On the Provable Suboptimality of Momentum SGD in Nonstationary Stochastic Optimization

Sharan Sahu*
ss4329@cornell.edu
Department of Statistics and Data Science
Cornell University

Cameron J. Hogan*
cjh337@cornell.edu
Department of Statistics and Data Science
Cornell University

Martin T. Wells
mtw1@cornell.edu
Department of Statistics and Data Science
Cornell University

January 18, 2026

Abstract

While momentum-based acceleration has been studied extensively in deterministic optimization problems, its behavior in nonstationary environments—where the data distribution and optimal parameters drift over time—remains under-explored. We analyze the tracking performance of Stochastic Gradient Descent (SGD) and its momentum variants (Polyak Heavy-Ball and Nesterov) under uniform strong convexity and smoothness in varying step-size regimes. We derive finite-time, expectation and high-probability bounds for the tracking error, establishing a sharp decomposition into three components: a transient initialization term, a noise-induced variance term, and a drift-induced tracking lag. Crucially, our analysis uncovers a fundamental trade-off: while momentum can suppress gradient noise, it incurs an explicit penalty on the tracking capability. We show that momentum can substantially amplify drift-induced tracking error, with amplification that becomes unbounded as the momentum parameter approaches one, formalizing the intuition that satisfying "stale" gradients hinders adaptation to rapid regime shifts. Complementing these upper bounds, we establish minimax lower bounds for dynamic regret under gradient-variation constraints. These lower bounds prove that the inertia-induced penalty is not an artifact of analysis but an information-theoretic barrier: in drift-dominated regimes, momentum creates an unavoidable "inertia window" that fundamentally degrades performance. Collectively, these results provide a definitive theoretical grounding for the empirical instability of momentum in dynamic environments and delineate the precise regime boundaries where SGD provably outperforms its accelerated counterparts.

1 Introduction

Consider the optimization problem posed by a strongly convex objective function $G : \mathbb{R}^d \to \mathbb{R}$ defined as follows:

$$\boldsymbol{\theta}^\star \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{X\sim\Pi}[g(\boldsymbol{\theta}, X)].$$

Here $g(\boldsymbol{\theta}, X)$ is a noise perturbed measurement of $G(\boldsymbol{\theta})$ and X is a random variable sampled from a distribution Π. This is a classic formulation of stochastic optimization which serves as the theoretical and practical backbone of modern machine learning and has been explored across many domains [BB07, Pow19, CLLZ24, WS21, GCPB16]. Over the last two decades, finite-time guarantees for stochastic approximation methods have matured substantially [ACDL14, MB11, LXZ19, NJLS09], but the predominant assumption is stationarity of Π i.e. the data are generated from a fixed distribution throughout the run of the algorithm. In many real systems—especially nonstationary time series—this assumption is violated due to regime changes, concept drift, or unmodeled exogenous effects and results in

*These authors contributed equally to this work.

the data distribution evolving over time as Π_t. These settings naturally connect to classical stochastic tracking in signal processing [KY97, Say03] and concept drift in online learning [Haz16, CDH21], where the goal is not to converge to a single static solution, but to track an unknown, time-varying sequence of minimizers as closely as possible.

The primary algorithmic tool for this tracking task, provided that g is differentiable, is Stochastic Gradient Descent (SGD) [RM51], especially when dealing with high-dimensional covariates X_t and the number of training data points is large. SGD replaces the unavailable population gradient by noisy gradients computed from samples (or mini-batches). Due to its simplicity, scalability to big data, and robustness to noise and uncertainty, stochastic gradient methods have become popular in large-scale optimization, machine learning, and data mining applications [Zha04, CBS14, SLJ+15].

Developing methods to accelerate the convergence of SGD has been studied extensively. Momentum methods, such as Polyak's heavy-ball (HB) [Pol63] and Nesterov acceleration (NAG) [Nes83], are ubiquitous in deep learning and are widely believed to reduce gradient noise and accelerate training [SMDH13]. It is widely known for a variety of strongly convex and nonconvex settings that SGD with momentum provides faster convergence and better generalization compared to SGD without momentum [RKAC+24, SI24, LGY20, CM20]. Yet this deterministic intuition becomes fragile in genuinely nonstationary regimes: when the data distribution Π_t drifts, past gradients may become systematically *stale*, so the same temporal averaging that suppresses noise can induce tracking lag and even instability, as observed empirically in online forecasting and suggested by prior steady-state analyses in the slow-adaptation regime.

These observations motivate the following central question in the nonstationary tracking setting:

When does momentum help—and when does it provably hurt*—in stochastic optimization under distribution shift? Concretely, can we (a) construct explicit drifting instances where Heavy-Ball/Nesterov incur an unavoidable* inertia window *and are uniformly worse than SGD, and (b) characterize the precise regime boundaries (in terms of drift, noise, conditioning, and momentum β) under which momentum improves variance reduction without sacrificing tracking accuracy?*

Our contributions. We answer this question for μ–strongly convex, L–smooth risks under predictable distribution shift and make the following advances:

1. **Finite-time tracking guarantees and a concrete momentum penalty.** We prove finite-time tracking bounds that cleanly separate three effects: forgetting of the initialization, an irreducible noise floor, and an irreducible tracking lag due to drift. Extending the analysis to Heavy-Ball and Nesterov, we obtain the same decomposition but show that momentum incurs an explicit penalty as β tends to 1: it becomes more sensitive to initialization, inflates the effective noise level, and requires a more conservative stepsize for stability. These effects are especially pronounced in ill-conditioned problems.

2. **Drift-adaptive high-probability bounds that expose a new drift–noise coupling.** Prior high-probability results often summarize nonstationarity by a single worst-case drift parameter, which cannot distinguish bursty regime shifts from persistent drift and ignores exponential forgetting. In contrast, our bounds are *time-resolved*: only recent drift meaningfully affects the guarantee. Moreover, the bounds reveal a *drift–noise interaction* mechanism: drift increases tracking mismatch, which in turn amplifies stochastic fluctuations. This structure provides a principled motivation for restart/windowing/forgetting rules that truncate stale history and thereby reduce both drift accumulation and variance inflation after regime changes.

3. **Minimax lower bounds and the "inertia window" as an information-theoretic barrier for momentum.** We establish minimax dynamic-regret lower bounds over a standard gradient-variation budget. The lower bound contains an explicit momentum-dependent term and formalizes an unavoidable *inertia window*: with large momentum, the method is provably "behind" a drifting optimizer for a nontrivial period, regardless of tuning. In the uniformly spread drift regime, our tuned upper bound matches the minimax rate and exhibits the sharp momentum penalty.

4. **Nonstationary numerical experiments that corroborates the predicted regime split.** We evaluate SGD, Heavy-Ball, and Nesterov on drifting quadratics, drifting linear/logistic regression, and a drifting teacher–student MLP. Across tasks, the results match the theory: momentum can help in near stationary/noise-dominated settings, while under genuine drift—especially with large β or ill-conditioning—momentum becomes brittle and SGD is uniformly more robust.

1.1 Related Work

1.1.1 Momentum and Acceleration in Stochastic Optimization

Developing methods to accelerate the convergence of SGD has been studied extensively. Momentum methods, such as Polyak's heavy-ball (HB) [Pol63] and Nesterov acceleration (NAG) [Nes83], are ubiquitous in deep learning and are widely believed to reduce gradient noise and accelerate training [SMDH13]. Adaptive variants such as Adagrad and Adam further modify the effective step-size online [DHS11, KB15, KB17, LH19, Zei12, LK24], and the general research landscape of second-order and filtering based methods has been studied greatly [Yan23, YGS+21, Oll19, Vuc18, DSC+22, ZRS+18, ZDT+21, ZJDC20, LLZ+25, GBLP25, ZZL+19, IKS22, LJH+20, JKR25, ZWU21, LXL19, GKS18, VMZ+25, YYC+25, YYC+25, Doz16], with methods being developed specifically for optimizing the training of large neural networks as of late [LSY+25, HZJ+25, BWAA18]. It is widely known for a variety of strongly convex and nonconvex settings that SGD with momentum provides faster convergence and better generalization compared to SGD without momentum [RKAC+24, SI24, LGY20, CM20]. While it is relatively well-known that a fixed momentum rate is not as effective as adaptive momentum [SMDH13], addressing how to tune momentum optimally is an active area of research. In practice, it is common to fix the momentum parameter to a standard value (e.g., $\beta = 0.9$) [KB17, CPM+17], or use a simple scheduler such as exponential momentum decay, cosine momentum, etc.

1.1.2 Limitations and Instabilities of Momentum

While popular, momentum does not always improve training [WMW+24, FWZ+23, DSLL24]. For online and adaptive learning, prior work suggests that the deterministic acceleration intuition does not directly transfer: in the slow-adaptation regime, momentum can be fundamentally equivalent to standard SGD with a re-scaled step-size, and can even degrade steady-state performance unless parameters are chosen carefully [YYS16]. A key challenge in nonstationary time series is that *past gradients become stale*: gradients computed under Π_{t-k} can systematically point in the wrong direction for the current objective G_t. This motivates discounting or forgetting historical gradient information [ZWL20]. Furthermore, empirical evidence in online forecasting has pointed to SGD with momentum not being as stable as SGD without momentum, especially as learning rates grow [AZF19]. In [JJMS25] they show that in high dimensional online settings, momentum can amplify noise and degrade the performance of SGD when step size is kept constant. In [DSLL24], they show that momentum degrades the stability of SGD when β approaches 1. Methods have been developed to tune momentum throughout the training process in a more stable manner, such as using local quadratic approximations [ZM19], combining different loss planes [TC25], momentum decay [CWLK22], or passive damping [LSZG19]. Although research has been done on understanding the effects of momentum theoretically [JL22, LK24, GLZX19, RKK18, Qia99, KS21], the noisy, time-dependent, non-stationary environment is understudied.

1.1.3 Optimization under Time-Dependent and Non-Stationary Environments

In [YYS16] they establish that in online stochastic optimization with constant step-sizes, momentum methods are essentially equivalent to standard SGD with a re-scaled learning rate at the steady-state. Their results indicate that momentum does not inherently lower the mean-square error (MSE) floor under persistent noise when the target is fixed. However, their analysis is primarily situated in the stationary regime; it does not account for the tracking lag and stability issues that arise when the underlying minimizer $\boldsymbol{\theta}_t^\star$ evolves as a non-stationary process. This is also supported in [KS21]. In [YPT17], they present a novel online gradient learning method to deal with time dependent data containing outliers by modifying the Adam Optimizer to suggest which data points are outliers. In [GTG+24], they do a soft parameter reset for neural network parameters in the non-stationary regime modeled as an Ornstein-Uhlenbeck process with an adaptive drift parameter that pulls weights toward a prior distribution. In [CDH23] they give non-asymptotic efficiency estimates for the tracking error of the proximal stochastic gradient method under time drift. In [SZZ26] they establish optimal estimation and regret bounds for SGD under temporally dependent data. [Vid25] studies the stability and convergence of momentum-based methods where the objective function is allowed to vary with time. They do not address non-stationarity explicitly nor do they prove any bounds on the tracking error. In [CZP19], they study online stochastic optimization under time-varying data distributions and derive dynamic regret bounds for SGD. In [ZZZZ24] they study online convex optimization in non-stationary environments and develop problem-dependent dynamic regret bounds that adapt to the variation of the optimal solution and the loss sequence over time.

2 Preliminaries

2.1 Notations

We will denote scalars and vectors by lowercase and bold lowercase letters, respectively, and denote matrices by uppercase boldface letters. We use calligraphic letters to denote sets, operators, or σ-algebras. For a vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we denote the ℓ_2 norm by $\|\mathbf{x}\|$. We also denote the inner product between two vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ as $\langle \mathbf{x}, \mathbf{x}' \rangle$. The transpose of a matrix $\mathbf{A}$ is indicated by $\mathbf{A}^\top$.

Given two sequences $\{a_m\}$ and $\{b_m\}$, we write $a_m = \mathcal{O}(b_m)$ if there exists a positive constant $0 < C < +\infty$ such that $a_m \le C b_m$, $a_m = \Theta(b_m)$ if there exists two positive constants $C_1, C_2 > 0$ such that $a_m \le C_1 b_m$ and $b_m \le C_2 a_m$, and $a_m = \Omega(b_m)$ if there exists a positive constant $0 < C < +\infty$ such that $a_m \ge C b_m$. We write $a_m \lesssim b_m$ if there exists a constant $C > 0$ that may depend only on fixed absolute constants (e.g., universal numerical constants) but not on m or the varying problem parameters such that $a_m \le C\, b_m$. We define $a_m \gtrsim b_m$ and $a_m \asymp b_m$ analogously.

For a function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$, we denote its gradient as $\nabla f(\mathbf{x})$. The minimum value of $f(\mathbf{x})$ is denoted as $\min f$ and the minimizer of $f(\mathbf{x})$ is denoted as $\mathbf{x}^*$. We use $\mathbb{E}[\cdot]$ to denote the expectation with respect to the underlying probability measure. We also denote $\mathcal{F}_t = \sigma(X_0, X_1, \ldots, X_t)$ to denote the natural filtration i.e. the σ-algebra generated by random variables $X_0, \cdots, X_t$.

2.2 Orlicz norm and conditional Orlicz norms

Let X be a real-valued random variable and let $\alpha \ge 1$. The Ψ_α–Orlicz norm is

$$\|X\|_{\Psi_\alpha} \;:=\; \inf\Big\{u > 0 :\; \mathbb{E}\exp\big(|X/u|^\alpha\big) \le 2\Big\}.$$

Let $\mathcal{F}$ be a σ-field. A random variable $K_\mathcal{F}$ is said to be $\mathcal{F}$-measurable if $K_\mathcal{F}$ is measurable with respect to $\mathcal{F}$ (equivalently, $K_\mathcal{F} \in \mathcal{F}$).

Definition 2.1 (Conditional Orlicz norm)**.** Let $\mathcal{F}$ be a σ-field and let $K_\mathcal{F} > 0$ be an $\mathcal{F}$-measurable random variable. We say that X satisfies $\|X \mid \mathcal{F}\|_{\Psi_\alpha} \le K_\mathcal{F}$ if and only if

$$\mathbb{E}\Big[\exp\Big(\big|X/K_\mathcal{F}\big|^\alpha\Big)\,\Big|\,\mathcal{F}\Big] \le 2, \qquad \text{a.s.} \tag{2.1}$$

Note that the conditional expectation in (2.1) is itself an $\mathcal{F}$-measurable random variable. For a random vector $\mathbf{X} \in \mathbb{R}^d$, define

$$\|\mathbf{X}\|_{\Psi_\alpha} \;:=\; \sup_{\mathbf{u}\in\mathbb{S}^{d-1}} \|\mathbf{u}^\top \mathbf{X}\|_{\Psi_\alpha}, \qquad \mathbb{S}^{d-1} := \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_2 = 1\}.$$

Definition 2.2 (Conditional Orlicz norm (vector))**.** Let $\mathcal{F}$ be a σ-field and let $K_\mathcal{F} > 0$ be $\mathcal{F}$-measurable. We say that $\mathbf{X}$ satisfies $\|\mathbf{X} \mid \mathcal{F}\|_{\Psi_\alpha} \le K_\mathcal{F}$ if and only if

$$\sup_{\mathbf{u}\in\mathbb{S}^{d-1}, \mathbf{u}\in\mathcal{F}} \mathbb{E}\Big[\exp\Big(\big|\mathbf{u}^\top \mathbf{X}/K_\mathcal{F}\big|^\alpha\Big)\,\Big|\,\mathcal{F}\Big] \le 2, \qquad \text{a.s.} \tag{2.2}$$

For a random matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, define

$$\|\mathbf{X}\|_{\Psi_\alpha} \;:=\; \sup_{\mathbf{u}\in\mathbb{S}^{d_1-1}} \sup_{\mathbf{v}\in\mathbb{S}^{d_2-1}} \|\mathbf{u}^\top \mathbf{X}\mathbf{v}\|_{\Psi_\alpha},$$

with conditional versions defined analogously (conditioning on $\mathcal{F}$ and allowing $\mathcal{F}$-measurable unit vectors in the suprema). These norms are well-defined and satisfy the usual norm properties (see Appendix D.1 for more details).

2.3 Problem setup

Let $G : \mathbb{R}^d \to \mathbb{R}$ be a strongly convex objective function. Recall the following non-stationary stochastic optimization problem:

$$\boldsymbol{\theta}_t^\star \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{X_t\sim\Pi_t}[g(\boldsymbol{\theta}, X_t)].$$

Here $g(\boldsymbol{\theta}, X_t)$ is a noise perturbed measurement of $G(\boldsymbol{\theta})$ and X_t is a random variable sampled from a distribution Π_t. Let $(\mathcal{F}_t)_{t\geq 1}$ be the natural filtration $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$. Throughout the entirety of our analysis, we will be imposing an assumption modeling stochasticity in the non-stationary setting:

Assumption 2.1 (Stochastic predictability framework)**.** *There exists a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Let $(X_t)_{t\geq 0}$ be an $\mathbb{F}$–adapted process, i.e., X_t is $\mathcal{F}_t$–measurable for all t. For each $t \geq 0$, define the conditional risk $F_{t+1}(\boldsymbol{\theta}) := \mathbb{E}\big[g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t\big]$, and let $\boldsymbol{\theta}^\star_{t+1}$ denote its (a.s. unique) minimizer: $\boldsymbol{\theta}^\star_{t+1} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} F_{t+1}(\boldsymbol{\theta})$. Assume the following hold for all $t \geq 0$:*

i) ***(Predictable target)*** *$\boldsymbol{\theta}^\star_{t+1}$ is $\mathcal{F}_t$–measurable.*

ii) ***(Algorithm adaptedness)*** *The iterate $\boldsymbol{\theta}_t$ is $\mathcal{F}_t$–measurable.*

iii) ***(Martingale difference noise)*** *Define the conditional mean gradient $m_{t+1}(\boldsymbol{\theta}) = \mathbb{E}\big[\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t\big]$, and the gradient noise $\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, X_{t+1}) - m_{t+1}(\boldsymbol{\theta})$. Then $\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta})$ is $\mathcal{F}_{t+1}$–measurable and satisfies $\mathbb{E}[\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}) \mid \mathcal{F}_t] = \mathbf{0}$ a.s. for all $\boldsymbol{\theta} \in \mathbb{R}^d$.*

Our parameter updates using the standard SGD update can then be written as

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \gamma_t \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_{t+1}) \\ &= \boldsymbol{\theta}_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t) - \gamma_t \xi_{t+1}(\boldsymbol{\theta}_t). \end{aligned} \tag{2.3}$$

For SGD with momentum, we will present a generalized momentum stochastic gradient method which can capture both Polyak's heavy-ball method and Nesterov's acceleration method as special cases. Consider the following with two momentum parameters $\beta_1, \beta_2 \in [0, 1)$:

$$\begin{aligned} \boldsymbol{\psi}_t &= \boldsymbol{\theta}_t + \beta_1(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\psi}_t - \gamma_t \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\psi}_t, X_{t+1}) + \beta_2(\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}). \end{aligned} \tag{2.4}$$

When $\beta_1 = 0$ and $\beta_2 = \beta$, then we recover Polyak's heavy-ball method. If $\beta_2 = 0$ and $\beta_1 = \beta$, we recover Nesterov's accelerated method. Thus to capture both these methods simultaneously, we assume $\beta_1 + \beta_2 = \beta$ and $\beta_1\beta_2 = 0$ for some constant fixed $\beta \in [0, 1)$. We adapt this formulation from [YYS16]. Using this, our parameter updates can be written as follows:

$$\begin{aligned} \boldsymbol{\psi}_t &= \boldsymbol{\theta}_t + \beta_1(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\psi}_t - \gamma_t m_{t+1}(\boldsymbol{\psi}_t) - \gamma_t \xi_{t+1}(\boldsymbol{\psi}_t) + \beta_2(\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}). \end{aligned} \tag{2.5}$$

Note in this setup that $\boldsymbol{\theta}_{t+1}$ is not $\mathcal{F}_t$ measurable. Rather if we define $\mathcal{F} \subset \mathcal{F}_t^+ = \sigma(X_0, \ldots, X_t, X_{t+1})$, then $\boldsymbol{\theta}_{t+1}$ is $\mathcal{F}_t^+$ measurable. In order to examine the convergence of standard and momentum based stochastic gradient methods, we will need to make some assumptions on the conditional mean gradient map $\boldsymbol{\theta} \mapsto m_t(\boldsymbol{\theta})$. This amounts to assuming that the map $\boldsymbol{\theta} \mapsto m_t(\boldsymbol{\theta})$ is uniformly μ-strongly monotone and Lipschitz. Note that this is equivalent to assuming that $\mathbb{E}_{X_t\sim\Pi_t}[g(\boldsymbol{\theta}, X_t)]$ is μ-strongly convex and has Lipschitz gradients (under mild technical conditions for which the Dominated Convergence Theorem holds). Under strong convexity, the minimizer $\boldsymbol{\theta}^\star_t$ is unique. It should be noted that these conditions are standard in the analysis of SGD and SGD with momentum in the literature [YYS16] and are satisfied by many problems of interest, especially when regularization is used (e.g. mean-square error risks, logistic risks, etc).

Assumption 2.2 (Uniform μ-strongly monotone)**.** *For all $t \geq 0$, there exists a constant $\mu > 0$ such that $\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$,*

$$\langle m_{t+1}(\boldsymbol{\theta}) - m_{t+1}(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle \geq \mu \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2.$$

Assumption 2.3 (Uniform Lipschitz continuity)**.** *For all $t \geq 0$, there exists a constant $L > 0$ such that $\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$,*

$$\|m_{t+1}(\boldsymbol{\theta}) - m_{t+1}(\boldsymbol{\theta}')\| \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

3 Theoretical Results

3.1 Bounds in expectation

In this section, we will obtain guarantees on the tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|$ in expectation. The intermediate steps used to get this bound will be useful to obtain high-probability bounds that we discuss later. Before proceeding, we will make a few assumptions regarding the second moments of the minimizer drift $\Delta_t = \|\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star\|$ and the gradient noise, simply for presentation. These assumptions are not essential for the result, but we adopt them to simplify the exposition.

Assumption 3.1 (Conditional second-moment bounds)**.** *There exist constants $\Delta, \sigma > 0$ such that for all $t \geq 0$,*

i) ***(Minimizer drift)*** *The minimizer drift Δ_t satisfies $\mathbb{E}\big[\Delta_t^2 \mid \mathcal{F}_t\big] \leq \Delta^2$ a.s.*

ii) ***(Gradient noise along iterates)*** *The gradient noise $\boldsymbol{\xi}_{t+1}$ satisfies $\mathbb{E}\big[\|\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)\|^2 \mid \mathcal{F}_t\big] \leq \sigma^2$ a.s.*

We can now establish the following theorem that establishes the expected tracking error for SGD in nonstationary time-series environments. We defer the proofs for this section to Appendix A.1.

Theorem 3.1 (Tracking error bound in expectation for SGD)**.** *Under Assumption 3.1, for $\forall t \geq 0$ and $\gamma \leq \min\{\mu/L^2, 1/L\}$, the following tracking error bound holds in expectation for SGD:*

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \leq \left(1 - \frac{\gamma\mu}{2}\right)^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{4\Delta^2}{\gamma^2\mu^2} + \frac{\sigma^2\gamma}{\mu}.$$

This bound is similar to [CDH21] as it consists of a *contraction* term that arises from optimization which decays linearly in t, a *drift/tracking* term that depends on the minimizer drift, and a *noise* term. These contributions are *irreducible* for constant stepsize: even as $t \to \infty$, SGD cannot converge arbitrarily close to $\boldsymbol{\theta}_t^\star$ because (i) the optimizer itself moves over time (drift), and (ii) stochastic gradients inject persistent variance (noise). Moreover, the two steady-state terms exhibit an explicit stepsize tradeoff: larger γ amplifies the noise floor, while smaller γ reduces the noise but worsens tracking of a moving target. Consequently, we can show that after a burn-in period which varies based on whether we use a constant stepsize or epoch-wise step-decay schedule, we can reach the irreducible floor of order $O\left(\sigma^2\gamma/\mu + \Delta^2/\gamma^2\mu^2\right)$.

Theorem 3.2 (Time to reach the asymptotic tracking error in expectation for SGD)**.** *Assume $\gamma_t \in (0, 1/(2L)]$ for all $t \geq 0$. For any constant $\gamma \in (0, 1/2L]$, define*

$$\mathcal{E}(\gamma) := \frac{\sigma^2\gamma}{\mu} + \frac{4\Delta^2}{\mu^2\gamma^2}, \qquad \gamma^\star \in \arg\min_{\gamma \in (0,1/2L]} \mathcal{E}(\gamma), \qquad \mathcal{E} := \mathcal{E}(\gamma^\star).$$

Then we have the following:

(i) ***(Constant learning rate).*** *If $\gamma_t \equiv \gamma^\star$, then for all $t \geq 0$,*

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \lesssim \mathcal{E} \text{ after time } t \lesssim \frac{1}{\mu\gamma^\star}\log\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{\mathcal{E}}\right).$$

(ii) ***(Step-decay schedule in the low drift-to-noise regime).*** *Suppose $\gamma^\star < 1/2L$ (equivalently, the minimizer of $\mathcal{E}(\gamma)$ is not at the smoothness cap), so that*

$$\gamma^\star = \left(\frac{8\Delta^2}{\mu\sigma^2}\right)^{1/3}, \qquad \mathcal{E} = 3\left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3}.$$

Define epochs $k = 0, 1, \ldots, K-1$ with

$$\gamma_0 := \frac{1}{2L}, \qquad \gamma_k := \frac{\gamma_{k-1} + \gamma^\star}{2} \quad (k \geq 1), \qquad K := 1 + \left\lceil \log_2\left(\frac{\gamma_0}{\gamma^\star}\right)\right\rceil,$$

and epoch lengths

$$T_0 := \left\lceil \frac{2}{\mu\gamma_0}\log\left(\frac{2\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{\mathcal{E}(\gamma_0)}\right)\right\rceil, \qquad T_k := \left\lceil \frac{2\log 4}{\mu\gamma_k}\right\rceil \quad (k \geq 1).$$

Run SGD with constant stepsize γ_k for T_k iterations in epoch k, starting from θ_0. Let $T := \sum_{k=0}^{K-1} T_k$ be the total horizon. Then the final iterate satisfies

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \lesssim \mathcal{E} \text{ after time } t \lesssim \frac{L}{\mu}\log\Big(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2}{\mathcal{E}}\Big) + \frac{\sigma^2}{\mu^2\mathcal{E}}.$$

Theorem 3.1 and Theorem 3.2 give us algorithmic guarantees for running SGD in non-stationary environments. These closely match the guarantees provided by [CDH21] and are identical to the efficiency estimates in the static setting [Lan11]. Since the velocity iterate $\boldsymbol{v}_t$ is tightly coupled with $\boldsymbol{\psi}_t$, a direct adaptation of standard SGD-style arguments is not straightforward. One must instead analyze a coupled recursion involving both $\boldsymbol{v}_t$ and $\boldsymbol{\psi}_t$, which typically necessitates a Lyapunov (stability) function approach. However, constructing an appropriate Lyapunov function in this setting is itself cumbersome, making the resulting analysis considerably more delicate. In light of Lemma 3.1, we can view SGD with momentum as a 2D dynamical system on extended state vectors. We defer the proof to Appendix A.2:

Lemma 3.1 (Extended 2D recursion for SGD with momentum)**.** *Under Assumption 2.2, Assumption 2.3, and $\beta_1+\beta_2 = \beta$ and $\beta_1\beta_2 = 0$ for fixed $\beta \in [0, 1)$, the SGD with momentum update equations can be transformed into the following extended recursion:*

$$\begin{bmatrix}\widehat{\boldsymbol{\theta}}_t \\ \check{\boldsymbol{\theta}}_t\end{bmatrix} = \begin{bmatrix} I_d - \frac{\gamma_t}{1-\beta}\mathbf{H}_{t-1} & \frac{\gamma_t\beta'}{1-\beta}\mathbf{H}_{t-1} \\ -\frac{\gamma_t}{1-\beta}\mathbf{H}_{t-1} & \beta I_d + \frac{\gamma_t\beta'}{1-\beta}\mathbf{H}_{t-1}\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\theta}}_{t-1} \\ \check{\boldsymbol{\theta}}_{t-1}\end{bmatrix} + \frac{1}{1-\beta}\begin{bmatrix}-(I_d - \gamma_t\mathbf{H}_{t-1})\Delta_{t-1} - \mathbf{K}_{t-1}\Delta_{t-2} \\ -(I_d - \gamma_t\mathbf{H}_{t-1})\Delta_{t-1} - \mathbf{K}_{t-1}\Delta_{t-2}\end{bmatrix} + \frac{\gamma_t}{1-\beta}\begin{bmatrix}\xi_{t+1}(\boldsymbol{\psi}_{t-1}) \\ \xi_{t+1}(\boldsymbol{\psi}_{t-1})\end{bmatrix}$$

where

$$\beta' \triangleq \beta\beta_1 + \beta_2 \tag{3.1}$$

$$\mathbf{H}_{t-1} \triangleq \int_0^1 \nabla^2 F_{t+1}(\boldsymbol{\theta}^\star_{t+1} + s(\boldsymbol{\psi}_t - \boldsymbol{\theta}^\star_{t+1}))ds \tag{3.2}$$

$$\mathbf{K}_{t-1} = -\beta I_d + \gamma_t\beta_1\mathbf{H}_{t-1}. \tag{3.3}$$

This characterization of SGD with momentum proves very useful as it allows us to decouple the interaction between the noise induced from stochastic gradients, the error being accumulated by using momentum, and the effect of momentum on the drift. We now will provide an expectation bound on the tracking error for SGD with momentum. One will note that we incur extra scaling factors of $(1-\beta)^{-1}$ and $(1-\beta)^{-2}$ which explains why momentum can be significantly worse than SGD in non-stationary environments.

Theorem 3.3 (Tracking error bound in expectation for SGD with momentum)**.** *Let Assumption 2.2, Assumption 2.3, and Assumption 3.1. Consider the momentum stochastic gradient method (2.5) and the extended 2D recursion (A.13). Then when the step-sizes γ_t satisfies*

$$\gamma_t \leq \frac{\mu(1-\beta)^2}{4L^2}, \tag{3.4}$$

the following tracking error bound holds in expectation for SGD with momentum:

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \leq \frac{4}{(1-\beta)^2}\rho^{2(t+1)}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2 + \frac{48(1+\beta+\gamma L)^2}{(1-\beta)^2}\cdot\frac{\Delta^2}{(1-\rho)^2} + \frac{48}{(1-\beta)^2}\cdot\frac{\sigma^2\gamma^2}{1-\rho^2}. \tag{3.5}$$

In particular taking $\rho = 1 - \gamma\mu/2(1-\beta)$, we obtain:

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \leq \frac{48}{(1-\beta)^2}\exp\left(-\frac{\gamma\mu}{1-\beta}(t+1)\right)\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2 + \frac{192(2+\beta)^2}{\gamma^2\mu^2}\Delta^2 + \frac{96\sigma^2\gamma}{\mu(1-\beta)}. \tag{3.6}$$

Theorem 3.3 admits the same three-way decomposition as SGD: an exponentially decaying *optimization contraction* plus two *irreducible* floors from drift and gradient noise, but with a sharper dependence on the momentum parameter. Notably, increasing momentum amplifies sensitivity to initialization through the prefactor $(1-\beta)^{-2}$ and inflates the steady-state noise-driven tracking error by $(1-\beta)^{-1}$, consistent with the interpretation of momentum as increasing the effective memory of stochastic gradients. Moreover, the exponential rate depends on $\gamma\mu/(1-\beta)$. Under standard stability choices $\gamma \lesssim 1/L$, one has $\gamma\mu \asymp 1/\kappa$ with $\kappa := L/\mu$ so ill-conditioned objectives can exhibit a long "burn-in"

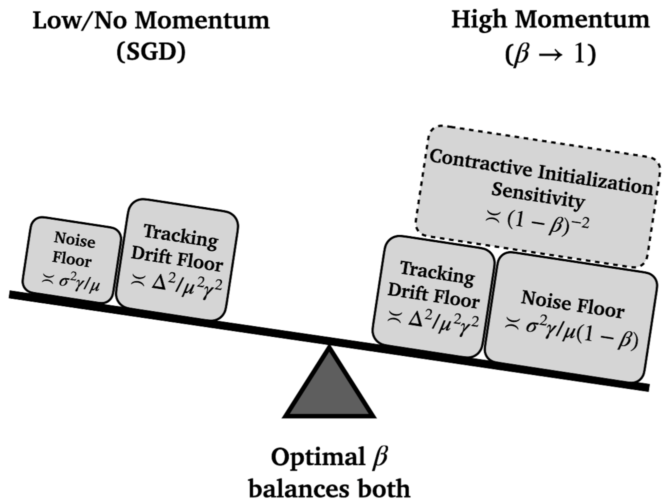


Figure 1: **Trade-off between drift tracking and stochastic noise under constant stepsize (expectation bounds).** For SGD, the expected tracking error decomposes into an exponentially decaying initialization term plus two irreducible steady-state floors: a noise floor $\asymp \sigma^2\gamma/\mu$ and a drift-induced tracking floor $\asymp \Delta^2/(\mu^2\gamma^2)$. For SGDM, momentum increases sensitivity to initialization by a factor $\asymp (1-\beta)^{-2}$ and inflates the noise floor by $\asymp (1-\beta)^{-1}$, while the drift floor retains the same γ^{-2} scaling (up to constants). Moreover, stability requires $\gamma \le \mu(1-\beta)^2/(4L^2)$, so large β can indirectly worsen drift tracking by forcing smaller admissible stepsizes. Together, these effects formalize when momentum helps in stationary regimes yet becomes fragile under drift, and when SGD is provably more robust.

before the contraction becomes negligible. This effect is particularly relevant for deep networks, whose curvature spectra are widely observed to be highly ill-conditioned with many near-flat directions (effectively small μ), so the initialization term can dominate over the short horizons typical of nonstationary regimes [SBL17, JKB+19, GKX19, Pap20]. Finally, while Nesterov-style acceleration improves deterministic gradient descent from $O(\kappa)$ to $O(\sqrt{\kappa})$ iterations on smooth strongly convex problems [Nes83, Nes14], our bound highlights that in drifting stochastic settings, performance is governed by tracking and variance floors rather than deterministic bias decay. Consequently, the classical $\sqrt{\kappa}$ acceleration effect does not need to take effect, and aggressive momentum can primarily raise the noise floor and prolong the burn-in time. This becomes clear in light of Theorem 3.4 which illustrates that momentum induces much longer burn-in periods:

Theorem 3.4 (Time to reach the asymptotic tracking error in expectation for SGD with momentum)**.** *Assume Assumption 2.2 Assumption 2.3, and Assumption 3.1 and let $\beta \in [0,1)$. Consider SGD with momentum (2.5). Assume a constant stepsize $\gamma_t \equiv \gamma$ satisfying $\gamma \le \mu(1-\beta)^2/(4L^2)$, and set*

$$\rho \;:=\; 1 - \frac{\gamma\mu}{2(1-\beta)} \in (0,1).$$

Define the (stepsize-dependent) steady-state tracking error

$$\mathcal{E}_\beta(\gamma) = \frac{192(2+\beta)^2}{\mu^2\gamma^2}\Delta^2 + \frac{96}{\mu(1-\beta)}\sigma^2\gamma, \quad \gamma^\star_\beta \in \arg\min_{\gamma\in(0,\ \mu(1-\beta)^2/(4L^2)]} \mathcal{E}_\beta(\gamma), \quad \mathcal{E}_\beta := \mathcal{E}_\beta(\gamma^\star_\beta). \tag{3.7}$$

Then:

(i) ***(Constant learning rate).*** *If $\gamma_t \equiv \gamma^\star_\beta$, then for all $t \ge 0$,*

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \;\lesssim\; \mathcal{E}_\beta \quad \text{after time} \quad t \;\lesssim\; \frac{1-\beta}{\mu\gamma^\star_\beta}\log\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2}{(1-\beta)^2\mathcal{E}_\beta}\right).$$

(ii) ***(Step-decay schedule with momentum restart).*** *Suppose $\gamma^\star_\beta < \mu(1-\beta)^2/(4L^2)$ (i.e., the minimizer of $\mathcal{E}_\beta(\gamma)$ is*

not at the stability cap). Define the epoch stepsizes

$$\gamma_0 := \frac{\mu(1-\beta)^2}{4L^2}, \qquad \gamma_k := \frac{\gamma_{k-1}+\gamma_\beta^\star}{2} \quad (k \geq 1), \qquad K := 1 + \left\lceil \log_2\left(\frac{\gamma_0}{\gamma_\beta^\star}\right)\right\rceil.$$

Define epoch lengths

$$T_0 := \left\lceil \frac{1-\beta}{\mu\gamma_0} \log\left(\frac{2\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{(1-\beta)^2\mathcal{E}_\beta(\gamma_0)}\right)\right\rceil, \qquad T_k := \left\lceil \frac{1-\beta}{\mu\gamma_k} \log 4\right\rceil \quad (k \geq 1).$$

Run SGD with momentum with constant stepsize γ_k for T_k iterations in epoch k, restarting the momentum buffer at the start of each epoch *(equivalently, set $v = 0$ or $\boldsymbol{\theta}_{t_k-1} = \boldsymbol{\theta}_{t_k}$ at epoch boundaries). Let $T := \sum_{k=0}^{K-1} T_k$ be the total horizon. Then the final iterate satisfies*

$$\mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \;\lesssim\; \mathcal{E}_\beta \quad \textit{after time} \quad T \;\lesssim\; \frac{L^2}{\mu^2(1-\beta)} \log\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{(1-\beta)^2\mathcal{E}_\beta}\right) \;+\; \frac{\sigma^2}{\mu^2\mathcal{E}_\beta},$$

up to universal numerical constants.

3.2 High probability bounds

To obtain high-probability guarantees on the tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|$, we will need to make a standard light-tail assumptions on the gradient noise [HLPR19, Lan11, CDH21]:

Assumption 3.2 (Conditional sub-Gaussian gradient noise along iterates)**.** *There exists a constant $\sigma > 0$ such that for all $t \geq 0$, $\left\|\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t) \mid \mathcal{F}_t\right\|_{\Psi_2} \leq \sigma$ a.s.*

It should be noted that prior high-probability tracking analyses often iterate an MGF recursion for $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2$, which necessitates light-tail assumptions on the drift $\Delta_t := \|\boldsymbol{\theta}_{t+1}^\star - \boldsymbol{\theta}_t^\star\|$ (e.g., conditional sub-exponentiality of Δ_t^2). Our proof instead uses a optional stopping-time argument for weighted martingale difference sums. Proceeding in this way will also highlight some insights that are missing from previous bounds for SGD in nonstationary settings.

Theorem 3.5 (High probability tracking error bound for SGD)**.** *Under Assumption 3.2, for all $t \in [T]$, $\gamma \leq \min\left\{\mu/L^2, 1/L\right\}$, and $\delta \in (0,1)$, the following tracking error bound holds for SGD with probability atleast $1-\delta$,*

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \lesssim \left(1 - \frac{\gamma\mu}{2}\right)^T \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{\mathfrak{D}_T}{\gamma\mu} + \left(d\sigma^2\gamma^2 + \frac{d^2\sigma^4\gamma^3}{\mu}\right)\log\frac{2T}{\delta} + \left(\frac{\sigma^2\gamma}{\mu} + \gamma^2\sigma^2\mathfrak{D}_t^{(2)}\right)\log\frac{2T}{\delta}$$

where $\mathfrak{D}_t := \sum_{\ell=0}^{t-1}(1-\gamma\mu/2)^{t-\ell-1}\Delta_\ell^2$ and $\mathfrak{D}_t^{(2)} := \sum_{\ell=0}^{t-1}(1-\gamma\mu/2)^{2(t-\ell-1)}\Delta_\ell^2$.

Prior high-probability guarantees for SGD in drifting environments typically control nonstationarity via a *uniform* drift bound, e.g., assuming light-tailed increments and summarizing drift by a single parameter Δ [CDH21]. Consequently, intermittent or localized nonstationarity (e.g., bursty regime shifts) is indistinguishable from persistent drift of magnitude Δ at all times, even though the dynamics exponentially forget older perturbations. In contrast, our bound is *drift-adaptive and time-resolved* and captures that (i) only recent nonstationarity drift affects the guarantee and (ii) drift amplifies stochastic fluctuations by increasing the tracking mismatch. This structure directly motivates restart/windowing/forgetting rules, which truncate or downweight the history and thereby reduce both drift accumulation and variance inflation after regime changes. As a consequence of Theorem 3.5, we can get guarantees similar to Theorem 3.2 by replicating the same argument. We exclude these results for SGD and SGDM for brevity.

Theorem 3.6 (High probability tracking error bound for SGDM)**.** *Under Assumption 3.2, for all $t \in [T]$, $\gamma \leq \min\left\{1/L, \mu(1-\beta)^2/4L^2\right\}$, and $\delta \in (0,1)$, provided one takes a zero momentum initialization $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0$, the following tracking error bound holds for SGD with probability atleast $1-\delta$,*

$$\begin{aligned}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \lesssim& \frac{2}{(1-\beta)^2}\exp\left(-\frac{\gamma^2\mu^2}{4(1-\beta)^2}T\right)\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{1}{\gamma\mu}\cdot\frac{1}{1-\beta}\mathfrak{D}_t^{\mathrm{mom}} + \frac{d\sigma^2}{\mu^2} \\ &+ \left(\frac{d\sigma^2\gamma^2}{(1-\beta)^2} + \frac{d^2\sigma^4\gamma^3}{\mu(1-\beta)^3}\right)\log\frac{2T}{\delta} + \left(\frac{\sigma^2}{\mu^2} + \frac{\sigma^2\gamma^2}{(1-\beta)^2}\mathfrak{D}_t^{\mathrm{mom},(2)}\right)\log\frac{2T}{\delta}\end{aligned} \tag{3.8}$$

where $\mathfrak{D}_t^{\mathrm{mom}} := \sum_{\ell \le T-1} \tilde{\rho}^{T-\ell-1} \|\mathbf{b}_\ell\|^2$, *and* $\mathfrak{D}_t^{\mathrm{mom},(2)} := \sum_{\ell \le T-1} \tilde{\rho}^{2(T-\ell-1)} \|\mathbf{b}_\ell\|^2$ *with* $\tilde{\rho} = 1 - \eta^2\mu^2/4$ *where* $\mathbf{b}_\ell := -(I_d - \gamma_t \mathbf{H}_{\ell-1})\Delta_{\ell-1} - \mathbf{K}_{\ell-1}\Delta_{\ell-2}$ *with* $\mathbf{H}_\ell, \mathbf{K}_\ell$ *defined as in Lemma 3.1.*

Our high-probability analysis reveals that while momentum accelerates convergence in stationary settings, it introduces a fundamental volatility in nonstationary settings that scales aggressively with the momentum parameter β. Specifically, the SGDM bound highlights three distinct failure modes absent in standard SGD. First, momentum introduces an effective temporal memory horizon of $(1-\beta)^{-1}$, creating an inertia-induced lag. Unlike SGD, which tracks the immediate drift, SGDM's tracking error depends on the acceleration of the minimizer ($\Delta_{\ell-1}$ and $\Delta_{\ell-2}$), causing it to overshoot targets during regime shifts. Second, we identify a critical drift-noise interaction term in the high-probability concentration scaling with $(1-\beta)^{-2}$. This pathwise coupling suggests that systematic tracking misalignment effectively "amplifies" the projection of stochastic gradient noise onto the error direction—a phenomenon obscured in expectation-based bounds and previous high-probability bounds [CDH21]. Finally, these effects are compounded in ill-conditioned regimes ($\kappa \gg 1$), where stability constraints force the step size to scale as $\gamma \lesssim (1-\beta)^2/\kappa L$. This creates a "stability-plasticity" dilemma: to prevent divergence, SGDM requires a step size so small that the transient decay rate $\exp(-\gamma^2\mu^2 T)$ becomes negligible, leaving the algorithm unable to forget initial conditions or react to drift. Consequently, for ill-conditioned, nonstationary problems, SGD serves as the strictly more robust estimator, avoiding the compounded variance and lag penalties inherent to the momentum mechanism.

