Asymptotic theory of SGD with a general learning-rate

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stochastic gradient descent (SGD) with polynomially decaying step-sizes has long underpinned theoretical analyses, yielding a broad spectrum of statistically attractive guarantees. Yet in practice, such schedules find rare use due to their prohibitively slow convergence, revealing a persistent gap between theory and empirical performance. In this paper, we introduce a unified framework that quantifies the uncertainty of online SGD under arbitrary learning-rate choices. In particular, we provide the first comprehensive convergence characterizations for two widely used but theoretically under-examined schemes—cyclical learning rates and linear decay to zero. Our results not only explain the observed behavior of these schedules but also facilitate principled tools for statistical inference and algorithm design. All theoretical findings are corroborated by extensive simulations across diverse settings.

1 Introduction

Stochastic Gradient Descent (SGD) has gained popularity in modern machine learning since the seminal work of Robbins and Monro [1951]. While its theoretical foundations are well established, the literature has largely focused on two standard step-size choices: *constant step-sizes*, which provide exponentially fast convergence to a biased stationary distribution and allow straightforward tuning, and *polynomially decaying step-sizes*, typically of the form $\eta_t \asymp t^{-\alpha}$, which offer statistical guarantees such as consistency and asymptotic normality [Chung, 1954, Sacks, 1958, Fabian, 1968, Ruppert, 1988, Polyak and Juditsky, 1992] and extend to many SGD variants [Poljak, 1964, Gadat and Panloup, 2023, Li et al., 2024b]. However, polynomially decaying schedules converge slowly in practice, while constant step-sizes require careful calibration to avoid divergence [Bengio, 2012]. Hybrid schemes combining these approaches have gained traction in deep learning [He et al., 2016, Smith and Topin, 2019], though learning rates are often chosen empirically or via hyper-parameter tuning over standard schedules [Wu et al., 2018]. This practical reliance leaves a notable gap in theoretical understanding of general step-size effects on SGD. The critical role of learning rates in stochastic approximation convergence has long been recognized [Spall, 2003, Nemirovski et al., 2008], underscoring the need for a unified theoretical framework encompassing a broader range of step-size strategies.

Recent applications have introduced a variety of learning rate schedules that, despite their empirical success, lack comprehensive theoretical support. For instance, Smith [2017] proposes several cyclical learning rate schemes that perform well in practice across standard neural network architectures. Another under-theorized category is that of finite-horizon schedules, where the learning rate depends explicitly on the total number of iterations. In high-dimensional linear regression, for example, Agrawalla et al. [2023] recommends a schedule of the form $\eta_{t,n} \propto \frac{\log n}{n}$. Among such schedules, the linearly decaying to zero (`Linear-D2Z`) learning rate has seen widespread use in training large-scale architectures. Detailed discussion on the relevant literature – though by no means exhaustive – is included in Section 1.3. Despite its prevalence, the non-asymptotic behavior of `Linear-D2Z` and related finite-horizon step-size policies remains poorly understood from a theoretical standpoint. In this work, we aim to bridge this theoretical gap by (i) developing a unified framework to show the non-asymptotic moment convergence, and (ii) explicitly characterizing the non-asymptotic behavior of the wide class of cyclical learning rates as well as `Linear-D2Z`.

1.1 SGD preliminaries

Before summarizing the main contributions of this article, we briefly introduce the stochastic gradient descent (SGD) problem and establish a consistent notational framework for the analysis. Consider the problem of minimizing a function $F : \mathbb{R}^d \to \mathbb{R}$, $F \in \mathcal{C}^1$, given by:

$$\theta^\star = \arg\min_{\theta \in \mathbb{R}^d} F(\theta),$$

and the corresponding SGD algorithm:

$$\theta_t = \theta_{t-1} - \eta_t \nabla f(\theta_{t-1}, \xi_t), \quad \theta_0 \in \mathbb{R}^d, \tag{1.1}$$

where η_t are step-sizes at t-th step, and $\xi_1, \xi_2, \ldots,$ are i.i.d. samples from some unknown distribution $\mathbb{P}_\xi$ such that $\mathbb{E}_{\xi\sim\mathbb{P}_\xi}[\nabla f(\theta_{t-1}, \xi_t)] = \nabla F(\theta)$ for all $\theta \in \mathbb{R}^d$. With this formulation in place, we now proceed to outlining the main contributions of this work.

1.2 Main contributions

This article presents a unified framework for deriving the non-asymptotic mean-squared error for arbitrary learning rate schedules. Our results not only encompass the known theory for polynomially decaying learning rates as a specific case, but also extend the asymptotic analysis to other commonly used learning rates that previously lacked theoretical support. Our main contributions are summarized below.

(1) In our Theorem 2.1, we prove a general bound on the mean error of the SGD iterates, $\mathbb{E}[|\theta_n - \theta^\star|^p]$ involving the step sizes $\{\eta_t\}$. In particular, under certain regularity conditions, we prove the following.

Theorem 1.1 (Theorem 2.1, informal)**.** *If $S_{s,t} = \sum_{j=s+1}^{t} \eta_j$ for $t > s$, then it holds that*

$$\mathbb{E}[|\theta_n - \theta^\star|^p] \lesssim \exp(-c_p S_{1,n})|\theta_0 - \theta^\star|^p + \sum_{j=1}^{n} \eta_j^2 \exp(-c_p S_{j,n}).$$

This result enables straightforward application to a wide range of learning rate schedules η_t, provided that the second term, $\sum_{j=1}^{n} \eta_j^2 \exp(-c_p S_{j,n})$, can be effectively controlled. For example, such bounds are typically tractable for many "approximately" polynomially decaying learning rates. Moreover, the result explicitly captures the influence of the initial point on the final error of the SGD iterates.

(2) A key aspect of Theorem 2.1 is that the step sizes η_j are allowed to depend on the total iteration count n, thereby accommodating finite-horizon learning rate schedules. Such schedules have rarely been studied in the context of mean-squared error. We specifically examine the widely used linearly decaying schedule $\eta_t = \eta(1 - t/n)$ as a representative case. In Theorem 2.2, we leverage Theorem 2.1 to characterize the non-asymptotic moment convergence of SGD iterates under this step-size rule. Although this schedule is common in practice, Theorem 2.2 provides, to the best of our knowledge, its first explicit, rigorous analysis in the online SGD setting. Furthermore, our approach can be readily extended to a broad class of finite-horizon schedules.

(3) When the learning rate is constant ($\eta_t \equiv \eta$) or cyclical ($\eta_t = \eta_{t \bmod T}$), Theorem 2.1 yields only an $O(1)$ bound, which offers only limited insight. While it is well established that, in the constant case, the SGD iterates converge to a stationary distribution, there is, to the best of our knowledge, no existing asymptotic theory for the cyclical learning rate setting. In Theorem 2.4, we address this gap by presenting a novel convergence result for SGD iterates under cyclical learning rate schedules.

Theorem 1.2 (Theorem 2.4, informal)**.** *For a cyclical learning rate $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)$, if not all of the η_k's are too big, then there exists a "cyclostationary" process π such that*

$$\theta_n \overset{w}{\Rightarrow} \pi, \text{ as } n \to \infty,$$

where $\overset{w}{\Rightarrow}$ denotes the convergence in distribution.

This result highlights a fundamental behavioral difference between SGD with cyclical learning rates and with constant learning rates. While the latter typically converges to a stationary distribution, resembling the behavior of a Markov chain, the former converges to a distinct type of non-stationary distribution exhibiting periodic patterns over time- formally known as cyclostationary distribution. To aid the reader's understanding, we also include a brief but formal discussion of cyclostationary processes.

(4) Our theoretical results are substantiated by extensive numerical exercises. Section 3.2 focuses on linearly decaying schedules, which empirically demonstrate both fast early convergence and low final error, consistent with Theorem 2.2, and justifying their practical appeal. On the other hand, Section 3.3 examines cosine schedules, where the learning rate follows a smooth periodic pattern; the resulting error exhibits cyclical behavior, particularly in the variance of its estimate, also complimenting Theorem 2.4. Some additional numerical exercises can be found in Appendix C. Specifically, Appendix C.4 collects five different learning schedules, and provides a comparative study that highlights both their behavior in the "transient" (i.e. with respect to initialization) phase, as well as their asymptotic behavior.

1.3 Related works

There exists a substantial, though primarily empirical, body of literature examining gradient descent and batched SGD in the context of neural networks. For instance, Wu et al. [2019] investigates a variety of step-size schedules, including exponentially decaying and time-inverted schemes. Among the many proposed strategies, our focus is on two broad and widely used classes: cyclical schedules and linearly decaying schedules. Since the introduction of the "triangular" learning rate by Smith [2017], periodic learning rate schemes—and their decaying variants such as cosine annealing—have become influential in training deep architectures like convolutional neural networks (CNNs) [Loshchilov and Hutter, 2017, Smith, 2023, Wang et al., 2023]. The periodic structure of these schedules allows for intermittent large steps (which encourage exploration) followed by smaller steps (which promote convergence), a behavior associated with so-called "super-convergence" as observed in both empirical and theoretical work Smith and Topin [2019], Oymak [2021].

