Convergence of Asynchronous Stochastic Gradient Descent for Polyak-Łojasiewicz Functions

Arijit Dey and Bahman Gharesifard, *Senior Member, IEEE*

***Abstract*— We analyze the convergence properties of Asynchronous Stochastic Gradient Descent (A-SGD) by explicitly incorporating the delay into the corresponding stochastic differential equation. Unlike prior approaches that rely on delay-compensation techniques or approximate stochastic differential delay equations, we propose a framework where delays are directly embedded into the optimization process. We establish exponential convergence guarantees under the Polyak–Łojasiewicz (PL) condition and derive a novel bound on steady-state behavior, demonstrating that A-SGD remains stable under realistic delay conditions. We validate our framework through experiments on synthetic functions satisfying the PL condition, showing improved convergence rates compared to existing methods. We further confirm the theoretical upper bound on an example of an over-parameterized neural network.**

I. Introduction

Many modern machine learning tasks involve extremely large datasets and models, which are too big or too time-consuming for a single computer to handle efficiently. A common solution is to split the work across multiple computational units, often called *workers*. Each worker receives some subset of the data or a portion of the computational task. By collaborating, the workers aim to find a parameter set that minimizes an objective function f. Formally, consider the problem of minimizing a function f, where $\theta \in \mathbb{R}^k$ are the parameters:

$$f(\theta) = \mathbb{E}_{x\sim\mathcal{D}}[f(\theta; x)].$$

Since the expectation is generally intractable (e.g., due to large or infinite dataset $\mathcal{D}$), we approximate it using a mini-batch of samples $\{x_i\}_{i=1}^{B}$. The standard Stochastic Gradient Descent (SGD) update is:

$$\theta_{t+1} = \theta_t - \eta_t \cdot \nabla_\theta \left(\frac{1}{B} \sum_{i=1}^{B} f(\theta_t; x_i) \right).$$

In synchronous SGD, we parallelize the computation using N workers. Each worker performs the following steps at iteration t: it receives the current parameters θ_t, which we call $\theta_t^{(j)}$. Next, worker j samples a mini-batch $\mathcal{B}_j$ and computes the *local gradient*:

$$g_t^{(j)} = \nabla_\theta \left(\frac{1}{|\mathcal{B}_j|} \sum_{x\in\mathcal{B}_j} f(\theta_t^{(j)}; x) \right).$$

Once all workers have computed their local gradients, they synchronize and compute the average gradient:

$$\bar{g}_t = \frac{1}{N} \sum_{j=1}^{N} g_t^{(j)}$$

The final step involves updating the model parameters θ:

$$\theta_{t+1} = \theta_t - \eta_t \cdot \bar{g}_t$$

Note that in a *synchronous* training setup, each worker j computes the gradient of f with respect to its parameter vector $\theta_t^{(j)} \in \mathbb{R}^k$, evaluates it on its local data, and then waits until *all* other workers have finished before aggregating their updates to form the new global parameter θ_t. Only after this joint update is completed does the next round of gradient computation begin. This synchronization mechanism guarantees that the local parameter $\theta_t^{(j)}$ of each worker coincides with the global parameter θ_t at time t. However, such rigid synchronization can be inefficient: if a single worker is delayed—due to slower hardware or network congestion—all other workers must remain idle until it completes.

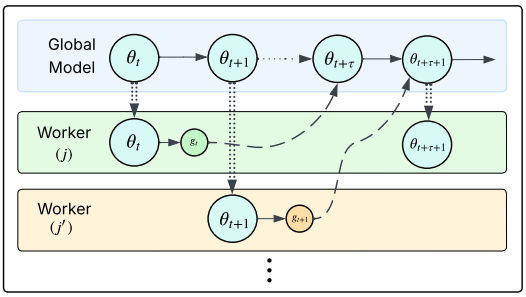

Fig. 1: A-SGD training process as discussed in [1][†].

Arijit Dey has an M.S. from the University of California, Los Angeles. Email: arijit.dey.1729@gmail.com

Bahman Gharesifard is with the Department of Mathematics and Statistics at Queen's University, Kingston, ON, Canada. Email: bahman.gharesifard@queensu.ca.

[†]Before a worker wants to add the gradient $\mathbf{g}_t$(based on the model at time t), to the global model, several other workers may already have added their gradients and the global model has been updated to $\theta_{t+\tau}$, where τ is the delay.

To mitigate this bottleneck, one can use A-SGD. In A-SGD, each worker computes its local gradient as soon as it finishes its current batch and sends the update to a central server (or parameter store) without waiting for others. This lets faster workers continue contributing updates immediately. However, it also means some workers may compute gradients using stale versions of the model. For example, a worker might base its gradient on the parameter θ_t at time t, but by the time the update is applied, the global model may already have advanced to $\theta_{t+\tau}$ for some delay τ.

Formally, let g_t be the gradient a worker computes using the model snapshot at time t. (See Figure 1) Because of asynchrony, the global model may have moved to $\theta_{t+\tau}$ before g_t is finally applied. This *delayed gradient* phenomenon makes the analysis of A-SGD more challenging than the synchronous case, because updates are no longer calculated from the *current* global model. The main goal of this work is to incorporate these inevitable delays into a modified framework—using an SDE–so that we can rigorously analyze and compensate for their impact: *Under reasonable delay assumptions, and with suitable learning rate, we prove convergence of A-SGD in expectation to a neighborhood of optima up to the persistent constant which scales quadratically with the delay.*

II. Previous Work

The continuous-time perspective of SGD [2], has inspired subsequent studies of delay effects and compensation in asynchronous optimization. For instance, [1] proposed an A-SGD algorithm with delay compensation, showing that careful adjustments can mitigate adverse effects of stale gradients. In a related direction, [3] examined noise-to-state stability in stochastic differential equations with persistent noise. The lock-free SGD framework *Hogwild!*, introduced in [4], pioneered subsequent research on asynchronous methods. An in-depth survey on delay and asynchrony in large-scale machine learning can be found in [5]. Additionally, real-world issues related to delayed updates in distributed deep networks are outlined in [6].

Our work is related to the convergence properties of A-SGD under delays. In [7], the error-runtime trade-off is examined and *K-async* algorithm and its variant *K-batch-async* are proposed , achieving balanced staleness and iteration speed. By relaxing the standard bounded-delay assumptions, recent work has identified conditions under which asynchronous methods can surpass synchronous counterparts in wall-clock efficiency. In a similar direction, [8] revisits the convergence of A-SGD with staleness, proving that for both convex and nonconvex objectives, convergence rates can be made independent of the maximum delay, depending only on the number of parallel workers. The analysis introduces a virtual iterate framework and employs delay-adaptive step sizes, showing that A-SGD can often outperform equivalent synchronous minibatch SGD in wall-clock time. Finally, in [9] *Ringmaster ASGD* is proposed which is a fully asynchronous scheme that filters out overly stale gradients via a delay threshold R and applies updates immediately on "fresh-enough" gradients, effectively prioritizing fast workers while disregarding outdated information. The authors prove it is the first A-SGD with optimal time complexity under arbitrary heterogeneous worker speeds–matching known lower bounds–thereby giving wall-clock optimality guarantees rather than iteration-only rates. This establishes a modern theoretical foundation that complements results under the Polyak–Łojasiewicz condition by removing overly pessimistic delay constraints.

In discrete time, [10] studied distributed and Federated learning to show improved convergence of $\mathcal{O}(\sigma^2\epsilon^{-2} + \sqrt{\tau_{\max}\tau_{\text{avg}}}\epsilon^{-1})$, where σ^2 is the noise variance, $\tau_{\max}$, τ_{avg} are the maximum and average delay, respectively, and ϵ is the suboptimality gap. Refining the bounds in [11], [12] establishes an improved rate of $\mathcal{O}(L\tau_{\text{avg}}/\epsilon + \sigma^2\epsilon^{-2})$ for L-smooth convex functions and further derives sharper iteration complexity bounds for proximal incremental aggregated gradient methods. Meanwhile, [13] established a rate of $\mathcal{O}(1/\sqrt{K})$ under bounded delay (where K is the number of iterations), and [14] extended this result to unbounded delays.

Another direction is the *fixed-time gradient dynamics* framework for continuous-time optimization [15], where time-varying coefficients enable convergence within a predetermined horizon, independent of initial conditions. These ideas motivate the incorporation of delay-awareness into stochastic optimization. Likewise, [16] explored block-coordinate descent under the Polyak-Łojasiewicz condition in discrete time.

A separate line of work [17] used SDDEs to approximate A-SGD, mainly bounding discretization errors and deriving rates via moment and energy analyses. Our approach departs from such approximations by embedding delay directly into the SDE, removing the need for auxiliary compensation techniques. Under the Polyak–Łojasiewicz condition, we prove exponential convergence and clarify the role of noise and learning rate. Unlike [17], which links noise to batch sizes and delays, our framework highlights their direct proportionality to the learning rate, yielding clearer insight into optimization stability.

On a completely different note, [18] addresses generalization in asynchronous training. Specifically, the authors prove non-vacuous generalization bounds for A-SGD under much weaker conditions than Lipschitz continuity. They establish on-average model stability results for A-SGD and quantify how factors such as delay, initialization, and sample size affect the excess generalization error. In [19], the authors explore fully asynchronous Local-SGD in large-scale language model training. The authors show that naive asynchronous training can suffer slower convergence due to stale gradients with momentum. They propose a delayed Nesterov momentum correction and adaptive local step schedules to counteract staleness.

III. Our Contributions

In this work, a continuous-time stochastic framework is developed to model A-SGD with delayed updates. By embedding gradient delays directly into the SDE–instead of relying on classical SDDEs or discrete compensation techniques–the approach naturally captures the role of delays and the accumulation of noise. Under the Polyak-Łojasiewicz (PL)

condition, this formulation yields an exponential convergence bound that quantifies how delays, learning rates, and stochastic noise jointly affect convergence to a steady-state neighborhood of the optimum. The analysis further reveals a direct linear dependence between the steady-state noise level and the learning rate, underscoring the practical trade-off between faster early-stage convergence and a larger long-term error floor. In contrast to prior treatments that impose rigid assumptions on bounded delays or gradient variances, the proposed continuous-time framework accommodates a broader range of delay processes and offers stronger stability guarantees. Lastly, it aligns well to the principles of input-to-state stability (ISS) and noise-to-state stability (NSS) in stochastic control systems, demonstrating that A-SGD converges to a noise-dependent neighborhood of the optimal parameter at an exponential rate despite the presence of delays. We verify our findings, in particular, on an example scenario of an over-parameterized neural network[2]

Auxiliary Implications: Over-parameterized neural networks often satisfy a PL-like condition [20], leading to more rapid convergence [21], and sometimes global convergence [22], [23]. Our framework remains valid when delays vanish, reproducing standard SGD analysis. This connects the known benefits of over-parameterization with A-SGD by linking delay effects to faster convergence under PL-like assumptions. In [24], the study of over-parameterized regimes is extended to multi-layer neural networks, showing that for sufficiently wide architectures, stochastic gradient descent can reach global minima of deep training objectives in polynomial time. The authors demonstrate that wide networks exhibit dynamics closely approximated by neural tangent kernels, leading to linear convergence of SGD under mild assumptions. This result complements the Polyak–Łojasiewicz framework by establishing that large network width can inherently induce PL-like optimization landscapes, even for highly non-linear deep models.

IV. Mathematical Formulation

In this work, we analyze the parameter update for each individual worker as a continuous-time SDE of the form:

$$d\theta(t) = -\eta(t)\nabla f(\theta(t-\tau(t)))dt + \eta(t)\sigma(\theta(t-\tau(t)))dW(t), \tag{1}$$

where $f : \mathbb{R}^k \mapsto \mathbb{R}$ is a real-valued differentiable function, $\eta(\cdot)$ is a learning rate to be specified later, $\theta(t) \in \mathbb{R}^k$, $\sigma^2(\theta(t))$ is the noise variance, $0 \leq \tau(t) \leq T < \infty$ is the delay and $W \in \mathbb{R}^k$ is a k-dimensional Brownian Motion. Let us first explain the meaning of the learning rate which we have used in the above equation.