3.3 Minimax Regret Bounds

We now discuss obtaining minimax regret (and subsequently tracking error) lower bounds for nonstationary stochastic optimization with distribution shift. To discuss this, we must reframe the problem. We consider nonstationary stochastic optimization in which the underlying sample loss is fixed but the data distribution shifts over time. This is in contrast to seminal work by [BGZ15, CWW19] which considers online nonstationary stochastic optimization where the loss function ℓ_t is time varying. Let $\Theta \subset \mathbb{R}^d$ be convex. At each time $t \in \{0, \ldots, T-1\}$ the algorithm chooses $\boldsymbol{\theta}_t \in \Theta$, then observes a fresh sample X_{t+1} whose (conditional) law may drift over time. This drift induces a time-varying *population risk*

$$F_{t+1}(\boldsymbol{\theta}) := \mathbb{E}[g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t], \qquad \boldsymbol{\theta}^\star_{t+1} \in \arg\min_{\theta \in \Theta} F_{t+1}(\boldsymbol{\theta}),$$

where $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$. We assume F_t is μ-strongly convex and L-smooth uniformly in t, so $\boldsymbol{\theta}^\star_t$ is a.s. unique. Define the conditional mean gradient $m_{t+1}(\boldsymbol{\theta}) = \mathbb{E}[\nabla g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t]$ and noise $\xi_{t+1}(\boldsymbol{\theta}) = \nabla g(\boldsymbol{\theta}, X_{t+1}) - m_{t+1}(\boldsymbol{\theta})$, so that $\mathbb{E}[\xi_{t+1}(\boldsymbol{\theta}) \mid \mathcal{F}_t] = 0$ a.s. We consider SGD and momentum updates driven by this noisy first-order feedback given by (2.3) and (2.5). We will measure performace via dynamic regret against the drifting minimizers,

$$\mathcal{R}_T^\pi(F) := \mathbb{E}\left[\sum_{t=0}^{T-1} \Big(F_{t+1}(\boldsymbol{\theta}_t) - F_{t+1}(\boldsymbol{\theta}^\star_{t+1})\Big)\right].$$

To obtain nontrivial guarantees, given a sequence of differentiable functions $f_1, \ldots, f_T : \Theta \to \mathbb{R}$, define the $L_{p,q}$ gradient-variational functional of $f = (f_1, \ldots, f_T)$ as

$$\mathrm{GVar}_{p,q}(f) := \begin{cases} \left(\frac{1}{T}\sum_{t=1}^{T-1} \|\nabla f_{t+1} - \nabla f_t\|_p^q\right)^{1/q} & 1 \le p \le \infty,\ 1 \le q < \infty \\ \max_{1 \le t \le T-1} \|\nabla f_{t+1} - \nabla f_t\|_p & q = \infty \end{cases} \tag{3.9}$$

where for any measurable function $f : \Theta \to \mathbb{R}$, we have

$$\|h\|_p := \begin{cases} \left(\int_{\mathcal{X}} \|h(x)\|_2^p \, dx\right)^{1/p} & p < \infty \\ \sup_{x \in \mathcal{X}} \|h(x)\|_2 & p = \infty. \end{cases} \tag{3.10}$$

We set a budget constraint defined by the function class

$$\mathcal{F}_{p,q}(V_T) := \left\{ f : \mathrm{GVar}_{p,q}(f) \le V_T \right\}. \tag{3.11}$$

This is again in contrast to [BGZ15, CWW19] who place constraints on the function values rather than the gradients. We do this since gradient variation naturally gives rise to distributional shift metrics such as the the Wasserstein distance

$\mathcal{W}_1(\Pi_{t+1}, \Pi_t)$ via Kantorovich-Rubenstein duality and the Total Variation (TV) distance. We also note that placing a gradient variation budget implies a budget on the minimizer drift by strong convexity. We now establish the following worst-case regret for noisy gradient feedback over $\mathcal{F}_{p,q}(V_T)$. We defer the proofs to Appendix C.

Theorem 3.7 (Minimax lower bound for strongly-convex function sequences using SGDM)**.** *Fix arbitrary $1 \le p < \infty$ and $1 \le q < \infty$. Suppose Assumption 2.2 and Assumption 2.3 hold. Consider the class Π_β of $\mathrm{SGDM}(\beta)$ policies with constant step size $\gamma \le c_0(1-\beta)^2/L$. Then there exists a class $\mathcal{F}_{p,q}(V_T)$ of μ-strongly convex, L-smooth function sequences whose gradient-variational functional budget satisfies $\mathrm{GVar}_{p,q}(f) \le \mathcal{V}_T$ such that*

$$\inf_{\pi \in \Pi_\beta} \sup_{F:\, \mathrm{GVar}_{p,q}(F) \le \mathcal{V}_T} \mathcal{R}_T^\pi(F) \gtrsim \max\Big\{(1-\beta)^{-2/(\alpha q+2)} \sigma^{4/(\alpha q+2)} \mu^{(\alpha q-2q-2)/(\alpha q+2)} \mathcal{V}_T^{2q/(\alpha q+2)} T^{\alpha q/(\alpha q+2)}, \\ (1-\beta)^{-2/(\alpha q)} \mu^{(\alpha q-2q-2)/(\alpha q)} L^{2/(\alpha q)} \mathcal{V}_T^{2/\alpha} T^{1-2/(\alpha q)}\Big\}. \tag{3.12}$$

where $\alpha = 1 + d/p$.

It should be noted that by strong convexity, this implies a lower bound on the tracking error (up to constant factors). Furthermore, it is worth noting that under uniformly spread drift (i.e., $q = 1$, $p = \infty$, and $\Delta \asymp \mathcal{V}_T/(\mu T)$), optimizing the constant stepsize in Theorem 3.4 yields a dynamic regret upper bound of the form $\mathcal{R}_T \lesssim (1-\beta)^{-2/3} \sigma^{4/3} \mu^{-1/3} \mathcal{V}_T^{2/3} T^{1/3}$ which matches the corresponding minimax lower bound in its dependence on $\mathcal{V}_T$, T, σ and crucially the momentum penalty $(1-\beta)^{-2/3}$.

The minimax regret bound consists of two terms. The first term corresponds to a statistical regime in which regret is forced primarily by noisy gradient feedback together with the gradient-variation budget. In this regime, performance is bottlenecked by information rather than by the algorithm's internal dynamics: no method can extract substantially more signal from the same noisy gradients under the same variation constraint. For SGDM, the dependence on $(1-\beta)$ in this term is comparatively mild (it enters through the stability-restricted tuning of the constant stepsize), reflecting that this regime is fundamentally governed by noise and variation rather than inertia. The second term is induced by a inertia-limited regime where delay dominates. Momentum reduces the impact of stochastic noise by averaging gradients over time, but this same averaging creates *inertia*: after a change in the environment, the iterate continues to follow stale gradient information. Thus, even under the best allowable tuning, high momentum creates an unavoidable "inertia window" of length $\Omega(\kappa/(1-\beta))$ where $\kappa = L/\mu$ under standard stability step-size restrictions.

These two terms have two concrete implications. First, in our upper bounds the drift-driven component of tracking error/regret is explicitly amplified as $\beta \uparrow 1$, whereas plain SGD does not incur this inertia amplification. Second, this phenomenon is not an artifact of analysis: the minimax lower bound constructs blockwise shifts (consistent with the gradient-variation budget) for which any $\mathrm{SGDM}(\beta)$ policy must spend τ_β steps per change in a transient regime where it is systematically misaligned with the current optimizer, thereby forcing an unavoidable regret penalty that deteriorates with β. Taken together, the theory explains the empirical dichotomy: in near-stationary regimes the noise term dominates and momentum can help, while under genuine distribution shift the inertia term dominates and SGD can be uniformly preferable because it adapts immediately rather than averaging gradients over time.

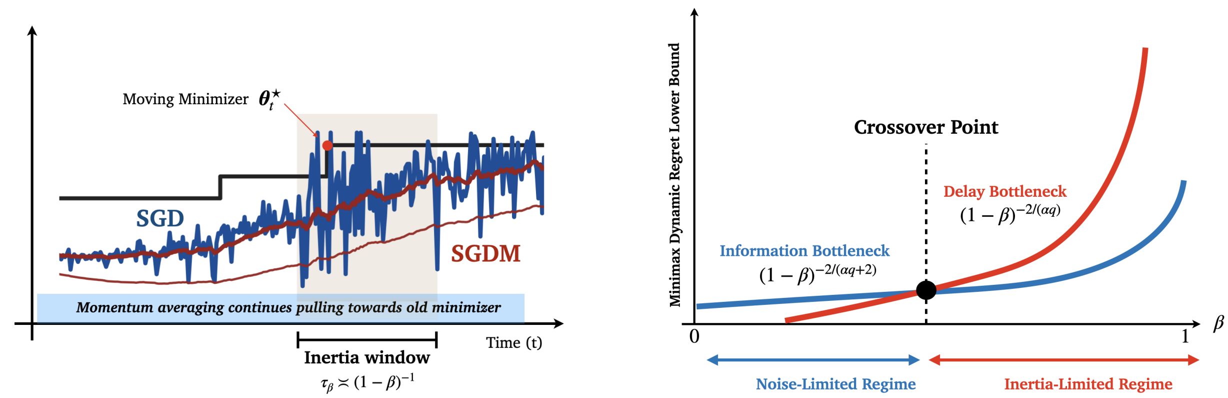


Figure 2: **Two minimax regimes and the inertia window.** **(a)** After a regime change (distribution shift), momentum averages past gradients: this reduces noise but induces *inertia*, so the iterate lags behind the drifting minimizer for an "inertia window" whose length grows with the momentum parameter. **(b)** The minimax lower bound in Theorem 3.7 is the maximum of two contributions: a noise/variation-limited term which is bottlenecked by information rather than inertia, and an inertia-limited term that worsens with momentum and dominates when delay is the bottleneck. The crossover separates a statistical regime from an inertia regime, explaining when momentum is provably worse than SGD.

4 Experimental Results

In this section, we empirically corroborate our theoretical findings by evaluating the tracking performance of SGD and momentum-based methods across a hierarchy of optimization tasks. We study online optimization under *drifting optima* by generating a stream of losses whose population minimizer $\theta_t^\star$ varies over time. We begin with a strongly convex quadratic objective, providing a direct bridge to our theoretical bounds where the impact of drift and noise can be isolated. We then consider three more advanced tasks: (i) linear regression (squared loss) with a drifting ground-truth weight vector $\theta_t^\star$, (ii) logistic regression (binary cross-entropy) with drifting $\theta_t^\star$, and (iii) a teacher–student two-layer MLP regression problem (MSE) with drifting teacher parameters. Across all settings, we simulate non-stationarity via a random walk of the target parameters. We compare SGD, Heavy-Ball (HB), and Nesterov accelerated gradient (NAG). For our last three experiments, we use a common base momentum $\beta = 0.9$ for HB/NAG and task/regime-specific base step sizes γ (reported in Table 2). Crucially, to reflect non-stationary stability constraints for momentum, we cap momentum step sizes by the theoretical bound

$$\gamma \;\leq\; \frac{\mu(1-\beta)^2}{4L^2},$$

which can be orders of magnitude smaller than the SGD stability range (e.g., $\gamma_{\max}^{\mathrm{mom}} = 2.5 \times 10^{-5}$ for $\kappa = 10$ and 2.5×10^{-9} for $\kappa = 1000$ when $\beta = 0.9$). This cap isolates the empirically observed *sluggishness* of momentum under drift: momentum must use a tiny stable step size and therefore adapts slowly to distribution shifts.

4.1 Strongly Convex Quadratic Minimizer

At each time step $t \geq 0$, we consider the time-varying objective function

$$f_t(\theta) = \frac{\mu}{2}\|\theta - \theta_t^\star\|^2, \tag{4.1}$$

where $\theta \in \mathbb{R}^d$ is the parameter vector and $\theta_t^\star \in \mathbb{R}^d$ denotes the (moving) minimizer at time t. We set $\mu = 1$. The function is μ-strongly convex with gradient $\nabla f_t(\theta) = \mu(\theta - \theta_t^\star)$. The sequence of optima $\{\theta_t^\star\}_{t\geq 0}$ is modeled as a random walk in $\mathbb{R}^d$:

$$\theta_{t+1}^\star \;=\; \theta_t^\star + \delta_{\mathrm{rw}} \frac{\mathbf{u}_t}{\|\mathbf{u}_t\|_2}, \qquad \mathbf{u}_t \sim \mathcal{N}(0, I_d), \tag{4.2}$$

which is a normalized random walk with step size δ_{rw}. This yields a cumulative drift scale $\sum_{t=1}^{T-1} \|\theta_{t+1}^\star - \theta_t^\star\|_2 = (T-1)\delta_{\mathrm{rw}}$, matching the "drift budget" viewpoint common in tracking analyses. In Table 1 we set $\delta_{\mathrm{rw}} = 0.01$, but results were robust across varying levels of δ_{rw}. At each time step, the algorithm observes a noisy gradient

$$G_t(\boldsymbol{\theta}) = \nabla f_t(\boldsymbol{\theta}) + \xi_t = \mu(\boldsymbol{\theta} - \boldsymbol{\theta}_t^\star) + \xi_t, \tag{4.3}$$

where the noise $\xi_t \sim \mathcal{N}(0, \sigma^2 I_d)$ is an i.i.d. Gaussian vector with variance σ^2 in each coordinate. The performance metric is the squared tracking error: $e_t = \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2$. We report the averaged tracking error $\mathbb{E}[e_t] = \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{\theta}_t^{(i)} - (\boldsymbol{\theta}_t^\star)^{(i)}\|^2$ as a function of time t, where (i) indexes independent runs. The experiment is averaged over 20 runs. These results were consistent across varying dimension values. In addition to Gaussian shifts, we also repeated the experiments with shifts drawn from a Student's t distribution, and observed qualitatively identical results.

4.2 Linear and Logistic Regression

Similar to the previous setup, to model non-stationarity we generate a time-varying population minimizer sequence $\{\boldsymbol{\theta}_t^\star\}_{t=1}^T$ via controlled drift. For linear and logistic regression, we set $\boldsymbol{\theta}_t^\star \in \mathbb{R}^d$ and evolve $\boldsymbol{\theta}_t^\star$ according to random walk in $\mathbb{R}^d$:

$$\boldsymbol{\theta}_{t+1}^\star = \boldsymbol{\theta}_t^\star + \delta_{\text{rw}} \frac{\mathbf{u}_t}{\|\mathbf{u}_t\|_2}, \qquad \mathbf{u}_t \sim \mathcal{N}(0, I_d), \tag{4.4}$$

which is a normalized random walk with step size δ_{rw}. Both of these tasks are run in two spectral regimes: *well-conditioned* ($\kappa = 10$) and *ill-conditioned* ($\kappa = 1000$). Conditioning is controlled by drawing covariates as $x = z\,\Sigma^{1/2}$ with $z \sim \mathcal{N}(0, I)$ and $\text{cond}(\Sigma) = \kappa$, so that the smallest and largest eigenvalues satisfy $\mu = 1$ and $L = \kappa$.

4.3 Teacher–student MLP drift.

For the teacher–student regression task, the population minimizer corresponds to the (time-varying) teacher network parameters. We also run this in both the well-conditioned and ill-conditioned regimes. Let $f_{\boldsymbol{\theta}}(x)$ be a two-layer ReLU MLP and let $\boldsymbol{\theta}_t^\star$ denote the teacher parameters at time t. We evolve $\boldsymbol{\theta}_t^\star$ analogously via either a continuous random-walk drift. Since parameter-space distances are not identifiable under hidden-unit permutation and ReLU scaling symmetries, we evaluate tracking primarily in *prediction space* i.e., the mean squared discrepancy between student and drifting teacher predictions on a fixed validation set.

4.4 Tables and Figures

Table 1: **Mean tracking error after 5000 iterations on a drifting strongly convex quadratic** ($d = 100$)**.** Boldface indicates the best (lowest) method within each $(\beta, \sigma^2, \gamma)$ setting. Across regimes, increasing β markedly degrades the steady-state tracking of HB/NAG—especially at larger γ and σ^2—highlighting inertia-induced lag under drift, while SGD remains comparatively robust.

	$\gamma = 0.01$					$\gamma = 0.05$					$\gamma = 0.10$			
β	σ^2	SGD	HB	NAG	β	σ^2	SGD	HB	NAG	β	σ^2	SGD	HB	NAG
0.50	0.1	1.036	**0.342**	0.349	0.50	0.1	**0.288**	0.504	0.523	0.50	0.1	**0.528**	1.054	0.914
0.50	0.5	1.235	**0.728**	0.745	0.50	0.5	**1.322**	2.453	2.435	0.50	0.5	**2.626**	5.273	4.848
0.50	0.8	1.305	**0.961**	1.019	0.50	0.8	**2.020**	4.022	3.941	0.50	0.8	**4.112**	8.505	7.813
0.90	0.1	1.029	0.497	**0.453**	0.90	0.1	**0.306**	2.472	1.732	0.90	0.1	**0.525**	5.172	2.596
0.90	0.5	**1.230**	2.358	2.247	0.90	0.5	**1.278**	12.298	9.194	0.90	0.5	**2.540**	27.411	13.433
0.90	0.8	**1.466**	3.899	3.721	0.90	0.8	**2.076**	20.372	14.604	0.90	0.8	**4.286**	40.666	21.376
0.95	0.1	1.039	1.051	**0.808**	0.95	0.1	**0.286**	4.866	2.641	0.95	0.1	**0.534**	9.981	3.331
0.95	0.5	**1.263**	5.144	4.114	0.95	0.5	**1.367**	24.077	13.073	0.95	0.5	**2.663**	51.361	17.378
0.95	0.8	**1.351**	7.696	6.801	0.95	0.8	**2.206**	42.031	20.205	0.95	0.8	**4.036**	79.822	27.169
0.99	0.1	**1.036**	4.837	2.619	0.99	0.1	**0.313**	25.674	4.224	0.99	0.1	**0.545**	54.250	4.414
0.99	0.5	**1.231**	23.669	13.096	0.99	0.5	**1.340**	133.720	21.351	0.99	0.5	**2.715**	243.978	24.099
0.99	0.8	**1.403**	38.802	21.038	0.99	0.8	**2.160**	198.361	34.105	0.99	0.8	**4.110**	401.372	38.317

Table 2: **Benchmark summary across linear regression, logistic regression, and teacher–student MLP under streaming drift.** Mean performance after 5000 iterations for SGD, Heavy-Ball (HB), and Nesterov (NAG) under two conditioning regimes (well-conditioned: $\kappa = 10$; ill-conditioned: $\kappa = 1000$). Arrows indicate the preferred direction (lower is better for loss/tracking/val MSE; higher is better for accuracy). For the teacher–student MLP, tracking is computed in *prediction space*. We report the training settings $(\sigma^2, \beta, \gamma)$ used in each regime (with γ the base step size). Overall, SGD is the most robust across conditioning, while HB/NAG deteriorate sharply in ill-conditioned regimes—especially on tracking metrics—consistent with inertia amplifying lag under nonstationary drift.

Linear Regression

Well-conditioned ($\kappa = 10$)

Settings: $\sigma^2 = 0.5,\ \beta = 0.9,\ \gamma = 0.1$

Method	Loss ↓	Track ↓
SGD	**1.192**	**0.174**
HB	1.449	0.493
NAG	1.447	0.494

Ill-conditioned ($\kappa = 1000$)

Settings: $\sigma^2 = 0.5,\ \beta = 0.9,\ \gamma = 0.001$

Method	Loss ↓	Track ↓
SGD	**1.262**	**0.168**
HB	56.910	0.898
NAG	56.902	0.898

Logistic Regression

Well-conditioned ($\kappa = 10$)

Settings: $\sigma^2 = 0.5,\ \beta = 0.9,\ \gamma = 0.5$

Method	Loss ↓	Track ↓	Acc ↑
SGD	**0.509**	**0.154**	**0.741**
HB	0.609	0.763	0.693
NAG	0.609	0.763	0.693

Ill-conditioned ($\kappa = 1000$)

Settings: $\sigma^2 = 0.5,\ \beta = 0.9,\ \gamma = 0.1$

Method	Loss ↓	Track ↓	Acc ↑
SGD	**0.173**	**0.518**	**0.924**
HB	0.690	0.902	0.758
NAG	0.690	0.902	0.759

Teacher–Student MLP

Well-conditioned ($\kappa = 10$)

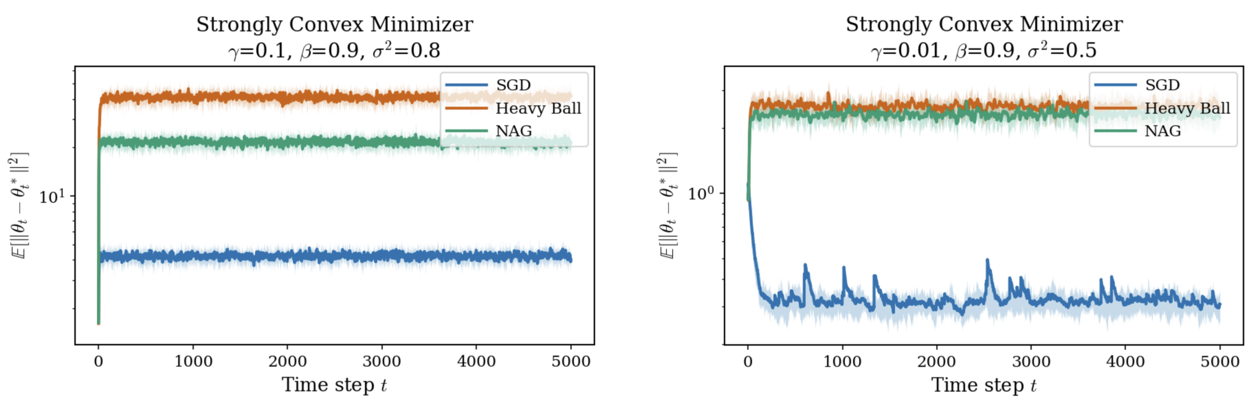
Settings: $\sigma^2 = 0.5,\ \beta = 0.9,\ \gamma = 0.06$

Method	Loss ↓	Pred. Track ↓	Val MSE ↓
SGD	**14.47**	**28.31**	**14.15**
HB	15.62	30.19	15.09
NAG	15.59	30.18	15.09

Ill-conditioned ($\kappa = 1000$)

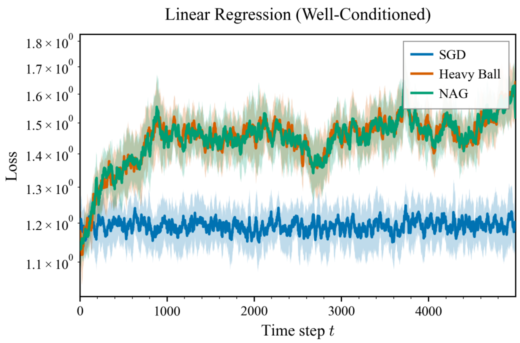
Settings: $\sigma^2 = 0.5,\ \beta = 0.9,\ \gamma = 0.02$

Method	Loss ↓	Pred. Track ↓	Val MSE ↓
SGD	**525.1**	**1064.5**	**532.3**
HB	640.5	1275.3	637.6
NAG	639.8	1275.3	637.6

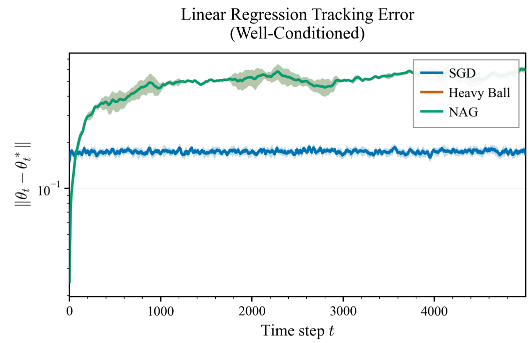


(a) $\gamma = 0.1, \beta = 0.9, \sigma^2 = 0.8$ (b) $\gamma = 0.05, \beta = 0.9, \sigma^2 = 0.5$

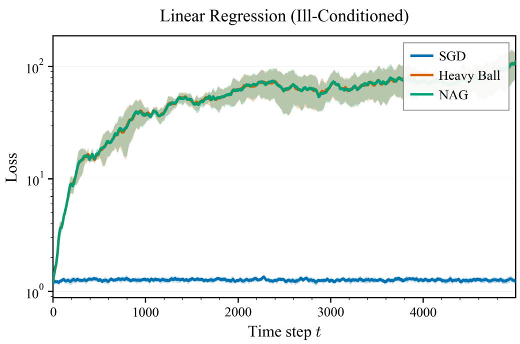
Figure 3: **Tracking a drifting minimizer under strong convexity.** Mean squared tracking error $\mathbb{E}\left[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|_2^2\right]$ versus time for SGD, Heavy-Ball (HB), and Nesterov (NAG) on a strongly convex quadratic with noisy gradients. Parameters $(\gamma, \beta, \sigma^2)$ are shown above each panel with shaded bands denoting ±1 std over random seeds. Across both regimes, momentum methods (HB/NAG) suppress short-term noise but exhibit a substantially larger steady-state tracking error—consistent with inertia-induced lag when the minimizer drifts—whereas SGD tracks the minimizer more closely but with higher variability.



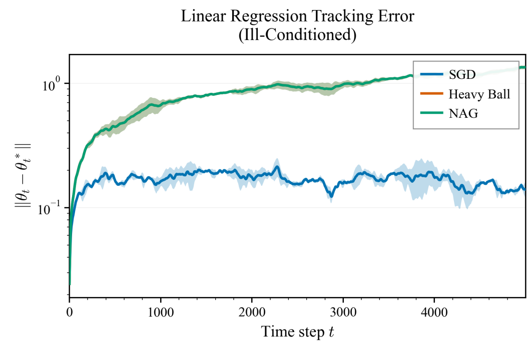
(a) LINEAR (WELL): loss.



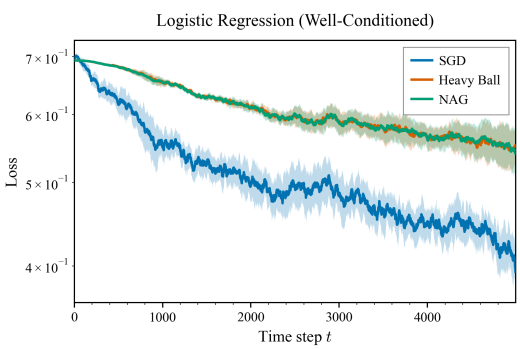
(b) LINEAR (WELL): tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|$.



(c) LINEAR (ILL): loss.

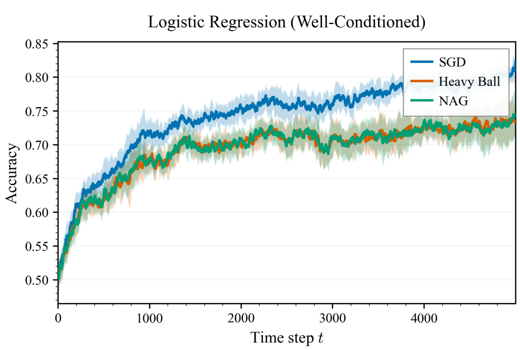


(d) LINEAR (ILL): tracking error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|$.

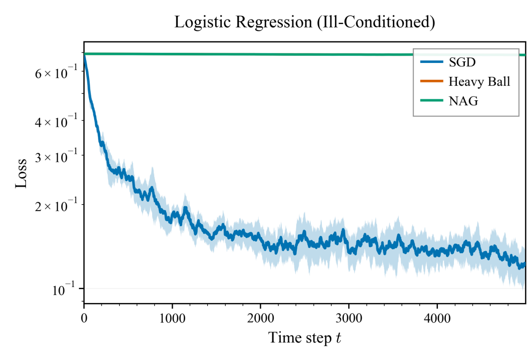


(e) LOGISTIC (WELL): loss.

(f) LOGISTIC (WELL): tracking error.

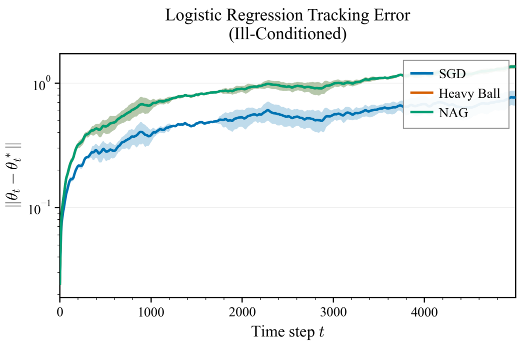


(g) LOGISTIC (WELL): accuracy.

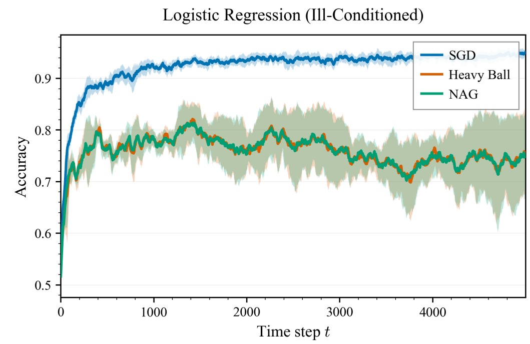


(h) LOGISTIC (ILL): loss.

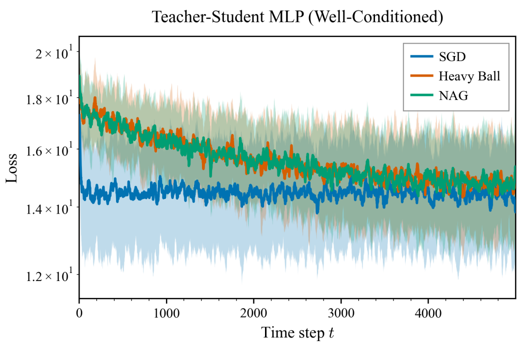
Figure 4: **Non-stationary benchmark suite.** Results for Linear and Logistic tasks. *(Continued on next page.)*



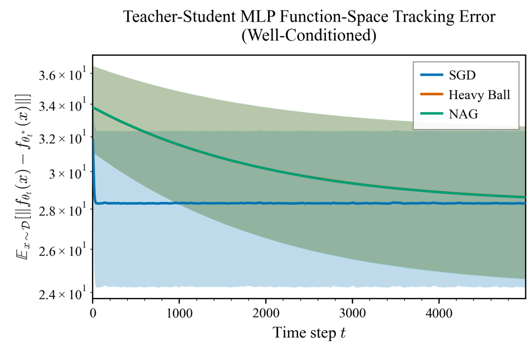
(i) Logistic (ill): tracking error.



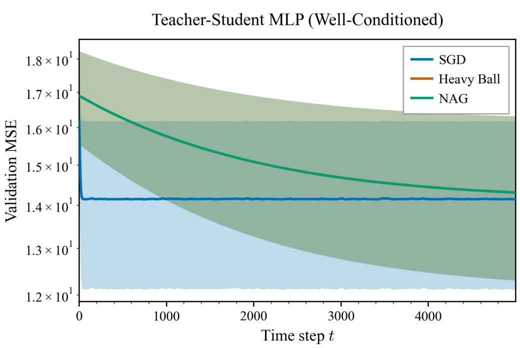
(j) Logistic (ill): accuracy.



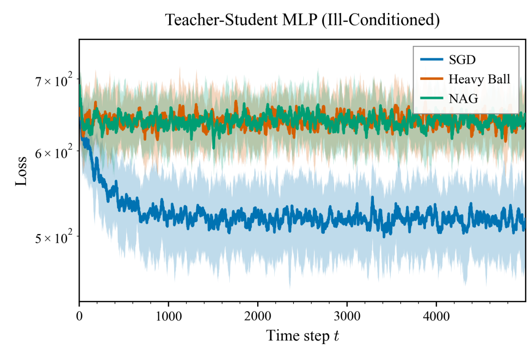
(k) MLP (well): training loss.



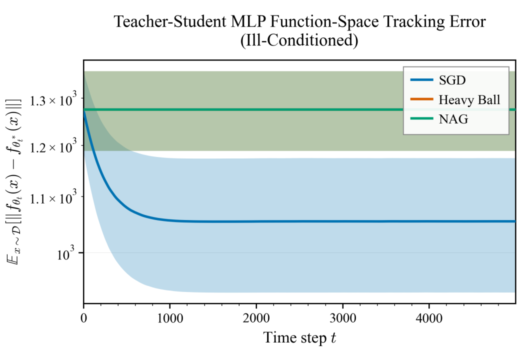
(l) MLP (well): prediction-space tracking.



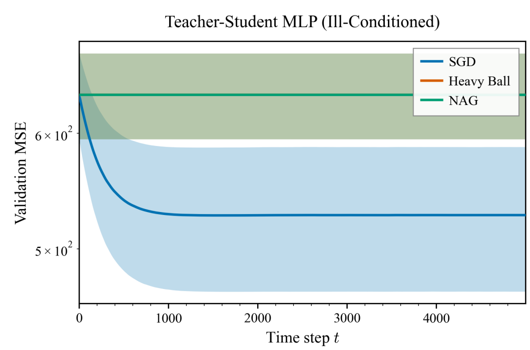
(m) MLP (well): validation MSE.



(n) MLP (ill): training loss.



(o) MLP (ill): prediction-space tracking.



(p) MLP (ill): validation MSE.

Figure 4: **Non-stationary benchmark suite (streaming) across linear, logistic, and teacher–student MLP tasks.** We report mean ± std curves over seeds under two conditioning regimes (well-conditioned: $\kappa = 10$; ill-conditioned: $\kappa = 1000$). For linear/logistic tasks, tracking error is measured in *parameter space* as $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|$. For the teacher–student MLP regression task, tracking is measured in *prediction space* (function space), i.e., $\mathbb{E}_{x\sim\mathcal{D}}[\|f_{\boldsymbol{\theta}_t}(x) - f_{\boldsymbol{\theta}_t^\star}(x)\|^2]$ estimated on a fixed validation set (reported as tracking error and validation MSE). Methods compared: SGD, Heavy-Ball (HB), and Nesterov (NAG), with step sizes respecting the regime-dependent stability caps used in our analysis.

5 Conclusion

Momentum is known to accelerate optimization in stationary problems, but its behavior under distribution shift is fundamentally different. In this paper we studied stochastic optimization when the population risk changes over time and the optimizer must track a drifting minimizer rather than converge to a fixed point. Our finite-time and high-probability analyses for SGD, Heavy-Ball, and Nesterov provide a unified picture: tracking performance decomposes into (i) a transient term that quantifies how quickly the method forgets initialization, (ii) an irreducible noise floor from stochastic gradients, and (iii) an irreducible tracking lag induced by drift. This decomposition makes the central trade-off explicit: temporal averaging can suppress noise, but it also makes gradients stale and slows adaptation to regime changes.

A key outcome is that momentum incurs a concrete and unavoidable *tracking penalty* as the momentum parameter increases. Large momentum increases sensitivity to initialization, inflates the effective variance level along the trajectory, and can force substantially more conservative step sizes for stability in ill-conditioned problems. Our high-probability bounds further show that drift can amplify stochastic fluctuations through trajectory-dependent mismatch, which directly motivates practical forgetting mechanisms (e.g., windowing or restarting) after regime shifts.

Complementing the upper bounds, we established minimax lower bounds for dynamic regret (and subsequently tracking error) under gradient-variation constraints. These results show that in drift-dominated regimes, momentum induces an information-theoretic *inertia window* during which the method is effectively behind the moving target, regardless of tuning. Across a nonstationary benchmark suite spanning convex and nonconvex tasks, the empirical results corroborate the predicted regime split: momentum can help in near-stationary, noise-dominated settings, but becomes brittle under genuine drift, especially with large momentum and ill-conditioning.

There are many future directions we would like to study. Our theory focuses on predictable drift under uniform strong convexity and smoothness, and uses first-order stochastic feedback. An important next step is to develop principled, drift-adaptive momentum schedules and restart rules with guarantees, and to extend the analysis to broader nonconvex settings. It would also be interesting to extend our high-probability analysis to include forms of heavier-tailed gradient noise such as sub-exponential or α-stable. It would also be valuable to study how adaptive preconditioning and second-order structure interact with drift, and to characterize when such methods can retain noise suppression benefits without incurring the inertia penalties identified here.

References

[ACDL14] Alekh Agarwal, Oliveier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15(32):1111–1133, 2014. 1

[AZF19] Sergul Aydore, Tianhao Zhu, and Dean Foster. *Dynamic local regret for non-convex online forecasting*. Curran Associates Inc., Red Hook, NY, USA, 2019. 1.1.2

[BB07] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. 1

[BGZ15] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, October 2015. 3.3, 3.3, C, C.1

[BWAA18] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 560–569. PMLR, 10–15 Jul 2018. 1.1.1

[CBS14] Volkan Cevher, Stephen Becker, and Mark Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Processing Magazine*, 31(5):32–43, 2014. 1

[CDH21] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under time drift: iterate averaging, step-decay schedules, and high probability guarantees. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11859–11869. Curran Associates, Inc., 2021. 1, 3.1, 3.1, 3.2, 3.2, 3.2, A.1

[CDH23] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under distributional drift. *Journal of Machine Learning Research*, 24(147):1–56, 2023. 1.1.3

[CLLZ24] Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for stochastic optimization via Kiefer-Wolfowitz methods. *Journal of the American Statistical Association*, 119(548):2972–2982, 2024. 1

[CM20] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 1, 1.1.1

[CPM+17] Jianmin Chen*, Xinghao Pan*, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous SGD, 2017. 1.1.1

[CT12] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012. C

[CWLK22] John Chen, Cameron Wolfe, Zhao Li, and Anastasios Kyrillidis. Demon: Improved neural network training with momentum decay. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3958–3962, 2022. 1.1.2

[CWW19] Xi Chen, Yining Wang, and Yu-Xiang Wang. Technical note—nonstationary stochastic optimization under lp,q-variation measures. *Operations Research*, 67(6):1752–1765, 2019. 3.3, 3.3, C, C.1, C.4

[CZP19] Xuanyu Cao, Junshan Zhang, and H. Vincent Poor. On the time-varying distributions of online stochastic optimization. In *2019 American Control Conference (ACC)*, pages 1494–1500, 2019. 1.1.3

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. 1.1.1

[Doz16] Timothy Dozat. Incorporating Nesterov momentum into Adam. In *ICLR 2016 Workshop Track*, 2016. Workshop track paper. 1.1.1

[DSC+22] Aram Davtyan, Sepehr Sameni, Llukman Cerkezi, Givi Meishvili, Adam Bielski, and Paolo Favaro. Koala: A Kalman optimization algorithm with loss adaptivity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6471–6479, Jun. 2022. 1.1.1

[DSLL24] Xiaoge Deng, Tao Sun, Dongsheng Li, and Xicheng Lu. Exploring the inefficiency of heavy ball as momentum parameter approaches 1. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 3899–3907. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track. 1.1.2

[FWZ+23] Jingwen Fu, Bohan Wang, Huishuai Zhang, Zhizheng Zhang, Wei Chen, and Nanning Zheng. When and why momentum accelerates SGD: An empirical study, 2023. arXiv:2306.09000. 1.1.2

[GBLP25] Antoine Godichon-Baggioni, Wei Lu, and Bruno Portier. Online estimation of the inverse of the hessian for stochastic optimization with application to universal stochastic newton algorithms, 2025. arXiv: 2401.10923. 1.1.1

[GCPB16] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 1

[GKS18] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850. PMLR, 10–15 Jul 2018. 1.1.1

[GKX19] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 09–15 Jun 2019. 3.1

[GLZX19] Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. *Understanding the role of momentum in stochastic gradient methods*. Curran Associates Inc., Red Hook, NY, USA, 2019. 1.1.2

[GS80] R. Graham and N. Sloane. Lower bounds for constant weight codes. *IEEE Transactions on Information Theory*, 26(1):37–43, 1980. C.4

[GTG+24] Alexandre Galashov, Michalis Titsias, András György, Clare Lyle, Razvan Pascanu, Yee Whye Teh, and Maneesh Sahani. Non-stationary learning of neural networks with automatic soft parameter reset. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1.1.3

[Haz16] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 08 2016. 1

[HLPR19] Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1579–1613. PMLR, 25–28 Jun 2019. 3.2

[HZJ+25] Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. In *The Thirteenth International Conference on Learning Representations*, 2025. 1.1.1

[IKS22] Wendyam Eric Lionel Ilboudo, Taisuke Kobayashi, and Kenji Sugimoto. Robust stochastic gradient descent with Student-t distribution based first-order momentum. *IEEE Transactions on Neural Networks and Learning Systems*, 33(3):1324–1337, 2022. 1.1.1

[JJMS25] Aukosh Jagannath, Taj Jones-McCormick, and Varnan Sarangian. High-dimensional limit theorems for SGD: Momentum and adaptive step-sizes, 2025. arXiv:2511.03952. 1.1.2

[JKB+19] Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *International Conference on Learning Representations*, 2019. 3.1

[JKR25] Arnulf Jentzen, Julian Kranz, and Adrian Riekert. PADAM: Parallel averaged Adam reduces the error for stochastic optimization in scientific machine learning, 2025. arXiv:2505.22085. 1.1.1

[JL22] Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning, 2022. 1.1.2

[KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1.1.1

[KB17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. arXiv:1412.6980. 1.1.1

[KS21] Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021. 1.1.2, 1.1.3

[KY97] Harold J. Kushner and G. George Yin. *Applications: Proofs of Convergence*, pages 251–272. Springer New York, New York, NY, 1997. 1

[Lan11] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:365 – 397, 2011. 3.1, 3.2

[LGY20] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. 1, 1.1.1

[LH19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1.1.1

[LJH+20] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020. 1.1.1

[LK24] Fangshuo Liao and Anastasios Kyrillidis. Provable accelerated convergence of nesterov's momentum for deep relu neural networks. In Claire Vernade and Daniel Hsu, editors, *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 732–784. PMLR, 25–28 Feb 2024. 1.1.1, 1.1.2

[LLZ+25] Xianliang Li, Jun Luo, Zhiwei Zheng, Hanxiao Wang, Li Luo, Lingkun Wen, Linlong Wu, and Sheng Xu. On the performance analysis of momentum method: A frequency domain perspective. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Representation Learning*, volume 2025, pages 47713–47734, 2025. 1.1.1

[LSY+25] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training, 2025. arXiv: 2502.16982. 1.1.1

[LSZG19] James Lucas, Shengyang Sun, Richard Zemel, and Roger Grosse. Aggregated momentum: Stability through passive damping. In *International Conference on Learning Representations*, 2019. 1.1.2

[LXL19] Liangchen Luo, Yuanhao Xiong, and Yan Liu. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations*, 2019. 1.1.1

[LXZ19] Hunter Lang, Lin Xiao, and Pengchuan Zhang. Using statistics to automate stochastic optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[MB11] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. 1

[Nes83] Y. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl Akad Nauk SSSR*, 269:543, 1983. 1, 1.1.1, 3.1

[Nes14] Yurii Nesterov. Introductory lectures on convex optimization - a basic course. In *Applied Optimization*, 2014. 3.1

[NJLS09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. 1

[Oll19] Yann Ollivier. The extended Kalman filter is a natural gradient descent in trajectory space, 2019. arXiv:1901.00696. 1.1.1

[Pap20] Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020. 3.1

[Pol63] B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. 1, 1.1.1

[Pow19] Warren B. Powell. A unified framework for stochastic optimization. *European Journal of Operational Research*, 275(3):795–821, 2019. 1

[Qia99] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. 1.1.2

[RKAC+24] Ali Ramezani-Kebrya, Kimon Antonakopoulos, Volkan Cevher, Ashish Khisti, and Ben Liang. On the generalization of stochastic gradient descent with momentum. *J. Mach. Learn. Res.*, 25(1), January 2024. 1, 1.1.1

[RKK18] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018. 1.1.2

[RM51] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. 1

[Say03] A.H. Sayed. *Fundamentals of Adaptive Filtering*. IEEE Press. Wiley, 2003. 1

[SBL17] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the Hessian in deep learning: Singularity and beyond, 2017. arXiv:1611.07476. 3.1

[SI24] Naoki Sato and Hideaki Iiduka. Role of momentum in smoothing objective function and generalizability of deep neural networks, 2024. 1, 1.1.1

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 1

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. 1, 1.1.1

[SZZ26] Yinan Shen, Yichen Zhang, and Wen-Xin Zhou. SGD with dependent data: Optimal estimation, regret, and inference, 2026. arXiv:2601.01371. 1.1.3, D.1

[TC25] Kristi Topollai and Anna Choromanska. Adaptive memory momentum via a model-based framework for deep learning optimization, 2025. arXiv:2510.04988. 1.1.2

[Vid25] Mathukumalli Vidyasagar. Convergence of momentum-based optimization algorithms with time-varying parameters, 2025. arXiv:2506.11904. 1.1.3

[VMZ+25] Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing shampoo using Adam for language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. 1.1.1

[Vuc18] James Vuckovic. Kalman gradient descent: Adaptive variance reduction in stochastic optimization, 2018. arXiv:1810.12273. 1.1.1

[WMW+24] Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In *The Twelfth International Conference on Learning Representations*, 2024. 1.1.2

[WS16] Yining Wang and Aarti Singh. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar.