In parallel, annealing-based strategies have also played a prominent role in optimization [Huang et al., 2017, Li et al., 2019, Nakkiran, 2020], with certain variants—such as geometrically decaying step-sizes—proven to be minimax optimal in convex settings [Ge et al., 2019]. Within this context, the linearly decaying to zero (`Linear-D2Z`) schedule has gained significant traction in applications involving highly non-smooth or complex optimization landscapes, including state-space models [Touvron et al., 2023], large language models [Devlin et al., 2019, Liu et al., 2019, Bergsma et al., 2025], and vision transformers [Wu et al., 2024]. Notably, several works advocate for a "knee schedule" [Howard and Ruder, 2018, Hoffmann et al., 2022, Iyer et al., 2023, Defazio et al., 2023, Hägele et al., 2024, Bergsma et al., 2025], which begins with a large learning rate (a "warm start") followed by a `Linear-D2Z` phase. Despite their widespread adoption, the asymptotic behavior of both cyclical and `Linear-D2Z` step-size schedules remains theoretically unexplored—even in relatively simple convex settings. This lack of theoretical understanding presents a significant barrier to rigorous statistical inference and uncertainty quantification, underscoring the need for systematic analysis.

1.4 Notations

In this paper, we denote the set $\{1, \dots, n\}$ by $[n]$. The d-dimensional Euclidean space is $\mathbb{R}^d$. For a vector $a \in \mathbb{R}^d$, $|a|$ denotes its Euclidean norm. For a random vector $X \in \mathbb{R}^d$, we denote $\|X\| := \sqrt{\mathbb{E}[|X|^2]}$. We also denote in-probability convergence, and stochastic boundedness by $o_{\mathbb{P}}$ and $O_{\mathbb{P}}$ respectively. We write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some constant $C > 0$, and $a_n \asymp b_n$ if $C_1 b_n \leq a_n \leq C_2 b_n$ for some constants $C_1, C_2 > 0$. Often we denote $a_n \lesssim b_n$ by $a_n = O(b_n)$. Additionally, if $a_n / b_n \to 0$, we write $a_n = o(b_n)$

2 Non-asymptotic moment convergence of SGD iterates with general step-sizes

This section is devoted to establishing the p-th moment convergence of SGD iterates (1.1) for any $p \geq 2$ with a general choice of learning rate. In particular, we allow for finite-horizon schedules; in the notation of Section 1.1, we allow $\eta_t \equiv \eta_{t,n}$. We note that this represents a significant improvement the existing body of literature that analyzes the statistical properties of SGD and its variants under different learning rate schedules. Before we discuss our main result, it is imperative to introduce the crucial technical assumptions behind our result.

2.1 Technical assumptions

We assume the following regularity assumptions.

Assumption 2.1. *The function F is μ-strongly convex, i.e. for a $\mu > 0$ and for all $x, y \in \mathbb{R}^d$, it holds that*

$$\langle \nabla F(x) - \nabla F(y), x - y \rangle \geq \mu |x - y|^2.$$

The strong-convexity assumption 2.1 can further be relaxed into the strong concordance assumption as follows:

Assumption 2.2 (Local strong convexity)**.** *There exists $\mu^\star > 0$ such that $\nabla_2 F(\theta^\star) \succeq \mu^\star$. Moreover, there exists a constant $C > 0$, and compact set $\Phi \subseteq \mathbb{R}^d$, such that for all $\theta_1, \theta_2 \in \Phi$, it holds that*

$$|\varphi'''(u)| \leq C\,|\theta_1 - \theta_2|\,\varphi''(u), \text{ where } \varphi : u \mapsto F\left(\theta_1 + u\left(\theta_2 - \theta_1\right)\right), u \in \mathbb{R}.$$

We remark that adoption of Assumption of 2.2 instead of Assumption 2.1 does not significantly alter any of our arguments; see Gu and Chen [2024] for details. For simplicity, we stick with Assumption 2.1.

Assumption 2.3. *For the noisy gradients $\nabla f(\cdot, \cdot)$ and some $p \geq 2$, there exists a constant $L_p > 0$ such that*

$$\mathbb{E}[|\nabla f(x, \xi) - \nabla f(y, \xi)|^p] \leq L_p^p |x - y|^p, \quad \text{for all } \ x, y \in \mathbb{R}^d.$$

In particular, for some constant $M_p > 0$, it holds that

$$(\mathbb{E}[|\nabla f(\theta^\star, \xi)|^p])^{1/p} =: M_p < \infty.$$

Assumption 2.3 entails that F is L_p-smooth by Hölder's inequality; in other words, for all $x, y \in \mathbb{R}^d$, it holds that

$$|\nabla F(x) - \nabla F(y)| \leq L_p |x - y|.$$

Assumptions 2.1 and 2.3 are standard features of statistical analysis of convex stochastic optimization, and have appeared extensively in Ruppert [1988], Polyak and Juditsky [1992], Bottou et al. [2018], Chen et al. [2020], Zhu et al. [2023], Wei et al. [2023], Li et al. [2024a]. With these standard regularity assumptions, we can introduce our general result.

2.2 A general moment convergence of SGD iterates

In this section, we introduce our main contribution – an umbrella result that furnishes a ready-made upper-bound of the SGD iterates (1.1) for any choice of learning rates. In particular, we have the following result.

Theorem 2.1 (L^p Convergence)**.** *Suppose that Assumptions 2.1 and 2.3 hold for some $p \geq 2$. Let $c_0 > 0$ be some universal constant. For any $t \geq 1$ such that η_t satisfies*

$$0 < \eta_t < \frac{2\mu}{(6p-5)L_p^2}, \tag{2.1}$$

we have,

$$\|\theta_t - \theta^\star\|_p^2 \leq \exp\Big\{ - c_0 \sum_{k=1}^{n} \eta_k \Big\} |\theta_0 - \theta^\star|^2 + 3(p-1)M_p^2 \sum_{j=1}^{n} \eta_j^2 \exp\Big\{ - c_0 \sum_{k=j+1}^{n} \eta_k \Big\}. \tag{2.2}$$

Theorem 2.1 is proved in appendix Section A.1. The bound (2.2) highlights the two key terms that the learning rates contribute in the moment bound. In particular, there is an inherent trade-off between the potential choices of step-sizes η_t that goes into determining the order of the p-th moment. We discuss this property in detail in the subsequent two remarks.

Remark 2.1 (Effect of initialization). Firstly, the $\exp$ term in (2.2) highlights that in order to neglect the effect of initialization, one must have $\sum_{k=1}^{n} \eta_k \to \infty$ as $n \to \infty$; in other words, the step-sizes cannot be too small. For example, Wu et al. [2019] discusses exponentially decaying step-sizes $\eta_t = \gamma^t$, whose performance heavily depends on the initial point even for large n, indicating that the effect of initialization cannot be ignored in this case.

Remark 2.2 (Effect of exploration). The second term $\sum_{j=1}^{n} \eta_j^2 \exp\Big\{ - c_0 \sum_{k=j+1}^{n} \eta_k \Big\}$ encodes the exploration property of the SGD iterates θ_t. Intuitively, if $\eta_n \to 0$ as $n \to \infty$, then this second term is also $o(1)$. Therefore, this term essentially ensures that η_j has to be decaying, and not all of them can be too big.

It is instructive to examine specific learning rate choices and their implications as reflected by the bound in (2.2). For example, with the commonly used polynomially decaying schedule $\eta_t \asymp t^{-\beta}$, the first term behaves like $\exp(-t^{1-\beta})$, while the second term is on the order of $O(\eta_t)$, recovering the classical mean square error (MSE) rate for this setting. In the following section, we apply Theorem 2.1 to analyze another important and theoretically less-explored finite-horizon schedule: the linearly decaying to zero (`Linear-D2Z`) learning rate.

2.3 Linear decaying step-sizes

As an important application of Theorem 2.1, consider the `Linear-D2Z` learning rate $\eta_t = \eta(1 - t/n)$. This learning schedule has recently been at the forefront of training large architectures, and its optimality properties have been investigated both theoretically [Defazio et al., 2023] and empirically Bergsma et al. [2025] in different context. Despite this interest, its non-asymptotic convergence rate remains unknown in the literature. Leveraging the bound in (2.2), we analyze the L_p convergence behavior of SGD under this learning rate schedule.

Theorem 2.2. *Recall θ_n from (1.1). Under the conditions of Theorem 2.1, we have*

$$\|\theta_n - \theta^\star\|_p^2 \le |\theta_0 - \theta^\star|^2 e^{-c_0 \eta (n-1)/2} + \frac{C}{\sqrt{n}},$$

where $C > 0$ is a universal constant.

Remark 2.3. Theorem 2.2 offers a remarkable insight into the behavior of the linearly decaying learning rate: it effectively combines the advantages of both constant and polynomially decaying step-sizes by being consistent and forgetting the initial condition at an exponential rate. Specifically, for any $c \in (0, 1)$ and all iterations $t \le \lfloor nc \rfloor$, the step size satisfies $\eta_t \ge \eta(1 - c)$; that is, the learning rate behaves like a constant for a substantial portion of the optimization process, providing a "warm start" and ensuring exponential decay relative to the initial point. Conversely, when $t = \lceil n - c_0 n^{1-c} \rceil$ for some $c > \frac{1}{2}$ and $c_0 > 0$, the step size satisfies $\eta_t = \eta c_0 n^{-c} \asymp t^{-c}$, mimicking a polynomially decaying schedule that yields the MSE of order $O(n^{-1/2})$. Therefore, by leveraging the strengths of both constant and polynomially decaying learning rates, the linearly decaying to zero (`Linear-D2Z`) schedule achieves a "best-of-both-worlds" effect. This theoretical insight is empirically validated in Section 2.3.

