{\theta}{1-\alpha} \to 0, \text{ as } \theta \to 0$$

Therefore, $L_\alpha(\theta) < \ell$ for all sufficiently small $\theta > 0$. By continuity of L_α (from its defining integral), the set $\{\theta > 0 :\ L_\alpha(\theta) = \ell\}$ is nonempty, and we may define the first hitting time

$$\tau := \inf\{\theta > 0 :\ L_\alpha(\theta) = \ell\}.$$

Then $\tau > 0$, $L_\alpha(\theta) < \ell$ for all $\theta \in (0, \tau)$, and $L_\alpha(\tau) = \ell$. By Lemma 18, $L'_\alpha(\tau) \ge 0$.

We have the following key observation about the structure of $L'_\alpha(\theta)$ when L_α passes the line $y = \ell$. Define for $\theta > 0$ the scalar function

$$\Psi(\theta) := 1 - \Big(2 - \frac{\alpha}{\theta}\Big)\ell.$$

Since L_α satisfies the ODE (9), at any $\theta > 0$ with $L_\alpha(\theta) = \ell$ we have

$$L'_\alpha(\theta) = \Psi(\theta). \tag{27}$$

A direct calculation shows $\Psi'(\theta) = -\frac{\alpha\ell}{\theta^2} < 0$, hence Ψ is strictly decreasing on $(0, \infty)$. Moreover, by the definition (26) of θ_ℓ we have

$$\Psi(\theta_\ell) = 1 - \Big(2 - \frac{\alpha}{\theta_\ell}\Big)\ell = 1 - \Big(2 - \Big(2 - \frac{1}{\ell}\Big)\Big)\ell = 1 - \frac{1}{\ell}\ell = 0.$$

Therefore,

$$\Psi(\theta) > 0 \ \text{ for } \ \theta < \theta_\ell, \qquad \Psi(\theta) = 0 \ \text{ for } \ \theta = \theta_\ell, \qquad \Psi(\theta) < 0 \ \text{ for } \ \theta > \theta_\ell. \tag{28}$$

Note that we cannot conclude that θ_ℓ is global maximizer because $L'_\alpha(\theta) = \Psi(\theta)$ only holds for $L_\alpha(\theta) = \ell$. However, this observation shows that every time when L_α passes the line $y = \ell$ before θ_ℓ, the function must be increasing, so it will not be able to go down and satisfy our assumption $L_\alpha(\theta_\ell) < \ell$. The detailed explanation is given below.

First, we claim that the first hit must occur strictly before θ_ℓ. Since $L_\alpha(\tau) = \ell$, (27) gives $L'_\alpha(\tau) = \Psi(\tau)$. Because $L'_\alpha(\tau) \ge 0$, we have $\Psi(\tau) \ge 0$. By (28), this implies $\tau \le \theta_\ell$. If $\tau = \theta_\ell$, then $L_\alpha(\theta_\ell) = \ell$, contradicting the hypothesis $L_\alpha(\theta_\ell) < \ell$. Hence

$$\tau < \theta_\ell. \tag{29}$$

By (28) again, $\Psi(\tau) > 0$, so $L'_\alpha(\tau) = \Psi(\tau) > 0$.

Next, we show that the first return produces a contradiction. Since $L_\alpha(\tau) = \ell$ and $L'_\alpha(\tau) > 0$, continuity implies there exists $\varepsilon > 0$ such that $L'_\alpha(\theta) > 0$ for all $\theta \in (\tau, \tau + \varepsilon)$, and therefore $L_\alpha(\theta) > \ell$ for all $\theta \in (\tau, \tau + \varepsilon)$. On the other hand, $L_\alpha(\theta_\ell) < \ell$ by hypothesis and $\tau < \theta_\ell$ by (29), so by continuity there exists at least one point in (τ, θ_ℓ) where L_α returns to the level ℓ. Define the first return time

$$\sigma := \inf\{\theta > \tau : \ L_\alpha(\theta) = \ell\}.$$

Then $\sigma \in (\tau, \theta_\ell)$, $L_\alpha(\sigma) = \ell$, and $L_\alpha(\theta) > \ell$ for all $\theta \in (\tau, \sigma)$. By Lemma 19, $L'_\alpha(\sigma) \le 0$.

But since $L_\alpha(\sigma) = \ell$, we have $L'_\alpha(\sigma) = \Psi(\sigma)$ by (27). Because $\sigma < \theta_\ell$, (28) gives $\Psi(\sigma) > 0$, hence $L'_\alpha(\sigma) > 0$. This contradicts $L'_\alpha(\sigma) \le 0$.

Therefore our assumption was false, and $L_\alpha(\theta) < \ell$ for all $\theta > 0$, i.e., $\sup_{\theta>0} L_\alpha(\theta) \le \ell$. ■

Lemma 29 (Series representation for L_α) *For $0 < \alpha < 1$ and $\theta > 0$,*

$$L_\alpha(\theta) = \theta e^{-2\theta} \sum_{n=0}^{\infty} \frac{(2\theta)^n}{n!\,(n+1-\alpha)}. \tag{30}$$

Proof With $w = 1 - u$, (7) becomes

$$L_\alpha(\theta) = \theta e^{-2\theta} \int_0^1 e^{2\theta w} w^{-\alpha}\, dw.$$

Expand $e^{2\theta w} = \sum_{n=0}^{\infty} \frac{(2\theta w)^n}{n!}$. For each N, the partial sums are nonnegative and increase pointwise to $e^{2\theta w}$, hence

$$e^{2\theta w} w^{-\alpha} = \lim_{N\to\infty} \sum_{n=0}^{N} \frac{(2\theta)^n}{n!}\, w^{n-\alpha} \quad \text{by monotone convergence.}$$

Apply monotone convergence to justify termwise integration:

$$\int_0^1 e^{2\theta w} w^{-\alpha}\, dw = \sum_{n=0}^{\infty} \frac{(2\theta)^n}{n!} \int_0^1 w^{n-\alpha}\, dw = \sum_{n=0}^{\infty} \frac{(2\theta)^n}{n!\,(n+1-\alpha)}.$$

Multiplying by $\theta e^{-2\theta}$ gives (30). ■

Corollary 30 (A particularly clean series at $\alpha = \frac{3}{4}$) *Let $\alpha_0 = \frac{3}{4}$. For $\theta > 0$,*

$$L_{\alpha_0}(\theta) = 4\theta e^{-2\theta} \sum_{n=0}^{\infty} \frac{(2\theta)^n}{n!\,(4n+1)}. \tag{31}$$

Proof Starting from (22),

$$L_{\alpha_0}(\theta) = 4\theta e^{-2\theta} \int_0^1 e^{2\theta y^4}\, dy.$$

Expand $e^{2\theta y^4} = \sum_{n=0}^{\infty} \frac{(2\theta)^n}{n!} y^{4n}$, and use monotone convergence (all terms nonnegative) to integrate termwise:

$$\int_0^1 e^{2\theta y^4}\, dy = \sum_{n=0}^{\infty} \frac{(2\theta)^n}{n!} \int_0^1 y^{4n}\, dy = \sum_{n=0}^{\infty} \frac{(2\theta)^n}{n!\,(4n+1)}.$$

Multiplying by $4\theta e^{-2\theta}$ gives (31). ■

Lemma 31 (Monotone ratios for the series coefficients) *Fix $\alpha \in (0,1)$ and $x > 0$, and define*

$$a_n := \frac{x^n}{n!\,(n+1-\alpha)} \qquad (n \geq 0).$$

Then the ratio $r_n := a_{n+1}/a_n$ is strictly decreasing for all integers $n \geq 1$.