2016. C.4, C.9

[WS21] Blake Woodworth and Nathan Srebro. An even more optimal stochastic optimization algorithm: Minibatching and interpolation learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 1

[Yan23] Xingyi Yang. Towards stochastic gradient variance reduction by solving a filtering problem, 2023. 1.1.1

[YGS+21] Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. ADAHESSIAN: An adaptive second order optimizer for machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10665–10673, May 2021. 1.1.1

[YPT17] Haimin Yang, Zhisong Pan, and Qing Tao. Robust and adaptive online time series prediction with long short-term memory. *Computational Intelligence and Neuroscience*, 2017:9478952, 2017. Article ID 9478952, 9 pages. 1.1.3

[YYC+25] Zhipeng Yao, Rui Yu, Guisong Chang, Ying Li, Yu Zhang, and Dazhou Li. Signal processing meets SGD: From momentum to filter, 2025. arXiv:2311.02818. 1.1.1

[YYS16] Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016. 1.1.2, 1.1.3, 2.3, 2.3, A.2

[ZDT+21] Juntang Zhuang, Yifan Ding, Tommy Tang, Nicha Dvornek, Sekhar C Tatikonda, and James Duncan. Momentum centering and asynchronous update for adaptive gradient methods. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28249–28260. Curran Associates, Inc., 2021. 1.1.1

[Zei12] Matthew D. Zeiler. Adadelta: An adaptive learning rate method, 2012. arXiv:1212.5701. 1.1.1

[Zha04] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 116, New York, NY, USA, 2004. Association for Computing Machinery. 1

[ZJDC20] Kaiwen Zhou, Yanghua Jin, Qinghua Ding, and James Cheng. Amortized Nesterov's momentum: A robust momentum and its application to deep learning. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 211–220. PMLR, 03–06 Aug 2020. 1.1.1

[ZM19] Jian Zhang and Ioannis Mitliagkas. Yellowfin and the art of momentum tuning. In A. Talwalkar, V. Smith, and M. Zaharia, editors, *Proceedings of Machine Learning and Systems*, volume 1, pages 289–308, 2019. 1.1.2

[ZRS+18] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1.1.1

[ZWL20] Yulai Zhang, Yuchao Wang, and Guiming Luo. A new optimization algorithm for non-stationary time series prediction based on recurrent neural networks. *Future Generation Computer Systems*, 102:738–745, 2020. 1.1.2

[ZWU21] Liu Ziyin, Zhikang T. Wang, and Masahito Ueda. Laprop: Separating momentum and adaptivity in Adam, 2021. arXiv:2002.04839. 1.1.1

[ZZL+19] Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations*, 2019. 1.1.1

[ZZZZ24] Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1–52, 2024. 1.1.3

Appendix

Table of Contents

A Proofs of tracking error bounds in expectation

For this section, we will assume the following hold for the minimizer drift and gradient noise second moments (Assumption 3.1):

Assumption A.1 (Conditional second-moment bounds)**.** *There exist constants $\Delta, \sigma > 0$ such that for all $t \geq 0$,*

i) ***(Minimizer drift)*** *The minimizer drift Δ_t satisfies $\mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_t\right] \leq \Delta^2$ a.s.*

ii) ***(Gradient noise along iterates)*** *The gradient noise $\boldsymbol{\xi}_{t+1}$ satisfies $\mathbb{E}\left[\|\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t)\|^2 \mid \mathcal{F}_t\right] \leq \sigma^2$ a.s.*

A.1 Proof for SGD tracking error bound

First we will prove a recursive relation for the tracking error that we will subsequently use for our expectation and high-probability bounds:

Lemma A.1 (Recursive relation for the SGD tracking error)**.** *For $\forall t \geq 0$ and $\gamma_t \leq \min\left\{\mu/L^2, 1/L\right\}$, the following recursive relation for the tracking error holds*

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \leq \left(1 - \frac{\gamma_t\mu}{2}\right)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 + \frac{2}{\gamma_t\mu}\|\Delta_t\|^2 + \gamma_t^2\|\xi_{t+1}(\boldsymbol{\theta}_t)\|^2 + M_{t+1}$$

where the martingale increment is $M_{t+1} = -2\gamma_t\langle d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t), \xi_{t+1}(\boldsymbol{\theta}_t)\rangle$ with $d_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}^\star$.

Proof of Lemma A.1. Before we proceed, let us recall some notation that we will use in our analysis. Let $(\mathcal{F}_t)_{t\geq 1}$ be the natural filtration $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$. Define $m_{t+1}(\boldsymbol{\theta}) = \mathbb{E}[\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t]$ to be the conditional expected gradient and the gradient noise as $\xi_{t+1}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, X_{t+1}) - m_{t+1}(\boldsymbol{\theta})$ so that $\mathbb{E}[\xi_{t+1}(\boldsymbol{\theta}) \mid \mathcal{F}_t] = 0$. Define $d_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}^\star$. Then notice that we can write the minimizer drift as follows:

$$\begin{aligned} d_t &= \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}^\star \\ &= \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star + \boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star \\ &= e_t + \Delta_t. \end{aligned}$$

Then from the SGD update rule (2.3), we have that

$$\begin{aligned} \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star &= \boldsymbol{\theta}_t - \gamma_t(m_{t+1}(\boldsymbol{\theta}_t) + \xi_{t+1}(\boldsymbol{\theta}_t)) - \boldsymbol{\theta}_{t+1}^\star \\ &= d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t) - \gamma_t\xi_{t+1}(\boldsymbol{\theta}_t). \end{aligned}$$

Taking the ℓ_2 norm of each side, we find that

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 = \|d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t)\|^2 - 2\gamma_t\langle d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t), \xi_{t+1}(\boldsymbol{\theta}_t)\rangle + \gamma^2\|\xi_{t+1}(\boldsymbol{\theta}_t)\|^2.$$

Define $\Phi_{t+1}(\boldsymbol{\theta}_t) = d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t)$. By definition we have

$$\|\Phi_{t+1}(\boldsymbol{\theta}_t)\|^2 = \|d_t\|^2 - 2\gamma\langle d_t, m_{t+1}(\boldsymbol{\theta}_t)\rangle + \gamma^2\|m_{t+1}(\boldsymbol{\theta}_t)\|^2.$$

By μ-strong monotonicity of m_t (Assumption 2.2), we have

$$\begin{aligned} \langle d_t, m_{t+1}(\boldsymbol{\theta}_t)\rangle &= \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}^\star, m_{t+1}(\boldsymbol{\theta}_t) - m_{t+1}(\boldsymbol{\theta}_{t+1}^\star)\rangle \\ &\geq \mu\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}^\star\|^2 \\ &= \mu\|d_t\|^2. \end{aligned} \tag{A.1}$$

where the first equality holds since $m_{t+1}(\boldsymbol{\theta}_{t+1}^\star) = 0$. We also have by Lipschitz continuity of m_t (Assumption 2.3) that $\|m_{t+1}(\boldsymbol{\theta}_t)\|^2 \leq L^2\|d_t\|^2$. Combining this with (A.1), we find that

$$\|\Phi_{t+1}(\boldsymbol{\theta}_t)\|^2 \leq (1 - 2\gamma\mu + \gamma^2L^2)\|d_t\|^2. \tag{A.2}$$

If we take $\gamma_t \le \mu/L^2$ for $\forall t \ge 0$, then we have that $1 - 2\gamma\mu + \gamma^2 L^2 \le 1 - \gamma\mu$. Let $\rho := 1 - \gamma\mu$. Now let the martingale increment be $M_{t+1} = -2\gamma_t \langle d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t), \xi_{t+1}(\boldsymbol{\theta}_t)\rangle$. Then combining our restriction on γ_t with (A.2), we find that

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le \rho \|d_t\|^2 + \gamma_t^2 \|\xi_{t+1}(\boldsymbol{\theta}_t)\|^2 + M_{t+1}. \tag{A.3}$$

Note that in the literature, one will usually apply Young's inequality to M_{t+1}, resulting in the right side being in terms of d_t and ξ_{t+1}. However, since d_t can be written in terms of Δ_t to relate back to the drift, this will result in us needing to apply an iterating MGF approach (see [CDH21] for details). We instead will apply Young's inequality of d_t. This will help us in avoiding any tail-assumptions on the drift, thus yielding a more generalizable high-probability bound. By Young's inequality, for any $\alpha > 0$, we have

$$\|d_t\|^2 \le (1 + \alpha)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2 + \left(1 + \frac{1}{\alpha}\right) \|\Delta_t\|^2. \tag{A.4}$$

Using (A.4) with (A.3), we find that

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le \rho(1 + \alpha)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2 + \rho\left(1 + \frac{1}{\alpha}\right) \|\Delta_t\|^2 + \gamma_t^2 \|\xi_{t+1}(\boldsymbol{\theta}_t)\|^2 + M_{t+1}. \tag{A.5}$$

Since we want $\rho(1 + \alpha) \le 1$ for convergence guarantees as $t \to \infty$, we take $\alpha = (1 - \rho)/2\rho$. Using this, we find

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le \left(\frac{1 + \rho}{2}\right) \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2 + \frac{\rho(1 + \rho)}{1 - \rho} \|\Delta_t\|^2 + \gamma_t^2 \|\xi_{t+1}(\boldsymbol{\theta}_t)\|^2 + M_{t+1}. \tag{A.6}$$

Substituting in $\rho := 1 - \gamma\mu$, we find

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le \left(1 - \frac{\gamma_t\mu}{2}\right) \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2 + \frac{(1 - \gamma_t\mu)(2 - \gamma_t\mu)}{\gamma\mu} \|\Delta_t\|^2 + \gamma_t^2 \|\xi_{t+1}(\boldsymbol{\theta}_t)\|^2 + M_{t+1}.$$

Note that μ-strong monotonicity and Lipschitz continuity of m_t jointly imply that $\mu \le L$. As a consequence of taking $\gamma_t \le \mu/L^2$, we have $\gamma_t\mu \le 1$ which implies $-3 + \gamma_t\mu \le 0$. This implies that

$$\begin{aligned}
\frac{(1 - \gamma_t\mu)(2 - \gamma_t\mu)}{\gamma\mu} &= \frac{2}{\gamma_t\mu} - 3 + \gamma_t\mu \\
&\le \frac{2}{\gamma_t\mu}.
\end{aligned}$$

Thus we can conclude that

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le \left(1 - \frac{\gamma_t\mu}{2}\right) \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_t\|^2 + \frac{2}{\gamma_t\mu} \|\Delta_t\|^2 + \gamma_t^2 \|\xi_{t+1}(\boldsymbol{\theta}_t)\|^2 + M_{t+1}.$$

□

Applying Lemma A.1 recursively and setting a constant step-size $\gamma_t = \gamma$, we obtain the following result:

Proposition A.1 (Final-iterate tracking error bound for SGD)**.** *Let $\gamma \le \min\{\mu/L^2, 1/L\}$. Then for $\forall t \ge 0$, the following bound holds:*

$$\begin{aligned}
\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \le{}& \left(1 - \frac{\gamma\mu}{2}\right)^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2 + \frac{2}{\gamma\mu} \sum_{\ell=0}^{t-1} \left(1 - \frac{\gamma\mu}{2}\right)^{t-\ell-1} \|\Delta_\ell\|^2 \\
&+ \gamma^2 \sum_{\ell=0}^{t-1} \left(1 - \frac{\gamma\mu}{2}\right)^{t-\ell-1} \|\xi_\ell(\boldsymbol{\theta}_\ell)\|^2 + \sum_{\ell=0}^{t-1} \left(1 - \frac{\gamma\mu}{2}\right)^{t-\ell-1} M_\ell
\end{aligned}$$

where $M_{t+1} := -2\gamma_t \langle d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t), \xi_{t+1}(\boldsymbol{\theta}_t)\rangle$ with $d_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star_{t+1}$.

We can finally obtain a tracking error bound in expectation for SGD using Assumption 3.1.

Corollary A.1 (Tracking error bound in expectation for SGD)**.** *Under Assumption 3.1, for* $\forall t \geq 0$ *and* $\gamma \leq \min\{\mu/L^2, 1/L\}$*, the following tracking error bound holds in expectation for SGD:*

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \leq \left(1 - \frac{\gamma\mu}{2}\right)^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2 + \frac{4\Delta^2}{\gamma^2\mu^2} + \frac{\sigma^2\gamma}{\mu}.$$

Proof of Corollary A.1. Recall that $M_{t+1} = -2\gamma_t\langle d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t), \xi_{t+1}(\boldsymbol{\theta}_t)\rangle$. Since $m_{t+1}(\boldsymbol{\theta}_t)$ is $\mathcal{F}_t$ measurable and $\mathbb{E}[\xi_{t+1}(\boldsymbol{\theta}_t) \mid \mathcal{F}_t] = 0$, we have that

$$\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = -2\gamma_t\langle d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t), \mathbb{E}[\xi_{t+1}(\boldsymbol{\theta}_t) \mid \mathcal{F}_t]\rangle = 0.$$

This implies that M_{t+1} is a martingale difference sequence (MDS). Thus by iterated expectation and using the fact that $\sum_{\ell\geq 0}\delta = 1/(1-\delta)$, we can conclude. □

Using Corollary A.1, we can obtain the following result which gives us a algorithmic sample complexity guarantee for SGD:

Corollary A.2 (Time to reach the asymptotic tracking error in expectation for SGD)**.** *Assume* $\gamma_t \in (0, 1/(2L)]$ *for all* $t \geq 0$*. For any constant* $\gamma \in (0, 1/2L]$*, define*

$$\mathcal{E}(\gamma) := \frac{\sigma^2\gamma}{\mu} + \frac{4\Delta^2}{\mu^2\gamma^2}, \qquad \gamma^\star \in \arg\min_{\gamma\in(0,1/2L]} \mathcal{E}(\gamma), \qquad \mathcal{E} := \mathcal{E}(\gamma^\star).$$

Then we have the following:

(i) ***(Constant learning rate).*** *If* $\gamma_t \equiv \gamma^\star$*, then for all* $t \geq 0$*,*

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \lesssim \mathcal{E} \text{ after time } t \lesssim \frac{1}{\mu\gamma^\star}\log\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2}{\mathcal{E}}\right).$$

(ii) ***(Step-decay schedule in the low drift-to-noise regime).*** *Suppose* $\gamma^\star < 1/2L$ *(equivalently, the minimizer of* $\mathcal{E}(\gamma)$ *is not at the smoothness cap), so that*

$$\gamma^\star = \left(\frac{8\Delta^2}{\mu\sigma^2}\right)^{1/3}, \qquad \mathcal{E} = 3\left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3}.$$

Define epochs $k = 0, 1, \ldots, K-1$ *with*

$$\gamma_0 := \frac{1}{2L}, \qquad \gamma_k := \frac{\gamma_{k-1} + \gamma^\star}{2} \quad (k \geq 1), \qquad K := 1 + \left\lceil \log_2\left(\frac{\gamma_0}{\gamma^\star}\right)\right\rceil,$$

and epoch lengths

$$T_0 := \left\lceil \frac{2}{\mu\gamma_0}\log\left(\frac{2\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2}{\mathcal{E}(\gamma_0)}\right)\right\rceil, \qquad T_k := \left\lceil \frac{2\log 4}{\mu\gamma_k}\right\rceil \quad (k \geq 1).$$

Run SGD with constant stepsize γ_k *for* T_k *iterations in epoch* k*, starting from* θ_0*. Let* $T := \sum_{k=0}^{K-1} T_k$ *be the total horizon. Then the final iterate satisfies*

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \lesssim \mathcal{E} \text{ after time } t \lesssim \frac{L}{\mu}\log\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2}{\mathcal{E}}\right) + \frac{\sigma^2}{\mu^2\mathcal{E}}.$$

Proof of Corollary A.2. From Corollary A.1, we have the following

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \leq \left(1 - \frac{\gamma\mu}{2}\right)^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2 + \frac{4\Delta^2}{\gamma^2\mu^2} + \frac{\sigma^2\gamma}{\mu}. \tag{A.7}$$

Proof of (i). By (A.7) and $\gamma = \gamma^\star$, using $\mathcal{E}(\gamma) := \sigma^2\gamma/\mu + 4\Delta^2/\mu^2\gamma^2$, we have:

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \leq \left(1 - \frac{\mu\gamma^\star}{2}\right)^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2 + \mathcal{E}(\gamma^\star).$$

If $t \geq (2/\mu\gamma^\star)\log(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2/\mathcal{E})$, then $\left(1 - \mu\gamma^\star/2\right)^t \leq \exp(-t\mu\gamma^\star/2) \leq \mathcal{E}/\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2$, and hence $\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \leq 2\mathcal{E}$.

Proof of (ii). Let $t_0 := 0$ and $t_{k+1} := t_k + T_k$. Define epoch iterates $X_k := \boldsymbol{\theta}_{t_k}$ and corresponding minimizers $X_k^\star := \boldsymbol{\theta}_{t_k}^\star$. Applying (A.7) inside epoch k with stepsize γ_k and length T_k yields

$$\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \leq \left(1 - \frac{\mu\gamma_k}{2}\right)^{T_k}\mathbb{E}\|X_k - X_k^\star\|^2 + \mathcal{E}(\gamma_k).$$

For $k \geq 1$, by the choice of T_k and $\log(1-x) \leq -x$,

$$\left(1 - \frac{\mu\gamma_k}{2}\right)^{T_k} \leq \exp\left(-\frac{\mu\gamma_k}{2}T_k\right) \leq \exp(-\log 4) = \frac{1}{4},$$

so for $k \geq 1$,

$$\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \leq \frac{1}{4}\,\mathbb{E}\|X_k - X_k^\star\|^2 + \mathcal{E}(\gamma_k). \tag{A.8}$$

Base epoch. By the definition of T_0, we have $\left(1 - \mu\gamma_0/2\right)^{T_0}\|X_0 - X_0^\star\|^2 \leq \mathcal{E}(\gamma_0)$, hence

$$\mathbb{E}\|X_1 - X_1^\star\|^2 \leq \left(1 - \frac{\mu\gamma_0}{2}\right)^{T_0}\|X_0 - X_0^\star\|^2 + \mathcal{E}(\gamma_0) \leq 2E(\gamma_0).$$

Induction. We claim that for all $k \geq 1$,

$$\mathbb{E}\|X_k - X_k^\star\|^2 \leq 2\mathcal{E}(\gamma_{k-1}).$$

Assume this holds for some $k \geq 1$. Using (A.8),

$$\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \leq \frac{1}{4}\cdot 2\mathcal{E}(\gamma_{k-1}) + \mathcal{E}(\gamma_k) = \frac{1}{2}\mathcal{E}(\gamma_{k-1}) + \mathcal{E}(\gamma_k).$$

Since $\gamma_k = (\gamma_{k-1} + \gamma^\star)/2$, we have $\gamma_{k-1} \leq 2\gamma_k$. For $E(\gamma) = A/\gamma^2 + B\gamma$ with $A = 4\Delta^2/\mu^2$ and $B = \sigma^2/\mu$, one checks that for all $\gamma \geq \gamma^\star$,

$$\mathcal{E}(2\gamma) \leq 2\mathcal{E}(\gamma).$$

Because $\gamma_k \geq \gamma^\star$, this gives $\mathcal{E}(\gamma_{k-1}) \leq \mathcal{E}(2\gamma_k) \leq 2\mathcal{E}(\gamma_k)$, and thus

$$\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \leq \frac{1}{2}\cdot 2\mathcal{E}(\gamma_k) + \mathcal{E}(\gamma_k) = 2\mathcal{E}(\gamma_k),$$

closing the induction. Hence $\mathbb{E}\|X_K - X_K^\star\|^2 \leq 2\mathcal{E}(\gamma_{K-1})$.

Next, by the definition of K,

$$\gamma_{K-1} - \gamma^\star = \frac{\gamma_0 - \gamma^\star}{2^{K-1}} \leq \frac{\gamma_0}{2^{K-1}} \leq \gamma^\star, \quad \text{so} \quad \gamma_{K-1} \leq 2\gamma^\star.$$

Therefore $\mathcal{E}(\gamma_{K-1}) \leq \mathcal{E}(2\gamma^\star) \leq 2\mathcal{E}(\gamma^\star) = 2\mathcal{E}$, yielding

$$\mathbb{E}\|e_T\|^2 = \mathbb{E}\|X_K - X_K^\star\|^2 \leq 2\mathcal{E}(\gamma_{K-1}) \leq 4\mathcal{E}.$$

Time bound. Since $\mathcal{E}(\gamma_0) \geq \mathcal{E}$, we have

$$T_0 \leq \frac{2}{\mu\gamma_0}\log\left(\frac{2\|X_0 - X_0^\star\|^2}{\mathcal{E}}\right) + 1 = \frac{4L}{\mu}\log\left(\frac{2\|X_0 - X_0^\star\|^2}{\mathcal{E}}\right) + 1.$$

Moreover, for $k \geq 1$, $\gamma_k \geq \gamma_0/2^{k+1}$, hence

$$\sum_{k=1}^{K-1}\frac{1}{\gamma_k} \leq \sum_{k=1}^{K-1}\frac{2^{k+1}}{\gamma_0} \leq \frac{2^{K+1}}{\gamma_0} \leq \frac{4}{\gamma^\star}.$$

Thus

$$\sum_{k=1}^{K-1} T_k \le \sum_{k=1}^{K-1} \Big(\frac{2\log 4}{\mu\gamma_k} + 1\Big) \le \frac{8\log 4}{\mu} \sum_{k=1}^{K-1} \frac{1}{\gamma_k} + O(K) \le \frac{32\log 4}{\mu\gamma^\star} + O(K).$$

In the low drift-to-noise regime, $\gamma^\star = (8\Delta^2/(\mu\sigma^2))^{1/3}$ and $E = 3(\Delta\sigma^2/\mu^2)^{2/3}$, so $1/\mu\gamma^\star = 3/2\cdot\sigma^2/\mu^2 E$. Substituting yields

$$\sum_{k=1}^{K-1} T_k \le 32\log 4 \cdot \frac{3}{2} \cdot \frac{\sigma^2}{\mu^2\mathcal{E}} + O(1) = 48\log 4 \cdot \frac{\sigma^2}{\mu^2\mathcal{E}} + O(1).$$

Combining with the bound on T_0 proves the stated horizon bound up to universal constants. □

A.2 Proof for SGDM tracking error bound

For SGD with momentum, we will take a different approach to prove bounds on the tracking error in expectation and with high-probability. We will instead work with a 2D state-space view of SGD with momentum by defining extended state vectors. We will first introduce the transformation matrices:

$$\mathbf{V} = \begin{bmatrix} I_d & -\beta I_d \\ I_d & -I_d \end{bmatrix}, \ \mathbf{V}^{-1} = \frac{1}{1-\beta} \begin{bmatrix} I_d & -\beta I_d \\ I_d & -I_d \end{bmatrix}. \tag{A.9}$$

Recall the following SGD with momentum updates (2.4):

$$\begin{aligned} \boldsymbol{\psi}_t &= \boldsymbol{\theta}_t + \beta_1(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\psi}_t - \gamma_t m_{t+1}(\boldsymbol{\psi}_t) - \gamma_t \xi_{t+1}(\boldsymbol{\psi}_t) + \beta_2(\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}). \end{aligned} \tag{A.10}$$

Define $\widetilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_t$. Define the transformed error vectors, each of size $2d \times 1$:

$$\begin{bmatrix} \widehat{\boldsymbol{\theta}}_t \\ \check{\boldsymbol{\theta}}_t \end{bmatrix} \triangleq \mathbf{V}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{\theta}}_t \\ \widetilde{\boldsymbol{\theta}}_{t-1} \end{bmatrix} = \frac{1}{1-\beta} \begin{bmatrix} \widetilde{\boldsymbol{\theta}}_t - \beta\widetilde{\boldsymbol{\theta}}_{t-1} \\ \widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{\theta}}_{t-1}. \end{bmatrix} \tag{A.11}$$

We can obtain an extended recursion 2D state-space matrix that captures the dynamics of SGD with momentum. This will prove very useful in our subsequent analysis to obtain expectation and high probability bounds without needing to recurse on the velocity vector $\mathbf{v}_t$ or using a Lyapunov stability function argument. Before we proceed with obtaining this extended recursion, we state an assumption that typically holds under sufficient conditions for the Dominated convergence theorem to hold:

Assumption A.2 (Interchanging conditional expectation and gradient)**.** *For each $t \ge 0$, define the conditional objective*

$$F_{t+1}(\boldsymbol{\theta}) := \mathbb{E}[g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t].$$

Assume F_{t+1} is differentiable and

$$\nabla F_{t+1}(\boldsymbol{\theta}) = \mathbb{E}[\nabla g(\boldsymbol{\theta}, X_{t+1}) \mid \mathcal{F}_t] = m_{t+1}(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \mathbb{R}^d,$$

almost surely. (Sufficient conditions are standard dominated-convergence hypotheses.)

Note that since we assumed m_{t+1} is μ-strongly monotone and Lipschitz continuous (Assumption 2.2 and Assumption 2.3), it follows that F_{t+1} is μ-strongly convex and L-smooth. Equivalently, this implies

$$\mu I_d \preceq \nabla^2 F_{t+1}(\boldsymbol{\theta}) \preceq L I_d \quad \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{A.12}$$

We now prove the extended 2D recursion. This proof largely follows [YYS16] with a similar form except for a matrix decoupling that arises from the minimizer drift.

Lemma A.2 (Extended 2D recursion for SGD with momentum)**.** *Under Assumption 2.2, Assumption 2.3, and $\beta_1+\beta_2 = \beta$ and $\beta_1\beta_2 = 0$ for fixed $\beta \in [0,1)$, the SGD with momentum update equations can be transformed into the following extended recursion:*

$$\begin{bmatrix}\widehat{\theta}_t \\ \check{\theta}_t\end{bmatrix} = \begin{bmatrix} I_d - \frac{\gamma_t}{1-\beta}\mathbf{H}_{t-1} & \frac{\gamma_t\beta'}{1-\beta}\mathbf{H}_{t-1} \\ -\frac{\gamma_t}{1-\beta}\mathbf{H}_{t-1} & \beta I_d + \frac{\gamma_t\beta'}{1-\beta}\mathbf{H}_{t-1}\end{bmatrix}\begin{bmatrix}\widehat{\theta}_{t-1} \\ \check{\theta}_{t-1}\end{bmatrix} + \frac{1}{1-\beta}\begin{bmatrix}-(I_d-\gamma_t\mathbf{H}_{t-1})\Delta_{t-1} - \mathbf{K}_{t-1}\Delta_{t-2} \\ -(I_d-\gamma_t\mathbf{H}_{t-1})\Delta_{t-1} - \mathbf{K}_{t-1}\Delta_{t-2}\end{bmatrix} + \frac{\gamma_t}{1-\beta}\begin{bmatrix}\xi_{t+1}(\boldsymbol{\psi}_{t-1}) \\ \xi_{t+1}(\boldsymbol{\psi}_{t-1})\end{bmatrix} \tag{A.13}$$

where

$$\beta' \triangleq \beta\beta_1 + \beta_2 \tag{A.14}$$

$$\mathbf{H}_{t-1} \triangleq \int_0^1 \nabla^2 F_{t+1}(\boldsymbol{\theta}^\star_{t+1} + s(\boldsymbol{\psi}_t - \boldsymbol{\theta}^\star_{t+1}))ds \tag{A.15}$$

$$\mathbf{K}_{t-1} = -\beta I_d + \gamma_t\beta_1\mathbf{H}_{t-1}. \tag{A.16}$$

Proof of Lemma A.2. From the SGD with momentum update equation (2.4), we have

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\psi}_t - \gamma_t m_{t+1}(\boldsymbol{\psi}_t) - \gamma_t\xi_{t+1}(\boldsymbol{\psi}_t) + \beta_2(\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}). \tag{A.17}$$

Let $\widetilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}^\star_t - \boldsymbol{\theta}_t$ and $\widetilde{\boldsymbol{\psi}}_t = \boldsymbol{\theta}^\star_{t+1} - \boldsymbol{\psi}_t$. Subtracting both sides of (A.17) from $\boldsymbol{\theta}^\star_t$, we get:

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\psi}}_t + \gamma_t m_{t+1}(\boldsymbol{\psi}_t) + \gamma_t\xi_{t+1}(\boldsymbol{\psi}_t) - \beta_2(\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}). \tag{A.18}$$

We can now appeal to the mean-value theorem:

$$\nabla F_{t+1}(\boldsymbol{\psi}_t) - \nabla F_{t+1}(\boldsymbol{\theta}^\star_{t+1}) = \left(\int_0^1 \nabla^2 F_{t+1}(\boldsymbol{\theta}^\star_{t+1} + s(\boldsymbol{\psi}_t - \boldsymbol{\theta}^\star_{t+1}))ds\right)(\boldsymbol{\psi}_t - \boldsymbol{\theta}^\star_{t+1}) \triangleq -\mathbf{H}_t\widetilde{\boldsymbol{\psi}}_t. \tag{A.19}$$

Since we have that $m_{t+1}(\boldsymbol{\theta}^\star_{t+1}) = 0$, this implies that $m_{t+1}(\boldsymbol{\psi}_t) = -\mathbf{H}_t\widetilde{\boldsymbol{\psi}}_t$. Another thing to note is that since $\widetilde{\boldsymbol{\psi}}_t = \boldsymbol{\theta}^\star_{t+1} - \boldsymbol{\psi}_t$, we have

$$\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1} = (\widetilde{\boldsymbol{\psi}}_t - \boldsymbol{\theta}^\star_{t+1}) - (\widetilde{\boldsymbol{\psi}}_{t-1} - \boldsymbol{\theta}^\star_{t+1}) = -\widetilde{\boldsymbol{\psi}}_t + \widetilde{\boldsymbol{\psi}}_{t-1}. \tag{A.20}$$

Combining (A.18), (A.20) and (A.19), we find

$$\begin{aligned}\widetilde{\boldsymbol{\theta}}_{t+1} &= \widetilde{\boldsymbol{\psi}}_t + \gamma_t m_{t+1}(\boldsymbol{\psi}_t) + \gamma_t\xi_{t+1}(\boldsymbol{\psi}_t) - \beta_2(\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}) \\ &= \widetilde{\boldsymbol{\psi}}_t - \gamma_t\mathbf{H}_t\widetilde{\boldsymbol{\psi}}_t + \beta_2(\widetilde{\boldsymbol{\psi}}_t - \widetilde{\boldsymbol{\psi}}_{t-1}) + \gamma_t\xi_{t+1}(\boldsymbol{\psi}_t) \\ &= ((1+\beta_2)I_d - \gamma_t\mathbf{H}_t)\widetilde{\boldsymbol{\psi}}_t - \beta_2\widetilde{\boldsymbol{\psi}}_{t-1} + \gamma_t\xi_{t+1}(\boldsymbol{\psi}_t).\end{aligned} \tag{A.21}$$

Now for any $k \le t$, let us define

$$\widetilde{\boldsymbol{\theta}}^{(t+1)}_k := \boldsymbol{\theta}^\star_{t+1} - \boldsymbol{\theta}_k. \tag{A.22}$$

Then we have that $\widetilde{\boldsymbol{\theta}}^{(t+1)}_{t+1} = \widetilde{\boldsymbol{\theta}}_{t+1}$. Using the lookahead from SGD with momentum, we have

$$\widetilde{\boldsymbol{\psi}}_t = \boldsymbol{\theta}^\star_{t+1} - \boldsymbol{\psi}_t = (\boldsymbol{\theta}^\star_{t+1} - \boldsymbol{\theta}_t) - \beta_1(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) = \widetilde{\boldsymbol{\theta}}^{(t+1)}_t - \beta_1(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}). \tag{A.23}$$

However also note that

$$\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1} = -(\boldsymbol{\theta}^\star_{t+1} - \boldsymbol{\theta}_t) + (\boldsymbol{\theta}^\star_{t+1} - \boldsymbol{\theta}_{t-1}) = -\widetilde{\boldsymbol{\theta}}^{(t+1)}_t + \widetilde{\boldsymbol{\theta}}^{(t+1)}_{t-1}. \tag{A.24}$$

Combining (A.23) and (A.24), we find

$$\widetilde{\boldsymbol{\psi}}_t = \widetilde{\boldsymbol{\theta}}^{(t+1)}_t - \beta_1\left(-\widetilde{\boldsymbol{\theta}}^{(t+1)}_t + \widetilde{\boldsymbol{\theta}}^{(t+1)}_{t-1}\right) = (1+\beta_1)\widetilde{\boldsymbol{\theta}}^{(t+1)}_t - \beta_1\widetilde{\boldsymbol{\theta}}^{(t+1)}_{t-1}. \tag{A.25}$$

Plugging in (A.25) into (A.21), we obtain the following:

$$\widetilde{\boldsymbol{\theta}}^{(t+1)}_{t+1} = \mathbf{J}_t\widetilde{\boldsymbol{\theta}}^{(t+1)}_t + \mathbf{K}_t\widetilde{\boldsymbol{\theta}}^{(t+1)}_{t-1} + L\widetilde{\boldsymbol{\theta}}^{(t+1)}_{t-2} + \gamma_t\xi_{t+1}(\boldsymbol{\psi}_t), \tag{A.26}$$

where

$$\begin{aligned}
\mathbf{J}_t &\triangleq (1+\beta_1)(1+\beta_2)I_d - \gamma_t(1+\beta_1)\mathbf{H}_t \\
\mathbf{K}_t &\triangleq -(\beta_1+\beta_2+\beta_1\beta_2)I_d + \gamma_t\beta_1\mathbf{H}_t \\
L &\triangleq \beta_1\beta_2.
\end{aligned} \tag{A.27}$$

Since we have that $\beta_1 + \beta_2 = \beta$ and $\beta_1\beta_2 = 0$, we can simplify this as

$$\begin{aligned}
\mathbf{J}_t &= (1+\beta)I_d - \gamma_t(1+\beta_1)\mathbf{H}_t \\
\mathbf{K}_t &= -\beta I_d + \gamma_t\beta_1\mathbf{H}_t \\
L &= 0.
\end{aligned} \tag{A.28}$$

This gives us a recursive bound based on an augmented shifted-error state. That is, it is written in a coordinate system relative to the minimizer $\boldsymbol{\theta}^{\star}_{t+1}$. We now convert this into a recursion for the tracking error $\widetilde{\boldsymbol{\theta}}_t$. Recall that the minimizer drift is defined as $\Delta_t = \boldsymbol{\theta}^{\star}_t - \boldsymbol{\theta}^{\star}_{t+1}$. Then we can write

$$\widetilde{\boldsymbol{\theta}}^{(t+1)}_t = \boldsymbol{\theta}^{\star}_{t+1} - \boldsymbol{\theta}_t = (\boldsymbol{\theta}^{\star}_t - \boldsymbol{\theta}_t) + (\boldsymbol{\theta}^{\star}_{t+1} - \boldsymbol{\theta}^{\star}_t) = \widetilde{\boldsymbol{\theta}}_t - \Delta_t. \tag{A.29}$$

Similarly we have

$$\widetilde{\boldsymbol{\theta}}^{(t+1)}_{t-1} = (\boldsymbol{\theta}^{\star}_{t-1} - \boldsymbol{\theta}_{t-1}) + (\boldsymbol{\theta}^{\star}_{t+1} - \boldsymbol{\theta}^{\star}_t + \boldsymbol{\theta}^{\star}_t - \boldsymbol{\theta}^{\star}_{t-1}) = \widetilde{\boldsymbol{\theta}}_{t-1} - (\Delta_{t-1} + \Delta_t). \tag{A.30}$$

Plugging (A.29) and (A.30) into (A.26), we get

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \mathbf{J}_t\widetilde{\boldsymbol{\theta}}_t + \mathbf{K}_t\widetilde{\boldsymbol{\theta}}_{t-1} + \mathbf{b}_t + \gamma_t\xi_{t+1}(\boldsymbol{\psi}_t), \tag{A.31}$$

where

$$\mathbf{b}_t \triangleq -(\boldsymbol{J}_t + \boldsymbol{K}_t)\Delta_t - \boldsymbol{K}_t\Delta_{t-1} = -(I_d - \gamma_t H_t)\Delta_t - \mathbf{K}_t\Delta_{t-1}. \tag{A.32}$$

It follows that we can write the extended recursion relation as:

$$\underbrace{\begin{bmatrix}\widetilde{\boldsymbol{\theta}}_{t+1}\\ \widetilde{\boldsymbol{\theta}}_t\end{bmatrix}}_{\triangleq\, \mathbf{z}_{t+1}} = \underbrace{\begin{bmatrix}\mathbf{J}_t & \mathbf{K}_t\\ I_d & 0\end{bmatrix}}_{\triangleq\, \mathbf{B}_t} \underbrace{\begin{bmatrix}\widetilde{\boldsymbol{\theta}}_t\\ \widetilde{\boldsymbol{\theta}}_{t-1}\end{bmatrix}}_{\triangleq\, \mathbf{z}_t} + \underbrace{\begin{bmatrix}\mathbf{b}_t\\ 0\end{bmatrix}}_{\triangleq\, \mathbf{u}_t} + \gamma_t \underbrace{\begin{bmatrix}\xi_{t+1}(\boldsymbol{\psi}_t)\\ 0\end{bmatrix}}_{\triangleq\, \boldsymbol{\eta}_{t+1}}. \tag{A.33}$$

Note that we can write $\mathbf{B}_t = \mathbf{P} - \mathbf{M}_t$ where

$$\mathbf{P} = \begin{bmatrix}(1+\beta)I_d & -\beta I_d\\ I_d & 0\end{bmatrix}, \quad \mathbf{M}_t = \begin{bmatrix}\gamma_t(1+\beta_1)\mathbf{H}_t & -\gamma_t\beta_1\mathbf{H}_t\\ 0 & 0\end{bmatrix}. \tag{A.34}$$

Now $\mathbf{P}$ has the following eigenvalue decomposition $\mathbf{P} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ where

$$\mathbf{V} = \begin{bmatrix}I_d & -\beta I_d\\ I_d & -I_d\end{bmatrix}, \quad \mathbf{V}^{-1} = \frac{1}{1-\beta}\begin{bmatrix}I_d & -\beta I_d\\ I_d & -I_d\end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix}I_d & 0\\ 0 & \beta I_d\end{bmatrix}. \tag{A.35}$$

Define the transformed "mode-splitting" coordinates as follows:

$$\mathbf{y}_t := \mathbf{V}^{-1}\mathbf{z}_t = \begin{bmatrix}\widehat{\boldsymbol{\theta}}_t\\ \check{\boldsymbol{\theta}}_t\end{bmatrix} = \frac{1}{1-\beta}\begin{bmatrix}\widetilde{\boldsymbol{\theta}}_t - \beta\widetilde{\boldsymbol{\theta}}_{t-1}\\ \widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{\theta}}_{t-1}\end{bmatrix}. \tag{A.36}$$

Using this transform with (A.31) gives us:

$$\begin{aligned}
\mathbf{y}_{t+1} &= \mathbf{V}^{-1}\mathbf{B}_t\mathbf{V}\mathbf{y}_t + \mathbf{V}^{-1}\mathbf{u}_t + \gamma_t\mathbf{V}^{-1}\boldsymbol{\eta}_{t+1} \\
&= \underbrace{(\mathbf{D} - \mathbf{V}^{-1}\mathbf{M}_t\mathbf{V})}_{\triangleq\, \widetilde{\mathbf{B}}_t}\mathbf{y}_t + \mathbf{V}^{-1}\mathbf{u}_t + \gamma_t\mathbf{V}^{-1}\boldsymbol{\eta}_{t+1}.
\end{aligned} \tag{A.37}$$

Finally we note that

$$\mathbf{D} - \mathbf{V}^{-1}\mathbf{M}_t\mathbf{V} = \begin{bmatrix} I_d - \frac{\gamma_t}{1-\beta}\mathbf{H}_t & \frac{\gamma_t\beta'}{1-\beta}\mathbf{H}_t \\ -\frac{\gamma_t}{1-\beta}\mathbf{H}_t & \beta I_d + \frac{\gamma_t\beta'}{1-\beta}\mathbf{H}_t \end{bmatrix}, \quad \mathbf{V}^{-1}\boldsymbol{\eta}_{t+1} = \frac{1}{1-\beta}\begin{bmatrix} \xi_{t+1}(\boldsymbol{\psi}_{t+1}) \\ \xi_{t+1}(\boldsymbol{\psi}_{t+1}) \end{bmatrix}. \tag{A.38}$$

This concludes the proof. □

In order to establish a convergence guarantee of the momentum stochastic gradient method, we will need to analyze the dynamics of the norm of the extended state vectors. What we aim to do is show a path-wise inequality (analogous to a Lyapunov stability function) of the form $\mathbf{s}_t \leq \mathbf{\Gamma}_t\mathbf{s}_t$ and then get a uniform $\mathbf{\Gamma}$ by upper bounding $\mathbf{\Gamma}_t, \forall t \geq 0$.

