

Time-uniform concentration bounds for iterative algorithms

Tuan Pham

Department of Statistics and Data Science, University of Texas, Austin

tuan.pham@utexas.edu

Alessandro Rinaldo

Department of Statistics and Data Science, University of Texas, Austin

alessandro.rinaldo@austin.utexas.edu

Purnamrita Sarkar

Department of Statistics and Data Science, University of Texas, Austin

purna.sarkar@austin.utexas.edu

Abstract

We develop a new framework for deriving time-uniform concentration bounds for the output of stochastic sequential algorithms satisfying certain recursive inequalities akin to those defining the almost-supermartingale processes introduced by [49]. Our approach is of wide applicability, and can be deployed in settings in which exponential supermartingale processes, required by prevailing methodologies for anytime-valid concentration inequalities, are not readily available. Our results can be viewed as quantitative versions of the classical Robbins–Siegmund Lemma. We demonstrate the effectiveness of our method by providing new and optimal time-uniform concentration bounds for Oja’s algorithm for streaming PCA, stochastic gradient descent, and stochastic approximations.

Contents

1	Introduction	2
1.1	Related results	5
2	Main results	5
2.1	Preliminaries	5
2.2	A quantitative Robbins-Siegmund’s Lemma and time-uniform concentration bounds	7
2.3	An explicit step size construction	10
3	Applications	11
3.1	SGD algorithms	11
3.2	Time-uniform bounds under (λ -SC)	12
3.3	Uniform error bound under (μ -smooth) and (P-L)	15
3.4	Oja’s algorithm for PCA	16
3.5	Robbins-Monroe scheme and a lower bound	20
4	Outline of the argument	22

5 Conclusion and discussion	23
6 Acknowledgments	24
7 Proof of Theorem 1	24
7.1 Step 1: Maximal inequality	25
7.2 Step 2: Induction argument	28
8 Proof of Proposition 1	33
9 Proof of Theorem 2	35
10 Proof of Theorem 3	37
11 Proof of Theorem 4	39
A Extension to the sub-Gaussian coefficients settings	44
A.1 Extension of Theorem 1	44
A.2 Extension of Proposition 1 and Theorem 2	45
A.3 Some technical results	45
A.4 Proof of Theorem 1A	49
B Technical proofs	50
C Bibliography	55

1 Introduction

Modern statistical and machine learning modeling and techniques are often applied to datasets whose size and complexity call for scalable iterative stochastic algorithms, which are naturally deployed in a sequential and often data-driven manner. Concretely, suppose that we are interested in estimating a parameter θ^* using an iterative algorithm that outputs a sequence of estimators $\{\hat{\theta}_t\}_{t=0,1,2,\dots}$ computed in a sequential manner - that is, $\hat{\theta}_t$ is an update of $\hat{\theta}_{t-1}$ after acquiring or processing a new observation. At each iterate t , the accuracy of the algorithm is measured by the (unobservable) quantity

$$\text{Loss}(\hat{\theta}_t, \theta_*),$$

for an appropriate loss function $\text{Loss}(\cdot, \cdot)$. The standard approach in the literature to address the convergence properties of $\{\hat{\theta}_t\}_{t=0,1,2,\dots}$ is to establish high-probability and concentration bounds for the loss that hold at a fixed and large deterministic number of iterations. Specifically, for any given value of the iterate T (say larger than some value T_0) and a probability parameter $\delta \in (0, 1)$, the corresponding *fixed-time* concentration bound takes the form

$$\mathbb{P}(\text{Loss}(\hat{\theta}_T, \theta_*) \geq r_{\text{fixed-time}}(T, \delta)) \leq \delta, \quad \forall T \geq T_0, \quad (1)$$

with $r_{\text{fixed-time}}(T, \delta)$ a monotonically vanishing function of T that depends on δ and properties of the data-generating distribution (e.g., the dimension of the support).

While useful in providing a certificate of convergence, fixed-time concentration bounds however offer only a point-wise (in time) and thus weak guarantee. Instead, as illustrated in the recent and growing literature on any-time bounds and confidence sequences, it is more desirable and statistically safer to establish *time-uniform* or *any-time* concentration bounds of the form

$$\mathbb{P}(\text{Loss}(\hat{\theta}_t, \theta_*) \geq r_{\text{any-time}}(t, \delta), \forall t \geq T_0) \leq \delta, \quad (2)$$

where the, deterministic or random, time boundary function $t \mapsto r_{\text{any-time}}(t, \delta)$ is also vanishing in t . Any-time concentration bounds deliver significant algorithmic and statistical advantages over fixed-time ones.

First, they yield a stronger and arguably more natural notion of algorithmic convergence, as they ensure that the sequence of estimators $\{\hat{\theta}_t\}_{t=0,1,\dots}$ will remain close to the target parameter simultaneously over all large iterates t , with high probability. While this is presumably how practitioners would typically regard the convergence behavior of a stochastic algorithm, such interpretation is not warranted by the fixed-time guarantee (1).

Secondly, from a statistical viewpoint, any-time bounds enjoy remarkable features over traditional inference methods that make them well suited to handle data adaptive decisions, e.g. when to terminate the data collection process, stop the algorithm, fine tune and adjust model parameters on the fly based on the values observed so far. Specifically, they are valid under arbitrary stopping rules [55, 54], i.e. they remain valid at any data-dependent stopping time. What is more, they also enjoy *post-hoc validity* at arbitrary random (not just stopping) times. This is a crucial property in practical scenarios where data collection – in the case of sequential algorithms, the estimation task itself – may be interrupted or retrospectively adjusted. For example, practitioners may face unforeseen budget cuts requiring them to terminate an experiment at an unplanned or random time. Despite this change from the initial plan, any-time concentration bounds remain valid. Finally, any-time bounds are adaptive to random sample sizes: one may stop early when a convergence criterion is met or continue beyond the planned sample size, all while maintaining correct coverage and without requiring any corrections. To reiterate, any-time bounds provide broad protection against data-adaptive decision. We emphasize that “data-peeking” practices of this sort are widespread and, arguably, a fairly natural way to carry out data analytic tasks and sequential estimation, to formulate and test scientific hypothesis and to design or adjust experiments. However, they are incompatible with and, in fact, often invalidate conventional statistical methodologies that are not designed to be data-adaptive.

Any-time concentration bounds are certainly not new in concept: they are rooted in statistical sequential analysis [47, 31] and have been deployed extensively in the study of bandit algorithms. However, due to recent breakthroughs in the construction of time-uniform, martingale-based concentration inequalities due to [21, 54, 20, 55, 39], the past few years have witnessed a flurry of new results and applications on a multitude of problems, from mean estimation, A/B testing, reinforcement learning, and bandit optimization [21, 54, 55, 20]; see also Section 1.1 for more details.

While powerful and broadly applicable, this current prevailing approach relies in a fundamental way on the availability of martingale and exponential supermartingale processes (and maximal inequalities thereof), which combines naturally with the use of the Cramer-Chernoff method for deriving high probability bounds. For the purpose of studying sequential algorithms, these martingale processes are constructed based on the sequence of values $\{\text{Loss}(\hat{\theta}_t, \theta_*)\}_{t=0,1,\dots}$, or related quantities. While this is possible in many problems, there are important scenarios in which constructing an exponential supermartingale in this manner may be difficult or suboptimal. A primary example is Oja’s algorithm for online PCA, which is used to compute principal eigenvectors of an unknown covariance matrix Σ in a sequential manner [44, 45] based on, say, i.i.d. samples $\{X_t\}_{t \geq 1}$, revealed one at a time. In detail, algorithm is initialized with a random unit vector $\hat{v}_0 \sim \text{Uni}(\mathbb{S}^{p-1})$, and, for a given a sequence of step sizes $\{\eta_t\}_{t \geq 1}$, implements the online update rule

$$\hat{v}_t := (\mathbf{I}_p + \eta_t X_t X_t^\top) \hat{v}_{t-1}, \quad \hat{v}_t \leftarrow \frac{\hat{v}_t}{\|\hat{v}_t\|}. \quad (3)$$

For this problem, one may consider the function $\text{Loss}(\hat{v}_t, v_0) = \sin^2(\hat{v}_t, v_0)$, where v_0 is the leading eigenvector of Σ . It is well known that, under mild conditions, Oja’s iterates $\{\hat{v}_t\}_{t \geq 0}$ converge to v_0 [44, 45]. The key advantage of Oja’s algorithm, which operates in an online manner by processing one observation at a time, over batch methods is its memory and computational efficiency: it requires only $O(p)$ memory and $O(np)$ time, where p is the dimension of the data. The convergence properties

of Oja's iterates have been extensively studied over the past decades. In particular, assuming an i.i.d. and Markovian sequence of data points from a centered distribution with covariance matrix Σ , Oja's iterate at a *fixed number of iterations* has been shown to concentrate tightly around the principal eigenvector of Σ . See, e.g., [23, 27] and Section 3.4 for additional references and a short discussion.

Generalizing existing sharp fixed-time concentration bounds for Oja's algorithm to time-uniform bounds, i.e. holding simultaneously over an infinite time course, appears to be non-trivial. The main difficulty lies in the form (3) of the updates, which consist of matrix product updates. Unlike matrix addition, matrix products do not directly lead to an exponential supermartingale process, a representation that enables concentration. Indeed, the best known matrix product concentration bounds, due to [22], were obtained using different means, namely uniform smoothness properties of the Schatten trace classes. As a result, to the best of our knowledge, techniques to derive sharp any-time concentration bounds for Oja's algorithm are not available in the literature.

In this article, we propose a new technique for constructing anytime-valid concentration bounds for sequential stochastic algorithms for which the loss function at the current iterate satisfies certain recursive inequalities that can be viewed as an adaptation of those defining *almost supermartingale* processes put forward in the seminal paper [49]; see (6) for a precise definition. Thus, our settings do not require a well-defined martingale or supermartingale structure. In particular, Oja's algorithm is one example falling within this class. The key difference from the original almost supermartingale process of Robbins and Siegmund is that the form of the recursion we assume will guarantee not only almost sure convergence of the process, but almost sure convergence to zero. Remarkably, this simple adaptation is enough to deduce any-time concentration bounds.

Let us briefly introduce our framework, deferring a rigorous formulation to Section 2. We consider a nonnegative process $\{L_t; t \geq 0\}$ adapted to some filtration $\{\mathcal{F}_t; t \geq 0\}$ satisfying a recursive inequality of the form

$$L_t \leq (1 - \eta_t)L_{t-1} + U_t, \quad (4)$$

where $\{\eta_t; t \geq 1\}$ is a deterministic sequence of *stepsizes* taking values in $(0, 1)$, and $\{U_t, t \geq 0\}$ is an adapted noise process whose magnitude can be controlled by polynomials in the step sizes and the previous iterate, i.e. such that

$$\begin{cases} |\mathbb{E}[U_t | \mathcal{F}_{t-1}]| &\lesssim \sum_{i=1}^m \eta_t^{a_i} L_{t-1}^{b_i}, \\ |U_t| &\lesssim \sum_{i=1}^m \eta_t^{c_i} L_{t-1}^{d_i}, \end{cases} \quad (5)$$

where $\{a_i, b_i, c_i, d_i\}_{1 \leq i \leq m}$ are positive constants and the symbol \lesssim indicates equality up to some deterministic quantities, depending on the problem parameters. One should think of L_t as the value of the loss function at the t th iterate of a sequential algorithm, i.e. in the notation introduced above, $L_t = \text{Loss}(\hat{\theta}_t, \theta_*)$. Inequality (4) quantifies how much the loss changes at time t , after one iteration with step size η_t . The noise process U_t can be thought of as an extra term that prevents L_t from being a supermartingale and needs to be controlled. Concrete examples falling under this framework – that is, satisfying conditions (4) and (5) – include Oja's algorithm, stochastic gradient descent and Robbins-Monroe scheme for stochastic approximations, which we will analyze in Section 3.1. A general strategy used in the literature on the convergence analysis of iterative algorithms is to leverage inequalities of the form (4) and (5) to control the process U_t appropriately and finally demonstrate that the loss function decreases in probabilistic sense. In this paper we take a further step and formulate of a general technique to deduce not only stochastic convergence to zero of the loss process but also any-time concentration bounds for processes satisfying (4) and (5). Specifically, we make the following contributions.

- In Theorem 1 we show that the information provided by (4) and (5) is sufficient to yield a time-uniform concentration bound vanishing at the rate $O(\log \log t/t)$, using step sizes $\eta_t = \Omega(1/t)$, provided that L_0 is sufficiently small with high probability.
- The LIL-style $O(\log \log t/t)$ concentration rate is optimal among all processes that satisfy (4) and (5) with step sizes $\eta_t = \Omega(1/t)$. This is the content of Theorem 4.
- We apply our results to derive time-uniform bounds for SGD in the strongly convex case or under the Polyak–Lojasiewicz condition (see Section 3.1), two versions of Oja’s algorithm (see Section 3.4) and Robbins–Monroe approximation scheme (see Section 3.5). To the best of our knowledge, ours are the first time-uniform concentration bounds for the output of Oja-style algorithms. Our any-time concentration bounds are as sharp as the best known fixed-time bound established by [23] for streaming PCA, save for the unavoidable $\log \log t/t$ term.
- Interestingly, when applied to SGD, for which martingale-based time-uniform concentration bounds up to a fixed time horizon already exist [see, e.g. 46], our technique delivers slightly sharper bounds, in addition to being applicable to an infinite time course; see Section 2.3.
- Our analysis is tailored to problems with bounded noise, a setting that is commonly assumed in the literature. Extensions to the cases involving unbounded noise, in which the coefficients on the right-hand side in (5) is random and has sub-Gaussian tail, are presented in supplement.

The rest of the paper is organized as follows. Background and related results are discussed in Section 2.1. The main results are presented in Section 2, with applications to various algorithms given in Sections 3.1, 3.4, and 3.5. The lower bound is presented in Section 3.5. Conclusions, remarks, and directions for future research are provided in Section 5. All proofs and technical results are deferred to the remaining sections of the paper.

1.1 Related results

The analysis of stochastic gradient algorithms has been studied extensively over the past few decades. Recent advances in this direction have established non-asymptotic bounds and optimality results in a variety of settings. An incomplete list of recent works on convergence and limit theorems for the last iterate includes [7, 16, 34, 10, 38].

In contrast, time-uniform convergence results are much less understood. To the best of the authors’ knowledge, this topic has only been explored in [46, 17]. The results in [46] provide time-uniform bounds for the iterates of SGD in a strongly convex setting over a fixed time horizon, while [17] establishes related results for stochastic momentum algorithms using Ville’s inequality. To the best of our knowledge, no existing work provides a time-uniform analysis of Oja’s algorithm. Even for the last iterate, most existing results are limited to the case of bounded data.

2 Main results

2.1 Preliminaries

In order to illustrate the rationale behind our formulation and results, it is first useful to recall the classical Robbins–Siegmund’s lemma [49], a result about the almost-sure convergence of non-negative almost-supermartingales. In detail, suppose $\{L_t; t \geq 1\}$ is a sequence of non-negative stochastic process adapted to a filtration $\{\mathcal{F}_t; t \geq 1\}$ such that

$$\mathbb{E}(L_{t+1}|\mathcal{F}_t) \leq (1 + a_t)L_t + b_t - c_t \quad (6)$$

where a_t, b_t, c_t are non-negative random weights that are adapted to \mathcal{F}_t . A stochastic process that satisfies (6) is called an “almost-supermartingale”. It is clear that when a_t, b_t, c_t are identically zero, (6) reduces to the classical concept of a supermartingale. Another useful variant of (6), introduced in [32], takes the form

$$L_{t+1} \leq (1 + a_t)L_t + b_{t+1} - c_{t+1} + w_t \xi_t \quad (7)$$

where $\{\xi_t, \mathcal{F}_t\}$ is a martingale difference sequence and a_t, b_t, c_t, w_t are non-negative random weights adapted to \mathcal{F}_t .

Robbins-Siegmund’s lemma states that any almost-supermartingale in the sense of (6) converges almost surely to a limit X on the event

$$\left\{ \sum_{t=1}^{\infty} a_t < \infty, \sum_{t=1}^{\infty} b_t < \infty \right\}.$$

The same type of convergence (but with more relaxed assumptions) also holds for (7), though without the requirement that $\sum_{t=1}^{\infty} b_t < \infty$.

The most common applications of Robbins-Siegmund’s lemma arise when $L_t := \mathcal{L}(\hat{\theta}_t, \theta)$ for some loss function \mathcal{L} , and $\hat{\theta}_t$ being a sequence of estimator that is obtained via the stochastic gradient descent (SGD) algorithms; see, for examples, [47, 49, 32] and the references therein. It is often the case that X_t serves as the Lyapunov function associated with the ODE of the SGD, which is obtained by letting the step sizes diminish to zero in the noiseless setting. However, as we will see later, there are examples when the Lyapunov functions are hard to construct explicitly, and other choices of L_t are needed.