2.4 Asymptotic convergence in distribution of cyclical step-sizes

Note that for constant or cyclical learning rate schemes, Theorem 2.1 can only guarantee an MSE bound of order $O(1)$. This naturally motivates a deeper investigation into the convergence properties of these schedules. Specifically, an SGD sequence as defined in (1.1) with a constant step size $\eta_t \equiv \eta$ can be interpreted as an aperiodic Markov chain. Under standard regularity conditions, it is well-known that the iterates θ_t converge weakly to a stationary distribution. However, as discussed, recent empirical work has highlighted the benefits of periodic or cyclical step-size schedules, such as $\eta_t \equiv \eta_{t \bmod T}$. In this setting, the time-varying learning rate breaks the asymptotic stationarity of the SGD chain. Nonetheless, the periodic structure of the step-size schedule induces a corresponding periodicity in the asymptotic behavior of the iterates. Such non-stationary processes, characterized by recurring statistical properties over time, are known as cyclostationary processes, which we briefly introduce below.

Definition 2.3 (Cyclostationary process)**.** *A stochastic process $\{X_t\}_{t\in\mathbb{R}}$ is said to be* cyclostationary *with period $T > 0$ if it holds that for all $s \in [T]$, and $i \in \mathbb{N}$, $\{X_i, \dots, X_{i+s}\} \overset{d}{=} \{X_{i+T}, \dots, X_{i+s+T}\}$.*

Cyclostationary process were introduced as a model of communications systems in Bennett [1958] and Franks [1969], later finding wide use in econometrics [Parzen and Pagano, 1979] as well as atmospheric sciences [Bloomfield et al., 1994] – the reader is encouraged to look into [Gardner et al., 1994, Napolitano, 2016], and the references therein for an introduction and a comprehensive list of all its applications. In the context of SGD, it is instructive to look at the iterative random function construction of the cyclostationary process, as introduced by Bonnerjee et al. [2024]:

$$X_t = g(\phi_t, \mathcal{F}_t),\ \mathcal{F}_t = \sigma(\varepsilon_s : s \le t),\ \phi_t = \phi_{t \bmod T},\ \text{for some period } T \in \mathbb{N}. \tag{2.3}$$

This representation suggests an immediate connection to the SGD iterates θ_t in (1.1), which, in the case of cyclical learning schedules, can be represented as

$$\theta_t = F_{\xi_t}(\eta_t, \theta_{t-1}),\ \eta_t = \eta_{t \bmod T},\ \text{for some period } T \in \mathbb{N}, \text{ with } F_\xi(\eta, \theta) = \theta - \eta \nabla f(\theta, \xi). \tag{2.4}$$

Equations (2.3) and (2.4) suggest an immediate connection between the cyclostationary process and SGD with cyclic learning rate, with the choice $\phi_t = \eta_t$ for $t \in [T]$. The following result, proved in appendix Section A.2, makes this connection precise by establishing a novel asymptotic convergence result.

Theorem 2.4. *Suppose that Assumptions 2.1 and 2.3 hold for some $p > 2$. Let $\rho_p(\gamma)^p := (1+\gamma L_p)^p - p\gamma L_p - p\mu\gamma$, $\gamma \in \mathbb{R}$, where μ and L_p are as in Assumptions 2.1 and 2.3 respectively. Consider a periodic step-size schedule with fixed period T. Then there exist T stationary processes $\pi_1, \dots, \pi_T$ such that for all $\boldsymbol{\eta} := (\eta_1, \dots, \eta_T) \in \mathbb{R}^T$ satisfying*

$$\rho_p(\eta_1) \dots \rho_p(\eta_T) < 1, \tag{2.5}$$

it holds that

$$\theta_{nT+i} \overset{w}{\Rightarrow} \pi_i \text{ as } n \to \infty \text{ for all } i \in [T]. \tag{2.6}$$

Moreover, if $\boldsymbol{\eta}$ further satisfies

$$\min_s \mathcal{J}_p(s) < 1,\ \text{with } \mathcal{J}_p(s) = \sum_{k=1}^{T} \prod_{j=1}^{k} \rho_p(\eta_{s+j})^p, \tag{2.7}$$

where $\eta_j = \eta_{j \bmod T}$ for $j > T$, then there exists a cyclostationary process π, such that

$$\theta_n \overset{w}{\Rightarrow} \pi \text{ as } n \to \infty. \tag{2.8}$$

Remark 2.4. Note that (2.5) ensures that $\rho_p(\eta_{s^\star}) < 1$, where $s^\star = \arg\min_s \mathcal{J}_p(s)$. In contrast to the SGD with constant learning rate, none of the conditions (2.5) and (2.7) presupposes that η_i's are required to satisfy $\rho_p(\eta_i) < 1$ for each $i \in [T]$. In particular, at least some of the η_i's may be taken to be large, which helps in faster convergence, which is also seen empirically in Figure 4. This result underpins the flexibility and the resulting popularity of the periodic step-size schedule over its constant counter-part, guaranteeing convergence under very mild conditions (2.5) and (2.7) respectively.

3 Simulation

To validate our theoretical analysis, we conduct an empirical study of SGD with various learning rate schedules. The goal is to assess how different step-size strategies affect convergence behavior and mean squared error, and how these compare with the theoretical predictions in Theorems 2.1, 2.2 and 2.4. For simplicity, all experiments use a simple linear regression model with known ground truth and are repeated across multiple Monte Carlo runs to estimate average performance and variability. In particular, Section 3.1 provides the model specifications. Section 3.2, we study the linearly decaying to zero (Linear-D2Z) schedule $\eta_t = \eta_0(1 - t/n)$, confirming its "best-of-both-worlds" performance—fast early convergence and diminishing final error—as predicted by Theorem 2.2. Finally, Section 3.3 examines cosine learning rate schedules of the form $\eta_t = \eta_0(1 + \cos(\pi t/T))$, where we observe periodic error fluctuations that empirically confirm the cyclostationary behavior predicted by Theorem 2.4. Additional empirical studies can be found in Appendix C, where we also include a particularly illuminating comparative study between the different schedules in Section C.4. All the code files are available in github.

3.1 Model specification

All the experiments are based on the following simple linear regression model:

$$y_i = \theta^{(0)} + \theta^{(1)} x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1) \text{ i.i.d.}, \; \theta^\star = (\theta^{(0)}, \theta^{(1)})^\top \in \mathbb{R}^2. \tag{3.1}$$

where $(x_i, y_i) \in \mathbb{R}^2$ denotes the observed data and $\theta^\star \in \mathbb{R}^2$ is the unknown parameter. The true parameter vector is fixed at $\theta^\star = (2, -3)^\top$ throughout all experiments. For all the subsequent simulation studies, we initialize the SGD chain at $(0, 0)^\top$, which provides sufficient distance for meaningful comparisons across different learning rate schedules, while not being so far away from the ground truth so that it fails to converge and denies us the full picture. Subsequently, we focus on an empirical evaluation of both the convergence trajectory and the distribution of the final error across different learning rate strategies.

3.2 `Linear-D2Z` rate

We consider `Linear-D2Z` schedules of the form $\eta_t = \eta_0(1 - t/n)$, which, despite their widespread use, have received comparatively little theoretical attention. Similar to Section 3.3, we let $\eta_0 \in \{0.01, 0.05, 0.1\}$, and for each experiment, the mean errors are estimated via $n_{iter} = 500$ many independent repetitions. Firstly, to analyze the non-asymptotic MSE of the end-term SGD iterates, θ_n, we run SGD on the same regression task with $n \in \{100, 200, \ldots, 10^4\}$, using 500 independent repetitions for each n. Figure **??** displays the terminal squared error, which decays polynomially with n, in line with Theorems 2.1 and 2.2.

The appeal of the `Linear-D2Z` schedule lies in its hybrid structure: as explained in Remark 2.3, early iterations benefit from relatively large step sizes, enabling rapid descent—potentially faster than a constant-rate scheme. Later, the schedule tapers off, reducing variance and yielding low final error. We numerically investigate this as follows. For a fixed $n = 10^4$, Figure 2 shows the first 100 iterations for $\eta_0 \in \{0.05, 0.1, 0.5\}$. Across all settings, the error drops sharply, even under the high-variance $\eta_0 = 0.5$ case, demonstrating the robustness of the approach. Later in training, as shown in Figure 3) the error decay appears nearly linear before plateauing, with the stabilization occurring earlier for larger η_0.