Corollary A.3 (Uniform stability for SGD with momentum)**.** *Let Assumption 2.2, Assumption 2.3 hold, and $\beta_1+\beta_2 = \beta$ and $\beta_1\beta_2 = 0$ for fixed $\beta \in [0, 1)$. Consider the momentum stochastic gradient method (2.5) and the extended 2D recursion (A.13). Then when the step-sizes γ_t satisfies*

$$\gamma_t \leq \frac{\mu(1-\beta)^2}{4L^2} \tag{A.39}$$

it holds that the mean-square values of the transformed error vectors evolve according to the recursive inequality below:

$$\begin{bmatrix} \|\widehat{\boldsymbol{\theta}}_{t+1}\|^2 \\ \|\check{\boldsymbol{\theta}}_{t+1}\|^2 \end{bmatrix} \leq \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} \|\widehat{\boldsymbol{\theta}}_{t}\|^2 \\ \|\check{\boldsymbol{\theta}}_{t}\|^2 \end{bmatrix}, \tag{A.40}$$

where

$$a = 1 - \frac{\gamma_t\mu}{1-\beta}, \quad b = \frac{\gamma_t\beta'^2L^2}{\mu(1-\beta)}, \quad c = \frac{2\gamma_t^2L^2}{(1-\beta)^3}, \quad d = \beta + \frac{2\gamma_t^2\beta'^2L^2}{(1-\beta)^3} \tag{A.41}$$

and the coefficient matrix is stable i.e.

$$\rho\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) < 1. \tag{A.42}$$

Proof of Corollary A.3. Define the energy vector as follows:

$$\mathbf{s}_t = \begin{bmatrix} \|\widehat{\boldsymbol{\theta}}_{t}\|^2 \\ \|\check{\boldsymbol{\theta}}_{t}\|^2 \end{bmatrix} \in \mathbb{R}^2_{\geq 0}. \tag{A.43}$$

We will show the pathwise inequality $\mathbf{s}_{t+1} \leq \mathbf{\Gamma}_t\mathbf{s}_t$ with explicit scalar entries a_t, b_t, c_t, d_t, and then get a uniform $\mathbf{\Gamma}$ by upper bounded $\mathbf{\Gamma}_t$ over $\forall t \geq 0$. From Lemma A.2, we find the first row can be written as

$$\widehat{\boldsymbol{\theta}}_{t+1} = \mathbf{A}_t\widehat{\boldsymbol{\theta}}_t + \mathbf{B}_t\check{\boldsymbol{\theta}}_t, \tag{A.44}$$

with $\mathbf{A}_t = I_d - \eta_t\mathbf{H}_t$ and $\mathbf{B}_t = \eta_t\beta'\mathbf{H}_t$ where $\eta_t = \gamma_t/(1-\beta)$ and $\beta' = \beta\beta_1 + \beta_2$. By Young's inequality, for $\tau > 0$

$$\|\widehat{\boldsymbol{\theta}}_{t+1}\|^2 \leq \frac{1}{1-\tau}\|\mathbf{A}_t\widehat{\boldsymbol{\theta}}_t\|^2 + \frac{1}{\tau}\|\mathbf{B}_t\check{\boldsymbol{\theta}}_t\|^2 \leq \frac{\|\mathbf{A}_t\|^2}{1-\tau}\|\widehat{\boldsymbol{\theta}}_t\|^2 + \frac{\|\mathbf{B}_t\|^2}{\tau}\|\check{\boldsymbol{\theta}}_t\|^2. \tag{A.45}$$

We will first impose that $\eta_t \leq 1/L$. Also recall that since $\mu I_d \preceq \nabla^2 F_{t+1}(\boldsymbol{\theta}) \preceq LI_d$, we immediately have that $\mu I_d \preceq \mathbf{H}_t \preceq LI_d$. Since we have defined $\mathbf{A}_t = I_d - \eta_t\mathbf{H}_t$, these facts imply the eigenvalues of $\mathbf{A}_t$ lie in $[1-\eta_t L, 1-\eta_t\mu] \subset [0, 1-\eta_t\mu]$. Hence we find that $\|\mathbf{A}_t\|_{\text{op}} = 1 - \eta_t\mu$. Likewise we find that $\|\mathbf{B}_t\|_{\text{op}} \leq \eta_t\beta' L$. Thus we get

$$\|\widehat{\boldsymbol{\theta}}_{t+1}\|^2 \leq \frac{(1-\eta_t\mu)^2}{1-\tau}\|\widehat{\boldsymbol{\theta}}_t\|^2 + \frac{\eta_t^2\beta'^2L^2}{\tau}\|\check{\boldsymbol{\theta}}_t\|^2. \tag{A.46}$$

Take $\tau = \eta_t\mu \in (0, 1)$(indeed this is true since $\eta_t \leq 1/L$ and $\mu \leq L$). Then we find that

$$\|\widehat{\boldsymbol{\theta}}_{t+1}\|^2 \leq \underbrace{(1-\eta_t\mu)}_{\triangleq a}\|\widehat{\boldsymbol{\theta}}_t\|^2 + \underbrace{\left(\eta_t\frac{\beta'^2L^2}{\mu}\right)}_{\triangleq b}\|\check{\boldsymbol{\theta}}_t\|^2. \tag{A.47}$$

We will now repeat this for the second row. We can write the second row as

$$\check{\boldsymbol{\theta}}_{t+1} = -\eta_t \mathbf{H}_t \widehat{\boldsymbol{\theta}}_t + (\beta I_d + \eta_t \beta' \mathbf{H}_t)\check{\boldsymbol{\theta}}_t = \beta \check{\boldsymbol{\theta}}_t + r_t, \tag{A.48}$$

where $r_t := \eta_t \mathbf{H}_t(\beta' \check{\boldsymbol{\theta}}_t - \widehat{\boldsymbol{\theta}}_t)$. By the convexity of the map $x \mapsto \|x\|^2$, we have that

$$\begin{aligned} \|\check{\boldsymbol{\theta}}_{t+1}\|^2 &= \|\beta \check{\boldsymbol{\theta}}_t + r_t\|^2 \\ &= \|\beta \check{\boldsymbol{\theta}}_t + (1-\beta) r_t/(1-\beta)^2\|^2 \\ &\leq \beta \|\check{\boldsymbol{\theta}}_t\|^2 + \frac{1}{1-\beta}\|r_t\|^2. \end{aligned} \tag{A.49}$$

Now we bound $\|r_t\|$:

$$\|r_t\| \leq \eta_t \|\mathbf{H}_t\| \|\beta' \cdot \check{\boldsymbol{\theta}}_t - \widehat{\boldsymbol{\theta}}_t\| \leq \eta_t L \left(\beta' \|\check{\boldsymbol{\theta}}_t\| + \|\widehat{\boldsymbol{\theta}}_t\|\right). \tag{A.50}$$

Using the inequality $(a+b)^2 \leq a^2 + b^2$, we conclude that

$$\|r_t\|^2 \leq 2\eta_t^2 L^2 \left(\beta'^2 \|\check{\boldsymbol{\theta}}_t\|^2 + \|\widehat{\boldsymbol{\theta}}_t\|^2\right). \tag{A.51}$$

Therefore we find that

$$\|\check{\boldsymbol{\theta}}_{t+1}\|^2 \leq \underbrace{\left(\frac{2\eta_t^2 L^2}{1-\beta}\right)}_{\triangleq c} \|\widehat{\boldsymbol{\theta}}_t\|^2 + \underbrace{\left(\beta + \frac{2\eta_t^2 \beta'^2 L^2}{1-\beta}\right)}_{\triangleq d} \|\check{\boldsymbol{\theta}}_t\|^2. \tag{A.52}$$

Collecting (A.47) and (A.52), we find

$$\begin{bmatrix} \|\widehat{\boldsymbol{\theta}}_{t+1}\|^2 \\ \|\check{\boldsymbol{\theta}}_{t+1}\|^2 \end{bmatrix} \leq \boldsymbol{\Gamma}_t \begin{bmatrix} \|\widehat{\boldsymbol{\theta}}_t\|^2 \\ \|\check{\boldsymbol{\theta}}_t\|^2 \end{bmatrix}, \tag{A.53}$$

with

$$\boldsymbol{\Gamma}_t := \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 - \eta_t \mu & \eta_t \frac{\beta'^2 L^2}{\mu} \\ \frac{2\eta_t^2 L^2}{1-\beta} & \beta + \frac{2\eta_t^2 \beta'^2 L^2}{1-\beta} \end{bmatrix}. \tag{A.54}$$

We will now examine the stability of the 2x2 coefficient matrix $\boldsymbol{\Gamma}_t$. First assume a constant upper bound

$$\eta := \sup_t \eta_t = \sup_t \frac{\gamma_t}{1-\beta}.$$

Define the uniform dominating matrix

$$\boldsymbol{\Gamma} := \begin{bmatrix} 1 - \eta \mu & \eta \frac{\beta'^2 L^2}{\mu} \\ \frac{2\eta^2 L^2}{1-\beta} & \beta + \frac{2\eta^2 \beta'^2 L^2}{1-\beta} \end{bmatrix}. \tag{A.55}$$

Then $\boldsymbol{\Gamma}_t \leq \boldsymbol{\Gamma}$ entrywise $\forall t \geq 0$. Hence we have that $\mathbf{s}_{t+1} \leq \boldsymbol{\Gamma} s_t \leq \boldsymbol{\Gamma}^{\alpha} s_{t-\alpha}$ for any $\alpha \leq t$. Since the spectral radius of a matrix is upper bounded by its 1-norm, we have that $\rho(\boldsymbol{\Gamma}) \leq \max\{a+c, b+d\}$, it suffices to compute the column sums and impose conditions on η to ensure that $\rho(\boldsymbol{\Gamma}) < 1$. Computing this, we find

$$\rho(\boldsymbol{\Gamma}) \leq \max\left\{1 - \eta\mu + \frac{2\eta^2 L^2}{1-\beta}, \beta + \eta \frac{\beta'^2 L^2}{\mu} + \frac{2\eta^2 \beta'^2 L^2}{1-\beta}\right\}. \tag{A.56}$$

We can choose a step-size η small enough to satisfy:

$$\begin{cases} \frac{\eta\mu}{2} > \frac{2\eta^2 L^2}{1-\beta} \\ \frac{\eta \beta'^2 L^2}{\mu} > \frac{2\eta^2 \beta'^2 L^2}{1-\beta} \\ 1 - \beta > \frac{2\eta \beta'^2 L^2}{\mu}, \end{cases} \tag{A.57}$$

which is equivalent to

$$\eta < \min\left\{\frac{\mu(1-\beta)}{4L^2}, \frac{1-\beta}{2\mu}, \frac{\mu(1-\beta)}{\beta'^2 L^2}\right\} = \frac{\mu(1-\beta)}{4L^2}. \tag{A.58}$$

Converting this back to γ using the relation $\eta = \gamma/(1-\beta)$, we find

$$\gamma < \min\left\{\frac{\mu(1-\beta)^2}{4L^2}, \frac{(1-\beta)^2}{2\mu}, \frac{\mu(1-\beta)^2}{\beta'^2 L^2}\right\} = \frac{\mu(1-\beta)^2}{4L^2}. \tag{A.59}$$

It then holds that

$$\rho(\mathbf{\Gamma}) < \max\left\{1 - \frac{\gamma\mu}{2(1-\beta)}, \beta + \frac{2\gamma L^2}{\mu(1-\beta)}\right\} \leq 1. \tag{A.60}$$

In this case, $\mathbf{\Gamma}$ is a stable matrix. This concludes the proof. □

From Lemma A.2, we established that we have a recursive relation of the form $\mathbf{y}_{t+1} = \widetilde{\boldsymbol{B}}_t \mathbf{y}_t + \mathbf{V}^{-1}\mathbf{u}_t + \gamma_t \mathbf{V}^{-1}\eta_{t+1}$. Define the original state matrices as $\mathbf{\Phi}(t,s) := \boldsymbol{B}_{t-1}\boldsymbol{B}_{t-2}\ldots\boldsymbol{B}_s$ and the transformed state transition matrices as $\widetilde{\mathbf{\Phi}}(t,s) := \widetilde{\boldsymbol{B}}_{t-1}\widetilde{\boldsymbol{B}}_{t-2}\ldots\widetilde{\boldsymbol{B}}_s$ for $\forall t > s$ and $\widetilde{\mathbf{\Phi}}(s,s) := I_d$. With the above established, we can convert the fact that $\mathbf{\Gamma}$ is stable into a bound on $\widetilde{\mathbf{\Phi}}$ and subsequently $\mathbf{\Phi}$. We will establish this below:

Corollary A.4 (Conversion from $\widetilde{\mathbf{\Phi}}$ to $\mathbf{\Phi}$)**.** *Let Assumption 2.2, Assumption 2.3 hold, and $\beta_1 + \beta_2 = \beta$ and $\beta_1\beta_2 = 0$ for fixed $\beta \in [0,1)$. Consider the momentum stochastic gradient method (2.5) and the extended 2D recursion (A.13). Then when the step-sizes γ_t satisfies*

$$\gamma_t \leq \frac{\mu(1-\beta)^2}{4L^2}, \tag{A.61}$$

then we have the following bound:

$$\|\mathbf{\Phi}(t,\alpha)\| \leq \frac{4}{1-\beta}\rho^{t-\alpha} \tag{A.62}$$

where $\rho := \sqrt{\rho(\mathbf{\Gamma})} < 1$.

Proof of Corollary A.4. First note that $\mathbf{y}_t = \widetilde{\mathbf{\Phi}}(t,\alpha)\mathbf{y}_\alpha$. Recall from Lemma A.2 that we have

$$\|\mathbf{y}_t\|^2 = \|\widehat{\boldsymbol{\theta}}_t\|^2 + \|\check{\boldsymbol{\theta}}_t\|^2 = \mathbf{1}^\top \mathbf{s}_t. \tag{A.63}$$

where $\mathbf{s}_t = [\widehat{\boldsymbol{\theta}}_t; \check{\boldsymbol{\theta}}_t]$. In Corollary A.3, we showed that $s_t \leq \mathbf{\Gamma}^\alpha s_{t-\alpha}$ or equivalently $s_t \leq \mathbf{\Gamma}^{t-\alpha} s_\alpha$ for $\forall \alpha \leq t$. This implies that

$$\|\mathbf{y}_t\|^2 = \mathbf{1}^\top \mathbf{s}_t \leq \mathbf{1}^\top \mathbf{\Gamma}^{t-\alpha} s_\alpha \leq \|\mathbf{\Gamma}^{t-\alpha}\| \mathbf{1}^\top s_\alpha = \|\mathbf{\Gamma}^{t-\alpha}\| \|\mathbf{y}_\alpha\|^2. \tag{A.64}$$

Now since $\|\mathbf{\Gamma}\|_1 \leq \rho(\mathbf{\Gamma}) < 1$, we have that $\|\mathbf{\Gamma}^n\|_1 \leq \|\mathbf{\Gamma}\|_1^n \leq \rho^n(\mathbf{\Gamma})$. Hence we find

$$\|\mathbf{y}_t\| \leq \rho^{t-\alpha}\|\mathbf{y}_\alpha\|. \tag{A.65}$$

Therefore this implies that $\|\widetilde{\mathbf{\Phi}}(t,\alpha)\| \leq \rho^{t-\alpha}$. Since we have $\mathbf{z}_t = \mathbf{V}\mathbf{y}_t$, we have that

$$\mathbf{\Phi}(t,\alpha) = \mathbf{B}_{t-1}\cdots\mathbf{B}_\alpha = \mathbf{V}\widetilde{\mathbf{\Phi}}(t,\alpha)\boldsymbol{V}^{-1}.$$

Hence we have

$$\|\mathbf{\Phi}(t,\alpha)\| \leq \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\|\widetilde{\mathbf{\Phi}}(t,\alpha)\| \leq \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\rho^{t-\alpha}. \tag{A.66}$$

From the (A.9), one can easily show that $\|\mathbf{V}[x;y]\| \leq \sqrt{2}(\|x\| + \|y\|) \leq 2\|[x;y]\|$. Hence we have that $\|\mathbf{V}\| \leq 2$. This also implies $\|\mathbf{V}\|^{-1} \leq 2/(1-\beta)$. We can thus conclude. □

In light of these results, we can finally prove an expectation bound on the tracking error of SGD with momentum:

Corollary A.5 (Tracking error bound in expectation for SGD with momentum)**.** *Let Assumption 2.2, Assumption 2.3, and Assumption 3.1. Consider the momentum stochastic gradient method (2.5) and the extended 2D recursion (A.13). Then when the step-sizes γ_t satisfies*

$$\gamma_t \leq \frac{\mu(1-\beta)^2}{4L^2}, \tag{A.67}$$

the following tracking error bound holds in expectation for SGD with momentum with constant stepsize $\gamma_t = \gamma$:

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \leq \frac{48}{(1-\beta)^2}\rho^{2(t+1)}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2 + \frac{48(1+\beta+\gamma L)^2}{(1-\beta)^2}\cdot\frac{\Delta^2}{(1-\rho)^2} + \frac{48}{(1-\beta)^2}\cdot\frac{\sigma^2\gamma^2}{1-\rho^2}. \tag{A.68}$$

In particular taking $\rho = 1 - \gamma\mu/2(1-\beta)$, we obtain:

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star_{t+1}\|^2 \leq \frac{48}{(1-\beta)^2}\exp\left(-\frac{\gamma\mu}{1-\beta}(t+1)\right)\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star_0\|^2 + \frac{192(2+\beta)^2}{\gamma^2\mu^2}\Delta^2 + \frac{96\sigma^2\gamma}{\mu(1-\beta)}. \tag{A.69}$$

Proof of Corollary A.5. For the analysis, we will denote $C_\beta := 4/(1-\beta)$. From Lemma A.2, we showed that the SGD with momentum dynamics follow a 2D recursive system as follows:

$$\underbrace{\begin{bmatrix}\widetilde{\boldsymbol{\theta}}_{t+1}\\ \widetilde{\boldsymbol{\theta}}_t\end{bmatrix}}_{\triangleq\, \mathbf{z}_{t+1}} = \underbrace{\begin{bmatrix}\mathbf{J}_t & \mathbf{K}_t\\ I_d & 0\end{bmatrix}}_{\triangleq\, \mathbf{B}_t}\underbrace{\begin{bmatrix}\widetilde{\boldsymbol{\theta}}_t\\ \widetilde{\boldsymbol{\theta}}_{t-1}\end{bmatrix}}_{\triangleq\, \mathbf{z}_t} + \underbrace{\begin{bmatrix}\mathbf{b}_t\\ 0\end{bmatrix}}_{\triangleq\, \mathbf{u}_t} + \gamma_t\underbrace{\begin{bmatrix}\xi_{t+1}(\boldsymbol{\psi}_t)\\ 0\end{bmatrix}}_{\triangleq\, \boldsymbol{\eta}_{t+1}} \tag{A.70}$$

where

$$\begin{aligned}\mathbf{J}_t &= (1+\beta)I_d - \gamma_t(1+\beta_1)\mathbf{H}_t\\ \mathbf{K}_t &= -\beta I_d + \gamma_t\beta_1\mathbf{H}_t\\ \mathbf{b}_t &= -(I_d - \gamma_t H_t)\Delta_t - \mathbf{K}_t\Delta_{t-1}.\end{aligned} \tag{A.71}$$

Iterating (A.70) gives us the identity:

$$\mathbf{z}_{t+1} = \boldsymbol{\Phi}(t+1,0)\mathbf{z}_0 + \sum_{k=0}^{t}\boldsymbol{\Phi}(t+1,k+1)\mathbf{u}_k + \sum_{k=0}^{t}\boldsymbol{\Phi}(t+1,k+1)\gamma_k\eta_{k+1}. \tag{A.72}$$

Define the following quantities:

$$A_{t+1} := \boldsymbol{\Phi}(t+1,0)\mathbf{z}_0,\quad D_{t+1} := \sum_{k=0}^{t}\boldsymbol{\Phi}(t+1,k+1)\mathbf{u}_k,\quad N_{t+1} := \sum_{k=0}^{t}\boldsymbol{\Phi}(t+1,k+1)\gamma_k\boldsymbol{\eta}_{k+1}. \tag{A.73}$$

Then we have that

$$\mathbf{z}_{t+1} = A_{t+1} + D_{t+1} + N_{t+1}. \tag{A.74}$$

Using the inequality $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, we have:

$$\mathbb{E}\|\mathbf{z}_{t+1}\|^2 \leq 3\mathbb{E}\|A_{t+1}\|^2 + 3\mathbb{E}\|D_{t+1}\|^2 + 3\mathbb{E}\|N_{t+1}\|^2. \tag{A.75}$$

We will now proceed with bounding each of these quantities.

Bounding $\mathbb{E}\|A_{t+1}\|^2$: From the stability established in Corollary A.4, we have

$$\|A_{t+1}\| = \|\boldsymbol{\Phi}(t+1,0)\mathbf{z}_0\| \leq C_\beta\rho^{t+1}\|\mathbf{z}_0\|. \tag{A.76}$$

Thus we can conclude that $\mathbb{E}\|A_{t+1}\|^2 \leq C_\beta^2\rho^{2(t+1)}\|\mathbf{z}_0\|^2$.

Bounding $\mathbb{E}\|D_{t+1}\|^2$: First simply note that $\|\mathbf{u}_t\| = \|\mathbf{b}_t\|$. Thus it is sufficient to bound $\|\mathbf{b}_t\|$. By definition, we have that $\mathbf{b}_t = -(I_d - \gamma_t H_t)\Delta_t - \mathbf{K}_t\Delta_{t-1}$. Since $\mu I_d \preceq \mathbf{H}_t \preceq L I_d$ and $\gamma_t \leq 1/L$, the eigenvalues of $I_d - \gamma_t \mathbf{H}_t$ lie in $[1-\gamma_t L, 1-\gamma_t\mu] \subset [0,1]$ so we have that $\|I_d - \gamma_t \mathbf{H}_t\| \leq 1$. Also note that since $\mathbf{K}_t = -\beta I_d + \gamma_t\beta_1\mathbf{H}_t$, we have $\|\mathbf{K}_t\| \leq \beta + \gamma_t L$. Thus putting everything together, we find

$$\|\mathbf{b}_t\| \leq \|I_d - \gamma_t\mathbf{H}_t\|\|\Delta_t\| + \|K_t\|\|\Delta_{t-1}\| \leq (1+\beta+\gamma_t L)\Delta. \tag{A.77}$$

Using triangle inequality and the bound on $\|\mathbf{b}_t\|$, we have

$$\|D_{t+1}\|^2 \leq \sum_{k=0}^{t}\|\mathbf{\Phi}(t+1,k+1)\|^2\|\mathbf{b}_k\|^2 \leq C_\beta^2\Delta^2(1+\beta+\gamma_t L)^2\left(\sum_{k=0}^{t}\rho^{t-k}\right)^2. \tag{A.78}$$

Using the fact that $\sum_{j\geq 0}\rho^j \leq 1/(1-\rho)$, we conclude that

$$\mathbb{E}\|D_{t+1}\|^2 \leq C_\beta^2(1+\beta+\gamma_t L)^2\frac{\Delta^2}{(1-\rho)^2}. \tag{A.79}$$

Bounding $\mathbb{E}\|N_{t+1}\|^2$: We will work with the recursion that N_{t+1} satisfies:

$$N_{t+1} = \mathbf{B}_t N_t + \gamma_t\boldsymbol{\eta}_{t+1} \tag{A.80}$$

with $N_0 = 0$. Multiplying by $\mathbf{V}^{-1}$, we find that

$$n_{t+1} = \widetilde{\mathbf{B}}_t n_t + \gamma_t\widetilde{\boldsymbol{\eta}}_{t+1} \tag{A.81}$$

where $n_{t+1} = \mathbf{V}^{-1}N_{t+1}$, $\widetilde{\mathbf{B}}_t = \mathbf{V}^{-1}\mathbf{B}_t\mathbf{V}$, and $\widetilde{\boldsymbol{\eta}}_{t+1} = \mathbf{V}^{-1}\boldsymbol{\eta}_{t+1}$. Now since $\widetilde{\mathbf{B}}_t n_t$ is $\mathcal{F}_t$ measurable and $\mathbb{E}[\widetilde{\boldsymbol{\eta}}_{t+1} \mid \mathcal{F}_t] = 0$, we have by conditioning and then iterated expectation that

$$\mathbb{E}[\|n_{t+1}\|^2] = \|\widetilde{\mathbf{B}}_t n_t\|^2 + \gamma_t^2\mathbb{E}[\|\widetilde{\boldsymbol{\eta}}_{t+1}\|^2]. \tag{A.82}$$

From the stability established in Corollary A.3, we have that $\|\widetilde{\mathbf{B}}_t n_t\|^2 \leq \rho\|n_t\|^2$. Also note that from Lemma A.2, we have an explicit form for $\widetilde{\boldsymbol{\eta}}_{t+1}$:

$$\widetilde{\boldsymbol{\eta}}_{t+1} = \frac{1}{1-\beta}\begin{bmatrix}\xi_{t+1}(\boldsymbol{\psi}_{t+1})\\ \xi_{t+1}(\boldsymbol{\psi}_{t+1})\end{bmatrix}. \tag{A.83}$$

Under the boundedness assumption (Assumption 3.1), we have that

$$\mathbb{E}[\|\widetilde{\boldsymbol{\eta}}_{t+1}\|^2] \leq \frac{2\sigma^2}{(1-\beta)^2}. \tag{A.84}$$

Thus we find that

$$\mathbb{E}[\|n_{t+1}\|^2] = \rho\|n_t\|^2 + \frac{2\gamma_t^2\sigma^2}{(1-\beta)^2}. \tag{A.85}$$

Unrolling this expression and using the fact that $N_0 = 0$, we find that

$$\mathbb{E}[\|n_{t+1}\|^2] \leq \frac{2\gamma_t^2\sigma^2}{(1-\beta)^2(1-\rho)^2}. \tag{A.86}$$

Thus we can conclude with:

$$\mathbb{E}\|N_{t+1}\|^2 = \|\mathbf{V}\|^2\mathbb{E}[\|n_{t+1}\|^2] \leq \frac{8\gamma_t^2\sigma^2}{(1-\beta)^2(1-\rho)^2}. \tag{A.87}$$

Combining everything and noting that $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \leq \|\mathbf{z}_t\|^2$, we conclude that

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \leq \frac{48}{(1-\beta)^2}\rho^{2(t+1)}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{48(1+\beta+\gamma L)^2}{(1-\beta)^2}\cdot\frac{\Delta^2}{(1-\rho)^2} + \frac{24}{(1-\beta)^2}\cdot\frac{\sigma^2\gamma^2}{1-\rho^2}. \tag{A.88}$$

From Corollary A.3, we can take $\rho = 1 - \gamma\mu/2(1-\beta)$ and use the inequality $(1-a)^{2(t+1)} \leq e^{-2a(t+1)}$ to conclude:

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \leq \frac{48}{(1-\beta)^2}\exp\left(-\frac{\gamma\mu}{1-\beta}(t+1)\right)\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{192(2+\beta)^2}{\gamma^2\mu^2}\Delta^2 + \frac{96\sigma^2\gamma}{\mu(1-\beta)}. \tag{A.89}$$

□

Using Corollary A.5, we can similarly obtain a algorithmic sample complexity guarantee for SGD:

Corollary A.6 (Time to reach the asymptotic tracking error in expectation for SGD with momentum)**.** *Assume Assumption 2.2 Assumption 2.3, and Assumption 3.1 and let $\beta \in [0, 1)$. Consider SGD with momentum (2.5). Assume a constant stepsize $\gamma_t \equiv \gamma$ satisfying $\gamma \leq \mu(1-\beta)^2/(4L^2)$, and set*

$$\rho \ := \ 1 - \frac{\gamma\mu}{2(1-\beta)} \in (0, 1).$$

Define the (stepsize-dependent) steady-state tracking error

$$\mathcal{E}_\beta(\gamma) = \frac{192(2+\beta)^2}{\mu^2\gamma^2}\,\Delta^2 + \frac{96}{\mu(1-\beta)}\,\sigma^2\gamma, \quad \gamma_\beta^\star \in \arg\min_{\gamma \in (0,\ \mu(1-\beta)^2/(4L^2)]} \mathcal{E}_\beta(\gamma), \quad \mathcal{E}_\beta := \mathcal{E}_\beta(\gamma_\beta^\star). \tag{A.90}$$

Then:

(i) ***(Constant learning rate).*** *If $\gamma_t \equiv \gamma_\beta^\star$, then for all $t \geq 0$,*

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \ \lesssim \ \mathcal{E}_\beta \quad \textit{after time} \quad t \ \lesssim \ \frac{1-\beta}{\mu\gamma_\beta^\star}\log\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{(1-\beta)^2\mathcal{E}_\beta}\right).$$

(ii) ***(Step-decay schedule with momentum restart).*** *Suppose $\gamma_\beta^\star < \mu(1-\beta)^2/(4L^2)$ (i.e., the minimizer of $\mathcal{E}_\beta(\gamma)$ is not at the stability cap). Define the epoch stepsizes*

$$\gamma_0 := \frac{\mu(1-\beta)^2}{4L^2}, \qquad \gamma_k := \frac{\gamma_{k-1} + \gamma_\beta^\star}{2} \quad (k \geq 1), \qquad K := 1 + \left\lceil \log_2\left(\frac{\gamma_0}{\gamma_\beta^\star}\right)\right\rceil.$$

Define epoch lengths

$$T_0 := \left\lceil \frac{1-\beta}{\mu\gamma_0}\log\left(\frac{2\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{(1-\beta)^2\mathcal{E}_\beta(\gamma_0)}\right)\right\rceil, \qquad T_k := \left\lceil \frac{1-\beta}{\mu\gamma_k}\log 4\right\rceil \quad (k \geq 1).$$

Run SGD with momentum with constant stepsize γ_k for T_k iterations in epoch k, restarting the momentum buffer at the start of each epoch *(equivalently, set $v = 0$ or $\theta_{t_k - 1} = \theta_{t_k}$ at epoch boundaries). Let $T := \sum_{k=0}^{K-1} T_k$ be the total horizon. Then the final iterate satisfies*

$$\mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \ \lesssim \ \mathcal{E}_\beta \quad \textit{after time} \quad T \ \lesssim \ \frac{L^2}{\mu^2(1-\beta)}\log\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{(1-\beta)^2\mathcal{E}_\beta}\right) \ + \ \frac{\sigma^2}{\mu^2\mathcal{E}_\beta},$$

up to universal numerical constants.

Proof of Corollary A.6. Fix any stepsize $\gamma \in (0, \gamma_{\max}]$ and define

$$\rho(\gamma) := 1 - \frac{\gamma\mu}{2(1-\beta)} \in (0, 1).$$

By Corollary A.5 with this choice of ρ and the inequality $\rho^{2(t+1)} \leq \exp(-\gamma\mu(t+1)/(1-\beta))$, we have for all $t \geq 0$:

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \ \leq \ \underbrace{\frac{48}{(1-\beta)^2}}_{\triangleq\, C_\beta}\exp\left(-\frac{\mu\gamma}{1-\beta}(t+1)\right)\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 \ + \ \mathcal{E}_\beta(\gamma). \tag{A.91}$$

Since the assumptions are *uniform in time* (same μ, L, Δ, σ for all t), the same inequality applies when the method is started at any time $s \geq 0$ with initial point θ_s and initial minimizer $\boldsymbol{\theta}_s^\star$, *provided the momentum buffer is reset at time* s. Concretely, if we set $v_s = 0$ (equivalently $\boldsymbol{\theta}_{s-1} = \boldsymbol{\theta}_s$), and run for T steps with constant stepsize γ, then (A.91) (with t replaced by $T-1$ and $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0^\star)$ replaced by $(\boldsymbol{\theta}_s, \boldsymbol{\theta}_s^\star)$) yields:

$$\mathbb{E}\|\boldsymbol{\theta}_{s+T} - \boldsymbol{\theta}_{s+T}^\star\|^2 \ \leq \ C_\beta\exp\left(-\frac{\mu\gamma}{1-\beta}T\right)\mathbb{E}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^\star\|^2 \ + \ \mathcal{E}_\beta(\gamma). \tag{A.92}$$

Proof of (i). Apply (A.91) with $\gamma = \gamma_\beta^\star$:

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \le C_\beta \exp\Big(-\frac{\mu\gamma_\beta^\star}{1-\beta}(t+1)\Big)\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \mathcal{E}_\beta.$$

If

$$t \;\ge\; \frac{1-\beta}{\mu\gamma_\beta^\star}\log\Big(\frac{C_\beta\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2}{\mathcal{E}_\beta}\Big),$$

then the transient term is at most $\mathcal{E}_\beta$ and hence the total is at most $2\mathcal{E}_\beta$. Substituting $C_\beta = 48/(1-\beta)^2$ yields the stated sufficient condition.

Proof of (ii). Let $t_0 := 0$ and $t_{k+1} := t_k + T_k$, and define the epoch iterates and epoch minimizers

$$X_k := \boldsymbol{\theta}_{t_k}, \qquad X_k^\star := \boldsymbol{\theta}_{t_k}^\star.$$

By construction, at the start of each epoch k we restart the momentum buffer (set $v_{t_k} = 0$), so we may apply the one-epoch bound (A.92) with $s = t_k$, $T = T_k$, and $\gamma = \gamma_k$:

$$\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \;\le\; C_\beta \exp\Big(-\frac{\mu\gamma_k}{1-\beta}T_k\Big)\mathbb{E}\|X_k - X_k^\star\|^2 \;+\; \mathcal{E}_\beta(\gamma_k). \tag{A.93}$$

Contraction for epochs $k \ge 1$. For $k \ge 1$, by the definition of T_k we have

$$T_k \;\ge\; \frac{1-\beta}{\mu\gamma_k}\log(4C_\beta),$$

and therefore

$$C_\beta \exp\Big(-\frac{\mu\gamma_k}{1-\beta}T_k\Big) \;\le\; C_\beta\exp(-\log(4C_\beta)) \;=\; \frac{1}{4}.$$

Substituting this into (A.93) yields, for all $k \ge 1$,

$$\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \;\le\; \frac{1}{4}\mathbb{E}\|X_k - X_k^\star\|^2 \;+\; \mathcal{E}_\beta(\gamma_k). \tag{A.94}$$

Base epoch ($k = 0$). By the definition of T_0,

$$T_0 \;\ge\; \frac{1-\beta}{\mu\gamma_0}\log\Big(\frac{2C_\beta\|X_0 - X_0^\star\|^2}{\mathcal{E}_\beta(\gamma_0)}\Big),$$

so

$$C_\beta \exp\Big(-\frac{\mu\gamma_0}{1-\beta}T_0\Big)\|X_0 - X_0^\star\|^2 \;\le\; \frac{1}{2}\mathcal{E}_\beta(\gamma_0).$$

Plugging into (A.93) at $k = 0$ gives

$$\mathbb{E}\|X_1 - X_1^\star\|^2 \;\le\; \frac{1}{2}\mathcal{E}_\beta(\gamma_0) \;+\; \mathcal{E}_\beta(\gamma_0) \;=\; \frac{3}{2}\mathcal{E}_\beta(\gamma_0) \;\le\; 2\,\mathcal{E}_\beta(\gamma_0). \tag{A.95}$$

Induction. We claim that for all $k \ge 1$,

$$\mathbb{E}\|X_k - X_k^\star\|^2 \;\le\; 2\,\mathcal{E}_\beta(\gamma_{k-1}). \tag{A.96}$$

The base case $k = 1$ follows from (A.95). Assume (A.96) holds for some $k \ge 1$. Applying (A.94) yields

$$\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \le \frac{1}{4}\cdot 2\mathcal{E}_\beta(\gamma_{k-1}) + \mathcal{E}_\beta(\gamma_k) = \frac{1}{2}\mathcal{E}_\beta(\gamma_{k-1}) + \mathcal{E}_\beta(\gamma_k).$$

Since $\gamma_k = (\gamma_{k-1} + \gamma_\beta^\star)/2$, we have $\gamma_{k-1} \le 2\gamma_k$. Write $\mathcal{E}_\beta(\gamma) = A_\beta/\gamma^2 + B_\beta\gamma$ with $A_\beta, B_\beta > 0$. For any $\gamma > 0$,

$$\mathcal{E}_\beta(2\gamma) = \frac{A_\beta}{4\gamma^2} + 2B_\beta\gamma \;\le\; 2\Big(\frac{A_\beta}{\gamma^2} + B_\beta\gamma\Big) = 2\mathcal{E}_\beta(\gamma),$$

so in particular

$$\mathcal{E}_\beta(\gamma_{k-1}) \;\le\; \mathcal{E}_\beta(2\gamma_k) \;\le\; 2\mathcal{E}_\beta(\gamma_k).$$

Therefore,

$$\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \le \frac{1}{2}\cdot 2\mathcal{E}_\beta(\gamma_k) + \mathcal{E}_\beta(\gamma_k) = 2\,\mathcal{E}_\beta(\gamma_k),$$

which proves (A.96) at $k+1$ and closes the induction. Hence $\mathbb{E}\|X_K - X_K^\star\|^2 \le 2\mathcal{E}_\beta(\gamma_{K-1})$. By the definition of K,

$$\gamma_{K-1} - \gamma_\beta^\star = \frac{\gamma_0 - \gamma_\beta^\star}{2^{K-1}} \le \frac{\gamma_0}{2^{K-1}} \le \gamma_\beta^\star, \quad \text{so} \quad \gamma_{K-1} \le 2\gamma_\beta^\star.$$

Thus,

$$\mathcal{E}_\beta(\gamma_{K-1}) \le \mathcal{E}_\beta(2\gamma_\beta^\star) \le 2\mathcal{E}_\beta(\gamma_\beta^\star) = 2\mathcal{E}_\beta,$$

and therefore

$$\mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 = \mathbb{E}\|X_K - X_K^\star\|^2 \le 2\mathcal{E}_\beta(\gamma_{K-1}) \le 4\mathcal{E}_\beta.$$

Time bound. Since $\mathcal{E}_\beta(\gamma_0) \ge \mathcal{E}_\beta$, we have

$$T_0 \le \frac{1-\beta}{\mu\gamma_0}\log\left(\frac{2C_\beta\|\theta_0 - \theta_0^\star\|^2}{\mathcal{E}_\beta}\right) + 1.$$

Moreover, for $k \ge 1$, $\gamma_k \ge \gamma_0/2^{k+1}$ (by the same argument as in the SGD proof), hence

$$\sum_{k=1}^{K-1}\frac{1}{\gamma_k} \le \sum_{k=1}^{K-1}\frac{2^{k+1}}{\gamma_0} \le \frac{2^{K+1}}{\gamma_0} \le \frac{4}{\gamma_\beta^\star}.$$

Therefore,

$$\sum_{k=1}^{K-1} T_k \le \sum_{k=1}^{K-1}\left(\frac{1-\beta}{\mu\gamma_k}\log(4C_\beta) + 1\right) \le \frac{(1-\beta)\log(4C_\beta)}{\mu}\sum_{k=1}^{K-1}\frac{1}{\gamma_k} + K \le \frac{4(1-\beta)\log(4C_\beta)}{\mu\gamma_\beta^\star} + K.$$

Combining with the bound on T_0 yields the stated explicit horizon bound. Finally, in the interior regime (unconstrained minimizer), $\mathcal{E}_\beta(\gamma) = A_\beta/\gamma^2 + B_\beta\gamma$ satisfies $\mathcal{E}_\beta = \frac{3}{2}B_\beta\gamma_\beta^\star$, so

$$\frac{1-\beta}{\mu\gamma_\beta^\star} = \frac{1-\beta}{\mu}\cdot\frac{B_\beta}{(2/3)\mathcal{E}_\beta} = \frac{1-\beta}{\mu}\cdot\frac{96\sigma^2/(\mu(1-\beta))}{(2/3)\mathcal{E}_\beta} = \frac{144\sigma^2}{\mu^2\mathcal{E}_\beta}.$$

Also $\gamma_0 = \mu(1-\beta)^2/(4L^2)$ implies $\frac{1-\beta}{\mu\gamma_0} = \frac{4L^2}{\mu^2(1-\beta)}$. Substituting yields the final bound on T. □

B Proofs of tracking error bounds with high probability

For this section, we will assume a light-tailed assumption on the gradient noise (Assumption 3.2)

Assumption B.1 (Conditional sub-Gaussian gradient noise along iterates)**.** *There exists a constant $\sigma > 0$ such that for all $t \ge 0$, $\left\|\boldsymbol{\xi}_{t+1}(\boldsymbol{\theta}_t) \mid \mathcal{F}_t\right\|_{\Psi_2} \le \sigma$ a.s.*

B.1 Proof for SGD high probability tracking error bound

We will prove the following high probability bound holds for the tracking error in SGD:

Theorem B.1 (High probability tracking error bound for SGD)**.** *Under Assumption 3.2, for all $t \in [T]$, $\gamma \le \min\{\mu/L^2, 1/L\}$, and $\delta \in (0,1)$, the following tracking error bound holds for SGD with probability atleast $1-\delta$,*

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \lesssim \left(1 - \frac{\gamma\mu}{2}\right)^T\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{\mathfrak{D}_T}{\gamma\mu} + \left(d\sigma^2\gamma^2 + \frac{d^2\sigma^4\gamma^3}{\mu}\right)\log\frac{2T}{\delta} + \left(\frac{\sigma^2\gamma}{\mu} + \gamma^2\sigma^2\max_{t\in[T]}\mathfrak{D}_t^{(2)}\right)\log\frac{2T}{\delta}$$

where $\mathfrak{D}_t := \sum_{\ell=0}^{t-1}(1-\gamma\mu/2)^{t-\ell-1}\|\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star\|^2$ and $\mathfrak{D}_t^2 := \sum_{\ell=0}^{t-1}(1-\gamma\mu/2)^{2(t-\ell-1)}\|\boldsymbol{\theta}_t^\star - \boldsymbol{\theta}_{t+1}^\star\|^2$.