Proof A direct computation gives

$$r_n = \frac{a_{n+1}}{a_n} = \frac{x}{n+1} \cdot \frac{n+1-\alpha}{n+2-\alpha}.$$

For $n \geq 1$,

$$\frac{r_{n+1}}{r_n} = \frac{n+1}{n+2} \cdot \frac{(n+2-\alpha)^2}{(n+1-\alpha)(n+3-\alpha)}.$$

Thus $r_{n+1} \leq r_n$ is equivalent to

$$(n+2)(n+1-\alpha)(n+3-\alpha) - (n+1)(n+2-\alpha)^2 \geq 0.$$

Expanding the left-hand side gives

$$n^2 + (3-2\alpha)n + (2-4\alpha+\alpha^2).$$

For $\alpha \in (0,1)$ this quadratic is strictly increasing for $n \geq 0$ and positive at $n = 1$:

$$1 + (3-2\alpha) + (2-4\alpha+\alpha^2) = \alpha^2 - 6\alpha + 6 \geq 1.$$

Hence it is positive for all $n \geq 1$, and therefore $r_{n+1} < r_n$ for all $n \geq 1$. ■

Appendix E. Exact Calculations for $L_{3/4}$ and $B_{\sup}(3/4)$

In this section we present the final calculations needed to complete the proof of the $\alpha = 3/4$ version of Lemma 11.

Lemma 32 (One-point bound for L_{α_0}) *With $\alpha_0 = \frac{3}{4}$ and $\ell_0 = \frac{497}{500}$,*

$$L_{\alpha_0}\left(\theta_{\ell_0}\right) < \ell_0.$$

Proof

(i) Rewrite $L_{\alpha_0}(\theta_{\ell_0})$ as a positive series

Let

$$\theta := \frac{1491}{1976}, \qquad x := 2\theta = \frac{1491}{988}.$$

By Corollary 30,

$$L_{\alpha_0}(\theta) = 4\theta e^{-x} \sum_{n=0}^{\infty} \frac{x^n}{n!\,(4n+1)}. \tag{32}$$

Define the positive series

$$S(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!\,(4n+1)}.$$

(ii) An explicit rational upper bound for e^{-x}

Since $e^x = \sum_{k=0}^{\infty} x^k/k!$ has positive terms for $x \geq 0$, for any integer $m \geq 0$,

$$e^x \geq P_m(x) := \sum_{k=0}^{m} \frac{x^k}{k!}.$$

Therefore

$$e^{-x} \leq \frac{1}{P_m(x)}. \tag{33}$$

(iii) Rational truncation and tail bound for $S(x)$

Fix $N \geq 0$ and write

$$S_N(x) := \sum_{n=0}^{N} \frac{x^n}{n!\,(4n+1)}.$$

Since $n \mapsto (4n+1)^{-1}$ is decreasing, for $n \geq N+1$,

$$\frac{1}{4n+1} \leq \frac{1}{4(N+1)+1} = \frac{1}{4N+5}.$$

Hence

$$S(x) \leq S_N(x) + \frac{1}{4N+5} \sum_{n=N+1}^{\infty} \frac{x^n}{n!}.$$

Assume $x < N+2$. For $n \geq N+1$,

$$\frac{x^{n+1}/(n+1)!}{x^n/n!} = \frac{x}{n+1} \leq \frac{x}{N+2} < 1,$$

so the exponential tail is bounded by a geometric series:

$$\sum_{n=N+1}^{\infty} \frac{x^n}{n!} \leq \frac{x^{N+1}}{(N+1)!} \cdot \frac{1}{1 - \frac{x}{N+2}}. \tag{34}$$

Combining,

$$S(x) \leq S_N(x) + \frac{1}{4N+5} \cdot \frac{x^{N+1}}{(N+1)!} \cdot \frac{1}{1 - \frac{x}{N+2}}. \tag{35}$$

(iv) Choose (N, m) and conclude

Combining (32), (33), and (35) yields

$$L_{\alpha_0}(\theta) \le 4\theta \cdot \frac{1}{P_m(x)} \cdot \left[S_N(x) + \frac{1}{4N+5} \cdot \frac{x^{N+1}}{(N+1)!} \cdot \frac{1}{1 - \frac{x}{N+2}} \right].$$

Take $N = 4$ and $m = 8$. Since $x = \frac{1491}{988} < 6 = N + 2$, the condition $x < N + 2$ holds.

Evaluating the right-hand side in exact rational arithmetic gives

$$Q_{4,8} := 4\theta \cdot \frac{1}{P_8(x)} \cdot \left[S_4(x) + \frac{1}{21} \cdot \frac{x^5}{5!} \cdot \frac{1}{1 - \frac{x}{6}} \right] = \frac{1287675113562193776446577567744}{1295590272667287121985809656211}.$$

Moreover,

$$\ell_0 - Q_{4,8} = \frac{7080873454481140365861526 4867}{64779513633364356099290482 8105500} > 0.$$

Hence $Q_{4,8} < \ell_0$, and therefore

$$L_{\alpha_0}(\theta_{\ell_0}) \le Q_{4,8} < \ell_0,$$

proving Lemma 32. ■

Lemma 33 (Bound for $B_{\sup}$) *With $\alpha_0 = \frac{3}{4}$, we have $B_{\sup}(\alpha_0) \le \frac{201}{200} \cdot \frac{497}{500} = \frac{99897}{100000} < 1$.*

Proof Fix $t \ge 300$ and $\rho \in (0, \frac{1}{2}]$. Write

$$(t+2)^{\alpha_0} B_t = \left(1 + \frac{2}{t}\right)^{\alpha_0} t^{\alpha_0} B_t.$$

By Lemma 15 with $u = \frac{2}{t}$,

$$\left(1 + \frac{2}{t}\right)^{\alpha_0} \le 1 + \alpha_0 \cdot \frac{2}{t} \le 1 + \alpha_0 \cdot \frac{2}{300} = 1 + \frac{2\alpha_0}{300}.$$

Using Lemma 32 and Lemma 28, we have $t^{\alpha_0} B_t \le \ell_0$, hence

$$(t+2)^{\alpha_0} B_t \le \left(1 + \frac{2\alpha_0}{300}\right)\ell_0 = \frac{201}{200} \cdot \frac{497}{500} = \frac{99897}{100000} < 1.$$

■

Appendix F. Remaining Details for the Proof of Lemma 11 with $\alpha = 3/4 + 0.001$

We will show that there exists an absolute constant $C > 0$ such that for all $t \ge 1000$, $K \ge C$, $\rho \in (0, 1/2]$, and $\alpha = 3/4 + 0.001$, we have:

$$\rho(1-2\rho)^t + K\rho \sum_{i=1}^{t} \frac{(1-2\rho)^{t-i}}{i^\alpha} \le \frac{K}{(t+2)^\alpha}.$$

and then we will show that we can also choose $C = 1000$.

