



On the Performance of Stochastic Gradient Methods with Momentum in Time-Varying Regimes

Enrico Bernardi¹ · Alberto Lanconelli¹ · Christopher S. A. Lauria¹

Received: 24 September 2025 / Accepted: 29 August 2026
© The Author(s) 2026

Abstract

We explore how Stochastic Gradient methods with momentum perform within a time-varying framework by establishing bounds on their tracking errors, specifically focusing on quadratic cases. Notably, we find that momentum methods achieve, in *high drift-to-noise regimes*, i.e., when the rate of change of the dynamic optimum prevails on the variance of the gradient noise, smaller neighborhood of convergence compared to stochastic gradient descent. To the best of our knowledge, this is the first proof that a momentum method can improve upon stochastic gradient descent's tracking error bounds in a time-varying setting. We then investigate, for a given learning rate, the optimal choice of the momentum parameter that minimizes the tracking error bound.

Keywords Stochastic gradient descent · Stochastic heavy ball · Nesterov accelerated · Optimization

Mathematics Subject Classification 65K05 · 62F12

1 Introduction

1.1 Optimization in the Static Case

Among the numerous optimization techniques available, Stochastic Gradient Descent (SGD) stands out for its simplicity, scalability, and effectiveness in handling large-scale data and high dimensional objective functions. Introduced by [23], SGD minimizes

The general content of the manuscript is not submitted, accepted, or published elsewhere, including conference proceedings.

✉ Christopher S. A. Lauria
christopher.lauria2@unibo.it

Enrico Bernardi
enrico.bernardi@unibo.it

Alberto Lanconelli
alberto.lanconelli2@unibo.it

¹ Dipartimento di Scienze Statistiche Paolo Fortunati, Università di Bologna, Bologna, Italy

an objective function by iteratively updating parameters using noisy estimates of the gradient, often computed over small, randomly sampled subsets of data. Formally, given an objective function of the form

$$l(x) := \mathbb{E}_X[f(x, X)], \quad x \in \mathbb{R}^d, \quad (1)$$

where X is a random variable representing the data distribution, and $f(x, X)$ is a stochastic function that depends on both the decision variable x and the random variable X , the SGD algorithm is defined by the following recursive equation

$$x_{t+1} = x_t - \alpha \nabla_x f(x_t, X_{t+1}), \quad (2)$$

where $\alpha > 0$ is the *learning rate* (or step size) while X_{t+1} represents a random sample drawn from the data distribution X . Thus, under mild regularity conditions, $\nabla_x f(x_t, X_{t+1})$ acts as an unbiased stochastic estimator of the gradient of the objective function, namely

$$\mathbb{E}_{X_{t+1}}[\nabla_x f(x, X_{t+1})] = \nabla_x l(x), \quad \text{for all } t \in \mathbb{N} \cup \{0\} \text{ and } x \in \mathbb{R}^d; \quad (3)$$

this property ensures that, on average, the stochastic gradient aligns with the true gradient, enabling the algorithm to converge in the appropriate setting.

Currently, the most widely used algorithms in practical applications are *accelerated* or *momentum methods*. Renowned for their effectiveness, these algorithms have been empirically found to increase the convergence speed. Indeed, momentum based gradient descent algorithms are well-known for their ability to enhance convergence rates in static, deterministic settings [20, 21]. As a foundational approach for modern adaptive gradient methods like Adam [15], understanding the properties of momentum algorithms has substantial practical implications.

Historically, the first momentum method to be introduced was the deterministic Heavy Ball (HB) method [21]. This method is defined by the recursive equation

$$x_{t+1} = x_t - \alpha \nabla_x \ell(x_t) + \gamma(x_t - x_{t-1}), \quad (4)$$

where $\gamma > 0$ is known as the *momentum parameter*. The initial convergence results for HB by [21] demonstrated its accelerated convergence (in a deterministic setting) for strongly convex quadratics. A global convergence analysis of HB was provided by [12], showing that the method applies to convex problems with a Lipschitz-continuous gradient, even without assuming strong convexity. Recently, [13] demonstrated that beyond quadratic cases, the HB method cannot achieve an accelerated worst-case convergence rate on the standard class of smooth, strongly convex problems in higher dimensions.

The stochastic version of the HB method, commonly referred to as the Stochastic Heavy Ball (SHB) method, extends its deterministic counterpart (4) to stochastic optimization settings:

$$x_{t+1} = x_t - \alpha \nabla_x f(x_t, X_{t+1}) + \gamma(x_t - x_{t-1}). \quad (5)$$

[11] established almost sure convergence of (5) for general non-convex coercive functions, also examining convex and strongly convex potentials. Their work provides non-asymptotic results for the SHB method. Some limitations were identified by [14], who demonstrated that simple problem instances exist where SHB fails to outperform SGD, even with optimal parameter settings. Subsequently, [18] established the first linear convergence rate for SHB on quadratic objectives.

Another renowned momentum method is Nesterov's Accelerated Gradient descent (NAG) that was developed by [20] to achieve an accelerated convergence rate for smooth convex optimization problems. The NAG method can be written in a single recursive equation as

$$x_{t+1} = x_t - \alpha \nabla_x \ell(x_t + \gamma(x_t - x_{t-1})) + \gamma(x_t - x_{t-1}),$$

where $\gamma > 0$ is again the momentum parameter.

Similarly as before, in the stochastic setting, we can define its counterpart which we call the Stochastic Nesterov's Accelerated (SNA) method. It can be written as

$$x_{t+1} = x_t - \alpha \nabla_x f(x_t + \gamma(x_t - x_{t-1}), X_{t+1}) + \gamma(x_t - x_{t-1}). \quad (6)$$

In the stochastic approximation framework, [2] showed that the SNA method converges to a neighborhood of the optimal point at the same accelerated rate as in the deterministic case. However, in the finite-sum setting, they proved that SNA may diverge with the standard choice of learning rate and momentum unless additional problem-specific conditions are met. Meanwhile, [27] studied the SNA method with a constant learning rate and diminishing momentum, demonstrating linear convergence when the gradient variance vanishes at a stationary point.

1.2 Optimization in the Dynamic Case

Traditional optimization techniques typically assume a static environment, where parameters and constraints remain constant over time. However, real-world systems are often dynamic, with underlying parameters that fluctuate as conditions change. Static optimization approaches can fall short in capturing these variations, potentially leading to suboptimal solutions in practice. As a result, there has been a growing interest in time-varying optimization, a paradigm that explicitly incorporates temporal changes into the optimization framework. Formally, the goal is to find or track the sequence of time-dependent optima defined as:

$$x_t^* = \arg \min_{x \in \mathcal{X}} \mathfrak{l}(t, x), \quad t \in \mathbb{N},$$

where $\mathcal{X} \subset \mathbb{R}^d$ is typically a convex set and $\{\mathfrak{l}(t, \cdot)\}_{t \in \mathbb{N}}$ denotes a sequence of objective functions, often assumed to be smooth and convex. Since we will be analyzing the stochastic setting, we consider sequences of the form

$$\mathfrak{l}(t, x) := \mathbb{E}_{X_t}[f(x, X_t)], \quad t \in \mathbb{N}, x \in \mathbb{R}^d; \quad (7)$$

here the data distribution X_t is allowed to change with t (cf. (1)).

This formulation is closely connected to adaptive stochastic approximation. A recursive estimator with a non-vanishing learning rate can continually react to a moving parameter, but faster adaptation also transmits more observation noise. The resulting trade-off between tracking ability and noise rejection is central in adaptive filtering and online system identification [17]. It also appears in continual learning, where a learner processes a nonstationary data stream and must remain plastic enough to follow distributional change while retaining stable performance on previously encountered data. Distributional-drift models provide an optimization-based description of this aspect of continual learning [9].

The time-varying setting has been the focus of extensive study, with foundational contributions by Polyak [21, Chapter 6], who was among the first to formally address the scenario in which the optimum evolves over time, a context he termed the non-stationary case. Polyak demonstrated that by applying a gradient descent algorithm using, at each time t , the gradient of the corresponding function in the objective function sequence, it was possible to track a minimum that evolved smoothly over time. Subsequently, numerous other works emerged, providing theoretical tracking error bounds for gradient-based algorithms [22], [5], [10].

In the context of time-varying systems, Simonetto et al. [25] were the first to explore prediction-correction algorithms, which include both a prediction step and a corrective adjustment (typically performed using gradient descent). More recently, Madden et al. [19] analyzed a general first-order framework, establishing a universal lower bound on tracking iterate error. Additionally, they proposed a ‘long-step’ method, demonstrating that it achieves this lower bound up to a fixed constant. For a review of time-varying optimization schemes see also [26].

Several studies have also addressed the *stochastic* time-varying setting. In such cases, the analysis typically assumes access to a stochastic gradient at each time t , which plays the role, as we mentioned above, of an unbiased estimate of the deterministic gradient for the corresponding objective function. Formally, the canonical SGD update rule at each iteration t becomes

$$x_{t+1} = x_t - \alpha \nabla_x f(x_t, X_{t+1}), \quad (8)$$

where now $\nabla_x f(x, X_{t+1})$ is an unbiased estimator of $\nabla_x l(t+1, x)$, i.e.

$$\mathbb{E}_{X_t}[\nabla_x f(x, X_t)] = \nabla_x l(t, x), \quad \text{for all } t \in \mathbb{N}, x \in \mathbb{R}^d. \quad (9)$$

We remark that, while the recursive scheme (8) is identical to (2), the unbiasedness condition (9) differs from (3) in the fact that the law of X_t may vary with t (thus producing a time-varying objective function $l(t, \cdot)$).

In [6], rigorous bounds on the tracking error of stochastic gradient algorithms like the one defined by (8) applied to a time-varying problem are established. In a statistical context, Lanconelli and Lauria [16] investigated tracking errors in stochastic gradient algorithms designed to track a time-varying parameter. This analysis involved comparing the derived error bounds to the variance of other statistical estimators. In [9], novel non-asymptotic convergence guarantees were provided for stochastic algorithms

with iterate averaging. The analysis focuses on establishing bounds that hold both in expectation and with high probability for problems exhibiting concept drift, where the underlying hypothesis may evolve over time.

Although recent work by Sahu et al. [24] studies fixed-momentum SGD for non-stationary strongly convex objectives and shows that momentum can incur an inertia penalty after distribution shifts, to date, relatively little attention has been devoted to the study of momentum algorithms in the time-varying settings. In this paper, we establish tracking error bounds for momentum algorithms within a time-varying framework, focusing on quadratic problems with a time-dependent optimum, where SHB is equivalent to SNA. Specifically for quadratic problems with a time-varying optimum, we show that momentum methods achieve, in *high drift-to-noise regimes*, i.e. when the rate of change of the dynamic optimum prevails on the variance of the gradient noise, smaller error bounds than SGD when using a fixed learning rate. While SHB or SNA do not reduce the convergence neighborhood of SGD in a static setting with a fixed learning rate when optimizing quadratic functions, they do so in a dynamic, time-varying context. To our knowledge, this is the first proof demonstrating that momentum methods can improve upon the tracking error bounds of SGD.

2 Statement of the Main Results

Our analysis focuses on quadratic objective functions, which are widely used in the theoretical study of optimization algorithms with momentum due to their analytical tractability [18], [11]. The quadratic objective functions we will consider can be written in the form

$$\mathfrak{l}(t, x) = \mathbb{E}_{X_t} \left[\frac{(x - X_t)^2}{2} \right], \quad t \in \mathbb{R}, x \in \mathbb{R} \quad (10)$$

and hence, following the notation in (7),

$$f(x, X_t) = \frac{(x - X_t)^2}{2} \quad t \in \mathbb{R};$$

(see [23] for a similar setting). Here, $\{X_t\}_{t \geq 1}$ is a sequence of independent random variables with

$$\mathbb{E}[X_t] = x_t^* \quad \text{and} \quad \mathbb{V}[X_t] = \sigma^2 \quad \text{for all } t \in \mathbb{N};$$

in particular, $\{x_t^*\}_{t \in \mathbb{N}}$ is a sequence of real numbers while σ^2 is a positive constant. Since the optimum of $x \mapsto \mathfrak{l}(t, \cdot)$ in (10) is achieved at x_t^* , the optimization problem related to the objective functions $\mathfrak{l}(t, x)$ is classified as time-varying.