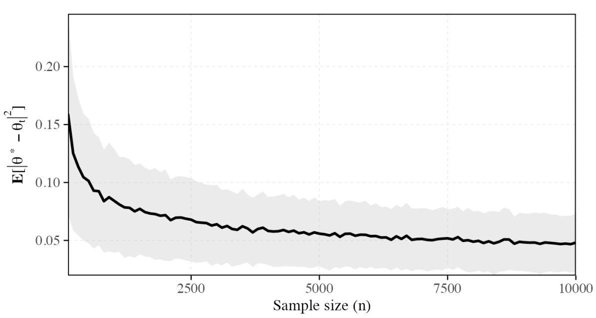

Figure 1: Plot of the terminal MSE estimate averaged over 500 SGD runs for $n = 100$ to $n = 10^4$.

3.3 Cosine learning rate

As an example of the cyclical learning rate, we employ the widely-used cosine scheduling. Specifically, we employ the schedule:

$$\eta_t = \eta_0 \left(1 + \cos\left(\frac{2\pi t}{T}\right)\right), \tag{3.2}$$

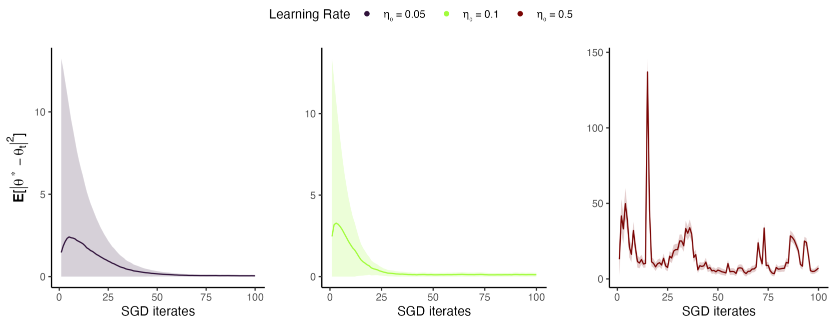

Figure 2: Plot of the MSE estimate averaged over 500 SGD runs for $n = 10^4$ iterations under a `Linear-D2Z` schedule, observing from step 1 to 100.

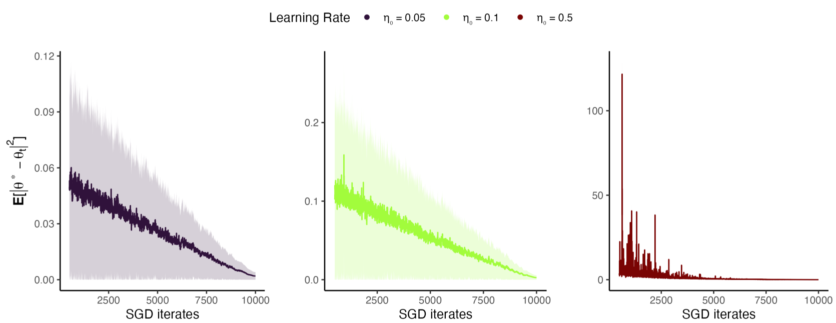

Figure 3: Plot of the MSE estimate averaged over 500 SGD runs for $n = 10^4$ iterations under a `Linear-D2Z` schedule, observing from step 500 onward.

where η_0 is the base learning rate and T denotes the period.

To assess the behavior induced by such schedules, we perform online SGD for $n = 10^4$ iterations with $\eta_0 \in \{0.01, 0.05, 0.1\}$ and $T = 3$, averaging results over $n_{iter} = 500$ independent trials. Figure 4 presents the resulting mean squared error (MSE) trajectories. Across all settings, the MSE exhibits an exponentially fast decay from the initial points before exhibiting persistent fluctuations about a steady-state level. The periodicity, as predicted by Theorem 2.4 is not apparent from this plot. To further probe this structure, we examine the *standard deviation* of the MSE across runs over the final 100 iterations in Figure 5. Despite autocorrelation between estimates (due to small step sizes), periodicity remains visible in the standard deviation curves. Even for $\eta_0 = 0.01$, where the process converges slowly, the periodicity of order 3 is easily discernible on the plot. This empirically confirms Theorem 2.4: despite the huge auto-correlation between successive iterations, with cyclical learning rates SGD iterates θ_t do not settle but oscillate periodically, reflecting the learning rate's structure. The oscillation's amplitude and frequency depend on η_0, with larger values causing stronger fluctuations and faster initial progress. This nuanced behavior highlights the balance between exploration and convergence enabled specifically by periodic schedules. For comparison, Figure 6 shows standard deviation curves under constant learning rates, which lack periodicity, confirming that the patterns in Figure 5 arise from the cosine schedule.

4 Conclusion

Sharp theoretical MSE bounds offer critical insights into the behavior of SGD for given learning rate schedules, yet most prior work has focused on polynomially decaying step sizes, often sacrificing convergence speed for statistical tractability. To the best of our knowledge, this paper is the first to systematically develop a unified framework that provides explicit MSE upper bounds for a broad class of learning rates. In particular, we establish novel convergence guarantees for cyclical and linearly decaying to zero (`Linear-D2Z`) learning rates—two popular but previously under-

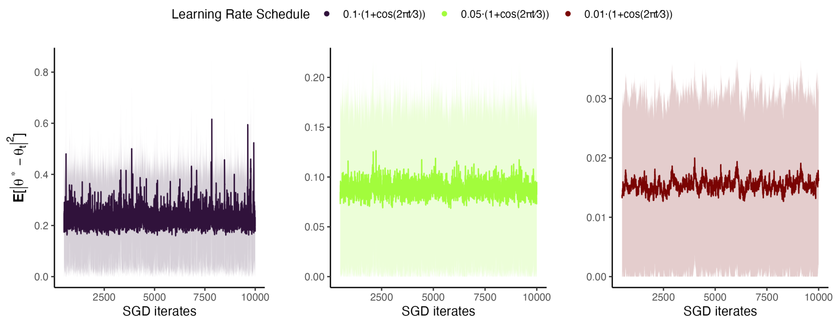

Figure 4: MSE estimates over 500 SGD runs (10^4 steps) with cosine learning rates, observing from step 500 onward. Periodic error fluctuations are evident, indicating cyclostationary dynamics.

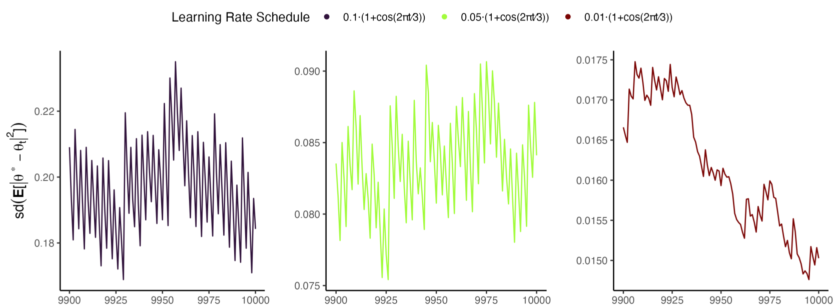

Figure 5: Standard deviation of the MSE over the final 100 iterations (across 500 runs) with cosine learning rates. Periodic fluctuation is clearly observed, even at a small η_0.

Figure 6: Standard deviation of the MSE over the final 100 iterations with constant learning rates. No periodicity is observed.

theorized choices—shedding light on their strong empirical performance. Our results motivate further exploration beyond the convex setting into non-convex and non-smooth landscapes, with an emphasis on understanding the statistical behavior of these schedules, including the potential for central limit theorems and refined uncertainty quantification. In this context, this work provides new insights into the practically important yet theoretically underexplored area of learning rate selection, and serves as a foundation for bridging practical success and theoretical understanding of SGD across diverse learning schedules regimes.

References

B. Agrawalla, K. Balasubramanian, and P. Ghosal. Statistical inference for linear functionals of online sgd in high-dimensional linear regression. *arXiv preprint arXiv:2302.09727*, 2023.

Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In G. Montavon, G. B. Orr, and K. Müller, editors, *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 437–478. Springer, 2012. doi: 10.1007/978-3-642-35289-8_26. URL https://doi.org/10.1007/978-3-642-35289-8_26.

W. R. Bennett. Statistics of regenerative digital transmission. *Bell System Tech. J.*, 37:1501–1542, 1958. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1958.tb01560.x. URL https://doi.org/10.1002/j.1538-7305.1958.tb01560.x.

S. Bergsma, N. S. Dey, G. Gosal, G. Gray, D. Soboleva, and J. Hestness. Straight to zero: Why linearly decaying the learning rate to zero works best for llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=hrOlBgHsMI.

P. Bloomfield, H. L. Hurd, and R. B. Lund. Periodic correlation in stratospheric ozone data. *J. Time Ser. Anal.*, 15(2):127–150, 1994. ISSN 0143-9782,1467-9892. doi: 10.1111/j.1467-9892.1994.tb00181.x. URL https://doi.org/10.1111/j.1467-9892.1994.tb00181.x.

S. Bonnerjee, S. Karmakar, and W. B. Wu. Gaussian approximation for nonstationary time series with optimal rate and explicit construction. *Ann. Statist.*, 52(5):2293–2317, 2024. ISSN 0090-5364,2168-8966. doi: 10.1214/24-aos2436. URL https://doi.org/10.1214/24-aos2436.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018. ISSN 0036-1445,1095-7200. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173.