For both formulations (6) and (7) of the quasi-supermartingale process, the almost sure limit X needs not be zero. Indeed, without any additional assumptions, this limit can be arbitrary. For example, consider a simple situation in the deterministic setting with

$$L_t := \prod_{k=1}^t \left(1 + \frac{1}{k^2}\right).$$

It is easy to check that L_t converges to a positive limit. Moreover, by changing the terms $1/k^2$ to Y_k/k^2 , for some sequence of random variables $\{Y_k, k \geq 1\}$, one can get any positive limiting distribution. To ensure almost sure convergence to a zero limit, we need slightly different conditions than (6) and (7). In particular, the ideal recursion should encode the information that L_t is decreasing on average. For that reason, we will consider a specific class of non-negative, almost-supermartingale as follows.

Lemma 1 (Simplified Robbins-Siegmund’s lemma). *Let $\{L_t\}_{t \geq 1}$ be an adapted, non-negative process with respect to the filtration $\{\mathcal{F}_t; t \geq 1\}$. Suppose there exists a positive, deterministic sequence $\{\eta_t; t \geq 1\}$ and a non-negative, adapted sequence β_t such that*

$$\mathbb{E}(L_{t+1} | \mathcal{F}_t) \leq (1 - \eta_t)L_t + \beta_t \quad (8)$$

where $\{\eta_t, \beta_t\}_{t \geq 1}$ satisfies

$$\begin{aligned} \sum_{t=1}^{\infty} \eta_t &= \infty; \\ \sum_{t=1}^{\infty} \beta_t &< \infty \text{ almost surely}; \\ \lim_{t \rightarrow \infty} \beta_t / \eta_t &= 0. \end{aligned}$$

Then, $\lim_t L_t = 0$, almost surely.

Let us compare (8) with (6) and (7). In the formulation (8), we have assumed that the negative term c_t is at least $-\eta_t L_t$. In other words, on average, the decrease from L_t to L_{t+1} should be a non-negligible proportion of L_t . Moreover, the extra gain obtained from β_t is required to be of smaller order than η_t .

In the applications we consider here, the sequence η_t is the step size and β_t is the error induced after one update of the algorithms. It turns out that recursion of the form (10) or (8) hold for many classes of problems, which will be explained in Sections 3.4, 3.1 and 3.5. Note that the assumptions in Lemma 1 implies that the sequence of step sizes η_t 's satisfies

$$\sum_t \eta_t = \infty \text{ and } \sum_t \eta_t^2 < \infty. \quad (9)$$

It is easy to see that the best possible choice of step sizes such that (9) holds are those that are of order $\Omega(1/t)$.

Below we show that if the recursive conditions (8) hold for process $\{L_t\}$ itself – and not just for the process of the conditional expectations $\{\mathbb{E}(L_t | \mathcal{F}_{t-1})\}$ – then it is possible to obtain time-uniform concentration bounds. This is the main contribution of the paper.

2.2 A quantitative Robbins-Siegmund's Lemma and time-uniform concentration bounds

In this section, we formulate a quantitative version of 1 in the form of a anytime-valid concentration bounds for the entire process $\{L_t\}$ vanishing at the optimal LIL rate $O\left(\frac{\log \log t}{t}\right)$ rate. Towards that goal, we impose the following conditions.

Assumption 1 (Recursion). *Let $\{L_t; t \geq 1\}$ be a non-negative process and $\{U_t; t \geq 0\}$ a noise process, both adapted to the filtration $\{\mathcal{F}_t; t \geq 0\}$. Let $\{\eta_t; t \geq 1\}$ be a positive deterministic sequence of stepsizes such that*

$$\frac{C}{t} < \eta_t < 1/C$$

for all $t \geq 1$ and some constant $C_1 > 0$. Almost surely, for some positive constants $C_1 < C, C_2, C_3$ and $\{A_i; B_i; a_i; b_i; c_i; d_i\}_{i=1}^m$, the recursive conditions

$$L_t \leq (1 - C_1 \eta_t) L_{t-1} + U_t, \quad (10)$$

$$\left| \mathbb{E}(U_t | \mathcal{F}_{t-1}) \right| \leq C_2 \cdot \eta_t^2 + \sum_{i=1}^m A_i \cdot \eta_t^{1+a_i} L_{t-1}^{b_i}, \quad (11)$$

$$|U_t| \leq C_3 \eta_t \sqrt{L_{t-1}} + \sum_{i=1}^m B_i \cdot \eta_t^{1/2+c_i} \cdot L_{t-1}^{d_i}, \quad (12)$$

hold.

The settings in (10), (11), and (12) may initially seem unnatural. However, they arise frequently in the analysis of iterative algorithms:

- **Stochastic Gradient Descent (SGD):** see Section 3.1 for a detailed discussion. Roughly speaking, if we define $L_t^{SGD} \equiv \|x_t - x^*\|^2$, then (10), (11), and (12) hold in the form

$$L_t^{SGD} \leq (1 - 2\lambda \eta_t) L_{t-1}^{SGD} + 2\eta_t Y_t^{SGD} + B^2 \eta_t^2,$$

where $\mathbb{E}[Y_t^{SGD} | \mathcal{F}_{t-1}] = 0$, $|Y_t^{SGD}| \leq B \sqrt{L_{t-1}^{SGD}}$, and B is a constant independent of t .

- **Oja's Algorithm:** a variant of SGD tailored for PCA, which will be discussed in detail in Section 3.4. If we set $L_t^{Oja} \equiv \sin^2(\hat{v}_t, v)$, where \hat{v}_t is the output of the algorithm and v is the true principal eigenvector, then

$$L_t^{Oja} \leq (1 - 2\rho\eta_t) L_{t-1}^{Oja} + 2\rho\eta_t (L_{t-1}^{Oja})^2 + Q_t^{Oja} + 5B^4\eta_t^2 + 2\eta_t^3 B^6,$$

where $\mathbb{E}[Q_t^{Oja} | \mathcal{F}_{t-1}] = 0$, $|Q_t^{Oja}| \leq B^2 \sqrt{L_{t-1}^{Oja}}$, and B, ρ are constants independent of t .

Another variant of the Oja's algorithm is the Krasulina-Oja's algorithm; see Section 3.4 for more details. In the same setting as above, and with L_t^{Kra} being the loss function between the iterates of Krasulina-Oja's algorithm and the true parameter, we have

$$L_t^{Kra} \leq (1 - 2\rho\eta_t) L_{t-1}^{Kra} + 2\rho\eta_t (L_{t-1}^{Kra})^2 + Q_t^{Kra} + 5B^4\eta_t^2,$$

where $\mathbb{E}[Q_t^{Kra} | \mathcal{F}_{t-1}] = 0$, $|Q_t^{Kra}| \leq B^2 \sqrt{L_{t-1}^{Kra}}$, and B, ρ are constants independent of t . Note that, unlike Oja's algorithm, this version does not include a term of order η_t^3 .

- **Robbins-Monro scheme:** One aims to find the root of a univariate function $M(x)$ based on noisy observations $Y(x)$; see Section 3.5 for details. Under the sub-polynomial condition on M , Proposition 5 shows that the loss $L_t^{RM} := \|x_t - x^*\|^2$ associated with the iterates of the Robbins-Monro scheme satisfies the recursion

$$L_t^{RM} \leq (1 - R_1\eta_t) L_{t-1}^{RM} + Q_t^{RM} + \eta_t^2 \cdot P(L_{t-1}^{RM})$$

where $\mathbb{E}(Q_t^{RM} | \mathcal{F}_{t-1}) = 0$, $|Q_t^{RM}| = O(\eta_t \sqrt{L_{t-1}^{RM}})$ and P is some polynomial with positive coefficients.

Let us explain the roles of the parameters appearing in (11) and (12). The two terms $C_2\eta_t^2$ in (11) and $C_3\eta_t\sqrt{L_{t-1}}$ are ubiquitous; they arise in all examples considered in this paper. The remaining terms in (11) and (12) are problem-dependent and may or may not appear, depending on the specific context.

We will later see that in the simplest setting where L_t corresponds to the squared error loss of iterates produced by SGD in the strongly convex regime, equations (11) and (12) reduce to (16) and (17), which will be studied in greater detail in Section 2.3. However, for more complex problems, especially in the non-convex case, additional terms with different exponents can arise in (11) and (12), as illustrated by the example of Oja's algorithm discussed above.

The following general result shows that even when the noise U_t is extremely complicated, given a sufficiently good initialization L_0 , one can always construct a time-uniform bound with asymptotically optimal width, scaling as $\log \log t/t$.

Theorem 1. Suppose the positive process $\{L_t\}_{t \geq 0}$ satisfies the recursion (10) with noise process $\{U_t\}_{t \geq 1}$ and filtration $\{\mathcal{F}_t\}_{t \geq 0}$ such that conditions (11) and (12) hold. Assume additionally that

$$\min_{1 \leq i \leq m} \{(a_i + b_i) \wedge (c_i + d_i)\} > 1.$$

Then, for any $\delta \in (0, e^{-2})$, there exist constants $M, r > 0$ (independent of t) and a step-size sequence $\{\eta_t\}_{t \geq 1}$ such that

$$\mathbb{P}\left(\forall t \geq 0 : L_t \leq M \cdot \frac{\log(\delta^{-1}) + \log \log(t+10)}{t+10}\right) \geq 1 - 2\delta, \quad (13)$$

provided that

$$\mathbb{P}(L_0 \leq r) \geq 1 - \delta. \quad (14)$$

Moreover, the step sizes satisfy $\eta_t \asymp 1/t$ as $t \rightarrow \infty$.

A few comments and explanations are in order.

On the choice of the step sizes η_t . In the statement of Theorem 1 above, the term $\eta_t \asymp 1/t$ is interpreted in the sense that there exists a positive constant C' independent of t , but depends on δ and other parameters in (10), (11) and (12), such that

$$\frac{C'}{t} \leq \eta_t \leq \frac{1}{C't}.$$

We do not have an explicit construction for η_t in the abstract settings of Theorem 1. However, explicit construction of the step sizes is possible if the bounds on the right-hand sides of (11) and (12) are simpler. Such settings are investigated in Section 2.3 below.

On the constants M and r . The constants M and r in the statement of Theorem 1 is hard to construct explicitly based on the settings in (11) and (12). However, the proof of Theorem 1 reveals that for small δ , M and r scale like $O(\text{polylog}(\delta^{-1}))$ and $O(1/\text{polylog}(\delta^{-1}))$, respectively, with respect to the parameters in (11) and (12).

On the form of our assumptions. Let us elaborate on the assumptions of Theorem 1. The condition (11) requires that the conditional mean of the noise can be controlled as a function of the step sizes and the previous iterate. The $O(\eta_t^2)$ term in (11) is standard and commonly appears in analyses of SGD. We allow smaller-order terms in (11) to account for more intricate examples, such as Oja's algorithm, which will be discussed in Section 3.4. Similarly, the condition (12) is used to control the magnitude of the noise based on the previous iteration. Heuristically, for a sequence of admissible step sizes in the sense of (9), the optimal convergence rate (ignoring logarithmic factors) is of order $O(1/t)$. Substituting $L_{t-1} = O(1/t)$ and $\eta_t = O(1/t)$ into the right-hand side of (11), we obtain

$$|\mathbb{E}[U_t | \mathcal{F}_{t-1}]| \leq O(t^{-2}) + \sum_{k=1}^m O(t^{-1-a_k-b_k}) = O(t^{-2}),$$

provided that $a_k + b_k > 1$ for all $k \in \{1, \dots, m\}$. Thus, condition (11) ensures that the conditional expectation of the noise U_t decays at rate $O(t^{-2})$. Similarly, condition (12) implies that the conditional variance of U_t is of order $O(t^{-3})$ under the same assumption. Informally, this suggests that

$$L_T \lesssim \sum_{t>T} |\mathbb{E}[U_t | \mathcal{F}_{t-1}]| + \sqrt{\sum_{t>T} \text{Var}(U_t | \mathcal{F}_{t-1})} \lesssim T^{-1}$$

with high probability, for a single iterate. Of course, this is only a heuristic argument, and rigorous justification requires more substantial analysis. The $\log \log T$ factor arises from a peeling argument, similar in spirit to the classical law of the iterated logarithm.

Extension to random constants. In (11) and (12), we assume that all parameters are constant. However, it is possible to extend Theorem 1 to a slightly more general setting where A_i and B_i are positive random variables, provided their conditional moment generating functions are sub-exponential with deterministic sub-exponential parameters. The step sizes stated in the statement of Theorem 1 are piecewise constant and asymptotically of order $O(1/t)$. However, in this abstract setting, we are unable to provide an explicit expression for the sequence $\{\eta_t; t \geq 1\}$. A special and important case where the step sizes can be made explicit is presented in Subsection 2.3.

Optimality of the bound. The LIL-style rate $\log \log t/t$ in (13) is sharp and matches that of the classical law of the iterated logarithm for partial sum of i.i.d. random variables. The same rate has been obtained in [21] in the context of empirical Bernstein's inequality. We also show that for stochastic approximation with squared-error loss, the same rate is optimal among all the choices of step sizes of order $\Omega(1/t)$; see Section 3.5 for more details. There have been several attempts to obtain quantitative versions of Robbins-Siegmund's lemma; see [43, 25, 36, 50] and the references therein. However, none of these works successfully capture the $\log \log t$ term in the bound. To fully

capture this subtle logarithmic term, we employ a novel, sharp error bound in short-time intervals and a ‘‘stitching’’ argument.

Relationship with supermartingales. A natural but naive approach would be to construct a non-negative supermartingale directly from the recursive estimate (10). To the best of our knowledge, such a construction is not applicable to the settings in Theorem 1 due to the highly nonlinear nature of (11) and (12). In the special case where the noises U_t ’s are deterministic and of order $O(\eta_t^2)$, it is indeed possible to construct a non-negative supermartingale directly; see [18, 32] and the references therein. However, this construction breaks down when the noises are random and more complicated, as it involves an infinite product starting at time n and running to infinity, which in turn leads to measurability issues.

On the necessity of condition (14). Interestingly, though the conditions stated in Theorem 1 guarantee that L_t has an almost sure limit when a_i ’s are greater than 1, they do *not* guarantee the convergence of L_t to 0 almost surely. For example, consider the recursion

$$L_t \leq (1 - \eta_t)L_{t-1} + 2\eta_t L_{t-1}^2 + \eta_t^2. \quad (15)$$

It is easy to see that the recursion above is of the form (10) and satisfies (11), (12) with $C_2 \equiv 1$, $C_3 \equiv 0$, $(A_1, a_1, b_1) \equiv (2, 0, 2)$, $(B_1, c_1, d_1) \equiv (2, 1/2, 2)$ and $(B_2, c_2, d_2) \equiv (1, 3/2, 0)$. However, if one takes a Bernoulli random variable Y such that $P(Y = 1) = 1/10$, then the process $L_t \equiv Y$ clearly satisfies (15) but

$$\mathbb{P}\left(\lim_{t \rightarrow \infty} L_t = 0\right) = 9/10.$$

The problem with this counter example is that L_0 can be large with non-negligible probability. Therefore, to rule out such counter examples, one should only expect a time-uniform bound of the form (13) under (14), where r is sufficiently small and can be determined by the parameters in (11) and (12). As a result, Theorem 1 captures the stable dynamic of a general class of iterated algorithms: once the output of the algorithm gets close to the minimizer, it will stay close to the minimizer within a neighborhood of optimal width $\log \log t/t$.

2.3 An explicit step size construction

Although Theorem 1 provides a general affirmative result regarding the time-uniform bound with optimal width, the parameters involved in the construction cannot be made explicit as a function of time t . To address this issue, we provide an explicit construction in a simpler setting that covers all the applications considered in the subsequent sections.

For simplicity, let us now focus on a simplified form of (11) and (12). Suppose the process $\{L_t; t \geq 0\}$ satisfies (10) such that the noise process U_t satisfying

$$\left| \mathbb{E}(U_t | \mathcal{F}_{t-1}) \right| \leq C_2 \cdot \eta_t^2; \quad (16)$$

$$|U_t| \leq C_3 \eta_t \sqrt{L_{t-1}} + C_2 \cdot \eta_t^2. \quad (17)$$

Conditions (16) and (17) correspond to the most common setting of SGD under a strongly convex function. This context will be further analyzed in Section 3.1 below. We first state a maximal inequality with explicit parameters.