Remark 2.1 For clarity of exposition, we develop our results in the one-dimensional setting. When the covariance matrix of X_t is diagonal the extension to $\mathbb{R}^d$ is direct, all statements carry over coordinate-wise without modification. For a general (full)

covariance matrix, the same conclusions obtains by applying a simple diagonalization argument to the covariance matrix Σ and leveraging the linearity of the recursive equations. Alternately, one can apply an appropriate preconditioning on the gradient (see [1]) —or perform a Newton update. That is, let $X_t \in \mathbb{R}^d$ be a random vector with $\mathbb{E}[X_t] = x_t^*$ and $\text{Cov}(X_t) = \Sigma \succ 0$. Consider the quadratic objective

$$\ell(t, x) = \frac{1}{2} \mathbb{E}[(x - X_t)^\top \Sigma^{-1} (x - X_t)] \quad (11)$$

The gradient and Hessian are

$$\nabla_x \ell(t, x) = \Sigma^{-1} (x - x_t^*), \quad \nabla_x^2 \ell(t, x) = \Sigma^{-1}. \quad (12)$$

Hence, a Newton update with step size $\eta > 0$ is

$$x_{t+1} = x_t - \alpha (\nabla_x^2 \ell(t+1, x_t))^{-1} \nabla_x \ell(x_t, X_{t+1}) = x_t - \alpha \Sigma \Sigma^{-1} (x_t - X_{t+1}) \quad (13)$$

$$= x_t - \alpha (x_t - X_{t+1}). \quad (14)$$

which coincides with the scalar update applied componentwise.

The SHB method has been extensively studied in the context of quadratic objective functions [18], [11], as these functions serve as a baseline model for understanding the behaviour of stochastic optimization algorithms. The focus on quadratic objectives is particularly useful for deriving precise theoretical results, as a more general framework can often lead to significant analytical complexity, especially in the stochastic setting.

For the specific case of quadratic objective functions defined in (10), the general SHB update equation (5) simplifies to

$$x_{t+1} = x_t - \alpha (x_t - X_{t+1}) + \gamma (x_t - x_{t-1}).$$

or equivalently,

$$x_{t+1} = (1 - \alpha)x_t + \alpha X_{t+1} + \gamma (x_t - x_{t-1}).$$

Notably, in the quadratic objective function case we are considering, the SNA method takes the form

$$x_{t+1} = (1 - \alpha)x_t + \alpha X_{t+1} + \gamma (1 - \alpha)(x_t - x_{t-1}).$$

that up to a reparametrization $\tilde{\gamma} = \gamma(1 - \alpha)$ is equivalent to the SHB method. Since we will be focusing on the quadratic case, we will henceforth refer to the algorithm under study simply as momentum methods.

To establish our primary result, we introduce a condition on the time-varying optimal solution we aim to track. This condition, which is commonly employed in the literature on time-varying optimization (see [21, Chapter 6], [26]), requires that the sequence of optima exhibit Lipschitz continuity over time.

Assumption 2.1 (Lipschitz continuity of the optima)

There exists a positive real constant K such that

$$|x_{t+1}^* - x_t^*| \leq K, \quad \text{for all } t \in \mathbb{N}.$$

We are now prepared to present our first main theorem that finds explicit error bounds on momentum methods. To make the result easier to state and interpret, we first introduce the notation used in its formulation. Let

$$E := \{(\alpha, \gamma) \in (0, 1) \times [0, 1) : 0 \leq \gamma < (1 - \sqrt{\alpha})^2\},$$

and, for $(\alpha, \gamma) \in E$, set

$$\begin{aligned} \Lambda(\alpha, \gamma) &:= 1 - \alpha - \gamma, & L(\alpha, \gamma) &:= 1 - \alpha + \gamma, \\ \mu(\alpha, \gamma) &:= \sqrt{L^2(\alpha, \gamma) - 4\gamma}, & \lambda_{1,2}(\alpha, \gamma) &:= \frac{L(\alpha, \gamma) \mp \mu(\alpha, \gamma)}{2}, \\ \rho_-(\alpha, \gamma) &:= \frac{2\gamma}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)}, & \rho_+(\alpha, \gamma) &:= \frac{2\gamma}{\Lambda(\alpha, \gamma) + \mu(\alpha, \gamma)}. \end{aligned}$$

Here λ_1 and λ_2 will be the two characteristic roots of the homogeneous momentum recursion.

Define coefficients

$$\begin{aligned} A(\alpha, \gamma) &:= \frac{\frac{\alpha^4}{\mu^2(\alpha, \gamma)} \rho_+^2(\alpha, \gamma) (1 + \rho_-(\alpha, \gamma))^2}{1 - \lambda_1^2(\alpha, \gamma)} + \frac{\frac{\alpha^4}{\mu^2(\alpha, \gamma)} \rho_-^2(\alpha, \gamma) (1 + \rho_+(\alpha, \gamma))^2}{1 - \lambda_2^2(\alpha, \gamma)} \\ &\quad - 2 \frac{\gamma}{1 - \gamma} \frac{\alpha^2}{\mu^2(\alpha, \gamma)}, \end{aligned} \quad (15)$$

and

$$\begin{aligned} B(\alpha, \gamma) &:= \left(\frac{\alpha}{\mu(\alpha, \gamma)} \frac{|(1 - \alpha)\rho_+(\alpha, \gamma) - \gamma|}{1 - \lambda_1(\alpha, \gamma)} \right. \\ &\quad \left. + \frac{\alpha}{\mu(\alpha, \gamma)} \frac{|(1 - \alpha)\rho_-(\alpha, \gamma) - \gamma|}{1 - \lambda_2(\alpha, \gamma)} \right)^2. \end{aligned} \quad (16)$$

Then we can state the following theorem, whose proof can be found in Sect. 5.

Theorem 2.1 Consider the momentum method

$$\begin{cases} x_{t+1} = (1 - \alpha)x_t + \alpha X_{t+1} + \gamma(x_t - x_{t-1}), & t \geq 2, \\ x_1 = X_1, \quad x_2 = X_2. \end{cases} \quad (17)$$

Under Assumption 2.1, the functions $A(\alpha, \gamma)$ and $B(\alpha, \gamma)$ defined above are analytic in E , and, for every $(\alpha, \gamma) \in E$,

$$\limsup_{t \rightarrow \infty} \mathbb{E}[(x_{t+1} - x_{t+1}^*)^2] \leq \sigma^2 A(\alpha, \gamma) + K^2 B(\alpha, \gamma). \quad (18)$$

The two terms in (18) separate the effects of observation noise and temporal drift: A measures the amplification of the noise variance, whereas B measures the amplification of the squared drift bound. The initialization $x_1 = X_1$, $x_2 = X_2$ does not affect the asymptotic bound. Indeed, two trajectories driven by the same observations but starting from different square-integrable initial values differ only by a solution of the homogeneous momentum recursion. Since both characteristic roots have modulus strictly smaller than one for $(\alpha, \gamma) \in E$, this difference decays geometrically in mean square. Thus initialization affects only the transient behavior and disappears from the asymptotic tracking error.

The special case $\gamma = 0$ is stated explicitly in the following remark.

Remark 2.2 The bound (18) with $\gamma = 0$ aligns with the tracking error bounds established for SGD in [4], i.e.

$$\limsup_{t \rightarrow \infty} \mathbb{E}[(x_{t+1} - x_{t+1}^*)^2] \leq \sigma^2 \frac{\alpha}{2 - \alpha} + K^2 \left(\frac{1 - \alpha}{\alpha} \right)^2,$$

thereby extending and generalizing their results. See also [8] for a similar tracking error bound.

Remark 2.3 Independence across time is used in the proof only to eliminate cross-covariance terms in the stochastic tracking error. For genuinely correlated observations the cross-covariance terms must be retained. Under standard short-range dependence conditions, such as absolutely summable autocovariances, stability still yields a finite asymptotic tracking bound, but the effective noise level depends on the full autocovariance sequence. See [3] for a treatment on how to proceed with correlated observations.

Due to the complexity of the expressions in the upper bound (18), analytically proving that the introduction of a momentum term improves the tracking error bounds is not straightforward. To address this, we now present an asymptotic expansion of the functions in (18) as γ approaches zero. For the proof see Sect. 6.

Theorem 2.2 The functions $A(\alpha, \gamma)$ and $B(\alpha, \gamma)$ in (18) have asymptotic developments, as γ goes to 0, given by

$$A(\alpha, \gamma) = \frac{\alpha}{2 - \alpha} + \frac{2\alpha(1 - \alpha)}{(2 - \alpha)^2} \gamma + \frac{2\alpha(1 + (1 - \alpha)^2)}{(2 - \alpha)^3} \gamma^2 + \mathcal{O}(\gamma^3),$$

and

$$B(\alpha, \gamma) = \left(\frac{1 - \alpha}{\alpha} - \frac{1}{\alpha} \gamma + \left(\frac{1}{1 - \alpha} + 2 \left(\frac{\alpha}{1 - \alpha} \right)^2 \right) \gamma^2 + \mathcal{O}(\gamma^3) \right)^2. \quad (19)$$

Combining Theorems 2.1 and 2.2 we can now prove that in *high drift-to-noise regimes*, i.e. for $\frac{\sigma^2}{K^2} \leq 1$, the algorithm with momentum performs better than the corresponding SGD scheme. This means that there exists a positive γ such that the upper bound (18) is smaller than the one with $\gamma = 0$.

Corollary 2.1 *If $\frac{\sigma^2}{K^2} \leq 1$, then for any $\alpha \in (0, 1)$ there exists $\gamma > 0$ such that*

$$\frac{\partial}{\partial \gamma} \left[\sigma^2 A(\alpha, \gamma) + K^2 B(\alpha, \gamma) \right] \Big|_{\gamma=0} < 0. \quad (20)$$

Consequently, for the values of α and $\gamma > 0$ such that condition (20) holds, we have

$$\sigma^2 A(\alpha, \gamma) + K^2 B(\alpha, \gamma) < \sigma^2 A(\alpha, 0) + K^2 B(\alpha, 0).$$

Proof Using the asymptotic expansions of $A(\alpha, \sigma)$ and $B(\alpha, \sigma)$ from Theorem 2.2 it is straightforward to see that condition (20) holds true if

$$\frac{\sigma^2}{K^2} < \frac{(2-\alpha)^2}{\alpha^3}. \quad (21)$$

Such inequality is always true since $\frac{\sigma^2}{K^2} \leq 1$ while $\frac{(2-\alpha)^2}{\alpha^3} > 1$ for all $\alpha \in (0, 1)$. $\square$

According to Corollary 2.1, for any $\alpha \in (0, 1)$, there exists $\gamma > 0$ such that the error bound of the momentum method is smaller than that of the corresponding SGD algorithm.

2.1 Interpretation of the Bound

For a general target sequence $\{x_t^*\}_{t \in \mathbb{N}}$ satisfying

$$|x_{t+1}^* - x_t^*| \leq K,$$

the right-hand side of the bound in Theorem 2.1 is a uniform guarantee over all admissible drift sequences. The assumption controls the size of each increment but not its temporal pattern; consequently, the bound need not be equal to the true asymptotic tracking error for a particular sequence. Thus, in this general setting, Corollary 2.1 establishes that momentum can improve the guaranteed tracking neighborhood, rather than asserting that it decreases the true tracking error for every admissible drift. For a linear drift, however, the true asymptotic error can be computed explicitly and compared directly with the bound revealing that the momentum method improves the exact tracking error in such a setting.