X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251–273, 2020. ISSN 0090-5364,2168-8966. doi: 10.1214/18-AOS1801. URL https://doi.org/10.1214/18-AOS1801.

K. L. Chung. On a stochastic approximation method. *Ann. Math. Statistics*, 25:463–483, 1954. ISSN 0003-4851. doi: 10.1214/aoms/1177728716. URL https://doi.org/10.1214/aoms/1177728716.

A. Defazio, A. Cutkosky, H. Mehta, and K. Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*, 2023.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

V. Fabian. On asymptotic normality in stochastic approximation. *Ann. Math. Statist.*, 39:1327–1332, 1968. ISSN 0003-4851. doi: 10.1214/aoms/1177698258. URL https://doi.org/10.1214/aoms/1177698258.

L. Franks. *Signal Theory*. Information theory series. Prentice-Hall, 1969. ISBN 9780138100773. URL https://books.google.com/books?id=f_dSAAAAMAAJ.

S. Gadat and F. Panloup. Optimal non-asymptotic analysis of the Ruppert-Polyak averaging stochastic algorithm. *Stochastic Process. Appl.*, 156:312–348, 2023. ISSN 0304-4149,1879-209X. doi: 10.1016/j.spa.2022.11.012. URL https://doi.org/10.1016/j.spa.2022.11.012.

W. A. Gardner et al. *Cyclostationarity in communications and signal processing*, volume 1. IEEE press New York, 1994.

R. Ge, S. M. Kakade, R. Kidambi, and P. Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14951–14962, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2f4059ce1227f021edc5d9c6f0f17dc1-Abstract.html.

J. Gu and S. X. Chen. Statistical inference for decentralized federated learning. *Ann. Statist.*, 52 (6):2931–2955, 2024. ISSN 0090-5364,2168-8966. doi: 10.1214/24-aos2452. URL https://doi.org/10.1214/24-aos2452.

A. Hägele, E. Bakouch, A. Kosson, L. B. Allal, L. von Werra, and M. Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8b970e15a89bf5d12542810df8eae8fc-Abstract-Conference.html.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. W. Rae, and L. Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1031. URL https://aclanthology.org/P18-1031/.

G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJYwwY9ll.

N. Iyer, V. Thejas, N. Kwatra, R. Ramjee, and M. Sivathanu. Wide-minima density hypothesis and the explore-exploit learning rate schedule. *J. Mach. Learn. Res.*, 24:65:1–65:37, 2023. URL https://jmlr.org/papers/v24/21-0549.html.

J. Li, Z. Lou, S. Richter, and W.-B. Wu. The stochastic gradient descent from a nonlinear time series perspective. *preprint*, 2024a.

T. Li, T. Xiao, and G. Yang. Revisiting the central limit theorems for the SGD-type methods. *Commun. Math. Sci.*, 22(5):1427–1454, 2024b. ISSN 1539-6746,1945-0796.

Y. Li, C. Wei, and T. Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11669–11680, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bce9abf229ffd7e570818476ee5d7dde-Abstract.html.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

P. Nakkiran. Learning rate annealing can provably help generalization, even for convex problems. *arXiv preprint arXiv:2005.07360*, 2020.

A. Napolitano. Cyclostationarity: New trends and applications. *Signal Processing*, 120:385–408, 2016. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2015.09.011.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008. ISSN 1052-6234,1095-7189. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

S. Oymak. Provable super-convergence with a large cyclical learning rate. *IEEE Signal Process. Lett.*, 28:1645–1649, 2021. doi: 10.1109/LSP.2021.3101131. URL https://doi.org/10.1109/LSP.2021.3101131.

E. Parzen and M. Pagano. An approach to modeling seasonally stationary time series. *Journal of Econometrics*, 9(1):137–153, 1979. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(79)90100-3. URL https://www.sciencedirect.com/science/article/pii/0304407679901003.

B. T. Poljak. Some methods of speeding up the convergence of iterative methods. *Ž. Vyčisl. Mat i Mat. Fiz.*, 4:791–803, 1964. ISSN 0044-4669.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

E. Rio. Moment Inequalities for Sums of Dependent Random Variables under Projective Conditions. *Journal of Theoretical Probability*, 22(1):146–163, Mar. 2009.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

D. Ruppert. Efficient estimations from a slowly convergent robbins-monro process. *Technical Report*, 1988.

J. Sacks. Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.*, 29:373–405, 1958. ISSN 0003-4851. doi: 10.1214/aoms/1177706619. URL https://doi.org/10.1214/aoms/1177706619.

L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, pages 464–472. IEEE Computer Society, 2017. doi: 10.1109/WACV.2017.58. URL https://doi.org/10.1109/WACV.2017.58.

L. N. Smith. General cyclical training of neural networks. *Adv. Artif. Intell. Mach. Learn.*, 3(1):958–976, 2023. doi: 10.54364/AAIML.2023.1157. URL https://doi.org/10.54364/aaiml.2023.1157.

L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

J. C. Spall. *Introduction to stochastic search and optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003. ISBN 0-471-33052-3. doi: 10.1002/0471722138. URL https://doi.org/10.1002/0471722138. Estimation, simulation, and control.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

W. Wang, C. M. Lee, J. Liu, T. Çolakoglu, and W. Peng. An empirical study of cyclical learning rate on neural machine translation. *Nat. Lang. Eng.*, 29(2):316–336, 2023. doi: 10.1017/S135132492200002X. URL https://doi.org/10.1017/S135132492200002X.

Z. Wei, W. Zhu, and W. B. Wu. Weighted averaged stochastic gradient descent: Asymptotic normality and optimality. *arXiv preprint arXiv:2307.06915*, 2023.

S. Wu, G. Zhang, and X. Liu. SwinSOD: Salient object detection using swin-transformer. *Image Vis. Comput.*, 146(105039):105039, June 2024.

X. Wu, R. Ward, and L. Bottou. Wngrad: Learn the learning rate in gradient descent. *arXiv preprint arXiv:1803.02865*, 2018.

Y. Wu, L. Liu, J. Bae, K. H. Chow, A. Iyengar, C. Pu, W. Wei, L. Yu, and Q. Zhang. Demystifying learning rate policies for high accuracy training of deep neural networks. In C. K. Baru, J. Huan, L. Khan, X. Hu, R. Ak, Y. Tian, R. S. Barga, C. Zaniolo, K. Lee, and Y. F. Ye, editors, *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019*, pages 1971–1980. IEEE, 2019. doi: 10.1109/BIGDATA47090.2019.9006104. URL https://doi.org/10.1109/BigData47090.2019.9006104.

W. Zhu, X. Chen, and W. B. Wu. Online covariance matrix estimation in stochastic gradient descent. *J. Amer. Statist. Assoc.*, 118(541):393–404, 2023. ISSN 0162-1459,1537-274X. doi: 10.1080/01621459.2021.1933498. URL https://doi.org/10.1080/01621459.2021.1933498.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims of the paper are accurately reflected in both the abstract and introduction. The paper proposes a general framework to theoretically derive the MSE of SGD iterates for a general class of learning schedules. All claims are theoretically proven and empirically validated throughout the work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All theoretical assumptions underlying the SGD setting are explicitly stated and discussed in detail in Section 2.1. These sections also elaborate on the limitations of the asymptotic analysis, particularly in relation to the two specific learning rate schedules examined in greater depth. These limitations are highlighted through dedicated remarks following the main theorems. Additional commentary on open questions and potential extensions is provided in the conclusion (Section 4).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the assumptions can be found in the theorem statements, and are discussed in main paper. All the proofs can be found in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The complete model specifications along with parameters, are provided in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the codes to reproduce the results can be found anonymously in github as well as in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details are provided in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, whenever the MSE of the SGD iterates is reported, an estimate of their variability/error-bars is reported in the form of a shade.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The experiments are lightweight and run quickly on a modern laptop.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research follows all ethical guidelines. No human data or ethically sensitive content is involved.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Since our work is theoretical in nature, we do not anticipate any negative impacts, and as such the paper does not include a dedicated speculative discussion of broader societal impacts in a separate section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release any models or datasets with high risk of misuse. All released components are synthetic and pose no privacy or safety risk.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All baseline methods are standard, and citations to prior work are provided with proper attribution and licensing notices.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces novel techniques to derive MSE of SGD iterates with any step-size choice, which are theoretically validated with extensive proofs. Moreover, it characterizes the convergence behavior for two widely-used, but theoretically less-understood learning schedules. All the accompanying empirical evidence is documented and released via a GitHub repository in anonymized form. Hyperparameters, dependencies, and usage instructions are all included.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: The paper does not involve crowdsourcing or human subject research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: No human subjects are involved in this research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: No large language models (LLMs) were used to produce, analyze, or verify the scientific content of this paper. All methods and results are original and independently verified using rigorous mathematical analysis and custom code.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Postponed Proofs

A.1 Proof of Theorem 2.1

Proof of Theorem 2.1. By recursively applying Lemma B.2, we have

$$\|\theta_n - \theta^\star\|_p^2 \le \prod_{k=1}^n (1 - c_0\eta_k)|\theta_0 - \theta^\star|^2 + 3(p-1)M_p^2 \sum_{j=1}^n \eta_j^2 \prod_{k=j+1}^n (1 - c_0\eta_k).$$