In this section, we define

$$A_{\sup}(\alpha) := \sup_{t\geq 1000} \sup_{\rho\in(0,\frac{1}{2}]} (t+2)^{\alpha} A_t(\rho), \qquad B_{\sup}(\alpha) := \sup_{t\geq 1000} \sup_{\rho\in(0,\frac{1}{2}]} (t+2)^{\alpha} B_t(\rho,\alpha). \quad (36)$$

The first term $A_{\sup}(\alpha)$ is easier to bound as before.

Lemma 34 *For every $\alpha \in (0,1]$, the function $t \mapsto f_\alpha(t)$ defined as*

$$f_\alpha(t) = \frac{(t+2)^\alpha}{2(t+1)}\Big(\frac{t}{t+1}\Big)^t$$

is strictly decreasing for all real $t > 0$. Consequently, we have $A_{\sup}(\alpha_0) < \frac{15251}{334150} < 0.046$.

Proof For real $t > 0$ write

$$\log f_\alpha(t) = \alpha \log(t+2) - \log\big(2(t+1)\big) + t\log\Big(\frac{t}{t+1}\Big).$$

Differentiating gives

$$\frac{d}{dt}\log f_\alpha(t) = \frac{\alpha}{t+2} - \log\Big(1+\frac{1}{t}\Big).$$

Using see Lemma 17 for $t > 0$, we have

$$\frac{d}{dt}\log f_\alpha(t) \leq \frac{\alpha}{t+2} - \frac{1}{t+1} \leq \frac{1}{t+2} - \frac{1}{t+1} < 0,$$

since $\alpha \leq 1$. Therefore $\log f_\alpha$ is strictly decreasing, and so is f_α. By Lemma 20, we have

$$\sup_{\rho\in(0,\frac{1}{2}]} (t+2)^{\alpha_0} A_t(\rho) = \frac{(t+2)^{\alpha_0}}{2(t+1)}\Big(\frac{t}{t+1}\Big)^t.$$

Therefore, we have $A_{\sup}(\alpha_0) < f_{\alpha_0}(1000) < f_{\alpha_0}(300)$. By Lemma 35, we have

$$A_{\sup}(\alpha_0) < f_{\alpha_0}(300) < \frac{15251}{334150} < 0.046$$

■

Lemma 35 *With $\alpha_0 = \frac{751}{1000}$,*

$$f_{\alpha_0}(300) \;<\; \frac{15251}{334150} \;<\; 0.046.$$

Proof By Lemma 20,

$$f_{\alpha_0}(300) = \frac{302^{751/1000}}{2\cdot 301}\Big(\frac{300}{301}\Big)^{300}.$$

Since $751/1000 = 1 - 249/1000$,

$$302^{751/1000} = 302\cdot 302^{-249/1000} \leq 302\cdot 301^{-249/1000},$$

because $x \mapsto x^{-249/1000}$ is decreasing and $302 > 301$. Hence

$$f_{\alpha_0}(300) \le \frac{151}{301} \cdot 301^{-249/1000} \cdot \Big(\frac{300}{301}\Big)^{300}. \tag{37}$$

(i) Bounding $301^{-249/1000}$. Write $301^{249/1000} = 301^{1/4-1/1000} = 301^{1/4}/301^{1/1000}$. We bound numerator and denominator separately. Since $(41/10)^4 = \frac{2825761}{10000} < 301$, we have $301^{1/4} > 41/10$. Also,

$$\Big(\frac{101}{100}\Big)^{100} = \sum_{k=0}^{100} \binom{100}{k} \frac{1}{100^k} > 1 + \binom{100}{1}\frac{1}{100} = 2,$$

so $(101/100)^{1000} > 2^{10} = 1024 > 301$, hence $301^{1/1000} < 101/100$. Therefore

$$301^{249/1000} = \frac{301^{1/4}}{301^{1/1000}} > \frac{41/10}{101/100} = \frac{410}{101}, \qquad \text{so} \qquad 301^{-249/1000} < \frac{101}{410}.$$

(ii) Bounding $(300/301)^{300}$. The standard inequality $(1-\frac{1}{n})^n < e^{-1}$ (for $n \ge 1$) gives

$$\Big(\frac{300}{301}\Big)^{301} < \frac{1}{e}.$$

Thus

$$\Big(\frac{300}{301}\Big)^{300} = \frac{301}{300}\Big(\frac{300}{301}\Big)^{301} < \frac{301}{300} \cdot \frac{1}{e}.$$

Using the 6-term lower bound $e > 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} = \frac{163}{60}$ yields $1/e < 60/163$, hence

$$\Big(\frac{300}{301}\Big)^{300} < \frac{301}{300} \cdot \frac{60}{163} = \frac{301}{815}.$$

(iii) Conclusion. Insert the bounds from (i)–(ii) into (37):

$$f_{\alpha_0}(300) < \frac{151}{301} \cdot \frac{101}{410} \cdot \frac{301}{815} = \frac{151 \cdot 101}{410 \cdot 815} = \frac{15251}{334150} < 0.046.$$

Finally, since $t_0 = 1000 > 300$, Lemma 34 implies $A_{\sup}(\alpha_0) = \sup_{t \ge 1000} f_{\alpha_0}(t) \le f_{\alpha_0}(300)$. ■

To bound the second term, we compare it with the function

$$L_\alpha(\theta) := \theta \int_0^1 e^{-2\theta u}(1-u)^{-\alpha}\, du. \tag{38}$$

as before.

Lemma 36 (One-point inequality at $\ell = \frac{499}{500}$) *Let $\alpha_0 = \frac{751}{1000}$ and $\ell = \frac{499}{500}$, with*

$$\theta_\ell = \frac{\alpha_0}{2 - \frac{1}{\ell}} = \frac{751/1000}{2 - 500/499} = \frac{374749}{498000}. \tag{39}$$

Then $L_{\alpha_0}(\theta_\ell) < \ell$.