Remark IV.1. (**Learning rate**): Suppose we have the stochastic differential equation:

$$dx = \eta\left(f(x)\,dt + g(x)\,dW(t)\right),$$

where η is a learning rate taken to be a constant, and $W(t)$ is standard Brownian motion. We define a new time variable t' such that

$$t' = \eta t \quad \Rightarrow \quad dt = \frac{dt'}{\eta},$$

and a new Brownian motion $\widetilde{W}(t')$ by

$$\widetilde{W}(t') = \sqrt{\eta}\,W(t),$$

so that

$$dW(t) = \frac{d\widetilde{W}(t')}{\sqrt{\eta}},$$

thanks to the scaling property of the Brownian motion. Substituting into the original equation:

$$\begin{aligned} dx &= \eta\left(f(x)\,\frac{dt'}{\eta} + g(x)\,\frac{d\widetilde{W}(t')}{\sqrt{\eta}}\right) \\ &= f(x)\,dt' + \sqrt{\eta}\,g(x)\,d\widetilde{W}(t'). \end{aligned}$$

The original SDE

$$dx = \eta\left(f(x)\,dt + g(x)\,dW(t)\right)$$

can be rewritten, after a change of time and Brownian motion, as

$$dx = f(x)\,dt' + \sqrt{\eta}\,g(x)\,d\widetilde{W}(t'),$$

which clarifies the claim. •

Throughout, for simplicity, we assume that $\sigma(\theta(t)) \equiv \sigma$ is a constant scalar; this can easily be generalized[3]. We therefore have that

$$d\theta(t) = -\eta(t)\nabla f(\theta(t-\tau(t)))dt + \eta(t)\sigma dW(t). \tag{2}$$

The main objective of this work is to analyze the convergence properties of (2).

V. Convergence for Polyak-Łojasiewicz Functions

We define a random variable along (2), which we term *burn-in time*, as:

$$\tau_M \coloneqq \inf\{t \geq 0 : f(\theta(t)) - f(\theta^*) \leq M\}. \tag{3}$$

Also, we define the exit time:

$$\tau_R^{\uparrow} \coloneqq \inf\{t \geq 0 : f(\theta(t)) - f(\theta^*) > R\}, \tag{4}$$

where $R > 0$. Consider a function $f \in \mathcal{C}^2$. We now state some of the assumptions that we use throughout.

Assumption 1. (**Polyak-Łojasiewicz condition**): There exists $\mu > 0$ such that for all $\theta \in \mathbb{R}^k$,

$$\frac{1}{2}\|\nabla f(\theta)\|^2 \geq \mu\left(f(\theta) - f(\theta^*)\right). \tag{5}$$

Assumption 2. (**Lipschitz gradients**): The gradient ∇f is L-Lipschitz continuous:

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq L\|\theta_1 - \theta_2\|, \quad \forall \theta_1, \theta_2 \in \mathbb{R}^k. \tag{6}$$

Assumption 3. (**Bounded delay**): The delay τ satisfies $\tau(t) \leq T < \infty$ for all $t \geq 0$.

[2]Our codes can be found in https://github.com/arijitcodespace/Asynchronous-SGD.

[3]for example, with $\sigma \in \mathbb{R}^k$, we replace σ^2 by $\mathrm{tr}\,(\sigma\sigma^\top)$.

Note that Assumption 3 is realistic, and although not uniformly assumed, for instance in [14], it has been widely used in the literature, e.g., [13]).

Assumption 4. (**Local gradient bound**): There exist $M, G_M \in \mathbb{R}$ such that $f(\theta)-f(\theta^*) \leq M \Rightarrow \|\nabla f(\theta)\| \leq G_M$, where,

$$G_M = \sup_{\theta: f(\theta)-f(\theta^*)\leq M} \|\nabla f(\theta)\| < \infty.$$

The finiteness of G_M is justified in the following assumption. This implication holds deterministically for all $\theta \in \mathbb{R}^k$.

Assumption 5. (**Bounded sublevel sets**): For all $R < \infty$, $\Omega_R := \{\theta : f(\theta) - f(\theta^*) \leq R\}$ is bounded. Thus, following from Assumption 2, we have that

$$G_R := \sup_{\theta\in\Omega_R} \|\nabla f(\theta)\| < \infty.$$

Assumption 6. (**Learning rate**): We assume that the learning-rate $\eta : [0, t_f] \mapsto \mathbb{R}_{>0}$ satisfies the following properties: Given a delay bound $T > 0$, a final time $0 < t_f < \infty$, and any $(K + 1)$ partition $(K \geq 1)$

$$0 = t_{-1} \leq t_0 < t_1 < \cdots < t_K = t_f,$$

which satisfies $t_{j+1} - t_j \geq T$ for $j \in \{0, \ldots, K - 1\}$, then,

1) *(Continuity)*: η is continuous on $[0, t_f]$;
2) *(Delayed monotonicity)*: η is non-increasing on the interval $[t_0, t_f]$.

Note that we do not require any monotonicity property on the interval $[0, t_0)$. The main result of this work is as follows:

Theorem 1. (**Convergence of A-SGD under PL condition**): Under Assumptions 1 – 5, and assuming that $\eta(t) = \eta$, where $\eta \in \mathbb{R}_{>0}$, the trajectories of (2) satisfy:

$$\mathbb{E}[f(\theta(t)) - f(\theta^*)] \leq (\mathbb{E}\left[f(\theta(\tau_M))\right] - f(\theta^*))\, e^{-\eta\mu(t-\tau_M)} + \kappa\left(1 - e^{-\eta\mu(t-\tau_M)}\right), \quad (7)$$

where

$$\kappa = \frac{\eta^2}{\mu}\left(L^2T^2\left(G_{2M}\right)^2 + L^2\sigma^2T\right) + \frac{\eta\sigma^2kL}{2\mu}$$

is called the *noise-margin*.

In practice, e.g., in neural networks, one often employs a time-varying learning rate. In what follows, we provide an upper bound for A-SGD in such settings.

Theorem 2. (**Convergence with time-varying** η): Let $t_f > 0$ and suppose that Assumptions 1 – 5 hold, and the learning rate η satisfies Assumption 6 with the specific

$$t_0 := \tau_M,$$

where τ_M is as in (3). For $K \geq 1$, consider the $(K + 1)$-partition of $[0, t_f]$ as presented by Assumption 6. The parameter trajectory governed by the SDE (2) satisfies the convergence bound:

$$\mathbb{E}[f(\theta(t)) - f(\theta^*)] \leq \mathbb{E}[f(\theta(t_0)) - f(\theta^*)]e^{-\mu\int_{t_0}^t \eta(s)ds} + \kappa(K) \quad (8)$$

for $t \geq t_0$ where,

$$\kappa(K) := \sum_{j=0}^{K-1} \frac{e^{-\mu\eta_{\max,\,j}\,t_0}}{\mu\eta_{\max,\,j}} \left(e^{\mu\left[\eta_{\max,\,j}-\eta_{\min,j}\right]t_{j+1}} e^{\mu\eta_{\min,j}t_j} - e^{\mu\left[\eta_{\max,\,j}+\eta_{\min,j}\right]t_j} e^{-\mu\eta_{\min,j}t_{j+1}} \right) \left(\frac{L^2}{2}\left[2\eta_{\max,j}^3T^2\left(G_{2M}\right)^2 + 2\eta_{\max,j}^3\sigma^2T\right] + \frac{\sigma^2}{2}\left|\max_{1\leq s\leq t_f} \operatorname{tr}\left[\nabla^2 f(\theta(s))\right]\right| \eta_{\max,j}^2 \right),$$

where $\eta_{\min,j}$ and $\eta_{\max,\,j}$ are, respectively, the minimum and maximum of η on the interval $[t_j, t_{j+1}]$, for all $j \in \{0, \ldots, K - 1\}$.

It is worth pointing out that by taking $K = 1$ and learning rate to be a constant $\eta(t) = \eta$, and further upper bounding

$$\left|\max_{1\leq s\leq t_f} \operatorname{tr}(\nabla^2(f(\theta(s)))\right| \leq kL,$$

then we recover Theorem 1 from Theorem 2. The result has several other implications, which we discuss below.

Remark V.1. (**Exponential descent under PL**): Once the learning rate decays to a small value, the dominant term in both (7) and (8) is

$$(f(\theta(t_0)) - f(\theta^\star))\, \exp\Big(-\mu\int_{t_0}^t \eta(s)\, ds\Big).$$

Therefore, for objective functions which satisfy the PL condition, the trajectories of A-SGD contract (with high probability) at the same exponential rate as *the ones for* synchronous SGD. The impact of the delay is shortening the admissible upper bound on η and as such, it does not affect the rate of contraction. •

Remark V.2. (**Noise margin versus step-size tail**): For *constant* or *block-wise constant* schedules, the accumulated variance of the Brownian term, $\int_{t_0}^{\infty} \eta^2(s)\, ds$, diverges. The residual integral in the proof (See Appendix IV) therefore stabilizes at a strictly positive value, yielding the fixed "noise-margin" κ in Theorem 2. •

Remark V.3. (**Practical speed-accuracy trade-off**): We observe from Theorem 2 that the learning rate only needs to decrease after some time t_0. As shown in the proof (Appendix IV), the interval $[0, t_0]$ corresponds to a *burn-in* phase during which the value gap $f(\theta(t)) - f(\theta^*)$ decays rapidly. In practice, one can set a relatively large learning rate in this phase so that the gap closes quickly (in fewer iterations) before the decay schedule takes effect. This drives the error rapidly into a ball of radius $\mathcal{O}(\eta\sigma^2)$ (see the dominant term in κ for $\eta(t) < 1$). The subsequent decay of the learning rate then gradually shrinks the radius of this ball, albeit at the cost of slower convergence in the late stage. •

Remark V.4. (**Delay robustness**): The coefficients of the delay terms (containing T) in the expression of κ reads:

$$\frac{L^2}{2}\left[2\eta_{\max,j}^2 T^2 \left(G_{2M}\right)^2 + 2\eta_{\max,j}^2 \sigma^2 T\right].$$

Once the learning rate begins to decay, the above term contributes less to the overall value of κ. Consequently, moderate asynchronous staleness affects only the constants in the analysis, without altering the qualitative behavior described above. •

VI. Experiments and Results

In this section, we present two kinds of experiments, one with a synthetic objective function and one with over-parameterized neural networks. We start our experiments from the simpler case of the synthetic objective function. We denote using $\theta^{(t)}$, the t-th iterate of θ. For brevity, we omit the algorithms for A-SGD simulation, Hessian Trace, and L-smoothness estimation for neural networks, which are fairly standard algorithms used in the literature and can be found in our repository listed before.

VI-A Synthetic Objective Function

We choose the objective function:

$$f(\theta) = \theta^\top Q\theta + \epsilon \sin(w^\top \theta),$$

for $Q \succ \mathbf{0}$, $\epsilon \geq 0$ and $w \neq \mathbf{0}$. This function satisfies the Polyak-Łojasiewicz condition for moderate values of ϵ and is additionally strongly convex for small values of ϵ. It can be verified that when f is strongly convex, the strong convexity constant μ (which is also the PL-constant) is given by (See Appendix V for a proof),

$$\mu = 2\lambda_{\min}(Q) - \epsilon\|w\|^2.$$

Additionally, the L-smoothness constant is also known and is given by,

$$L = 2\lambda_{\max}(Q) + \epsilon\|w\|^2,$$

where $\lambda_{\min}(Q)$ and $\lambda_{\max}(Q)$ are the smallest and largest eigenvalues of Q, respectively.

Setting: Throughout this section, we focus solely on a strongly convex objective and compare our bound to others relevant in the literature. We verify the upper bound proposed by Theorem 2 for a time-varying learning rate schedule. The results are plotted in Figure 2 and Figure 3. For [17, Theorem 4.2], to obtain an upper bound of $f(\theta^{(t)}) - f(\theta^*)$, we simply multiply their upper bound by a factor of $L/2$.