The proof is completed by noting that

$$\prod_{k=1}^n (1 - c_0\eta_k) \le \exp\Big\{ - c_0 \sum_{k=1}^n \eta_k \Big\}.$$

□

A.2 Proof of Theorem 2.4

Proof of Theorem 2.4. Suppose $F_{X,\gamma}(\theta) = \theta - \gamma\nabla f(\theta, \xi)$ encodes the iterative random function governing the SGD trajectory (1.1). Fix $i \in [T]$. For (2.6), observe that

$$\theta_{nT+i} = F^i_{(\xi_{(n-1)T+i+1},\ldots,\xi_{nT+i})}(\theta_{(n-1)T+i}, \boldsymbol{\eta}),$$

where, for $\mathbf{X} = (X_1, \ldots, X_T)$, we define $F^i_{\mathbf{X}}(\theta, \boldsymbol{\eta}) := F_{X_1,\eta_{i+1}} \circ \ldots \circ F_{X_T,\eta_{i+T}}(\theta)$, with $\eta_{i+s} = \eta_{i+s \bmod T}$ with slight abuse of notation. Applying Theorem 2.2 of Li et al. [2024a] successively on the function compositions of F^i, it holds that

$$\|F^i_{\mathbf{X}}(\theta, \boldsymbol{\eta}) - F^i_{\mathbf{X}}(\theta', \boldsymbol{\eta})\| \le \rho_p(\eta_1) \ldots \rho_p(\eta_T),$$

from which, (2.6) follows in light of (2.5) and Theorem 2.2 of Li et al. [2024a]. To ensure (2.8), note that if π is a cyclostationary process with period T defined on $\mathbb{R}^d$, then $X \sim \pi$ iff $(X_1, \ldots, X_T) \sim \tilde{\pi}$ for some stationary process π on $\mathbb{R}^{d\times T}$. Therefore, it is enough to show

$$(\theta_{nT+s} : \ldots : \theta_{(n+1)T+s}) \to \tilde{\pi} \tag{A.1}$$

for some stationary process $\tilde{\pi}$ on $\mathbb{R}^{d\times T}$, and some $s \in [T]$. Choose $s = s^\star$ such that

$$s^\star = \arg\min_{s\in[T]} \sum_{k=1}^T \prod_{j=1}^k \rho_p(\eta_{s+j})^p$$

Define

$$\tilde{F}_{\mathbf{X}}(\boldsymbol{\Theta}, \boldsymbol{\eta}) = (F_{X_1,\eta_{s+1}}(\theta_t), F_{X_2,\eta_{s+2}} \circ F_{X_1,\eta_{s+1}}(\theta_t), \ldots, F_{X_T,\eta_{s+T}} \circ \cdots \circ F_{X_1,\eta_{s+1}}(\theta_t)),$$

where

$$\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_T) \in \mathbb{R}^{d\times T}.$$

Clearly, one derives

$$\begin{aligned}\|\tilde{F}_{\mathbf{X}}(\boldsymbol{\Theta}, \boldsymbol{\eta}) - F^i_{\mathbf{X}}(\boldsymbol{\Theta}', \boldsymbol{\eta})\|_p^p &= \sum_{k=1}^T \|F_{X_k,\eta_{s+k}} \circ \cdots F_{X_1,\eta_{s+1}}(\theta_t) - F_{X_k,\eta_k} \circ \cdots F_{X_1,\eta_1}(\theta'_t)\|_p^p \\ &\le \|\theta_t - \theta'_t\|_p^p \sum_{k=1}^T \prod_{j=1}^k \rho_p(\eta_{s+j})^p. \end{aligned} \tag{A.2}$$

Writing (1.1) as

$$(\theta_{nT+s+1,\ldots,\theta_{(n+1)T+s}}) = \tilde{F}_{\xi_{nT+s+1,\ldots,\xi_{(n+1)T+s}}}((\theta_{(n-1)T+s+1,\ldots,\theta_{nT+s}}), \boldsymbol{\eta}),\ n \ge 1$$

yet another application of Theorem 2.2 of Li et al. [2024a] yields (A.1) in light of (A.2) and (2.7). □

A.3 Proof of Corollary 2.2

Proof of Corollary 2.2. The first term in (2.2) becomes $e^{-c_0\eta(n-1)/2}$. For the second term, set $m = n - j$. Then $m = 0, 1, \ldots, n-1$, and

$$\eta_j = \eta\frac{n-j}{n} = \eta\frac{m}{n}, \qquad \sum_{k=j+1}^{n}\eta_k = \sum_{k=0}^{m-1}\eta\frac{k}{n} = \frac{\eta}{n}\frac{(m-1)m}{2} = \frac{\eta}{2n}(m^2 - m).$$

Let $S_n = \sum_{j=1}^{n}\eta_j^2\exp\Big\{-c_0\sum_{k=j+1}^{n}\eta_k\Big\}$, then since $m \le n$,

$$S_n = \sum_{m=0}^{n-1}\Big(\eta\frac{m}{n}\Big)^2\exp\Big\{-c_0\frac{\eta}{2n}(m^2-m)\Big\} \le \eta^2\exp(\frac{c_0\eta}{2})\sum_{m=0}^{n-1}\frac{m^2}{n^2}\exp\Big\{-\frac{c_0\eta m^2}{2n}\Big\}.$$

The sum can be further bounded by the integration. Since the function $x^2\exp\{c_0\eta x^2/2\}$ is eventually decreasing with x, we have

$$\sum_{m=0}^{n-1}\frac{m^2}{n^2}\exp\Big\{-\frac{c_0\eta m^2}{2n}\Big\} \le \frac{C'}{\sqrt{n}}\int_0^\infty x^2\exp\Big\{-\frac{c_0\eta x^2}{2}\Big\}dx = O(1/\sqrt{n})$$

where C' is a universal constant, and the inequality is obtained by substituting $x = \frac{m^2}{n}$. This completes the proof. ☐

B Auxiliary Section

In this section we collect two crucial auxiliary results that contribute towards the proof of Theorem 2.1.

Lemma B.1 (Rio's inequality [Rio, 2009]). *Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$ be two random vectors such that $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^p < \infty$ for some $p \ge 2$. Then we have*

$$\|X+Y\|_p^2 \le \|X\|_p^2 + (p-1)\|Y\|_p^2.$$

Lemma B.2. *Consider the SGD iterates $\{\theta_t\}_{t\ge1}$ in (1.1). Suppose that Assumptions 2.1 and 2.3 for $p \ge 2$. Then, for some constant $c_0 > 0$ such that for all $t \ge 1$,*

$$c_0 \le \min\Big\{\frac{1}{\eta_t}, 2\mu - (6p-5)L_p^2\eta_t\Big\},$$

we have, for all $t \ge 1$,

$$\|\theta_t - \theta^\star\|_p^2 \le (1 - c_0\eta_t)\|\theta_{n-1} - \theta^\star\|_p^2 + 3(p-1)\eta_t^2 M_p^2.$$

B.1 Proof of Lemma B.2

Proof of Lemma B.2. Since ξ_t, for $t \ge 1$, are i.i.d. random samples, it follows from the tower rule that

$$\mathbb{E}[\nabla f(\theta_{n-1},\xi_n) - \nabla F(\theta_{n-1}) \mid \theta_{n-1}] = 0.$$

Therefore, by applying Rio's inequality in Lemma B.1, for $p \ge 2$, we have

$$\begin{aligned}\|\theta_n - \theta^\star\|_p^2 &\le \|\theta_{n-1} - \theta^\star - \eta_n\nabla F(\theta_{n-1})\|_p^2 + (p-1)\eta_n^2\|\nabla f(\theta_{n-1},\xi_n) - \nabla F(\theta_{n-1})\|_p^2 \\ &=: \mathbb{I}_1 + \mathbb{I}_2.\end{aligned}$$

We shall bound the two parts $\mathbb{I}_1$ and $\mathbb{I}_2$ separately. For the first part $\mathbb{I}_1$, note that $\nabla F(\theta^\star) = 0$ and by the triangle inequality, we have

$$\begin{aligned}\mathbb{I}_1 &= \|\theta_{n-1} - \theta^\star - \eta_n\nabla F(\theta_{n-1})\|_p^2 \\ &= \Big\|\big\langle\theta_{n-1} - \theta^\star, \theta_{n-1} - \theta^\star\big\rangle - 2\eta_n\big\langle\theta_{n-1} - \theta^\star, \nabla F(\theta_{n-1}) - \nabla F(\theta^\star)\big\rangle \\ &\quad + \eta_n^2\big\langle\nabla F(\theta_{n-1}) - \nabla F(\theta^\star), \nabla F(\theta_{n-1}) - \nabla F(\theta^\star)\big\rangle\Big\|_{p/2} \\ &\le \Big\|\big\langle\theta_{n-1} - \theta^\star, \theta_{n-1} - \theta^\star\big\rangle - 2\eta_n\big\langle\theta_{n-1} - \theta^\star, \nabla F(\theta_{n-1}) - \nabla F(\theta^\star)\big\rangle\Big\|_{p/2} \\ &\quad + \eta_n^2\big\|\nabla F(\theta_{n-1}) - \nabla F(\theta^\star)\big\|_p^2.\end{aligned}$$