Indeed, suppose that

$$x_t^* = c + bt, \quad X_t = x_t^* + \varepsilon_t,$$

where the random variables ε_t are independent, centered, and have variance σ^2 . Let $e_t = x_t - x_t^*$. Since $x_{t+1}^* - x_t^* = b$, the tracking error of SHB in this case satisfies

$$e_{t+1} = (1 - \alpha + \gamma)e_t - \gamma e_{t-1} - (1 - \alpha - \gamma)b + \alpha \varepsilon_{t+1}.$$

For $0 < \alpha < 1$ and $0 \leq \gamma < 1$, this is a stable second-order autoregressive, or AR(2), recursion: e_{t+1} is a linear function of the two preceding errors e_t and e_{t-1} , a constant forcing term, and the new innovation $\alpha \varepsilon_{t+1}$; see, for example, [7, Chapter 3]. Here stability means that the two roots of the characteristic polynomial

$$p(r) = r^2 - \phi_1 r - \phi_2, \quad \phi_1 = 1 - \alpha + \gamma, \quad \phi_2 = -\gamma,$$

lie strictly inside the unit disk. For an AR(2) recursion this is equivalent to

$$1 - \phi_1 - \phi_2 > 0, \quad 1 + \phi_1 - \phi_2 > 0, \quad 1 + \phi_2 > 0.$$

In the present case,

$$1 - \phi_1 - \phi_2 = \alpha > 0, \quad 1 + \phi_1 - \phi_2 = 2 - \alpha + 2\gamma > 0, \quad 1 + \phi_2 = 1 - \gamma > 0.$$

Consequently, the effect of the initial conditions decays geometrically, and the recursion admits a unique covariance-stationary solution with finite variance. The stationary mean is

$$m_\gamma = -\frac{1 - \alpha - \gamma}{\alpha} b.$$

And the Yule–Walker equations give the stationary variance

$$v_\gamma = \sigma^2 \frac{\alpha(1 + \gamma)}{(1 - \gamma)[2(1 + \gamma) - \alpha]}.$$

Thus the initialization vanishes asymptotically and we have

$$\lim_{t \rightarrow \infty} \mathbb{E}[(x_t - x_t^*)^2] = \sigma^2 \frac{\alpha(1 + \gamma)}{(1 - \gamma)[2(1 + \gamma) - \alpha]} + b^2 \left(\frac{1 - \alpha - \gamma}{\alpha} \right)^2. \quad (22)$$

Where equation (22) follows from $\mathbb{E}[e_t^2] = v_\gamma + m_\gamma^2$.

One can notice that the exact error for linear drift and the upper bound of Theorem 2.1 agree at $\gamma = 0$ and have the same first derivative:

$$\begin{aligned} \left. \frac{\partial}{\partial \gamma} [\sigma^2 A(\alpha, \gamma) + b^2 B(\alpha, \gamma)] \right|_{\gamma=0} &= \left. \frac{\partial}{\partial \gamma} \lim_{t \rightarrow \infty} \mathbb{E}[(x_t - x_t^*)^2] \right|_{\gamma=0} \\ &= 2(1 - \alpha) \left[\frac{\sigma^2 \alpha}{(2 - \alpha)^2} - \frac{b^2}{\alpha^2} \right]. \end{aligned}$$

This derivative is negative whenever

$$\frac{\sigma^2}{b^2} < \frac{(2-\alpha)^2}{\alpha^3}.$$

In particular, the condition $\sigma^2/b^2 \leq 1$ guarantees that a sufficiently small positive momentum parameter decreases the exact asymptotic tracking error relative to SGD in the linear-drift model.

While the momentum parameter γ has been widely recognized for enhancing the speed of convergence in various settings, this result, to the best of our knowledge, represents the first instance where a stochastic momentum method has been shown to improve the *neighborhood* of convergence in a time-varying setting.

2.2 Beyond Quadratic Objectives: Local Mechanism and Limitations

The exact results above use more than convexity: they rely on the fact that, after subtracting the moving optimum, the algorithm is a time-invariant linear recursion. This may not be the case for a general smooth objective.

Consider a sequence of differentiable objectives $\ell_t : \mathbb{R}^d \rightarrow \mathbb{R}$, with minimizers x_t^* . Using similar notation to the one used in the proofs

$$e_t := x_t - x_t^*, \quad d_{t+1} := x_{t+1}^* - x_t^*, \quad H_{t+1} := \nabla^2 \ell_{t+1}(x_{t+1}^*),$$

a Taylor expansion around x_{t+1}^* gives

$$\nabla \ell_{t+1}(x_t) = H_{t+1}(e_t - d_{t+1}) + r_{t+1}.$$

Where we are calling r_{t+1} the nonlinear Taylor remainder. Consequently, the tracking error of a stochastic Heavy Ball update satisfies

$$e_{t+1} = (I - \alpha H_{t+1} + \gamma I)e_t - \gamma e_{t-1} + \gamma d_t + (\alpha H_{t+1} - I)d_{t+1} - \alpha \xi_{t+1} - \alpha r_{t+1}. \quad (23)$$

Equation (23) makes precise which part of the quadratic mechanism survives. Its linear part is again a second-order tracking system driven by drift and stochastic noise. The final term is absent for quadratic objectives and prevents the proof of Theorem 2.1 from being an exact global analysis in the general nonquadratic case. If the Hessian is ρ -Lipschitz in a neighborhood of the optimum, then

$$\|r_{t+1}\| \leq \frac{\rho}{2} \|x_t - x_{t+1}^*\|^2 = \frac{\rho}{2} \|e_t - d_{t+1}\|^2.$$

And the nonlinear remainder is second order in the distance from the point where the gradient is linearized. From this point additional conditions may be imposed on r_{t+1} to attempt to extend the results presented in this paper; however, such extensions are beyond the scope of the present work.

Remark 2.4 Suppose that we have smooth strongly convex objective functions such that

$$\mu I \preceq H_t \preceq LI$$

uniformly in t , and that the Hessian is Lipschitz. If the parameters (α, γ) give the linearized recursion a uniform stability margin, that is, perturbations decay geometrically at a rate bounded away from one, uniformly in time and if drift and noise keep the iterates in a sufficiently small neighborhood of x_t^* , then the Taylor remainder is of higher order than the linear terms. In this regime one expects the quadratic bias–variance mechanism to remain predictive. A rigorous extension would additionally require controlling higher moments of the tracking error and showing that the nonlinear remainder can be absorbed by the stability margin but strong convexity and smoothness alone do not provide this conclusion globally for a fixed Heavy Ball parameter.

Remark 2.5 In the case of an ill-conditioned Hessian a Newton-type update may be used in place of a standard gradient-descent step, which can be interpreted as preconditioning the gradient with an appropriate matrix. Preconditioning replaces the stochastic gradient g by Pg , using a positive definite matrix P to rescale directions having different curvatures. In our multidimensional quadratic setting,

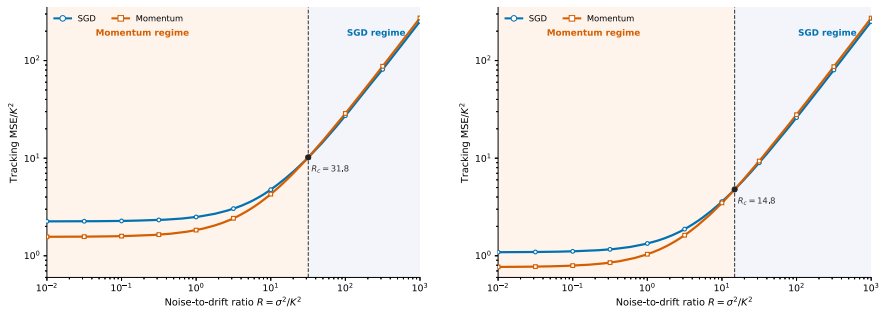
$$g(x, X_t) = \Sigma^{-1}(x - X_t), \quad H = \nabla^2 \ell = \Sigma^{-1}.$$

The Newton step preconditioner $P = H^{-1} = \Sigma$ therefore gives $Pg(x, X_t) = x - X_t$. Hence the effective Hessian is the identity and the preconditioned SHB recursion is precisely the scalar recursion applied componentwise. Without preconditioning, diagonalizing $H = Q \operatorname{diag}(\lambda_1, \dots, \lambda_d) Q^\top$ gives effective learning rate $\alpha \lambda_i$ in the i -th direction. If $\lambda_{\max}/\lambda_{\min}$ is large, stability in the steep directions constrains α , leaving slow tracking and potentially large lag in the flat directions. Momentum tuned to improve the latter may instead cause oscillations or amplify noise in the former. Thus, without preconditioning, performance depends on the Hessian spectrum and on the directions of drift and noise. To address this difficulty, one may use direction-dependent learning rates $\alpha = (\alpha_1, \dots, \alpha_d)$, with different learning rates assigned to each dimension.

Remark 2.6 In the case of nonconvex objectives there may be several competing minima, indefinite curvature, bifurcations, or disappearing minimizer branches. In such a setting the target x_t^* may not be unique or continuous, and momentum may cause transitions between basins. Thus, the present tracking-error theory does not address the general nonconvex case.

3 Numerical Experiments

We now examine the main qualitative predictions of the theory numerically. We first test the drift–noise transition and the dependence on the momentum parameter in the one-dimensional quadratic model. We then observe the effects of dimension in a



(a) **Linear drift.** For $x_t^* = a + t$, momentum has lower error below $R_c \simeq 31.8$, while SGD has lower error above the crossover.
 (b) **Nonlinear drift.** For $x_t^* = A \sin(2\pi(t + 1/2)/64)$, with $A = [2 \sin(\pi/64)]^{-1}$, the crossover is $R_c \simeq 14.8$.

Fig. 1 Drift–noise phase transition. Errors are normalized by K^2 . Solid curves are the exact asymptotic MSEs and markers are estimates from 3,000 paired Monte Carlo runs. Momentum reduces drift-induced lag but has a larger stochastic-variance coefficient, giving a transition from a drift-dominated momentum regime to a noise-dominated SGD regime. The sufficient condition $R \leq 1$ used in the theoretical result lies inside the observed momentum regime for both trajectories. The different crossover points also show that a bound on the one-step drift alone does not encode its full temporal structure

multidimensional quadratic and, finally, test whether the same phenomenon persists for a nonquadratic logistic regression objective. In all experiments, SGD and SHB are evaluated on paired realizations of the stochastic inputs.

3.1 Quadratic Objectives

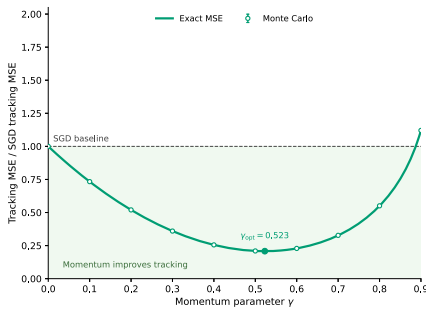
For the first three experiments we use

$$\ell_t(x) = \frac{1}{2} \mathbb{E} \|x - X_t\|^2, \quad X_t = x_t^* + \varepsilon_t,$$

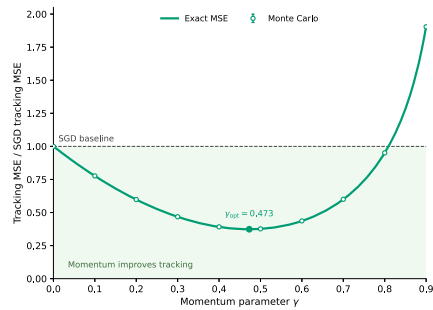
with independent Gaussian observation noise. The solid curves below are the exact asymptotic errors obtained from analytically solving the corresponding linear recursions for the error $e_t := x_t - x_t^*$ as in Subsect. 2.2. While the markers are produced by Monte Carlo estimates.

Drift–noise phase transition. We begin with the central prediction that momentum is beneficial when drift dominates stochastic noise, whereas SGD becomes preferable when noise is sufficiently large. We fix $\alpha = 0.4$, $\gamma = 0.1$, and normalize the targets so that $\sup_t \|x_{t+1}^* - x_t^*\| = K = 1$. We then vary $R = \sigma^2/K^2$ for a constant-velocity target and for a bounded sinusoidal target.

Dependence on the momentum parameter. For the same constant-velocity and sinusoidal targets considered in Fig. 1, we fix $R = \sigma^2/K^2 = 1$, and vary the momentum parameter γ . Each error is divided by the exact SGD error for the same target. This experiment illustrates both the gain from moderate momentum and the deterioration caused by excessive momentum (see Fig. 2).

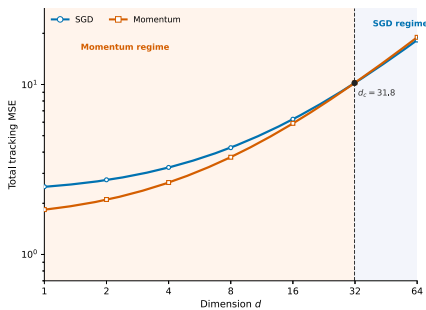


(a) **Linear drift.** The exact minimum occurs at $\gamma_{\text{opt}} \simeq 0.523$, where the asymptotic MSE is approximately 0.208 times the SGD error.