For simplicity of simulation, in the case of synthetic functions, we further upper bound κ, given in Theorem 2 by upper bounding

$$\left|\max_{\theta\in\mathbb{R}^k} \operatorname{tr}(\nabla^2 f(\theta))\right| \leq kL.$$

We do this to avoid calculating the maximum trace of the Hessian throughout the simulation.

Results: We perform two experiments, the parameters of which are given in Table I and Table II respectively. The results are shown in Figure 2 and Figure 3. It should be noted that the trajectories shown in the figures are the empirical mean trajectory across a number parallel simulations. For the second experiment, we chose a time-varying noise variance $\sigma^2(t)$ instead of a fixed variance, and to plot the upper bound, we used the mean variance across iterations. The trajectory and the corresponding upper bound obtained, therefore, improves substantially the ones obtained in [17, Theorem 4.2]. For the learning rate schedules, see Figure 2 (Right) and Figure 3 (Right), we keep the learning rate fixed during the initial phase and start the decay only when the difference between optimum and current objective value has fallen below the threshold given in Table I and Table II.

Remark VI.1. (**Variance schedule in Experiment 2**): For Experiment 2, we chose a variance schedule instead of a fixed variance, i.e., the variance is time-varying:

$$\sigma^2(t) \equiv \operatorname{tr}(\sigma(t)\sigma(t)^\top) = 10^{-6}\left(e^{-5\times10^{-3}i} + 10^{-10}\right),$$

where i denotes the index of iteration; $i \in \{1, 2, \dots\}$. But while calculating the upper bound, we used the empirical mean of the schedule defined above:

$$\bar{\sigma}^2 = \frac{10^{-6}}{t_f}\left(\sum_{i=0}^{t_f-1} e^{-5\times10^{-3}i} + 10^{-10}\right).$$

•

VI-B "Over-parameterized" Neural Networks

In this section, we validate our upper bound given by Theorem 2 on over-parameterized neural networks. Roughly speaking, a neural network is called "over-parameterized" if the number of parameters is larger than the number of data points. From [20], we know that this class of neural networks tends to satisfy the PL condition on the loss landscape.

Setting: We chose the CIFAR-100 [25] dataset for our experiments. This contains $50,000$ RGB images, each of size $32 \times 32 \times 3$. We choose our neural network to be VGG-16 [26]. To estimate the gradient variance, we use the gradient of a 4096-batch as a reference when obtaining the estimated gradient with `batch_size` samples. During the burn-in time, we opt for a learning rate schedule with a warm-up that turned out to be empirically sound. The PL-constant μ is estimated by using:

$$\mu = \min_\theta \frac{\|\nabla f(\theta)\|^2}{2f(\theta)},$$

since $f(\theta^*) = 0$. In other words, the largest slope of the line above which lie all the points in scatter-plot of $f(\theta)$ vs $\|\nabla f(\theta)\|^2$. It must be noted that we train a auxiliary network (with the same architecture as in the experiments) while estimating the PL-constant.

To estimate the minimum L-smoothness constant, which is defined as the largest eigenvalue of the Hessian for twice differentiable functions, we use the power iteration method to determine the largest eigenvalue of the Hessian at each iteration during a proxy training and then take the minimum over all the largest eigenvalues. We also estimate the trace of the Hessian, using Hutchinson's method [27]. It must be noted

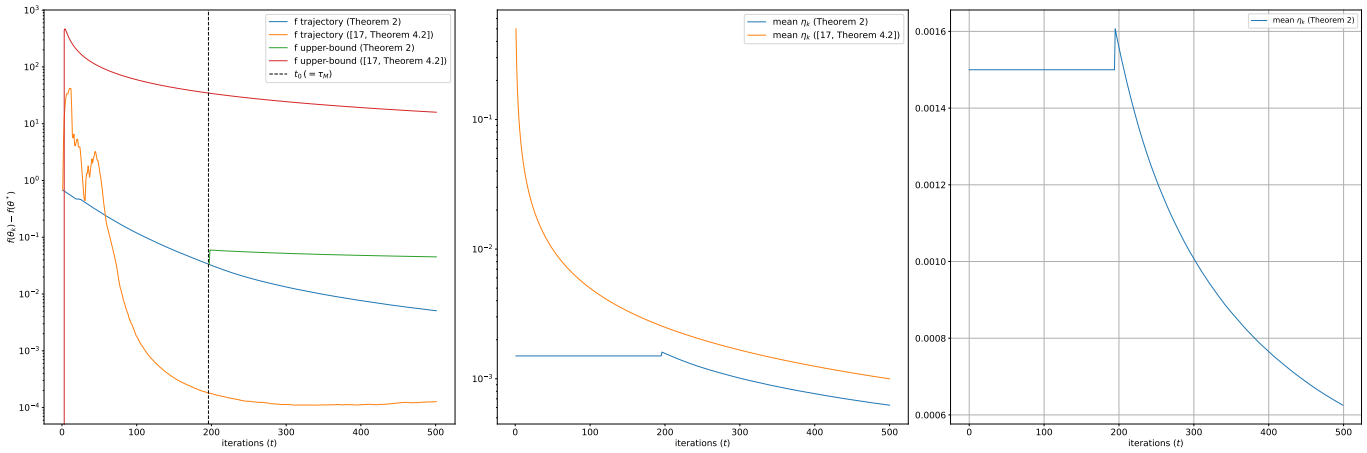

Fig. 2: Experiment 1: Trajectories for A-SGD and corresponding upper bounds for Theorem 2 and [17, Theorem 4.2] (Left); learning rate schedules in log-scale (Middle); learning rate for Theorem 2 (linear scale) (Right).

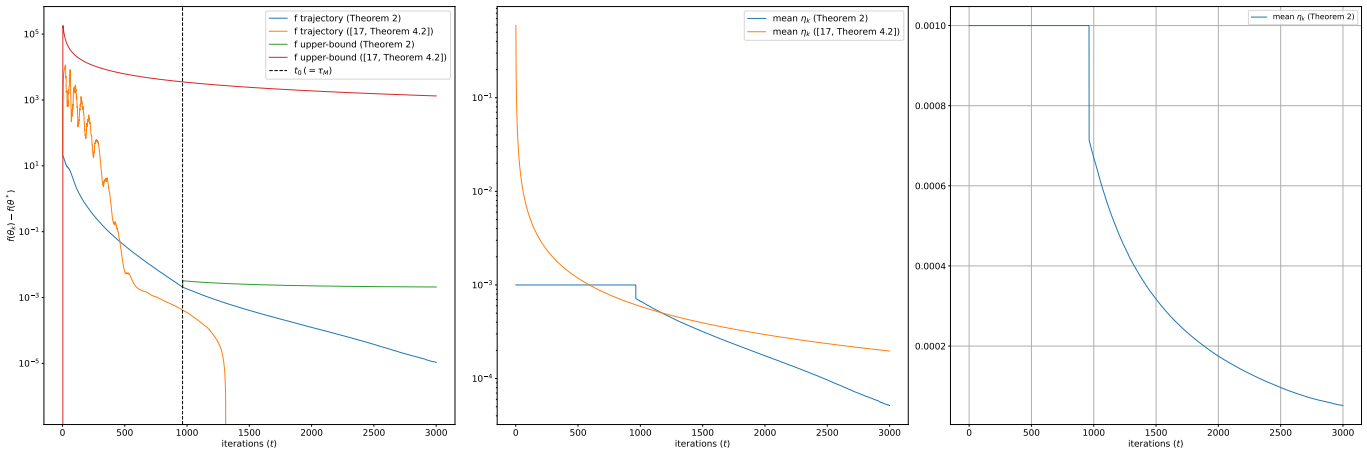

Fig. 3: Experiment 2: Trajectories for A-SGD and corresponding upper bounds for Theorem 2 and [17, Theorem 4.2] (Left) learning rate schedules, Upper and lower bounds (Middle); learning rate for Theorem 2 (linear scale) (Right).

Parameter	**Theorem 2**	**[17, Theorem 4.2]**
Number of parallel A-SGDs	100	100
Initial norm $\|\theta_0\|$	0.5	0.5
Dimension	50	50
$\lambda_{\max}(Q)$	5.0	5.0
$\lambda_{\min}(Q)$	1.0	1.0
ϵ	10^{-4}	10^{-4}
iterations (t_f)	500	500
Maximum Delay (iterations)	20	20
σ^2 or $\mathrm{tr}(\sigma\sigma^\top)$	10^{-8}	10^{-8}
fraction of initial value gap after which learning rate decay starts	0.05	–

TABLE I: Comparison of Parameters in Theorem 2 and [17, Theorem 4.2], Experiment 1.

that the L-smoothness constant, the PL-constant μ, does not depend on the kind of optimizer used. It only depends on the loss landscape, and hence, to get near the optimum, we used the Adam optimizer with momentum and estimated these parameters. However, the term $\max_{1\le s\le t_f} \mathrm{tr}(\nabla^2 f(\theta(s)))$ does depend on the trajectory the optimizer takes and, consequently, is also dependent on the choice of the optimizer itself. To empirically estimate the trace and to reduce computation overhead by doing two A-SGDs (one to estimate the trace and one for simulation), we do the trace calculation along with the calculation of μ and L. We note that this is an empirical estimate, and to be taken with that in mind.

Finally, to simulate A-SGD, we use momentum during training. This is because large networks like VGG-16 are

Parameter	Theorem 2	[17, Theorem 4.2]
Number of parallel A-SGDs	200	200
Initial norm $\|\theta_0\|$	2.0	2.0
Dimension	100	100
$\lambda_{\max}(Q)$	10.0	10.0
$\lambda_{\min}(Q)$	1.0	1.0
ϵ	10^{-2}	10^{-2}
iterations (t_f)	3000	3000
Maximum Delay (iterations)	50	50
σ^2 or $\mathrm{tr}(\sigma\sigma^\top)$	See Remark VI.1	See Remark VI.1
fraction of initial value gap after which learning rate decay starts	5×10^{-4}	–

TABLE II: Comparison of Parameters in Theorem 2 and [17, Theorem 4.2], Experiment 2.

Parameter	Theorem 2
Dimension (k)	138M
Maximum Delay (iterations)	10
σ^2 or $\mathrm{tr}(\sigma\sigma^\top)$	10^{-8}
iterations (t_f)	25000
learning rate schedule during burn-in	$\begin{cases} 2\times10^{-8}t + 10^{-4}, & t \le 10000 \\ 3\times10^{-4}\exp\left(-4\times10^{-5}(t-10000)\right), & t > 10000 \end{cases}$ $t \to$ index of iteration; $t \in \{1, 2, \dots\}$
$\max \mathrm{tr}(\nabla^2 f(\theta))$	26250
μ	0.1273
L	33
f^*	0
fraction of initial value gap after which burn-in stops	0.3
G_{2M}	4.14
batch size	1024
Apply Momentum	True

TABLE III: Parameter Settings for Experiment 3.

known to settle/oscillate at poor local minima while trained without momentum-based optimizers [28], [29] This is *not* in contradiction to our theory because momentum-based optimization leads to better minima and thus the upper bound given by Theorem 2 still holds with some slack. Now, to simulate Theorem 2 we partition the time interval $[0, t_f]$ into a K-partition $t_0, t_1, \dots, t_{K-1}$ such that $t_{j+1} - t_j > T$. We perform two experiments on this network, the parameters of which are provided in Table III.

Results: Figure 4 shows the upper bound and loss trajectory under the settings of Table III. The upper bound is valid after the process has reached a preset threshold (given in Table III), i.e., after the process has entered the set Ω_M for a chosen M. Since our bound is derived for vanilla A-SGD, intuitively and practically, it makes sense that we observed the bound obeys momentum-based A-SGDs; this is because the loss trajectories for momentum-based techniques tend to remain lower than those without momentum (at least for convex objectives).