By applying Assumption 2.1 to the first term and Assumption 2.3 to the second term, we can obtain

$$\mathbb{I}_1 \le (1 - 2\eta_n\mu + \eta_n^2 L_p^2)\|\theta_{n-1} - \theta^\star\|_p^2.$$

Regarding the second part $\mathbb{I}_2$, since $\nabla F(\theta^\star) = 0$, we have

$$\begin{aligned}&\|\nabla f(\theta_{n-1}, \xi_n) - \nabla F(\theta_{n-1})\|_p \\ &\le \|\nabla f(\theta_{n-1}, \xi_n) - \nabla f(\theta^\star, \xi_n)\|_p + \|\nabla F(\theta_{n-1}) - \nabla F(\theta^\star)\|_p + \|\nabla f(\theta^\star, \xi_n)\|_p.\end{aligned}$$

Hence, by Assumption 2.3, we can achieve

$$\|\nabla f(\theta_{n-1}, \xi_n) - \nabla F(\theta_{n-1})\|_p^2 \le 6L_p^2\|\theta_{n-1} - \theta^\star\|_p^2 + 3\|\nabla f(\theta^\star, \xi_n)\|_p^2.$$

Combining results from $\mathbb{I}_1$ and $\mathbb{I}_2$, we can obtain

$$\|\theta_n - \theta^\star\|_p^2 \le (1 - 2\eta_n\mu + (6p-5)\eta_n^2 L_p^2)\|\theta_{n-1} - \theta^\star\|_p^2 + 3(p-1)\eta_n^2\|\nabla f(\theta^\star, \xi_n)\|_p^2.$$

This can directly lead to the desired inequality. □

C Additional simulations

In this section, we collect some additional simulation studies, complimenting the numerical exercises of Section 3. In particular, in Appendix C.1, we begin with constant learning rates ($\eta_t \equiv \eta$), which serve as a baseline and exhibit the expected bias-variance tradeoff, with final error scaling linearly in η. Appendix C.2 turns to polynomially decaying learning rates, $\eta_t = \eta_0 t^{-\beta}$, and demonstrates how different values of $\beta \in (1/2, 1)$ influence the tradeoff between fast initial descent and long-run variance control, consistent with the structure of Theorem 2.1. Moreover, Appendix C.3 examines alternating schedules that interleave two polynomial decay rates or combine a constant rate with a decaying one, highlighting how the more aggressive schedule tends to dominate long-run behavior. Finally, Appendix C.4 includes a detailed comparison study of the different step-sizes considered in this paper, with particular focus into the initial phase, as well as asymptotic behavior upon convergence.

C.1 Constant learning rate

To ground our analysis, we start with the familiar case of constant learning rates—long favored for their simplicity, but known to encode a fundamental tradeoff. We test three fixed values, $\eta = 0.1, 0.05$, and 0.01, and track their performance over 10^4 SGD iterations. The patterns are predictable but instructive: larger step sizes yield faster initial progress, yet settle into higher-variance regimes; smaller ones move more cautiously, but converge closer to the optimum with lower final error.

Because the early dynamics often involve rapid error reduction—sometimes several orders of magnitude—we focus the MSE plot on later stages: starting from $t = 100$ for $\eta = 0.1$ and 0.05, and from $t = 500$ for $\eta = 0.01$. This lets us zoom in on the asymptotic behavior, where the long-term effects of each learning rate become more clearly visible.

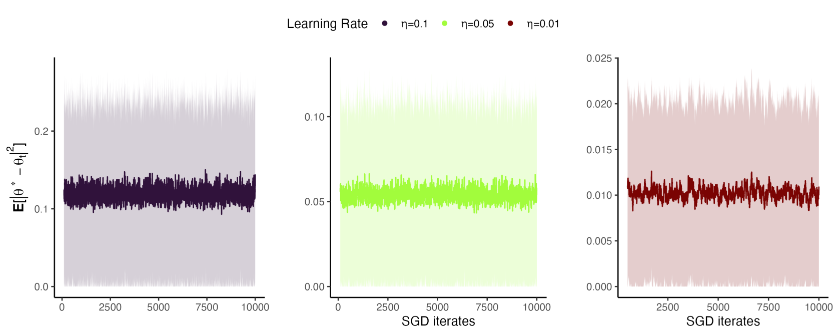

Figure 7: MSE estimates over 500 SGD runs (10^4 steps) with constant learning rates.

The empirical trajectories in Figure 7 reflect this tradeoff. Larger learning rates lead to faster early reduction in error but exhibit higher variance and stabilize farther from the optimum. In contrast, smaller learning rates result in slower progress but achieve significantly lower terminal error.

To complement this, Figure 8 plots the final mean squared error against a dense grid of fixed learning rates ranging from $\eta = 0.01$ to $\eta = 0.1$. The trend is unmistakable: terminal MSE scales linearly with η, matching the $O(\eta)$ variance bound predicted by theory. This reinforces the fundamental tension in fixed-rate SGD: speed comes at the cost of noise, and there's no single value of η that avoids the tradeoff entirely.

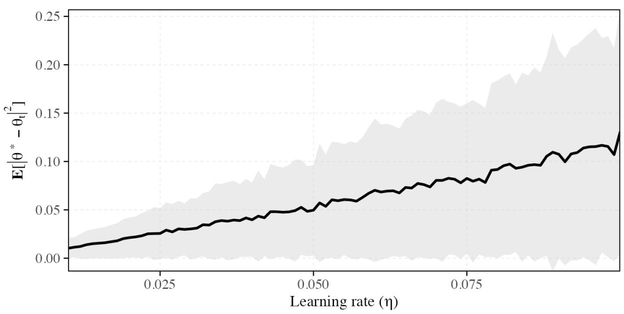

Figure 8: Plot of the terminal MSE estimate averaged over 500 SGD runs for 10^4 steps, for constant values for η between 0.01 and 0.1.

C.2 Polynomially decaying learning rate

We now turn to the classical regime of polynomially decaying learning rates. These schedules take the form $\eta_t = \eta_0 t^{-\beta}$, where $\eta_0 > 0$ and $\beta \in (0, 1]$ controls the rate of decay. The general non-asymptotic error bound given in Theorem 2.1 applies to this setting directly, and allows us to capture the tradeoffs these schedules induce. In particular, when $\beta > \frac{1}{2}$, the sum $\sum_{t=1}^{\infty} \eta_t^2$ converges, ensuring that variance contributions decay to zero; whereas the condition $\beta < 1$ guarantees $\sum_{t=1}^{\infty} \eta_t = \infty$, which is necessary for the bias to vanish. These facts jointly imply that SGD with $\beta \in (\frac{1}{2}, 1)$ is consistent and convergent, with error rates depending sensitively on the balance between the two terms in the bound.

To explore these effects empirically, we simulate SGD with learning rates of the form $\eta_t = \eta_0 t^{-\beta}$ using two values of β: 0.505 and 0.75, each tested with base rates $\eta_0 = 0.1, 0.05, 0.01$. Figure 9 shows the mean squared error over time for these settings, averaged over 500 independent runs. The results highlight the central tradeoff: smaller values of β yield faster initial descent but larger long-run fluctuations, while larger β dampen early progress but reduce terminal error. This qualitative pattern matches the structure of Theorem 2.1, in which the exponential forgetting term dominates early on, and the variance accumulation term becomes decisive in the long run.

Because the early dynamics often involve rapid error reduction—sometimes several orders of magnitude—we focus the MSE plot on later stages: starting from $t = 1000$ for $\eta = 0.1, \beta = 0.505$, $t = 4000$ for $\eta = 0.05, \beta = 0.505$ and $t = 5000$ for $\beta = 0.75$. This lets us zoom in on the asymptotic behavior, where the long-term effects of each learning rate become more clearly visible.

An interesting feature of these experiments is the relative importance of β compared to η_0. While smaller base rates do modestly influence early error and variance, the dominant effect stems from the decay exponent. The case $\beta = 0.505$, being just above the variance threshold, achieves a strong balance between speed and consistency—converging faster than $\beta = 0.75$ while eventually achieving comparably low error. This behavior reflects the non-asymptotic structure predicted by Theorem 2.1, which separates the error into two components: a bias decay term, of the form $\exp(-c \sum_{t=1}^{n} \eta_t)$, and a cumulative variance term, of the form $\sum_{j=1}^{n} \eta_j^2 \exp(-c \sum_{t=j+1}^{n} \eta_t)$. For polynomial learning rates, the bias term vanishes polynomially in n, and the variance term converges to zero if $\beta > \frac{1}{2}$, but only slowly. These dynamics explain the empirical behavior observed: $\beta = 0.505$ gives faster early convergence due to slower bias decay, while $\beta = 0.75$ leads to more effective long-run averaging, with reduced variance. The simulations thus concretely illustrate the tension between forgetting and fluctuation that the theory encodes, and validate the asymptotics of polynomial decay schedules as captured by Theorem 2.1.