Proof Set $x := 2\theta_\ell$. By (39),

$$x = \frac{374749}{249000} = \frac{301}{200} + \frac{1}{62250}.$$

In particular $x > \frac{301}{200}$. Also $x < \frac{753}{500}$ since $\frac{753}{500} - \frac{374749}{249000} = \frac{49}{49800} > 0$. Let

$$x_+ := \frac{753}{500}, \qquad \theta_+ := \frac{x_+}{2} = \frac{753}{1000}.$$

Then $\theta_\ell < \theta_+$, $e^{-x} < e^{-301/200}$, and by (30),

$$L_{\alpha_0}(\theta_\ell) = \theta_\ell e^{-x} \sum_{n=0}^{\infty} \frac{x^n}{n!\,(n+1-\alpha_0)} < \theta_+\, e^{-301/200} \sum_{n=0}^{\infty} \frac{x_+^n}{n!\,(n+1-\alpha_0)}. \tag{40}$$

Define

$$a_n := \frac{x_+^n}{n!\,(n+1-\alpha_0)} \qquad (n \geq 0), \qquad S(x_+) := \sum_{n=0}^{\infty} a_n.$$

(i) A rational bound on $S(x_+)$. A direct calculation gives the first five terms

$$a_0 = \frac{1000}{249}, \quad a_1 = \frac{1506}{1249}, \quad a_2 = \frac{567009}{1124500}, \quad a_3 = \frac{15813251}{90250000}, \quad a_4 = \frac{107166402027}{2124500000000}.$$

Moreover,

$$r_4 = \frac{a_5}{a_4} = \frac{x_+}{5} \cdot \frac{5-\alpha_0}{6-\alpha_0} = \frac{3199497}{13122500} < \frac{1}{4},$$

since $4 \cdot 3199497 = 12797988 < 13122500$. By Lemma 31, $r_n \leq r_4$ for all $n \geq 4$, hence

$$\sum_{n\geq 4} a_n \leq a_4 \sum_{j\geq 0} r_4^j \leq a_4 \sum_{j\geq 0} \Big(\frac{1}{4}\Big)^j = \frac{4}{3} a_4.$$

Therefore

$$S(x_+) \leq a_0 + a_1 + a_2 + a_3 + \frac{4}{3}a_4 = \frac{800429153543344037119501}{134108154748420125000000}. \tag{41}$$

(ii) A rational bound on $e^{-301/200}$. Let $P_8(y) := \sum_{j=0}^{8} \frac{y^j}{j!}$. Since $e^y = P_8(y) + \sum_{j\geq 9} \frac{y^j}{j!} > P_8(y)$ for $y > 0$, we have $e^{-y} < 1/P_8(y)$. For $y = \frac{301}{200}$ one computes

$$P_8\Big(\frac{301}{200}\Big) = \frac{66414558043759180589143}{14745600000000000000000}, \qquad \text{so} \qquad e^{-301/200} < \frac{14745600000000000000000}{66414558043759180589143}. \tag{42}$$

(iii) Conclude $L_{\alpha_0}(\theta_\ell) < \ell$. Combining (40), (41), and (42) yields the explicit bound

$$\begin{aligned} &L_{\alpha_0}(\theta_\ell) \\ <&\frac{753}{1000} \cdot \frac{14745600000000000000000}{66414558043759180589143} \cdot \frac{800429153543344037119501}{134108154748420125000000} \\ =&\frac{237000387179893775381686224027648000000}{237512902071476980327802708758559405 41}. \end{aligned}$$

To show this is $< \ell = \frac{499}{500}$, it suffices to check the strict integer inequality

$$\begin{aligned}&499 \cdot 237512902071476980327802708758559405 41\\ &- 500 \cdot 237000387179893775381686224027648000 00\\ =&187445437201254927304396566971432995 9 > 0,\end{aligned}$$

which completes the proof. ∎

Lemma 37 (A bound on $B_{\sup}(\alpha_0)$) *With $\alpha_0 = \frac{751}{1000}$ and $\ell = \frac{499}{500}$,*

$$B_{\sup}(\alpha_0) \le \frac{249874749}{250000000} \qquad \textit{and hence} \qquad 1 - B_{\sup}(\alpha_0) \ge \frac{125251}{250000000}.$$

Proof Fix $t \ge 1000$ and $\rho \in (0, \frac{1}{2}]$. If $\rho \in (0, \frac{1}{2})$, Lemma 25 and Lemma 28 give $t^{\alpha_0} B_t(\rho, \alpha_0) \le \ell$. If $\rho = \frac{1}{2}$, Lemma 26 gives $t^{\alpha_0} B_t(\frac{1}{2}, \alpha_0) = \frac{1}{2} \le \ell$.

Thus, for all $\rho \in (0, \frac{1}{2}]$,

$$(t+2)^{\alpha_0} B_t(\rho, \alpha_0) = \left(1 + \frac{2}{t}\right)^{\alpha_0} t^{\alpha_0} B_t(\rho, \alpha_0) \le \left(1 + \frac{2}{t}\right)^{\alpha_0} \ell.$$

Since $\alpha_0 \in (0, 1)$, Bernoulli's inequality gives $(1+u)^{\alpha_0} \le 1 + \alpha_0 u$ for $u \ge 0$. With $u = 2/t$ and $t \ge 1000$,

$$\left(1 + \frac{2}{t}\right)^{\alpha_0} \le 1 + \frac{2\alpha_0}{t} \le 1 + \frac{2\alpha_0}{1000} = 1 + \frac{751}{500000} = \frac{500751}{500000}.$$

Therefore

$$(t+2)^{\alpha_0} B_t(\rho, \alpha_0) \le \frac{500751}{500000} \cdot \frac{499}{500} = \frac{249874749}{250000000}.$$

Taking the supremum over $t \ge 1000$ and $\rho \in (0, \frac{1}{2}]$ yields the stated bound on $B_{\sup}(\alpha_0)$. The bound on $1 - B_{\sup}(\alpha_0)$ is immediate. ∎

Lemma 38 (Requirement for K) *With the bounds from Lemma 35 and Lemma 37,*

$$\frac{A_{\sup}(\alpha_0)}{1 - B_{\sup}(\alpha_0)} < \frac{76255000000}{837052433} < 100 < 1000.$$

Proof Using $A_{\sup}(\alpha_0) < \frac{15251}{334150}$ and $1 - B_{\sup}(\alpha_0) \ge \frac{125251}{250000000}$,

$$\frac{A_{\sup}(\alpha_0)}{1 - B_{\sup}(\alpha_0)} < \frac{\frac{15251}{334150}}{\frac{125251}{250000000}} = \frac{76255000000}{837052433}.$$

Finally,

$$\frac{76255000000}{837052433} < 100 \iff 76255000000 < 100 \cdot 837052433 = 83705243300,$$

which is immediate. ∎

Appendix G. Lower Bound for $\|\mathbf{N}_t\|$

In this section, we give an upper bound for the parameter α that is admissible to make the inequality in Lemma 11 hold and then give a lower bound for $\|\mathbf{N}_t\|$.