Remark VI.2. (**Adaptive local bounds and a lower-envelope certificate**): Our analysis is local: it certifies the trajectory once it enters a chosen sublevel set Ω_M. This allows practitioners tailor M or G_M, to the phase of training, compute a small family of inexpensive certificates, and then take their pointwise lower envelope to produce a tight, global-in-time esti-

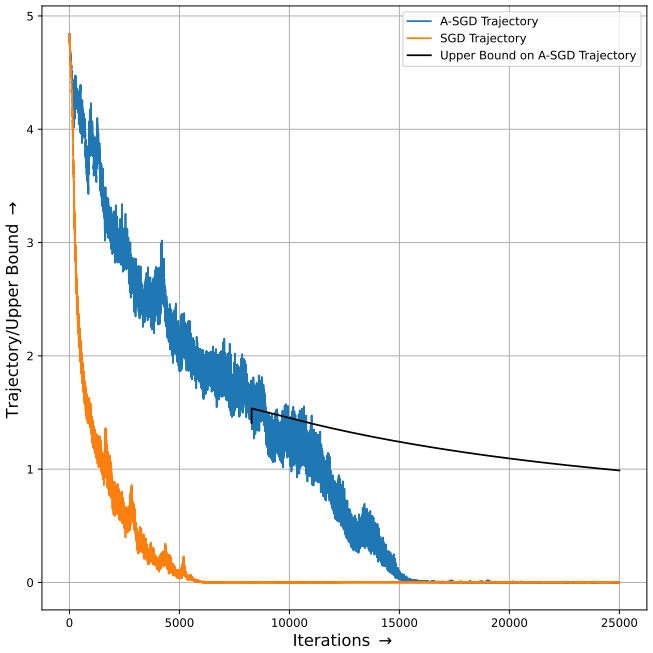

Fig. 4: Experiment 3: Upper bound (Black), loss trajectory of A-SGD (Blue), and loss trajectory of SGD (Orange) of the neural network.

mate. In contrast, fully non-local bounds (e.g., [17, Theorem 4.2]) are necessarily more conservative and may not capture the late-stage decay seen in practice. Operationally, one can pre-compute bounds at K checkpoints–from initialization up to the initial gradient norm–and use the lower envelope as a robust worst-case curve. See Figure 5. •

VII. Conclusion and Future Scope

In this work, we derived an approach to analyze the performance of A-SGD on the class of Polyak–Łojasiewicz functions and demonstrated its performance without the explicit need for delay compensation. Using an appropriate choice of the learning rate schedule, we provided an upper bound for the underlying A-SGD process that is tighter than the previously known upper bound when restricted to the class of strongly convex functions. We demonstrated the effectiveness of the

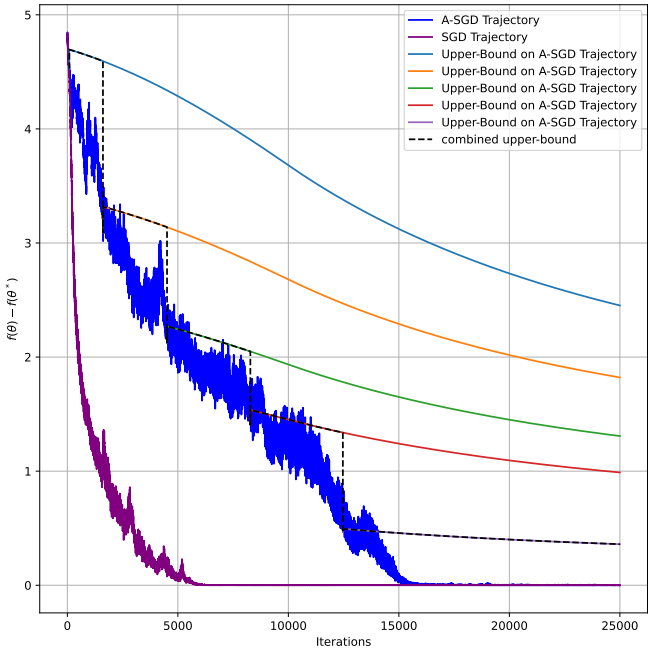

Fig. 5: Minimum/combined upper bound (dashed) of the five calculated upper bounds (solid).

bound in a deep neural network experiment in an over-parameterized regime, which tends to satisfy a variant of the PL-condition. Furthermore, we showed in an experiment that our bound can be applied to momentum-based optimization settings. Further refinements and tightening of the proposed bound require knowledge of the structure imposed on the loss landscape for deep neural networks. To this end, we close the discussion with a broad future scope to expand the development to other, wider classes of functions—for example, the Kurdyka–Łojasiewicz class of functions.

[1] S. Zheng, Q. Meng, T. Wang, W. Chen, N. Yu, Z.-M. Ma, and T.-Y. Liu, "Asynchronous stochastic gradient descent with delay compensation," in *International Conference on Machine Learning*. PMLR, 2017, pp. 4120–4129.

[2] J. Sirignano and K. Spiliopoulos, "Stochastic gradient descent in continuous time," *SIAM Journal on Financial Mathematics*, vol. 8, no. 1, pp. 933–961, 2017.

[3] D. Mateos-Nunez and J. Cortes, "p th moment noise-to-state stability of stochastic differential equations with persistent noise," *SIAM Journal on Control and Optimization*, vol. 52, no. 4, pp. 2399–2421, 2014.

[4] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[5] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[6] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang *et al.*, "Large scale distributed deep networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[7] S. Dutta, J. Wang, and G. Joshi, "Slow and stale gradients can win the race," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 1012–1024, 2021.

[8] K. Mishchenko, F. Bach, M. Even, and B. E. Woodworth, "Asynchronous sgd beats minibatch sgd under arbitrary delays," *Advances in Neural Information Processing Systems*, vol. 35, pp. 420–433, 2022.

[9] A. Maranjyan, A. Tyurin, and P. Richtárik, "Ringmaster asgd: The first asynchronous sgd with optimal time complexity," *arXiv preprint arXiv:2501.16168*, 2025.

[10] A. Koloskova, S. U. Stich, and M. Jaggi, "Sharper convergence guarantees for asynchronous sgd for distributed and federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 202–17 215, 2022.

[11] A. Cohen, A. Daniely, Y. Drori, T. Koren, and M. Schain, "Asynchronous stochastic optimization robust to arbitrary delays," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9024–9035, 2021.

[12] H. R. Feyzmahdavian and M. Johansson, "Asynchronous iterations in optimization: New sequence results and sharper algorithmic guarantees," *Journal of Machine Learning and Research*, vol. 24, no. 158, pp. 1–75, 2023.

[13] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[14] X. Zhang, J. Liu, and Z. Zhu, "Taming convergence for asynchronous stochastic gradient descent with unbounded delay in non-convex learning," in *IEEE Conf. on Decision and Control*. IEEE, 2020, pp. 3580–3585.

[15] L. T. Nguyen, X. Yu, A. Eberhard, and C. Li, "Fixed-time gradient dynamics with time-varying coefficients for continuous-time optimization," *IEEE Transactions on Automatic Control*, vol. 68, no. 7, pp. 4383–4390, 2022.

[16] K. Yazdani and M. Hale, "Asynchronous parallel nonconvex optimization under the polyak-łojasiewicz condition," *IEEE Control Systems Letters*, vol. 6, pp. 524–529, 2021.

[17] L. He, Q. Meng, W. Chen, Z.-M. Ma, and T.-Y. Liu, "Differential equations for modeling asynchronous algorithms," *arXiv preprint arXiv:1805.02991*, 2018.

[18] X. Deng, T. Sun, S. Li, D. Li, and X. Lu, "Stability and generalization of asynchronous sgd: Sharper bounds beyond lipschitz and smoothness," *Advances in Neural Information Processing Systems*, vol. 37, pp. 7675–7713, 2024.

[19] B. Liu, R. Chhaparia, A. Douillard, S. Kale, A. A. Rusu, J. Shen, A. Szlam, and M. Ranzato, "Asynchronous local-sgd training for language modeling," *arXiv preprint arXiv:2401.09135*, 2024.

[20] C. Liu, L. Zhu, and M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks," *Applied and Computational Harmonic Analysis*, vol. 59, pp. 85–116, 2022.

[21] K. A. Sankararaman, S. De, Z. Xu, W. R. Huang, and T. Goldstein, "The impact of neural network overparameterization on gradient confusion and stochastic gradient descent," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8469–8479.

[22] S. Arora, N. Cohen, N. Golowich, and W. Hu, "A convergence analysis of gradient descent for deep linear neural networks," *arXiv preprint arXiv:1810.02281*, 2018.

[23] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 322–332.

[24] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 242–252.

[25] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] M. F. Hutchinson, "A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines," *Communications in Statistics-Simulation and Computation*, vol. 18, no. 3, pp. 1059–1076, 1989.

[28] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*. pmlr, 2013, pp. 1139–1147.

[29] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.

[30] K. Itô, "109. stochastic integral," *Proceedings of the Imperial Academy*, vol. 20, no. 8, pp. 519–524, 1944.

Appendix I
Bounded Burn-in Time

We first state some lemmas that will be useful in proving Theorem 1 and Theorem 2.

Lemma 1. (**Bounded burn-in time**): Under Assumptions 1–3 and Assumption 5, there exists a deterministic constant $M_\star$ such that for every $M \geq M_\star$ and a constant learning rate η,

$$\mathbb{E}[\tau_M] \leq T_{\text{burn}} < \infty.$$

We analyze burn-in by working inside a *safe* sublevel set. First, an exponential supermartingale shows that leaving this set within any fixed window is unlikely. Second, while we remain inside, the PL condition gives a uniform negative drift of the objective, and the delay contributes only a bounded penalty; thus, over a short, fixed block of time, there is a constant, strictly positive chance of dropping below level M. Finally, we tile time into such blocks and restart the argument at block boundaries (invoking strong Markovianity). Each block is an *attempt* with the same success probability, so the number of attempts has a geometric tail, and the expected burn-in time is finite.
Now we begin the proof.

Proof. Let,

$$g_t \coloneqq \nabla f(\theta(t)),$$

and,

$$t_R \coloneqq t \wedge \tau_R^{\uparrow}.$$

Also, let R be such that $R > f(\theta(0)) - f(\theta^*)$.
Delay-Window Bounds: For any $t \geq 0$, we have

$$\begin{aligned}\left\| \int_{t-\tau(t)}^{t} g_{s-\tau(s)} ds \right\|^2 &\leq (\tau(t))^2 \cdot \sup_{u \in [t-\tau(t),t]} \left\| g_{u-\tau(u)} \right\|^2 \\ &\leq T^2 G_R^2 \end{aligned} \tag{9}$$

where G_R is as defined in Assumption 5.
Recent-Iterate Difference in Expectation: For any $t \geq 0$, we have using

$$\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2,$$

for $x, y \in \mathbb{R}^k$, and (2) that,

$$\begin{aligned}\|\theta(t_R) - \theta(t_R - \tau(t_R))\|^2 \leq 2\eta^2 \left\| \int_{t_R-\tau(t_R)}^{t_R} g_{s-\tau(s)} ds \right\|^2 \\ + 2\eta^2\sigma^2 \|W(t_R) - W(t_R - \tau(t_R))\|^2.\end{aligned}$$

Using (9) and the fact:

$$\mathbb{E}[\|W(t) - W(t-\Delta)\|^2] = \Delta,$$

we have:

$$\begin{aligned}\mathbb{E}\left[\|\theta(t_R) - \theta(t_R - \tau(t_R))\|^2\right] &\leq 2\eta^2 T^2 G_R^2 + 2\eta^2\sigma^2 T \\ &=: D_R \end{aligned} \tag{10}$$

Exponential Supermartingale: For any $\lambda \in \left(0, (\eta\sigma^2)^{-1}\right)$, define

$$a_\lambda \coloneqq \frac{\lambda\eta}{2}(1 - \lambda\eta\sigma^2) \quad (\geq 0),$$

$$B_t \coloneqq \frac{\lambda\eta}{2} L^2 \int_0^{t_R} \|\theta(s) - \theta(s - \tau(s))\|^2 ds + \frac{1}{2}\lambda\eta^2\sigma^2 k L t_R. \tag{11}$$