Proof of Theorem 3.5. First recall by Proposition A.1 that we have

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \le \left(1 - \frac{\gamma\mu}{2}\right)^T \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{2\mathfrak{D}_T}{\gamma\mu} + \underbrace{\gamma^2 \sum_{\ell=0}^{T-1} \left(1 - \frac{\gamma\mu}{2}\right)^{T-\ell-1} \|\xi_\ell(\boldsymbol{\theta}_\ell)\|^2}_{\text{(a)}} + \underbrace{\sum_{\ell=0}^{T-1} \left(1 - \frac{\gamma\mu}{2}\right)^{T-\ell-1} M_\ell}_{\text{(b)}} \tag{B.1}$$

where $M_{t+1} := -2\gamma_t \langle d_t - \gamma_t m_{t+1}(\boldsymbol{\theta}_t), \xi_{t+1}(\boldsymbol{\theta}_t)\rangle$ with $d_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}^\star$. It remains to bound (a) and (b).

Bounding part (a): First notice the following equivalence:

$$\|\xi_\ell(\boldsymbol{\theta}_t)\|^2 = \mathbb{E}[\|\xi_\ell(\boldsymbol{\theta}_t)\|^2 \mid \mathcal{F}_{\ell-1}] + V_\ell, \quad V_\ell := \|\xi_\ell(\boldsymbol{\theta}_t)\|^2 - \mathbb{E}[\|\xi_\ell(\boldsymbol{\theta}_t)\|^2 \mid \mathcal{F}_{\ell-1}]. \tag{B.2}$$

First we bound V_ℓ. Note that V_ℓ is a martingale difference sequence (MDS) and sub-exponential with $\|V_\ell \mid \mathcal{F}_{\ell-1}\|_{\Psi_1} \lesssim d\sigma^2$ (Lemma D.1). Define $\rho := (1 - \gamma\mu/2)$. Then for fixed $t \le T$, let $Z_\ell^{(t)} := \gamma^2 \rho^{t-\ell-1} V_\ell$. Then we have that $\|Z_\ell^{(t)} \mid \mathcal{F}_{\ell-1}\|_{\Psi_1} \lesssim \gamma^2 \rho^{t-\ell-1} d\sigma^2$. We also have the following that hold:

$$\sum_{\ell=0}^{T-1} \gamma^4 \rho^{2(t-\ell-1)} d^2\sigma^4 \lesssim \frac{d^2\sigma^4\gamma^3}{\mu}, \quad \max_{0\le\ell\le T-1} \gamma^2 \rho^{T-\ell-1} d\sigma^2 \lesssim d\sigma^2\gamma^2. \tag{B.3}$$

By Bernstein's inequality for sub-exponential (Lemma D.5), we have:

$$\mathbb{P}\left(\sum_{\ell=0}^{T-1} Z_\ell^{(t)} \ge s\right) \lesssim \exp\left(-\min\left\{\frac{s^2\mu}{d^2\sigma^4\gamma^3}, \frac{s}{\gamma^2 d\sigma^2}\right\}\right). \tag{B.4}$$

Take $s = (d\sigma^2\gamma^2 + \frac{d^2\sigma^2\gamma^3}{\mu})\log(2T/\delta)$. Furthermore since we have $\|\xi_\ell(\boldsymbol{\theta}_\ell) \mid \mathcal{F}_{\ell-1}\|_{\Psi_2} \le \sigma$, we have $\mathbb{E}[\|\xi_\ell(\boldsymbol{\theta}_\ell)^2\| \mid \mathcal{F}_{\ell-1}] \le d\sigma^2$ (Lemma D.4). Thus we have:

$$\gamma^2 \sum_{\ell=0}^{T-1} \rho^{t-\ell-1} \mathbb{E}[\|\xi_\ell(\boldsymbol{\theta}_\ell)^2\| \mid \mathcal{F}_{\ell-1}] \lesssim \frac{d\sigma^2\gamma}{\mu}. \tag{B.5}$$

Combining everything, we obtain the event $\mathcal{E}_\xi(t)$ that with probability atleast $1 - \delta/2T$,

$$\gamma^2 \sum_{\ell=0}^{T-1} \left(1 - \frac{\gamma\mu}{2}\right)^{t-\ell-1} \|\xi_\ell(\boldsymbol{\theta}_t)\|^2 \lesssim \frac{d\sigma^2\gamma}{\mu} + (d\sigma^2\gamma^2 + \frac{d^2\sigma^2\gamma^3}{\mu})\log(2T/\delta). \tag{B.6}$$

A union bound over $t = 1, \ldots, T$ gives the event $\mathcal{E}_\xi = \cap_{t\le T} \mathcal{E}_\xi(t)$ with $\mathbb{P}(\mathcal{E}_\xi) \ge 1 - \delta/2$. It remains to bound (b).

Bounding part (b): Recall that $M_{\ell+1} := -2\gamma_\ell \langle d_\ell - \gamma_\ell m_{\ell+1}(\boldsymbol{\theta}_\ell), \xi_{\ell+1}(\boldsymbol{\theta}_\ell)\rangle$ with $d_\ell = \boldsymbol{\theta}_\ell - \boldsymbol{\theta}_{\ell+1}^\star$. Define $a_\ell := d_\ell - \gamma\, m_{\ell+1}(\boldsymbol{\theta}_\ell)$ Since a_ℓ is $\mathcal{F}_\ell$–measurable and $\mathbb{E}[\xi_{\ell+1}(\boldsymbol{\theta}_\ell) \mid \mathcal{F}_\ell] = 0$, we have $\mathbb{E}[M_{\ell+1} \mid \mathcal{F}_\ell] = 0$ and thus $(M_{\ell+1})_{\ell\ge0}$ is a martingale difference sequence (MDS).

By Assumption 3.2, for any $\mathcal{F}_\ell$–measurable unit vector u, $\|u^\top \xi_{\ell+1}(\boldsymbol{\theta}_\ell) \mid \mathcal{F}_\ell\|_{\Psi_2} \le \sigma$ a.s. Hence, for any $\mathcal{F}_\ell$–measurable vector v, $\langle v, \xi_{\ell+1}(\boldsymbol{\theta}_\ell)\rangle$ is conditionally sub-Gaussian given $\mathcal{F}_\ell$, $\|\langle v, \xi_{\ell+1}(\boldsymbol{\theta}_\ell)\rangle \mid \mathcal{F}_\ell\|_{\Psi_2} \le \sigma\|v\|$. Moreover, by the contraction inequality proved in Lemma A.1 (using $\gamma \le \mu/L^2$), $\|a_\ell\|^2 = \|d_\ell - \gamma m_{\ell+1}(\boldsymbol{\theta}_\ell)\|^2 \le \|d_\ell\|^2$. Combining these yields

$$\|M_{\ell+1} \mid \mathcal{F}_\ell\|_{\Psi_2} \le 2\gamma\,\sigma\,\|a_\ell\| \le 2\gamma\,\sigma\,\|d_\ell\|. \tag{B.7}$$

Fix a (deterministic) radius $\mathfrak{B} > 0$:

$$\mathfrak{B} := \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{1}{\gamma\mu}\max_{t\in[T]} \mathfrak{D}_t + \frac{d\sigma^2\gamma}{\mu} + \left(d\sigma^2\gamma^2 + \frac{d^2\sigma^4\gamma^3}{\mu}\right)\log\frac{2T}{\delta} + \left(\frac{\sigma^2\gamma}{\mu} + \gamma^2\sigma^2 \max_{t\in[T]} \mathfrak{D}_t^{(2)}\right)\log\frac{2T}{\delta}, \tag{B.8}$$

and define the stopping time

$$\tau := \inf \left\{ t \in [T] : \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 > \mathfrak{B} \right\}. \tag{B.9}$$

Define the stopped increments $M_{\ell+1}^\tau := M_{\ell+1} \mathbf{1}\{\ell < \tau\}$. Since $\mathbf{1}\{\ell < \tau\}$ is $\mathcal{F}_\ell$–measurable and $\mathbb{E}[M_{\ell+1} \mid \mathcal{F}_\ell] = 0$,

$$\mathbb{E}[M_{\ell+1}^\tau \mid \mathcal{F}_\ell] = \mathbf{1}\{\ell < \tau\} \mathbb{E}[M_{\ell+1} \mid \mathcal{F}_\ell] = 0, \tag{B.10}$$

so $(M_{\ell+1}^\tau)_{\ell \geq 0}$ is also an MDS. On the event $\{\ell < \tau\}$ we have $\|e_\ell\|^2 \leq \mathfrak{B}$, hence by triangle inequality,

$$\|d_\ell\|^2 = \|e_\ell + \Delta_\ell\|^2 \lesssim \mathfrak{B} + \|\Delta_\ell\|^2. \tag{B.11}$$

Plugging (B.11) into (B.7) yields the localized conditional sub-Gaussian bound:

$$\|M_{\ell+1}^\tau \mid \mathcal{F}_\ell\|_{\Psi_2} \leq 2\gamma\sigma \, \|d_\ell\| \mathbf{1}\{\ell < \tau\} \lesssim \gamma\sigma \, \sqrt{\mathfrak{B} + \|\Delta_\ell\|^2} \, \mathbf{1}\{\ell < \tau\}. \tag{B.12}$$

Fix an evaluation time $t \in \{1, \ldots, T\}$ and set $\rho := 1 - \gamma\mu/2 \in (0,1)$. Define the deterministic weights $w_\ell^{(t)} := \rho^{t-\ell-1}$ for $\ell = 0, \ldots, t-1$ and the weighted stopped increments

$$Z_{\ell+1}^{(t)} := w_\ell^{(t)} \, M_{\ell+1}^\tau = \rho^{t-\ell-1} M_{\ell+1}^\tau, \qquad \ell = 0, \ldots, t-1.$$

By (B.10) and the fact that $w_\ell^{(t)}$ is deterministic, $\mathbb{E}[Z_{\ell+1}^{(t)} \mid \mathcal{F}_\ell] = 0$, i.e., $(Z_{\ell+1}^{(t)})_{\ell=0}^{t-1}$ is an MDS. Let $S_k^{(t)} := \sum_{\ell=0}^{k-1} Z_{\ell+1}^{(t)}$ denote its partial sums ($k = 0, 1, \ldots, t$). Define the predictable scale

$$v_\ell^{(t)} := \gamma\sigma \, \rho^{t-\ell-1} \sqrt{\mathfrak{B} + \|\Delta_\ell\|^2}, \qquad \ell = 0, \ldots, t-1,$$

so that (B.12) implies $\|Z_{\ell+1}^{(t)} \mid \mathcal{F}_\ell\|_{\Psi_2} \leq v_\ell^{(t)}$. A standard implication of conditional Ψ_2 control is the conditional mgf bound: there exists an absolute constant $c > 0$ such that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\big(\lambda Z_{\ell+1}^{(t)}\big) \,\Big|\, \mathcal{F}_\ell\Big] \leq \exp\big(c\lambda^2 (v_\ell^{(t)})^2\big) \qquad \text{a.s.} \tag{B.13}$$

Define the predictable variance proxy $V_k^{(t)} := \sum_{\ell=0}^{k-1} (v_\ell^{(t)})^2$ and the exponential process

$$Y_k(\lambda) := \exp\Big(\lambda S_k^{(t)} - c\lambda^2 V_k^{(t)}\Big), \qquad k = 0, 1, \ldots, t.$$

Using (B.13) we have

$$\mathbb{E}[Y_{k+1}(\lambda) \mid \mathcal{F}_k] = Y_k(\lambda) \, \mathbb{E}\Big[\exp\big(\lambda Z_{k+1}^{(t)} - c\lambda^2 (v_k^{(t)})^2\big) \,\Big|\, \mathcal{F}_k\Big] \leq Y_k(\lambda),$$

so $(Y_k(\lambda))_{k=0}^t$ is a nonnegative $(\mathcal{F}_k)$–supermartingale. Now apply optional sampling at the bounded stopping time $\tau \wedge t$ (Lemma D.6):

$$\mathbb{E}\big[Y_{\tau\wedge t}(\lambda)\big] \leq \mathbb{E}[Y_0(\lambda)] = 1. \tag{B.14}$$

Since $Z_{\ell+1}^{(t)}$ already includes the factor $\mathbf{1}\{\ell < \tau\}$, we have $S_{\tau\wedge t}^{(t)} = S_t^{(t)} = \sum_{\ell=0}^{t-1} \rho^{t-\ell-1} M_{\ell+1}^\tau$. Therefore, by Markov's inequality and (B.14), for any $s > 0$,

$$\mathbb{P}\left(\sum_{\ell=0}^{t-1} \rho^{t-\ell-1} M_{\ell+1}^\tau \geq s\right) \leq \exp\Big(-\lambda s + c\lambda^2 V_t^{(t)}\Big).$$

Optimizing over λ yields the sub-Gaussian tail

$$\mathbb{P}\left(\sum_{\ell=0}^{t-1} \rho^{t-\ell-1} M_{\ell+1}^\tau \geq s\right) \lesssim \exp\left(-\frac{s^2}{c\, V_t^{(t)}}\right). \tag{B.15}$$

By definition of $v_\ell^{(t)}$,

$$V_t^{(t)} = \sum_{\ell=0}^{t-1} (v_\ell^{(t)})^2 \lesssim \gamma^2\sigma^2 \sum_{\ell=0}^{t-1} \rho^{2(t-\ell-1)} \big(\mathfrak{B} + \|\Delta_\ell\|^2\big) = \gamma^2\sigma^2 \left(\mathfrak{B} \sum_{k=0}^{t-1} \rho^{2k} + \sum_{\ell=0}^{t-1} \rho^{2(t-\ell-1)} \|\Delta_\ell\|^2\right).$$

Using $\sum_{k=0}^{t-1} \rho^{2k} \leq \frac{1}{1-\rho^2} \leq \frac{1}{1-\rho} = \frac{2}{\gamma\mu}$, we obtain

$$V_t^{(t)} \lesssim \gamma^2\sigma^2 \left(\frac{\mathfrak{B}}{\gamma\mu} + \mathfrak{D}_t^{(2)}\right), \qquad \mathfrak{D}_t^{(2)} := \sum_{\ell=0}^{t-1} \rho^{2(t-\ell-1)} \|\Delta_\ell\|^2. \tag{B.16}$$

Plugging (B.16) into (B.15) and choosing s so that the right-hand side equals $\delta/(2T)$, we obtain the event

$$\mathcal{E}_M(t) := \left\{ \sum_{\ell=0}^{t-1} \rho^{t-\ell-1} M_{\ell+1}^\tau \lesssim \gamma\sigma \sqrt{\left(\frac{\mathfrak{B}}{\mu}\gamma + \gamma^2\sigma^2\, \mathfrak{D}_t^{(2)}\right) \log \frac{2T}{\delta}} \right\}, \tag{B.17}$$

holding with $\mathbb{P}(\mathcal{E}_M(t)) \geq 1 - \delta/(2T)$. A union bound over $t = 1, \dots, T$ yields

$$\mathbb{P}\Big(\bigcap_{t=1}^T \mathcal{E}_M(t)\Big) \geq 1 - \delta/2.$$

Applying Young's inequality $xy \leq \frac{1}{4}x^2 + y^2$ with

$$x := \sqrt{\mathfrak{B}}, \qquad y := \sigma\sqrt{\frac{\gamma}{\mu} \log \frac{2T}{\delta}},$$

and again with $x := \sqrt{\mathfrak{B}}$ and $y := \gamma\sigma\sqrt{\mathfrak{D}_t^{(2)} \log \frac{2T}{\delta}}$, we obtain that on the event $\bigcap_{t=1}^T \mathcal{E}_M(t)$,

$$\sum_{\ell=0}^{t-1} \rho^{t-\ell-1} M_{\ell+1}^\tau \;\lesssim\; \mathfrak{B} \;+\; \frac{\sigma^2\gamma}{\mu} \log \frac{2T}{\delta} \;+\; \gamma^2\sigma^2\, \mathfrak{D}_t^{(2)} \log \frac{2T}{\delta}, \qquad \forall t \in [T]. \tag{B.18}$$

Concluding the bound: Let $\mathcal{E} := \mathcal{E}_\xi \cap \bigcap_{t=1}^T \mathcal{E}_M(t)$; then $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. Working on $\mathcal{E}$, apply the stopped recursion (Lemma A.1), we have that

$$\begin{aligned}
\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 &\lesssim \left(1 - \frac{\gamma\mu}{2}\right)^T \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{\mathfrak{D}_T}{\gamma\mu} + \left(d\sigma^2\gamma^2 + \frac{d^2\sigma^4\gamma^3}{\mu}\right) \log \frac{2T}{\delta} + \left(\frac{\sigma^2\gamma}{\mu} + \gamma^2\sigma^2 \max_{t\in[T]} \mathfrak{D}_t^{(2)}\right) \log \frac{2T}{\delta}. \\
&\lesssim \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + + \frac{1}{\gamma\mu} \max_{0\leq t\leq T-1} \mathfrak{D}_t + \left(d\sigma^2\gamma^2 + \frac{d^2\sigma^4\gamma^3}{\mu}\right) \log \frac{2T}{\delta} + \left(\frac{\sigma^2\gamma}{\mu} + \gamma^2\sigma^2 \max_{t\in[T]} \mathfrak{D}_t^{(2)}\right) \log \frac{2T}{\delta}.
\end{aligned} \tag{B.19}$$

since $(1 - \gamma\mu/2)^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 \leq \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2$ and $\mathfrak{D}_t \leq \max_{0\leq t\leq T-1} \mathfrak{D}_t$. Thus the right-side is less than $\mathfrak{B}$. Hence, on $\mathcal{E}$, we have that $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \lesssim \mathfrak{B}$ for all $t \in [T]$ which forces $\tau = T + 1$ on $\mathcal{E}$ by definition of τ. Hence $M_{\ell+1}^\tau = M_{\ell+1}$ for all $\ell \leq T - 1$ on $\mathcal{E}$, i.e., the stopped bounds coincide with the original process. Taking $t = T$ above yields $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 = \|e_T\|^2 \leq \mathfrak{B}$ on $\mathcal{E}$. Since $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$, the claimed high-probability bound follows. □

B.2 Proof for SGDM high probability tracking error bound

Before we prove the SGDM high probability bound, we will first establish a recursive relation for the SGDM tracking error that we will use to prove the high probability bound, synonymous to Proposition A.1.

Proposition B.1 (Final iterate recursive relation for the SGDM tracking error)**.** *For all $t \geq 0$ and fixed $\gamma = \gamma_t \leq \min\{1/L, \mu(1-\beta)^2/4L^2\}$, the following recursive relation for the tracking error holds provided one takes a zero momentum initialization $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0$:*

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \lesssim \frac{2}{(1-\beta)^2} \tilde{\rho}^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{1}{\gamma\mu} \cdot \frac{1}{1-\beta} \mathfrak{D}_t^{\mathrm{mom}} + \frac{2\gamma^2}{(1-\beta)^2} \sum_{\ell=0}^t \tilde{\rho}^{t-\ell-1} \|\xi_{\ell+1}(\psi_\ell)\|^2 + \sum_{\ell=0}^t \tilde{\rho}^{t-\ell-1} M_{\ell+1}$$

where $\tilde{\rho} = 1 - \eta^2\mu^2/4$, $\eta = \gamma/(1-\beta)$, $\mathfrak{D}_t^{\mathrm{mom}} := \sum_{\ell\leq t} \tilde{\rho}^{t-\ell-1} \|\mathbf{b}_\ell\|^2$, and the martingale increment is $M_{\ell+1} := 2\eta\langle \widetilde{\mathbf{B}}_\ell \mathbf{y}_\ell + \mathbf{r}_\ell, \boldsymbol{\zeta}_{\ell+1}\rangle$ where $\mathbf{b}_\ell, \widetilde{\mathbf{B}}_\ell,\ \mathbf{y}_\ell,\ \mathbf{r}_\ell,\ \boldsymbol{\zeta}_{\ell+1}$ are defined in Lemma A.2.

Proof of Proposition B.1. Define $\eta := \gamma/(1-\beta)$. From Lemma A.2, we have the transformed "mode-splitting" coordinates as follows:

$$\mathbf{y}_t := \mathbf{V}^{-1}\mathbf{z}_t = \begin{bmatrix} \widehat{\boldsymbol{\theta}}_t \\ \check{\boldsymbol{\theta}}_t \end{bmatrix} = \frac{1}{1-\beta}\begin{bmatrix} \widetilde{\boldsymbol{\theta}}_t - \beta\widetilde{\boldsymbol{\theta}}_{t-1} \\ \widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{\theta}}_{t-1} \end{bmatrix}. \tag{B.20}$$

Using this transform with (A.31) gives us:

$$\underbrace{\begin{bmatrix} \widehat{\boldsymbol{\theta}}_t \\ \check{\boldsymbol{\theta}}_t \end{bmatrix}}_{\triangleq\ \mathbf{y}_t} = \underbrace{\begin{bmatrix} I_d - \frac{\gamma_t}{1-\beta}\mathbf{H}_{t-1} & \frac{\gamma_t\beta'}{1-\beta}\mathbf{H}_{t-1} \\ -\frac{\gamma_t}{1-\beta}\mathbf{H}_{t-1} & \beta I_d + \frac{\gamma_t\beta'}{1-\beta}\mathbf{H}_{t-1} \end{bmatrix}}_{\triangleq\ \widetilde{\mathbf{B}}_t} \begin{bmatrix} \widehat{\boldsymbol{\theta}}_{t-1} \\ \check{\boldsymbol{\theta}}_{t-1} \end{bmatrix} + \underbrace{\frac{1}{1-\beta}\begin{bmatrix} -(I_d - \gamma_t\mathbf{H}_{t-1})\Delta_{t-1} - \mathbf{K}_{t-1}\Delta_{t-2} \\ -(I_d - \gamma_t\mathbf{H}_{t-1})\Delta_{t-1} - \mathbf{K}_{t-1}\Delta_{t-2} \end{bmatrix}}_{\triangleq\ \mathbf{r}_t} + \underbrace{\frac{\gamma_t}{1-\beta}\begin{bmatrix} \xi_{t+1}(\boldsymbol{\psi}_{t-1}) \\ \xi_{t+1}(\boldsymbol{\psi}_{t-1}) \end{bmatrix}}_{\triangleq\ \eta\zeta_{t+1}} \tag{B.21}$$

where

$$\beta' \triangleq \beta\beta_1 + \beta_2 \tag{B.22}$$

$$\mathbf{H}_{t-1} \triangleq \int_0^1 \nabla^2 F_{t+1}(\boldsymbol{\theta}^\star_{t+1} + s(\boldsymbol{\psi}_t - \boldsymbol{\theta}^\star_{t+1}))ds \tag{B.23}$$

$$\mathbf{K}_{t-1} = -\beta I_d + \gamma_t\beta_1\mathbf{H}_{t-1}. \tag{B.24}$$

Define $E_t := \|\mathbf{y}_t\|^2 = \|\widehat{\boldsymbol{\theta}}_t\|^2 + \|\check{\boldsymbol{\theta}}_t\|^2$. From the momentum initialization, it follows that $\widehat{\boldsymbol{\theta}}_t = (\widetilde{\boldsymbol{\theta}}_0 - \beta\widetilde{\boldsymbol{\theta}}_0)/(1-\beta)$ and $\check{\boldsymbol{\theta}}_0 = 0$. This implies that

$$E_0 := \|\mathbf{y}_0\|^2 \le \frac{2}{(1-\beta)^2}\|\widetilde{\boldsymbol{\theta}}_0\|^2 \tag{B.25}$$

where the inequality holds from triangle inequality and $\beta \in (0,1)$. Another fact to note is that by definition of $\mathbf{y}_t$, one sees that $\widetilde{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_t - \beta\check{\boldsymbol{\theta}}_t$. This implies

$$\|\widetilde{\boldsymbol{\theta}}_t\|^2 \le \|\widehat{\boldsymbol{\theta}}_t\|^2 + \beta^2\|\check{\boldsymbol{\theta}}_t\|^2 \le E_t \tag{B.26}$$

where the first inequality holds from triangle inequality and $\beta \in (0,1)$. Thus it suffices to upper bound E_t.

From (B.21), we have that

$$\mathbf{y}_{t+1} = \widetilde{\mathbf{B}}_t\mathbf{y}_t + \mathbf{r}_t + \eta\zeta_{t+1}. \tag{B.27}$$

Expanding the square, we get

$$\|\mathbf{y}_{t+1}\|^2 = \|\widetilde{\mathbf{B}}_t\mathbf{y}_t + \mathbf{r}_t\|^2 + 2\eta\langle\widetilde{\mathbf{B}}_t\mathbf{y}_t + \mathbf{r}_t, \zeta_{t+1}\rangle + \eta^2\|\zeta_{t+1}\|^2. \tag{B.28}$$

We will first upper bound $\|\widetilde{\mathbf{B}}_t\mathbf{y}_t + \mathbf{r}_t\|^2$. Applying Young's inequality with $\alpha = \eta\mu/2 \in (0,1)$, we have

$$\begin{aligned} \|\widetilde{\mathbf{B}}_t\mathbf{y}_t + \mathbf{r}_t\|^2 &\le \left(1 + \frac{\eta\mu}{2}\right)\|\widetilde{\mathbf{B}}_t\mathbf{y}_t\|^2 + \left(1 + \frac{2}{\eta\mu}\right)\|\mathbf{r}_t\|^2 \\ &\le \left(1 + \frac{\eta\mu}{2}\right)\rho\|\mathbf{y}_t\|^2 + \left(1 + \frac{2}{\eta\mu}\right)\|\mathbf{r}_t\|^2. \end{aligned} \tag{B.29}$$

where the last inequality holds from the stability condition (Corollary A.3). Now observe that

$$\left(1 + \frac{\eta\mu}{2}\right)\rho = \left(1 + \frac{\eta\mu}{2}\right)\left(1 - \frac{\eta\mu}{2}\right) = 1 - \frac{\eta^2\mu^2}{4} \in (0,1) \tag{B.30}$$

Now define $M_{t+1} := 2\eta\langle\widetilde{\mathbf{B}}_t\mathbf{y}_t + \mathbf{r}_t, \zeta_{t+1}\rangle$. Since $\mathbf{B}_t\mathbf{y}_t + \mathbf{r}_t$ is $\mathcal{F}_t$ measurable and $\mathbb{E}[\zeta_{t+1} \mid \mathcal{F}_t] = 0$, we see that M_{t+1} is a MDS. Furthermore note that since $\|\zeta_{t+1}\|^2 = 2\|\xi_{t+1}(\boldsymbol{\psi}_t)\|^2$, we have

$$\eta^2\|\zeta_{t+1}\|^2 = 2\eta^2\|\xi_{t+1}(\boldsymbol{\psi}_t)\|^2. \tag{B.31}$$

Combining everything, we get the following recursive relation:

$$E_{t+1} \le \tilde{\rho}E_t + \left(1 + \frac{2}{\eta\mu}\right)\|\mathbf{r}_t\|^2 + 2\eta^2\|\xi_{t+1}(\boldsymbol{\psi}_t)\|^2 + M_{t+1}. \tag{B.32}$$

Unrolling this recursion, we find

$$E_{t+1} \leq \tilde{\rho}^t E_0 + \left(1 + \frac{2}{\eta\mu}\right) \sum_{\ell=0}^{t} \tilde{\rho}^{t-\ell-1} \|\mathbf{r}_\ell\|^2 + 2\eta^2 \sum_{\ell=0}^{t} \tilde{\rho}^{t-\ell-1} \|\xi_{\ell+1}(\boldsymbol{\psi}_\ell)\|^2 + \sum_{\ell=0}^{t} \tilde{\rho}^{t-\ell-1} M_{\ell+1}. \tag{B.33}$$

We can now simplify the drift term using $\mathbf{r}_\ell = [\mathbf{b}_\ell; \mathbf{b}_\ell]/(1-\beta)$ (A.13). This gives us

$$\sum_{\ell=0}^{t} \tilde{\rho}^{t-\ell-1} \|\mathbf{r}_\ell\|^2 = \frac{2}{(1-\beta)^2} \sum_{\ell=0}^{t} \tilde{\rho}^{t-\ell-1} \|\mathbf{r}_\ell\|^2 := \frac{2}{(1-\beta)^2} \mathfrak{D}_t^{\mathrm{mom}}. \tag{B.34}$$

We can also use $\eta = \gamma/(1-\beta)$ and (B.25) to get

$$E_{t+1} \leq \frac{2}{(1-\beta)^2} \tilde{\rho}^t \|\widetilde{\boldsymbol{\theta}}_0\|^2 + \left(1 + \frac{2}{\eta\mu}\right) \frac{2}{(1-\beta)^2} \mathfrak{D}_t^{\mathrm{mom}} + \frac{2\gamma^2}{(1-\beta)^2} \sum_{\ell=0}^{t} \tilde{\rho}^{t-\ell-1} \|\xi_{\ell+1}(\boldsymbol{\psi}_\ell)\|^2 + \sum_{\ell=0}^{t} \tilde{\rho}^{t-\ell-1} M_{\ell+1}. \tag{B.35}$$

Since $1 + 2/\eta\mu \lesssim 1/\eta\mu$, we have

$$\left(1 + \frac{2}{\eta\mu}\right) \frac{2}{(1-\beta)^2} \mathfrak{D}_t^{\mathrm{mom}} \lesssim \frac{1-\beta}{\gamma\mu} \frac{1}{(1-\beta)^2} \mathfrak{D}_t^{\mathrm{mom}} = \frac{1}{\gamma\mu} \cdot \frac{1}{1-\beta} \mathfrak{D}_t^{\mathrm{mom}}. \tag{B.36}$$

This lets us conclude using (B.26). □

We can now establish a high-probability bound on the tracking error for SGDM. This proof will be similar to Theorem 3.5.

Theorem B.2 (High probability tracking error bound for SGDM)**.** *Under Assumption 3.2, for all $t \in [T]$, $\gamma \leq \min\left\{1/L, \mu(1-\beta)^2/4L^2\right\}$, and $\delta \in (0,1)$, provided one takes a zero momentum initialization $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0$, the following tracking error bound holds for SGD with probability atleast $1-\delta$,*

$$\begin{aligned} \|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \lesssim & \frac{2}{(1-\beta)^2} \exp\left(-\frac{\gamma^2\mu^2}{4(1-\beta)^2} T\right) \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{1}{\gamma\mu} \cdot \frac{1}{1-\beta} \mathfrak{D}_t^{\mathrm{mom}} + \frac{d\sigma^2}{\mu^2} \\ & + \left(\frac{d\sigma^2\gamma^2}{(1-\beta)^2} + \frac{d^2\sigma^4\gamma^3}{\mu(1-\beta)^3}\right) \log(2T/\delta) + \left(\frac{\sigma^2}{\mu^2} + \frac{\sigma^2\gamma^2}{(1-\beta)^2} \mathfrak{D}_t^{\mathrm{mom},(2)}\right) \log\frac{2T}{\delta} \end{aligned} \tag{B.37}$$

where $\mathfrak{D}_t^{\mathrm{mom}} := \sum_{\ell \leq T-1} \tilde{\rho}^{T-\ell-1} \|\mathbf{b}_\ell\|^2$, and $\mathfrak{D}_t^{\mathrm{mom},(2)} := \sum_{\ell \leq T-1} \tilde{\rho}^{2(T-\ell-1)} \|\mathbf{b}_\ell\|^2$ with $\tilde{\rho} = 1 - \eta^2\mu^2/4$ where $\mathbf{b}_\ell := -(I_d - \gamma_t \mathbf{H}_{\ell-1})\Delta_{\ell-1} - \mathbf{K}_{\ell-1}\Delta_{\ell-2}$ with $\mathbf{H}_\ell, \mathbf{K}_\ell$ defined as in Lemma A.2.

Proof of Theorem B.2. First recall by Proposition B.1 that we have,

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \lesssim \frac{2}{(1-\beta)^2} \tilde{\rho}^T \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{1}{\gamma\mu} \cdot \frac{1}{1-\beta} \mathfrak{D}_t^{\mathrm{mom}} + \underbrace{\frac{2\gamma^2}{(1-\beta)^2} \sum_{\ell=0}^{T-1} \tilde{\rho}^{T-\ell-1} \|\xi_{\ell+1}(\psi_\ell)\|^2}_{\text{(a)}} + \underbrace{\sum_{\ell=0}^{T-1} \tilde{\rho}^{T-\ell-1} M_{\ell+1}}_{\text{(b)}} \tag{B.38}$$

where $\tilde{\rho} = 1 - \eta^2\mu^2/4$, $\eta = \gamma/(1-\beta)$, $\mathfrak{D}_t^{\mathrm{mom}} := \sum_{\ell \leq T-1} \tilde{\rho}^{T-\ell-1} \|\mathbf{b}_\ell\|^2$, and the martingale increment is $M_{\ell+1} := 2\eta \langle \widetilde{\mathbf{B}}_\ell \mathbf{y}_\ell + \mathbf{r}_\ell, \boldsymbol{\zeta}_{\ell+1} \rangle$ where $\mathbf{b}_\ell, \widetilde{\mathbf{B}}_\ell$, $\mathbf{y}_\ell$, $\mathbf{r}_\ell$, $\boldsymbol{\zeta}_{\ell+1}$ are defined in Lemma A.2. It remains to bound (a) and (b).

Bounding part (a): First notice the following equivalence:

$$\|\xi_\ell(\boldsymbol{\theta}_t)\|^2 = \mathbb{E}[\|\xi_\ell(\boldsymbol{\theta}_t)\|^2 \mid \mathcal{F}_{\ell-1}] + V_\ell, \quad V_\ell := \|\xi_\ell(\boldsymbol{\theta}_t)\|^2 - \mathbb{E}[\|\xi_\ell(\boldsymbol{\theta}_t)\|^2 \mid \mathcal{F}_{\ell-1}]. \tag{B.39}$$

First we bound V_ℓ. Note that V_ℓ is a martingale difference sequence (MDS) and sub-exponential with $\|V_\ell \mid \mathcal{F}_{\ell-1}\|_{\Psi_1} \lesssim d\sigma^2$ (Lemma D.1). Then for fixed $t \leq T$, let $Z_\ell^{(t)} := 2\gamma^2 \tilde{\rho}^{t-\ell-1} V_\ell/(1-\beta)^2$. Then we have that $\|Z_\ell^{(t)} \mid \mathcal{F}_{\ell-1}\|_{\Psi_1} \lesssim 2\gamma^2 \tilde{\rho}^{t-\ell-1} d\sigma^2/(1-\beta)^2$. We also have the following that hold:

$$\sum_{\ell=0}^{T-1} \frac{2\gamma^4}{(1-\beta)^4} \tilde{\rho}^{2(T-\ell-1)} d\sigma^2 \lesssim \frac{d^2\sigma^4\gamma^3}{\mu(1-\beta)^3}, \quad \max_{0 \leq \ell \leq T-1} \frac{2\gamma^2}{(1-\beta)^2} \tilde{\rho}^{T-\ell-1} d\sigma^2 \lesssim \frac{d\sigma^2\gamma^2}{(1-\beta)^2}. \tag{B.40}$$

By Bernstein's inequality for sub-exponential (Lemma D.5), we have:

$$\mathbb{P}\left(\sum_{\ell=0}^{T-1} Z_\ell^{(t)} \geq s\right) \lesssim \exp\left(-\min\left\{\frac{s^2\mu(1-\beta)^3}{d^2\sigma^4\gamma^3}, \frac{s(1-\beta)^2}{d\sigma^2\gamma^2}\right\}\right). \tag{B.41}$$

Take $s = (d\sigma^2\gamma^2/(1-\beta)^2)\log(T/2\delta) + (d^2\sigma^4\gamma^3/\mu(1-\beta)^3)\log(T/2\delta)$. Then with probability atleast $1-\delta/2T$, we have

$$\sum_{\ell=0}^{T-1} Z_\ell^{(t)} \lesssim \left(\frac{\gamma^2}{(1-\beta)^2}d\sigma^2\right)\log(2T/\delta) + \left(\frac{d^2\sigma^4\gamma^3}{\mu(1-\beta)^3}\right)\log(2T/\delta). \tag{B.42}$$

Furthermore since we have $\|\xi_\ell(\boldsymbol{\theta}_\ell) \mid \mathcal{F}_{\ell-1}\|_{\Psi_2} \leq \sigma$, we have $\mathbb{E}[\|\xi_\ell(\boldsymbol{\theta}_\ell)^2\| \mid \mathcal{F}_{\ell-1}] \leq d\sigma^2$ (Lemma D.4). Thus we have:

$$\frac{2\gamma^2}{(1-\beta)^2}\sum_{\ell=0}^{t-1}\tilde{\rho}^{T-\ell-1}\mathbb{E}[\|\xi_\ell(\boldsymbol{\theta}_t)\|^2 \mid \mathcal{F}_{\ell-1}] \lesssim \frac{\gamma^2}{(1-\beta)^2(1-\rho)}d\sigma^2. \tag{B.43}$$

Since $1-\tilde{\rho} = \eta^2\mu^2/4 = \gamma^2\mu^2/4(1-\beta)^2$, we have

$$\frac{\gamma^2}{(1-\beta)^2}\cdot\frac{1}{(1-\rho)} \asymp \frac{1}{\mu^2}. \tag{B.44}$$

Combining everything, we obtain the event $\mathcal{E}_\xi(t)$ that with probability atleast $1-\delta/2T$,

$$\frac{2\gamma^2}{(1-\beta)^2}\sum_{\ell=0}^{T-1}\tilde{\rho}^{T-\ell-1}\|\xi_{\ell+1}(\boldsymbol{\psi}_\ell)\|^2 \lesssim \frac{d\sigma^2}{\mu^2} + \left(\frac{d\sigma^2\gamma^2}{(1-\beta)^2} + \frac{d^2\sigma^4\gamma^3}{\mu(1-\beta)^3}\right)\log(2T/\delta). \tag{B.45}$$

A union bound over $t = 1,\ldots,T$ gives the event $\mathcal{E}_\xi = \cap_{t\leq T}\mathcal{E}_\xi(t)$ with $\mathbb{P}(\mathcal{E}_\xi) \geq 1-\delta/2$. It remains to bound (b).

Bounding part (b): Recall that $M_{\ell+1} := 2\eta\langle\widetilde{\mathbf{B}}_\ell\mathbf{y}_\ell + \mathbf{r}_\ell, \zeta_{\ell+1}\rangle$ where $\mathbf{b}_\ell, \widetilde{\mathbf{B}}_\ell,\ \mathbf{y}_\ell,\ \mathbf{r}_\ell,\ \zeta_{\ell+1}$ are defined in Lemma A.2. Define $a_\ell := \widetilde{\mathbf{B}}_\ell\mathbf{y}_\ell + \mathbf{r}_\ell \in \mathbb{R}^{2d}$. Since a_ℓ is $\mathcal{F}_\ell$–measurable and $\mathbb{E}[\zeta_{\ell+1}(\boldsymbol{\theta}_t) \mid \mathcal{F}_\ell] = 0$, we have $\mathbb{E}[M_{\ell+1} \mid \mathcal{F}_\ell] = 0$ and thus $(M_{\ell+1})_{\ell\geq 0}$ is a martingale difference sequence (MDS).