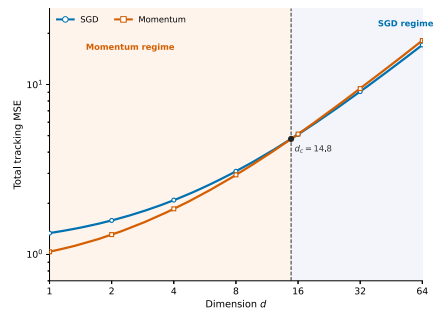


(b) **Nonlinear drift.** The exact cycle-averaged error is minimized at $\gamma_{\text{opt}} \simeq 0.473$, where it is approximately 0.374 times the SGD error.

Fig. 2 Dependence on γ . We use $\alpha = 0.4$, $K = 1$, and $R = \sigma^2/K^2 = 1$. Solid curves are exact and markers are obtained from 3,000 paired runs. In both cases the tracking error is U-shaped: momentum first reduces the drift-induced bias, but its stochastic-variance cost eventually dominates. The different minimizers show that the preferred momentum depends on the temporal form of the drift, not only on K and σ^2



(a) **Linear drift.** For $x_t^* = t\mathbf{1}_d/\sqrt{d}$, momentum has lower total MSE below $d_c \simeq 31.8$, after which SGD is preferable.



(b) **Nonlinear drift.** For $x_t^* = A \sin(2\pi(t + 1/2)/64)\mathbf{1}_d/\sqrt{d}$, the crossover occurs at $d_c \simeq 14.8$.

Fig. 3 Dimension dependence. We use $\alpha = 0.4$, $\gamma = 0.1$, and dimensions $d = 1, \dots, 64$. Solid curves are the exact total asymptotic MSEs, while markers are obtained from 1,500 paired runs. With fixed per-coordinate noise variance, total noise energy grows linearly with d , eventually reversing the momentum advantage

Dependence on dimension. To isolate dimension from conditioning, we consider the same quadratic objective in $\mathbb{R}^d$ with $\varepsilon_t \sim \mathcal{N}(0, I_d)$, so that the Hessian is the identity. We keep the total drift norm fixed at $K = 1$, while the per-coordinate observation variance remains one. Consequently, $\mathbb{E}\|\varepsilon_t\|^2 = d$, and increasing dimension moves the problem toward a noise-dominated regime.

In these experiments, the noise variance is fixed in each coordinate, so the total noise energy grows as d . If the total noise energy were instead held fixed by taking $\varepsilon_t \sim \mathcal{N}(0, I_d/d)$, the isotropic quadratic model would exhibit no dependence on dimension. Thus, Fig. 3a and b illustrate the effect of accumulating independent noise across dimensions, not an intrinsic deterioration caused by high dimension itself.

3.2 Beyond Quadratics: Time-Varying Logistic Regression

We finally test the phenomenon on a standard nonquadratic statistical objective. In particular we choose as a toy model the case of a logistic regression objective function

$$f(\theta; z, y) = \log(1 + e^{z^\top \theta}) - yz^\top \theta,$$

where $z \sim \mathcal{N}(0, I_S)$ and, conditionally on z ,

$$\mathbb{P}(y = 1 \mid z) = s(z^\top \theta_t^*), \quad s(u) = \frac{1}{1 + e^{-u}}.$$

The correctly specified population logistic risk is minimized at θ_t^* . We let this minimizer rotate according to

$$\theta_t^* = 1.5(\cos(2\pi t/512), \sin(2\pi t/512), 0, 0, 0).$$

Unlike the quadratic model, the Hessian

$$\mathbb{E}\left[s(z^\top \theta)(1 - s(z^\top \theta))zz^\top\right]$$

depends on the current parameter. The experiment is therefore an empirical test of whether the tracking mechanism persists beyond the constant-Hessian setting (see Fig. 4).

4 A Special Choice for γ

A natural question that arises at this point is: what value of γ should be selected in the general setting where the exact dynamics of the drift are unknown? In this section we provide a heuristic discussion for the choice of a guaranteed-improvement choice of γ . To this aim, let us set

$$\mathcal{E}_S(\alpha, \sigma, K) := \sigma^2 \left(\frac{\alpha}{2 - \alpha} \right) + K^2 \left(\frac{1 - \alpha}{\alpha} \right)^2 \quad (24)$$

and

$$\mathcal{E}_{HB}(\alpha, \gamma, \sigma, K) := \sigma^2 \left(\frac{\alpha}{2 - \alpha} + \frac{2\alpha(1 - \alpha)}{(2 - \alpha)^2} \gamma \right) + K^2 \left(\frac{1 - \alpha}{\alpha} - \frac{1}{\alpha} \gamma \right)^2; \quad (25)$$

the expression in (24) stands for the upper bound of the SGD method while (25) corresponds to the upper bound of the momentum algorithm where we have truncated the expressions obtained in Theorem 2.2 at first order in γ . Accordingly, the calculations below concern this surrogate and not the exact momentum bound.

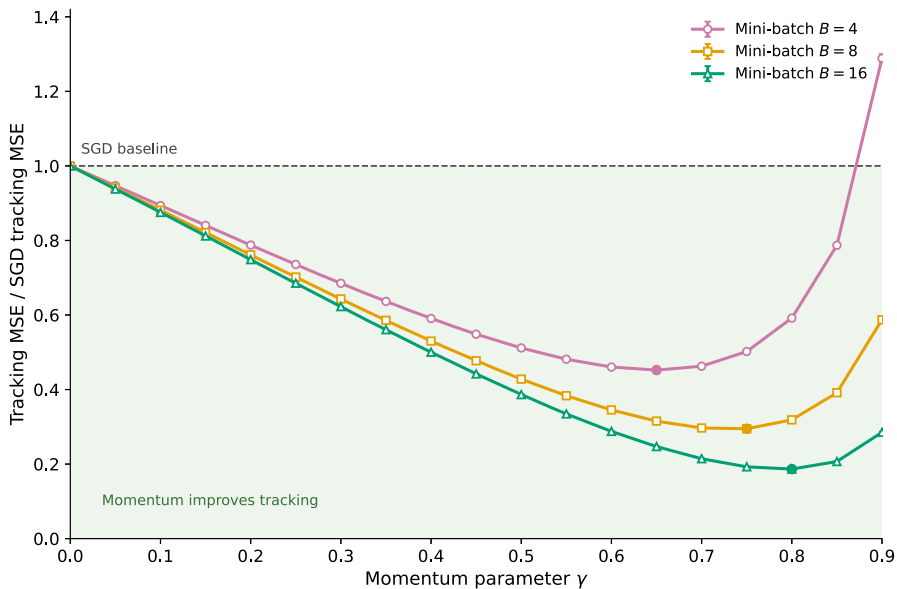


Fig. 4 Time-varying logistic regression. SGD and SHB use fresh mini-batch stochastic gradients with step size $\alpha = 0.1$. The vertical axis is $\mathbb{E}\|\theta_t - \theta_t^*\|^2$ divided by the SGD error for the same mini-batch size. Curves are obtained from 300 paired runs, and filled markers identify the best tested γ . Momentum improves tracking over a broad range for all three mini-batch sizes. As the mini-batch size grows and gradient noise decreases, the best tested momentum shifts from 0.65 to 0.75 and 0.80, while the corresponding relative MSEs decrease from 0.452 to 0.295 and 0.187. Excessive momentum again worsens tracking. These results provide empirical, rather than theorem-derived, evidence that the momentum effect persists for a nonquadratic objective

Let $R := \frac{\sigma^2}{K^2}$. We suppose $\alpha \in (0, 1)$ is fixed and find the value of γ that maximizes the gap

$$G(\alpha, \gamma, \sigma, K) := \mathcal{E}_S(\alpha, \sigma, K) - \mathcal{E}_{HB}(\alpha, \gamma, \sigma, K).$$

Notice that $G(\alpha, \gamma, \sigma, K) > 0$ is equivalent to

$$\gamma \left[\left(\frac{2(1-\alpha)}{\alpha^2} - R \frac{2\alpha(1-\alpha)}{(2-\alpha)^2} \right) - \frac{1}{\alpha^2} \gamma \right] > 0$$

or alternatively,

$$\gamma \left[2(1-\alpha) \left(1 - R \frac{\alpha^3}{(2-\alpha)^2} \right) - \gamma \right] > 0. \quad (26)$$

The maximum value of γ for which (26) holds is

$$\gamma_*(\alpha) = (1-\alpha) \left(1 - R \frac{\alpha^3}{(2-\alpha)^2} \right). \quad (27)$$

Before we plug this value in the algorithm described in Theorem 2.1 we have however to check whether it lies in the parameter interval covered by that result. That is we

need to check whether $\gamma_*(\alpha) < \gamma_e(\alpha)$, where we have defined

$$\gamma_e(\alpha) := (1 - \sqrt{\alpha})^2,$$

the right end point of the region where the two eigenvalues of the matrix M in the proof of Theorem 2.1 are distinct. Recalling that $0 < R \leq 1$, it is a simple calculation to check that

$$\gamma_*(\alpha) > \gamma_e(\alpha) \quad \text{for all } \alpha \in (0, 1).$$

Thus the maximizer of the truncated surrogate lies beyond the interval and so the γ that maximizes the distance between the upper bounds of SGD and momentum methods is

$$\gamma_e(\alpha) = (1 - \sqrt{\alpha})^2 \quad \text{for all } \alpha \in (0, 1).$$

This choice does not establish that $\gamma_e(\alpha)$ minimizes the true tracking error but it facilitates a more structured and manageable formulation of the tracking error bounds, although the matrix M is no longer diagonalizable and is instead similar to a Jordan block of order 2. This observation necessitates a modification of the proof technique used in the preceding theorems to derive the tracking error bounds. The refined approach is presented in the proof of the following theorem.

Theorem 4.1 *Consider the momentum method*

$$\begin{cases} x_{t+1} = (1 - \alpha)x_t + \alpha X_{t+1} + (1 - \sqrt{\alpha})^2(x_t - x_{t-1}), & t \geq 2 \\ x_1 = X_1, x_2 = X_2. \end{cases}$$

Then, there exist two functions $A(\alpha)$ and $B(\alpha)$ analytic in $(0, 1)$ such that for all $\alpha \in (0, 1)$ one has

$$\limsup_{t \rightarrow \infty} \mathbb{E}[(x_{t+1} - x_{t+1}^*)^2] \leq \sigma^2 A(\alpha) + K^2 B(\alpha); \quad (28)$$

furthermore $A(\alpha)$ and $B(\alpha)$ can be written explicitly as

$$A(\alpha) = \sqrt{\alpha} \frac{2 - 2\sqrt{\alpha} + \alpha}{(2 - \sqrt{\alpha})^3}$$

and

$$B(\alpha) = \frac{4(1 - \sqrt{\alpha})^2}{\alpha}.$$

Theorem 4.1 gives a closed-form tracking-error bound for this explicit choice of momentum. More importantly, its improvement over SGD can be verified directly, without relying on the truncated surrogate. Indeed, writing $s = \sqrt{\alpha}$, comparison with (24) shows that the bound in Theorem 4.1 is strictly smaller than the SGD bound whenever

$$R < R_{\text{crit}}(\alpha) := \frac{(3s + 1)(2 - s)^3(2 - s^2)}{4s^5}. \quad (29)$$

For $s \in (0, 1)$, we have $3s + 1 > 4s$, $2 - s > 1$, and $2 - s^2 > 1$, and therefore $R_{\text{crit}}(\alpha) > 1$. Consequently, the assumption $R = \sigma^2/K^2 \leq 1$ guarantees that $\gamma_e = (1 - \sqrt{\alpha})^2$ produces a strictly smaller tracking-error bound than SGD. Thus, although the optimizer of the original bound or of the true tracking error may occur at another value, γ_e is a simple *guaranteed-improvement choice* in a general setting.

Remark 4.1 The condition

$$\frac{\sigma^2}{K^2} \leq 1$$

compares two per-update scales: the standard deviation σ of the observation noise and the bound K on the one-step displacement of the optimum. It is a simple sufficient condition, rather than a sharp characterization of when momentum is beneficial. In particular, the numerical experiments show that the momentum advantage can persist for $\sigma^2/K^2 > 1$, with the crossover depending on the algorithm parameters and on the temporal structure of the drift. Moreover, because K is a worst-case bound, the terminology “drift-dominated” is most informative when K is chosen as a tight or physically meaningful drift scale.