We note that $B_t \geq 0$. Define the stopped exponential process

$$X_\lambda(t) = \exp\left(\lambda[f(\theta(t_R)) - f(\theta^*)] + a_\lambda \int_0^{t_R} \|g_s\|^2 ds - B_t\right). \tag{12}$$

Let,

$$Y_t \coloneqq \lambda[f(\theta(t_R)) - f(\theta^*)] + a_\lambda \int_0^{t_R} \|g_s\|^2 ds - B_t,$$

so that $X_\lambda(t) = e^{Y_t}$. Thus,

$$dY_t = \lambda df(\theta(t_R)) + a_\lambda \|g_{t_R}\|^2 dt - dB_t.$$

Using Itô's Lemma [30] to obtain $df(\theta(t_R))$, we get:

$$\begin{aligned}dY_t = \Big[&- \lambda\eta g_{t_R}^\top g_{t_R-\tau} + \frac{1}{2}\lambda\eta^2\sigma^2 \operatorname{tr}(\nabla^2 f(\theta(t_R))) \\ &+ a_\lambda \|g_{t_R}\|^2 \Big] dt - dB_t + \lambda dM(t_R)\end{aligned}$$

where $dM(t)$ represents the martingale term and

$$dM(t_R) = \eta\sigma g_{t_R}^\top dW(t). \tag{13}$$

Now using,

$$x^\top y = \frac{1}{2}\left[\|x\|^2 + \|y\|^2 - \|x-y\|^2\right], \tag{14}$$

for $x, y \in \mathbb{R}^k$, we simplify the above SDE as:

$$\begin{aligned}dY_t = \Big[&- \frac{\lambda\eta}{2}\Big(\|g_{t_R}\|^2 + \|g_{t_R-\tau(t_R)}\|^2 - \|g_{t_R} - g_{t_R-\tau(t_R)}\|^2\Big) \\ &+ \frac{1}{2}\lambda\eta^2\sigma^2 \operatorname{tr}(\nabla^2 f(\theta(t_R))) + a_\lambda \|g_{t_R}\|^2 \Big] dt - dB_t + \lambda dM(t_R).\end{aligned} \tag{15}$$

Now, we write the quadratic variation as

$$d\langle Y\rangle_t = \lambda^2 d\langle M\rangle_t = \lambda^2\eta^2\sigma^2 \|g_{t_R}\|^2 dt.$$

Now,

$$dX_\lambda(t_R) = X_\lambda(t_R)\left(dY_t + \frac{1}{2} d\langle Y\rangle_t\right). \tag{16}$$

Using (13) and (15) in (16), we get

$$\begin{aligned}dX_\lambda(t_R) = X_\lambda(t_R)\Bigg[\Bigg(&- \frac{\lambda\eta}{2}\Big(\|g_{t_R}\|^2 + \|g_{t_R-\tau(t_R)}\|^2 \\ &- \|g_{t_R} - g_{t_R-\tau(t_R)}\|^2\Big) + \frac{1}{2}\lambda\eta^2\sigma^2 \operatorname{tr}(\nabla^2 f(\theta(t_R))) \\ &+ a_\lambda \|g_{t_R}\|^2 - \dot{B}_t + \frac{1}{2}\lambda^2\eta^2\sigma^2 \|g_{t_R}\|^2\Bigg) dt + \lambda dM(t_R)\Bigg],\end{aligned} \tag{17}$$

where,

$$\begin{aligned}\dot{B}_t &= \frac{dB}{dt} \\ &= \frac{\lambda\eta}{2} L^2 \|\theta(t_R) - \theta(t_R - \tau(t_R))\|^2 + \frac{1}{2}\lambda\eta^2\sigma^2 k L. \end{aligned} \tag{18}$$

Call the coefficient of dt in (17) as $D(t)$. Then, by our choice of a_λ from (11), we have that

$$-\frac{\lambda\eta}{2}\|g_{t_R}\|^2 + a_\lambda \|g_{t_R}\|^2 + \frac{1}{2}\lambda^2\eta^2\sigma^2 \|g_{t_R}\|^2 = 0. \tag{19}$$

By Assumption 2, for any $t \geq 0$,

$$\begin{aligned} \left\|g_t - g_{t-\tau(t)}\right\| &\leq L\|\theta(t) - \theta(t-\tau(t))\| \\ \operatorname{tr}(\nabla^2 f(\theta(t))) &\leq kL. \end{aligned} \tag{20}$$

Using (19), (20) in (17), we get

$$D(t) \leq -\frac{\lambda\eta}{2}\left\|g_{t_R-\tau(t_R)}\right\|^2 + \frac{\lambda\eta}{2}L^2\|\theta(t_R) - \theta(t_R - \tau(t_R))\|^2 + \frac{1}{2}\lambda\eta^2\sigma^2 kL - \dot{B}_t.$$

Using (18), this simplifies to

$$D(t) \leq -\frac{\lambda\eta}{2}\left\|g_{t_R-\tau(t_R)}\right\|^2 \leq 0.$$

So,

$$\mathbb{E}[X_\lambda(t)] \leq \mathbb{E}[X(0)] = \exp\left(\lambda\left[f(\theta(0)) - f(\theta^*)\right]\right). \tag{21}$$

We therefore have the decomposition

$$dX_\lambda(t) = X_\lambda(t)\lambda dM(t) + X_\lambda(t)\underbrace{D(t)}_{\leq 0}dt,$$

and

$$\mathbb{E}[dX_\lambda(t)|\mathcal{F}_t] < 0 \quad \text{for } \lambda \in (0, (\eta\sigma^2)^{-1}).$$

Thus for any $\lambda \in (0, (\eta\sigma^2)^{-1})$, the process $X_\lambda(t)$ is a non-negative supermartingale.

High Probability Localization: Define the stopped process

$$X_\lambda^{\mathrm{st}}(t) \coloneqq X_\lambda(t_R).$$

Note that X_λ^{st} stays a non-negative supermartingale. On the event $\{\tau_R^\uparrow \leq t\}$, we have $t_R = \tau_R^\uparrow$ and $f(\theta(t_R)) - f(\theta^*) \geq R$. Since the integral in (12) is non-negative,

$$X_\lambda^{\mathrm{st}}(t) \geq \exp(\lambda R - B_t)\mathbb{1}_{\{\tau_R^\uparrow \leq t\}} \tag{22}$$

Taking expectations and using (21),

$$\begin{aligned} \exp\left(\lambda\left[f(\theta(0)) - f(\theta^*)\right]\right) &\geq \mathbb{E}[X_\lambda^{\mathrm{st}}(t)] \\ &\geq \exp(\lambda R)\mathbb{E}\left[e^{-B_t}\mathbb{1}_{\{\tau_R^\uparrow \leq t\}}\right] \\ &\geq \exp(\lambda R)\mathbb{P}(\tau_R^\uparrow \leq t), \end{aligned}$$

since $B_t \geq 0$. Hence, we get a time-uniform tail bound:

$$\mathbb{P}(\tau_R^\uparrow \leq t) \leq \exp\left[\lambda[f(\theta(0)) - f(\theta^*)] - \lambda R\right], \tag{23}$$

for all $t \geq 0$ and $\lambda \in (0, (\eta\sigma^2)^{-1}]$.

Finite Burn-In Time: Now we show $\mathbb{E}[\tau_M] < \infty$. Using Itô's Lemma and (14) as before, we obtain:

$$\begin{aligned} df(\theta(t)) = &-\frac{\eta}{2}\|g_t\|^2 dt - \frac{\eta}{2}\left\|g_{t-\tau(t)}\right\|^2 dt \\ &+ \frac{\eta}{2}\left\|g_t - g_{t-\tau(t)}\right\|^2 dt \\ &+ \frac{1}{2}\eta^2\sigma^2 \operatorname{tr}(\nabla^2 f(\theta(t)))dt + dM(t) \end{aligned}$$

where $dM(t)$ is as defined in (13). Using (20) and dropping the non-positive term $-\frac{\eta}{2}\left\|g_{t-\tau(t)}\right\|^2 dt$ we have,

$$df(\theta(t)) \leq -\frac{\eta}{2}\|g_t\|^2 dt + \frac{\eta}{2}L^2\|\theta(t) - \theta(t-\tau(t))\|^2 dt + \frac{1}{2}\eta^2\sigma^2 kL dt + dM(t).$$

Using $V(t) = f(\theta(t)) - f(\theta^*)$ and $\mathbb{E}[dM(t)] = 0$, we have,

$$\frac{d}{dt}\mathbb{E}[V(t)] \leq -\frac{\eta}{2}\mathbb{E}[\|g_t\|^2] + \frac{\eta}{2}L^2\mathbb{E}\left[\|\theta(t) - \theta(t-\tau(t))\|^2\right] + \frac{1}{2}\eta^2\sigma^2 kL. \tag{24}$$

By Assumption 1,

$$\|g_t\|^2 \geq 2\mu V(t).$$

Thus,

$$-\frac{\eta}{2}\mathbb{E}[\|g_t\|^2] \leq -\eta\mu\mathbb{E}[V(t)]. \tag{25}$$

Now using (25) and (10) in (24), we have

$$\frac{d}{dt}\mathbb{E}[V(t)] \leq -\eta\mu\mathbb{E}[V(t)] + \frac{\eta L^2}{2}D_R + \frac{1}{2}\eta^2\sigma^2 kL. \tag{26}$$

Let,

$$C_R \coloneqq \frac{\eta L^2}{2}D_R + \frac{1}{2}\eta^2\sigma^2 kL. \tag{27}$$

Now define for $\alpha > 0$,

$$Z(t) \coloneqq e^{\alpha t}V(t \wedge \tau_M \wedge \tau_R^\uparrow).$$

Thus,

$$\frac{d}{dt}\mathbb{E}[Z(t)] \leq e^{\alpha t}\Big[(\alpha - \eta\mu)\mathbb{E}[V(t)\mathbb{1}_{\{t<\tau_M\wedge\tau_R^\uparrow\}}] + C_R\mathbb{P}(t < \tau_M \wedge \tau_R^\uparrow)\Big].$$

On $\{t < \tau_M\}$, we have $V(t) \geq M$; so by Markov's inequality,

$$\mathbb{P}(t < \tau_M \wedge \tau_R^\uparrow) \leq \frac{1}{M}\mathbb{E}[V(t)\mathbb{1}_{\{t<\tau_M\wedge\tau_R^\uparrow\}}].$$

Thus,

$$\frac{d}{dt}\mathbb{E}[Z(t)] \leq e^{\alpha t}\left(\alpha - \eta\mu + \frac{C_R}{M}\right)\mathbb{E}[V(t)\mathbb{1}_{\{t<\tau_M\wedge\tau_R^\uparrow\}}].$$

Choose,

$$\alpha = \eta\mu - \frac{2C_R}{M} > 0, \tag{28}$$

to make $\frac{d}{dt}\mathbb{E}[Z(t)] < 0$. Thus, we conclude

$$\mathbb{E}[Z(t)] < \mathbb{E}[Z(0)] = V(0) \quad \forall\, t \geq 0. \tag{29}$$

Therefore, for any $s \geq 0$,

$$\begin{aligned} \mathbb{P}(\tau_M \wedge \tau_R^\uparrow \geq s) &\leq \frac{\mathbb{E}[Z(s)]}{Me^{\alpha s}} \\ &\leq \frac{V(0)}{M}e^{-\alpha s}. \end{aligned} \tag{30}$$

Integrating from $s = 0$ to ∞ gives,

$$\mathbb{E}[\tau_M \wedge \tau_R^\uparrow] \leq \frac{V(0)}{\alpha M} < \infty. \tag{31}$$

Now, for the supermartingale X_λ^{st}, we already have a time uniform exit bound:

$$\delta_R(s) \coloneqq \mathbb{P}(\tau_R^\uparrow < s) \leq e^{\lambda(V(0)-R)} \in (0, 1) \tag{32}$$

for all $\lambda \in (0, (\eta\sigma^2)^{-1})$. Fix such a λ and

$$R > V(0) + \frac{1}{\lambda}\log 2, \tag{33}$$

so that $\delta_R < 1/2$. Next, define a window length

$$t_\star \coloneqq \frac{1}{\alpha}\log\left(\frac{2V(0)}{M}\right) \tag{34}$$

so that,

$$\mathbb{P}(\tau_M \wedge \tau_R^\uparrow > t_\star) \le \frac{1}{2}. \tag{35}$$