As we will see shortly in Section 3.1, not only the maximal inequality in Proposition 1 can be used to derive time-uniform bounds, it can also improve the existing last iteration bound for the classical SGD. Let $\{L_t; t \geq 0\}$ satisfies (10) such that the noise process U_t satisfying (16) and (17), then

Proposition 1 (maximal inequality). *Assume that $\eta_t = 2 / (C_1(t + L))$ for all $t \in [t_0 + 1, t_1]$ and some positive integer $L \geq 3$. Suppose $A \subset \{L_{t_0} \leq a\}$ such that $\mathbb{P}(A) > 0$. Then, for any $\delta > 0$, one has*

$$\mathbb{P}\left(\exists t \in [t_0 + 1, t_1] : L_t \geq \frac{M(t_1 - t_0) \log(\delta^{-1})}{(t+3)^2} \middle| A\right) \leq \frac{\delta}{\mathbb{P}(A)}$$

where

$$M = \frac{31.5 \times (L - 1)}{L} \max \left\{ \frac{aL(L - 1)}{\log(\delta^{-1})(t_1 - t_0)}, \frac{C_2}{C_1^2 \log(\delta^{-1})}, \frac{C_2}{C_1^2 \sqrt{\log(\delta^{-1})}}, \frac{C_3^2}{C_1^2} \right\}.$$

Proposition 1 gives a maximal inequality for the sequence $\{L_t\}$ over the time interval $[t_0, t_1]$ when the step sizes are set to be $\eta_t = 2/(C_1(t + L))$. If one sets $t_0 \equiv 0$ and suppose that $L_{t_0} \leq a$ almost surely, then for all δ sufficiently small, we have

$$L_{t_1} \leq 31.5 \times \frac{L - 1}{L} \cdot \frac{C_3^2 \log(\delta^{-1})}{C_1^2 t_1} \quad (18)$$

with probability at least $1 - \delta$.

In the context of SGD, (18) implies that the last iterate is of order $1/t$, which is an improvement of some results in literature. This will be discussed in more details in Section 3.1. Proposition 1 allows us obtain the time-uniform bound in the same settings.

Theorem 2. *Consider the settings in Proposition 1. Then,*

$$\mathbb{P} \left(\exists t \geq 0 : L_t \geq 31.5 \times K \max \left\{ \frac{aL}{\log(\delta^{-1})}, \frac{C_2}{C_1^2}, \frac{C_3^2}{C_1^2} \right\} \frac{\log(\delta^{-1}) + 2 \log \log(t + 9)}{t + L} \right) \geq 1 - 2\delta$$

if

$$\mathbb{P}(L_0 \leq a) \geq 1 - \delta,$$

and

$$K := \max \{L - 2; 32\} \cdot (\mathbf{1}_{\{L \geq 32\}} + 32 \cdot \mathbf{1}_{\{L \leq 31\}}). \quad (19)$$

Theorem 2 provides a time-uniform bound for any process $\{L_t\}_{t \geq 0}$ satisfying conditions (10), (16), and (17), given an initial upper bound a on L_0 that holds with probability at least $1 - \delta$. In contrast to Theorem 1, we do not require a to be sufficiently small, and both the constants and step sizes in the construction are made explicit. This should not be surprising: under the conditions (10), (16), and (17), one can show that L_t converges to 0 almost surely for any admissible sequence of step sizes $\{\eta_t; t \geq 1\}$ in the sense of (9). The requirement that a be sufficiently small, as in Theorem 1, arises only in the *inhomogeneous* setting, where the right-hand side contains terms of order η_t^k for some $k < 2$.

3 Applications

As anticipated, we will exemplify the general time-uniform concentration bounds of Theorem 1 in three important applications.

3.1 SGD algorithms

We consider standard settings for the analysis of SGD algorithms, which we briefly review; see, e.g., [42, 46] (see also the references therein). Throughout, \mathcal{X} denotes a compact, convex set in \mathbb{R}^d with non-empty interior and $F : \mathcal{X} \rightarrow \mathbb{R}$ a real valued function on it. We will consider several assumptions on F .

- *Strong convexity.* F is said to be λ -strongly convex (λ -SC) if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x}, \nabla F(\mathbf{x}) \rangle + \frac{\lambda}{2} \|\mathbf{y} - \mathbf{x}\|^2. \quad (\lambda\text{-SC})$$

- *Smoothness.* F is said to be μ -smooth if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x}, \nabla F(\mathbf{x}) \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2. \quad (\mu\text{-smooth})$$

- *Polyak-Lojasiewicz condition.* F is said to satisfy the Polyak-Lojasiewicz (PL) condition with parameter $\tau > 0$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$\|\nabla F(\mathbf{x})\|^2 \geq \tau (F(\mathbf{x}) - F(\mathbf{x}^*)) \quad (\text{P-L})$$

where $\mathbf{x}^* \in \text{int}(\mathcal{X})$ is the global minimum.

We are interested in solving the optimization problem

$$\min_{\mathbf{x} \in \mathcal{X}} F(\mathbf{x}) \quad (20)$$

where $F(\mathbf{x}) := \mathbb{E}(f(\mathbf{x}, \xi))$, ξ is a random vector from a distribution P_ξ supported in \mathbb{R}^{d_1} and $f(\mathbf{x}, \xi)$ is a convex function on $\mathcal{X} \times \mathbb{R}^{d_1}$, differentiable in \mathbf{x} for almost surely all realizations of ξ . Under this assumption, it is easy to check that F is a convex differentiable function on \mathcal{X} .

In the SGD framework, one has access to a series of unbiased estimators of the true gradient of F by sampling ξ repeatedly. Denote by $\Pi_{\mathcal{X}} : \mathbb{R}^d \rightarrow \mathcal{X}$ the projection map

$$\Pi_{\mathcal{X}}(\mathbf{y}) := \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|.$$

It is well-known that $\Pi_{\mathcal{X}}$ is well-defined and is a 1-Lipchitz map when \mathcal{X} is a compact convex set. The projected SGD algorithm for solving the problem (20) is defined iteratively as

$$\mathbf{x}_t := \Pi_{\mathcal{X}}\left(\mathbf{x}_{t-1} - \eta_t g(\mathbf{x}_{t-1}, \xi_t)\right) \quad (21)$$

where $\{\xi_t; t \geq 1\}$ is a sequence of i.i.d. samples from P_ξ , $\mathbf{x}_0 \in \mathcal{X}$ is a starting point and $g \in \mathbb{R}^d$ and noise vector satisfying

$$\mathbb{E}\left(g(\mathbf{x}_{t-1}, \xi_t) \mid \xi_1, \xi_2, \dots, \xi_{t-1}\right) = \nabla F(\mathbf{x}_{t-1}),$$

and

$$\max\{\|g - \nabla F\|; \|g\|\} \leq B$$

almost surely, for some positive constant $B > 0$.

3.2 Time-uniform bounds under (λ -SC)

Suppose the function F satisfies (λ -SC). Letting $L_t^{SGD} = \|\mathbf{x}_t - \mathbf{x}^*\|^2$, we immediately have the following recursion.

Proposition 2. *For any $t \geq 1$, we have*

$$L_t^{SGD} \leq (1 - 2\lambda\eta_t) L_{t-1}^{SGD} + 2\eta_t Y_t^{SGD} + B^2 \eta_t^2 \quad (22)$$

where $\mathbb{E}(Y_t^{SGD} | \mathcal{F}_{t-1}) = 0$ and $|Y_t| \leq B \sqrt{L_{t-1}^{SGD}}$.

From Proposition 2, we observe that the process $\{L_t; t \geq 0\}$ satisfies a recursion of the form (10), with a noise process that satisfies conditions (16) and (17), along with the corresponding constants

$$C_1 \equiv 2\lambda; C_2 \equiv B^2; C_3 \equiv 2B.$$

An application of Theorem 2 immediately yields time-uniform concentration bound for SGD.

Corollary 1. Set the step sizes $\eta_t = 1/(\lambda(t+32))$, then

$$\mathbb{P}\left(\forall t \geq 1 : L_t^{SGD} \leq 1008 \cdot \frac{B^2}{\lambda^2} \cdot \frac{\log(\delta^{-1}) + 2 \log \log(t+9)}{t+32}\right) \geq 1 - \delta$$

for any $\delta > 0$.

Corollary 1 is a direct application of Theorem 2 and the fact that

$$L_t^{SGD} \leq \frac{B^2}{\lambda^2}$$

almost surely, for all $t \geq 1$. This simple fact follows from the boundedness assumption and the strong convexity assumption; see Lemma 5 in [46] for a proof. We can see that Corollary 1 provides a time-uniform bound with optimal width with a larger constant than the convergence rate for the last iterate and an extra $\log \log$ factor, which is optimal and can not be removed.

It is helpful to compare the previous bound with an analogous one given in Proposition 1 of [46], which, under the same settings, states that for a fixed time $T \geq 3$, and setting $\eta_t = \frac{1}{\lambda t}$,

$$\mathbb{P}\left(\forall 1 \leq t \leq T : L_t^{SGD} \leq 624 \cdot \frac{B^2}{\lambda^2} \frac{(\log(\delta^{-1}) + \log \log T)}{t}\right) \geq 1 - \delta. \quad (23)$$

The previous bound is, of course, not anytime-valid, as it requires the specification of a terminal time T . In contrast, 1 provides time-uniform guarantees while exhibiting the same dependence on the model-related parameters $\frac{B^2}{\lambda^2}$.

Not only our method gives any-time concentration bound, it can also yield insight on the last iteration of SGD. To be more precise, we have

Corollary 2. Set $\eta_t = 1/(\lambda(t+3))$. Then, for any $\delta > 0$ and $t \geq 1$,

$$\mathbb{P}\left(L_t^{SGD} \leq \frac{21B^2}{\lambda^2} \cdot \frac{\log(\delta^{-1})}{t+3}\right) \geq 1 - \delta.$$

In comparison with (23), our result in Corollary 2 improves the convergence rate of the last iterate by a factor of $\log \log T$. In the next subsection, we will consider a non-convex setting involving the popular P-L condition.

Let us end this subsection with an example regarding least square.

Sequential ridge regression via SGD

To illustrate the effectiveness and generality of our time-uniform bounds, we show how Corollary 1 immediately yields anytime concentration bounds of optimal width based on the SGD iterates for the problem of estimating in a sequential manner the parameters of a standard regression linear model. In detail, we assume that we observe sequentially a stream of i.i.d. of response/covariates pairs $(\{y_t, \mathbf{x}_t\}_{t=1,2,\dots})$ obeying the linear regression model

$$y_t = \langle \boldsymbol{\theta}^*, \mathbf{x}_t \rangle + \xi_t. \quad (24)$$

Above, the unknown parameter $\boldsymbol{\theta}^*$ belongs to a known convex set $\Omega \subset \mathbb{R}^d$ such that $\text{diam}(\Omega) = D$ and the random covariates and noise are bounded: for all t ,

$$\max\{\|\mathbf{x}_t\|, |\xi_t|\} \leq B \quad \text{almost surely.}$$

We are interested in estimating θ^* using the sequential ridge regression estimator obtained as the output of the SGD iterates applied to the minimization problem

$$\min_{\theta \in \Omega} \left\{ \frac{1}{2} \mathbb{E}[(y - \langle \theta, \mathbf{x} \rangle)^2] + \frac{\lambda}{2} \|\theta\|^2 \right\},$$

where $\lambda \geq 0$ is a fixed regularization parameter and the pair (y, \mathbf{m}) is a draw from the data generating distribution obeying the linear model (24). The case $\lambda = 0$ corresponds to the least squares minimization.

In this setting, the SGD algorithm initialized at an arbitrary $\hat{\theta}_0$ yields the update rule

$$\hat{\theta}_t = \text{Proj}_{\Omega} \left(\hat{\theta}_{t-1} - \eta_t \mathbf{x}_t (\mathbf{x}_t^\top \hat{\theta}_{t-1} - y_t) + \lambda \hat{\theta}_{t-1} \right), \quad (25)$$

for a given sequence of step sizes $\{\eta_t\}_{t \geq 1}$. Then, assuming that

$$\lambda_{\min} := \lambda_{\min}(\mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^\top]) > 0,$$

with the choice of step sizes $\eta_t = 2/(\lambda_{\min}(t+32))$, Corollary 1 implies that, with probability at least $1 - \delta$ uniformly over all $t \geq 1$,

$$\|\hat{\theta}_t - \theta^*\|^2 \leq \frac{\lambda^2 \|\theta^*\|^2}{\lambda_{\min}^2} + 1008 \cdot \frac{B_1^2}{\lambda_{\min}^2} \cdot \frac{\log(\delta^{-1}) + 2 \log \log(t+9)}{t+32}, \quad (26)$$

where

$$B_1 := B^2 D + B^2 + \lambda D + \lambda \|\theta^*\|.$$

Note that when the penalty parameter λ is zero, the bound does not depend on $\|\theta^*\|$. This is not unexpected, since the SGD procedures in this case converges to the ordinary least squares estimator, which is unbiased in well-specified settings.

To the best of our knowledge, the above bound is new in the SGD literature, both because of its time-uniform form and because of its optimal rate of decay in t . It is illustrative to compare the above bound with existing anytime concentration bound for the online ridge estimator obtained by [1, 2]; see also [33, 8, 40] and the references therein. Let

$$\begin{aligned} \hat{\theta}_t^{\text{Ridge}} &:= (\mathbf{x}_{1:t}^\top \mathbf{x}_{1:t} + \lambda I)^{-1} \mathbf{x}_{1:t}^\top \mathbf{y}_{1:t}, \\ \bar{V}_t &:= \lambda I + \sum_{s=1}^t \mathbf{x}_s \mathbf{x}_s^\top, \\ \|\hat{\theta}_t^{\text{Ridge}} - \theta^*\|_{\bar{V}_t} &:= \sqrt{\langle \hat{\theta}_t^{\text{Ridge}} - \theta^*, \bar{V}_t (\hat{\theta}_t^{\text{Ridge}} - \theta^*) \rangle}. \end{aligned}$$

Here, $\mathbf{x}_{1:t}$ denotes the matrix with rows $\mathbf{x}_1^\top, \dots, \mathbf{x}_t^\top$. Then, Theorem 2 in [1] establishes that

$$\mathbb{P} \left(\forall t \geq 1 : \|\hat{\theta}_t^{\text{Ridge}} - \theta^*\|_{\bar{V}_t} \leq B \sqrt{d \log \left(\frac{1 + t B^2 / \lambda}{\delta} \right) + \sqrt{\lambda} \|\theta^*\|} \right) \geq 1 - \delta. \quad (27)$$

Though both concentration bounds (26) and (27) are anytime-valid, they differ in several important aspects. First, the bound by [1] is with respect to the Mahalanobis distance using the sample covariance matrix, as it is based on a self-normalization approach [13, 14] that is specifically well-suited to the linear regression settings and remain valid under minimal assumptions on the data generating distribution; in particular, the data need not be i.i.d. and can be chosen arbitrarily in a data-adaptive manner. In contrast, our bound was obtained as corollary to a general time-uniform concentration bound for almost-super martingale-type processes (namely, Theorem 2) that was not specifically designed for linear regression.

The conditions we impose for our bound, namely that the errors and the covariates are bounded and drawn in an i.i.d. manner, are of course restrictive compared to ones assumed by [1]. However, it is worth mentioning that they are in fact standard conditions in the literature on SGD, a methodology that should not be expected to work well in data-adaptive settings where the next observation is chosen in a predictable manner¹

A second important difference is that our bound (26) is concerned with a sequential, SGD-based estimator $\hat{\theta}_t$ given in (25), a one-pass procedure with a computational complexity $O(d)$ per iteration. In contrast, the more expensive online ridge estimator in (27) is based on the inverse of the regularized Gram matrix \bar{V}_t at every t , which, using the Sherman-Morrison formula, can be computed as a rank-one-update with $O(d^2)$ operations at each iteration.

Finally, in terms of scaling in t , the (squared) bound in (27) is of order $(\log t)/t$, while our bound in (26) is of order $(\log \log t)/t$, thereby recovering the optimal logarithmic dependence, as suggested by Theorem 4 below. Recently, in the context of deriving anytime valid concentration inequalities for self-normalized multivariate processes, [56] have sharpened the results in [1, 2] and recovered the optimal scaling $(\log \log t)/t$ for i.i.d. data; see Theorem 5.2 of therein. Thus, our result implies that, in the i.i.d settings in which SGD is usually deployed, the one-pass SGD-based estimator (25) achieves the same rate as the estimator computed using the full covariance matrix, while being more efficient in terms of computational cost.

3.3 Uniform error bound under $(\mu\text{-smooth})$ and $(\mathbf{P}\text{-L})$

We now turn to the analysis of SGD in non-convex problems, where we assume that $(\mu\text{-smooth})$ and $(\mathbf{P}\text{-L})$ hold. In this setting, we are interested in estimating the minimal value $F(\mathbf{x}^*)$ of the objective function at any minimizer \mathbf{x}^* .