Theorem 39 (Upper bound on the admissible exponent) *Let*

$$\alpha_0 := \frac{753}{1000} = 0.753.$$

Fix any $\alpha \geq \alpha_0$. Then for every $K > 0$ and every integer $t_0 \geq 1$, there exist an integer $t \geq t_0$ and a value $\rho \in (0, \frac{1}{2})$ such that

$$\rho(1-2\rho)^t + K\rho \sum_{i=1}^{t} \frac{(1-2\rho)^{t-i}}{i^\alpha} \leq \frac{K}{(t+2)^\alpha}. \tag{43}$$

fails, i.e.,

$$\rho(1-2\rho)^t \ + K\rho \sum_{i=1}^{t} \frac{(1-2\rho)^{t-i}}{i^\alpha} \ > \frac{K}{(t+2)^\alpha}.$$

In particular, if there exist constants $K > 0$ and $T \geq 1$ such that (43) holds for all integers $t \geq T$ and all $\rho \in (0, \frac{1}{2}]$, then necessarily

$$\alpha < \frac{753}{1000} = 0.753.$$

To prove Theorem 39, we make the following preparations.

Recall that

$$A_t(\rho) := \rho(1-2\rho)^t \quad \text{and} \quad B_t(\rho, \alpha) := \rho \sum_{i=1}^{t} \frac{(1-2\rho)^{t-i}}{i^\alpha}.$$

and then the inequality is equivalent to

$$(t+2)^\alpha A_t(\rho) + K(t+2)^\alpha B_t(\rho, \alpha) \leq K.$$

The first reduction is to observe that in order to make the inequality (43) fail, it suffices to show that $(t+2)^\alpha B_t(\rho, \alpha)$ is large.

Lemma 40 (Reducing to the second term) *For every $t \in \mathbb{N}$, $\rho \in (0, \frac{1}{2}]$, and $\alpha > 0$, one has $A_t(\rho) \geq 0$. Consequently, if (43) holds for some $t \in \mathbb{N}$, $\rho \in (0, \frac{1}{2}]$, $K > 0$, and $\alpha > 0$, then necessarily*

$$(t+2)^\alpha B_t(\rho, \alpha) \leq 1. \tag{44}$$

Proof Since $\rho \in (0, \frac{1}{2}]$, we have $1 - 2\rho \in [0, 1)$, hence $(1-2\rho)^t \geq 0$ and therefore $A_t(\rho) = \rho(1-2\rho)^t \geq 0$. If (43) holds, then subtracting the nonnegative term $A_t(\rho)$ from the left-hand side gives

$$K(t+2)^\alpha B_t(\rho, \alpha) \leq K.$$

Dividing by $K > 0$ gives the desired result. ■

The next reduction is to observe that if α_0 makes the inequality (43) fail, then any $\alpha > \alpha_0$ will also make the inequality fail, because of the monotonicity below.

Lemma 41 (Monotonicity of $t^\alpha B_t(\rho,\alpha)$ in α) *Fix $t \in \mathbb{N}$ and $\rho \in (0, \frac{1}{2}]$. Then the map*

$$\alpha \longmapsto t^\alpha B_t(\rho,\alpha)$$

is nondecreasing on $[0,\infty)$. Equivalently, if $0 \le \alpha_1 \le \alpha_2$, then

$$t^{\alpha_1} B_t(\rho,\alpha_1) \le t^{\alpha_2} B_t(\rho,\alpha_2). \tag{45}$$

Proof Write $r := 1 - 2\rho \in [0,1)$. Then

$$t^\alpha B_t(\rho,\alpha) = \rho \sum_{i=1}^{t} r^{t-i} \Big(\frac{t}{i}\Big)^\alpha.$$

For each $i \in \{1,\dots,t\}$ we have $t/i \ge 1$, so the function $\alpha \mapsto (t/i)^\alpha$ is nondecreasing on $[0,\infty)$. Every coefficient ρr^{t-i} is nonnegative, hence the whole sum is nondecreasing in α. This proves (45). ∎

For $\alpha \in (0,1)$ and $\theta > 0$, recall that in section 4 we define

$$L_\alpha(\theta) = \theta \int_0^1 e^{-2\theta u}(1-u)^{-\alpha}\, du. \tag{46}$$

to get upper bound for the quantity $t^\alpha B_t(\rho,\alpha)$.

Now to lower bound $t^\alpha B_t(\rho,\alpha)$, for $\alpha \in (0,1)$ and $\theta > 0$, we define a similar quantity

$$\Gamma_{\alpha,t}(\theta) = \theta \int_0^1 e^{-2\theta u}\Big(1 - u + \frac{1}{t}\Big)^{-\alpha} du. \tag{47}$$

Unlike $L_\alpha(\theta)$, the integral $\Gamma_{\alpha,t}(\theta)$ is a proper integral.

Lemma 42 (Discrete-to-continuous lower bound) *Fix $\alpha \in (0,1)$, $t \in \mathbb{N}$, and $\rho \in (0, \frac{1}{2})$. Then*

$$t^\alpha B_t(\rho,\alpha) \ge \Gamma_{\alpha,t}\Big(\frac{t}{2}\big(-\log(1-2\rho)\big)\Big). \tag{48}$$

Proof As in the proof of Lemma 25, let $r = 1-2\rho \in (0,1), \delta = -\log r > 0$, and $\theta = \frac{\delta t}{2} = \frac{t}{2}\big(-\log(1-2\rho)\big)$, and write $B_t(\rho,\alpha)$ in the integral form

$$\begin{aligned} B_t(\rho,\alpha) &= \frac{1}{2}\sum_{k=0}^{t-1} \frac{1}{(t-k)^\alpha} \int_k^{k+1} \delta e^{-\delta x}\, dx \\ &= \frac{1}{2}\sum_{k=0}^{t-1} \int_k^{k+1} \delta e^{-\delta x}(t-k)^{-\alpha}\, dx. \end{aligned} \tag{49}$$

Since we want a lower bound for $t^\alpha B_t(\rho,\alpha)$, we want to lower bound the factor $(t-k)^{-\alpha}$ to make it not depend on k. Fix $k \in \{0,1,\dots,t-1\}$ and $x \in [k,k+1]$. Then we have

$$t - x + 1 \ge t - (k+1) + 1 = t - k.$$

Because $\alpha > 0$, the function $y \mapsto y^{-\alpha}$ is decreasing on $(0, \infty)$, so

$$(t-k)^{-\alpha} \geq (t-x+1)^{-\alpha}.$$

Multiplying by the nonnegative factor $\delta e^{-\delta x}$ and integrating over $[k, k+1]$ gives

$$\int_k^{k+1} \delta e^{-\delta x}(t-k)^{-\alpha}\,dx \geq \int_k^{k+1} \delta e^{-\delta x}(t-x+1)^{-\alpha}\,dx.$$

Summing over k and using (49), we obtain

$$\begin{aligned} B_t(\rho,\alpha) &\geq \frac{1}{2}\sum_{k=0}^{t-1}\int_k^{k+1} \delta e^{-\delta x}(t-x+1)^{-\alpha}\,dx \\ &= \frac{1}{2}\int_0^t \delta e^{-\delta x}(t-x+1)^{-\alpha}\,dx. \end{aligned} \tag{50}$$