Let $s > 0$. Then,

$$\{\tau_M < s\}^c = \{\tau_M \ge s\} \subseteq \{\tau_M \wedge \tau_R^\uparrow \ge s\} \cup \{\tau_R^\uparrow \le s\}. \tag{36}$$

Using union bound,

$$\mathbb{P}(\tau_M < s) \ge 1 - \mathbb{P}(\tau_M \wedge \tau_R^\uparrow \ge s) + \mathbb{P}(\tau_R^\uparrow \le s). \tag{37}$$

Substituting using (31) and (32) we get

$$\mathbb{P}(\tau_M < s) \ge 1 - \frac{V(0)}{M}e^{-\alpha s} - e^{\lambda[V(0)-R]}.$$

For $s = t_\star$, we get

$$p_R \coloneqq \mathbb{P}(\tau_M \le t_\star) \ge 1 - \frac{1}{2} - \delta_R = \frac{1}{2} - \delta_R > 0$$

by our choice of R from (33). Now we build an upper-bound for $\mathbb{E}[\tau_M]$. Let,

$$S_0 = 0; \quad S_{n+1} = S_n + t_\star, \quad n \ge 0.$$

Define the Bernoulli indicators

$$\Xi_n \coloneqq \mathbb{1}_{\{\tau_m \le S_{n+1}\}} - \mathbb{1}_{\{\tau_m \le S_n\}},$$

and,

$$N \coloneqq \min\{n \ge 0 : \Xi_n = 1\}.$$

Then, $\tau_M \le S_{N+1} = (N+1)t_\star$. We claim:

$$\mathbb{P}(\tau_M \le S_{n+1} | \mathcal{F}_{S_n}) \ge p_R \quad \text{on } \{V(S_n) \le R\} \cap \{\tau_M > S_n\},$$

where $n \in \{1, 2, \ldots, N\}$. Indeed, if we define

$$\beta(t) \coloneqq \{\theta(t+u)\}_{u \in [-T,0]},$$

then $\beta(t)$ is strong Markov. So by the strong Markov property, conditional on $\mathcal{F}_{S_n}$, the post-S_n evolution is the same SDE started from $\beta(S_n)$. Again using a similar union bound as in (37), we get

$$\begin{aligned}\mathbb{P}(\tau_M \le S_{n+1} | \mathcal{F}_{S_n}) \ge 1 - \mathbb{P}(\tau_M \wedge \tau_R^\uparrow > S_{n+1} | \mathcal{F}_{S_n}) \\ - \mathbb{P}(\tau_R^\uparrow \le S_{n+1} | \mathcal{F}_{S_n}).\end{aligned} \tag{38}$$

Note that $S_{n+1} - S_n = t_\star$. By a similar localization argument as before, we reach

$$\begin{aligned}\mathbb{P}(\tau_M \wedge \tau_R^\uparrow > S_{n+1} | \mathcal{F}_{S_n}) &\le \frac{V(S_n)}{M}e^{-\alpha t_\star} \\ &\le \frac{R}{M}e^{-\alpha t_\star}.\end{aligned} \tag{39}$$

Using the supermartingale argument for escaping level sets, we reach

$$\begin{aligned}\mathbb{P}(\tau_R^\uparrow \le S_n + t_\star) &\le \exp\left(\lambda[V(S_n) - R]\right) &(40)\\ &= \exp\left(-\lambda[R - V(S_n)]\right) &(41)\\ &= e^{-\lambda\Delta} \quad (\Delta = R - V(S_n) \ge 0). &(42)\end{aligned}$$

Thus, the escape probability is strictly less than 1 whenever $V(S_n) < R$. Using (39) and (40) in (38), we get

$$\begin{aligned}\mathbb{P}(\tau_M \le S_{n+1} | \mathcal{F}_{S_n}) &\ge 1 - \frac{R}{M}e^{-\alpha t_\star} - e^{-\lambda(R - V(S_n))} \\ &\ge p_R.\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{E}[\Xi_n | \mathcal{F}_{S_n}] &= \mathbb{P}(S_n < \tau_M \le S_{n+1} | \mathcal{F}_{S_n}) \\ &= \mathbb{1}_{\{\tau_M > S_n\}}\mathbb{P}(\tau_M \le S_{n+1} | \mathcal{F}_{S_n}).\end{aligned}$$

Thus, on $\{\tau_M \le S_n\}$, the expectation is 0. So we are left with

$$\mathbb{P}(\Xi_n = 1 | \mathcal{F}_{S_n}) \ge p_R, \tag{43}$$

on the event $\{V(S_n) \le R\} \cap \{\tau_M > S_n\}$ as claimed.
Now, call block n an *attempt* if $\{V(S_n) \le R\} \cap \{\tau_M > S_n\}$ holds. (43) says that attempt-by-attempt the conditional success probability is $\ge p_R$. Define the index of the first successful block as

$$N \coloneqq \inf\{n \ge 0 : \tau_M \le S_n + t_\star\}.$$

WLOG, assume that the blocks are started only when inside the interior (i.e., $V(S_n) < R$ on $\{\tau_M > S_n\}$ for every n). Otherwise, we increase R (say, $R \leftarrow 2R$). Let $q_R = 1 - p_R$. For $n \ge 0$, set

$$A_n \coloneqq \{\tau_M > S_n\} \cap \{V(S_n) < R\} \in \mathcal{F}_{S_n}.$$

By construction, A_n holds whenever block n begins. From above, we know

$$\mathbb{E}[\Xi_n | \mathcal{F}_{S_n}] \ge p_R \mathbb{1}_{\{A_n\}}.$$

Since $\{N > n\} \subseteq A_n$ (if success has not occurred by the end of block n, then we are alive at its start and in the interior), we have

$$\begin{aligned}\mathbb{P}(N > n+1) &= \mathbb{E}[\mathbb{1}_{\{N>n\}}\mathbb{P}(\Xi_n = 0 | \mathcal{F}_{S_n})] \\ &\le \mathbb{E}[\mathbb{1}_{\{N>n\}}(1 - p_R)] \\ &= q_R \mathbb{P}(N > n).\end{aligned}$$

So, by induction

$$\mathbb{P}(N > m) \le q_R^{m+1}, \quad m \ge 0.$$

Thus,

$$\begin{aligned}\mathbb{E}[N] &= \sum_{m \ge 0}\mathbb{P}(N > m) \\ &\le \sum_{m \ge 0} q_R^{m+1} \\ &= \frac{1 - p_R}{p_R}.\end{aligned}$$

Finally, $\tau_M \le S_n + t_\star \le (N+1)t_\star$ and

$$\begin{aligned}\mathbb{E}[\tau_M] &\le \mathbb{E}[N+1]t_\star \\ &\le t_\star \frac{1 - p_R}{p_R} \\ &< \frac{t_\star}{p_R} < \infty.\end{aligned}$$

Consequently, from (28), we have

$$M_\star = \frac{2C_R}{\eta\mu}.$$

□

We note below that the result obtained in Lemma 1 holds, with an idential proof, even when the parameter η depend on time.

Corollary 1. (**Bounded burn-in time for time-varying learning rate**): If $\eta \equiv \eta(t)$ then with $\alpha \equiv \alpha(t)$ such that $\alpha, \eta : [0, \infty) \mapsto \mathbb{R}_{>0}$, we have,

$$\mathbb{E}[\tau_M] < \infty.$$

APPENDIX II
STAYING IN LOCAL SUBLEVEL SET

Lemma 2. (**Staying in local sublevel set**): Fix $M > 0$, and a horizon $H > 0$, and a failure level $\delta \in (0, 1)$. Assume for $t_0 > 0$ that $f(\theta(t_0)) - f(\theta^*) < M$ and that the delayed history over the last T units of time is already inside a slightly larger set Ω_{2M}, i.e.,

$$\theta(u) \in \Omega_{2M} \quad \forall\, u \in [t_0 - T, t_0],$$

where,

$$\Omega_{2M} \coloneqq \{\theta : f(\theta) - f(\theta^*) \le 2M\}.$$

Then there exists $\bar{\eta} > 0$ such that for all $\eta \in (0, \bar{\eta}]$,

$$\mathbb{P}\left(\sup_{t\in[t_0,t_0+H]} f(\theta(t)) - f(\theta^*) > 2M\right) \le \delta$$

Proof. Partition $[t_0, t_0 + H]$ into $J \coloneqq \lceil H/h \rceil$ sub-intervals of length $h \in (0, \min\{1, T\}]$ with grid points $t_j = t_0 + jh$. By (2), for any $t \in [t_j, t_{j+1}]$,

$$\theta(t) - \theta(t_j) = -\eta \int_{t_j}^{t} \nabla f(\theta(s - \tau(s))) + \eta\sigma \left(W(t) - W(t_j)\right). \quad (44)$$

On the event $\{\theta(s - \tau(s)) \in \Omega_{2M}\, \forall s \in [t_j, t_{j+1}]\}$, we have $\|\nabla f(\theta(s - \tau(s)))\| \le G_{2M}$, hence

$$\|\theta(t) - \theta(t_j)\| \le \eta \left(hG_{2M} + \sigma S_j\right),$$

where,

$$S_j \coloneqq \sup_{u\in[0,h]} \|W(t_j + u) - W(t_j)\|.$$

Note that Assumptions 2 and 5 guarantee $G_{2M} < \infty$.
By descent lemma for L-smooth functions,

$$\begin{aligned} |f(\theta(t)) - f(\theta(t_j))| &\le \|\nabla f(\theta(t_j))\|\|\theta(t) - \theta(t_j)\| \\ &\quad + \frac{L}{2}\|\theta(t) - \theta(t_j)\|^2 \\ &\le a\eta R_j + \frac{L}{2}\eta^2 R_j^2, \end{aligned} \quad (45)$$

where,

$$a \coloneqq G_{2M}, \quad R_j \coloneqq hG_{2M} + \sigma S_j.$$

Now, define

$$m(\eta, S) \coloneqq a\eta(hG_{2M} + \sigma S) + \frac{L}{2}\eta^2 \left(hG_{2M} + \sigma S\right)^2.$$

With $V(t) \coloneqq f(\theta(t)) - f(\theta^*)$, we have from (45) that,

$$\sup_{t\in[t_j,t_{j+1}]} V(t) \le V(t_j) + m(\eta, S_j). \quad (46)$$

By Gaussian concentration, $\exists c_1, c_2 > 0$ such that,

$$\mathbb{P}\left(S_j \ge \alpha\sqrt{h}\right) \le c_1 e^{-c_2\alpha^2}, \quad \forall\, \alpha > 0.$$

Choose,

$$\alpha \coloneqq \sqrt{\frac{1}{c_2}\log\left(\frac{2c_1 J}{\delta}\right)}.$$

Then a union bound over $j = 0, \ldots, J-1$, leads to $S_j \le \alpha\sqrt{h}$ for all j with probability at least $1 - \delta/2$. On this event (i.e., the event $\{S_j \le \alpha\sqrt{h}\}$), we have

$$m(\eta, S_j) \le m_\star(\eta) \quad (47)$$

$$\coloneqq a\eta\left(hG_{2M} + \sigma\alpha\sqrt{h}\right) + \frac{L}{2}\eta^2\left(hG_{2M} + \sigma\alpha\sqrt{h}\right)^2. \quad (48)$$