This regime is plausible when the latent parameter changes appreciably between successive updates while observations or stochastic gradients have moderate variance. Examples include transient periods following a regime change, adaptive filtering or online system identification with rapidly varying coefficients, and tracking of time-varying signals or channels. These are classical settings in which an algorithm must balance tracking ability against noise rejection [17]. A similar regime can arise in continual-learning problems when concept drift is appreciable and mini-batch gradients are sufficiently accurate.

5 Proof of Theorem 2.1

We will divide the proof of Theorem 2.1 in two steps.

5.1 Step One

We rewrite equation (17) as

$$\begin{cases} x_{t+1} = (1 - \alpha)x_t + \alpha X_{t+1} + \gamma B_t, \\ B_{t+1} = x_{t+1} - x_t; \\ x_1 = X_1, \quad x_2 = X_2; \end{cases} \quad (30)$$

moreover, setting for all $t \geq 1$

$$\Lambda_t = x_t - x_t^*, \quad Z_t = X_t - x_t^* \quad \text{and} \quad K_{t+1} := x_{t+1}^* - x_t^*,$$

system (30) becomes

$$\begin{cases} \Lambda_{t+1} = (1 - \alpha)\Lambda_t + \alpha Z_{t+1} + \gamma B_t - (1 - \alpha)K_{t+1}, \\ B_{t+1} = \Lambda_{t+1} - \Lambda_t + K_{t+1}; \\ \Lambda_1 = Z_1, \quad \Lambda_2 = Z_2; \end{cases} \quad (31)$$

We now set $V_t = (\Lambda_t, B_t)^T$; this transforms equation (31) into

$$\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} V_{t+1} = \begin{bmatrix} 1 - \alpha & \gamma \\ -1 & 0 \end{bmatrix} V_t + \begin{bmatrix} \alpha - 1 \\ 1 \end{bmatrix} K_{t+1} + \begin{bmatrix} \alpha \\ 0 \end{bmatrix} Z_{t+1}, \quad (32)$$

and upon multiplying (32) on the left by $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ one gets

$$V_{t+1} = M V_t + w_1 K_{t+1} + w_2 Z_{t+1}, \quad (33)$$

where

$$M = \begin{bmatrix} 1 - \alpha & \gamma \\ -\alpha & \gamma \end{bmatrix}, \quad w_1 = \begin{bmatrix} \alpha - 1 \\ \alpha \end{bmatrix}, \quad \text{and} \quad w_2 = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}.$$

The matrix M is diagonalizable, with eigenvalues $\lambda_i, i = 1, 2$ given below, and there exists Q such that $Q^{-1}MQ = \Delta$, with Δ diagonal. The transpose M^T can also be diagonalized as $(Q^T)M^T(Q^T)^{-1} = \Delta$. In order to lighten the final expressions of our result we recall the definitions of some auxiliary functions in $\alpha, \gamma \in (0, 1)$:

$$\begin{aligned} \Lambda(\alpha, \gamma) &:= 1 - \alpha - \gamma, \quad L(\alpha, \gamma) := 1 - \alpha + \gamma, \\ \mu^2(\alpha, \gamma) &:= 1 + \alpha^2 + \gamma^2 - 2\alpha\gamma - 2\alpha - 2\gamma, \\ \rho_-(\alpha, \gamma) &:= \frac{2\gamma}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)}, \quad \rho_+(\alpha, \gamma) := \frac{2\gamma}{\Lambda(\alpha, \gamma) + \mu(\alpha, \gamma)}, \\ \lambda_1(\alpha, \gamma) &:= \frac{L(\alpha, \gamma) - \mu(\alpha, \gamma)}{2}, \quad \lambda_2(\alpha, \gamma) := \frac{L(\alpha, \gamma) + \mu(\alpha, \gamma)}{2}. \end{aligned} \quad (34)$$

It is an easy computation to check that the diagonalizer $Q = Q(\alpha, \gamma)$ is

$$Q = \begin{pmatrix} -\rho_+(\alpha, \gamma) & -\rho_-(\alpha, \gamma) \\ 1 & 1 \end{pmatrix} \quad (35)$$

with inverse

$$Q^{-1} = \frac{\alpha}{\mu(\alpha, \gamma)} \begin{pmatrix} 1 & \rho_-(\alpha, \gamma) \\ -1 & -\rho_+(\alpha, \gamma) \end{pmatrix}, \quad (36)$$

and diagonal matrix

$$\Delta = \begin{pmatrix} \lambda_1(\alpha, \gamma) & 0 \\ 0 & \lambda_2(\alpha, \gamma) \end{pmatrix} \quad (37)$$

In the region E defined in the statement of Theorem 2.1 the eigenvalues of M satisfy the inequalities $0 \leq \lambda_1(\alpha, \gamma) \leq \lambda_2(\alpha, \gamma) < 1$. Notice in addition that the quantity

$\mu^2(\alpha, \gamma)$ will be non-negative if and only if $(\alpha, \gamma) \in E$, i.e.

$$0 \leq \gamma \leq \gamma_e = (1 - \sqrt{\alpha})^2.$$

5.2 Step Two

Let us now multiply (33) on the left by Q^{-1} to get

$$\hat{V}_{t+1} = Q^{-1}MQ\hat{V}_t + \hat{w}_1K_{t+1} + \hat{w}_2Z_{t+1}, \quad (38)$$

with

$$\hat{V}_t := Q^{-1}V_t, \quad \hat{w}_1 := Q^{-1}w_1 \quad \text{and} \quad \hat{w}_2 := Q^{-1}w_2.$$

Notice that (38) is equivalent to two scalar recursive equations

$$\hat{V}_{t+1}^{(i)} = \lambda_i \hat{V}_t^{(i)} + \hat{w}_1^{(i)}K_{t+1} + \hat{w}_2^{(i)}Z_{t+1},$$

whose solutions, assuming for simplicity that $\hat{V}_1 = 0$ (the initial condition is not relevant in our analysis since we focus on the asymptotic behaviour of the solution), are given by

$$\hat{V}_{t+1}^{(i)} = \hat{w}_1^{(i)} \sum_{h=1}^t K_{h+1} \lambda_i^{t-h} + \hat{w}_2^{(i)} \sum_{h=1}^t Z_{h+1} \lambda_i^{t-h}, \quad i = 1, 2. \quad (39)$$

Our aim is to estimate the quantity

$$\mathbb{E}[(x_{t+1} - x_{t+1}^*)^2]; \quad (40)$$

since

$$x_{t+1} - x_{t+1}^* = \left\langle \begin{bmatrix} x_{t+1} - x_{t+1}^* \\ B_{t+1} \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\rangle = \langle V_{t+1}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \rangle,$$

and $V_t = Q\hat{V}_t$, we see that

$$x_{t+1} - x_{t+1}^* = \langle \hat{V}_{t+1}, g \rangle,$$

where we set $g := Q^T \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Therefore using (39) we obtain

$$\begin{aligned} x_{t+1} - x_{t+1}^* &= \sum_{i=1}^2 g^{(i)} \hat{V}_{t+1}^{(i)} \\ &= \sum_{i=1}^2 g^{(i)} \left(\hat{w}_1^{(i)} \sum_{h=1}^t K_{h+1} \lambda_i^{t-h} + \hat{w}_2^{(i)} \sum_{h=1}^t Z_{h+1} \lambda_i^{t-h} \right) \\ &= D_t + R_t, \end{aligned}$$

where

$$D_t := \sum_{h=1}^t \left(\sum_{i=1}^2 g^{(i)} \hat{w}_1^{(i)} \lambda_i^{t-h} \right) K_{h+1}, \quad (41)$$

and

$$R_t := \sum_{h=1}^t \left(\sum_{i=1}^2 g^{(i)} \hat{w}_2^{(i)} \lambda_i^{t-h} \right) Z_{h+1}.$$

Thus, exploiting the fact that the random variables Z_h 's are independent and have mean zero, we can evaluate the quantity (40) as

$$\mathbb{E}[(x_{t+1} - x_{t+1}^*)^2] = D_t^2 + \sigma^2 \sum_{h=1}^t \left(\sum_{i=1}^2 g^{(i)} \hat{w}_2^{(i)} \lambda_i^{t-h} \right)^2. \quad (42)$$

The second term in the right hand side of (42) can be elaborated as follows:

$$\begin{aligned} \sum_{h=1}^t \left(\sum_{i=1}^2 g^{(i)} \hat{w}_2^{(i)} \lambda_i^{t-h} \right)^2 &= \sum_{l=0}^{t-1} \left(g^{(1)} \hat{w}_2^{(1)} \lambda_1^l + g^{(2)} \hat{w}_2^{(2)} \lambda_2^l \right)^2 \\ &= (g^{(1)} \hat{w}_2^{(1)})^2 \sum_{l=0}^{t-1} \lambda_1^{2l} + (g^{(2)} \hat{w}_2^{(2)})^2 \sum_{l=0}^{t-1} \lambda_2^{2l} \\ &\quad + 2g^{(1)} \hat{w}_2^{(1)} g^{(2)} \hat{w}_2^{(2)} \sum_{l=0}^{t-1} \gamma^l \\ &= (g^{(1)} \hat{w}_2^{(1)})^2 \frac{1 - \lambda_1^{2t}}{1 - \lambda_1^2} + (g^{(2)} \hat{w}_2^{(2)})^2 \frac{1 - \lambda_2^{2t}}{1 - \lambda_2^2} \\ &\quad + 2g^{(1)} \hat{w}_2^{(1)} g^{(2)} \hat{w}_2^{(2)} \frac{1 - \gamma^t}{1 - \gamma}, \end{aligned}$$

recalling that $\lambda_1 \times \lambda_2 = \gamma$. On the other hand, writing (41) as

$$D_t = \sum_{l=0}^{t-1} \left(\sum_{i=1}^2 g^{(i)} \hat{w}_1^{(i)} \lambda_i^l \right) K_{t-l+1} = \sum_{l=0}^{t-1} K_{t-l+1} \left(g^{(1)} \hat{w}_1^{(1)} \lambda_1^l + g^{(2)} \hat{w}_1^{(2)} \lambda_2^l \right),$$

and using the bound $|K_j| \leq K$, for all $j \geq 2$, yields

$$|D_t| \leq K \sum_{l=0}^{t-1} \left(|g^{(1)} \hat{w}_1^{(1)}| \lambda_1^l + |g^{(2)} \hat{w}_1^{(2)}| \lambda_2^l \right),$$

that is

$$|D_t| \leq K |g^{(1)} \hat{w}_1^{(1)}| \frac{1 - \lambda_1^t}{1 - \lambda_1} + K |g^{(2)} \hat{w}_1^{(2)}| \frac{1 - \lambda_2^t}{1 - \lambda_2}.$$

Combining all the previous inequalities we can conclude that

$$\begin{aligned} & \limsup_{t \rightarrow \infty} \mathbb{E}[(x_{t+1} - x_{t+1}^*)^2] \\ & \leq K^2 \left(|g^{(1)} \hat{w}_1^{(1)}| \frac{1 - \lambda_1^t}{1 - \lambda_1} + |g^{(2)} \hat{w}_1^{(2)}| \frac{1 - \lambda_2^t}{1 - \lambda_2} \right)^2 \\ & + \sigma^2 \left((g^{(1)} \hat{w}_2^{(1)})^2 \frac{1 - \lambda_1^{2t}}{1 - \lambda_1^2} + (g^{(2)} \hat{w}_2^{(2)})^2 \frac{1 - \lambda_2^{2t}}{1 - \lambda_2^2} \right. \\ & \left. + 2g^{(1)} \hat{w}_2^{(1)} g^{(2)} \hat{w}_2^{(2)} \frac{1 - \gamma^t}{1 - \gamma} \right). \end{aligned}$$

Using (34), (35) and (36) we have that

$$g = \begin{bmatrix} -\rho_+(\alpha, \gamma) \\ -\rho_-(\alpha, \gamma) \end{bmatrix},$$

and that

$$\hat{w}_1 = \frac{\alpha}{\mu} \begin{bmatrix} \alpha - 1 + \alpha \rho_-(\alpha, \gamma) \\ 1 - \alpha - \alpha \rho_+(\alpha, \gamma) \end{bmatrix}, \quad \hat{w}_2 = \frac{\alpha}{\mu} \begin{bmatrix} \alpha + \alpha \rho_-(\alpha, \gamma) \\ -\alpha - \alpha \rho_+(\alpha, \gamma) \end{bmatrix}.$$