Given the sequence $\{\mathbf{x}_t, t \geq 0\}$ of projected SGD iterations in (21) we let

$$L_t^{PL} := F(\mathbf{x}_t) - F(\mathbf{x}^*).$$

Similar to Proposition 2, the corresponding recursion in this case is given next.

Proposition 3. *For any $t \geq 1$, we have*

$$L_{t+1}^{PL} \leq (1 - \tau\eta_t)L_t^{PL} + \eta_t Y_t^{PL} + 2\mu B^2 \cdot \eta_t^2 \quad (28)$$

where

$$Y_t^{PL} := \langle \nabla F(\mathbf{x}_t), \nabla F(\mathbf{x}_t) - \hat{g}(\mathbf{x}_t) \rangle.$$

Moreover, $\mathbb{E}(Y_t^{PL} | \mathcal{F}_{t-1}) = 0$ and $|Y_t^{PL}| \leq B\sqrt{\mu} \cdot \sqrt{L_t^{PL}}$.

Then, Proposition 1 and Theorem 2 yield the following fixed and anytime-valid concentration bounds.

Corollary 3. *Recall τ in $(\mathbf{P}\text{-L})$. Consider the optimization problem (20) with the iterates (21).*

(i) *Set the step sizes $\eta_t = 2/(\tau(t+3))$. For any $\delta \in (0, e^{-4})$ and $t \geq 3/\log(\delta^{-1})$,*

$$\mathbb{P}\left(F(\mathbf{x}_t) - F(\mathbf{x}^*) \geq \frac{21\mu B^2}{\tau^2} \cdot \frac{\log(\delta^{-1})}{t+3}\right) \leq \delta.$$

(ii) *Set the step sizes $\eta_t = 2/(\tau(t+32))$. Then, uniformly over all time $t \geq 0$,*

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) \leq 1008 \cdot \max\left\{\frac{128B^2}{\tau \log(\delta^{-1})}, \frac{2B^2\mu}{\tau^2}\right\} \cdot \frac{\log(\delta^{-1}) + 2\log\log(t+9)}{t+32}$$

with probability at least $1 - \delta$.

¹Indeed, consistency of the SGD iterates is predicated on the assumption that the gradient of the target function can be estimated unbiasedly using a new observation, which rules out virtually any non-i.i.d. data streams.

Corollary 3 is an immediate application of Proposition 1 and 2 and the fact that for all $t \geq 0$,

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) \leq \frac{1}{\tau} \|\nabla F(\mathbf{x}_t)\|^2 \leq \frac{4B^2}{\tau}$$

which follows from (P-L).

To the best of our knowledge, the time-uniform bound for the error under (P-L) and (μ -smooth) in Corollary 3 is the first in the literature. Regarding the convergence of the last iterate, a similar result but without explicit constants was recently obtained by [38].

3.4 Oja's algorithm for PCA

Principal Component Analysis (PCA) is one of the most fundamental problems in multivariate analysis, where the goal is to estimate a low-dimensional subspace from observed data. In this section, we focus on the sequential setting, in which estimation is performed as data points arrive one by one. Accordingly, the results are presented in the fixed-dimensional setting. However, as we shall see, the effect of dimensionality appears only through a logarithmic factor, making the proposed method well-suited for large datasets.

Let $X_1, X_2, \dots, X_n \in \mathbb{R}^p$ be i.i.d. centered random vectors with covariance matrix Σ such that:

- $\|X_i\| \leq B$ almost surely.
- $\lambda_1 > \lambda_2$, where $\{\lambda_i\}_{i=1}^p$ are the eigenvalues of Σ sorted in descending order.

The goal of PCA is to estimate the principal component, i.e., the eigenvector associated with the largest eigenvalue λ_1 . In optimization terms, this corresponds to solving

$$\mathbf{v}_1 := \underset{\|\mathbf{v}\|=1}{\operatorname{argmin}} -\mathbf{v}^\top \Sigma \mathbf{v}. \quad (29)$$

In statistical literature, a natural approach is to use the leading eigenvector of the empirical covariance matrix,

$$\hat{\Sigma}_n := \frac{1}{n} \sum_{k=1}^n \mathbf{X}_k \mathbf{X}_k^\top.$$

However, this method is computationally expensive for large datasets and it requires storing all data points, which is often impractical due to memory constraints.

Oja's algorithm, initially proposed in [44, 45], is an efficient method for solving (29) that requires only $O(p)$ memory and has a runtime of $O(np)$. Oja's idea was to recast the problem (29) as an optimization problem and apply stochastic gradient descent (SGD) to solve it iteratively. Interestingly, the objective function is concave, and hence highly non-convex. Nevertheless, the optimization landscape is simple and free of spurious local minima, making this one of the simplest non-convex problems that can be solved efficiently using SGD, both in theory and in practice.

An incomplete list of early results includes [24, 15, 19, 5, 51, 41, 4]. More recent works, which establish optimal rates and extensions to various settings, can be found in [23, 37, 28, 3, 12, 26]; see also the references therein. However, to the best of our knowledge, there are no existing results on deriving any-time bounds or confidence sequences for the sequence of estimators $\{\hat{\mathbf{v}}_t; t \geq 1\}$ defined above. A natural approach to obtaining such bounds is to apply matrix concentration inequalities. However, such methods often yield suboptimal convergence rates in t . This is not surprising: achieving the optimal convergence rate for Oja's algorithm requires a two-stage step-size schedule (see [23] for details): a chaotic phase with constant step sizes, during which the algorithm explores the sphere to locate the region containing the eigenvector, followed by a stable phase in which the algorithm

achieves the $O(t^{-1})$ convergence rate. Consequently, concentration inequalities applied to a single step-size schedule of the form $\eta_t = A/(t + B)$ are unlikely to yield the optimal rate.

To describe Oja's algorithm, we first reformulate (29) as a stochastic optimization problem, where the population covariance matrix is replaced by its empirical counterpart:

$$\hat{\mathbf{v}}_n := \underset{\|\mathbf{v}\|=1}{\operatorname{argmin}} -\mathbf{v}^\top \hat{\Sigma}_n \mathbf{v}. \quad (30)$$

Note that the optimization problem (30) is *non-convex*. However, the landscape of the objective function is relatively simple: its population version (29) has no spurious local minima. In fact, this is one of the simplest non-convex problems that can be solved efficiently. In what follows, we describe the original Oja's algorithm and a common variant, known as the Krasulina-Oja algorithm.

(i) Oja's algorithm. Initialize a vector $\hat{\mathbf{v}}_0 \sim \text{Uni}(\mathbb{S}^{p-1})$. Let $\{\eta_t\}_{t \geq 1}$ be a sequence of step sizes. The update rule of Oja's algorithm is given by:

$$\hat{\mathbf{v}}_t := (\mathbf{I}_p + \eta_t \mathbf{X}_t \mathbf{X}_t^\top) \hat{\mathbf{v}}_{t-1}, \quad \hat{\mathbf{v}}_t \leftarrow \frac{\hat{\mathbf{v}}_t}{\|\hat{\mathbf{v}}_t\|}. \quad (\text{Oja})$$

Under the standard assumption (9) on the step sizes, it can be shown that Oja's algorithm converges almost surely to either \mathbf{v}_1 or $-\mathbf{v}_1$. Moreover, under (9), the normalization step can be omitted in practice, and it suffices to normalize only the output in the last iteration.

(ii) Krasulina-Oja's algorithm. Initialize a vector $\hat{\mathbf{v}}_0 \sim \text{Uni}(\mathbb{S}^{p-1})$. Let $\{\eta_t\}_{t \geq 1}$ be a sequence of step sizes. The update rule of Krasulina-Oja's algorithm is given by:

$$\hat{\mathbf{v}}_t = \hat{\mathbf{v}}_{t-1} + \eta_t \underbrace{\left[\mathbf{X}_t - \frac{\mathbf{y}_t}{\|\hat{\mathbf{v}}_{t-1}\|^2} \cdot \hat{\mathbf{v}}_{t-1} \right]}_{z_t} \quad (\text{Kra-Oja})$$

where

$$\mathbf{y}_t := \mathbf{X}_t^\top \hat{\mathbf{v}}_{t-1}. \quad (31)$$

The assumption that the \mathbf{X}_i 's are centered is not essential and can be removed with a simple modification. For example, in (Oja), instead of updating with a single data point at each iteration, one may use two data points and modify the update rule as

$$\hat{\mathbf{v}}_t := \left[\mathbf{I}_p + \frac{\eta_t}{\sqrt{2}} (\mathbf{A}_t - \mathbf{B}_t)(\mathbf{A}_t - \mathbf{B}_t)^\top \right] \hat{\mathbf{v}}_{t-1},$$

where \mathbf{A} and \mathbf{B} are two independent copies of \mathbf{X}_1 and are independent of the history up to time $t - 1$. This modification ensures that the update is centered, regardless of the original mean. For the sake of simplicity, we will assume throughout the rest of the paper that the data are centered.

The difference between (Kra-Oja) and (Oja) is the second order correction term in z_t . One advantage of such a term is that the update is orthogonal to the previous iteration. It is also known that under the admissible condition (9), the algorithm (Kra-Oja) output a sequence $\{\hat{\mathbf{v}}_t; t \geq 1\}$ such that

$$\lim_{t \rightarrow \infty} \sin^2 \left(\frac{\hat{\mathbf{v}}_t}{\|\hat{\mathbf{v}}_t\|}, \mathbf{v}_1 \right) = 0.$$

Here the sine-squared error loss $\sin^2(\mathbf{x}, \mathbf{y})$ is defined as

$$\sin^2(\mathbf{x}, \mathbf{y}) = 1 - \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$.

The convergence analysis of (Kra-Oja) and (Oja) has been extensively studied in the literature. Convergence results for the principal eigenvector can be found in [15, 51, 52], along with additional references therein. In contrast, results for k -PCA are relatively scarce; see [3, 23] and the references therein. Recent developments in Oja's algorithm include a bootstrap approximation for the sine-squared error [37] and a variant designed specifically for sparse PCA settings [28].

Despite the extensive literature on the convergence analysis of the final iteration of Oja's algorithm, little is known about proving time-uniform bounds for its iterates. There are two fundamental difficulties in establishing such a result. First, the dynamics of the algorithm are unstable during the initial iterations, providing little useful information about the principal eigenvector. Second, the outputs in (Oja) and (Kra-Oja) involve products of independent random matrices, which lack an immediately exploitable martingale structure. Indeed, while a *fixed* number of products of independent random matrices does exhibit concentration properties, as recently demonstrated by [22], it is not obvious how to extend such sharp bounds to a random number of products.

It is worth noting that the $(1 - \delta)$ time-uniform bound for Oja's algorithm cannot be obtained by simply applying a union bound over fixed-time convergence results. The difficulty arises because the step sizes themselves depend on δ , which alters the solution path if one attempts to apply last-iterate convergence results sequentially over a collection of probability levels. In contrast, the step-size construction in Section 3.1, corresponding to the SGD case, is independent of δ .

The first difficulty can be resolved by using the smoothing technique in [23] to show that after an initial uncertainty phase of order $T_0 := \Omega(B\delta^{-2}(\lambda_1 - \lambda_2)^{-2})$ steps, the output will be within a ball of radius $(T - T_0)^{-1/2}$ of the principal eigenvector. However, the second difficulty appears to be intractable using existing techniques, as constructing a non-negative supermartingale solely based on the matrix product remains difficult.

We will now demonstrate how Theorem 2 can be used to resolve the second difficulty. Let us first start with an useful observation: without loss of generality, one can assume Σ to be a diagonal matrix, i.e.

$$\Sigma = D = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_p\}. \quad (32)$$

Indeed, for an arbitrary Σ , one can write $\Sigma = ODO^\top$, for some orthogonal matrix O . Thus, the rotated data OX_1, \dots, OX_n has covariance matrix D . Furthermore, note that (Oja) implies

$$O\hat{v}_t := (\mathbf{I}_p + \eta_t X_t X_t^\top) O\hat{v}_{t-1}.$$

This means that the output with respect to the rotated data is also rotated by the same matrix O . Thus, if the algorithm converges for D , it also converges for Σ . The same argument applies for (Kra-Oja) as well. Therefore, we assume (32) and $v_1 = e_1$ from now on. The next lemma characterizes the improvement per iteration when one applies either (Oja) or (Kra-Oja) on a new data point.

At the moment, our analysis requires the data to be bounded. Although our framework for proving Theorems 1 and 2 can be extended to cases where the coefficients are sub-Gaussian or even sub-exponential, the coefficients in the recursion of Oja's algorithm are sub-Weibull, even for Gaussian data. This arises due to the presence of terms of order B^4 , which correspond to $\|X_t\|^4$.

Proposition 4. *Let $\rho := \lambda_1 - \lambda_2$ be the eigengap. Suppose \hat{v}_n be the output from either (Oja) or (Kra-Oja). Put*

$$L_t^{Oja} := \sin^2\left(\frac{\hat{v}_t}{\|\hat{v}_t\|}, v_1\right).$$

Then,

$$L_t^{Oja} \leq \begin{cases} (1 - 2\rho\eta_t) L_{t-1}^{Oja} + 2\rho\eta_t (L_{t-1}^{Oja})^2 + Q_t^{Oja} + 4B^4\eta_t^2 & \text{for (Kra-Oja),} \\ (1 - 2\rho\eta_t) L_{t-1}^{Oja} + 2\rho\eta_t (L_{t-1}^{Oja})^2 + Q_t^{Oja} + (5B^4 + 2\eta_t B^6)\eta_t^2 & \text{for (Oja).} \end{cases} \quad (33)$$

for all $t \geq 1$, where Q_t satisfies

$$\mathbb{E}\left(Q_t^{Oja} \middle| \mathcal{F}_{t-1}\right) = 0 \quad \text{and} \quad |Q_t^{Oja}| \leq 8B^2\eta_t \sqrt{L_{t-1}}.$$

The recursions for both (Kra-Oja) and (Oja) share the same form, differing only in the term involving B in η_t^2 . However, a key distinction arises between (33) and (22): the Oja algorithm includes an additional term, $\eta_t S_{t-1}^2$, which is absent in standard SGD settings. Heuristically, S_{t-1} can be large in the initial steps, leading to minimal improvement in the next iterations. This is the reason why most analyses of Oja's algorithm are divided into two phases: an initial unstable phase, where the algorithm's dynamics are erratic, followed by a stable phase, where the dynamics decay at the optimal $O(t^{-1})$ rate. Since the analyses of (Oja) and (Kra-Oja) are the same, we will only focus on (Kra-Oja) in what follows. Our main result in this section is the following theorem.

Theorem 3. *Let $\rho := \lambda_1 - \lambda_2$ be the eigengap. Recall the settings in Proposition 4. Suppose $\mathbb{P}(L_0 \leq 1/4) \geq 1 - \delta^3$ and define*

$$L = \max \left\{ \left[128B^4 \log(\delta^{-1})^2 / \rho^2 \right]; 32 \right\}.$$

Choose step sizes as $\eta_t = \frac{2}{\rho(t+L)}$. Then, with probability at least $1 - 2(e+1)\delta$,

$$L_t^{Oja} \leq \max \left\{ \frac{252L}{\log(\delta^{-1})}; \frac{1008B^4}{\rho^2} \right\} \cdot \frac{\log(\delta^{-1}) + 2 \log \log(t+9)}{t+L}$$

uniformly over all $t \geq 0$.

The constant L in the step sizes η_t control the magnitude of each update uniformly and guarantees that, with high probability, all iterates do not leave a neighborhood of radius $1/2$ (with respect to the sine squared distance) around the true eigenvector.

It should be noted that Theorem 3 does not follow directly from Theorem 2. Instead, we derive Theorem 3 by analyzing an auxiliary process that coincides with L_t on certain “good events” and is set to zero otherwise. This construction necessitates the assumption $\mathbb{P}(L_0 \leq 1/4) \geq 1 - \delta^3$, which introduces a cubic dependence on δ , in contrast to the linear dependence required in Theorem 1. The cubic dependence on δ can be relaxed to a $(1+\varepsilon)$ -dependence for any $\varepsilon > 0$, at the cost of introducing an additional constant factor $C_\varepsilon > 0$ in the bound, where C_ε depends only on ε . For the sake of clarity and simplicity, we state the time-uniform bound with a cubic dependence on δ .