Now, we use induction over sub-intervals to show that the entire trajectory from t_0 to $t_0 + H$ lies within Ω_{2M} with probability at least $1 - \delta/2$.
We have $V(t_0) \le M$. Because $h \le T$ and the initial delayed history on $[t_0 - T, t_0]$ lies in Ω_{2M}, Assumption 3, implies $\theta(s - \tau(s)) \in \Omega_{2M}$ for all $s \in [t_j, t_{j+1}]$. Choose,

$$\eta \le \bar{\eta} \coloneqq \min\left\{\frac{M}{2Ja\left(hG_{2M} + \sigma\alpha\sqrt{h}\right)}, \sqrt{\frac{M}{JL\left(hG_{2M} + \sigma\alpha\sqrt{h}\right)^2}}\right\},$$

so that,

$$Jm_\star(\eta) \le M.$$

Let,

$$V_j = \sup_{t\in[t_0,t_j]} V(t), \quad j \in \{0, \ldots, J\}.$$

On the event where all S_j are bounded, (46) and (47) gives

$$\sup_{t\in[t_j,t_{j+1}]} V(t) \le V(t_j) + m_\star(\eta) \le V_j + m_\star(\eta). \quad (49)$$

Therefore,

$$V_{j+1} = \sup\left\{V_j, \sup_{t\in[t_j,t_{j+1}]} V(t)\right\} \le V_j + m_\star(\eta)$$

Inducting from $V_0 = V(t_0) \le M$ yields,

$$V_j \le M + jm_\star(\eta) \quad \text{for all } j.$$

Taking $j = J$, we get

$$\sup_{t\in[t_0,t_0+H]} V(t) = V_J \le M + Jm_\star(\eta) \le 2M$$

with probability at least $1 - \delta/2$. On the complementary event (i.e., $\{S_j > \alpha\sqrt{h}\}$) we pay probability at most $\delta/2$. Combining both leads to the desired bound. □

APPENDIX III
PROOF OF THEOREM 1

Now, in proving the theorem, we split the SDE solution into two phases: the *burn-in* phase and the *local* phase. Intuitively, after initializing θ, say at $\theta(0)$, the burn-in phase corresponds to the regime where the SDE drives θ *rapidly* towards the optimum; in this phase the gradients are relatively large. Once the process enters the local phase, the parameter vector stabilizes near the optimum, so the gradients become correspondingly small. This decomposition allows us to bound the deviation of θ from the optimum in terms of the *local* gradient norm, i.e., the gradient bound valid in the local phase. Without such a refinement, one would need to impose a uniform bound on the gradient norm across all phases, which would in turn yield a larger steady-state noise margin.
We now state the proof formally.

Proof. Let V be as before. Then, by Lemma 1, $\mathbb{E}[\tau_M] \leq T_{\text{burn}}$ where τ_M is as in (3). The arguments to follow are to be thought of as being along a sample path of $V(t)$. In this sense, even if we do not state this explicitly, the statements are meant to hold almost surely. Since $\tau_M < \infty$, hence for $t > \tau_M$, we have that, $f(\theta(t)) - f(\theta^*) \leq 2M$. Since $f \in \mathcal{C}^2$, we obtain

$$\|\nabla f(\theta(t))\| \leq G_{2M},$$

for some $G_{2M} > 0$. Once the process has entered the sub-level set Ω_M, consider the stopping time $\rho := \inf\{t \geq \tau_M : V(t) > 2M\}$. Define the stopped process $\tilde{\theta}(t) := \theta(t \wedge \rho)$. By Lemma 2 applied with horizon $H = T$ and the delayed-history condition at time τ_M, with high-probability the segment $\{\tilde{\theta}(s - \tau(s))\}_{s\in[\tau_M,t]}$ stays in Ω_{2M} on every window of length T. Hence all local bounds below hold along $[\tau_M, t \wedge \rho]$. Finally, since the exit probability has a time-uniform tail (see the supermartingale bound in Lemma 1), we let $\rho \to \infty$ to obtain the claimed estimate for $\mathbb{E}[V(t)]$.
Now we derive an ODE-type inequality on $V(t)$ in Ω_M. Using Itô's Lemma [30] we have that,

$$\begin{aligned} dV(t) &= -\eta\nabla f(\theta(t))^\top \nabla f(\theta(t-\tau(t)))dt \\ &\quad + \eta\sigma\nabla f(\theta(t))^\top dW(t) + \frac{1}{2}\eta^2\sigma^2 \operatorname{tr}[\nabla^2 f(\theta(t))]dt \end{aligned} \tag{50}$$

Taking expectations and using $\operatorname{tr}[\nabla^2 f(\theta(t))] \leq kL$ where $\theta(t) \in \mathbb{R}^k$ we get,

$$\frac{d}{dt}\mathbb{E}[V(t)] \leq -\eta\mathbb{E}\left[\nabla f(\theta(t))^\top \nabla f(\theta(t-\tau(t)))\right] + \frac{1}{2}\eta^2\sigma^2 kL, \tag{51}$$

where we have also used the fact that the Brownian term is a martingale and has zero mean. Now we lower-bound the inner-product $\nabla f(\theta(t))^\top \nabla f(\theta(t-\tau(t)))$. Using the identity,

$$a \cdot b = \frac{1}{2}\left(\|a\|^2 + \|b\|^2\right) - \frac{1}{2}\|a-b\|^2,$$

where $a, b \in \mathbb{R}^k$ and applying expectations on (50) we have,

$$\begin{aligned} \mathbb{E}[\nabla f(\theta(t))^\top \nabla f(\theta(t-\tau(t)))] &= \frac{1}{2}\mathbb{E}\left[\|\nabla f(\theta(t))\|^2\right] \\ &\quad + \frac{1}{2}\mathbb{E}\left[\|\nabla f(\theta(t-\tau(t)))\|^2\right] \\ &\quad - \frac{1}{2}\mathbb{E}\left[\|\nabla f(\theta(t)) - \nabla f(\theta(t-\tau(t)))\|^2\right]. \end{aligned}$$

Using (5) to lower-bound the first two terms and Lipschitz continuity to upper bound the third term, we get

$$\begin{aligned} \mathbb{E}[\nabla f(\theta(t))^\top \nabla f(\theta(t-\tau(t)))] &\geq \mu\mathbb{E}\left[V(t)\right] + \mu\mathbb{E}\left[V(t-\tau(t))\right] \\ &\quad - \frac{1}{2}L^2\mathbb{E}\left[\|\theta(t) - \theta(t-\tau(t))\|^2\right]. \end{aligned} \tag{52}$$

Substituting the bound in (52) to (51) and dropping $-\eta\mu\mathbb{E}[V(t-\tau(t))]$ we conclude that,

$$\begin{aligned} \frac{d}{dt}\mathbb{E}[V(t)] &\leq -\eta\mu\mathbb{E}[V(t)] + \frac{\eta L^2}{2}\mathbb{E}\left[\|\theta(t) - \theta(t-\tau(t))\|^2\right] \\ &\quad + \frac{\eta^2\sigma^2 kL}{2}. \end{aligned} \tag{53}$$

We further bound $\|\theta(t) - \theta(t-\tau(t))\|^2$ as follows:

$$\begin{aligned} \theta(t) - \theta(t-\tau(t)) &= \int_{t-\tau(t)}^{t} -\eta\nabla f(u-\tau(u))du \\ &\quad + \int_{t-\tau(t)}^{t} \eta\sigma dW(u). \end{aligned}$$

Using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$,

$$\begin{aligned} \|\theta(t) - \theta(t-\tau(t))\|^2 &\leq 2\left\|\int_{t-\tau(t)}^{t} -\eta\nabla f(u-\tau(u))du\right\|^2 \\ &\quad + 2\left\|\int_{t-\tau(t)}^{t} \eta\sigma dW(u)\right\|^2, \end{aligned} \tag{54}$$

we bound each of the two terms above. First we bound the drift term.

$$\begin{aligned} \left\|\int_{t-\tau(t)}^{t} -\eta\nabla f(u-\tau(u))du\right\| &\leq \int_{t-\tau(t)}^{t} \eta\|\nabla f(u-\tau(u))\|du \\ &\leq \eta G_{2M} T. \end{aligned} \tag{55}$$

For the noise term, we note that,

$$\left\|\int_{t-\tau(t)}^{t} \eta\sigma dW(u)\right\|^2 \leq \eta^2\sigma^2\tau(t) \leq \eta^2\sigma^2 T. \tag{56}$$

Combining the drift and noise parts ((55), (56)) and substituting in (54), we obtain:

$$\|\theta(t) - \theta(t-\tau(t))\|^2 \leq 2\eta^2 T^2 \left(G_{2M}\right)^2 + 2\eta^2\sigma^2 T =: \Delta_{\text{local}}. \tag{57}$$

If $\theta(\cdot) \in \Omega_{2M}$ around time t, then

$$\|\theta(t) - \theta(t-\tau(t))\| \leq \Delta_{\text{local}}.$$

Thus following from the inequality in (53), we get

$$\frac{d}{dt}\mathbb{E}[V(t)] \leq -\eta\mu\mathbb{E}[V(t)] + \frac{\eta L^2}{2}\Delta_{\text{local}} + \frac{\eta^2\sigma^2 kL}{2}.$$

Denoting by,

$$\beta_{\text{local}} = \frac{\eta L^2}{2}\Delta_{\text{local}} + \frac{\eta^2\sigma^2 kL}{2}, \quad \text{and} \quad \xi = \eta\mu,$$

we have that,

$$\frac{d}{ds}\mathbb{E}[V(s)] \leq -\xi\mathbb{E}[V(s)] + \beta_{\text{local}}$$

for $s \geq \tau_M$. Integrating from $s = \tau_M$ to $s = t$ gives:

$$\mathbb{E}[V(t)] \leq \left(\mathbb{E}[V(\tau_M)] - \frac{\beta_{\text{local}}}{\xi}\right) e^{-\xi(t-\tau_M)} + \frac{\beta_{\text{local}}}{\xi}.$$

Finally defining

$$\kappa := \frac{\beta_{\text{local}}}{\xi} = \frac{\eta^2}{\mu}\left(L^2T^2\left(G_{2M}\right)^2 + L^2\sigma^2T\right) + \frac{\eta\sigma^2kL}{2\mu},$$

yields the result. □

Appendix IV
Proof Outline of Theorem 2

Proof. By Corollary 1, we have that $\mathbb{E}[\tau_M] < \infty$, where τ_M is as in (3). As a result, $\theta(t)$ enters the sublevel set $\Omega_M := \{\theta : V(t) < M\}$. Next we partition $[t_0, t_f]$, into blocks $[t_j, t_{j+1}]$ with $t_0 = \tau_M$ and $t_{j+1} - t_j \geq T$ as in Assumption 6. On each block, we apply Lemma 2 with horizon $t_{j+1} - t_j$. This yields that with probability at least $1 - \delta_j$, that $\{\theta(s - \tau(s))\}_{s\in[t_j,t_{j+1}]} \subset \Omega_{2M}$; hence the local smoothness/gradient bounds hold on the entire block. Choosing a failure budget with $\sum_j \delta_j \leq \delta$, we obtain the same but block-wise differential inequality as in Theorem 1:

$$\begin{aligned} dV(t) = &-\eta(t)\nabla f(\theta(t))^\top \nabla f(\theta(t-\tau(t)))dt \\ &+ \eta(t)\sigma\nabla f(\theta(t))^\top dW(t) \\ &+ \frac{1}{2}\eta(t)^2\sigma^2 \operatorname{tr}[\nabla^2 f(\theta(t))]dt \end{aligned}$$

for $t \geq \tau_M$. Take the expectation to obtain

$$\begin{aligned} \mathbb{E}[dV(t)] = &\eta(t)\mathbb{E}[\nabla f(\theta(t))^\top \nabla f(\theta(t-\tau(t)))]dt \\ &+ \frac{1}{2}\eta(t)^2\sigma^2\mathbb{E}\left[\operatorname{tr}\left[\nabla^2 f(\theta(t))\right]\right]dt. \end{aligned}$$

From here onward, we follow exactly along the lines of the proof of Theorem 1, as we decompose the inner product into sums of norms using

$$a^\top b = \frac{1}{2}\left(\|a\|^2 + \|b\|^2\right) - \frac{1}{2}\|a-b\|^2,$$

and use (5) to lower-bound the first two terms and L- smoothness to upper bound the norm difference squared, to obtain

$$\begin{aligned} \|\theta(t) - \theta(t-\tau(t))\|^2 &\leq 2\eta(t)^2T^2\left(G_{2M}\right)^2 + 2\eta(t)^2\sigma^2T \\ &=: \Delta_{\text{local}}(t). \end{aligned} \tag{58}$$

Proceeding further in a similar fashion as in the proof of Theorem 1, we arrive at the differential inequality:

$$\begin{aligned} \frac{d}{dt}\mathbb{E}[V(t)] \leq &-\mu\eta(t)\mathbb{E}[V(t)] + \frac{1}{2}L^2\Delta_{\text{local}}(t)\eta(t) \\ &+ \frac{\sigma^2}{2}\left|\max_{1\leq s\leq t_f} \operatorname{tr}(\nabla^2 f(\theta(s)))\right| \cdot \eta^2(t). \end{aligned} \tag{59}$$

A sufficient condition for negative drift would be to ensure the RHS of (59) to be negative.