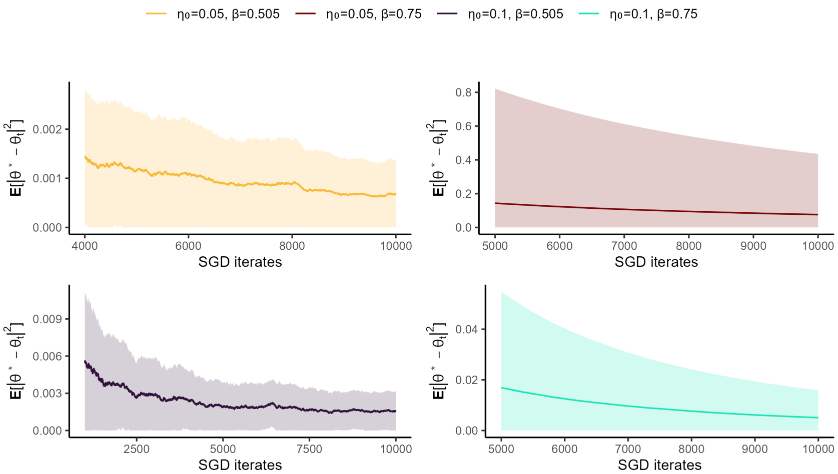

Figure 9: MSE estimates over 500 SGD runs (10^4 steps) with polynomially decaying learning rates.

C.3 Alternating polynomially decaying learning rate

Building on our observations from the previous experiments, we explore combining the strengths of both polynomially decaying and cyclical approaches. In particular, the cyclical schedules demonstrate superior initial convergence through their ability to take larger steps early in optimization, while polynomial decay provides better asymptotic properties and lower final error. To this end, we tried two forms of dual schedules:

1. Set a base rate η_0 and alternate between two polynomial decay rates or a polynomial decay rate and a constant rate. Using two values for η_0 $(0.1, 0.05)$, a constant rate and two exponents for polynomial decay rates $\beta = 0.505, 0.75$ gives us six total combinations. In all of these cases, we present the plots for the mixture of a decay schedule and a constant rate from $t = 500$ and for the mixture of two polynomial decay schedules from $t = 1000$, again to observe the general trend rather than the initial decay by several orders of magnitude in a short amount of time.
2. Set a fixed polynomial decay rate and alternate between two base rates, a "large" one and a "small" one. Plotted from $t = 2000$ for the same consideration as above.

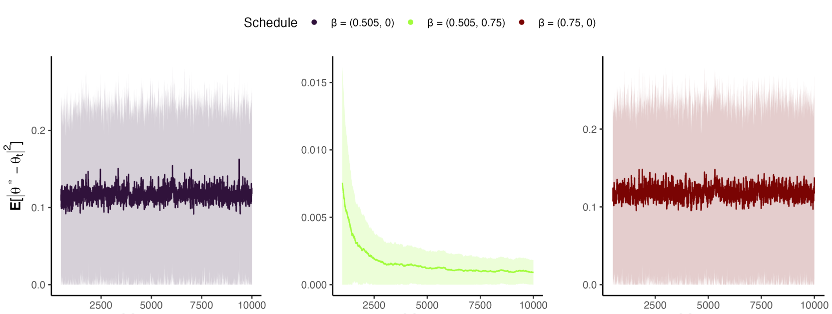

Figure 10: Plot of the MSE estimate averaged over 500 SGD runs for 10^4 steps.

For both $\eta_0 = 0.1$ and $\eta_0 = 0.05$, if the constant learning rate schedule is one of the two included, the process seems to behave in the same way over the long run. However, when we alternate between $\beta = 0.505$ and $\beta = 0.75$, we get an outcome more similar to just selecting $\beta = 0.505$. In all of these cases, the larger learning rate plays the dominant role in determining the convergence rate and terminal error of the process.

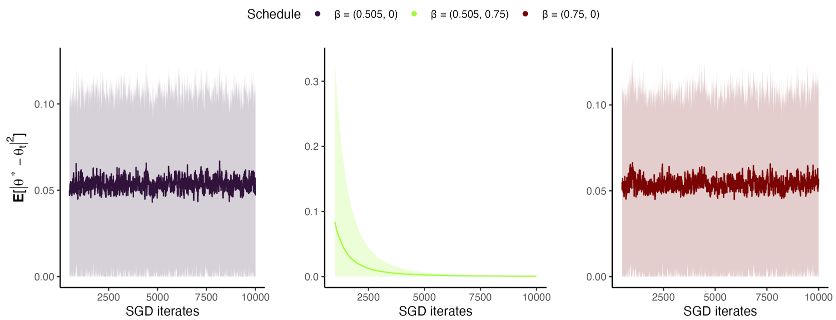

Figure 11: Plot of the MSE estimate averaged over 500 SGD runs for 10^4 steps.

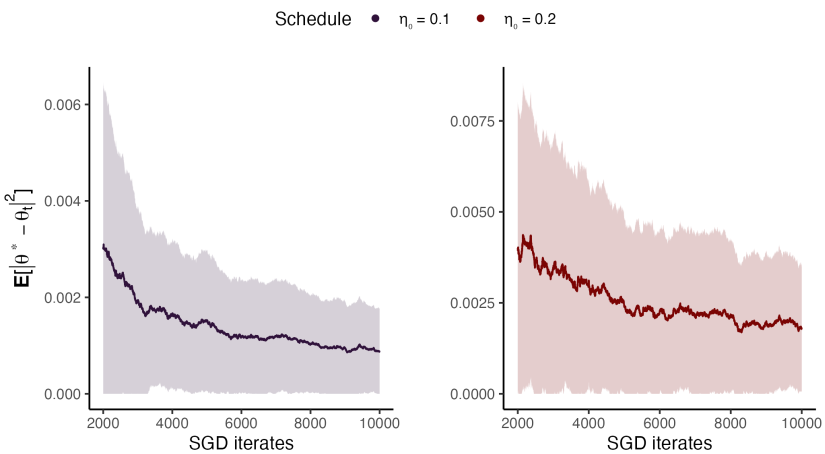

Figure 12: Plot of the MSE estimate averaged over 500 SGD runs for 10^4 steps.

The two plots look almost identical and even have error rates that seem to scale sublinearly with η_0. Compared with the earlier plots in this section and our observations in section 3.3, we have empirical evidence that in polynomial decay and mixed polynomial decay regimes, the choice of the rate of decay β is far more important than the choice of the base rate η_0.

C.4 Comparison of different learning rates

In this section, we numerically investigate the comparative performance of the different learning rate schedules analyzed in this paper. Specifically, we plot the estimates of $\mathbb{E}[|\theta^* - \theta_n|^2]$ against n for $n = 10^4$, and five different learning rate schedules corresponding to Sections 2.3-3.3, along with the additional studies in the appendix C. For careful comparison of both the effect of initialization as well as behavior at convergence, we investigate the MSE of θ_n for the initial 100 iterates, in Figure 13, as well as the final 100 iterates, in Figure 14.

In the early stages of the process, the models that have a non-decaying rate (constant learning rate and cosine schedule), as well as the `Linear-D2Z` model, which as discussed in Sections 2.3-3.3, exhibits superlinear convergence speed away from initialization, while the polynomial decay and mixed polynomial decay model both move more slowly. This is almost reversed in the later stages of the process - while all five models have converged to a small error, `Linear-D2Z` and the two polynomial decay-based models end up with a far smaller error estimate compared to the constant learning rate and the cosine schedule, which demonstrate a consistent bias. This demonstrates well-known theoretical results - larger learning rates converge faster but at the cost of a larger terminal bias. `Linear-D2Z` manages to reach an acceptably small error rather quickly and also enjoys a small terminal error due to being defined in a way that starts off with large step sizes that then taper off quickly, reinforcing its view as a "best-of-both-worlds" learning schedule.

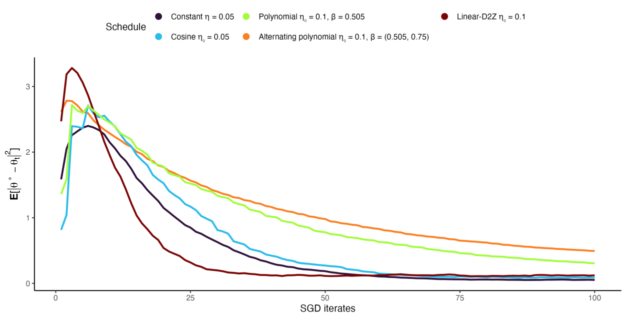

Figure 13: Plot of the evolution of the MSE estimate for the first 100 SGD iterations for five learning rate schedules.

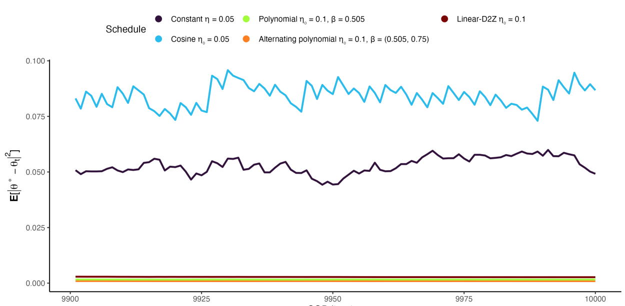

Figure 14: Plot of the evolution of the MSE estimate for the last 100 SGD iterations for five learning rate schedules.