By Assumption 3.2, for any $\mathcal{F}_\ell$–measurable unit vector u, $\|u^\top\zeta_{\ell+1}(\boldsymbol{\theta}_t) \mid \mathcal{F}_\ell\|_{\Psi_2} \leq \sigma$ a.s. Hence, for any $\mathcal{F}_\ell$–measurable vector v, $\langle v, \zeta_{\ell+1}(\boldsymbol{\theta}_t)\rangle$ is conditionally sub-Gaussian given $\mathcal{F}_\ell$, $\|\langle v, \zeta_{\ell+1}(\boldsymbol{\theta}_\ell)\rangle \mid \mathcal{F}_\ell\|_{\Psi_2} \leq \sigma\|v\|$. This yields

$$\|M_{\ell+1} \mid \mathcal{F}_\ell\|_{\Psi_2} \leq 2\eta\,\sigma\,\|a_\ell\|. \tag{B.46}$$

Since $\|a_\ell\|$ depends on $\mathbf{y}_\ell$, which is random and can, a priori, be unbounded, we cannot directly apply a uniform sub-Gaussian martingale concentration inequality. We will need $\|a_\ell\|$ to be bounded on the event we want to concentrate over so we will instead proceed with an optional stopping time argument. Fix a (deterministic) radius $\mathfrak{B} > 0$:

$$\begin{aligned}\mathfrak{B} :=& \frac{2}{(1-\beta)^2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{1}{\gamma\mu}\cdot\frac{1}{1-\beta}\max_{t\in[T]}\mathfrak{D}_t^{\mathrm{mom}} + \frac{d\sigma^2}{\mu^2}\\ &+ \left(\frac{d\sigma^2\gamma^2}{(1-\beta)^2} + \frac{d^2\sigma^4\gamma^3}{\mu(1-\beta)^3}\right)\log(2T/\delta) + \left(\frac{\sigma^2}{\mu^2} + \frac{\sigma^2\gamma^2}{(1-\beta)^2}\max_{t\in[T]}\mathfrak{D}_t^{\mathrm{mom},(2)}\right)\log\frac{2T}{\delta}\end{aligned}$$

and define the stopping time

$$\tau := \inf\left\{t \in [T] : \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\star\|^2 > \mathfrak{B}\right\}. \tag{B.47}$$

Define the stopped increments $M_{\ell+1}^\tau := M_{\ell+1}\mathbf{1}\{\ell < \tau\}$. Since $\mathbf{1}\{\ell < \tau\}$ is $\mathcal{F}_\ell$–measurable and $\mathbb{E}[M_{\ell+1} \mid \mathcal{F}_\ell] = 0$,

$$\mathbb{E}[M_{\ell+1}^\tau \mid \mathcal{F}_\ell] = \mathbf{1}\{\ell < \tau\}\mathbb{E}[M_{\ell+1} \mid \mathcal{F}_\ell] = 0, \tag{B.48}$$

so $(M_{\ell+1}^\tau)_{\ell\geq 0}$ is also an MDS. On the event $\{\ell < \tau\}$ we have $\|\mathbf{y}_\ell\|^2 \leq \mathfrak{B}$, hence by triangle inequality,

$$\|a_\ell\| = \|\widetilde{\mathbf{B}}_\ell\mathbf{y}_\ell + \mathbf{r}_\ell\| \leq \|\mathbf{B}_\ell\mathbf{y}_\ell\| + \|\mathbf{r}_\ell\| \lesssim \sqrt{\mathfrak{B}} + \|\mathbf{r}_\ell\|. \tag{B.49}$$

Hence this yields a localized Ψ_2 bound:

$$\|M_{\ell+1} \mid \mathcal{F}_\ell\|_{\Psi_2} \lesssim 2\eta\sigma(\sqrt{\mathfrak{B}} + \|\mathbf{r}_\ell\|)\mathbf{1}\{\ell < \tau\}. \tag{B.50}$$

Fix an evaluation time $t \in \{1, \dots, T\}$ and set $\tilde{\rho} := 1-\gamma^2\mu^2/4 \in (0, 1)$. Define the deterministic weights $w_\ell^{(t)} := \tilde{\rho}^{\,t-\ell-1}$ for $\ell = 0, \dots, t-1$ and the weighted stopped increments

$$Z_{\ell+1}^{(t)} := w_\ell^{(t)}\, M_{\ell+1}^\tau = \tilde{\rho}^{\,t-\ell-1} M_{\ell+1}^\tau, \qquad \ell = 0, \dots, t-1.$$

By (B.48) and the fact that $w_\ell^{(t)}$ is deterministic, $\mathbb{E}[Z_{\ell+1}^{(t)} \mid \mathcal{F}_\ell] = 0$, i.e., $(Z_{\ell+1}^{(t)})_{\ell=0}^{t-1}$ is an MDS. Let $S_k^{(t)} := \sum_{\ell=0}^{k-1} Z_{\ell+1}^{(t)}$ denote its partial sums ($k = 0, 1, \dots, t$). Define the predictable scale

$$v_\ell^{(t)} := \eta\sigma\, \tilde{\rho}^{\,t-\ell-1}(\sqrt{\mathfrak{B}} + \|\mathbf{r}_\ell\|), \qquad \ell = 0, \dots, t-1,$$

so that (B.50) implies $\|Z_{\ell+1}^{(t)} \mid \mathcal{F}_\ell\|_{\Psi_2} \le v_\ell^{(t)}$. A standard implication of conditional Ψ_2 control is the conditional mgf bound: there exists an absolute constant $c > 0$ such that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\big(\lambda Z_{\ell+1}^{(t)}\big) \,\Big|\, \mathcal{F}_\ell\Big] \le \exp\big(c\lambda^2 (v_\ell^{(t)})^2\big) \qquad \text{a.s.} \tag{B.51}$$

Define the predictable variance proxy $V_k^{(t)} := \sum_{\ell=0}^{k-1}(v_\ell^{(t)})^2$ and the exponential process

$$Y_k(\lambda) := \exp\Big(\lambda S_k^{(t)} - c\lambda^2 V_k^{(t)}\Big), \qquad k = 0, 1, \dots, t.$$

Using (B.51) we have

$$\mathbb{E}[Y_{k+1}(\lambda) \mid \mathcal{F}_k] = Y_k(\lambda)\, \mathbb{E}\Big[\exp\big(\lambda Z_{k+1}^{(t)} - c\lambda^2 (v_k^{(t)})^2\big) \,\Big|\, \mathcal{F}_k\Big] \le Y_k(\lambda),$$

so $(Y_k(\lambda))_{k=0}^t$ is a nonnegative $(\mathcal{F}_k)$–supermartingale. Now apply optional sampling at the bounded stopping time $\tau \wedge t$ (Lemma D.6):

$$\mathbb{E}\big[Y_{\tau\wedge t}(\lambda)\big] \le \mathbb{E}[Y_0(\lambda)] = 1. \tag{B.52}$$

Since $Z_{\ell+1}^{(t)}$ already includes the factor $\mathbf{1}\{\ell < \tau\}$, we have $S_{\tau\wedge t}^{(t)} = S_t^{(t)} = \sum_{\ell=0}^{t-1} \tilde{\rho}^{t-\ell-1} M_{\ell+1}^\tau$. Therefore, by Markov's inequality and (B.52), for any $s > 0$,

$$\mathbb{P}\left(\sum_{\ell=0}^{t-1} \tilde{\rho}^{t-\ell-1} M_{\ell+1}^\tau \ge s\right) \le \exp\Big(-\lambda s + c\lambda^2 V_t^{(t)}\Big).$$

Optimizing over λ yields the sub-Gaussian tail

$$\mathbb{P}\left(\sum_{\ell=0}^{t-1} \tilde{\rho}^{t-\ell-1} M_{\ell+1}^\tau \ge s\right) \lesssim \exp\left(-\frac{s^2}{c\, V_t^{(t)}}\right). \tag{B.53}$$

By definition of $v_\ell^{(t)}$,

$$V_t^{(t)} = \sum_{\ell=0}^{t-1} (v_\ell^{(t)})^2 \lesssim \eta^2\sigma^2 \sum_{\ell=0}^{t-1} \tilde{\rho}^{2(t-\ell-1)}(\mathfrak{B} + \|\mathbf{r}_\ell\|^2) = \eta^2\sigma^2 \left(\mathfrak{B}\sum_{k=0}^{t-1}\tilde{\rho}^{2k} + \sum_{\ell=0}^{t-1}\tilde{\rho}^{2(t-\ell-1)}\|\mathbf{r}_\ell\|^2\right).$$

Using $\sum_{k=0}^{t-1}\tilde{\rho}^{2k} \le \frac{1}{1-\tilde{\rho}^2} \le \frac{1}{1-\tilde{\rho}} = \frac{4(1-\beta)^2}{\gamma^2\mu^2}$ and $\eta = \gamma/(1-\beta)$, we get

$$V_t^{(t)} \lesssim \frac{\sigma^2}{\mu^2}\mathfrak{B} + \frac{\sigma^2\gamma^2}{(1-\beta)^2}\mathfrak{D}_t^{(2)}, \qquad \mathfrak{D}_t^{\mathrm{mom},(2)} := \sum_{\ell=0}^{t-1} \tilde{\rho}^{2(t-\ell-1)}\|\mathbf{r}_\ell\|^2. \tag{B.54}$$

Choose $s > 0$ such that $\exp\left(-s^2/c\, V_t(t)\right) = \delta/2T$. Then with probability at least $1 - \delta/(2T)$,

$$\sum_{\ell=0}^{t-1} \tilde{\rho}^{\,t-\ell-1} M_{\ell+1}^{\tau} \;\lesssim\; \sqrt{V_t^{(t)}\;\log\frac{T}{\delta}}. \tag{B.55}$$

and therefore using (B.54) and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$,

$$\sum_{\ell=0}^{t-1} \tilde{\rho}^{\,t-\ell-1} M_{\ell+1}^{\tau} \;\lesssim\; \sqrt{\frac{\sigma^2}{\mu^2}\mathfrak{B}\,\log\frac{2T}{\delta}} + \sqrt{\frac{\sigma^2\gamma^2}{(1-\beta)^2}\mathfrak{D}_t^{\mathrm{mom},(2)}\;\log\frac{2T}{\delta}}. \tag{B.56}$$

Apply Young's inequality $\sqrt{\mathcal{B}}\,u \le \frac{1}{4}\mathcal{B} + u^2$ to the first term with $u := \sqrt{\frac{\sigma^2}{\mu^2}\log\frac{2T}{\delta}}$, and similarly to the second term:

$$\sum_{\ell=0}^{t-1} \rho^{\,t-\ell-1} M_{\ell+1}^{\tau} \;\lesssim\; \mathfrak{B} + \frac{\sigma^2}{\mu^2}\log\frac{2T}{\delta} + \frac{\sigma^2\gamma^2}{(1-\beta)^2}\mathfrak{D}_t^{\mathrm{mom},(2)}\;\log\frac{2T}{\delta}. \tag{B.57}$$

Call this event $\mathcal{E}_M(t)$. A union bound over $t \in [T]$ yields

$$\mathcal{E}_M := \bigcap_{t=1}^{T}\mathcal{E}_M(t) \qquad \text{with} \qquad \mathbb{P}(\mathcal{E}_M) \ge 1 - \delta/3. \tag{B.58}$$

Concluding the bound: Let $\mathcal{E} := \mathcal{E}_\xi \cap \mathcal{E}_M$ with $\mathbb{P}(\mathcal{E}) \ge 1-\delta$. Working on $\mathcal{E}$, apply the stopped recursion (Proposition B.1), we have that

$$\begin{aligned}
\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 \lesssim& \frac{2}{(1-\beta)^2}\tilde{\rho}^T\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{1}{\gamma\mu}\cdot\frac{1}{1-\beta}\mathfrak{D}_t^{\mathrm{mom}} + \frac{d\sigma^2}{\mu^2} \\
&+ \left(\frac{d\sigma^2\gamma^2}{(1-\beta)^2} + \frac{d^2\sigma^4\gamma^3}{\mu(1-\beta)^3}\right)\log(2T/\delta) + \left(\frac{\sigma^2}{\mu^2} + \frac{\sigma^2\gamma^2}{(1-\beta)^2}\mathfrak{D}_t^{\mathrm{mom},(2)}\right)\log\frac{2T}{\delta} \\
\lesssim& \frac{2}{(1-\beta)^2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star\|^2 + \frac{1}{\gamma\mu}\cdot\frac{1}{1-\beta}\max_{t\in[T]}\mathfrak{D}_t^{\mathrm{mom}} + \frac{d\sigma^2}{\mu^2} \\
&+ \left(\frac{d\sigma^2\gamma^2}{(1-\beta)^2} + \frac{d^2\sigma^4\gamma^3}{\mu(1-\beta)^3}\right)\log(2T/\delta) + \left(\frac{\sigma^2}{\mu^2} + \frac{\sigma^2\gamma^2}{(1-\beta)^2}\max_{t\in[T]}\mathfrak{D}_t^{\mathrm{mom},(2)}\right)\log\frac{2T}{\delta}
\end{aligned} \tag{B.59}$$

since $\tilde{\rho} \le 1$ and $\mathfrak{D}_t^{\mathrm{mom}} \le \max_{0\le t\le T-1}\mathfrak{D}_t^{\mathrm{mom}}$. Thus the right-side is less than $\mathfrak{B}$. Hence, on $\mathcal{E}$, we have that $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{t+1}^\star\|^2 \lesssim \mathfrak{B}$ for all $t \in [T]$ which forces $\tau = T+1$ on $\mathcal{E}$ by definition of τ. Hence $M_{\ell+1}^{\tau} = M_{\ell+1}$ for all $\ell \le T-1$ on $\mathcal{E}$, i.e., the stopped bounds coincide with the original process. Taking $t = T$ above yields $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^\star\|^2 = \|e_T\|^2 \le \mathfrak{B}$ on $\mathcal{E}$. Since $\mathbb{P}(\mathcal{E}) \ge 1-\delta$, the claimed high-probability bound follows. We conclude using the inequality $\rho^{2(t+1)} \le \exp(-\gamma\mu(t+1)/(1-\beta))$. □

C Minimax lower bounds for SGDM under nonstationary strongly-convex losses

Before we proceed with obtaining the minimax lower bound, we must first reduce the problem of lower bounding regret to the problem of lower bounding the success probability of testing a sequence of functions. We do this as there are information theoretic tools such as Fano's lemma [CT12] that be used. We first give the definition that measures the "discrepancy" between two functions $f, \widetilde{f} : \Theta \to \mathbb{R}$:

$$\chi(f,\widetilde{f}) := \inf_{\boldsymbol{\theta}\in\Theta}\max\{f(\boldsymbol{\theta}) - f^\star, \widetilde{f}(\boldsymbol{\theta}) - \widetilde{f}^\star\} \quad \text{where} \quad f^\star = \inf_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta}), \quad \widetilde{f}^\star = \inf_{\boldsymbol{\theta}\in\Theta}\widetilde{f}(\boldsymbol{\theta}).$$

This measures characterizes the best regret one can achieve without knowing whether f or $\widetilde{f}$ is the underlying function. This allows us to obtain a reduction from regret minimization to testing problems. This is synonymous to the reduction used by [BGZ15, CWW19].

C.1 From regret minimization to testing

Fix $1 \le p < \infty$, $1 \le q \le \infty$, and $V_T > 0$. Let ϕ^G denote the noisy gradients feedback model. Consider any finite packing $\Theta = \{f^{(1)}, \dots, f^{(M)}\} \subseteq \mathcal{F}_{p,q}(\mathcal{V}_T)$ where each $f^{(i)} = (f_1^{(i)}, \dots, f_T^{(i)})$ is a length-T sequence of convex losses. The following reduction formalizes that *uniformly small regret* over $\mathcal{F}_{p,q}(\mathcal{V}_T)$ implies the existence of a *hypothesis test* that identifies the true sequence $f \in \Theta$ with constant probability. We include a proof for completeness but note that this result is standard in [BGZ15, CWW19].

Lemma C.1 (Reduction from regret to testing)**.** *Let $\Theta \subseteq \mathcal{F}_{p,q}(\mathcal{V}_T)$ be finite, and let $\chi(\cdot,\cdot)$ be a nonnegative per-round separation functional. Suppose there exists an admissible policy π such that*

$$\sup_{f \in \mathcal{F}_{p,q}(\mathcal{V}_T)} \mathcal{R}_T^{\pi,\phi^G}(f) \;\le\; \frac{1}{9} \inf_{\substack{f,\tilde{f} \in \Theta \\ f \neq \tilde{f}}} \sum_{t=1}^T \chi(f_t, \tilde{f}_t). \tag{C.1}$$

Then there exists an estimator (measurable decision rule) mapping the interactions and feedback $\{(\boldsymbol{\theta}_{t-1}, \phi_t^G)\}_{t=1}^T \longmapsto \widehat{f} \in \Theta$ such that

$$\sup_{f \in \Theta} \mathbb{P}_f\left[\widehat{f} \neq f\right] \;\le\; \frac{1}{3}, \tag{C.2}$$

where $\mathbb{P}_f$ denotes the distribution induced by the policy π and the feedback model ϕ^G when the underlying function sequence is f.

Proof of Lemma C.1. Fix any admissible policy π satisfying (C.1). For a realized actions and feedback produced by π, write the (realized) dynamic regret of π on a sequence f as

$$\mathcal{R}_T^\pi(f) \;=\; \sum_{t=1}^T \big(f_t(\boldsymbol{\theta}_{t-1}) - f_t^\star\big), \qquad f_t^\star := \inf_{\boldsymbol{\theta} \in \Theta} f_t(\boldsymbol{\theta}),$$

where $\{\boldsymbol{\theta}_{t-1}\}_{t=1}^T$ are the actions generated by π and the expectation in $\sup_{f \in \mathcal{F}_{p,q}(V_T)} \mathcal{R}_T^\pi(f)$ is over the feedback noise under $\mathbb{P}_f$. Let

$$\Delta \;:=\; \inf_{\substack{f,\tilde{f} \in \Theta \\ f \neq \tilde{f}}} \sum_{t=1}^T \chi(f_t, \tilde{f}_t).$$

By assumption (C.1), for every $f \in \mathcal{F}_{p,q}(V_T)$ we have

$$\mathbb{E}_f\left[\mathcal{R}_T^\pi(f)\right] \;\le\; \frac{1}{9}\Delta.$$

Applying Markov's inequality to the nonnegative random variable $\mathcal{R}_T^\pi(f)$ yields, for every $f \in \Theta$,

$$\mathbb{P}_f\left(\mathcal{R}_T^\pi(f) \;\le\; \frac{1}{3}\Delta\right) \;\ge\; 1 - \frac{\mathbb{E}_f[\mathcal{R}_T^\pi(f)]}{\Delta/3} \;\ge\; 1 - \frac{(1/9)\Delta}{\Delta/3} \;=\; \frac{2}{3}. \tag{C.3}$$

Define the estimator $\widehat{f}$ as the empirical risk minimizer over Θ evaluated on the realized action sequence $\{\boldsymbol{\theta}_{t-1}\}_{t=1}^T$:

$$\widehat{f} \;\in\; \arg\min_{g \in \Theta} \sum_{t=1}^T \big(g_t(\boldsymbol{\theta}_{t-1}) - g_t^\star\big). \tag{C.4}$$

By definition of $\widehat{f}$,

$$\sum_{t=1}^T \big(\widehat{f}_t(\boldsymbol{\theta}_{t-1}) - \widehat{f}_t^\star\big) \;\le\; \sum_{t=1}^T \big(f_t(\boldsymbol{\theta}_{t-1}) - f_t^\star\big) \;=\; \mathcal{R}_T^\pi(f). \tag{C.5}$$

Now condition on the event $\mathcal{E}_f := \left\{\mathcal{R}_T^\pi(f) \le \frac{1}{3}\Delta\right\}$ which has probability at least $2/3$ under $\mathbb{P}_f$ by (C.3). On $\mathcal{E}_f$, we can upper bound the separation between $\widehat{f}$ and f as follows. First, by the definition of χ assumed in the lemma statement (nonnegative per-round separation),

$$\chi(\widehat{f}_t, f_t) \le \inf_{\boldsymbol{\theta}\in\Theta} \max\left\{\widehat{f}_t(\boldsymbol{\theta}) - \widehat{f}_t^\star,\ f_t(\boldsymbol{\theta}) - f_t^\star\right\} \le \max\left\{\widehat{f}_t(\boldsymbol{\theta}_{t-1}) - \widehat{f}_t^\star,\ f_t(\boldsymbol{\theta}_{t-1}) - f_t^\star\right\}, \tag{C.6}$$

and hence, summing over t and using $\max\{a, b\} \le a + b$ for $a, b \ge 0$,

$$\begin{aligned}
\sum_{t=1}^T \chi(\widehat{f}_t, f_t) &\le \sum_{t=1}^T \max\left\{\widehat{f}_t(\boldsymbol{\theta}_{t-1}) - \widehat{f}_t^\star,\ f_t(\boldsymbol{\theta}_{t-1}) - f_t^\star\right\} \\
&\le \sum_{t=1}^T \left(\widehat{f}_t(\boldsymbol{\theta}_{t-1}) - \widehat{f}_t^\star\right) + \sum_{t=1}^T \left(f_t(\boldsymbol{\theta}_{t-1}) - f_t^\star\right) \\
&\le 2\sum_{t=1}^T \left(f_t(\boldsymbol{\theta}_{t-1}) - f_t^\star\right) = 2\mathcal{R}_T^\pi(f),
\end{aligned} \tag{C.7}$$

where in the last step we used (C.5). On $\mathcal{E}_f$, (C.7) gives

$$\sum_{t=1}^T \chi(\widehat{f}_t, f_t) \le 2\mathcal{R}_T^\pi(f) \le \frac{2}{3}\Delta. \tag{C.8}$$

Finally, we show that (C.8) forces $\widehat{f} = f$. Indeed, if $\widehat{f} \ne f$, then by definition of Δ,

$$\sum_{t=1}^T \chi(\widehat{f}_t, f_t) \ge \inf_{\substack{g,\tilde{g}\in\Theta \\ g\ne\tilde{g}}} \sum_{t=1}^T \chi(g_t, \tilde{g}_t) = \Delta,$$

which contradicts (C.8). Therefore, on the event $\mathcal{E}_f$ we must have $\widehat{f} = f$. Combining with (C.3), we conclude

$$\mathbb{P}_f(\widehat{f} \ne f) \le \mathbb{P}_f(\mathcal{E}_f^c) \le \frac{1}{3},$$

and taking the supremum over $f \in \Theta$ proves (C.2). □

Now let $D_{\mathrm{KL}}(P \,||\, Q) = \int \log \frac{dP}{dQ} dP$ denote the Kullback-Leibler divergence between two distributions P and Q. We now introduce the following version of Fano's inequality:

Lemma C.2 (Fano's inequality)**.** *Let $\Theta = \{\theta_1, \ldots, \theta_M\}$ be a finite parameter set with $|\Theta| = M$. For each $\theta \in \Theta$, let P_θ denote the distribution of the observations under parameter θ. Suppose there exists a constant $0 < \beta < \infty$ such that $D_{\mathrm{KL}}(P_\theta \,\|\, P_{\theta'}) \le \beta$ for all $\theta, \theta' \in \Theta$. Then*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{P}_\theta\left[\widehat{\theta} \ne \theta\right] \ge 1 - \frac{\beta + \log 2}{\log M}.$$

With Lemma C.1 and Lemma C.2, obtaining a minimax lower bound on the regret reduces to finding a "hard" subset $\Theta \subseteq \mathcal{F}_{p,q}(\mathcal{V}_T)$ such that we can lower bound $\inf_{f,\tilde{f}\in\Theta} \sum_{t=1}^T \chi(f_t, \tilde{f}_t)$ and upper bound $\sup_{f,\tilde{f}\in\Theta} D_{\mathrm{KL}}(P_f \,\|\, P_{\tilde{f}})$. We proceed by constructing two μ-strongly convex, $O(\mu)$-smooth losses f_+ and f_- whose minimizers are separated by $2a$ along e_1, but whose gradients differ only inside a ball of radius $\asymp a$. This localization makes $\|\nabla f_+ - \nabla f_-\|_{L^p(\Theta)}$ of order $\mu a^{1+d/p}$ so we can switch between f_+ and f_- many times while staying within the gradient-variation budget. We then build a J-block family of sequences $F_{1:T}^u$ indexed by $u \in \{\pm 1\}^J$ where each block uses either f_+ or f_- and $\Delta_T := \lfloor T/J \rfloor$ for $1 \le J \le T$ to be determined later. If an algorithm cannot reliably infer u from the noisy gradients, it must play near the wrong minimizer on a constant fraction of the time indices, costing $\Omega(\mu a^2)$ loss per round. Hence the regret is $\Omega(\mu a^2 T)$ times the testing error. Finally we can upper bound the mutual information using Gaussian KL chain rule, apply Fano's inequality, and choose Δ_T to conclude.

C.2 Constructing hard smooth convex functions with localized gradients

We will proceed with constructing hard μ-strongly convex, $O(\mu)$-smooth losses f_+ and f_-. Fix a smooth bump $\psi : \mathbb{R}^d \to [0,1]$ such that $\psi(\boldsymbol{\theta}) = 1$ for $\|\boldsymbol{\theta}\| \leq 1/2$ and $\psi(\boldsymbol{\theta}) = 0$ for $\|\boldsymbol{\theta}\| \geq 1$ with $\|\nabla\psi(\boldsymbol{\theta})\| \leq C_\psi$ and $\|\nabla^2\psi(\boldsymbol{\theta})\|_{\text{op}} \leq C_\psi$ for all $\boldsymbol{\theta} \in \Theta$. For scale $r > 0$, define $\psi_r(\boldsymbol{\theta}) := \psi(\boldsymbol{\theta}/r)$. Then we have that $\|\nabla\psi_r(\boldsymbol{\theta})\| \leq C_\psi/r$ and $\|\nabla^2\psi_r(\boldsymbol{\theta})\|_{\text{op}} \leq C_\psi/r^2$. Let $e_1 = (1, 0, \ldots, 0)$. Choose parameters $a > 0$ and $r > 0$ with $a \leq r/4$. Define on $\mathbb{R}^d$ the following function:

$$f_u(\boldsymbol{\theta}) := \frac{\mu}{2}\|\boldsymbol{\theta}\|^2 - u \cdot \mu a \langle \boldsymbol{\theta}, e_1 \rangle \psi_r(\boldsymbol{\theta}), \quad u \in \{+1, -1\}. \tag{C.9}$$

We now show that this function is strongly convex and smooth on all of $\mathbb{R}^d$

Lemma C.3 (Smoothness and strong convexity of f_u)**.** *There exists a universal constant $c_1 > 0$ such that if $a/r \leq c_1$, then each f_u is $\mu/4$ strongly convex and 2μ-smooth on all of $\mathbb{R}^d$.*

Proof of Lemma C.3. We can compute the gradient as follows:

$$\nabla_{\boldsymbol{\theta}} f_u(\boldsymbol{\theta}) = \mu\boldsymbol{\theta} - u \cdot \mu a \left(e_1\psi_r(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}, e_1 \rangle \nabla\psi_r(\boldsymbol{\theta})\right). \tag{C.10}$$

Similarly we can compute the Hessian as:

$$\nabla^2 f_u(\boldsymbol{\theta}) = \mu I_d - u \cdot \mu a \left(e_1(\nabla\psi_r(\boldsymbol{\theta}))^\top + (\nabla\psi_r(\boldsymbol{\theta}))e_1^\top + \langle \boldsymbol{\theta}, e_1 \rangle \nabla^2\psi_r(\boldsymbol{\theta})\right). \tag{C.11}$$

Using $\|e_1(\nabla\psi_r(\boldsymbol{\theta}))^\top\| \leq \|\nabla\psi_r(\boldsymbol{\theta})\|$, the derivative bounds for the bump function, and $|\langle \boldsymbol{\theta}, e_1 \rangle| \leq \|\boldsymbol{\theta}\| \leq r$ whenever $\psi_r(\boldsymbol{\theta}) \neq 0$, we get for all $\theta \in \Theta$:

$$\|\nabla^2 f_u(\boldsymbol{\theta}) - \mu I_d\|_{\text{op}} \leq \mu a \left(2\|\nabla\psi_r(\boldsymbol{\theta}\| + |\langle \boldsymbol{\theta}, e_1 \rangle| \, \|\nabla^2\psi_r(\boldsymbol{\theta}\|_{\text{op}})\right) \leq \frac{3\mu a C_\psi}{r}. \tag{C.12}$$

So if $a/r \leq c_1 := 1/(12C_\psi)$, we have $\|\nabla^2 f_u(\boldsymbol{\theta}) - \mu I_d\|_{\text{op}} \leq \mu/4$. Thus we can conclude. □

We now show that with this construction, the discrepancy between these functions is lower bounded by μa^2.

Lemma C.4 (Minimizers and discrepancy of the two-point construction)**.** *Assume $a \leq r/4$ and $a/r \leq c_1$, where $c_1 > 0$ is the constant from Lemma C.3. Let f_u be defined in Equation (C.9) for $u \in \{+1, -1\}$, and assume that Θ contains the line segment $\{tae_1 : t \in [-1, 1]\}$.*

*(i) (**Minimizers.**) Each f_u admits a unique minimizer $\boldsymbol{\theta}_u^\star$, and $\boldsymbol{\theta}_u^\star = uae_1$.*

*(ii) (**Discrepancy.**) The two-point discrepancy satisfies $\chi(f_+, f_-) \geq \mu a^2/8$.*

Proof. On the Euclidean ball $\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq r/2\}$, we have $\psi_r(\boldsymbol{\theta}) = 1$ and $\nabla\psi_r(\boldsymbol{\theta}) = 0$ by construction of the bump. Therefore, for all $\|\boldsymbol{\theta}\|_2 \leq r/2$, $\nabla f_u(\boldsymbol{\theta}) = \mu\boldsymbol{\theta} - u\mu ae_1$. Setting $\nabla f_u(\boldsymbol{\theta}) = 0$ yields $\boldsymbol{\theta} = uae_1$, and this point indeed lies in the region where the above simplification holds since $\|uae_1\|_2 = a \leq r/4 < r/2$. By Lemma C.3, each f_u is $\mu/4$–strongly convex on $\mathbb{R}^d$, hence admitting a unique global minimizer, which must coincide with its unique stationary point. Thus $\boldsymbol{\theta}_u^\star = uae_1$ and this proves the first claim.

To prove the second claim, by $\mu/4$–strong convexity of f_u, for every $\boldsymbol{\theta} \in \Theta$,

$$f_u(\boldsymbol{\theta}) - f_u(\boldsymbol{\theta}_u^\star) \;\geq\; \frac{\mu}{8}\|\boldsymbol{\theta} - \boldsymbol{\theta}_u^\star\|_2^2.$$

Define $d_+(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - ae_1\|_2$ and $d_-(\boldsymbol{\theta}) := \|\boldsymbol{\theta} + ae_1\|_2$. Then

$$\max\{d_+(\boldsymbol{\theta})^2, d_-(\boldsymbol{\theta})^2\} \;\geq\; \frac{1}{2}\big(d_+(\boldsymbol{\theta})^2 + d_-(\boldsymbol{\theta})^2\big).$$

Moreover, by the parallelogram identity,

$$d_+(\boldsymbol{\theta})^2 + d_-(\boldsymbol{\theta})^2 = \|\boldsymbol{\theta} - ae_1\|_2^2 + \|\boldsymbol{\theta} + ae_1\|_2^2 = 2\|\boldsymbol{\theta}\|_2^2 + 2\|ae_1\|_2^2 \geq 2a^2,$$

so $\max\{d_+(\boldsymbol{\theta})^2, d_-(\boldsymbol{\theta})^2\} \geq a^2$. Combining with the strong-convexity lower bound gives, for all $\boldsymbol{\theta} \in \Theta$,

$$\max\{f_+(\boldsymbol{\theta}) - f_+(\boldsymbol{\theta}_+^\star),\ f_-(\boldsymbol{\theta}) - f_-(\boldsymbol{\theta}_-^\star)\} \ \geq\ \frac{\mu}{8}\max\{d_+(\boldsymbol{\theta})^2, d_-(\boldsymbol{\theta})^2\} \ \geq\ \frac{\mu a^2}{8}.$$

Taking $\inf_{\boldsymbol{\theta}\in\Theta}$ of the left-hand side yields $\chi(f_+, f_-) \geq \mu a^2/8$. Finally, since $\boldsymbol{\theta}_u^\star = uae_1 \in \Theta$ by the segment assumption, we have $\min_{\boldsymbol{\theta}\in\Theta} f_u(\boldsymbol{\theta}) = f_u(\boldsymbol{\theta}_u^\star)$ for each u. This proves the second claim. □

Next we show that this localization makes $\|\nabla f_+ - \nabla f_-\|_{L^p(\Theta)}$ of order $\mu a^{1+d/p}$ so we can switch between f_+ and f_- many times while staying within the gradient-variation budget.

Lemma C.5 (Localized gradient difference)**.** *Assume $r \asymp a$ (i.e., $c_- a \leq r \leq c_+ a$ for universal constants $c_-, c_+ > 0$). Then there exists a universal constant $C > 0$ such that*

$$\|\nabla f_+ - \nabla f_-\|_{L^p(\Theta)} \ \leq\ C\,\mu\,a\,r^{d/p} \ \asymp\ C\,\mu\,a^{1+d/p}.$$

Equivalently, writing $\alpha := 1 + d/p$, we have $\|\nabla f_+ - \nabla f_-\|_{L^p(\Theta)} \lesssim \mu a^\alpha$.

Proof of Lemma C.5. From Lemma C.3, we explicitly computed $\nabla_\theta f_u$. Using this, we find

$$\nabla f_+(\boldsymbol{\theta}) - \nabla f_-(\boldsymbol{\theta}) = -2\mu a\Big(e_1\,\psi_r(\boldsymbol{\theta}) + \langle\boldsymbol{\theta}, e_1\rangle\,\nabla\psi_r(\boldsymbol{\theta})\Big).$$

The right-hand side vanishes whenever $\psi_r(\boldsymbol{\theta}) = 0$, hence it is supported on $B(0,r)$. On this support, $|\langle\boldsymbol{\theta}, e_1\rangle| \leq \|\boldsymbol{\theta}\|_2 \leq r$, and by the bump derivative bounds $\|\nabla\psi_r(\boldsymbol{\theta})\|_2 \leq C_\psi/r$. Therefore, for all $\boldsymbol{\theta} \in \Theta$,

$$\|\nabla f_+(\boldsymbol{\theta}) - \nabla f_-(\boldsymbol{\theta})\|_2 \leq 2\mu a\Big(\|e_1\|_2\,|\psi_r(\boldsymbol{\theta})| + |\langle\boldsymbol{\theta}, e_1\rangle|\,\|\nabla\psi_r(\boldsymbol{\theta})\|_2\Big) \leq 2\mu a\Big(1 + r\cdot\frac{C_\psi}{r}\Big) \leq C_0\,\mu a,$$

for a universal constant $C_0 > 0$, and the quantity is 0 outside $B(0,r)$. Hence

$$\|\nabla f_+ - \nabla f_-\|_{L^p(\Theta)} = \Big(\int_\Theta \|\nabla f_+(\boldsymbol{\theta}) - \nabla f_-(\boldsymbol{\theta})\|_2^p\,dx\Big)^{1/p} \leq (C_0\mu a)\,|\Theta \cap B(0,r)|^{1/p} \leq (C_0\mu a)\,|B(0,r)|^{1/p}.$$

Using $|B(0,r)| = c_d r^d$ yields

$$\|\nabla f_+ - \nabla f_-\|_{L^p(\Theta)} \leq (C_0\mu a)\,(c_d r^d)^{1/p} = C\,\mu\,a\,r^{d/p}.$$

If additionally $r \asymp a$, then $a\,r^{d/p} \asymp a^{1+d/p}$, giving the claimed scaling. □

A simple consequence of Lemma C.5 and constructing f_u through a smooth bump function is that we can obtain pointwise localization as stated in the following result:

Lemma C.6 (Pointwise localization)**.** *Let f_u be defined by (C.9). Then for all $\theta \in \Theta$,*

$$\Big\|\nabla f_+(\boldsymbol{\theta}) - \nabla f_-(\boldsymbol{\theta})\Big\|_2^2 \ \leq\ 4C\mu^2 a^2 \mathbf{1}\{\|\boldsymbol{\theta}\|_2 < r\}, \tag{C.13}$$

for some universal $C > 0$.

Proof of Lemma C.6. From Lemma C.5, we found that

$$\nabla f_+(\boldsymbol{\theta}) - \nabla f_-(\boldsymbol{\theta}) = -2\mu a\Big(e_1\,\psi_r(\boldsymbol{\theta}) + \langle\boldsymbol{\theta}, e_1\rangle\,\nabla\psi_r(\boldsymbol{\theta})\Big).$$

The right-hand side vanishes whenever $\psi_r(\boldsymbol{\theta}) = 0$, hence it is supported on $B(0,r)$. On this support, $|\langle\boldsymbol{\theta}, e_1\rangle| \leq \|\boldsymbol{\theta}\|_2 \leq r$, and by the bump derivative bounds $\|\nabla\psi_r(\boldsymbol{\theta})\|_2 \leq C_\psi/r$. Using the triangle inequality, we find

$$\Big\|\nabla f_+(\boldsymbol{\theta}) - \nabla f_-(\boldsymbol{\theta})\Big\|_2^2 \leq 4C\mu^2 a^2 \mathbf{1}\{\|\boldsymbol{\theta}\|_2 < r\}$$

which is exactly (C.13) after absorbing constants. □

C.3 Inducing sharp momentum penalty into the noise-dominated term

We now show that the noise-dominated (information-theoretic) minimax term inherits sharp dependence on the momentum parameter β. The key is to exploit that, under noisy gradient feedback, the KL divergence between two environments is proportional to the cumulative squared mean shift observed along the algorithm's trajectory. For our localized bump construction, this mean shift is supported on a small region. Define $\mathrm{Occ}(r) := \sum_{t=0}^{T-1} \mathbf{1}\{\|\boldsymbol{\theta}_t\|_2 \le r\}$. We will show that by tuning the bump radius into a rare-visit regime whose visitation probability depends on β, the expected occupation time $\mathbb{E}^{\pi}[\mathrm{Occ}(r)]$ scales with $(1-\beta)^{-1}$ which will in turn cause the KL to scales as if the noise variance were inflated from σ^2 to $\sigma^2/(1-\beta)$. This yields the desired $(1-\beta)$–penalty in the statistical (noise-limited) term.

Before we prove this claim, recall our localized construction (C.9). Since $\psi_r(\boldsymbol{\theta}) = 0$ and $\nabla\psi_r(\boldsymbol{\theta}) = 0$ whenever $\|\boldsymbol{\theta}\|_2 \ge r$, we have the identity $\nabla f_u(\boldsymbol{\theta}) = \mu\,\boldsymbol{\theta}$ for all $\|\boldsymbol{\theta}\|_2 \ge r$ and all $u \in \{+1,-1\}$. Thus, outside the bump region, the SGDM iterate evolves according to the same linear recursion (independent of u). This is the mechanism we exploit: the visitation frequency of the bump is controlled by the stationary spread of this linear system, and this spread depends sharply on β. We will focus on Polyak Heavy-Ball but we note that this analysis can be extended to Nesterov acceleration.

We now quantify the fundamental reason the occupation time depends on momentum: in the presence of gradient noise, Heavy-Ball SGDM exhibits a stationary covariance that scales as $(1-\beta)^{-1}$.

Lemma C.7 (Exact stationary variance of Heavy-Ball on a quadratic; $(1-\beta)^{-1}$ inflation)**.** *Fix $\mu > 0$, $\gamma > 0$, $\beta \in [0,1)$, and $\sigma > 0$. Consider the one-dimensional Heavy-Ball recursion on the quadratic $x \mapsto \frac{\mu}{2}x^2$ with additive Gaussian gradient noise:*

$$x_{t+1} = (1+\beta-\gamma\mu)\,x_t - \beta\,x_{t-1} - \gamma\,\zeta_{t+1}, \qquad \zeta_{t+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2), \tag{C.14}$$

for $t \ge 0$, with some deterministic initialization $(x_0, x_{-1}) \in \mathbb{R}^2$. Define the state vector $S_t := (x_t, x_{t-1})^\top \in \mathbb{R}^2$ and the matrices

$$A := \begin{pmatrix} 1+\beta-\gamma\mu & -\beta \\ 1 & 0 \end{pmatrix}, \qquad B := \begin{pmatrix} -\gamma \\ 0 \end{pmatrix}. \tag{C.15}$$

Then (C.14) is equivalently the linear stochastic system

$$S_{t+1} = AS_t + B\zeta_{t+1}. \tag{C.16}$$

Assume the stability condition

$$\rho(A) < 1, \tag{C.17}$$

where $\rho(\cdot)$ denotes spectral radius. Equivalently, the roots of $\lambda^2 - (1+\beta-\gamma\mu)\lambda + \beta = 0$ lie strictly inside the open unit disk.

Then:

(i) The Markov chain $(S_t)_{t\ge0}$ admits a unique stationary distribution π_∞, which is a centered Gaussian $\mathcal{N}(0,\Sigma_\infty)$ on $\mathbb{R}^2$.

(ii) The stationary covariance Σ_∞ is the unique positive semidefinite solution of the discrete-time Lyapunov equation

$$\Sigma = A\Sigma A^\top + \sigma^2 BB^\top. \tag{C.18}$$

(iii) Writing $\Sigma_\infty = \left(\begin{smallmatrix} v & c \\ c & v \end{smallmatrix}\right)$ with $v = \mathrm{Var}_{\pi_\infty}(x_t)$ and $c = \mathrm{Cov}_{\pi_\infty}(x_t, x_{t-1})$, we have the exact closed form

$$v = \frac{(1+\beta)\gamma^2\sigma^2}{(1-\beta)\Big((1+\beta)^2 - (1+\beta-\gamma\mu)^2\Big)} = \frac{(1+\beta)\gamma\sigma^2}{(1-\beta)\mu\big(2(1+\beta)-\gamma\mu\big)}. \tag{C.19}$$

(iv) In the stable regime one necessarily has $\gamma\mu < 2(1+\beta)$, so the denominator in (C.19) is positive. Moreover, if additionally $\gamma\mu \le 1+\beta$ (a common small-step regime), then

$$v \;\ge\; \frac{\gamma\sigma^2}{4\mu(1-\beta)}. \tag{C.20}$$

Proof. We proceed in four steps: (1) existence/uniqueness of a stationary law, (2) characterization via a Lyapunov equation, (3) explicit solution of the $(1, 1)$ entry, and (4) extraction of the $(1 - \beta)^{-1}$ lower bound.

Step 1: (S_t) is an affine Gaussian Markov chain; existence and uniqueness of a stationary law under $\rho(A) < 1$. Fix any deterministic $S_0 \in \mathbb{R}^2$. Iterating (C.16) gives the explicit representation

$$S_t = A^t S_0 + \sum_{k=1}^{t} A^{t-k} B\zeta_k, \qquad t \geq 1. \tag{C.21}$$

Since $(\zeta_k)_{k\geq 1}$ are independent Gaussians and (C.21) is an affine map of the vector $(\zeta_1, \dots, \zeta_t)$, it follows that each S_t is Gaussian in $\mathbb{R}^2$. Moreover, $\mathbb{E}[\zeta_k] = 0$ yields

$$\mathbb{E}[S_t] = A^t S_0. \tag{C.22}$$

Under (C.17), there exists a matrix norm $\|\cdot\|$ and constants $C \geq 1$, $\rho \in (0, 1)$ such that $\|A^t\| \leq C\rho^t$ for all $t \geq 0$ (standard consequence of $\rho(A) < 1$). Hence $\mathbb{E}[S_t] \to 0$ as $t \to \infty$. Next compute the covariance. Let $\Sigma_t := \mathrm{Cov}(S_t)$. From (C.16), using independence of ζ_{t+1} from $\sigma(S_0, \zeta_1, \dots, \zeta_t)$, we have

$$\Sigma_{t+1} = A\Sigma_t A^\top + \sigma^2 BB^\top, \qquad t \geq 0, \tag{C.23}$$

with $\Sigma_0 = 0$ if S_0 is deterministic. Iterating (C.23) yields the series representation

$$\Sigma_t = \sum_{j=0}^{t-1} A^j \,(\sigma^2 BB^\top)\,(A^j)^\top \;+\; A^t \Sigma_0 (A^t)^\top. \tag{C.24}$$

The second term vanishes as $t \to \infty$ because $\|A^t\| \to 0$. The first term converges absolutely in operator norm: indeed, using $\|A^j\| \leq C\rho^j$ and $\|B\| < \infty$,

$$\left\|A^j \,(\sigma^2 BB^\top)\,(A^j)^\top\right\| \leq \sigma^2 \|A^j\|^2 \|B\|^2 \leq \sigma^2 C^2 \rho^{2j} \|B\|^2,$$

and $\sum_{j\geq 0} \rho^{2j} < \infty$. Therefore $\Sigma_t \to \Sigma_\infty$ where

$$\Sigma_\infty = \sum_{j=0}^{\infty} A^j \,(\sigma^2 BB^\top)\,(A^j)^\top. \tag{C.25}$$

Define $\pi_\infty := \mathcal{N}(0, \Sigma_\infty)$. From (C.21)–(C.25), the law of S_t converges weakly to π_∞, and π_∞ is stationary since substituting (C.25) into the covariance recursion yields (C.18) (verified in Step 2).

For uniqueness, if π is any stationary distribution with finite second moment, then letting $\Sigma_\pi := \mathrm{Cov}_\pi(S)$, stationarity and (C.16) imply Σ_π solves (C.18). Under $\rho(A) < 1$, the Lyapunov equation (C.18) has a unique positive semidefinite solution (see Step 2), hence $\Sigma_\pi = \Sigma_\infty$; since the chain is Gaussian and linear, this identifies π as $\mathcal{N}(0, \Sigma_\infty)$. (Equivalently, any stationary law must be Gaussian because the transition kernel is Gaussian and linear, and the unique second moment determines it.)