Let us compare the any-time-valid bound of Theorem 3 with the fixed-time bound of Theorem 2.3 in [23]. Below, the notation $\tilde{\Theta}(a_n)$ means asymptotically equivalent up to logarithm factors of a_n . Their results state that for

$$T_0 = \tilde{\Theta}\left(\frac{B^4}{\delta^2\rho^2}\right); \beta = \tilde{\Theta}\left(\frac{B^4}{\delta^2\rho^2}\right); \eta_t = \begin{cases} \tilde{\Theta}\left(\frac{1}{\rho T_0}\right) & \text{for } t \leq T_0; \\ \tilde{\Theta}\left(\frac{1}{\rho(t+\beta-T_0)}\right) & \text{for } t \geq T_0 + 1, \end{cases}$$

we have

$$\sin^2\left(\frac{\hat{\mathbf{v}}_t}{\|\hat{\mathbf{v}}_t\|}, \mathbf{v}_1\right) \lesssim \sqrt{\frac{\beta+1}{\beta+T}}$$

for all fixed $T > T_0$, with probability at least $1 - \delta$.

Both results share the same dependence on B^4/ρ^2 , up to an additional factor of order $\text{polylog}(\delta^{-1})$. The extra $\log \log t$ term is expected and represents the mild cost of achieving a time-uniform bound. The step-size schedule adopted here is also similar to that used in the stable phase of [23] in the stable phase $T > T_0$, which exhibits a factor of order ρ^{-1} .

To get a time-uniform bound from Theorem 3 with a “cold” start, it is crucial to determine the first time T_0 at which L_{T_0} falls below $1/4$, as required by Theorem 3. Determining T_0 depends on the specific structure of the problem, and sharp estimates generally cannot be obtained from an almost-supermartingale of the form (10). This is not surprising, as the form of (10) reduces the analysis of the dynamics of a multi-dimensional system to that of a one-dimensional process. As a result, information about the initial phase of the dynamics is lost. To derive T_0 , we follow the scheme proposed in [23]. Recall that \hat{v}_t is the output of the Oja’s algorithm according to either (Oja) or (Kra-Oja). The starting time T_0 has to be chosen in such a way that

$$\mathbb{P}\left(\sin^2(\hat{v}_{T_0}, v_1) \leq \frac{1}{4}\right) \geq 1 - \delta^3.$$

According to [23], we get

$$T_0 = O\left(\frac{B^4}{\rho^2} \cdot \log\left(\frac{B^2}{\rho\delta}\right)\right) + \underbrace{\tilde{\Theta}\left(\frac{B^4}{\delta^6\rho^2}\right)}_{H_0}$$

where the corresponding sequence of step sizes η_t can be determined as

$$\eta_t = \begin{cases} \tilde{\Theta}\left(\frac{1}{\rho H_0}\right), & \text{if } t \leq H_0, \\ \Theta\left(\frac{1}{\rho(\beta+t-H_0)}\right), & \text{if } t \geq H_0. \end{cases}$$

If a uniformly distributed initialization \hat{v}_0 is used, then the bound in Theorem 3 should be shifted by a factor of T_0 , that is, by replacing t with $t + T_0$. Note that T_0 is the sum of two terms. The second term, denoted H_0 , corresponds to the initial chaotic phase during which the algorithm explores a sufficiently small region likely to contain the true minimum. In this phase, [23] employs a constant step-size scheme. The first term in the expression for T_0 corresponds to the stable phase, where the error decays at the optimal rate and the step sizes are chosen to be of order $1/t$.

3.5 Robbins-Monroe scheme and a lower bound

In this section, we show that the rates obtained in Theorem 1 is sharp by constructing an explicit example that one can not improve the rate for any choice of step sizes that asymptotically of order $\Omega(1/t)$. The goal of this section is solely to demonstrate the sharpness of the rates, and we therefore stick to a one-dimensional example. We will establish a lower bound in the context of the Robbins–Monro scheme, a stochastic gradient-type algorithm for estimating the root of an equation based on observing unbiased realizations of the underlying function. It was introduced in [48] and has been studied extensively in the last few decades. Basic properties and convergence analyses of the Robbins–Monro scheme can be found in [6]; see also [30, 29] and the references therein.

Consider the classical Robbin-Monroe scheme in dimension one, that is to find the root θ of the function $M(x)$ through an unbiased estimator $Y(x)$ of $M(x)$. The Robbins-Monroe scheme works by initializing a value X_0 and update recursively by the rule

$$X_{k+1} := X_k - \eta_{k+1} Y(X_k).$$

We make the following assumptions on Y , M and $\{X_n; n \geq 1\}$:

- $Y(X_n)$ is distributed as $Y(x)$ conditionally on

$$X_n = x, X_{n-1}, Y(X_{n-1}), X_{n-2}, \dots, X_1, Y(X_1).$$

In other words,

$$Y(X_n) \Big| X_n = x, X_{n-1}, Y(X_{n-1}), \dots, Y(X_0), X_0 \stackrel{d}{=} Y(x). \quad (34)$$

- $\mathbb{E}Y(x) = M(x)$ and $\text{Var}(Y(x)) = \sigma^2(x)$ and $\sup_x \sigma^2(x) < \infty$.
- $M(\theta) = 0$ and $M'(x) \geq R > 0$ for almost everywhere $x \in \mathbb{R}$.
- $M(x)$ is sub-polynomial, that is

$$|M(x)| \leq P(|x - \theta|) \quad (35)$$

for some polynomial P with positive coefficients.

The sub-polynomial condition on M is more general than the commonly imposed sub-linear condition in the literature. The sub-linear condition corresponds to the special case where $P(x) = Ax + B$ for some positive constants A and B . In general, the Robbins-Monro scheme may not yield consistent solutions under the sub-polynomial condition. However, Theorem 1 guarantees consistency provided that the initialization is sufficiently close to the root θ .

Put $L_t^{RM} := |X_t - \theta|^2$. In what follows, we consider the additive noise structure $Y(x) = M(x) + \xi_x$, for a collection of i.i.d. centered, unit variance random variables ξ_x such that

$$|\xi| \leq R_1 \quad (36)$$

for some constant $R_1 > 0$. We will assume that the distribution of ξ_x 's is continuous to avoid the technical issue that the iterates form a loop with positive probability.

It is easy to check that such a noise structure satisfies (34). With this setting, we have the recursion

Proposition 5. *Under the settings described above, for all $t \geq 1$,*

$$L_t^{RM} \leq (1 - 2R\eta_t) L_{t-1}^{RM} + Q_t^{RM} + 2\eta_t^2 \cdot \left(P^2 \left(\sqrt{L_{t-1}^{RM}} \right) + R_1^2 \right)$$

where $\mathbb{E}(Q_t^{RM} | \mathcal{F}_{t-1}) = 0$, $|Q_t^{RM}| \leq 2\eta_t R_1 \sqrt{S_{t-1}}$ and P is the polynomial in (35).

Proposition 5 asserts that the iterates of the Robbins-Monroe scheme satisfy a recursion of a form compatible with our main Assumption 1. Thus, the results in Theorem 1 also hold in this case. Since our focus is on proving a lower bound, we will not attempt to derive an anytime bound in this setting.

Next, we show that in the simple case where P is linear, the typical extreme magnitude of the iterates under squared-error loss is exactly of order $\log \log t/t$.

Theorem 4. *Suppose $P(x) = Ax + B$ for some positive constants A and B , where $P(x)$ is as in (35). Assume that the sequence of step sizes $\{\eta_t; t \geq 1\}$ satisfy*

$$\frac{L_1}{t} \leq \eta_t \leq \frac{L_2}{t}$$

for some positive constants L_1, L_2 . Then,

$$\mathbb{P}\left(\limsup_{t \rightarrow \infty} \left(\frac{t \cdot L_t^{RM}}{\log \log t} \right) \geq L\right) = 1$$

where

$$L := \frac{\sqrt{L_1}}{4 [1 + L_2 \cdot \log(8) \cdot M'(\theta)]}. \quad (37)$$

Theorem 4 implies that the typical extreme magnitude of $|X_t - \theta|^2$ cannot be asymptotically smaller than order $\log \log t/t$, for any choice of step sizes of order $\Omega(t^{-1})$. In other words, the construction given in the proof of Theorem 1 yields a time-uniform bound of optimal width in a fairly general setting.

4 Outline of the argument

Let us demonstrate our technique by explaining the proof of Proposition 1. Assume $t_0 = 0$ for simplicity. By choosing step sizes of the form $\eta_t = 2/C_1(t + L)$, the recursion can be rewritten as

$$L_t \leq b_t(X_0 + U_t^*),$$

where

$$\begin{aligned} b_t &:= \frac{L(L-1)}{(t+L-1)(t+L)}, \\ U_t^* &:= \frac{U_t}{b_t} + \frac{U_{t-1}}{b_{t-1}} + \cdots + \frac{U_1}{b_1} = \frac{U_t}{b_t} + U_{t-1}^*. \end{aligned}$$

Note that U_t^* is a weighted sum of U_t . To obtain a time-uniform bound on the interval $[0, t_1]$, it suffices to derive a maximal inequality of the form

$$\mathbb{P}\left(\max_{1 \leq k \leq t_1} U_k^* \geq x_\delta\right) \leq \delta.$$

for some $x_\delta > 0$ depending on δ . To do this, we exploit the recursive structure of the problem. Heuristically, we have

$$\begin{array}{ccccccc} L_0 \text{ small} & \xrightarrow[\text{by (11) + (12)}]{} & U_1 \text{ small} & \xrightarrow[\text{by (10)}]{} & L_1 \text{ small} & \xrightarrow[\text{by (11) + (12)}]{} & \dots \end{array}$$

In order to fully exploit this heuristic, we make use of the following simple observation:

Lemma 2. *Let (M_t) be an adapted process with the corresponding filtration $\{\mathcal{F}_t; t \geq 1\}$. Let $\varepsilon > 0$ and let τ be a stopping time such that*

$$\mathbb{P}\left(\tau \geq t+1 \mid \max_{1 \leq k \leq t} M_k \leq \varepsilon\right) = 1$$

for all $1 \leq t \leq T$. Then

$$\mathbb{P}\left(\max_{1 \leq t \leq T} M_t > \varepsilon\right) = \mathbb{P}\left(\max_{1 \leq t \leq T} M_{\tau \wedge t} > \varepsilon\right).$$

The lemma above has been used in the analysis of Oja's algorithm by [11]. To use it, define the stopping time

$$\tau = \inf\{t \geq t_0 : L_t > r_t\}$$

for some sequence $\{r_t\}$ that serves as tuning parameters to obtain sharp thresholds.

Observe that the stopped process $U_{\tau \wedge t}^*$ has manageable conditional mean and magnitude, given by

$$\begin{aligned} U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^* &= \mathbf{1}_{\{\tau \geq t\}} \cdot (U_t^* - U_{t-1}^*) \\ &= \frac{\mathbf{1}_{\{\tau \geq t\}}}{b_t} \cdot U_t. \end{aligned}$$

Thus, using (16) and (17), and noting that $L_t \leq r_t$ on the event $\{\tau \leq t\}$, we get that

$$|U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^*| = \frac{\mathbf{1}_{\{\tau \geq t\}}}{b_t} \cdot |U_t| \leq \frac{2C_3(t+L-1)\sqrt{r_{t-1}}}{C_1(L-1)L} + \frac{4C_2(t+L-1)}{C_1^2(t+L)(L-1)L}$$

and

$$\left|\mathbb{E}\left[U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^* \mid \mathcal{F}_{t-1}\right]\right| \leq \frac{4C_2}{C_1^2(t+L)} \cdot \frac{t+L-1}{(L-1)L}.$$

A standard application of Azuma–Hoeffding’s inequality along with a telescoping argument then yields

$$\mathbb{P}\left(\max_{1 \leq k \leq t_1} U_{k \wedge \tau}^* \geq y(\delta, r_t)\right) \leq \delta$$

for some function $y(\delta, r_t)$ of δ and r_t . We then tune the sequence $\{r_t\}$ in a way such that the lemma above applies to deduce that

$$\mathbb{P}\left(\max_{1 \leq k \leq t_1} U_k^* \geq y(\delta, r_t)\right) = \mathbb{P}\left(\max_{1 \leq k \leq t_1} U_{k \wedge \tau}^* \geq y(\delta, r_t)\right) \leq \delta.$$

The conclusion of Proposition 1 follows immediately.

The proof of Theorem 1 also follows from the same scheme, but requires a different choice of the tuning sequence r_t to accommodate piecewise-constant step sizes. The general form of (11) and (12) makes it quite challenging to construct an explicit step-size schedule. Finally, note that the $\log \log$ term comes from a peeling argument: we apply Proposition 1 repeatedly on small intervals.

One major advantage of the approach above is that it can yield a maximal inequality by fully exploiting the hierarchical structure of the recursion, while requiring only simple concentration inequalities in the analysis. We believe that a similar type of maximal inequality can also be obtained by using the mixture method, as in [20, 21] (see also the references therein). However, it is not clear to us which choices of the prior would yield a comparable bound involving only C_3^2/C_1^2 and C_2/C_1^2 .

5 Conclusion and discussion

In this paper, we develop a new technique based on a variant of the classical Robbins–Siegmund lemma [49] for obtaining time-uniform bounds for a broad class of almost supermartingale-like processes, where the remainder terms at each iteration are bounded by a polynomially growing function of the step sizes and previous iterates. We prove that the rates achieved by our technique are optimal and provide explicit constructions of step size schedules for processes with simple landscapes (i.e., without spurious local minima). As applications, we use the method to derive time-uniform bounds for projected SGD, Oja’s algorithm, and the Robbins–Monro scheme. We conclude by outlining several potential extensions of our framework that present interesting problems for future research:

- **Streaming k -PCA.** In our analysis of Oja’s algorithm, we considered only the estimation of the principal eigenvector. A natural extension is to handle the streaming k -PCA problem by deriving a recursive estimate analogous to Proposition 4. We believe the argument will be similar, but more technically involved. The key step would be to derive an analog of Proposition 4.
- **Relaxed moment conditions.** Our results rely in a fundamental way on the bounds (11) and (12) of our recursive Assumption 1, which deliver the optimal $\log \log t/t$ scaling. A natural weakening of Assumption 1 is to allow for the coefficients A_i ’s and B_i ’s in (11) and (12) to be random. Determining the moment conditions on those coefficients that still guarantee a $\log \log t/t$ rate is both interesting and challenging. Although our results remain valid (possibly with worse constants) when the coefficients are conditionally sub-Gaussian (see Appendix A in the supplement for more details), it remains unclear whether the same holds under sub-exponential, or even the more challenging sub-Weibull, settings.

Notably, in the simplified case of (16) and (17), the same rate (with possibly worse constants) still holds when C_3 is sub-Gaussian and C_2 is sub-exponential—matching the SGD rate under sub-Gaussian error recently setting studied in [38]. We believe that settings involving noise distributions with tails heavier than sub-exponential can be addressed using a combination of truncation and path coupling arguments. This represents an interesting direction for future work.

- **Low-rank constrained least squares.** Our technique can be applied to other problems of similar structure. For example, the low-rank constrained least squares framework of [15] encompasses several important tasks, including matrix completion, phase retrieval, and subspace tracking.
- **Stochastic heavy-ball algorithms.** Another promising direction is to extend our results to stochastic heavy-ball methods; see, for example, [50, 35] and references therein. The main technical challenge is that these algorithms involve an additional momentum term at each step, leading to a two-dimensional process rather than a scalar one.

6 Acknowledgments

We thank Shubhangshu Shekhar for productive discussions during the initial phase of the project and Aaditya Ramdas for helpful comments and for pointing out the reference [56]. TP and PS gratefully acknowledge NSF grants 2217069 and 2019844.

7 Proof of Theorem 1

The following result is a variant of Lemma 6.9 in [11], previously used in the analysis of certain Oja-type algorithms, and plays a crucial role in our proof. It shows that the crossing probability of a process can be controlled by that of a stopped process via a suitably chosen stopping time.