This means that

$$\begin{aligned} & \left(|g^{(1)} \hat{w}_1^{(1)}| \frac{1 - \lambda_1^t}{1 - \lambda_1} + |g^{(2)} \hat{w}_1^{(2)}| \frac{1 - \lambda_2^t}{1 - \lambda_2} \right)^2 \\ & \leq \left(\left(\frac{\alpha}{\mu(\alpha, \gamma)} |\gamma + (\alpha - 1) \rho_+(\alpha, \gamma)| \frac{1}{1 - \lambda_1} + \left(\frac{\alpha}{\mu(\alpha, \gamma)} \frac{|\gamma + (\alpha - 1) \rho_-(\alpha, \gamma)|}{1 - \lambda_2(\alpha, \gamma)} \right) \right)^2 \right). \end{aligned} \quad (43)$$

By the same token we get that

$$\begin{aligned} & (g^{(1)} \hat{w}_2^{(1)})^2 \frac{1 - \lambda_1^{2t}}{1 - \lambda_1^2} + (g^{(2)} \hat{w}_2^{(2)})^2 \frac{1 - \lambda_2^{2t}}{1 - \lambda_2^2} + 2g^{(1)} \hat{w}_2^{(1)} g^{(2)} \hat{w}_2^{(2)} \frac{1 - \gamma^t}{1 - \gamma} \\ & \leq \frac{\frac{\alpha^4}{\mu^2(\alpha, \gamma)} \rho_+^2(\alpha, \gamma) (1 + \rho_-(\alpha, \gamma))^2}{1 - \lambda_1^2(\alpha, \gamma)} \\ & + \frac{\frac{\alpha^4}{\mu^2(\alpha, \gamma)} \rho_-^2(\alpha, \gamma) (1 + \rho_+(\alpha, \gamma))^2}{1 - \lambda_2^2(\alpha, \gamma)} - 2 \frac{\gamma}{1 - \gamma} \frac{\alpha^2}{\mu^2(\alpha, \gamma)}. \end{aligned} \quad (44)$$

Calling $B(\alpha, \gamma)$ the right hand side of (43) and $A(\alpha, \gamma)$ the right hand side of (44) we have proven (15) and (16) and thus Theorem (2.1).

6 Proof of Theorem 2.2

The proof of Theorem 2.2 is rather long so again we divide it in several steps.

6.1 Step One

Referring to the notation of Sect. 5.1, it is easy to check that

$$Mw_1 = (1 - \alpha)w_1 + \gamma w_2 \quad \text{and} \quad Mw_2 = -\alpha w_1 + \gamma w_2.$$

Since we are assuming $\alpha > 0$, w_1 and w_2 form a basis of $\mathbb{R}^2$ so that we can always have for some real coefficients g_l, h_l

$$M^l w_1 = g_l w_1 + h_l w_2, \tag{45}$$

for every $l \in \mathbb{N}$: $g_0 = 1, h_0 = 0$ and $g_1 = 1 - \alpha, h_1 = \gamma$. We want to compute these coefficients g_l, h_l for every $l \in \mathbb{N}$. By comparing the $l + 1$ powers from (45) we get

$$\begin{cases} g_{l+1} = (1 - \alpha)g_l - \alpha h_l \\ h_{l+1} = \gamma(g_l + h_l) \end{cases}, \tag{46}$$

and (46) can be written as

$$\zeta_{l+1} = M^T \zeta_l \quad \text{with} \quad \zeta_l = (g_l, h_l)^T. \tag{47}$$

From (47) we deduce that

$$\zeta_l = (M^T)^l \zeta_0 = (M^T)^l \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{48}$$

Therefore we also have from the preceding form of Q

$$Q^T = \begin{pmatrix} -\rho_+(\alpha, \gamma) & 1 \\ -\rho_-(\alpha, \gamma) & 1 \end{pmatrix},$$

and

$$(Q^{-1})^T = \frac{\alpha}{\mu(\alpha, \gamma)} \begin{pmatrix} 1 & -1 \\ \rho_-(\alpha, \gamma) & -\rho_+(\alpha, \gamma) \end{pmatrix}.$$

Using (48) we get

$$(M^T)^l = (Q^T)^{-1} \Delta^l Q^T, \tag{49}$$

with Δ given by (37). Therefore we get from (48) and (49)

$$g_l(\alpha, \gamma) = \frac{\alpha}{\mu(\alpha, \gamma)} (\rho_-(\alpha, \gamma) \lambda_2^l(\alpha, \gamma) - \rho_+(\alpha, \gamma) \lambda_1^l(\alpha, \gamma)), \tag{50}$$

and

$$h_l(\alpha, \gamma) = \frac{\alpha}{\mu(\alpha, \gamma)} (\lambda_2^l(\alpha, \gamma) - \lambda_1^l(\alpha, \gamma)) \rho_-(\alpha, \gamma) \rho_+(\alpha, \gamma), \quad (51)$$

for every $l \in \mathbb{N}$. Also notice that

$$\rho_-(\alpha, \gamma) \rho_+(\alpha, \gamma) = \frac{\gamma}{\alpha}.$$

6.2 Step Two

Let us consider (33) written like

$$V_{t+1} = M V_t + S_{t+1},$$

with $S_t = w_1 K_t + w_2 Z_t$. Then the recursive equation is solved explicitly

$$V_t = M^{t-1} V_1 + \sum_{h=1}^t M^{t-h} S_h. \quad (52)$$

Since we are interested in estimating the first component of V_t we multiply (52) scalarly by $e_1^T = (1, 0)$. Thus we get

$$\Lambda_t = \langle M^t V_1, e_1 \rangle + \sum_{h=1}^t \langle M^{t-h} S_h, e_1 \rangle.$$

Therefore we have

$$\Lambda_t = \langle M^t V_1, e_1 \rangle + \sum_{h=1}^t \langle w_1 K_h + w_2 Z_h, (M^T)^{t-h} e_1 \rangle.$$

Using (48) we get

$$\begin{aligned} \Lambda_t &= \langle M^t V_1, e_1 \rangle + \sum_{l=0}^{t-1} K_{t-l} ((\alpha - 1) g_l(\alpha, \gamma) + \alpha h_l(\alpha, \gamma)) \\ &\quad + \sum_{l=0}^{t-1} \alpha (g_l(\alpha, \gamma) + h_l(\alpha, \gamma)) Z_{t-l} \\ &= v_1 + v_2 + v_3 \end{aligned} \quad (53)$$

We now proceed to estimate all the terms coming from squaring Λ_t and subsequently applying the $\mathbb{E}$ operator to them.

6.3 Step Three

Let us examine v_2 ; we have:

$$v_2 = \sum_{l=0}^{t-1} K_{t-l} \left((\alpha - 1) \frac{\alpha}{\mu(\alpha, \gamma)} (\rho_-(\alpha, \gamma) \lambda_2^l(\alpha, \gamma) - \rho_+(\alpha, \gamma) \lambda_1^l(\alpha, \gamma)) + \frac{\alpha\gamma}{\mu(\alpha, \gamma)} (\lambda_2^l(\alpha, \gamma) - \lambda_1^l(\alpha, \gamma)) \right). \quad (54)$$

Let us regroup (54) in this way

$$v_2 = \left(\frac{\alpha\gamma}{\mu(\alpha, \gamma)} \left(\frac{2(\alpha - 1)}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)} + 1 \right) \right) \sum_{l=0}^{t-1} K_{t-l} \lambda_2^l(\alpha, \gamma) - \left(\frac{\alpha\gamma}{\mu(\alpha, \gamma)} \left(\frac{2(\alpha - 1)}{\Lambda(\alpha, \gamma) + \mu(\alpha, \gamma)} + 1 \right) \right) \sum_{l=0}^{t-1} K_{t-l} \lambda_1^l(\alpha, \gamma). \quad (55)$$

Let us set

$$\Phi_2(\alpha, \gamma) := \frac{\alpha\gamma}{\mu(\alpha, \gamma)} \left(\frac{2(\alpha - 1)}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)} + 1 \right), \quad (56)$$

$$\Phi_1(\alpha, \gamma) := \frac{\alpha\gamma}{\mu(\alpha, \gamma)} \left(\frac{2(\alpha - 1)}{\Lambda(\alpha, \gamma) + \mu(\alpha, \gamma)} + 1 \right); \quad (57)$$

it is easily checked that

$$\Phi_2(\alpha, \gamma) < \Phi_1(\alpha, \gamma) < 0,$$

in the region where $\mu^2 > 0$. In the same region v_2 is analytic in (α, γ) .

Our aim now is write Taylor asymptotic developments of all the functions involved in the computation of v_j , $j = 2, 3$, uniform in $\alpha \in (0, 1)$ when γ is close to 0 in the region $0 \leq \gamma < (1 - \sqrt{\alpha})^2$, so that we can measure the effect of "turning on" γ in the estimate for $\mathbb{E}[(\Lambda_t)^2]$.

6.4 Step Four

We now list a number of straightforward Taylor expansions for all the relevant functions and their combinations. The following computations are easy albeit tedious:

$$\begin{aligned} \mu(\alpha, \gamma) &= (1 - \alpha) - \frac{1 + \alpha}{1 - \alpha} \gamma - \frac{2\alpha}{(1 - \alpha)^3} \gamma^2 + \mathcal{O}(\gamma^3), \\ \frac{1}{\mu(\alpha, \gamma)} &= \frac{1}{1 - \alpha} + \frac{1 + \alpha}{1 - \alpha} \gamma + \frac{1 + 4\alpha + \alpha^2}{(1 - \alpha)^5} \gamma^2 + \mathcal{O}(\gamma^3), \\ \Lambda(\alpha, \gamma) - \mu(\alpha, \gamma) &= \frac{2\alpha}{1 - \alpha} \gamma + \frac{2\alpha}{(1 - \alpha)^3} \gamma^2 + \frac{2\alpha(1 + \alpha)}{(1 - \alpha)^5} \gamma^3 + \mathcal{O}(\gamma^4), \end{aligned} \quad (58)$$

$$\begin{aligned}\Lambda(\alpha, \gamma) + \mu(\alpha, \gamma) &= 2(1 - \alpha) - \frac{2}{1 - \alpha}\gamma - \frac{2\alpha}{(1 - \alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3), \\ L(\alpha, \gamma) - \mu(\alpha, \gamma) &= \frac{2}{1 - \alpha}\gamma + \frac{2\alpha}{(1 - \alpha)^3}\gamma^2 + \frac{2\alpha(1 + \alpha)}{(1 - \alpha)^5}\gamma^3 + \mathcal{O}(\gamma^4), \\ L(\alpha, \gamma) + \mu(\alpha, \gamma) &= 2(1 - \alpha) - \frac{2\alpha}{1 - \alpha}\gamma - \frac{2\alpha}{(1 - \alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3),\end{aligned}\quad (59)$$

$$\lambda_1(\alpha, \gamma) = \frac{1}{1 - \alpha}\gamma + \frac{\alpha}{(1 - \alpha)^3}\gamma^2 + \frac{\alpha(1 + \alpha)}{(1 - \alpha)^5}\gamma^3 + \mathcal{O}(\gamma^4),$$

$$\lambda_2(\alpha, \gamma) = (1 - \alpha) - \frac{\alpha}{1 - \alpha}\gamma - \frac{\alpha}{(1 - \alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3),$$

$$\frac{1}{1 - \lambda_2(\alpha, \gamma)} = \frac{1}{\alpha} - \frac{1}{\alpha(1 - \alpha)}\gamma - \frac{1}{(1 - \alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3),\quad (60)$$

$$\frac{1}{1 - \lambda_2^2(\alpha, \gamma)} = \frac{1}{\alpha(2 - \alpha)} - \frac{2}{\alpha(2 - \alpha)^2}\gamma - \frac{4 - 3\alpha}{(1 - \alpha)^3(2 - \alpha)}\gamma^2 + \mathcal{O}(\gamma^3).$$

Recalling that $\rho_- = \frac{2\gamma}{\Lambda - \mu}$ we get

$$\rho_-(\alpha, \gamma) = \frac{1 - \alpha}{\alpha} \left(1 - \frac{1}{(1 - \alpha)^2}\gamma - \frac{\alpha}{(1 - \alpha)^4}\gamma^2 + \mathcal{O}(\gamma^3) \right);$$

similarly, from $\rho_+ = \frac{2\gamma}{\Lambda + \mu}$ we see that

$$\rho_+(\alpha, \gamma) = \frac{1}{1 - \alpha}\gamma + \frac{1}{(1 - \alpha)^3}\gamma^2 + \frac{1 + \alpha}{(1 - \alpha)^5}\gamma^3 + \mathcal{O}(\gamma^4).$$