Set $x = tu$, so $u \in [0,1]$ and $dx = t\,du$. Then (50) becomes

$$\begin{aligned} B_t(\rho,\alpha) &\geq \frac{1}{2}\int_0^1 \delta e^{-\delta tu}\big(t - tu + 1\big)^{-\alpha} t\,du \\ &= \frac{1}{2}\,\delta\,t\int_0^1 e^{-\delta tu}\Big(t\Big(1-u+\frac{1}{t}\Big)\Big)^{-\alpha}\,du \\ &= \frac{1}{2}\,\delta\,t^{1-\alpha}\int_0^1 e^{-\delta tu}\Big(1-u+\frac{1}{t}\Big)^{-\alpha}\,du. \end{aligned} \tag{51}$$

Multiplying (51) by t^α gives

$$t^\alpha B_t(\rho,\alpha) \geq \frac{1}{2}\,\delta\,t\int_0^1 e^{-\delta tu}\Big(1-u+\frac{1}{t}\Big)^{-\alpha}\,du.$$

Now $\delta t = 2\theta$, so $e^{-\delta tu} = e^{-2\theta u}$ and $\frac{1}{2}\delta t = \theta$. Hence

$$t^\alpha B_t(\rho,\alpha) \geq \theta\int_0^1 e^{-2\theta u}\Big(1-u+\frac{1}{t}\Big)^{-\alpha}\,du = \Gamma_{\alpha,t}(\theta) = \Gamma_{\alpha,t}\Big(\frac{t}{2}\big(-\log(1-2\rho)\big)\Big),$$

which is exactly (48). ∎

Although to lower bound $t^\alpha B_t(\rho,\alpha)$ we use $\Gamma_{\alpha,t}(\theta)$ instead of $L_\alpha(\theta)$, the following convergence result allows us to transform from $\Gamma_{\alpha,t}(\theta)$ to $L_\alpha(\theta)$.

Lemma 43 (Monotone convergence) *Fix $\alpha \in (0,1)$ and $\theta > 0$. Then the sequence $t \mapsto \Gamma_{\alpha,t}(\theta)$ is nondecreasing and*

$$\lim_{t\to\infty} \Gamma_{\alpha,t}(\theta) = L_\alpha(\theta). \tag{52}$$

Proof Fix $u \in [0,1)$. As t increases, the quantity $1-u+\frac{1}{t}$ decreases. Since $\alpha > 0$, the function $x \mapsto x^{-\alpha}$ is decreasing on $(0,\infty)$, so $\Big(1-u+\frac{1}{t}\Big)^{-\alpha}$ is increasing in t and

$$\lim_{t\to\infty}\Big(1-u+\frac{1}{t}\Big)^{-\alpha} = (1-u)^{-\alpha}$$

Multiplying by the nonnegative factor $e^{-2\theta u}$ preserves monotonicity, so by the Monotone Convergence Theorem, we have

$$\lim_{t\to\infty}\int_0^1 e^{-2\theta u}\Big(1-u+\frac{1}{t}\Big)^{-\alpha}\,du=\int_0^1 e^{-2\theta u}(1-u)^{-\alpha}\,du.$$

Multiplying by $\theta>0$ gives (52). ■

Combining above reductions, to show that there exist an integer $t\geq t_0$ and a value $\rho\in(0,\frac{1}{2})$ such that (43) fails with $\alpha=\alpha_0$, we only need to show $L_{\alpha_0}(\theta_0)>1$ for some properly chosen θ_0. To this end, we choose $\theta_0=\frac{3}{4}$ and use the series approximation

$$L_\alpha(\theta)=\theta e^{-2\theta}\sum_{n=0}^{\infty}\frac{(2\theta)^n}{n!\,(n+1-\alpha)}.$$

from Lemma 29 to give explicit lower bound for $L_{\alpha_0}(\theta_0)$.

Lemma 44 (Lower bound for L) *With $\alpha_0=\frac{753}{1000}$ and $\theta_0=\frac{3}{4}$, one has*

$$L_{\alpha_0}(\theta_0)>1.$$

Proof by Lemma 29, we have

$$L_{\alpha_0}(\theta_0)=\theta_0 e^{-2\theta_0}\sum_{n=0}^{\infty}\frac{(2\theta_0)^n}{n!\,(n+1-\alpha_0)}=\theta_0 e^{-\frac{3}{2}}S(x_0).$$

where $x_0=2\theta_0=\frac{3}{2}$ and $S(x_0)=\sum_{n=0}^{\infty}\frac{x_0^n}{n!\,(n+1-\alpha_0)}=\sum_{n=0}^{\infty}\frac{(2\theta_0)^n}{n!\,(n+1-\alpha_0)}$.

We first estimate $e^{-3/2}$ by Taylor expansion. Apply Taylor's theorem with Lagrange remainder to $f(x)=e^{-x}$ at $x=\frac{3}{2}$, expanded at 0, through degree 9. There exists $\xi\in(0,\frac{3}{2})$ such that

$$e^{-3/2}=\sum_{k=0}^{9}\frac{(-\frac{3}{2})^k}{k!}+\frac{e^{-\xi}}{10!}\Big(\frac{3}{2}\Big)^{10}. \tag{53}$$

The remainder is strictly positive, so

$$e^{-3/2}>\sum_{k=0}^{9}\frac{(-\frac{3}{2})^k}{k!}. \tag{54}$$

A direct rational computation gives

$$\sum_{k=0}^{9}\frac{(-\frac{3}{2})^k}{k!}=\frac{511775}{2293760}=\frac{102355}{458752}.$$

Therefore, we have

$$e^{-3/2}>\frac{102355}{458752}.$$

Next, we bound $S(x_0)$. All terms in the series are positive, so it is enough to keep the first six terms and conclude

$$S(x_0) > \sum_{n=0}^{5} \frac{x_0^n}{n!\,(n+1-\alpha_0)}. \tag{55}$$

Since $\alpha_0 = \frac{753}{1000}$ and $x_0 = \frac{3}{2}$, the first six terms are

$$\begin{aligned}
\frac{x_0^0}{0!(1-\alpha_0)} &= \frac{1000}{247},\\
\frac{x_0^1}{1!(2-\alpha_0)} &= \frac{1500}{1247},\\
\frac{x_0^2}{2!(3-\alpha_0)} &= \frac{375}{749},\\
\frac{x_0^3}{3!(4-\alpha_0)} &= \frac{1125}{6494},\\
\frac{x_0^4}{4!(5-\alpha_0)} &= \frac{3375}{67952},\\
\frac{x_0^5}{5!(6-\alpha_0)} &= \frac{225}{18656}.
\end{aligned}$$

We now use the following six strict rational lower bounds:

$$\begin{aligned}
\frac{1000}{247} &> \frac{2024}{500},\\
\frac{1500}{1247} &> \frac{601}{500},\\
\frac{375}{749} &> \frac{500}{999},\\
\frac{1125}{6494} &> \frac{433}{2500},\\
\frac{3375}{67952} &> \frac{149}{3000},\\
\frac{225}{18656} &> \frac{301}{25000}.
\end{aligned}$$