Lemma 3. *Let M_t be an adapted process with the corresponding filtration $\{\mathcal{F}_t; t \geq 1\}$. Let $\varepsilon > 0$ and τ be a stopping time with respect to $\{\mathcal{F}_t; t \geq 1\}$ such that*

$$\mathbb{P}\left(\tau \geq t + 1 \mid \max_{1 \leq k \leq t} M_k \leq \varepsilon\right) = 1 \quad (38)$$

for all $1 \leq t \leq T$. Then,

$$\mathbb{P}\left(\max_{1 \leq t \leq T} M_t > \varepsilon\right) = \mathbb{P}\left(\max_{1 \leq t \leq T} M_{\tau \wedge t} > \varepsilon\right).$$

Proof of Lemma 3. One direction is obvious because

$$\max_{1 \leq t \leq T} M_t \geq \max_{1 \leq t \leq \tau \wedge T} M_t = \max_{1 \leq t \leq T} M_{\tau \wedge t}$$

which implies

$$\mathbb{P}\left(\max_{1 \leq t \leq T} M_t > \varepsilon\right) \geq \mathbb{P}\left(\max_{1 \leq t \leq T} M_{\tau \wedge t} > \varepsilon\right).$$

To prove the other direction, write

$$\begin{aligned} \mathbb{P}\left(\max_{1 \leq t \leq T} M_{\tau \wedge t} > \varepsilon\right) &= \sum_{k=1}^T \mathbb{P}\left(\tau = k, \max_{1 \leq i \leq k} M_i > \varepsilon\right) + \mathbb{P}\left(\tau \geq T + 1, \max_{1 \leq i \leq T} M_i > \varepsilon\right) \\ &= \sum_{k=1}^T \left[\mathbb{P}(\tau = k) - \underbrace{\mathbb{P}\left(\tau = k, \max_{1 \leq i \leq k} M_i \leq \varepsilon\right)}_{=0} \right] + \mathbb{P}\left(\tau \geq T + 1, \max_{1 \leq i \leq T} M_i > \varepsilon\right) \\ &= \sum_{k=1}^T \mathbb{P}(\tau = k) + \mathbb{P}\left(\tau \geq T + 1, \max_{1 \leq i \leq T} M_i > \varepsilon\right) \\ &= \mathbb{P}\left(\max_{1 \leq i \leq T} M_i > \varepsilon\right) - \underbrace{\mathbb{P}\left(\tau \leq T, \max_{1 \leq i \leq T} M_i > \varepsilon\right)}_{\geq 0} + \mathbb{P}(\tau \leq T). \end{aligned}$$

The proof is completed. \square

Lemma 4 (Azuma-Hoeffding's maximal inequality). *Let M_t be a process adapted to the filtration $\{\mathcal{F}_t; t \geq 0\}$ and such that $M_0 = 0$ almost surely. For every t , let μ_t and σ_t^2 satisfy*

$$|M_t - M_{t-1}| \leq c_t \quad (39)$$

$$\left| \mathbb{E}(M_t - M_{t-1} | \mathcal{F}_{t-1}) \right| \leq \mu_t. \quad (40)$$

Then, for all $T \geq 1$ and $\delta \in (0, 1)$,

$$P\left(\exists t \in [1, T] : M_t \geq \sqrt{2}V_T(\delta) + \sum_{i=1}^t \mu_i\right) \leq \delta,$$

where

$$V_T(\delta) := \sqrt{\log\left(\frac{1}{\delta}\right) \sum_{t=1}^T c_t^2} \quad (41)$$

Proof of Lemma 4. Define $d_0 := 0$ and $d_k := M_k - M_{k-1}$. It is easy to check that

$$M_k = \sum_{i=1}^k [d_i - \mathbb{E}(d_i | \mathcal{F}_{i-1})] + \sum_{i=1}^k \mathbb{E}(d_i | \mathcal{F}_{i-1}).$$

Let $h_i := d_i - \mathbb{E}(d_i | \mathcal{F}_{i-1})$. It is easy to check that $\{h_i; 1 \leq i \leq T\}$ forms a martingale difference sequence and

$$\begin{aligned} |\mathbb{E}(d_i | \mathcal{F}_{i-1})| &\leq \mu_i; \\ \mathbb{E}(e^{\lambda h_i} | \mathcal{F}_{i-1}) &\leq e^{\lambda^2 c_i^2 / 2} \end{aligned}$$

almost surely, for all $\lambda > 0$. Note that the second estimate in the display above follows from Chernoff's bound and the fact that $|h_i| \leq 2c_i$, which is due to (39). Thus, we have

$$\begin{aligned} &\mathbb{P}\left(\exists t \in [1, T] : M_t \geq \sqrt{2}V_T(\delta) + \sum_{i=1}^t \mu_i\right) \\ &\leq \mathbb{P}\left(\exists t \in [1, T] : \sum_{k=1}^t h_k \geq \sqrt{2}V_T(\delta)\right) \\ &= \mathbb{P}\left(\exists t \in [1, T] : \sum_{k=1}^t h_k \geq \sqrt{2 \log\left(\frac{1}{\delta}\right) \cdot \sum_{k=1}^T c_k^2}\right) \\ &\leq \exp\left(-\frac{2 \log\left(\frac{1}{\delta}\right) \cdot \sum_{k=1}^T c_k^2}{2 \sum_{k=1}^T c_k^2}\right) = \delta \end{aligned}$$

where the first inequality follows from (40), the second inequality follows from Azuma-Hoeffding's inequality. The proof is completed. \square

7.1 Step 1: Maximal inequality

The following result is our main technical result. It gives a sharp, short-time maximal inequality for a stochastic process that satisfies (10) with constant step sizes in a finite interval.

Lemma 5. Suppose X_0, X_1, \dots, X_n is a sequence of random variables that satisfy **ALE: Do you mean Assumption 1?** (10) with constant step sizes $\eta_n = \eta$. Let $r \in (0, 1)$ and $\delta \in (0, 1)$, and put

$$D_\delta(r) := \frac{8D}{(1 - C_1\eta)^n} \cdot \left(\sqrt{\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\eta r} + \sum_{i=1}^m \eta^{c_i} r^{d_i} \right) + \eta + \sum_{i=1}^m \eta^{a_i} r^{b_i} \right), \quad (42)$$

where

$$D := \max_{1 \leq i \leq m} \left\{ \frac{A_i}{C_1} \right\} \vee \max_{1 \leq i \leq m} \left\{ \frac{\sqrt{m+1} \cdot B_i}{\sqrt{C_1}} \right\} \vee \frac{C_2}{C_1} \vee \frac{C_3 \sqrt{m+1}}{\sqrt{C_1}}. \quad (43)$$

Then, for any $\varepsilon_0 > 0$, we have that

$$\mathbb{P}(\exists t \in [1, n] : X_t \geq (1 - C_1\eta)^t (\varepsilon_0 + D_\delta), A) \leq \delta$$

whenever $\varepsilon_0 + D_\delta(r) \leq r$ and $A \subset \{X_0 \leq \varepsilon_0\}$ is such that $\mathbb{P}(A) > 0$.

Proof of Lemma 5. For each $t \geq 1$, define

$$U_t^* := (1 - C_1\eta)^{-t} \left[U_t + (1 - C_1\eta) \cdot U_{t-1} + \cdots + (1 - C_1\eta)^t U_0 \right]$$

and set $U_0^* = 0$. Note that with this notation, we have

$$U_t^* = (1 - C_1\eta)^{-t} U_t + U_{t-1}^*; \quad (44)$$

$$X_t \leq (1 - C_1\eta)^t (X_0 + U_t^*). \quad (45)$$

Our strategy is to establish via Lemma 3 the maximal inequality

$$\mathbb{P}\left(\max_{1 \leq t \leq n} U_t^* \geq D_\delta | A\right) \leq \frac{\delta}{\mathbb{P}(A)}. \quad (46)$$

Towards that end, we will first show that

$$\mathbb{P}\left(\max_{1 \leq t \leq n} U_{\tau \wedge t}^* \geq D_\delta | A\right) \leq \frac{\delta}{\mathbb{P}(A)}, \quad (47)$$

where τ is the stopping time

$$\tau = \inf \{t \geq 0 : X_t \geq r\}.$$

Note that, for any arbitrary stopping time τ , we have the relation

$$U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^* = \mathbf{1}_{\{\tau \geq t\}} (U_t^* - U_{t-1}^*).$$

Thus,

$$\begin{aligned} |U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^*| &= |\mathbf{1}_{\{\tau \geq t\}} (U_t^* - U_{t-1}^*)| \leq c_t; \\ \left| \mathbb{E}\left(U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^* \mid \mathcal{F}_{t-1}\right) \right| &= \underbrace{\left| \mathbf{1}_{\{\tau \geq t\}} \mathbb{E}\left(U_t^* - U_{t-1}^* \mid \mathcal{F}_{t-1}\right) \right|}_{\text{by using (44)}} \\ &= (1 - C_1\eta)^{-t} |\mathbb{E}(U_t^* | \mathcal{F}_t)| \leq \mu_t, \end{aligned}$$

where

$$\begin{aligned} \mu_t &:= (1 - C_1\eta)^{-t} \cdot \left(C_2 \eta^2 + \sum_{k=1}^m A_k \cdot \eta^{1+a_k} r^{b_k} \right); \\ c_t &:= (1 - C_1\eta)^{-t} \cdot \left(C_3 \eta \sqrt{r} + \sum_{k=1}^m B_k \eta^{1/2+c_k} r^{d_k} \right). \end{aligned}$$

By Lemma 4, we obtain that

$$\mathbb{P}\left(\exists t \in [1, n] : U_{\tau \wedge t}^* \geq 8 \sqrt{\log(\delta^{-1}) \cdot \sum_{k=1}^n c_k^2 + \sum_{k=1}^t \mu_k}\right) \leq \delta.$$

Next, one can easily check that

$$\begin{aligned} \sum_{k=1}^t \mu_k &= \left(C_2 \cdot \eta^2 + \sum_{k=1}^m A_k \cdot \eta^{1+a_k} r^{b_k}\right) \sum_{k=1}^t (1 - C_1 \eta)^{-k} \\ &= \left(C_2 \cdot \eta^2 + \sum_{k=1}^m A_k \cdot \eta^{1+a_k} r^{b_k}\right) (1 - C_1 \eta)^{-1} \frac{1}{1/(1 - C_1 \eta) - 1} \left[\frac{1}{(1 - C_1 \eta)^t} - 1\right] \\ &\leq \left(C_2 \eta^2 + \sum_{k=1}^m A_k \cdot \eta^{1+a_k} r^{b_k}\right) \frac{(1 - C_1 \eta)^{-t}}{C_1 \eta} \\ &\leq D \left(\eta + \sum_{k=1}^m \eta^{a_k} r^{b_k}\right) (1 - C_1 \eta)^{-n} \end{aligned}$$

where D is defined in (43). Similarly,

$$\begin{aligned} \sum_{k=1}^n c_k^2 &\leq (m+1) \cdot \left(\eta^2 r + \sum_{k=1}^m \eta^{2c_i+1} r^{2d_i}\right) \cdot \frac{C_3^2 \vee (\max_{1 \leq k \leq m} B_k^2)}{2C_1 \eta - C_1^2 \eta^2} \cdot (1 - C_1 \eta)^{-2n} \\ &\leq (m+1) \cdot \underbrace{\frac{C_3^2 \vee (\max_{1 \leq k \leq m} B_k^2)}{C_1}}_{\leq D^2} \cdot \left(\eta r + \sum_{k=1}^m \eta^{2c_i} r^{2d_i}\right) \cdot (1 - C_1 \eta)^{-2n} \\ &\leq D^2 \left(\eta r + \sum_{k=1}^m \eta^{2c_i} r^{2d_i}\right) \cdot (1 - C_1 \eta)^{-2n}. \end{aligned}$$

Consequently,

$$\mathbb{P}_A(\exists t \in [1, n] : U_{\tau \wedge t}^* \geq D_\delta) \leq \frac{\delta}{\mathbb{P}(A)}$$

where \mathbb{P}_A denotes the conditional probability on A and

$$D_\delta := \frac{8D}{(1 - C_1 \eta)^n} \cdot \left(\sqrt{\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\eta r} + \sum_{i=1}^m \eta^{c_i} r^{d_i}\right) + \eta + \sum_{i=1}^m \eta^{a_i} r^{b_i}\right).$$

Note that the inequality above implies (47) since

$$(1 - C_1 \eta)^{-t} \leq (1 - C_1 \eta)^{-n}.$$

To deduce (46) from (47), we need to check that

$$\mathbb{P}_A\left(\tau \geq k+1 \mid \max_{1 \leq t \leq k} U_t^* \leq D_\delta\right) = 1.$$

Note that Lemma 3 is being applied to the conditional measure \mathbb{P}_A instead of \mathbb{P} . The last display is true since conditionally on $\{\max_{1 \leq t \leq k} U_t^* \leq D_\delta\} \cap A$, (45) gives

$$X_k \leq (1 - C_1 \eta)^k (X_0 + U_k^*) \leq (1 - C_1 \eta)^k (\varepsilon_0 + D_\delta) \leq r$$

since $\varepsilon_0 + D_\delta \leq r$ and $A \subset \{X_0 \leq \varepsilon_0\}$. Consequently,

$$\mathbb{P}\left(\max_{1 \leq t \leq n} U_t^* \geq D_\delta \mid A\right) \leq \frac{\delta}{\mathbb{P}(A)}.$$

The proof is completed. \square

7.2 Step 2: Induction argument

Let $h_i = h_0 \cdot 2^{-i}$ for some constant h_0 to be determined later. The value h_0 correspond to r in the statement of Theorem 1. We want to derive a sequence of epochs $\{t_i; i \geq 0\}$ such that with probability at least $1 - 2\delta$, $L_t \leq h_i$ for all $i \geq 0$ and $t \in [t_{i-1} + 1, t_i]$.

To visualize the argument below, it is helpful to refer to Figure 1, where we plot the width of the confidence sequence over time, assuming $h_0 = 1$. The error dynamic of our analysis can be summarized as follows

$$\underbrace{t_0 = 0}_{h_0} \rightarrow \underbrace{t_0 + 1}_{\leq(1-C_1\eta_0)2h_0} \rightarrow \cdots \rightarrow \underbrace{t_1 - 1}_{\leq(1-C_1\eta_0)^{t_1-t_0-1}2h_0} \rightarrow \underbrace{t_1}_{\leq 2h_0/4=h_1} \rightarrow \cdots$$

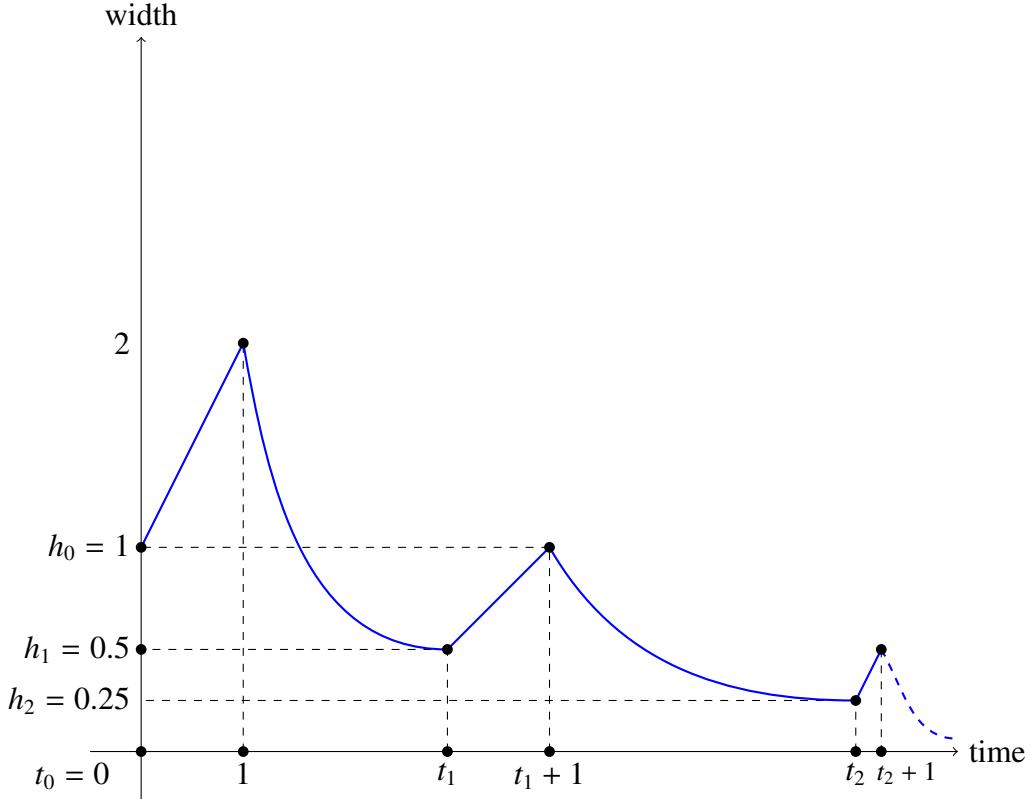


Figure 1: Errors induced by adaptive step sizes

The width of the confidence sequence will be a function f such that $a_i \leq f(t_i)$. Put

$$t_0 := 0, \\ \delta_i := \frac{\delta}{(i+10)^2}.$$

For a small constant $\kappa \leq 1$ to be chosen later, let us define

$$\eta_i := \frac{\kappa h_{i-1}}{\log(\delta_i^{-1})} \quad (48)$$

and choose the epochs t_i such that

$$\frac{1}{16} \leq (1 - C_1 \eta_i)^{t_i - t_{i-1}} \leq \frac{1}{8}. \quad (49)$$

The step sizes are chosen to be the constant η_i on the interval $[t_{i-1} + 1, t_i]$. In what follows, we will determine (κ, h_0) by applying Lemma 5 repeatedly. We will also require

$$\sup_{i \geq 1} \eta_i \leq \frac{1}{16C_1}. \quad (50)$$

Step 2a: Maximal inequality. Let us show that for all $i \geq 1$,

$$\mathbb{P}\left(\exists t \in [t_{i-1} + 1, t_i] : L_t \geq 4(1 - C_1\eta_i)^{t-t_{i-1}} h_i \mid L_{t_{i-1}} \leq h_{i-1}\right) \leq \frac{\delta_i}{\mathbb{P}(L_{t_{i-1}} \leq h_{i-1})} \quad (51)$$

for some choices of κ in (48) and a_0 that depend only on the parameters in (11) and (12).