Substituting for $\Delta_{\text{local}}(t)$ in (59) we obtain a cubic inequality The roots of this cubic are:

$$\eta_1(t) = 0, \tag{60a}$$

$$\eta_2(t) = \frac{-\sigma^2 H_{\max} + \sqrt{\sigma^4 H_{\max}^2 + 16\mu\mathbb{E}[V(t)]a}}{2a}, \tag{60b}$$

$$\eta_3(t) = \frac{-\sigma^2 H_{\max} - \sqrt{\sigma^4 H_{\max}^2 + 16\mu\mathbb{E}[V(t)]a}}{2a}, \tag{60c}$$

where,

$$H_{\max} = \left|\max_{1\leq s\leq t_f} \operatorname{tr}(\nabla^2 f(\theta(s)))\right|$$

and,

$$a = L^2T^2(G_{2M})^2 + L^2\sigma^2T.$$

Since $\eta(t) > 0$, thus for the RHS of (59) to be negative, we must have:

$$0 < \eta(t) < \eta_2(t),$$

for all $t \geq \tau_M$ (refer to Remark IV.1 for an interpretation).

Now continuing with the analysis, if $t \in [t_j, t_{j+1}]$, then using from (58), we have,

$$\begin{aligned} \|\theta(t) - \theta(t-\tau(t))\|^2 \leq &\, 2\eta_{\max,j}^2T^2\left(G_{2M}\right)^2 \\ &+ 2\eta_{\max,j}^2\sigma^2T =: \Delta_{\text{local}}^{(j)}, \end{aligned}$$

and we have,

$$\begin{aligned} \frac{d}{dt}\mathbb{E}[V(t)] \leq &-\mu\eta(t)\mathbb{E}[V(t)] + \\ &\underbrace{\frac{1}{2}L^2\Delta_{\text{local}}^{(j)}\eta(t) + \frac{\sigma^2}{2}\left|\max_{1\leq s\leq t_f} \operatorname{tr}(\nabla^2 f(\theta(s)))\right| \cdot \eta(t)^2}_{=:\beta(t)}. \end{aligned}$$

This is a linear differential inequality. With an integrating factor of $e^{\mu\int_{t_0}^t \eta(s)ds}$ we can solve it as:

$$\begin{aligned} \mathbb{E}[V(t)] \leq &\, \mathbb{E}[V(t_0)]e^{-\mu\int_{t_0}^t \eta(s)ds} \\ &+ e^{-\mu\int_{t_0}^t \eta(s)ds}\int_{t_0}^t e^{\mu\int_{t_0}^u \eta(s)ds}\beta(u)du. \end{aligned}$$

We upper bound the second term in the RHS above. We have,

$$e^{-\mu\int_{t_0}^t \eta(s)ds}\int_{t_0}^t e^{\mu\int_{t_0}^u \eta(s)ds}\beta(u)du = \sum_{j=0}^{K-1} R(j)$$

where

$$R(j) = e^{-\mu\int_{t_j}^{t_{j+1}} \eta(s)ds}\int_{t_j}^{t_{j+1}} e^{\mu\int_{t_0}^u \eta(s)ds}\beta(u)du.$$

with $t_K = t_f$. Since each summand of the above sum is positive, we can upper bound the sum by upper bounding each summand, which we do as follows.

On each interval $[t_j, t_{j+1}]$, since $t_{j+1} - t_j > T$, we have $\beta(u) \leq \beta_{\max,j}$ where,

$$\beta_{\max,j} := \frac{1}{2}L^2\Delta_{\text{local}}^{(j)}\eta_{\max,\,j} + \frac{\sigma^2}{2}\left|\max_{1\leq s\leq t_f} \operatorname{tr}(\nabla^2 f(\theta(s)))\right| \cdot \eta_{\max,j}^2.$$

We draw the following observations:

$$e^{-\mu\int_{t_j}^{t_{j+1}} \eta(s)\,ds} \leq e^{-\mu\eta_{\min,j}(t_{j+1}-t_j)},$$

and that

$$\int_{t_j}^{t_{j+1}} e^{\mu \int_{t_0}^{u} \eta(s)\, ds} \beta(u)\, du \leq \beta_{\max,j} e^{-\mu \eta_{\max,\, j}\, t_0} \cdot \int_{t_j}^{t_{j+1}} e^{\mu \eta_{\max,\, j} u}\, du = \frac{\beta_{\max,j}}{\mu \eta_{\max,\, j}} e^{-\mu \eta_{\max,\, j}\, t_0} \cdot \left(e^{\mu \eta_{\max,\, j} t_{j+1}} - e^{\mu \eta_{\max,\, j} t_j} \right).$$

Using both inequalities above, we have that

$$R(j) \leq \frac{\beta_{\max,j}}{\mu \eta_{\max,\, j}} e^{-\mu \eta_{\max,\, j}\, t_0} \left(e^{\mu \left[\eta_{\max,\, j} - \eta_{\min,j} \right] t_{j+1}} e^{\mu \eta_{\min,j} t_j} - e^{\mu \left[\eta_{\max,\, j} + \eta_{\min,j} \right] t_j} e^{-\mu \eta_{\min,j} t_{j+1}} \right). \tag{61}$$

We claim that the upper bound in (61) is positive. Note that

$$e^{\mu \left[\eta_{\max,\, j} - \eta_{\min,j} \right] t_{j+1}} e^{\mu \eta_{\min,j} t_j} > e^{\mu \left[\eta_{\max,\, j} + \eta_{\min,j} \right] t_j} e^{-\mu \eta_{\min,j} t_{j+1}},$$

whenever $t_{j+1} > t_j$ which is trivially true.
As a result, we have,

$$\sum_{j=0}^{K-1} R(j) \leq \sum_{j=0}^{K-1} \frac{\beta_{\max,j}}{\mu \eta_{\max,\, j}} e^{-\mu \eta_{\max,\, j}\, t_0} \left(e^{\mu \left[\eta_{\max,\, j} - \eta_{\min,j} \right] t_{j+1}} e^{\mu \eta_{\min,j} t_j} - e^{\mu \left[\eta_{\max,\, j} + \eta_{\min,j} \right] t_j} e^{-\mu \eta_{\min,j} t_{j+1}} \right).$$

Finally substituting $\beta_{\max,j}$, yields the result. □

Remark IV.1. (**Learning rate**): It can be verified that in the absence of delay ($T = 0$), we have a quadratic inequality instead of a cubic inequality. The condition for a strictly negative drift is then:

$$0 < \eta(t) < \frac{2\mu \mathbb{E}[V(t)]}{\sigma^2 H_{\max}}.$$

This comes both from taking the limit $T \to 0$ in (60b) as well as solving explicitly for $\eta(t)$ with the modified equation:

$$\frac{d}{dt}\mathbb{E}[V(t)] \leq -\mu \eta(t) \mathbb{E}[V(t)] + \frac{\sigma^2}{2} H_{\max} \eta(t)^2.$$

•

APPENDIX V
SUPPLEMENTARY PROOFS

Recall from Experiments and Results that the synthetic function we chose for our experiments was:

$$f(\theta) = \theta^\top Q \theta + \epsilon \sin(w^\top \theta). \tag{62}$$

In what follows, we derive the strong-convexity constant μ and the L-smoothness constant L for this function.

V-A Derivation of strong-convexity constant

Define,

$$R_Q(v) := \frac{v^\top Q v}{\|v\|^2},$$

to be the Rayleigh-quotient for a vector v for a symmetric positive-definite matrix Q.
The gradient and Hessian of f are given respectively as:

$$\nabla f(\theta) = 2Q\theta + \epsilon \cos(w^\top \theta) w, \tag{63a}$$

$$\nabla^2 f(\theta) = 2Q - \epsilon \sin(w^\top \theta) w w^\top. \tag{63b}$$

Pick any unit vector v. Using,

$$v^\top (w w^\top) v = (v^\top w)^2,$$

we have from (63b),

$$v^\top \left(\nabla^2 f(\theta) \right) v = 2R_Q(v) - \epsilon \sin(w^\top \theta)(v^\top w)^2.$$

Now using,

$$R_Q(v) \geq \lambda_{\min}(Q),$$

and,

$$\sin(\cdot) \in [-1, +1],$$

we can lower-bound $v^\top (\nabla^2 f(\theta)) v$ as,

$$v^\top \left(\nabla^2 f(\theta) \right) v \geq 2\lambda_{\min}(Q) - \epsilon (v^\top w)^2 \geq 2\lambda_{\min}(Q) - \epsilon \|w\|^2,$$

where the last inequality follows from Cauchy-Schwarz. Therefore provided,

$$\epsilon < \frac{2\lambda_{\min}(Q)}{\|w\|^2},$$

the Hessian is uniformly positive-definite and

$$\mu = 2\lambda_{\min}(Q) - \epsilon \|w\|^2.$$

V-B Derivation of L-smoothness constant

Now we upper bound $v^\top (\nabla^2 f(\theta)) v$, which in a similar fashion like the lower-bound happens to be

$$v^\top (\nabla^2 f(\theta)) v \leq 2\lambda_{\max}(Q) + \epsilon \|w\|^2.$$

Thus the operator-norm of the Hessian satisfies

$$\left\| \nabla^2 f(\theta) \right\|_2 \leq L = 2\lambda_{\max}(Q) + \epsilon \|w\|^2,$$

for all θ. Thus f is L-smooth.

Arijit Dey received the B.E. degree in Instrumentation and Electronics Engineering, in 2021 from Jadavpur University, India, M.Tech in Signal Processing from Indian Institute of Science, India in 2023 followed by an M.S. in Electrical and Computer Engineering from University of California, Los Angeles in 2024.

Bahman Gharesifard (Senior Member, IEEE) received the B.Sc. degree in Mechanical Engineering, in 2002, and the M.Sc. degree in Control and Dynamics, in 2005, from Shiraz University, Iran, and Ph.D. in Mathematics from Queen's University, Canada, in 2009. He is a Professor with the Department of Mathematics and Statistics at Queen's University, where has been appointed since 2013. He was a Professor with the Electrical and Computer Engineering Department at the University of California, Los Angeles from 2021 to 2024, where he was the Area Director for Signals and Systems 2023-2024. He was an Alexander von Humboldt research fellow with the Institute for Systems Theory and Automatic Control at the University of Stuttgart in 2019-2020. He held postdoctoral positions with the Department of Mechanical and Aerospace Engineering at University of California, San Diego (2009-2012) and with the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign (2012-2013). He received the 2019 CAIMS-PIMS Early Career Award, jointly awarded by the Canadian Applied and Industrial Math Society and the Pacific Institute for the Mathematical Sciences, an Alexander von Humboldt Foundation research fellowship for experienced researchers in 2019, an NSERC Discovery Accelerator in 2019, the SIAG/CST Best SICON Paper Prize in 2021, and the Canadian Society for Information Theory (CSIT) Best Paper Award in 2022. He has served on the Conference Editorial Board of the IEEE Control Systems Society, as an Associate Editor for the IEEE CONTROL SYSTEM LETTERS and IEEE TRANSACTIONS ON NETWORK CONTROL SYSTEMS. His research interests include systems and controls, machine learning, distributed control and optimization, social and economic networks, game theory, geometric control theory, geometric mechanics, and Riemannian geometry.