Each is verified by cross-multiplication:

$$\begin{aligned}
1000 \cdot 500 = 500000 &> 499928 = 2024 \cdot 247,\\
1500 \cdot 500 = 750000 &> 749447 = 601 \cdot 1247,\\
375 \cdot 999 = 374625 &> 374500 = 500 \cdot 749,\\
1125 \cdot 2500 = 2812500 &> 2811902 = 433 \cdot 6494,\\
3375 \cdot 3000 = 10125000 &> 10124848 = 149 \cdot 67952,\\
225 \cdot 25000 = 5625000 &> 5615456 = 301 \cdot 18656.
\end{aligned}$$

Hence

$$S(x_0) > \frac{2024}{500} + \frac{601}{500} + \frac{500}{999} + \frac{433}{2500} + \frac{149}{3000} + \frac{301}{25000}. \tag{56}$$

The right-hand side simplifies exactly to

$$\frac{18685693}{3121875}.$$

Finally,

$$\frac{18685693}{3121875} > \frac{299}{50} \iff 18685693 \cdot 50 > 3121875 \cdot 299,$$

and indeed

$$934284650 > 933440625.$$

Therefore $S(x_0) > \frac{299}{50}$.

Since $\theta_0 = \frac{3}{4} > 0$, combining all the estimates above, we obtain

$$L_{\alpha_0}(\theta_0) > \frac{3}{4} \cdot \frac{102355}{458752} \cdot \frac{299}{50} = \frac{91812435}{91750400} > 1.$$

This proves the lemma. ■

Proof [Proof of Theorem 39]

Let $\alpha_0 = \frac{753}{1000}$ and $\theta_0 = \frac{3}{4}$. By Lemma 44, we have

$$L_{\alpha_0}(\theta_0) > 1.$$

Now, consider arbitrary $K > 0$ and $t_0 > 0$. By Lemma 43, the sequence $\Gamma_{\alpha_0,t}(\theta_0)$ is nondecreasing in t and converges to $L_{\alpha_0}(\theta_0)$. Since the limit is greater than 1, there exists an integer $t \geq t_0$ such that

$$\Gamma_{\alpha_0,t}(\theta_0) > 1. \tag{57}$$

Now we set

$$\rho = \rho_t(\theta_0) = \frac{1 - e^{-2\theta_0/t}}{2} \in (0, \tfrac{1}{2}).$$

By Lemma 42, we have

$$t^{\alpha_0} B_t(\rho, \alpha_0) \geq \Gamma_{\alpha_0,t}(\theta_0) > 1.$$

Therefore

$$B_t(\rho, \alpha_0) > \frac{1}{t^{\alpha_0}}. \tag{58}$$

By monotonicity (Lemma 41), we also have

$$t^{\alpha} B_t(\rho, \alpha) \geq t^{\alpha_0} B_t(\rho, \alpha_0) > 1.$$

Hence

$$B_t(\rho, \alpha) > \frac{1}{t^{\alpha}}. \tag{59}$$

Since $t + 2 > t$, we have $(t+2)^{\alpha} > t^{\alpha}$, and therefore

$$\frac{1}{t^{\alpha}} > \frac{1}{(t+2)^{\alpha}}.$$

Combining this with (59), we get

$$B_t(\rho, \alpha) > \frac{1}{(t+2)^\alpha}.$$

Therefore, the necessary condition in Lemma 40 cannot hold.

■

To prove the lower bound for the matrix recursion $\mathbf{N}_t$, we first derive the following basic properties.

Lemma 45 (Eigenvalue recursion) *Let*

$$\bar{\mathbf{M}} = \mathbf{U}\mathbf{D}\mathbf{U}^\top, \qquad \mathbf{D} = \operatorname{diag}(\rho_1, \dots, \rho_n), \qquad 1 \geq \rho_1 \geq \cdots \geq \rho_n \geq 0.$$

Define the following matrix recursion:

$$\mathbf{N}_0 = \bar{\mathbf{M}}, \qquad \mathbf{N}_{t+1} = \mathbf{N}_t(\mathbf{I} - 2\bar{\mathbf{M}}) + \|\mathbf{N}_t\| \cdot \bar{\mathbf{M}}. \tag{60}$$

Then the followings hold:

(i) Every $\mathbf{N}_t$ is of the form

$$\mathbf{N}_t = \mathbf{U} \operatorname{diag}(\lambda_{1,t}, \dots, \lambda_{n,t})\mathbf{U}^\top,$$

with $\lambda_{k,0} = \rho_k$ and

$$\lambda_{k,t+1} = (1 - 2\rho_k)\lambda_{k,t} + \rho_k \mu_t, \qquad \textit{where} \qquad \mu_t = \max_{1 \leq i \leq n} \lambda_{i,t} = \|\mathbf{N}_t\|. \tag{61}$$

(ii) $\mathbf{N}_t \succeq 0$ for all $t \geq 0$.

(iii) For every $t \geq 0$,

$$\mathbf{N}_t \succeq \bar{\mathbf{M}}(\mathbf{I} - \bar{\mathbf{M}})^t. \tag{62}$$

Consequently,

$$\|\mathbf{N}_t\| \geq \max_{1 \leq k \leq n} \rho_k (1 - \rho_k)^t. \tag{63}$$

Proof We first prove (i). Because $\mathbf{N}_0 = \bar{\mathbf{M}}$ is a polynomial in $\bar{\mathbf{M}}$, the claim is true at $t = 0$. Assume $\mathbf{N}_t = p_t(\bar{\mathbf{M}})$ for some real polynomial p_t. Then

$$\mathbf{N}_{t+1} = p_t(\bar{\mathbf{M}})(\mathbf{I} - 2\bar{\mathbf{M}}) + \|\mathbf{N}_t\| \, \bar{\mathbf{M}},$$

which is again a polynomial in $\bar{\mathbf{M}}$. Thus, by induction, every $\mathbf{N}_t$ is a polynomial in $\bar{\mathbf{M}}$ and therefore diagonal in the eigenbasis of $\bar{\mathbf{M}}$. Writing

$$\mathbf{N}_t = \mathbf{U} \operatorname{diag}(\lambda_{1,t}, \dots, \lambda_{n,t})\mathbf{U}^\top,$$

and substituting this into (60), we obtain

$$\lambda_{k,t+1} = (1 - 2\rho_k)\lambda_{k,t} + \rho_k \|\mathbf{N}_t\|, \qquad k = 1, \dots, n.$$

Since $\mathbf{N}_t$ is symmetric and diagonal in the basis $\mathbf{U}$, its operator norm is the maximum absolute value of its eigenvalues. We will prove below that $\mathbf{N}_t \succeq 0$ for all t, so the operator norm is simply the largest eigenvalue:

$$\|\mathbf{N}_t\| = \max_k \lambda_{k,t} = \mu_t.$$

To prove (ii), we observe that this is true for $t = 0$. Assuming $\mathbf{N}_t \succeq 0$, observe that $\mathbf{N}_t$ commutes with $\bar{\mathbf{M}}$ and

$$\begin{aligned}\mathbf{N}_{t+1} &= \mathbf{N}_t(\mathbf{I} - 2\bar{\mathbf{M}}) + \|\mathbf{N}_t\|\,\bar{\mathbf{M}} \\ &= \mathbf{N}_t(\mathbf{I} - \bar{\mathbf{M}}) + (\|\mathbf{N}_t\|\,\mathbf{I} - \mathbf{N}_t)\bar{\mathbf{M}}.\end{aligned}$$

The commuting matrices $\mathbf{N}_t$, $\mathbf{I} - \bar{\mathbf{M}}$, $\|\mathbf{N}_t\|\,\mathbf{I} - \mathbf{N}_t$, and $\bar{\mathbf{M}}$ are all positive semidefinite, hence so are the products $\mathbf{N}_t(\mathbf{I} - \bar{\mathbf{M}})$ and $(\|\mathbf{N}_t\|\,\mathbf{I} - \mathbf{N}_t)\bar{\mathbf{M}}$. Their sum is therefore positive semidefinite, so $\mathbf{N}_{t+1} \succeq 0$.