To choose κ, a_0 such that (51) holds, we apply Lemma 5 to $\{L_t\}$ to get

$$\mathbb{P}\left(\exists t \in [t_{i-1} + 1, t_i] : L_t \geq (1 - C_1\eta_i)^t (h_{i-1} + D(r_i)) \mid L_{t_{i-1}} \leq h_{i-1}\right) \leq \frac{\delta_i}{\mathbb{P}(L_{t_{i-1}} \leq h_{i-1})}$$

whenever

$$h_{i-1} + D(r_i) \leq r_i. \quad (52)$$

Here $D(r_i)$ is defined to be

$$D(r_i) := \frac{8D}{(1 - C_1\eta_i)^{t_i - t_{i-1}}} \cdot \left[\sqrt{\log\left(\frac{1}{\delta_i}\right)} \cdot \left(\sqrt{\eta_i r_i} + \sum_{k=1}^m \eta_i^{c_k} r_i^{d_k} \right) + \eta_i + \sum_{k=1}^m \eta_i^{a_k} r_i^{b_k} \right]$$

where D is the constant in (43).

One does not have to worry about the existence of r_i in (52). Let us fix the choice $r_i = 2h_{i-1}$ and choose κ in (48) and a_0 such that $h_{i-1} + D(r_i) \leq r_i$ for all $i \geq 1$. Observe that

$$\begin{aligned} \sqrt{\log(\delta_i^{-1})} \sqrt{\eta_i r_i} &= \sqrt{2\kappa} h_{i-1}; \\ \sqrt{\log(\delta_i^{-1})} \eta_i^{c_k} r_i^{d_k} &= \sqrt{\log(\delta_i^{-1})} \underbrace{\left(\frac{\kappa}{2 \log(\delta_i^{-1})} \right)^{c_k}}_{\leq 1} 2^{c_k+d_k} \cdot h_{i-1}^{c_k+d_k-1} \cdot h_{i-1} \\ &\leq \left[2^{c_k+d_k} \cdot \frac{\sqrt{\log(\delta_i^{-1})}}{2^{(i-1)(c_k+d_k-1)}} \cdot h_0^{c_k+d_k-1} \right] h_{i-1}; \\ \eta_i^{a_k} r_i^{b_k} &= \underbrace{\left(\frac{\kappa}{2 \log(\delta_i^{-1})} \right)^{a_k}}_{\leq 1} \cdot 2^{a_k+b_k} \cdot h_{i-1}^{a_k+b_k-1} \cdot h_{i-1} \\ &\leq \left[2^{a_k+b_k} \cdot h_0^{a_k+b_k-1} \right] h_{i-1}. \end{aligned}$$

Note that $(1 - C_1\eta_i)^{t_i - t_{i-1}} \geq 1/16$ due to (49). Thus, to get (52), we need to choose K and a_0 such that

$$128D \left[\sqrt{2\kappa} + \sum_{k=1}^m 2^{c_k+d_k} \cdot \frac{\sqrt{\log(\delta_i^{-1})}}{2^{(i-1)(c_k+d_k-1)}} \cdot h_0^{c_k+d_k-1} + \frac{\kappa}{\log(\delta_i^{-1})} + \sum_{k=1}^m 2^{a_k+b_k} \cdot h_0^{a_k+b_k-1} \right] \leq 1.$$

It suffices to pick κ and a_0 such that

$$\begin{aligned} \sqrt{2\kappa} &\leq \frac{1}{(2m+2)128D}; \\ h_0^{c_k+d_k-1} &\leq \frac{1}{(2m+2)128D} \cdot 2^{-c_k-d_k} \cdot \min_{i \geq 1} \left\{ \frac{2^{(i-1)(c_k+d_k-1)}}{\sqrt{\log(\delta_i^{-1})}} \right\}, \forall k \in [1, m]; \\ \kappa &\leq \frac{1}{(2m+2)128D}; \\ h_0^{a_k+b_k-1} &\leq \frac{2^{-a_k-b_k}}{(2m+2)128D}, \forall k \in [1, m]. \end{aligned}$$

From the first and third inequalities, it is clear that one can choose κ as a function of m, D :

$$\kappa := \min \left\{ 1; \frac{1}{2} \left[\frac{1}{(2m+2)128D} \right]^2; \frac{1}{(2m+2)128D} \right\}.$$

Moreover, one can also choose h_0 such that the second and last inequalities also hold because

$$\min_{i \geq 1} \left\{ \frac{2^{(i-1)(c_k+d_k-1)}}{\sqrt{\log(\delta_i^{-1})}} \right\} > 0 \text{ and } (c_k + d_k - 1) \wedge (a_k + b_k - 1) > 0$$

for all $1 \leq k \leq m$. Such a specific choice is

$$h_0 := \min_{1 \leq k \leq m} \left\{ \exp \left(\frac{A_k(\delta)}{c_k + d_k - 1} \right) \wedge \exp \left(\frac{B_k}{a_k + b_k - 1} \right) \wedge \frac{\log(\delta^{-1})}{16C_1\kappa} \right\}$$

where the last term is due to the constraint on step sizes (50) and

$$\begin{aligned} A_k(\delta) &:= \log \left(\frac{1}{(2m+2)128D} \cdot 2^{-c_k-d_k} \cdot \min_{i \geq 1} \left\{ \frac{2^{(i-1)(c_k+d_k-1)}}{\sqrt{\log(\delta_i^{-1})}} \right\} \right); \\ B_k &:= \log \left(\frac{2^{-a_k-b_k}}{(2m+2)128D} \right). \end{aligned}$$

Consequently, we have (51) since $h_{i-1} + D(r_i) \leq r_i = 4h_i$. We now set $r \equiv h_0$, where r is the constant in the conclusion of Theorem 1. From now on, we assume that

$$\mathbb{P}(L_0 \leq h_0) \geq 1 - \delta.$$

Step 2b: Valid coverage of the confidence sequence. Define

$$r(i, t) := \begin{cases} 4(1 - C_1\eta_i)^{t-t_{i-1}} h_i, & \text{for } t \in [t_{i-1} + 1, t_i], \\ h_i, & \text{for } t = t_i. \end{cases}$$

Note that $r(i, t_i) = h_i$ for all $i \geq 0$ by the definition above. Put

$$r_t^* := \begin{cases} r(i, t) & \text{for } t \in [t_{i-1} + 1, t_i], \\ h_0 & \text{for } t = 0. \end{cases} \quad (53)$$

By using the union bound and (51), we have

$$\begin{aligned}\mathbb{P}(\forall t \geq 0 : L_t \leq r_t^*) &\geq 1 - \mathbb{P}(L \leq a_0) - \sum_{i=1}^{\infty} \mathbb{P}\left(\{\exists t \in [t_{i-1} + 1, t_i] : L_t \geq r_t^*\} \cap \{L_{t_{i-1}} \leq r_{t_{i-1}}^*\}\right) \\ &\geq 1 - \delta - \sum_{i=1}^{\infty} \delta_i \geq 1 - 2\delta.\end{aligned}$$

Step 2c: Width of the confidence sequence. We will show that there exists a constant M not depending on t such that

$$r_t^* \leq M \cdot \frac{\log(\delta^{-1}) + \log \log(t+10)}{t+10} \quad (54)$$

for all $t \geq 0$, where r^* is defined as in (53).

Let us first show that

$$\sup_{k \geq 0} \left\{ h_k \cdot \frac{t_k + 10}{\log(\delta^{-1}) + \log \log(t_k + 10)} \right\} < \infty.$$

From (49), we have

$$\frac{\log(16)}{\log\left(\frac{1}{1-C_1\eta_i}\right)} \geq t_i - t_{i-1} \geq \frac{\log(8)}{\log\left(\frac{1}{1-C_1\eta_i}\right)}.$$

Note that there exists a universal constant $c_0 > 0$ small enough such that

$$\frac{c_0}{x} \leq \frac{1}{\log(1+x)} \leq \frac{1}{c_0 x}$$

for all $x \in (0, 1/15)$.

Write

$$t_k = \sum_{i=1}^k t_i - t_{i-1} \leq \sum_{i=1}^k \frac{\log(16)}{\log\left(\frac{1}{1-C_1\eta_i}\right)} \quad (55)$$

$$\begin{aligned}&= \sum_{i=1}^k \frac{\log(16)}{\log\left(1 + \frac{C_1\eta_i}{1-C_1\eta_i}\right)} \\ &\leq \log(16) \cdot \sum_{i=1}^k \frac{1}{c_0} \cdot \frac{1 - C_1\eta_i}{C_1\eta_i} \\ &\leq \frac{\log(16)}{c_0 C_1 \kappa} \cdot \sum_{i=1}^k h_{i-1}^{-1} (\log(\delta^{-1}) + 2 \log(i+1)) \\ &= \frac{\log(16)h_0}{2c_0 C_1 \kappa} \cdot \sum_{i=1}^k 2^i (\log(\delta^{-1}) + 2 \log(i+1)) \\ &\leq \frac{\log(16)h_0}{c_0 C_1 \kappa} \cdot 2^k (\log(\delta^{-1}) + 2 \log(k+1)).\end{aligned} \quad (56)$$

Similarly,

$$\begin{aligned}t_k &\geq \log(16)c_0 \cdot \sum_{i=1}^k \frac{1 - C_1\eta_i}{C_1\eta_i} \\ &\geq \frac{15 \log(16)c_0}{16} \cdot \sum_{i=1}^k h_{i-1}^{-1} (\log(\delta^{-1}) + 2 \log(i+1)) \\ &\geq \frac{15 \log(16)c_0 h_0}{32} \cdot 2^k (\log(\delta^{-1}) + 2 \log(k+1)).\end{aligned} \quad (57)$$

Now, observe that

$$\begin{aligned} h_k \cdot \frac{t_k + 10}{\log(\delta^{-1}) + \log \log(t_k + 10)} &\leq \frac{2^{-k} \cdot h_0 \left[\frac{\log(16)h_0}{c_0 C_1 \kappa} \cdot 2^k (\log(\delta^{-1}) + 2 \log(k+1)) \right]}{\log(\delta^{-1}) + \log \log(t_k + 10)} \\ &\leq \frac{\log(16)h_0^2}{c_0 C_1 \kappa} \cdot \frac{\log(\delta^{-1}) + 2 \log(k+1)}{\log(\delta^{-1}) + \log \log(t_k + 10)}. \end{aligned}$$

If $k+1 \geq \sqrt{\delta} \exp\left(\frac{2}{\log(16)c_0h_0}\right)$, we have

$$\begin{aligned} &\log(\delta^{-1}) + \log \log(t_k + 10) \\ &> \log(\delta^{-1}) + \log \log(t_k) \\ &\geq \log(\delta^{-1}) + \log \left[(k+1) \log 2 + \underbrace{\log\left(\frac{15 \log(16)c_0h_0}{64} (\log(\delta^{-1}) + 2 \log(k+1))\right)}_{\geq 0} \right] \\ &\geq \log(\delta^{-1}) + \log(k+1) + \log \log 2. \end{aligned}$$

Therefore,

$$\begin{aligned} h_k \cdot \frac{t_k + 10}{\log(\delta^{-1}) + \log \log(t_k + 10)} &\leq \frac{\log(16)h_0^2}{c_0 C_1 \kappa} \cdot \frac{\log(\delta^{-1}) + 2 \log(k+1)}{\log(\delta^{-1}) + \log(k+1) + \log \log 2} \\ &\leq \frac{\log(16)h_0^2}{2c_0 C_1 \kappa} \end{aligned}$$

for all $k \geq \sqrt{\delta} \exp\left(\frac{2}{\log(16)c_0h_0}\right) - 1$.

Consequently,

$$\sup_{k \geq 0} \left\{ h_k \cdot \frac{t_k + 10}{\log(\delta^{-1}) + \log \log(t_k + 10)} \right\} = H < \infty$$

for some $H > 0$ that depends only on h_0 and the parameters in (11) and (12).

Now, suppose $t \in [t_{k-1} + 1, t_k]$ for some $k \geq 1$. Recall r_t^* in (53), we have

$$\begin{aligned} r_t^* &= 4(1 - C_1 \eta_k)^{t-t_{k-1}} h_i \\ &\leq 4h_i \\ &\leq 4H \cdot \frac{\log(\delta^{-1}) + \log \log(t_k + 10)}{t_k + 10} \\ &\leq 4H \cdot \frac{\log(\delta^{-1}) + \log \log(t+10)}{t+10} \end{aligned}$$

by using Lemma 9. Thus, the constant M in (54) can be chosen to be $4H$.

Final step: Showing $\eta_t \asymp 1/t$. Finally, we prove the statement regarding the asymptotic scaling of η_t . For $t \in [t_{k-1} + 1, t_k]$, write

$$\begin{aligned} \eta_t &= \frac{2\kappa h_0}{2^k [\log(\delta^{-1}) + 2 \log(k+1)]} \\ &\leq \frac{2\kappa h_0}{\log(16)h_0} \cdot c_0 C \kappa \cdot \frac{1}{t_k} \\ &\leq \frac{2c_0 C \kappa^2}{\log(16)} \cdot \frac{1}{t}, \end{aligned}$$

where we use (56) in the second line. For the lower bound, suppose $k \geq 3$. Then, by using (57),

$$\begin{aligned}\eta_t &= \frac{2kh_0}{2^k [\log(\delta^{-1}) + 2\log(k+1)]} \\ &= \frac{\kappa h_0}{2^{k-1} [\log(\delta^{-1}) + 2\log k]} \cdot \underbrace{\frac{\log(\delta^{-1}) + 2\log k}{\log(\delta^{-1}) + 2\log(k+1)}}_{\geq 1/4} \\ &\geq \frac{32\kappa h_0}{15\log(16)c_0 h_0} \cdot \frac{1}{t_{k-1}} \cdot \frac{1}{4} \\ &\geq \frac{8\kappa}{15\log(16)c_0} \cdot \frac{1}{t}.\end{aligned}$$

The proof is completed. \square

8 Proof of Proposition 1

Rewrite (10) as

$$L_t \leq b_t (X_{t_0} + U_t^*)$$

where

$$\begin{aligned}b_t &:= \frac{(t_0 + L - 1)(t_0 + L)}{(t_1 + L - 1)(t_1 + L)} \\ U_t^* &:= \frac{U_t}{b_t} + \frac{U_{t-1}}{b_{t-1}} + \cdots + \frac{U_1}{b_1} = \frac{U_t}{b_t} + U_{t-1}^*.\end{aligned}$$

Let us give a tail bound on $\max_{t_0+1 \leq t \leq t_1} U_k^*$. Choose a time-dependent stopping rule τ :

$$\tau = \inf \{t \geq t_0 : X_t > r_t\}$$

for some sequence $\{r_t; t_0 \leq t \leq t_1\}$ to be chosen later.