Step 2: Lyapunov characterization and uniqueness of the stationary covariance. We now justify (ii) rigorously. Let Σ be any solution to (C.18). Vectorize: $\mathrm{vec}(A\Sigma A^\top) = (A \otimes A)\mathrm{vec}(\Sigma)$, so (C.18) becomes

$$\big(I - A \otimes A\big)\mathrm{vec}(\Sigma) = \sigma^2 \mathrm{vec}(BB^\top). \tag{C.26}$$

Under $\rho(A) < 1$, we have $\rho(A \otimes A) = \rho(A)^2 < 1$, hence $I - A \otimes A$ is invertible and (C.26) admits a unique solution. This proves uniqueness of Σ_∞. Moreover, the Neumann series yields

$$\mathrm{vec}(\Sigma_\infty) = \sum_{j=0}^{\infty} (A \otimes A)^j \sigma^2 \mathrm{vec}(BB^\top) = \sigma^2 \sum_{j=0}^{\infty} \mathrm{vec}\big(A^j BB^\top (A^j)^\top\big),$$

which is exactly (C.25) upon unvectorizing. Hence (C.25) is the unique PSD solution of (C.18).

Step 3: explicit computation of $v = \mathrm{Var}_{\pi_\infty}(x_t)$ from (C.18). Write

$$\Sigma_\infty = \begin{pmatrix} v & c \\ c & w \end{pmatrix}, \qquad v = \mathbb{E}_{\pi_\infty}[x_t^2], \quad c = \mathbb{E}_{\pi_\infty}[x_t x_{t-1}], \quad w = \mathbb{E}_{\pi_\infty}[x_{t-1}^2].$$

Stationarity of (S_t) implies x_t and x_{t-1} have the same marginal law under π_∞, hence $w = v$. Thus,

$$\Sigma_\infty = \begin{pmatrix} v & c \\ c & v \end{pmatrix}. \tag{C.27}$$

Next note from (C.15) that $\sigma^2 BB^\top = \sigma^2 \left(\begin{smallmatrix} \gamma^2 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Plugging (C.27) into (C.18), we compute $A\Sigma_\infty A^\top$ explicitly.

Let $a := 1 + \beta - \gamma\mu$ so that $A = \left(\begin{smallmatrix} a & -\beta \\ 1 & 0 \end{smallmatrix}\right)$. First,

$$A\Sigma_\infty = \begin{pmatrix} a & -\beta \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v & c \\ c & v \end{pmatrix} = \begin{pmatrix} av - \beta c & ac - \beta v \\ v & c \end{pmatrix}.$$

Then

$$A\Sigma_\infty A^\top = \begin{pmatrix} av - \beta c & ac - \beta v \\ v & c \end{pmatrix} \begin{pmatrix} a & 1 \\ -\beta & 0 \end{pmatrix} = \begin{pmatrix} a^2 v - 2a\beta c + \beta^2 v & av - \beta c \\ av - \beta c & v \end{pmatrix}.$$

Now impose (C.18):

$$\begin{pmatrix} v & c \\ c & v \end{pmatrix} = \begin{pmatrix} a^2 v - 2a\beta c + \beta^2 v & av - \beta c \\ av - \beta c & v \end{pmatrix} + \sigma^2 \begin{pmatrix} \gamma^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

Equating the $(1, 2)$ entries gives

$$c = av - \beta c \quad \Longleftrightarrow \quad (1 + \beta)c = av \quad \Longleftrightarrow \quad c = \frac{a}{1 + \beta} v. \tag{C.28}$$

Equating the $(1, 1)$ entries gives

$$v = (a^2 + \beta^2)v - 2a\beta c + \gamma^2\sigma^2. \tag{C.29}$$

Substitute (C.28) into (C.29):

$$v = (a^2 + \beta^2)v - 2a\beta \left(\frac{a}{1 + \beta} v\right) + \gamma^2\sigma^2 = \left(a^2 + \beta^2 - \frac{2a^2\beta}{1 + \beta}\right) v + \gamma^2\sigma^2.$$

Rearrange this as:

$$\left(1 - a^2 - \beta^2 + \frac{2a^2\beta}{1 + \beta}\right) v = \gamma^2\sigma^2.$$

Solving for v yields

$$v = \frac{(1 + \beta)\gamma^2\sigma^2}{(1 - \beta)\big((1 + \beta)^2 - a^2\big)}.$$

Recalling $a = 1 + \beta - \gamma\mu$ gives (C.19), and the alternative expression follows from

$$(1 + \beta)^2 - (1 + \beta - \gamma\mu)^2 = \gamma\mu\big(2(1 + \beta) - \gamma\mu\big).$$

Step 4: positivity and the $(1 - \beta)^{-1}$ lower bound. We first note that the denominator in (C.19) is strictly positive under stability. Indeed, stability implies in particular that $|a| < 1 + \beta$ for the AR(2) polynomial (equivalently, the roots lie in the unit disk), hence $(1 + \beta)^2 - a^2 > 0$. More explicitly,

$$(1 + \beta)^2 - a^2 = \gamma\mu\big(2(1 + \beta) - \gamma\mu\big),$$

so positivity is equivalent to $\gamma\mu < 2(1 + \beta)$, which is necessary for stability. Now assume $\gamma\mu \le 1 + \beta$. Then $2(1 + \beta) - \gamma\mu \le 2(1 + \beta)$ and also $1 + \beta \le 2$. Using the second form in (C.19),

$$v = \frac{(1 + \beta)\gamma\sigma^2}{(1 - \beta)\mu\big(2(1 + \beta) - \gamma\mu\big)} \ \ge\ \frac{(1 + \beta)\gamma\sigma^2}{(1 - \beta)\mu \cdot 2(1 + \beta)} = \frac{\gamma\sigma^2}{4\mu(1 - \beta)}.$$

This completes the proof. □

Lemma C.7 implies that for sufficiently small r, the process spends only an $O(1-\beta)$ fraction of time inside the bump. This is the mechanism that injects momentum dependence into the KL. We track the scalar process $x_t := \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_u^\star, e_1 \rangle$ because the bump perturbation ψ (and hence the discrepancy between $u = +1$ and $u = -1$) acts only along the direction e_1. After centering by $\boldsymbol{\theta}_u^\star$, this coordinate evolves as a one-dimensional Heavy–Ball AR(2) recursion, whose stationary variance satisfies $\mathrm{Var}(x_\infty) \asymp \gamma\sigma^2/\mu(1-\beta)$. This enables sharp β-dependent small-ball and occupation bounds for the bump region.

Lemma C.8 (Beta-dependent occupation bound in the rare-visit regime)**.** *Fix $T \in \mathbb{N}$ and $r > 0$. Let $(\boldsymbol{\theta}_t)_{t=0}^{T-1}$ be any $\mathbb{R}^d$-valued process and let $e_1 \in \mathbb{R}^d$ denote the first standard basis vector. Define the scalar projection*

$$x_t := \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_u^\star, e_1 \rangle, \qquad t = 0, \dots, T-1,$$

for some (fixed) center $\boldsymbol{\theta}_u^\star \in \mathbb{R}^d$, and define the (radius-r) occupation count

$$\mathrm{Occ}_T(r) := \sum_{t=0}^{T-1} \mathbf{1}\{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_u^\star\|_2 \le r\}.$$

Assume that for each $t = 0, \dots, T-1$, the random variable x_t is Gaussian with mean $m_t \in \mathbb{R}$ and variance $v_t := \mathrm{Var}(x_t)$ satisfying the uniform lower bound

$$v_t \;\ge\; v_{\min} \;>\; 0, \qquad \textit{for all } t = 0, \dots, T-1. \tag{C.30}$$

Then there exists a universal constant $C_1 > 0$ such that

$$\mathbb{E}[\mathrm{Occ}_T(r)] \;\le\; C_1\, T \cdot \frac{r}{\sqrt{v_{\min}}}. \tag{C.31}$$

In particular, if x_t follows a stable Heavy-Ball recursion on the quadratic core so that

$$v_{\min} \;\ge\; \frac{\gamma\sigma^2}{4\mu(1-\beta)} \tag{C.32}$$

(as in (C.20)), then for a (possibly different) universal constant $C_1 > 0$,

$$\mathbb{E}[\mathrm{Occ}_T(r)] \;\le\; C_1\, T \cdot r \sqrt{\frac{\mu(1-\beta)}{\gamma\sigma^2}}. \tag{C.33}$$

Consequently, choosing the bump radius in the rare-visit regime

$$r \;:=\; c_r\, \sigma \sqrt{\frac{\gamma(1-\beta)}{\mu}}, \tag{C.34}$$

for a sufficiently small universal constant $c_r > 0$, yields

$$\mathbb{E}[\mathrm{Occ}_T(r)] \;\le\; C_2\, T(1-\beta), \tag{C.35}$$

for a universal constant $C_2 > 0$.

Proof. We prove (C.31) first. (C.33) and (C.35) then follow by substitution. For any t, the inclusion

$$\{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_u^\star\|_2 \le r\} \;\subseteq\; \{|\langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_u^\star, e_1 \rangle| \le r\} = \{|x_t| \le r\} \tag{C.36}$$

holds deterministically, because $|\langle v, e_1 \rangle| \le \|v\|_2$ for all $v \in \mathbb{R}^d$. Therefore, taking indicators and summing over t gives the pointwise domination

$$\mathrm{Occ}_T(r) = \sum_{t=0}^{T-1} \mathbf{1}\{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_u^\star\|_2 \le r\} \;\le\; \sum_{t=0}^{T-1} \mathbf{1}\{|x_t| \le r\}. \tag{C.37}$$

Taking expectations yields

$$\mathbb{E}[\mathrm{Occ}_T(r)] \;\leq\; \sum_{t=0}^{T-1} \mathbb{P}(|x_t| \leq r). \tag{C.38}$$

Fix $t \in \{0, \ldots, T-1\}$. By assumption, $x_t \sim \mathcal{N}(m_t, v_t)$ with $v_t \geq v_{\min}$ (this is also true for our case by Lemma C.7). Using Lemma D.7 with (C.38):

$$\mathbb{E}[\mathrm{Occ}_T(r)] \leq \sum_{t=0}^{T-1} \sqrt{\frac{2}{\pi}}\, \frac{r}{\sqrt{v_{\min}}} = T\sqrt{\frac{2}{\pi}}\, \frac{r}{\sqrt{v_{\min}}}.$$

This is (C.31) with $C_1 = \sqrt{2/\pi}$. If additionally (C.32) holds, then

$$\frac{1}{\sqrt{v_{\min}}} \leq \sqrt{\frac{4\mu(1-\beta)}{\gamma\sigma^2}} \lesssim \sqrt{\frac{\mu(1-\beta)}{\gamma\sigma^2}},$$

and substituting into (C.31) yields (C.33) (after absorbing the numerical factor 2 into the universal constant C_1). Finally, substitute the rare-visit choice (C.34) into (C.33):

$$\mathbb{E}[\mathrm{Occ}_T(r)] \;\leq\; C_1\, T \cdot c_r \sigma \sqrt{\frac{\gamma(1-\beta)}{\mu}} \cdot \sqrt{\frac{\mu(1-\beta)}{\gamma\sigma^2}} = (C_1 c_r)\, T(1-\beta).$$

Setting $C_2 := C_1 c_r$ proves (C.35). □

C.4 Constructing a J-block hard family

We now build a J-block packing family of nonstationary sequences $F^u_{1:T}$ indexed by sign vectors $u \in \{\pm 1\}^{\Delta_T}$, where each block uses one of the two base losses $\{f_+, f_-\}$. Fix an integer $1 \leq J \leq T$ (to be tuned later) and set $\Delta_T := \lfloor T/J \rfloor$. Partition the horizon $[T] := \{1, \ldots, T\}$ into J disjoint consecutive batches (blocks) $B_1, \ldots, B_J$ such that each B_j has cardinality either Δ_T or $\Delta_T + 1$ and $\bigcup_{j=1}^{J} B_j = [T]$. For each block $j \in [J]$, let $|B_j|$ denote its length and write its indices as

$$B_j = \{t_j(1), t_j(2), \ldots, t_j(|B_j|)\}, \qquad t_j(1) < \cdots < t_j(|B_j|).$$

Let $\{\pm 1\}^J$ be the class of sign vectors of length J, and let $\mathcal{U} \subset \{\pm 1\}^J$. For any $u = (u_1, \ldots, u_J) \in \mathcal{U}$, define the nonstationary loss sequence $F^u_{1:T}$ by holding the loss fixed within each block,

$$F^u_t(\cdot) \;:=\; f_{u_j}(\cdot) \qquad \text{for all } t \in B_j,\; j \in [J]. \tag{C.39}$$

Thus F^u can change only at block boundaries while encoding one bit per block. To apply Fano's method we require a large subset $\mathcal{U} \subset \{\pm 1\}^J$ whose elements are well separated in Hamming distance. We obtain such a set via a constant-weight code construction. This is the same construction used by [CWW19] and originates from [GS80] and [WS16] gave an explicit lower bound which we cite below. For simplicity assume J is even (the odd case follows by restricting to $J-1$ coordinates).

Lemma C.9 ([WS16], Lemma 9)**.** *Suppose $J \geq 2$ is even. There exists a subset $\mathcal{I} \subset \{0,1\}^J$ such that (i) every $i \in \mathcal{I}$ has exactly $J/2$ ones, i.e. $\sum_{j=1}^{J} i_j = J/2$; and (ii) every pair $i \neq i' \in \mathcal{I}$ satisfies $\Delta_H(i, i') \geq J/16$, where Δ_H is Hamming distance. Moreover,* $\log|\mathcal{I}| \geq 0.0625\, J$.

By Lemma C.9, any two indices $u \neq v$ disagree on at least $J/8$ blocks; equivalently, the sequences F^u and F^v differ on at least $\sum_{j:\, u_j \neq v_j} |B_j| \geq (J/8)\Delta_T \gtrsim T$ time indices, which is precisely the separation needed to convert a constant testing error (via Fano) into an $\Omega(T)$ regret gap. We next bound the gradient-variation functional $\mathrm{GVar}_{p,q}$ for the J-block construction.

Lemma C.10 ($\mathrm{GVar}_{p,q}$ bound for the J-block construction)**.** *Let $1 \leq J \leq T$ and set $\Delta_T := \lfloor T/J \rfloor$. Consider the J disjoint consecutive blocks $B_1, \ldots, B_J$ with $|B_j| \in \{\Delta_T, \Delta_T + 1\}$ and the blockwise-constant sequence $F^u_{1:T}$ defined by (C.39) for some $u \in \{\pm 1\}^J$. Then, for every such u,*

$$\mathrm{GVar}_{p,q}(F^u_{1:T}) \;\leq\; C\, \mu\, a^{\alpha}\, J^{1/q}, \qquad \alpha := 1 + \frac{d}{p}, \tag{C.40}$$

where $C > 0$ is a universal constant (depending only on the bump construction in ψ_r through constants such as C_ψ in Lemma C.5).

Proof. By construction, F_t^u is constant on each block B_j, hence the gradient can change only at a boundary between consecutive blocks. More precisely, if t is a boundary time with $t \in B_j$ and $t+1 \in B_{j+1}$, then by (C.39),

$$\nabla F_{t+1}^u - \nabla F_t^u \;=\; \nabla f_{u_{j+1}} - \nabla f_{u_j}.$$

If $u_{j+1} = u_j$, the right-hand side is 0. If $u_{j+1} \neq u_j$, then $\nabla f_{u_{j+1}} - \nabla f_{u_j} = \pm(\nabla f_+ - \nabla f_-)$ and therefore, by Lemma C.5,

$$\left\|\nabla F_{t+1}^u - \nabla F_t^u\right\|_{L^p(\Theta)} \;\leq\; \|\nabla f_+ - \nabla f_-\|_{L^p(\Theta)} \;\leq\; C\,\mu\,a^\alpha.$$

There are at most $J-1$ such boundaries, and all other times contribute zero. Hence, for $1 \leq q < \infty$,

$$\mathrm{GVar}_{p,q}(F_{1:T}^u) = \left(\sum_{t=1}^{T-1} \left\|\nabla F_{t+1}^u - \nabla F_t^u\right\|_{L^p(\Theta)}^q\right)^{1/q} \;\leq\; \Big((J-1)\,(C\mu a^\alpha)^q\Big)^{1/q} \;\leq\; C\,\mu\,a^\alpha\,J^{1/q},$$

after adjusting C by a universal factor. The case $q = \infty$ is analogous and yields $\mathrm{GVar}_{p,\infty}(F_{1:T}^u) \leq C\,\mu\,a^\alpha$ since the maximum increment occurs at a boundary. ◻

We now show that if the algorithm cannot identity u, it must incur regret $\gtrsim \mu a^2$. This proof is rather elementary and follows simply from the definition of the discrepancy measure and Lemma C.4. We include it for completeness.

Lemma C.11 (Per-round separation)**.** *Let $f_+, f_- : \Theta \to \mathbb{R}$ be the two base losses, with minimizers $\boldsymbol{\theta}_+^\star \in \arg\min_\Theta f_+$ and $\boldsymbol{\theta}_-^\star \in \arg\min_\Theta f_-$. Define the discrepancy*

$$\chi(f_+, f_-) \;:=\; \inf_{\boldsymbol{\theta}\in\Theta} \max\left\{f_+(\boldsymbol{\theta}) - f_+(\boldsymbol{\theta}_+^\star),\; f_-(\boldsymbol{\theta}) - f_-(\boldsymbol{\theta}_-^\star)\right\}.$$

Then for every $\boldsymbol{\theta} \in \Theta$,

$$\big(f_+(\boldsymbol{\theta}) - f_+(\boldsymbol{\theta}_+^\star)\big) + \big(f_-(\boldsymbol{\theta}) - f_-(\boldsymbol{\theta}_-^\star)\big) \;\geq\; \chi(f_+, f_-). \tag{C.41}$$

In particular, under the conditions of Lemma C.4, we have

$$\big(f_+(\boldsymbol{\theta}) - f_+(\boldsymbol{\theta}_+^\star)\big) + \big(f_-(\boldsymbol{\theta}) - f_-(\boldsymbol{\theta}_-^\star)\big) \;\geq\; \chi(f_+, f_-) \;\geq\; \frac{\mu a^2}{8} \qquad \forall\,\boldsymbol{\theta} \in \Theta. \tag{C.42}$$

Proof. Fix any $\boldsymbol{\theta} \in \Theta$. By the definition of $\chi(f_+, f_-)$,

$$\max\left\{f_+(\boldsymbol{\theta}) - f_+(\boldsymbol{\theta}_+^\star),\; f_-(\boldsymbol{\theta}) - f_-(\boldsymbol{\theta}_-^\star)\right\} \;\geq\; \chi(f_+, f_-).$$

Using the elementary inequality $A + B \geq \max\{A, B\}$ for all real numbers A, B, we obtain (C.41). The strengthened bound (C.42) follows immediately by combining (C.41) with Lemma C.4, which gives $\chi(f_+, f_-) \geq \mu a^2/8$. ◻

This immediately gives us a lower bound on the quantity $\inf_{f,\tilde{f}\in\Theta} \sum_{t=1}^T \chi(f_t, \tilde{f}_t)$

Corollary C.1 (Blockwise accumulation of discrepancy)**.** *Fix $1 \leq J \leq T$ and set $\Delta_T := \lfloor T/J \rfloor$. Let $B_1, \ldots, B_J$ be a partition of $[T]$ into consecutive blocks with $|B_j| \in \{\Delta_T, \Delta_T + 1\}$. For any $u, v \in \{\pm 1\}^J$, define the blockwise-constant sequences $F_{1:T}^u$ and $F_{1:T}^v$ via (C.39), i.e.,*

$$F_t^u \equiv f_{u_j} \quad \textit{and} \quad F_t^v \equiv f_{v_j} \qquad \textit{for all } t \in B_j,\; j \in [J].$$

Assume the per-round separation bound holds:

$$\chi(f_+, f_-) \;\geq\; \frac{\mu a^2}{8}. \tag{C.43}$$

Then for any $u, v \in \{\pm 1\}^J$,

$$\sum_{t=1}^{T} \chi(F_t^u, F_t^v) \;\ge\; \frac{\mu a^2}{8} \sum_{j=1}^{J} |B_j| \, \mathbf{1}\{u_j \ne v_j\} \;\ge\; \frac{\mu a^2}{8} \Delta_T \, d_H(u, v), \tag{C.44}$$

where $d_H(u, v) := \sum_{j=1}^{J} \mathbf{1}\{u_j \ne v_j\}$ is the Hamming distance. In particular, if $\mathcal{U} \subset \{\pm 1\}^J$ satisfies $d_H(u, v) \ge J/8$ for all distinct $u, v \in \mathcal{U}$, then for all $u \ne v$ in $\mathcal{U}$,

$$\sum_{t=1}^{T} \chi(F_t^u, F_t^v) \;\ge\; c_0 \, \mu a^2 T, \tag{C.45}$$

for a universal constant $c_0 > 0$.

Proof of Corollary C.1. Fix two distinct codewords $u, v \in \mathcal{U}$ and recall that the horizon is partitioned into disjoint blocks $B_1, \ldots, B_J$, each of size $|B_j| \in \{\Delta_T, \Delta_T + 1\}$ with $\Delta_T := \lfloor T/J \rfloor$. Fix an arbitrary block B_j and (w.l.o.g.) assume $|B_j| = \Delta_T$; the blocks of length $\Delta_T + 1$ can be absorbed into constants since T and J are large.

If $u_j = v_j$, then $F_t^u \equiv F_t^v$ throughout B_j, hence $\chi(F_t^u, F_t^v) = 0$ for all $t \in B_j$. If instead $u_j \ne v_j$, then along the entire block we have $(F_t^u, F_t^v) \in \{(f_+, f_-), (f_-, f_+)\}$, and therefore by Lemma C.11,

$$\chi(F_t^u, F_t^v) \;=\; \chi(f_{u_j}, f_{v_j}) \;\ge\; \chi(f_+, f_-) \;\ge\; \frac{\mu a^2}{8} \qquad \text{for all } t \in B_j.$$

Summing over the block yields the blockwise contribution

$$\sum_{t \in B_j} \chi(F_t^u, F_t^v) \;\ge\; m \cdot \frac{\mu a^2}{8} \qquad \text{whenever } u_j \ne v_j.$$

Consequently, summing over all mismatched blocks gives

$$\sum_{t=1}^{T} \chi(F_t^u, F_t^v) \;=\; \sum_{j:\, u_j \ne v_j} \; \sum_{t \in B_j} \chi(F_t^u, F_t^v) \;\ge\; m \cdot \Delta_H(u, v) \cdot \frac{\mu a^2}{8},$$

where $\Delta_H(u, v) := \big|\{j \in [J] : u_j \ne v_j\}\big|$ is the Hamming distance. Finally, by the Varshamov–Gilbert packing property of $\mathcal{U}$ (Lemma C.9), any distinct $u \ne v$ satisfy $d_H(u, v) \ge K/8$, and since $m = \lfloor T/K \rfloor$ we obtain

$$\sum_{t=1}^{T} \chi(F_t^u, F_t^v) \;\ge\; \frac{J}{8} \cdot \left\lfloor \frac{T}{J} \right\rfloor \cdot \frac{\mu a^2}{8} \;\ge\; c_0 \, \mu a^2 T,$$

for a universal constant $c_0 > 0$ (absorbing the floor effect into constants). In words, two sequences that disagree on a constant fraction of blocks incur a constant per-round discrepancy on each such block, and this discrepancy accumulates over $\Theta(T)$ rounds. □

It remains to upper bound $\sup_{f, \tilde{f} \in \Theta} D_{\mathrm{KL}}(P_f \,\|\, P_{\tilde{f}})$. Using the noisy gradient feedback model, this follows simply from a Gaussian KL chain rule.

Lemma C.12 (KL control under Gaussian gradient noise)**.** *Fix any admissible policy π and two environments $u, v \in \mathcal{U}$. Let P_u^π and P_v^π denote the joint laws of $Z_{1:T} := (Y_1, \ldots, Y_T)$ generated under environments u and v, respectively, when the learner follows π. Assume the (conditional) Gaussian gradient feedback model*

$$Y_t \;=\; \nabla F_t(\boldsymbol{\theta}_{t-1}) + \varepsilon_t, \qquad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_d),$$

where $\boldsymbol{\theta}_{t-1}$ is $\sigma(Y_{1:t-1})$–measurable under π. Then the KL divergence satisfies the chain-rule identity

$$D_{\mathrm{KL}}(P_u^\pi \,\|\, P_v^\pi) \;=\; \frac{1}{2\sigma^2} \sum_{t=1}^{T} \mathbb{E}_u\!\left[\left\|\nabla F_t^u(\boldsymbol{\theta}_{t-1}) - \nabla F_t^v(\boldsymbol{\theta}_{t-1})\right\|_2^2\right]. \tag{C.46}$$

Moreover, for the block construction with base losses $\{f_+, f_-\}$, we have the uniform bound

$$D_{\mathrm{KL}}(P_u^\pi \,\|\, P_v^\pi) \;\le\; C\,\frac{\mu^2 a^2}{\sigma^2}\,T, \tag{C.47}$$

for a universal constant $C > 0$ (using $\sup_{\theta\in\Theta} \|\nabla f_+(\boldsymbol{\theta}) - \nabla f_-(\boldsymbol{\theta})\|_2 \le C'\mu a$ from Lemma C.5).

Proof of Lemma C.12. We compute $D_{\mathrm{KL}}(P_u^\pi \| P_v^\pi)$ by conditioning on the learner's past. Let $\mathcal{F}_{t-1} := \sigma(Y_1, \ldots, Y_{t-1})$ denote the natural filtration. By admissibility of π, the iterate $\boldsymbol{\theta}_{t-1}$ is $\mathcal{F}_{t-1}$–measurable, and under environment $w \in \{u, v\}$, the conditional law of Y_t given $\mathcal{F}_{t-1}$ is Gaussian:

$$Y_t \,\big|\, \mathcal{F}_{t-1} \;\sim\; \mathcal{N}\big(\nabla F_t^w(\boldsymbol{\theta}_{t-1}),\, \sigma^2 I_d\big).$$

Using the chain rule for KL divergence,

$$D_{\mathrm{KL}}(P_u^\pi \| P_v^\pi) = \sum_{t=1}^{T} \mathbb{E}_u\left[D_{\mathrm{KL}}\big(P_u^\pi(Y_t \mid \mathcal{F}_{t-1}) \,\big\|\, P_v^\pi(Y_t \mid \mathcal{F}_{t-1})\big)\right]. \tag{C.48}$$

Since the two conditional distributions in (C.48) are Gaussians with the same covariance $\sigma^2 I_d$, their conditional KL is

$$D_{\mathrm{KL}}\Big(\mathcal{N}(m_1, \sigma^2 I_d) \,\big\|\, \mathcal{N}(m_2, \sigma^2 I_d)\Big) = \frac{1}{2\sigma^2}\|m_1 - m_2\|_2^2.$$

Substituting $m_1 = \nabla F_t^u(\boldsymbol{\theta}_{t-1})$ and $m_2 = \nabla F_t^v(\boldsymbol{\theta}_{t-1})$ into (C.48) yields (C.46). By Lemma C.6, we have that $\big\|\nabla f_+(\boldsymbol{\theta}) - \nabla f_-(\boldsymbol{\theta})\big\|_2^2 \;\le\; 4C\mu^2 a^2 \mathbf{1}\{\|\boldsymbol{\theta}\|_2 < r\}$. This results in

$$D_{\mathrm{KL}}(P_u^\pi \,\|\, P_v^\pi) \le \frac{2C\mu^2 a^2}{\sigma^2} \sum_{t=1}^{T} \mathbb{E}_u[\mathbf{1}\{\|\boldsymbol{\theta}_{t-1}\|_2 < r\}] = \frac{2C\mu^2 a^2}{\sigma^2}\mathbb{E}_u[\mathrm{Occ}_T(r)].$$

Combining this with Lemma C.8, we conclude that

$$D_{\mathrm{KL}}(P_u^\pi \,\|\, P_v^\pi) \le \frac{2C'\mu^2 a^2}{\sigma^2/(1-\beta)}T.$$

□

C.5 Obtaining a minimax lower bound in the statistical/variation-limited regime

We can get a bound on $\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{P}_\theta\left[\widehat{\theta} \neq \theta\right]$ via Fano's inequality:

Corollary C.2 (Fano lower bound via KL control). *Fix any admissible policy π and let $\mathcal{U} \subset \{\pm 1\}^J$ be a packing set with*

$$\log|\mathcal{U}| \;\ge\; \frac{J}{8}\log 2. \tag{C.49}$$

Let U be uniformly distributed on $\mathcal{U}$, and let $Z_{1:T} := (Y_1, \ldots, Y_T)$ denote the transcript generated under environment U when the learner follows policy π; write P_u^π for the law of $Z_{1:T}$ under environment $u \in \mathcal{U}$. Then for any estimator $\widehat{U} = \widehat{U}(Z_{1:T})$,

$$\mathbb{P}(\widehat{U} \neq U) \;\ge\; 1 - \frac{I(U; Z_{1:T}) + \log 2}{\log|\mathcal{U}|}. \tag{C.50}$$

In particular, if the Gaussian-gradient KL bound of Lemma C.12 holds so that

$$\sup_{u,v\in\mathcal{U}} D_{\mathrm{KL}}(P_u^\pi \,\|\, P_v^\pi) \;\le\; C\,\frac{\mu^2 a^2}{\sigma^2/(1-\beta)}\,T, \tag{C.51}$$

then $I(U; Z_{1:T}) \le C\,\mu^2 a^2 T/(\sigma^2/(1-\beta))$. Consequently, if

$$C\,\frac{\mu^2 a^2}{\sigma^2/(1-\beta)}\,T \;\le\; \frac{1}{16}\,J\log 2, \tag{C.52}$$

then every estimator $\widehat{U}$ obeys the constant error lower bound

$$\mathbb{P}(\widehat{U} \neq U) \;\geq\; \frac{1}{2}. \tag{C.53}$$

Proof of Corollary C.2. Combining Lemma D.8 with (C.51) yields

$$I(U; Z_{1:T}) \;\leq\; \frac{1}{|\mathcal{U}|^2} \sum_{u,v \in \mathcal{U}} C \, \frac{\mu^2 a^2}{\sigma^2/(1-\beta)} T \;=\; C \, \frac{\mu^2 a^2}{\sigma^2/(1-\beta)} T.$$

Substituting this into (C.50) and using (C.49) gives

$$\mathbb{P}(\widehat{U} \neq U) \;\geq\; 1 - \frac{C \, \mu^2 a^2 T/(\sigma^2/(1-\beta)) + \log 2}{(J/8) \log 2}.$$

Under the condition (C.52), the numerator is at most $(J/16)\log 2 + \log 2$, hence for all $J \geq 16$ the right-hand side is at least $1/2$ after adjusting constants (and for bounded J this only changes constants in the final lower bound). This yields (C.53). □

We can finally conclude with Lemma C.1 and the upper and lower bounds we have on the KL divergence and discrepancy measure χ respectively.

Lemma C.13 (Parameter tuning for the information-limited lower bound)**.** *Fix $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, and set $\alpha := 1 + d/p$. Consider the J-block construction $\{F^u_{1:T} : u \in \mathcal{U}\}$ with block length $\Delta_T := \lfloor T/J \rfloor$ and base losses $\{f_+, f_-\}$. Assume:*

(i) ***Gradient-variation bound:*** *from Lemma C.10, we have* $\mathrm{GVar}_{p,q}(F^u_{1:T}) \;\leq\; C \, \mu \, a^\alpha \, J^{1/q}, \quad \forall u \in \mathcal{U}$;

(ii) ***Gaussian-gradient KL bound:*** *from Lemma C.12, we have* $\sup_{u,v \in \mathcal{U}} D_{\mathrm{KL}}(P^\pi_u \,\|\, P^\pi_v) \;\leq\; C\mu^2 a^2 T/(\sigma^2/(1-\beta)), \quad \forall \pi$;

(iii) ***Packing size lower bound:*** *from Lemma C.9, we have the packing size lower bound* $\log |\mathcal{U}| \geq c\, J$ *for a universal constant $c > 0$ (e.g., from Varshamov–Gilbert or constant-weight codes).*

Let $\mathcal{V}_T > 0$ be the variation budget. Choose

$$a \;:=\; \left(\frac{\mathcal{V}_T}{C \, \mu \, J^{1/q}} \right)^{1/\alpha}. \tag{C.54}$$

Then $\mathrm{GVar}_{p,q}(F^u_{1:T}) \leq \mathcal{V}_T$ for all $u \in \mathcal{U}$. Moreover, if J additionally satisfies

$$C \, \frac{\mu^2 a^2}{\sigma^2/(1-\beta)} T \;\leq\; \frac{c}{2} J, \tag{C.55}$$

then Fano's method yields a constant testing error (hence a constant separation in regret) for the family $\{F^u_{1:T} : u \in \mathcal{U}\}$ under any policy. In particular, substituting (C.54) into (C.55) shows it suffices to take

$$J \;\gtrsim\; \left(\frac{\mu^{2-2/\alpha}}{\sigma^2/(1-\beta)} \, \mathcal{V}_T^{2/\alpha} \, T \right)^{\alpha q/(\alpha q + 2)}, \tag{C.56}$$

and for this choice one obtains the scaling

$$\inf_{\pi \in \Pi_\beta} \; \sup_{F : \mathrm{GVar}_{p,q}(F) \leq \mathcal{V}_T} \mathcal{R}^\pi_T(F) \;\gtrsim\; (1-\beta)^{-2/(\alpha q+2)} \sigma^{4/(\alpha q+2)} \mu^{(\alpha q - 2q - 2)/(\alpha q+2)} \mathcal{V}_T^{2q/(\alpha q+2)} T^{\alpha q/(\alpha q+2)}, \tag{C.57}$$

where $\alpha = 1 + d/p$.

Proof of Lemma C.13. First note that by Corollary C.2, we have that $\mathbb{P}(\widehat{U} \neq U) \geq \frac{1}{2}$. Subsequently invoking Lemma C.1, we conclude that there does not exist an admissible policy $\pi \in \Pi_\beta$ such that $\sup_{f \in \mathcal{F}_{p,q}(\mathcal{V}_T)} \mathcal{R}^{\pi,\phi^G}_T(f) \;\leq$

$\frac{1}{9}\inf_{f,\tilde{f}\in\Theta}\sum_{t=1}^{T}\chi(f_t,\tilde{f}_t)$. This gives us a lower bound for the minimax regret. Now we must just enforce the gradient variational budget and the information constraint. To enforce the gradient variational budget, we have by Lemma C.10 that the condition $\mathrm{GVar}_{p,q}(F^u_{1:T}) \le \mathcal{V}_T$ for all $u \in \mathcal{U}$ is ensured whenever $C\mu a^\alpha J^{1/q} \le \mathcal{V}_T$, which is exactly achieved by the choice (C.54).

To enforce the information (Fano/KL) constraint,, we have by Lemma C.12, for any policy π and any $u, v \in \mathcal{U}$,

$$D_{\mathrm{KL}}(P^\pi_u \,\|\, P^\pi_v) \;\le\; C\,\frac{\mu^2 a^2}{\sigma^2/(1-\beta)}\,T.$$

Combining this uniform pairwise bound with the standard mutual-information upper bound (Lemma D.8), we get

$$I(U; Z_{1:T}) \;\le\; C\,\frac{\mu^2 a^2}{\sigma^2/(1-\beta)}\,T.$$

Since $\log|\mathcal{U}| \ge cJ$, imposing (C.55) makes $I(U; Z_{1:T})$ a small constant fraction of $\log|\mathcal{U}|$. Fano's inequality then yields a constant lower bound on the minimax probability of misidentifying U, uniformly over all estimators $\widehat{U}(Z_{1:T})$.

Substituting (C.54) into (C.55) gives

$$C\,\frac{\mu^2 T}{\sigma^2/(1-\beta)}\left(\frac{\mathcal{V}_T}{C\mu J^{1/q}}\right)^{2/\alpha} \;\lesssim\; J,$$

or equivalently,

$$\frac{\mu^{2-2/\alpha}}{\sigma^2/(1-\beta)}\,\mathcal{V}_T^{2/\alpha}\,T \;\lesssim\; J^{1+\frac{2}{\alpha q}}.$$

Rearranging yields (C.56). Finally, the quantity that drives the regret separation in the information-limited regime is $\mu a^2 T$. Using (C.54), plugging $J \asymp J_\star$ from (C.56), and simplifying exponents yields we get

$$\inf_{\pi\in\Pi_\beta}\;\sup_{F:\,\mathrm{GVar}_{p,q}(F)\le\mathcal{V}_T}\;\mathcal{R}^\pi_T(F) \;\gtrsim\; (1-\beta)^{-2/(\alpha q+2)}\sigma^{4/(\alpha q+2)}\mu^{(\alpha q-2q-2)/(\alpha q+2)}\mathcal{V}_T^{2q/(\alpha q+2)}T^{\alpha q/(\alpha q+2)}.$$

□

C.6 Obtaining a minimax lower bound in the inertia-limited regime

Our information-theoretic construction (packing + Fano/KL) yields a lower bound that is driven by *statistical indistinguishability* of environments and therefore captures the dependence on the noise level σ and the variation budget $\mathcal{V}_T$. However, it does not by itself explain the empirically dominant failure mode of momentum under drift: *inertia*. To isolate this mechanism, we analyze SGDM (heavy-ball) on the simplest strongly convex and smooth objective, $\phi_u(z) = \frac{\mu}{2}(z - ua)^2$, where the minimizer jumps between $\pm a$. We prove the momentum-specific lower bound by explicitly analyzing SGDM on a single block switch and showing it takes $\tau_\beta := \Omega(\kappa/(1-\beta))$ steps to reduce the tracking error by a constant factor under a step size restriction. Crucially, this part does not rely on hiding information; the learner may instantly know the new function. The lower bound comes from the algorithmic constraint of SGDM under stability. We will focus on the Polyak heavy-ball method of momentum ($\beta_1 = 0, \beta_2 = \beta$). However a similar analysis can be carried out for Nesterov and yield a synonymous result.

Proposition C.1 (SGDM on a 1D quadratic)**.** *Fix $u \in \{\pm 1\}$ and consider the one-dimensional μ–strongly μ-smooth convex quadratic $\phi_u(z) := \frac{\mu}{2}(z - ua)^2$, $z \in \mathbb{R}$. Let heavy-ball SGDM (Polyak momentum) with step size $\gamma > 0$ and momentum $\beta \in [0, 1)$ evolve as*

$$z_{t+1} = z_t + \beta(z_t - z_{t-1}) - \gamma\big(\mu(z_t - ua) + \eta_t\big), \qquad \eta_t \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.} \tag{C.58}$$

Define the error $e_t := z_t - ua$. Then the mean error satisfies the deterministic linear recursion

$$\begin{pmatrix}\mathbb{E}[e_{t+1}]\\ \mathbb{E}[e_t]\end{pmatrix} = \underbrace{\begin{pmatrix}1+\beta-\gamma\mu & -\beta\\ 1 & 0\end{pmatrix}}_{\triangleq A}\begin{pmatrix}\mathbb{E}[e_t]\\ \mathbb{E}[e_{t-1}]\end{pmatrix}, \tag{C.59}$$

and hence $\|\binom{\mathbb{E}[e_{t+1}]}{\mathbb{E}[e_t]}\|_2 \le \|A\|_2^t \|\binom{\mathbb{E}[e_1]}{\mathbb{E}[e_0]}\|_2$. *Moreover, let* $\lambda_{\max}$ *denote the eigenvalue of* A *with maximal modulus. If*

$$0 < \gamma\mu \le \frac{1-\beta}{4}, \tag{C.60}$$

then A *has two real eigenvalues in* $(0,1)$ *and there exists a universal constant* $c > 0$ *such that*

$$|\lambda_{\max}| \;\ge\; 1 - c\,\frac{\gamma\mu}{1-\beta}. \tag{C.61}$$

Consequently, for any initialization with $\mathbb{E}[e_0] = \mathbb{E}[e_{-1}] = a$*, there exists a universal constant* $C \ge 1$ *such that for all* $t \ge 0$*,*

$$|\mathbb{E}[e_t]| \;\ge\; C^{-1}|\lambda_{\max}|^t\, a \;\ge\; C^{-1}\Big(1 - c\,\frac{\gamma\mu}{1-\beta}\Big)^t a. \tag{C.62}$$

In particular, defining the response time

$$\tau_\beta \;:=\; \min\Big\{t \ge 0 :\; \Big(1 - c\,\frac{\gamma\mu}{1-\beta}\Big)^t \le \frac{1}{2}\Big\}, \tag{C.63}$$

we have the scaling

$$\tau_\beta \;\asymp\; \frac{1-\beta}{\gamma\mu}. \tag{C.64}$$

Finally, Jensen's inequality yields the expected suboptimality lower bound

$$\mathbb{E}\big[\phi_u(z_t) - \phi_u(ua)\big] = \frac{\mu}{2}\mathbb{E}[e_t^2] \;\ge\; \frac{\mu}{2}\big(\mathbb{E}[e_t]\big)^2 \;\gtrsim\; \mu a^2\Big(1 - c\,\frac{\gamma\mu}{1-\beta}\Big)^{2t}. \tag{C.65}$$

Proof of Proposition C.1. Subtracting ua from (C.58) gives

$$e_{t+1} = (1+\beta-\gamma\mu)e_t - \beta e_{t-1} - \gamma\eta_t. \tag{C.66}$$

Taking expectations and using $\mathbb{E}[\eta_t] = 0$ yields the homogeneous recursion

$$\mathbb{E}[e_{t+1}] = (1+\beta-\gamma\mu)\mathbb{E}[e_t] - \beta\mathbb{E}[e_{t-1}],$$

which is equivalently the two-dimensional linear system (C.59) with matrix A as stated. The eigenvalues of A are the roots of

$$\lambda^2 - (1+\beta-\gamma\mu)\lambda + \beta = 0. \tag{C.67}$$

Under (C.60), the discriminant is nonnegative so $\lambda_\pm \in \mathbb{R}$. Moreover, $\lambda_\pm \in (0,1)$ in this regime. Let $\lambda_{\max} := \max\{\lambda_+, \lambda_-\}$. To lower bound $\lambda_{\max}$, write $\delta := 1-\beta \in (0,1]$ and $\varepsilon := \gamma\mu$. Then

$$\lambda_{\max} = 1 - \frac{\delta+\varepsilon}{2} + \frac{1}{2}\sqrt{(\delta-\varepsilon)^2 + 2\varepsilon\delta}.$$

Using $\sqrt{x+y} \ge \sqrt{x} + \frac{y}{2\sqrt{x+y}}$ with $x = (\delta-\varepsilon)^2$ and $y = 2\varepsilon\delta$, and the bound $\sqrt{(\delta-\varepsilon)^2 + 2\varepsilon\delta} \le \delta + \varepsilon$, one obtains (for $\varepsilon \le \delta/4$) a universal $c > 0$ such that

$$\lambda_{\max} \;\ge\; 1 - c\,\frac{\varepsilon}{\delta} \;=\; 1 - c\,\frac{\gamma\mu}{1-\beta},$$

which is (C.61). Since A is diagonalizable with eigenvalues in $(0,1)$, the solution of (C.59) can be written as a linear combination of λ_+^t and λ_-^t. For the symmetric initialization $\mathbb{E}[e_0] = \mathbb{E}[e_{-1}] = a$, the dominant mode is aligned with $\lambda_{\max}$, hence there exists a constant $C \ge 1$ (depending only on the conditioning of the eigenbasis, and thus universal in the regime (C.60)) such that (C.62) holds. The response time bound (C.64) follows immediately from the definition (C.63) and the standard approximation $\log(1-x) \asymp -x$ for $x \in (0, 1/2)$, applied with $x = c\,\gamma\mu/(1-\beta)$. Finally, since $\phi_u(z_t) - \phi_u(ua) = \frac{\mu}{2}e_t^2$ and $\mathbb{E}[e_t^2] \ge (\mathbb{E}[e_t])^2$, plugging (C.62) into this inequality yields (C.65). □

We now use this as the block length $\Delta_T \asymp \tau_\beta$. Since in our J-block construction, the minimizer is constant within a block and flips between blocks. If the blocks are much longer than τ_β, SGDM has time to settle near the new minimizer after each switch. Then the regret per block is mostly transient and doesn't accumulate strongly. On the other hand, if the block size is much shorter than τ_β, SGDM will never be able to catch up. Taking $\Delta_T \asymp \tau_\beta$ will give us a constant per-round suboptimality throughout the block (up to constants), hence resulting in $\Omega(\mu a^2 \Delta_T)$ regret contribution per block. We formalize this in the following result:

Theorem C.1 (From response time to regret under block switching)**.** *Consider the J-block construction of nonstationary losses $F^u_{1:T}$ over $[T]$ with blocks $B_1, \dots, B_J$ of lengths $|B_j| \in \{\Delta_T, \Delta_T + 1\}$, $\Delta_T = \lfloor T/J \rfloor$, and $F^u_t \equiv f_{u_j}$ for $t \in B_j$, where $u \in \{\pm 1\}^J$. Assume the base losses are the translated quadratics $\phi_\pm$ above (embedded along e_1 if $d > 1$), so that within each block the unique minimizer is $x^\star_{u_j} = u_j a$. Run heavy-ball SGDM with parameters (γ, β) satisfying the stability cap*

$$\gamma \;\le\; c_0 \frac{(1-\beta)^2}{L}, \tag{C.68}$$

for a universal constant $c_0 > 0$. Let τ_β be the response time from Proposition C.1, so in particular

$$\tau_\beta \;\gtrsim\; \frac{L}{\mu(1-\beta)}. \tag{C.69}$$

Choose the block length on the order of the response time,

$$\Delta_T \;\asymp\; \tau_\beta, \qquad \text{equivalently} \qquad J \;\asymp\; \frac{T}{\tau_\beta}. \tag{C.70}$$

Then there exists a universal constant $c > 0$ such that for any $u \in \{\pm 1\}^J$ with a sign flip at each block boundary (e.g., $u_{j+1} = -u_j$), the expected dynamic regret of heavy-ball SGDM satisfies

$$\mathcal{R}^{\pi_{\mathrm{HB}}}_T(F^u) \;\ge\; c\,\mu a^2 T. \tag{C.71}$$

Moreover, if the block family is tuned to satisfy the gradient-variation budget $\mathrm{GVar}_{p,q}(F^u_{1:T}) \le \mathcal{V}_T$ via Lemma C.10, i.e.

$$a \;\asymp\; \left(\frac{\mathcal{V}_T}{\mu\, J^{1/q}}\right)^{1/\alpha}, \qquad \alpha := 1 + \frac{d}{p}, \tag{C.72}$$

then substituting $J \asymp T/\tau_\beta$ into (C.71) yields

$$\mathcal{R}^{\pi_{\mathrm{HB}}}_T(F^u) \;\gtrsim\; \mu^{1-2/\alpha}\, \mathcal{V}_T^{2/\alpha}\, \tau_\beta^{2/(\alpha q)}\, T^{1-(2/\alpha q)}. \tag{C.73}$$

Finally, using (C.69) gives the explicit inertia-dependent lower bound

$$\mathcal{R}^{\pi_{\mathrm{HB}}}_T(F^u) \;\gtrsim\; \mu^{1-2/\alpha}\, \mathcal{V}_T^{2/\alpha} \left(\frac{L}{\mu(1-\beta)}\right)^{2/(\alpha q)} T^{1-(2/\alpha q)}. \tag{C.74}$$

Proof of Theorem C.1. The proof has three steps: (i) a single sign flip creates an $\Omega(a)$ initialization error (in the extended state) relative to the *new* minimizer, (ii) over a time window of length $\Theta(\tau_\beta)$ this error cannot contract by more than a constant factor, and (iii) strong convexity converts this persistent distance into $\Omega(\mu a^2)$ per-round regret, which then accumulates across blocks.