We are now ready to start analyzing the coefficients in (57) and (56).

$$\Phi_1(\alpha, \gamma) = \frac{\alpha\gamma}{\mu(\alpha, \gamma)} \left(\frac{2(\alpha - 1)}{\Lambda(\alpha, \gamma) + \mu(\alpha, \gamma)} + 1 \right),$$

and using the previous expansion it is readily seen that

$$\Phi_1(\alpha, \gamma) = -\frac{\alpha}{(1 - \alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3).$$

Now let us study

$$\Phi_2(\alpha, \gamma) = \frac{\alpha\gamma}{\mu(\alpha, \gamma)} \left(\frac{2(\alpha - 1)}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)} + 1 \right).$$

Since what follows is crucial we need to compute everything very carefully:

$$\frac{2(\alpha - 1)}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)} + 1 = \frac{-1 + \alpha - \gamma - \mu(\alpha, \gamma)}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)} = -\frac{L(\alpha, \gamma) + \mu(\alpha, \gamma)}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)}. \quad (61)$$

This means

$$\begin{aligned}\Phi_2(\alpha, \gamma) &= -\frac{\alpha\gamma}{\mu(\alpha, \gamma)} \frac{L(\alpha, \gamma) + \mu(\alpha, \gamma)}{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)} \\ &= -\frac{\alpha}{\mu(\alpha, \gamma)} \frac{L(\alpha, \gamma) + \mu(\alpha, \gamma)}{\frac{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)}{\gamma}} \\ &= -\alpha \frac{\left(2(1-\alpha) - \frac{2\alpha}{1-\alpha}\gamma - \frac{2\alpha}{(1-\alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3)\right)}{D(\alpha, \gamma)},\end{aligned}$$

where

$$D(\alpha, \gamma) = \mu(\alpha, \gamma) \frac{\Lambda(\alpha, \gamma) - \mu(\alpha, \gamma)}{\gamma}.$$

Using the previous asymptotic expansions we have that

$$\begin{aligned}D(\alpha, \gamma) &= \left((1-\alpha) - \frac{1+\alpha}{1-\alpha}\gamma - \frac{2\alpha}{(1-\alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3)\right) \\ &\quad \times \left(\frac{2\alpha}{1-\alpha} + \frac{2\alpha}{(1-\alpha)^3}\gamma + \frac{2\alpha(1+\alpha)}{(1-\alpha)^5}\gamma^2 + \mathcal{O}(\gamma^3)\right).\end{aligned}$$

Combining the asymptotic developments of the numerator and of the denominator we see that

$$\Phi_2(\alpha, \gamma) = -(1-\alpha) + \mathcal{O}(\gamma^2)$$

in other words no linear term in γ appears. The coefficient of γ^2 is $\frac{1}{(1-\alpha)^3}$. Going back to (55) we can write

$$v_2 = \left(-(1-\alpha) + \frac{1}{(1-\alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3)\right) \sum_{l=0}^{t-1} K_{t-l} \lambda_2^l \quad (62)$$

$$+ \left(\frac{\alpha}{(1-\alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3)\right) \sum_{l=0}^{t-1} K_{t-l} \lambda_1^l. \quad (63)$$

From (62) we derive the estimate of the absolute value of v_2 :

$$|v_2| \leq |(1-\alpha) + \frac{1}{(1-\alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3)| K \frac{1}{1-\lambda_2} + K \left(\frac{\alpha}{(1-\alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3)\right). \quad (64)$$

Notice that by (60)

$$\frac{1}{1-\lambda_2(\alpha, \gamma)} = \frac{1}{\alpha} - \frac{1}{\alpha(1-\alpha)}\gamma - \frac{1}{(1-\alpha)^3}\gamma^2 + \mathcal{O}(\gamma^3)$$

and also that $\lambda_1(\alpha, 0) = 0$ therefore combining the two asymptotic developments we end up with

$$|v_2| \leq K \left| \frac{1-\alpha}{\alpha} - \frac{1}{\alpha} \gamma + \left(\frac{1}{1-\alpha} + 2\left(\frac{\alpha}{1-\alpha}\right)^2 \right) \gamma^2 + \mathcal{O}(\gamma^3) \right|.$$

We can therefore define

$$B(\alpha, \gamma) = \left(\frac{1-\alpha}{\alpha} - \frac{1}{\alpha} \gamma + \left(\frac{1}{1-\alpha} + 2\left(\frac{\alpha}{1-\alpha}\right)^2 \right) \gamma^2 + \mathcal{O}(\gamma^3) \right)^2,$$

and conclude that

$$v_2^2 \leq B(\alpha, \gamma).$$

The analyticity of $B(\alpha, \gamma)$ is clear from (64).

6.5 Step Five

Now from (53) we get

$$\Lambda_t^2 = (v_1 + v_2 + v_3)^2 = v_1^2 + v_2^2 + v_3^2 + 2v_1v_2 + 2v_2v_3 + 2v_3v_1. \quad (65)$$

Let us examine v_3^2 :

$$v_3^2 = \left(\sum_{l=0}^{t-1} \alpha(g_l(\alpha, \gamma) + h_l(\alpha, \gamma)) Z_{t-l} \right)^2 = \alpha^2 \sum_{l=0}^{t-1} (g_l(\alpha, \gamma) + h_l(\alpha, \gamma))^2 Z_{t-l}^2 \\ + \text{terms with zero expectation}.$$

From equations (50) and (51) we have

$$g_l + h_l = \frac{\alpha}{\mu(\alpha, \gamma)} (\rho_-(\alpha, \gamma) \lambda_2^l(\alpha, \gamma) - \rho_+(\alpha, \gamma) \lambda_1^l(\alpha, \gamma)) \\ + \frac{\gamma}{\mu(\alpha, \gamma)} (\lambda_2^l(\alpha, \gamma) - \lambda_1^l(\alpha, \gamma)). \quad (66)$$

We can see that in (66) with $l = 0$ we have

$$g_0 + h_0 = \frac{\alpha}{\mu} (\rho_- - \rho_+) = 1,$$

as it is easily verified. Thus we have

$$v_3^2 = \alpha^2 \left(Z_t^2 + \sum_{l=1}^{t-1} (g_l(\alpha, \gamma) + h_l(\alpha, \gamma))^2 Z_{t-l}^2 \right).$$

Since we also interested in the first order error in γ we need to be a little more precise. When $l \geq 1$ we have that

$$(g_l(\alpha, \gamma) + h_l(\alpha, \gamma))^2 = \left[\left(\frac{\alpha \rho_-(\alpha, \gamma)}{\mu(\alpha, \gamma)} + \frac{\gamma}{\mu(\alpha, \gamma)} \right) \lambda_2^l(\alpha, \gamma) - \left(\frac{\alpha \rho_+(\alpha, \gamma)}{\mu(\alpha, \gamma)} + \frac{\gamma}{\mu(\alpha, \gamma)} \right) \lambda_1^l(\alpha, \gamma) \right]^2.$$

Now, notice that $\rho_+(\alpha, \gamma) = \mathcal{O}(\gamma)$ as well as λ_1 . More precisely we have

$$\frac{\alpha \rho_+(\alpha, \gamma)}{\mu(\alpha, \gamma)} + \frac{\gamma}{\mu(\alpha, \gamma)} = \frac{1}{(1-\alpha)^2} \gamma + \frac{1+2\alpha}{(1-\alpha)^4} \gamma^2 + \mathcal{O}(\gamma^3),$$

which gives us

$$\left(\frac{\alpha \rho_+(\alpha, \gamma)}{\mu(\alpha, \gamma)} + \frac{\gamma}{\mu(\alpha, \gamma)} \right) \lambda_1(\alpha, \gamma) = \frac{1}{(1-\alpha)^3} \gamma^2 (1 + \mathcal{O}(\gamma)),$$

and the term

$$\left(\frac{\alpha \rho_+(\alpha, \gamma)}{\mu(\alpha, \gamma)} + \frac{\gamma}{\mu(\alpha, \gamma)} \right) \lambda_1^l(\alpha, \gamma),$$

is $\mathcal{O}(\gamma^{l+1})$ for every $l \geq 2$. Utilizing all the previous asymptotic expansions we have that $(\frac{\alpha \rho_-}{\mu} + \frac{\gamma}{\mu})^2$ has the following asymptotic development

$$\left(\frac{\alpha \rho_-(\alpha, \gamma)}{\mu(\alpha, \gamma)} + \frac{\gamma}{\mu(\alpha, \gamma)} \right)^2 = 1 + 2 \frac{1}{(1-\alpha)^2} \gamma + \frac{4\alpha+3}{(1-\alpha)^4} \gamma^2 + \mathcal{O}(\gamma^3).$$

This implies that modulo $\mathcal{O}(\gamma^3)$ we can write when $l = 1$

$$(g_1(\alpha, \gamma) + h_1(\alpha, \gamma))^2 = \left(\left(1 + 2 \frac{1}{(1-\alpha)^2} \gamma + \frac{4\alpha+3}{(1-\alpha)^4} \gamma^2 \right) \lambda_2(\alpha, \gamma) - \frac{2}{(1-\alpha)^2} \gamma^2 \right) + \mathcal{O}(\gamma^3),$$

and that modulo $\mathcal{O}(\gamma^3)$ we can write when $l \geq 2$

$$(g_l(\alpha, \gamma) + h_l(\alpha, \gamma))^2 = \left(1 + 2 \frac{1}{(1-\alpha)^2} \gamma + \frac{4\alpha+3}{(1-\alpha)^4} \gamma^2 \right) \lambda_2^2(\alpha, \gamma) + \mathcal{O}(\gamma^3),$$

so that

$$\mathbb{E}[v_3^2] \leq \sigma^2 \alpha^2 \left(1 + \left(1 + 2 \frac{1}{(1-\alpha)^2} \gamma + \frac{4\alpha+3}{(1-\alpha)^4} \gamma^2 \right) \sum_{l=1}^{\infty} \lambda_2^{2l} - \frac{2}{(1-\alpha)^2} \gamma^2 \right) + \mathcal{O}(\gamma^3),$$

which implies

$$\mathbb{E}[v_3^2] \leq \sigma^2 \alpha^2 \left(\frac{1}{1 - \lambda_2^2} + \left(2 \frac{1}{(1 - \alpha)^2} \gamma + \frac{4\alpha + 3}{(1 - \alpha)^4} \gamma^2 \right) \frac{\lambda_2^2}{1 - \lambda_2^2} - \frac{2}{(1 - \alpha)^2} \gamma^2 \right) + \mathcal{O}(\gamma^3).$$

But we recall that $\frac{\lambda_2^2(\alpha, \gamma)}{1 - \lambda_2^2(\alpha, \gamma)^2} = \frac{(1 - \alpha)^2}{2\alpha - \alpha^2} - \frac{2}{\alpha(2 - \alpha)^2} \gamma + \mathcal{O}(\gamma^2)$, and that

$$\frac{1}{1 - \lambda_2(\alpha, \gamma)^2} = \frac{1}{\alpha(2 - \alpha)} - \frac{2}{\alpha(2 - \alpha)^2} \gamma - \frac{4 - 3\alpha}{(1 - \alpha)^3(2 - \alpha)} \gamma^2 + \mathcal{O}(\gamma^3),$$

eventually yielding

$$\mathbb{E}[v_3^2] \leq \sigma^2 \left(\frac{\alpha}{2 - \alpha} + \frac{2\alpha(1 - \alpha)}{(2 - \alpha)^2} \gamma + \frac{2\alpha(1 + (1 - \alpha)^2)}{(2 - \alpha)^3} \gamma^2 + \mathcal{O}(\gamma^3) \right) = \sigma^2 A(\alpha, \gamma).$$

6.6 Step Six

Going back to equation (65) and applying the expectation $\mathbb{E}$ operator we can compute

$$\mathbb{E}[\Lambda_t^2] = \mathbb{E}[v_1^2] + v_2^2 + \mathbb{E}[v_3^2] + 2\mathbb{E}[v_1 v_2]. \quad (67)$$