To prove (iii), we recall the recursion

$$\mathbf{N}_{t+1} = \mathbf{N}_t(\mathbf{I} - \bar{\mathbf{M}}) + (\|\mathbf{N}_t\|\,\mathbf{I} - \mathbf{N}_t)\bar{\mathbf{M}}. \tag{64}$$

The second term on the right-hand side of (64) is positive semidefinite, so

$$\mathbf{N}_{t+1} \succeq \mathbf{N}_t(\mathbf{I} - \bar{\mathbf{M}}). \tag{65}$$

Since $\mathbf{N}_0 = \bar{\mathbf{M}}$, iterating (65) gives

$$\mathbf{N}_t \succeq \bar{\mathbf{M}}(\mathbf{I} - \bar{\mathbf{M}})^t,$$

which is (62). Now diagonalize in the basis of $\bar{\mathbf{M}}$. The eigenvalues of $\bar{\mathbf{M}}(\mathbf{I} - \bar{\mathbf{M}})^t$ are exactly $\rho_k(1 - \rho_k)^t$, so (62) implies

$$\lambda_{k,t} \geq \rho_k(1 - \rho_k)^t, \qquad k = 1, \ldots, n.$$

Taking the maximum over k yields (63). ■

With all the previous preparations, we can now prove the lower bound for the matrix recursion $\mathbf{N}_t$.

Theorem 46 (Lower bound for the matrix recursion) *Let*

$$\alpha_0 = \frac{753}{1000}, \qquad \theta_0 = \frac{3}{4}, \qquad \rho_m = \frac{1 - e^{-2\theta_0/m}}{2} \quad (m \in \mathbb{N}).$$

Then there exist integers $t_{46} \geq 1$ and a constant $c_{46} > 0$ with the following property: for every horizon $T \geq t_{46}$, if one defines the diagonal matrix

$$\bar{\mathbf{M}}^{(T)} := \operatorname{diag}(\rho_1, \rho_2, \ldots, \rho_T) \in \mathbb{R}^{T \times T}$$

and lets $(\mathbf{N}_t^{(T)})_{t \geq 0}$ be the recursion (60) associated with $\bar{\mathbf{M}}^{(T)}$, then

$$\|\mathbf{N}_t^{(T)}\| \geq \frac{c_{46}}{(t+1)^{\alpha_0}} \qquad \textit{for every } 0 \leq t \leq T. \tag{66}$$

Proof By Lemmas 43 and 44, there exists an integer $t_1 \geq 1$ such that

$$\Gamma_{\alpha_0,t}(\theta_0) \geq 1 \qquad \text{for every } t \geq t_1. \tag{67}$$

Fix $T \geq t_1$, and write

$$\mu_t = \|\mathbf{N}_t^{(T)}\|, \qquad \lambda_{k,t} = \lambda_{k,t}^{(T)}.$$

We will prove that

$$\mu_t \geq c_1(t+1)^{-\alpha_0} \qquad (0 \leq t \leq T) \tag{68}$$

for some constant $c_1 > 0$ independent of T.

The shift by $+1$ is convenient for two reasons. First, the theorem is stated starting at $t = 0$. Second, in the induction step the estimate at time $i-1$ becomes

$$\mu_{i-1} \geq c_1 i^{-\alpha_0},$$

which matches exactly the power appearing in $B_t(\rho_t, \alpha_0)$.

By Lemma 45, the eigenvalues satisfy

$$\lambda_{k,t+1} = (1-2\rho_k)\lambda_{k,t} + \rho_k \mu_t,$$

and therefore, after iteration,

$$\lambda_{k,t} = \rho_k(1-2\rho_k)^t + \rho_k \sum_{i=1}^{t} (1-2\rho_k)^{t-i}\mu_{i-1}. \tag{69}$$

We first handle the finitely many times $0 \leq t < t_1$ by choosing a sufficiently small constant $c_1 > 0$. More precisely, by Lemma 45 (iii), we have

$$\mu_t \geq \rho_1(1-\rho_1)^t \qquad (t \geq 0).$$

Hence

$$c_1 = \min_{0 \leq s < t_1} (s+1)^{\alpha_0}\rho_1(1-\rho_1)^s > 0$$

is well defined and independent of T. Consequently,

$$\mu_t \geq c_1(t+1)^{-\alpha_0} \qquad (0 \leq t < t_1). \tag{70}$$

We now prove (68) for $t \geq t_1$ by induction. Fix $t \in \{t_1, \ldots, T\}$, and assume that

$$\mu_s \geq c_1(s+1)^{-\alpha_0} \qquad (0 \leq s < t)$$

or equivalently,

$$\mu_{i-1} \geq c_1 i^{-\alpha_0} \qquad (1 \leq i \leq t).$$

Using (69) with $k = t$, dropping the nonnegative first term, and then using the induction hypothesis, we obtain

$$\mu_t \geq \lambda_{t,t} \geq \rho_t \sum_{i=1}^{t} (1-2\rho_t)^{t-i}\mu_{i-1} \geq c_1\, \rho_t \sum_{i=1}^{t} (1-2\rho_t)^{t-i} i^{-\alpha_0} = c_1 B_t(\rho_t, \alpha_0).$$

Now

$$1 - 2\rho_t = e^{-2\theta_0/t}, \qquad \text{so} \qquad \frac{t}{2}\big(-\log(1-2\rho_t)\big) = \theta_0.$$

Therefore Lemma 42, together with (67), yields

$$t^{\alpha_0} B_t(\rho_t, \alpha_0) \geq \Gamma_{\alpha_0,t}(\theta_0) \geq 1.$$

Hence

$$B_t(\rho_t, \alpha_0) \geq t^{-\alpha_0} \geq (t+1)^{-\alpha_0},$$

and so

$$\mu_t \geq c_1 (t+1)^{-\alpha_0}.$$

which completes the induction.

We have thus shown that (66) holds for every $0 \leq t \leq T$. Therefore the theorem follows with

$$t_{46} = t_1, \qquad c_{46} = c_1.$$

■