Step 1: A flip induces an $\Omega(a)$ error at the start of the next block. Consider a boundary between two consecutive blocks B_j and B_{j+1} at which the sign flips, $u_{j+1} = -u_j$. Let

$$x^\star_j := u_j a, \qquad x^\star_{j+1} := u_{j+1} a = -u_j a, \qquad \text{so that} \qquad |x^\star_{j+1} - x^\star_j| = 2a.$$

Let the first time index in block B_{j+1} be t_0 and define the post-switch error variables

$$e_t := x_t - x^\star_{j+1}, \qquad t \in B_{j+1},$$

(and analogously for $e_{t_0-1} = x_{t_0-1} - x^\star_{j+1}$). By the reverse triangle inequality,

$$|e_{t_0-1}| = \left|x_{t_0-1} - x^\star_{j+1}\right| \geq |x^\star_j - x^\star_{j+1}| - \left|x_{t_0-1} - x^\star_j\right| = 2a - \left|x_{t_0-1} - x^\star_j\right|. \tag{C.75}$$

Thus, whenever the iterate has achieved even a moderate accuracy on the previous block, say $|x_{t_0-1} - x^\star_j| \leq a$, we obtain $|e_{t_0-1}| \geq a$. In particular, for the alternating-sign choice in the theorem (a flip at every boundary), either the algorithm fails to track the previous minimizer by time $t_0 - 1$—which already incurs regret on block B_j—or else it necessarily starts block B_{j+1} with initial error at least a relative to the new minimizer. This dichotomy is what allows us to lower bound regret block-by-block. To make this quantitative in a way compatible with the heavy-ball state, define the extended state

$$s_t := \begin{bmatrix} e_t \\ e_{t-1} \end{bmatrix}.$$

There exists a universal constant $c_{\text{init}} > 0$ such that at the beginning of each block with a flip,

$$\|s_{t_0-1}\|_2 \;\geq\; c_{\text{init}}\, a. \tag{C.76}$$

We will use (C.76) as the initial condition for the lag argument in Step 2.

Step 2: Over $\Theta(\tau_\beta)$ steps the error cannot shrink by more than a constant factor. On block B_{j+1} the loss is the fixed quadratic $f_{u_{j+1}}(x) = \frac{\mu}{2}(x - x^\star_{j+1})^2$, so $\nabla f_{u_{j+1}}(x) = \mu(x - x^\star_{j+1}) = \mu e_t$. The heavy-ball recursion therefore yields the homogeneous linear recurrence

$$e_{t+1} = (1 + \beta - \gamma\mu)e_t - \beta e_{t-1}, \qquad t \in B_{j+1}, \tag{C.77}$$

and in state form

$$s_{t+1} = A\, s_t, \qquad A = \begin{bmatrix} 1 + \beta - \gamma\mu & -\beta \\ 1 & 0 \end{bmatrix}. \tag{C.78}$$

By definition of the response time τ_β from Proposition C.1, there exists a universal constant $c_{\text{lag}} \in (0, 1)$ such that for all $k \leq c_{\text{lag}}\tau_\beta$, the contraction of the deterministic dynamics is at most a constant factor, i.e.

$$\|s_{t_0-1+k}\|_2 \;\geq\; \frac{1}{2}\,\|s_{t_0-1}\|_2. \tag{C.79}$$

Combining (C.76) and (C.79) gives that throughout the first $c_{\text{lag}}\tau_\beta$ steps of the new block,

$$\|s_{t_0-1+k}\|_2 \;\geq\; \frac{c_{\text{init}}}{2}\, a, \qquad k = 0, 1, \ldots, \lfloor c_{\text{lag}}\tau_\beta \rfloor. \tag{C.80}$$

In particular, since $\|s_t\|_2 \geq |e_t|$, we have $|e_{t_0+k}| \gtrsim a$ on the same window.

Step 3: Strong convexity turns persistent distance into regret and accumulates across blocks. For the quadratic loss on block B_{j+1},

$$f_{u_{j+1}}(x_t) - f_{u_{j+1}}(x^\star_{j+1}) = \frac{\mu}{2}\, e_t^2 \;\geq\; \frac{\mu}{2}\left(|e_t|\right)^2. \tag{C.81}$$

By (C.80), for each $k \leq \lfloor c_{\text{lag}}\tau_\beta \rfloor$ we have $|e_{t_0+k}| \gtrsim a$, hence the per-round regret on that window is

$$f_{t_0+k}(x_{t_0+k}) - f_{t_0+k}(x^\star_{t_0+k}) \;=\; f_{u_{j+1}}(x_{t_0+k}) - f_{u_{j+1}}(x^\star_{j+1}) \;\gtrsim\; \mu a^2. \tag{C.82}$$

Summing (C.82) over $k = 0, \ldots, \lfloor c_{\text{lag}}\tau_\beta \rfloor$ gives an expected regret contribution of order $\mu a^2 \tau_\beta$ per flipped block:

$$\sum_{t \in B_{j+1}} \left(f_t(x_t) - f_t(x^\star_t)\right) \;\geq\; \sum_{k=0}^{\lfloor c_{\text{lag}}\tau_\beta \rfloor} \left(f_{t_0+k}(x_{t_0+k}) - f_{t_0+k}(x^\star_{t_0+k})\right) \;\gtrsim\; \mu a^2\, \tau_\beta. \tag{C.83}$$

With the block-length choice $\Delta_T \asymp \tau_\beta$ in (C.70), this becomes $\gtrsim \mu a^2 \Delta_T$ per block. Since the construction flips at every boundary, a constant fraction of the J blocks contribute this amount, and therefore

$$\mathcal{R}_T^{\pi_{\mathrm{HB}}}(F^u) = \sum_{t=1}^{T} \big(f_t(x_t) - f_t(x_t^\star)\big) \;\gtrsim\; J \cdot \mu a^2 \Delta_T \;\asymp\; \mu a^2 T,$$

which proves (C.71).

Under the J-block construction, Lemma C.10 gives $\mathrm{GVar}_{p,q}(F^u_{1:T}) \lesssim \mu a^\alpha J^{1/q}$. Enforcing $\mathrm{GVar}_{p,q}(F^u_{1:T}) \le \mathcal{V}_T$ yields (C.72) (absorbing constants). Plugging (C.72) into (C.71) gives

$$\mathcal{R}_T^{\pi_{\mathrm{HB}}}(F^u) \;\gtrsim\; \mu T\Big(\frac{\mathcal{V}_T}{\mu J^{1/q}}\Big)^{2/\alpha} = \mu^{1-2/\alpha}\, \mathcal{V}_T^{2/\alpha}\, T\, J^{-2/(\alpha q)}.$$

Using $J \asymp T/\tau_\beta$ from (C.70) yields (C.73). Finally, the stability cap (C.68) implies $\gamma \lesssim (1-\beta)^2/L$, and combining this with $\tau_\beta \asymp (1-\beta)/(\gamma\mu)$ from Proposition C.1 yields $\tau_\beta \gtrsim L/(\mu(1-\beta))$, i.e. (C.69). Substituting this lower bound into (C.73) gives (C.74). □

Since both the statistical and inertia constructions produce μ-strongly convex, μ-smooth functions satisfying $\mathrm{GVar}_{p,q} \le \mathcal{V}_T$, the restricted minimax regret over Π_β is atleast the maximum of the two lower bounds. This completes the proof of the result.

D Technical Lemmas

D.1 Conditional Orlicz norms

In this section, we will introduce the definition of conditional Orlicz norm. These properties follow from [SZZ26].

Essential supremum/infimum. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let Φ be a collection of real-valued random variables on it. A random variable $\phi^\star$ is called the *essential supremum* of Φ, denoted $\operatorname{ess\,sup}\Phi$, if: (i) $\phi^\star \ge \phi$ a.s. for all $\phi \in \Phi$, and (ii) whenever ψ satisfies $\psi \ge \phi$ a.s. for all $\phi \in \Phi$, then $\psi \ge \phi^\star$ a.s. The *essential infimum* is defined by $\operatorname{ess\,inf}\Phi := -\operatorname{ess\,sup}(-\Phi)$. If Φ is *directed upward* (i.e., for any $\phi, \tilde\phi \in \Phi$ there exists $\psi \in \Phi$ with $\psi \ge \phi \vee \tilde\phi$ a.s.), then there exists an increasing sequence $\phi_1 \le \phi_2 \le \cdots$ in Φ such that $\operatorname{ess\,sup}\Phi = \lim_{n\to\infty}\phi_n$ a.s.

Theorem D.1 (Basic properties)**.** *Fix $\alpha \ge 1$ and let X be a real-valued random variable. For a sub-σ-field $\mathcal{G} \subseteq \mathcal{F}$, define the set of admissible (random) scales*

$$\Phi_\alpha(X \mid \mathcal{G}) \;:=\; \Big\{\phi : \Omega \to (0,\infty)\ \mathcal{G}\text{-measurable} \;:\; \mathbb{E}\big[\exp\big(|X/\phi|^\alpha\big) \mid \mathcal{G}\big] \le 2 \text{ a.s.}\Big\}.$$

The conditional Ψ_α–Orlicz norm *of X given $\mathcal{G}$ is*

$$\|X \mid \mathcal{G}\|_{\Psi_\alpha} \;:=\; \operatorname{ess\,inf}\, \Phi_\alpha(X \mid \mathcal{G}).$$

Let $\alpha \ge 1$ and $\mathcal{G} \subseteq \mathcal{F}$. Then $\|\cdot \mid \mathcal{G}\|_{\Psi_\alpha}$ is well-defined (a.s. finite whenever $\Phi_\alpha(X \mid \mathcal{G}) \ne \emptyset$) and satisfies:

(i) ***Positive homogeneity:*** *for any scalar a, $\|aX \mid \mathcal{G}\|_{\Psi_\alpha} = |a| \cdot \|X \mid \mathcal{G}\|_{\Psi_\alpha}$ a.s.*

(ii) ***Definiteness:*** *$\|X \mid \mathcal{G}\|_{\Psi_\alpha} = 0$ a.s. iff $X = 0$ a.s.*

(iii) ***Triangle inequality:*** *$\|X + Y \mid \mathcal{G}\|_{\Psi_\alpha} \le \|X \mid \mathcal{G}\|_{\Psi_\alpha} + \|Y \mid \mathcal{G}\|_{\Psi_\alpha}$ a.s.*

(iv) ***Normalization:***

$$\mathbb{E}\left[\exp\left(\left|\frac{X}{\|X \mid \mathcal{G}\|_{\Psi_\alpha}}\right|^\alpha\right) \,\middle|\, \mathcal{G}\right] \le 2 \qquad \text{a.s.}$$

Proof. Items (i) and the reverse implication of (ii) are immediate from the definition. For the converse direction of (ii), assume $\|X \mid \mathcal{G}\|_{\Psi_\alpha} = 0$ a.s. By definition of essential infimum, for every $\varepsilon > 0$ there exists a $\mathcal{G}$-measurable $\phi_\varepsilon \in (0, \varepsilon]$ such that $\mathbb{E}[\exp(|X/\phi_\varepsilon|^\alpha) \mid \mathcal{G}] \le 2$ a.s. On the event $\{|X| > 0\}$, we have $|X/\phi_\varepsilon| \to \infty$ as $\varepsilon \downarrow 0$, so

$\exp(|X/\phi_\varepsilon|^\alpha) \to \infty$. By conditional Fatou, $\mathbb{E}[\liminf_{\varepsilon\downarrow 0}\exp(|X/\phi_\varepsilon|^\alpha) \mid \mathcal{G}] \le \liminf_{\varepsilon\downarrow 0}\mathbb{E}[\exp(|X/\phi_\varepsilon|^\alpha) \mid \mathcal{G}] \le 2$, which forces $\mathbb{P}(|X| > 0) = 0$, i.e., $X = 0$ a.s.

For (iii), fix $\varepsilon > 0$ and set $u := \|X \mid \mathcal{G}\|_{\Psi_\alpha} + \varepsilon$ and $v := \|Y \mid \mathcal{G}\|_{\Psi_\alpha} + \varepsilon$. By definition of essential infimum, u, v are admissible scales, hence $\mathbb{E}[\exp(|X/u|^\alpha) \mid \mathcal{G}] \le 2$ and $\mathbb{E}[\exp(|Y/v|^\alpha) \mid \mathcal{G}] \le 2$ a.s. Let $s := u/(u+v)$ (note s is $\mathcal{G}$-measurable and $s \in (0,1)$). Then

$$\frac{|X+Y|}{u+v} \;\le\; s\cdot\frac{|X|}{u} + (1-s)\cdot\frac{|Y|}{v}.$$

Since $t \mapsto t^\alpha$ and $\exp(\cdot)$ are convex and increasing on $\mathbb{R}_+$, we have

$$\exp\Big(\big(sA + (1-s)B\big)^\alpha\Big) \;\le\; \exp(sA^\alpha + (1-s)B^\alpha) \;\le\; s\,e^{A^\alpha} + (1-s)\,e^{B^\alpha} \qquad (A, B \ge 0).$$

Applying this with $A = |X|/u$ and $B = |Y|/v$ and taking conditional expectations yields

$$\mathbb{E}\Bigg[\exp\bigg(\Big|\frac{X+Y}{u+v}\Big|^\alpha\bigg) \,\Big|\, \mathcal{G}\Bigg] \le s\,\mathbb{E}\big[e^{(|X|/u)^\alpha} \mid \mathcal{G}\big] + (1-s)\,\mathbb{E}\big[e^{(|Y|/v)^\alpha} \mid \mathcal{G}\big] \le 2.$$

Thus $u + v \in \Phi_\alpha(X+Y \mid \mathcal{G})$, so by taking the essential infimum over admissible scales and sending $\varepsilon \downarrow 0$ we obtain (iii).

For (iv), note that $\Phi_\alpha(X \mid \mathcal{G})$ is directed downward (equivalently, $-\Phi_\alpha$ is directed upward), so there exists a decreasing sequence $\{\phi_n\}_{n\ge1} \subset \Phi_\alpha(X \mid \mathcal{G})$ with $\phi_n \downarrow \|X \mid \mathcal{G}\|_{\Psi_\alpha}$ a.s. By conditional Fatou,

$$\mathbb{E}\Bigg[\exp\bigg(\Big|\frac{X}{\|X \mid \mathcal{G}\|_{\Psi_\alpha}}\Big|^\alpha\bigg) \,\Big|\, \mathcal{G}\Bigg] \le \liminf_{n\to\infty}\mathbb{E}\Bigg[\exp\bigg(\Big|\frac{X}{\phi_n}\Big|^\alpha\bigg) \,\Big|\, \mathcal{G}\Bigg] \le 2,$$

as claimed. □

For a random vector $\mathbf{Z} \in \mathbb{R}^d$, we use the standard sub-Gaussian extension

$$\|\mathbf{Z} \mid \mathcal{G}\|_{\Psi_2} \;:=\; \sup_{\mathbf{u}\in\mathbb{S}^{d-1}} \|\langle \mathbf{u}, \mathbf{Z}\rangle \mid \mathcal{G}\|_{\Psi_2}.$$

For a random matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, define similarly $\|\mathbf{M} \mid \mathcal{G}\|_{\Psi_1} := \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{d-1}} \|\mathbf{u}^\top\mathbf{M}\mathbf{v} \mid \mathcal{G}\|_{\Psi_1}$.

Lemma D.1 (Conditional "sub-Gaussian × sub-Gaussian ⇒ sub-exponential")**.** *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ be random vectors and $\mathcal{G} \subseteq \mathcal{F}$. Then*

$$\|\mathbf{X}\mathbf{Y}^\top \mid \mathcal{G}\|_{\Psi_1} \;\le\; \|\mathbf{X} \mid \mathcal{G}\|_{\Psi_2}\cdot\|\mathbf{Y} \mid \mathcal{G}\|_{\Psi_2} \qquad \textit{a.s.}$$

Proof. For any $u, v \in \mathbb{S}^{d-1}$, $\mathbf{u}^\top\mathbf{X}\mathbf{Y}^\top\mathbf{v} = \langle \mathbf{u}, \mathbf{X}\rangle\langle \mathbf{v}, \mathbf{Y}\rangle$. The scalar Orlicz inequality $\|\mathbf{U}\mathbf{V} \mid \mathcal{G}\|_{\Psi_1} \le \|\mathbf{U} \mid \mathcal{G}\|_{\Psi_2}\|\mathbf{V} \mid \mathcal{G}\|_{\Psi_2}$ holds by the usual Young-type argument (applied conditionally), and taking the suprema over $\mathbf{u}, \mathbf{v}$ gives the claim. □

We will repeatedly use the following consequences of the definition and the tower property of conditional expectation.

Lemma D.2 (Monotonicity and conditioning rules)**.** *(i) If $\mathcal{G}_1 \subseteq \mathcal{G}_2$ and $\|X \mid \mathcal{G}_2\|_{\Psi_2} \le K$ a.s. for a constant K, then $\|X \mid \mathcal{G}_1\|_{\Psi_2} \le K$ a.s.*

(ii) If $|Y| \le K$ a.s. for a constant K, then $\|XY \mid \mathcal{G}\|_{\Psi_\alpha} \le K\|X \mid \mathcal{G}\|_{\Psi_\alpha}$ a.s.

(iii) If X is independent of $\mathcal{G}$, then $\|X \mid \mathcal{G}\|_{\Psi_\alpha} = \|X\|_{\Psi_\alpha}$ a.s.

(iv) If $X \perp Y \mid \mathcal{G}$, then $\mathbb{E}[XY \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}]\,\mathbb{E}[Y \mid \mathcal{G}]$ a.s., and $\mathbb{E}[X \mid Y, \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}]$ a.s.

Proof. (i) By assumption, $\mathbb{E}[\exp(|X/K|^2) \mid \mathcal{G}_2] \le 2$ a.s.; taking $\mathcal{G}_1$-conditional expectations and using the tower property yields $\mathbb{E}[\exp(|X/K|^2) \mid \mathcal{G}_1] \le 2$ a.s., hence $\|X \mid \mathcal{G}_1\|_{\Psi_2} \le K$. (ii) If ϕ is admissible for X given $\mathcal{G}$, then $K\phi$ is admissible for XY given $\mathcal{G}$ since $|XY|/(K\phi) \le |X|/\phi$. Taking essential infima yields the claim. (iii) If $X \perp \mathcal{G}$, then $\mathbb{E}[\exp(|X/\phi|^\alpha) \mid \mathcal{G}] = \mathbb{E}[\exp(|X/\phi|^\alpha)]$ for any constant ϕ, so the conditional and unconditional admissible scales coincide a.s. (iv) this follows from the standard properties of conditional independence. □

Lemma D.3 (A frequently used specialization)**.** *Let $\mathcal{G}_1 \subseteq \mathcal{G}_2$. If $\|X \mid \mathcal{G}_1\|_{\Psi_2} \le \lambda$ a.s. and $\|\xi \mid \mathcal{G}_2\|_{\Psi_2} \le K$ a.s. for constants λ, K, then*

$$\|\xi X \mid \mathcal{G}_1\|_{\Psi_1} \;\le\; K\lambda \qquad a.s.$$

Proof. By Lemma D.2(i), $\|\xi \mid \mathcal{G}_1\|_{\Psi_2} \le K$ a.s. Apply Lemma D.1 (in the scalar case) conditioned on $\mathcal{G}_1$ to obtain $\|\xi X \mid \mathcal{G}_1\|_{\Psi_1} \le \|\xi \mid \mathcal{G}_1\|_{\Psi_2} \cdot \|X \mid \mathcal{G}_1\|_{\Psi_2} \le K\lambda$ a.s. □

Lemma D.4 (Conditional Ψ_2 control implies conditional second moments)**.** *Let $\mathcal{F}$ be a σ-field and let $K_{\mathcal{F}} > 0$ be $\mathcal{F}$-measurable.*

*i) **(Scalar).** If $\mathbb{E}\Big[\exp\Big(|X|^2/K_{\mathcal{F}}^2\Big) \mid \mathcal{F}\Big] \le 2$ a.s, then $\mathbb{E}[|X|^2 \mid \mathcal{F}] \le K_{\mathcal{F}}^2$ a.s*

*ii) **(Vector).** Let $\mathbf{X} \in \mathbb{R}^d$. If $\sup_{\mathbf{u}\in\mathbb{S}^{d-1},\, \mathbf{u}\in\mathcal{F}} \mathbb{E}\Big[\exp\Big(|\mathbf{u}^\top \mathbf{X}|^2/K_{\mathcal{F}}^2\Big) \mid \mathcal{F}\Big] \le 2$ a.s , then $\mathbb{E}[\mathbf{X}\mathbf{X}^\top \mid \mathcal{F}] \preceq K_{\mathcal{F}}^2 I_d$ a.s and hence $\mathbb{E}[\|\mathbf{X}\|_2^2 \mid \mathcal{F}] \le dK_{\mathcal{F}}^2$ a.s*

Proof. We first prove the scalar case. By $e^y \ge 1 + y$ for $y \ge 0$,

$$2 \;\ge\; \mathbb{E}\left[1 + \frac{|X|^2}{K_{\mathcal{F}}^2} \;\Big|\; \mathcal{F}\right] = 1 + \frac{1}{K_{\mathcal{F}}^2}\mathbb{E}[|X|^2 \mid \mathcal{F}] \quad \text{a.s.,}$$

so $\mathbb{E}[|X|^2 \mid \mathcal{F}] \le K_{\mathcal{F}}^2$.

The vector case follows very similarly. Fix any $\mathcal{F}$-measurable $\mathbf{u} \in \mathbb{S}^{d-1}$ and apply the scalar part to $\mathbf{u}^\top\mathbf{X}$ to get $\mathbb{E}[(\mathbf{u}^\top\mathbf{X})^2 \mid \mathcal{F}] \le K_{\mathcal{F}}^2$ a.s. Since $\mathbf{u}^\top\mathbb{E}[\mathbf{X}\mathbf{X}^\top \mid \mathcal{F}]\mathbf{u} = \mathbb{E}[(\mathbf{u}^\top\mathbf{X})^2 \mid \mathcal{F}]$, this yields $\mathbf{u}^\top\mathbb{E}[\mathbf{X}\mathbf{X}^\top \mid \mathcal{F}]\mathbf{u} \le K_{\mathcal{F}}^2$ for all $\mathbf{u} \in \mathbb{S}^{d-1}$, i.e., $\mathbb{E}[\mathbf{X}\mathbf{X}^\top \mid \mathcal{F}] \preceq K_{\mathcal{F}}^2 I_d$. Taking traces gives $\mathbb{E}[\|\mathbf{X}\|_2^2 \mid \mathcal{F}] = \mathrm{tr}(\mathbb{E}[\mathbf{X}\mathbf{X}^\top \mid \mathcal{F}]) \le \mathrm{tr}(K_{\mathcal{F}}^2 I_d) = dK_{\mathcal{F}}^2$. □

D.2 Martingale Concentration Inequalities

Lemma D.5 (Bernstein inequality for sub-exponential martingale differences)**.** *Let $(\mathcal{F}_t)_{t=0}^T$ be a filtration and let $Z_1, \ldots, Z_T$ be real-valued random variables such that $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$ a.s. $(t = 1, \ldots, T)$. Assume that $Z_t \mid \mathcal{F}_{t-1}$ is conditionally sub-exponential in the sense that $\big\|Z_t \mid \mathcal{F}_{t-1}\big\|_{\Psi_1} \le K_t$ a.s., $(t = 1, \ldots, T)$, where each K_t is a (deterministic) constant. Then there exist absolute constants $c, C > 0$ such that for all $s \ge 0$,*

$$\mathbb{P}\left(\sum_{t=1}^T Z_t \ge s\right) \;\le\; \exp\left(-\min\left\{\frac{s^2}{C\sum_{t=1}^T K_t^2}\,,\; \frac{s}{C\max_{1\le t\le T} K_t}\right\}\right).$$

Proof. Fix $\lambda > 0$. By Markov's inequality,

$$\mathbb{P}\left(\sum_{t=1}^T Z_t \ge s\right) = \mathbb{P}\left(\exp\Big(\lambda\sum_{t=1}^T Z_t\Big) \ge e^{\lambda s}\right) \le e^{-\lambda s}\,\mathbb{E}\exp\Big(\lambda\sum_{t=1}^T Z_t\Big). \tag{D.1}$$

Let $S_t := \sum_{i=1}^t Z_i$ with $S_0 := 0$. Using tower property and $\mathcal{F}_{T-1}$-measurability of S_{T-1},

$$\mathbb{E}e^{\lambda S_T} = \mathbb{E}\Big[e^{\lambda S_{T-1}}\,\mathbb{E}\big[e^{\lambda Z_T} \mid \mathcal{F}_{T-1}\big]\Big].$$

We now use the standard sub-exponential mgf bound in conditional form: there exist absolute constants $c, C > 0$ such that for any random variable X with $\|X\|_{\Psi_1} \le K$,

$$\mathbb{E}[e^{\lambda X}] \le \exp(C\lambda^2K^2) \quad \text{for all } 0 \le \lambda \le c/K.$$

Applying this conditional on $\mathcal{F}_{T-1}$ (and using $\|Z_T \mid \mathcal{F}_{T-1}\|_{\Psi_1} \le K_T$ a.s.), we obtain that for $0 \le \lambda \le c/K_T$,

$$\mathbb{E}\big[e^{\lambda Z_T} \mid \mathcal{F}_{T-1}\big] \le \exp(C\lambda^2K_T^2) \quad \text{a.s.}$$

Hence, for $0 \le \lambda \le c/K_T$,

$$\mathbb{E}e^{\lambda S_T} \le \exp(C\lambda^2K_T^2)\,\mathbb{E}e^{\lambda S_{T-1}}.$$

Iterating this argument for $t = T, T-1, \dots, 1$ yields that for

$$0 \le \lambda \le \frac{c}{\max_{1\le t\le T} K_t},$$

we have

$$\mathbb{E}\exp\Big(\lambda \sum_{t=1}^{T} Z_t\Big) = \mathbb{E}e^{\lambda S_T} \le \exp\left(C\lambda^2 \sum_{t=1}^{T} K_t^2\right). \tag{D.2}$$

Combining (D.1) and (D.2) gives, for all admissible λ,

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_t \ge s\right) \le \exp\left(-\lambda s + C\lambda^2 \sum_{t=1}^{T} K_t^2\right).$$

Choose

$$\lambda := \min\left\{\frac{s}{2C\sum_{t=1}^{T} K_t^2}\, ,\ \frac{c}{\max_{1\le t\le T} K_t}\right\}.$$

Substituting this choice into the previous bound yields

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_t \ge s\right) \le \exp\left(-\min\left\{\frac{s^2}{C\sum_{t=1}^{T} K_t^2}\, ,\ \frac{s}{C\max_{1\le t\le T} K_t}\right\}\right),$$

after adjusting absolute constants, which proves the claim. □

D.3 Martingale Theory

Lemma D.6 (Optional stopping / sampling theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a filtration $(\mathcal{F}_t)_{t\ge 0}$, and let $(M_t)_{t\ge 0}$ be an integrable $(\mathcal{F}_t)$-martingale, i.e. $\mathbb{E}[|M_t|] < \infty$ and $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = M_t$ a.s. for all $t \ge 0$. Let τ be an $(\mathcal{F}_t)$-stopping time.*

(i) ***(Bounded stopping time).*** *If $\tau \le T$ a.s. for some deterministic $T < \infty$, then*

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0].$$

(ii) ***(Unbounded stopping time via truncation).*** *Assume $\mathbb{E}[|M_{\tau\wedge n}|] < \infty$ for all n and that $\{M_{\tau\wedge n}\}_{n\ge 1}$ is uniformly integrable. Then*

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0].$$

More generally, if (M_t) is an integrable supermartingale (resp. submartingale), then in either *(i) or (ii) we have $\mathbb{E}[M_\tau] \le \mathbb{E}[M_0]$ (resp. $\mathbb{E}[M_\tau] \ge \mathbb{E}[M_0]$).*

Proof. We first show that the stopped process is a martingale:

Step 1: Show that the stopped process is a martingale. Fix a stopping time τ and define the stopped process

$$M_t^\tau := M_{t\wedge\tau}, \qquad t \ge 0.$$

We claim $(M_t^\tau)_{t\ge 0}$ is an $(\mathcal{F}_t)$-martingale (assuming integrability). Fix $t \ge 0$. Since $\{\tau \le t\} \in \mathcal{F}_t$, we can split on the events $\{\tau \le t\}$ and $\{\tau > t\}$:

$$\begin{aligned}
\mathbb{E}[M_{(t+1)\wedge\tau} \mid \mathcal{F}_t] &= \mathbb{E}[M_\tau \mathbf{1}\{\tau \le t\} + M_{t+1}\mathbf{1}\{\tau > t\} \mid \mathcal{F}_t] \\
&= M_\tau \mathbf{1}\{\tau \le t\} + \mathbf{1}\{\tau > t\}\, \mathbb{E}[M_{t+1} \mid \mathcal{F}_t] \\
&= M_\tau \mathbf{1}\{\tau \le t\} + \mathbf{1}\{\tau > t\} M_t \\
&= M_{t\wedge\tau}.
\end{aligned}$$

Thus, $\mathbb{E}[M_{(t+1)\wedge\tau} \mid \mathcal{F}_t] = M_{t\wedge\tau}$ a.s., so $(M_{t\wedge\tau})$ is a martingale.

Step 2: Bounded τ. If $\tau \le T$ a.s., then $(T \wedge \tau) = \tau$ and by the martingale property of $(M_{t\wedge\tau})$ from Step 1,

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_{T\wedge\tau}] = \mathbb{E}[M_0],$$

which proves (i).

Step 3: Unbounded τ under uniform integrability. Let $\tau_n := \tau \wedge n$. Each τ_n is a bounded stopping time, so by (i),

$$\mathbb{E}[M_{\tau_n}] = \mathbb{E}[M_0] \qquad \text{for all } n \ge 1.$$

Moreover, $\tau_n \uparrow \tau$ and $M_{\tau_n} \to M_\tau$ a.s. (since $\tau_n = \tau$ for all large n on $\{\tau < \infty\}$). If $\{M_{\tau_n}\}_{n\ge 1}$ is uniformly integrable, then $\mathbb{E}[M_{\tau_n}] \to \mathbb{E}[M_\tau]$ as $n \to \infty$, hence

$$\mathbb{E}[M_\tau] = \lim_{n\to\infty} \mathbb{E}[M_{\tau_n}] = \mathbb{E}[M_0],$$

which proves (ii). The super/submartingale case follows identically. □

D.4 Gaussian Concentration Inequalities

Lemma D.7 (Gaussian small-ball bound from a variance lower bound)**.** *Let X be a real-valued Gaussian random variable with* $\mathrm{Var}(X) \ge v_{\min} > 0$*. Then for every $r > 0$,*

$$\mathbb{P}(|X| \le r) \;\le\; \sqrt{\frac{2}{\pi}}\,\frac{r}{\sqrt{v_{\min}}}. \tag{D.3}$$

Moreover, the bound is sharp in the sense that equality holds in (D.3) whenever $X \sim \mathcal{N}(0, v_{\min})$ and $r \downarrow 0$ (up to the first-order Taylor expansion).

Proof. Since X is Gaussian, there exist $m \in \mathbb{R}$ and $v > 0$ such that $X \sim \mathcal{N}(m, v)$. By assumption, $v = \mathrm{Var}(X) \ge v_{\min}$. Let $f_{m,v}$ denote the Lebesgue density of $\mathcal{N}(m, v)$, i.e.,

$$f_{m,v}(x) := \frac{1}{\sqrt{2\pi v}} \exp\!\left(-\frac{(x-m)^2}{2v}\right), \qquad x \in \mathbb{R}. \tag{D.4}$$

For any Borel set $A \subseteq \mathbb{R}$, the probability of X falling in A is given by

$$\mathbb{P}(X \in A) = \int_A f_{m,v}(x)\,dx, \tag{D.5}$$

where the integral is with respect to Lebesgue measure. We apply (D.5) to the interval $A = [-r, r]$:

$$\mathbb{P}(|X| \le r) = \mathbb{P}(X \in [-r, r]) = \int_{-r}^{r} f_{m,v}(x)\,dx. \tag{D.6}$$

The key observation is that (D.6) can be upper bounded by the length of the interval times the supremum of the density on that interval:

$$\int_{-r}^{r} f_{m,v}(x)\,dx \;\le\; \int_{-r}^{r} \sup_{y\in\mathbb{R}} f_{m,v}(y)\,dx = (2r)\,\sup_{y\in\mathbb{R}} f_{m,v}(y). \tag{D.7}$$

Thus, it suffices to upper bound $\sup_{y\in\mathbb{R}} f_{m,v}(y)$ in terms of $v_{\min}$. From (D.4), the exponential factor satisfies

$$0 < \exp\!\left(-\frac{(x-m)^2}{2v}\right) \le 1 \qquad \text{for all } x \in \mathbb{R},$$

with equality if and only if $x = m$. Therefore,

$$\sup_{y\in\mathbb{R}} f_{m,v}(y) = f_{m,v}(m) = \frac{1}{\sqrt{2\pi v}}. \tag{D.8}$$

Since $v \geq v_{\min}$, we have $\frac{1}{\sqrt{v}} \leq \frac{1}{\sqrt{v_{\min}}}$ and hence

$$\sup_{y \in \mathbb{R}} f_{m,v}(y) = \frac{1}{\sqrt{2\pi v}} \leq \frac{1}{\sqrt{2\pi v_{\min}}}. \tag{D.9}$$

Substituting (D.9) into (D.7) and then into (D.6) yields

$$\mathbb{P}(|X| \leq r) \leq (2r) \cdot \frac{1}{\sqrt{2\pi v_{\min}}} = \sqrt{\frac{2}{\pi}} \frac{r}{\sqrt{v_{\min}}},$$

which is exactly (D.3). For $X \sim \mathcal{N}(0, v_{\min})$, the density at 0 is $(2\pi v_{\min})^{-1/2}$, so as $r \downarrow 0$,

$$\mathbb{P}(|X| \leq r) = \int_{-r}^{r} f_{0,v_{\min}}(x)\, dx = 2r f_{0,v_{\min}}(0) + o(r) = \sqrt{\frac{2}{\pi}} \frac{r}{\sqrt{v_{\min}}} + o(r),$$

showing that the constant $\sqrt{2/\pi}$ is tight to first order. □

D.5 Information Theory

Lemma D.8 (Mutual information bounded by average pairwise KL). *Let $\mathcal{U} = \{u^{(1)}, \ldots, u^{(M)}\}$ be a finite set and let U be uniformly distributed on $\mathcal{U}$. For each $u \in \mathcal{U}$, let P_u be a probability distribution on a measurable space $(\mathcal{Z}, \mathcal{A})$, and let Z be a random variable such that (U, Z) is generated by $U \sim \mathrm{Unif}(\mathcal{U})$ and $Z \mid (U = u) \sim P_u$. Define the mixture distribution*

$$\bar{P} := \frac{1}{M} \sum_{u \in \mathcal{U}} P_u.$$

Then we have the following bound for the mutual information:

$$I(U; Z) = \frac{1}{M} \sum_{u \in \mathcal{U}} D_{\mathrm{KL}}(P_u \,\|\, \bar{P}) \leq \frac{1}{M^2} \sum_{u,v \in \mathcal{U}} D_{\mathrm{KL}}(P_u \,\|\, P_v). \tag{D.10}$$

Proof of Lemma D.8. Write $M := |\mathcal{U}|$. By definition of mutual information and the uniform prior,

$$I(U; Z) = \sum_{u \in \mathcal{U}} \frac{1}{M} \int \log\!\left(\frac{dP_u}{d\bar{P}}(z)\right) dP_u(z) = \frac{1}{M} \sum_{u \in \mathcal{U}} D_{\mathrm{KL}}(P_u \,\|\, \bar{P}), \tag{D.11}$$

where $\bar{P} = \frac{1}{M} \sum_{v \in \mathcal{U}} P_v$ is the marginal law of Z. Fix $u \in \mathcal{U}$. Since $\bar{P} = \frac{1}{M} \sum_{v \in \mathcal{U}} P_v$, we have for P_u-a.e. z,

$$\log \bar{p}(z) = \log\!\left(\frac{1}{M} \sum_{v \in \mathcal{U}} p_v(z)\right),$$

where p_v are densities w.r.t. any common dominating measure (the argument is identical using Radon–Nikodym derivatives). Because $\log(\cdot)$ is concave, Jensen's inequality yields

$$\log\!\left(\frac{1}{M} \sum_{v \in \mathcal{U}} p_v(z)\right) \geq \frac{1}{M} \sum_{v \in \mathcal{U}} \log p_v(z). \tag{D.12}$$

Equivalently, $-\log \bar{p}(z) \leq \frac{1}{M} \sum_{v \in \mathcal{U}} (-\log p_v(z))$. Plugging this into the KL definition gives

$$\begin{aligned} D_{\mathrm{KL}}(P_u \,\|\, \bar{P}) &= \int \big(\log p_u(z) - \log \bar{p}(z)\big)\, p_u(z)\, dz \\ &\leq \int \left(\log p_u(z) - \frac{1}{M} \sum_{v \in \mathcal{U}} \log p_v(z)\right) p_u(z)\, dz \\ &= \frac{1}{M} \sum_{v \in \mathcal{U}} \int \big(\log p_u(z) - \log p_v(z)\big)\, p_u(z)\, dz = \frac{1}{M} \sum_{v \in \mathcal{U}} D_{\mathrm{KL}}(P_u \,\|\, P_v). \end{aligned} \tag{D.13}$$

Averaging (D.13) over $u \sim \mathrm{Unif}(\mathcal{U})$ and using (D.11) yields

$$I(U; Z) = \frac{1}{M} \sum_{u \in \mathcal{U}} D_{\mathrm{KL}}(P_u \,\|\, \bar{P}) \;\leq\; \frac{1}{M} \sum_{u \in \mathcal{U}} \frac{1}{M} \sum_{v \in \mathcal{U}} D_{\mathrm{KL}}(P_u \,\|\, P_v) = \frac{1}{M^2} \sum_{u,v \in \mathcal{U}} D_{\mathrm{KL}}(P_u \,\|\, P_v),$$

which is exactly (D.10). $\square$