From (53) $v_1 = \langle M^t V_1, e_1 \rangle$. Since the spectrum of M and of M^T is in $(0, 1)$ any induced norm of M^t goes to 0 when $t \rightarrow \infty$. Indeed, from $M = Q^{-1} \Delta Q$ we have $\|M^n\| = \|Q^{-1}(\Delta)^n Q\| \leq \|Q^{-1}\| \|Q\| \max\{\lambda_1^n, \lambda_2^n\} = C \max\{\lambda_1^n, \lambda_2^n\}$. From (67) we conclude that

$$\limsup_{t \rightarrow \infty} \mathbb{E}[\Lambda_t^2] \leq \sigma^2 A(\alpha, \gamma) + K^2 B(\alpha, \gamma),$$

with

$$A(\alpha, \gamma) = \left(\frac{\alpha}{2 - \alpha} + \frac{2\alpha(1 - \alpha)}{(2 - \alpha)^2} \gamma + \frac{2\alpha(1 + (1 - \alpha)^2)}{(2 - \alpha)^3} \gamma^2 + \mathcal{O}(\gamma^3) \right)$$

$$\text{and } B(\alpha, \gamma) = \left(\frac{1 - \alpha}{\alpha} - \frac{1}{\alpha} \gamma + \left(\frac{1}{1 - \alpha} + 2 \left(\frac{\alpha}{1 - \alpha} \right)^2 \right) \gamma^2 + \mathcal{O}(\gamma^3) \right)^2.$$

6.7 Proof of Theorem 4.1

The special value $\gamma = (1 - \sqrt{\alpha})^2$ is chosen in such a way as to maximize the negative contribution of $B(\alpha, \gamma)$ in (19), thus causing the function $\mu(\alpha, \gamma) = \sqrt{1 + \alpha^2 + \gamma^2 - 2\alpha\gamma - 2\alpha - 2\gamma}$ in the statement of Theorem 2.1 to vanish and hence generating an eigenvalue $\lambda(\alpha) = 1 - \sqrt{\alpha}$ of multiplicity 2 in the matrix M . The computation proceeds formally as in Sect. 6.1 until we get to

$$\zeta_l = (M^T)^l \zeta_0 = (M^T)^l \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

where now the matrix M is given by

$$M = \begin{bmatrix} 1 - \alpha & (1 - \sqrt{\alpha})^2 \\ -\alpha & (1 - \sqrt{\alpha})^2 \end{bmatrix}$$

It is easy to verify that M has a unique double eigenvalue $\lambda = 1 - \sqrt{\alpha}$ and that $M - \lambda I_2$ has a Jordan block of size 2. Then, we cannot diagonalize M but we can bring to its Jordan form via the matrix S

$$S = \begin{bmatrix} 1 - \frac{1}{\sqrt{\alpha}} & -\frac{1}{\alpha} \\ 1 & 0 \end{bmatrix}$$

and we can check that

$$M = SJS^{-1} \quad \text{where } J = \begin{bmatrix} 1 - \sqrt{\alpha} & 1 \\ 0 & 1 - \sqrt{\alpha} \end{bmatrix} = \begin{bmatrix} \lambda(\alpha) & 1 \\ 0 & \lambda(\alpha) \end{bmatrix}$$

A straightforward computation yields that

$$(M^T)^l \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \lambda^l(\alpha) - \alpha(1 - \frac{1}{\sqrt{\alpha}})l\lambda^{l-1}(\alpha) \\ (1 - \frac{1}{\sqrt{\alpha}})\lambda^l(\alpha) + (\alpha - \sqrt{\alpha})(-\frac{1}{\alpha} + l\lambda^{l-1}(\alpha)(1 - \frac{1}{\sqrt{\alpha}})) \end{bmatrix}$$

In a similar way to (50) and (51) we get, with $l \geq 0$

$$g_l(\alpha) = \lambda^l(\alpha) - (\alpha - \sqrt{\alpha})l\lambda^{l-1}(\alpha)$$

and

$$h_l(\alpha) = (1 - \sqrt{\alpha})^2 l\lambda^{l-1}(\alpha)$$

The same identity as (53) holds

$$\begin{aligned} \Lambda_t &= \langle M^t V_1, e_1 \rangle + \sum_{l=0}^{t-1} K_{t-l}((\alpha - 1)g_l(\alpha) + \alpha h_l(\alpha)) + \sum_{l=0}^{t-1} \alpha(g_l(\alpha) + h_l(\alpha))Z_{t-l} \\ &= v_1 + v_2 + v_3 \end{aligned}$$

To estimate v_2 we consider

$$(\alpha - 1)g_l(\alpha) + \alpha h_l(\alpha) = -(1 - \alpha)\lambda^l(\alpha) - \sqrt{\alpha}(1 - \sqrt{\alpha})^2 l\lambda^{l-1}(\alpha)$$

Then

$$|v_2| \leq K \sum_{l=0}^{\infty} \left((1 - \alpha)\lambda^l(\alpha) + \sqrt{\alpha}(1 - \sqrt{\alpha})^2 l\lambda^{l-1}(\alpha) \right) = K \frac{2(1 - \sqrt{\alpha})}{\sqrt{\alpha}}$$

and we finally get

$$|v_2|^2 \leq K^2 \frac{4(1 - \sqrt{\alpha})^2}{\alpha}.$$

Notice how

$$\frac{4(1 - \sqrt{\alpha})^2}{\alpha} < \left(\frac{1 - \alpha}{\alpha} \right)^2;$$

we set

$$B(\alpha) = \frac{4(1 - \sqrt{\alpha})^2}{\alpha}$$

We now estimate $\mathbb{E}[v_3^2]$:

$$\mathbb{E}[v_3^2] = \alpha^2 \sigma^2 \sum_{l=0}^{t-1} (g_l(\alpha) + h_l(\alpha))^2$$

We check that

$$g_l(\alpha) + h_l(\alpha) = \lambda^l(\alpha) - (\alpha - \sqrt{\alpha})l\lambda^{l-1}(\alpha) + (1 - \sqrt{\alpha})^2 l\lambda^{l-1}(\alpha) = (1 + l)\lambda^l(\alpha)$$

Recall now the simple summation identity

$$\sum_{k=0}^{\infty} (1 + k)^2 x^{2k} = \frac{1}{1 - x^2} + \frac{2x^2}{(1 - x^2)^2} + \frac{x^2(1 + x^2)}{(1 - x^2)^3}$$

valid when $|x| < 1$, then we have with

$$A(\alpha) = \sqrt{\alpha} \frac{2 - 2\sqrt{\alpha} + \alpha}{(2 - \sqrt{\alpha})^3}$$

that

$$\mathbb{E}[v_3^2] \leq \sigma^2 A(\alpha)$$

The final statement in (28) is then reached exactly as in the proof of Theorem 2.2.

Author Contributions All authors contributed equally to the conception, design, data collection, analysis, and writing of this manuscript. All authors read and approved the final version.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Amari, S.-I.: Natural gradient works efficiently in learning. *Neural Comput.* **10**(2), 251–276 (1998)
2. Assran, M., Rabbat, M.: On the convergence of Nesterov's accelerated gradient method in stochastic settings (2020). [arXiv:2002.12414](https://arxiv.org/abs/2002.12414)
3. Bernardi, E., Lanconelli, A., Lauria, C.S.: A novel theoretical framework for exponential smoothing (2024). [arXiv:2403.04345](https://arxiv.org/abs/2403.04345) arXiv preprint
4. Bernardi, E., Lanconelli, A., Lauria, C.S.A.: Tracking time varying parameters via online simplified maximum likelihood. *J. Optim. Theory Appl.* (2025). <https://doi.org/10.1007/s10957-025-02716-2>
5. Bernstein, A., Dall'Anese, E., Simonetto, A.: Online primal-dual methods with measurement feedback for time-varying convex optimization. *IEEE Trans. Signal Process.* **67**(8), 1978–1991 (2019)
6. Besbes, O., Gur, Y., Zeevi, A.: Non-stationary stochastic optimization. *Oper. Res.* **63**(5), 1227–1244 (2015)
7. Brockwell, P.J., Davis, R.A.: Time series: theory and methods. Springer, New York (2009)
8. Cutler, J., Drusvyatskiy, D., Harchaoui, Z.: Stochastic optimization under time drift: iterate averaging, step-decay schedules, and high probability guarantees. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (eds), *Advances in Neural Information Processing Systems*, vol. 34, pp. 11859–11869. Curran Associates, Inc. (2021). URL <https://proceedings.neurips.cc/paper/2021/file/62e7f2e090fe150ef8deb4466fdc81b3-Paper.pdf>
9. Cutler, J., Drusvyatskiy, D., Harchaoui, Z.: Stochastic optimization under distributional drift. *J. Mach. Learn. Res.* **24**(147), 1–56 (2023)
10. Dixit, R., Bedi, A.S., Tripathi, R., Rajawat, K.: Online learning with inexact proximal online gradient descent algorithms. *IEEE Trans. Signal Process.* **67**(5), 1338–1352 (2019)
11. Gadat, S., Panloup, F., Saadane, S.: Stochastic heavy ball. *Electronic J. Stat.* **12**, 09 (2018). <https://doi.org/10.1214/18-EJS1395>
12. Ghadimi, E., Feyzmahdavian, H.R., Johansson, M.: Global convergence of the heavy-ball method for convex optimization. In 2015 European control conference (ECC), pp. 310–315. IEEE (2015)
13. Goujaud, B., Taylor, A., Dieuleveut, A.: Provable non-accelerations of the heavy-ball method (2023). [arXiv:2307.11291](https://arxiv.org/abs/2307.11291) arXiv preprint
14. Kidambi, R., Netrapalli, P., Jain, P., Kakade, S.: On the insufficiency of existing momentum schemes for stochastic optimization. In 2018 Information Theory and Applications Workshop (ITA), pp. 1–9. IEEE (2018)
15. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. *International Conference on Learning Representations*, vol. 12 (2014)
16. Lanconelli, A., Lauria, C.S.: Maximum likelihood with a time varying parameter. *Stat. Pap.* (2024). <https://doi.org/10.1007/s00362-023-01497-y>
17. Ljung, L., Gunnarsson, S.: Adaptation and tracking in system identification—a survey. *Automatica* **26**(1), 7–21 (1990). [https://doi.org/10.1016/0005-1098\(90\)90154-A](https://doi.org/10.1016/0005-1098(90)90154-A)
18. Loizou, N., Richtárik, P.: Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Comput. Optim. Appl.* **77**(3), 653–710 (2020)
19. Madden, L., Becker, S., Dall'Anese, E.: Bounds for the tracking error of first-order online optimization methods. *J. Optim. Theory Appl.* **189**, 437–457 (2021)
20. Nesterov, Y.: *Introductory Lectures on Convex Optimization: A Basic Course*, 1st edn. Springer, Incorporated (2014)
21. Polyak, B.T.: *Introduction to Optimization* (Translations Series in Mathematics and Engineering). Optimization Software (1987). ISBN 0911575146. <https://www.amazon.com/dp/0911575146>

com/Introduction-Optimization-Translations-Mathematics-Engineering/dp/0911575146?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimb05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0911575146

22. Popkov, A.Y.: Gradient methods for nonstationary unconstrained optimization problems. *Autom. Remote. Control.* **66**(6), 883–891 (2005). <https://doi.org/10.1007/s10513-005-0132-z>
23. Robbins, H., Monro, S.: A stochastic approximation method. *Ann. Math. Stat.* **22**(3), 400–407 (1951). (<http://www.jstor.org/stable/2236626>)
24. Sahu, S., Hogan, C.J., Wells, M.T.: On the provable suboptimality of momentum SGD in nonstationary stochastic optimization. In: *Proceedings of the 43rd International Conference on Machine Learning* (2026)
25. Simonetto, A., Mokhtari, A., Koppel, A., Leus, G., Ribeiro, A.: A class of prediction-correction methods for time-varying convex optimization. *IEEE Trans. Signal Process.* **64**(17), 4576–4591 (2016). <https://doi.org/10.1109/TSP.2016.2568161>
26. Simonetto, A., Dall'Anese, E., Paternain, S., Leus, G., Giannakis, G.B.: Time-varying convex optimization: time-structured algorithms and applications. *Proc. IEEE* **108**(11), 2032–2048 (2020)
27. Vaswani, S., Bach, F., Schmidt, M.: Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204. PMLR (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.