Consider the stopped process

$$\max_{t_0+1 \leq t \leq t_1} U_{\tau \wedge t}^*.$$

Easily,

$$\begin{aligned}|U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^*| &= \frac{\mathbf{1}_{\{\tau \geq t\}}}{b_t} \cdot |U_t| \\ &\leq \frac{(t+L-1)(t+L)}{(t_0+L-1)(t_0+L)} \cdot \left(\frac{2C_3 \sqrt{r_{t-1}}}{C_1(t+L)} + \frac{4C_2}{C_1^2(t+L)^2} \right) \\ &= \underbrace{\frac{2C_3(t+L-1) \sqrt{r_{t-1}}}{C_1(t_0+L-1)(t_0+L)}}_{c_t} + \underbrace{\frac{4C_2(t+L-1)}{C_1^2(t+L)(t_0+L-1)(t_0+L)}}_{c_t}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^* | \mathcal{F}_{t-1}] &\leq \frac{4C_2}{C_1^2(t+L)^2} \cdot \frac{(t+L-1)(t+L)}{(t_0+L-1)(t_0+L)} \\ &= \underbrace{\frac{4C_2(t+L-1)}{C_1^2(t+L)(t_0+L-1)(t_0+L)}}_{\mu_t}.\end{aligned}$$

By using Azuma-Hoeffding maximal inequality, we have

$$\mathbb{P}\left(\max_{t_0+1 \leq t \leq t_1} U_{\tau \wedge t}^* \geq D(\delta) \mid A\right) \leq \frac{\delta}{\mathbb{P}(A)}$$

where

$$D(\delta) := \frac{4C_2(t_1 - t_0)}{C_1^2(t_0 + L - 1)(t_0 + L)} + \sqrt{4 \log(\delta^{-1}) \cdot \left(\frac{16C_2^2(t_1 - t_0)}{C_1^4(t_0 + L - 1)^2(t_0 + L)^2} + \sum_{t=t_0+1}^{t_1} \frac{4C_3^2(t + L - 1)^2 r_{t-1}}{C_1^2(t_0 + L - 1)^2(t_0 + L)^2} \right)}.$$

Now, let us tune the sequence $\{r_t\}$ in a way such that

$$\mathbb{P}\left(\max_{t_0+1 \leq t \leq t_1} U_t^* \geq D(\delta) \mid A\right) = \mathbb{P}\left(\max_{t_0+1 \leq t \leq t_1} U_{\tau \wedge t}^* \geq D(\delta) \mid A\right).$$

By Lemma 3, the display above holds if for any $1 \leq t \leq T$,

$$\mathbb{P}\left(\tau \geq t + 1 \mid \max_{1 \leq k \leq t} U_k^* \leq D(\delta), A\right) = 1.$$

To get this, it suffices to choose r_k 's in a way such that

$$b_t(a_0 + D(\delta)) \leq r_t$$

for all $t_0 \leq t \leq t_1$.

For $t \in [t_0 + 1, t_1]$, let us take

$$r_t = \frac{M(t_1 - t_0) \log(\delta^{-1})}{(t + L)^2}$$

where the number M is chosen to be large enough and will be specified shortly. For any $t \in [t_0 + 1, t_1]$, we have

$$\begin{aligned} & \frac{(t_0 + L - 1)(t_0 + L)}{(t + L - 1)(t + L)} (a_0 + D(\delta)) \\ & \leq \frac{(t_0 + L - 1)(t_0 + L)}{(t + L - 1)(t + L)} \left(a_0 + \frac{4C_2(t_1 - t_0)}{C_1^2(t_0 + L - 1)(t_0 + L)} \right) \\ & \quad + 2\sqrt{\log(\delta^{-1})} \cdot \frac{(t_0 + L - 1)(t_0 + L)}{(t + L - 1)(t + L)} \left(\frac{4C_2 \sqrt{t_1 - t_0}}{C_1^2(t_0 + L - 1)(t_0 + L)} + \frac{2C_3 \sqrt{M \log(\delta^{-1})}(t_1 - t_0)}{C_1(t_0 + L - 1)(t_0 + L)} \right) \end{aligned}$$

To make sure that the display above is smaller than $r_t = \frac{M(t_1 - t_0) \log(\delta^{-1})}{(t + L)^2}$, we need

$$\begin{aligned} & a_0(t_0 + L - 1)(t_0 + L) + \frac{4C_2(t_1 - t_0)}{C_1^2} + \frac{8C_2 \sqrt{(t_1 - t_0) \log(\delta^{-1})}}{C_1^2} \\ & + \frac{4C_3 \sqrt{M} \log(\delta^{-1})(t_1 - t_0)}{C_1} \leq \frac{t + L - 1}{t + L} \cdot M(t_1 - t_0) \log(\delta^{-1}) \end{aligned}$$

for all $t \in [t_0 + 1, t_1]$.

By observing that $(t + L - 1)/(t + L) \geq (L - 1)/L$ for all $t \geq 0$, it suffices to pick M such that

$$\begin{aligned} x \cdot \frac{L}{L-1} M(t_1 - t_0) \log(\delta^{-1}) &\geq a_0(t_0 + L - 1)(t_0 + L); \\ y \cdot \frac{L}{L-1} M(t_1 - t_0) \log(\delta^{-1}) &\geq \max \left\{ \frac{4C_2(t_1 - t_0)}{C_1^2}; \frac{8C_2 \sqrt{(t_1 - t_0) \log(\delta^{-1})}}{C_1^2} \right\}; \\ z \cdot \frac{L}{L-1} M(t_1 - t_0) \log(\delta^{-1}) &\geq \frac{4C_3 \sqrt{M}(t_1 - t_0) \log(\delta^{-1})}{C_1}. \end{aligned}$$

for some positive numbers x, y, z such that $x + y + z = 1$. Thus, we can pick M as

$$\begin{aligned} M &= \frac{L-1}{L} \cdot \min_{\substack{x,y,z>0 \\ x+y+z=1}} \left\{ \max \left\{ \frac{1}{x}; \frac{8}{y}; \frac{16}{z^2} \right\} \right\} \\ &\times \max \left\{ \frac{a_0(t_0 + L - 1)(t_0 + L)}{\log(\delta^{-1})(t_1 - t_0)}; \frac{C_2}{C_1^2 \log(\delta^{-1})}; \frac{C_2}{C_1^2 \sqrt{\log(\delta^{-1})}}; \frac{C_3^2}{C_1^2} \right\}. \end{aligned}$$

By Lemma 11, we have

$$\min_{\substack{x,y,z>0 \\ x+y+z=1}} \left\{ \max \left\{ \frac{1}{x}; \frac{8}{y}; \frac{16}{z^2} \right\} \right\} \leq 31.5.$$

Thus, we can pick

$$M = \frac{31.5(L-1)}{L} \cdot \max \left\{ \frac{a_0(t_0 + L - 1)(t_0 + L)}{\log(\delta^{-1})(t_1 - t_0)}; \frac{C_2}{C_1^2 \log(\delta^{-1})}; \frac{C_2}{C_1^2 \sqrt{\log(\delta^{-1})}}; \frac{C_3^2}{C_1^2} \right\}.$$

The proof is completed. \square

9 Proof of Theorem 2

Let us pick $t_k = \alpha^{k+1} - \alpha$, where $\alpha = \max\{L-1, 33\}$. We will apply Lemma 1 repeatedly on the intervals $[t_{k-1} + 1, t_k]$, for $k \geq 1$ and $\delta_i = \delta/(i+1)^2$. With a specified a_0 , define

$$\begin{aligned} M_k &:= 31.5 \times \max \left\{ \frac{a_{k-1}(t_{k-1} + L - 1)(t_{k-1} + L)}{\log(\delta_k^{-1})(t_k - t_{k-1})}; \frac{C_2}{C_1^2}; \frac{C_3^2}{C_1^2} \right\}; \\ r(k, t) &:= \frac{M_k(t_k - t_{k-1}) \log(\delta_k^{-1})}{(t + L)^2} \text{ for } t \in [t_{k-1} + 1, t_k]; \\ a_{k+1} &:= r(k+1, t_{k+1}). \end{aligned}$$

It is useful to think of this as $a \rightarrow M \rightarrow r \rightarrow a$. By Lemma 1, for any $k \geq 1$, we have

$$\mathbb{P}(\exists t \in [t_{k-1} + 1, t_k] : L_t \geq r(k, t), L_{t_{k-1}} \leq a_{k-1}) \leq \delta_k.$$

Therefore, by using the union bound,

$$\mathbb{P}\left(\forall k \geq 1 : \max_{t \in [t_{k-1} + 1, t_k]} \frac{L_t}{r(k, t)} \leq 1\right) \geq 1 - \sum_{k \geq 0} \delta_k \geq 1 - 2\delta.$$

Next, let us show that by induction on k such that

$$a_k \leq A \cdot \frac{\log(\delta_k^{-1})}{t_k + L}$$

where

$$A := 31.5 \times \max \left\{ \frac{a_0 L}{\log(\delta^{-1})}, \frac{C_2}{C_1^2}, \frac{C_3^2}{C_1^2} \right\}.$$

For $k = 0$, the statement above is true since

$$a_0 \leq \frac{31.5 \times a_0 L}{\log(\delta^{-1})} \cdot \frac{\log(\delta_0^{-1})}{t_0 + L} = 31.5 \times a_0.$$

Assume the statement is true for k , let us consider $k + 1$. Note that

$$\begin{aligned} a_{k+1} &\leq A \cdot \frac{\log(\delta_{k+1}^{-1})}{t_{k+1} + L} \\ \iff \frac{M_{k+1} (t_{k+1} - t_k) \log(\delta_{k+1}^{-1})}{(t_{k+1} + L)^2} &\leq A \cdot \frac{\log(\delta_{k+1}^{-1})}{t_{k+1} + L} \\ \iff M_{k+1} \cdot \frac{t_{k+1} - t_k}{t_{k+1} + L} &\leq A \\ \iff \frac{31.5 \times a_k (t_k + L - 1)(t_k + L)}{\log(\delta_{k+1}^{-1})(t_{k+1} - t_k)} \cdot \frac{t_{k+1} - t_k}{t_{k+1} + L} &\leq A. \end{aligned}$$

By the induction hypothesis,

$$\begin{aligned} \frac{31.5 \times a_k (t_k + L - 1)(t_k + L)}{\log(\delta_{k+1}^{-1})(t_{k+1} - t_k)} \cdot \frac{t_{k+1} - t_k}{t_{k+1} + L} &= \frac{31.5 \times a_k (t_k + L - 1)(t_k + L)}{\log(\delta_{k+1}^{-1})(t_{k+1} + L)} \\ &\leq 31.5 \times \frac{A \log(\delta_k^{-1})}{\log(\delta_{k+1}^{-1})} \cdot \frac{(t_k + L - 1)(t_k + L)}{(t_{k+1} + L)(t_k + L)} \\ &\leq 31.5 \times A \cdot \frac{1}{\alpha} \leq A. \end{aligned}$$

Note that this also gives

$$\begin{aligned} \sup_{k \geq 1} M_k &\leq 31.5 \times \max \left\{ A \cdot \sup_{k \geq 1} \left\{ \frac{\log(\delta_{k-1}^{-1})}{\log(\delta_k^{-1})} \cdot \frac{(t_{k-1} + L - 1)(t_{k-1} + L)}{(\alpha - 1)(t_{k-1} + \alpha)(t_{k-1} + \alpha)} \right\}; \frac{C_2}{C_1^2}, \frac{C_3^2}{C_1^2} \right\} \\ &\leq A \end{aligned}$$

where we use Lemma 10 to get

$$\frac{t_{k-1} + L - 1}{(\alpha - 1)(t_{k-1} + \alpha)} \leq \frac{1}{31.5}$$

in the last line. To get the final convergence rate, note that for $t \in [t_{k-1} + 1, t_k]$,

$$\begin{aligned} r(k, t) &\leq A \cdot \frac{(\alpha - 1)(t_{k-1} + \alpha) [\log(\delta^{-1}) + 2 \log(k + 1)]}{(t + L)^2} \\ &\leq A \cdot (\alpha - 1) \cdot (\mathbf{1}_{\{L \geq 32\}} + 33 \cdot \mathbf{1}_{\{L \leq 31\}}) \cdot \frac{\log(\delta^{-1}) + 2 \log(k + 1)}{t + L}. \end{aligned}$$

Moreover, for all $k \geq 1$,

$$\begin{aligned} \log(k + 1) &\leq \log \log(33^k - 33 + 10) \\ &= \log \log(t_{k-1} + 10) \leq \log \log(t + 9). \end{aligned}$$

Thus, for $t \notin \{t_k; k \geq 0\}$, we have

$$r(k, t) \leq KA \cdot \frac{\log(\delta^{-1}) + 2 \log \log(t+9)}{t+L}.$$

where

$$K := (\alpha - 1) \cdot (\mathbf{1}_{\{L \geq 32\}} + 33 \cdot \mathbf{1}_{\{L \leq 31\}}).$$

Finally, at one of the epochs t_k ($k \geq 1$), we have

$$\begin{aligned} a_k &\leq A \cdot \frac{\log(\delta^{-1}) + 2 \log(k+1)}{t_k+L} \\ &\leq A \cdot \frac{\log(\delta^{-1}) + 2 \log \log(t_k+9)}{t_k+L}. \end{aligned}$$

Consequently,

$$\mathbb{P}\left(\exists t \geq 0 : L_t \geq 31.5 \times K \max \left\{ \frac{a_0 L}{\log(\delta^{-1})}, \frac{C_2}{C_1^2}, \frac{C_3^2}{C_1^2} \right\} \frac{\log(\delta^{-1}) + 2 \log \log(t+9)}{t+L} \right) \geq 1 - 2\delta.$$

The proof is completed. \square

10 Proof of Theorem 3

In what follows, S_t stands for L_t^{Oja} for simplicity. We split the proof of Theorem 3 into two steps.

Step 1: Confidence sequence for a modified process. Define the process

$$\begin{aligned} S_t^* &:= S_0; \\ S_t^* &:= S_t \cdot \prod_{i=0}^{t-1} \mathbf{1}_{\{S_i \leq 1/2\}}. \end{aligned}$$

Observe that for any $t \geq 1$,

$$\begin{aligned} S_t^* &\leq \prod_{i=0}^{t-1} \mathbf{1}_{\{S_i \leq 1/2\}} \cdot \left[(1 - 2\rho\eta_t) S_{t-1} + 2\rho\eta_t S_{t-1}^2 + Q_t + 4B^4\eta_t^2 \right] \\ &\leq (1 - 2\rho\eta_t) S_{t-1}^* \cdot \mathbf{1}_{\{S_{t-1} \leq 1/2\}} + 2\rho\eta_t (S_{t-1}^*)^2 \cdot \mathbf{1}_{\{S_{t-1} \leq 1/2\}} \\ &\quad + Q_t \cdot \prod_{i=0}^{t-1} \mathbf{1}_{\{S_i \leq 1/2\}} + 4B^4\eta_t^2 \\ &\leq (1 - 2\rho\eta_t) S_{t-1}^* + \rho\eta_t S_{t-1}^* + Q_t^* + 4B^4\eta_t^2 \\ &= (1 - \rho\eta_t) S_{t-1}^* + Q_t^* + 4B^4\eta_t^2 \end{aligned}$$

where

$$Q_t^* := Q_t \cdot \prod_{i=0}^{t-1} \mathbf{1}_{\{S_i \leq 1/2\}}.$$

Note that we still have

$$\mathbb{E}(Q_t^* | \mathcal{F}_{t-1}) = 0$$

and

$$|Q_t^*| \leq \eta_t B^2 \sqrt{S_{t-1}^*}.$$

Therefore, by applying Theorem 2 with the constants $C_1 \equiv \rho$, $C_2 \equiv 4B^4$, $C_3 \equiv B^2$ and $a \equiv 1/4$, and the choice of step sizes $\eta_t = 2/\rho(t+L)$, we obtain that

$$\mathbb{P}\left(\forall t \geq 0 : S_t^* \leq 31.5 \times K \cdot \max\left\{\frac{L}{4\log(\delta^{-1})}; \frac{B^4}{\rho^2}\right\} \frac{\log(\delta^{-1}) + 2\log\log(t+9)}{t+L}\right) \geq 1 - 2\delta \quad (58)$$

where $K = \max\{L-2; 32\} \cdot (\mathbf{1}_{\{L \geq 32\}} + 32 \cdot \mathbf{1}_{\{L \leq 31\}})$. In other words, (58) reduces to

$$\mathbb{P}\left(\forall t \geq 0 : S_t^* \leq 1008 \cdot \max\left\{\frac{L}{4\log(\delta^{-1})}; \frac{B^4}{\rho^2}\right\} \frac{\log(\delta^{-1}) + 2\log\log(t+9)}{t+L}\right) \geq 1 - 2\delta,$$

for $L \geq 32$. Note that the statement above holds because

$$\mathbb{P}(S_0^* \geq 1/4) = \mathbb{P}(S_0 \geq 1/4) \geq 1 - \delta^3 \geq 1 - \delta.$$

Step 2: The original process is close to the modified process. Let us show that if one sets

$$L = \left[128B^4 \log(\delta^{-1})^2 / \rho^2 \right]$$

and $\eta_t = 2/\rho(t+L)$, then

$$\mathbb{P}(\forall t \geq 0 : S_t = S_t^*) \geq 1 - 2e\delta.$$

By using Proposition 4, for any $t \geq 1$, we have that

$$\begin{aligned} \mathbb{E}[\exp(\lambda S_t) | \mathcal{F}_{t-1}] &\leq \mathbb{E}[\exp(\lambda S_{t-1} + \lambda Q_t + 4B^4 \lambda \eta_t^2) | \mathcal{F}_{t-1}] \\ &\leq \exp(S_{t-1} + 4B^4 \eta_t^2) \mathbb{E}[\exp(\lambda Q_t) | \mathcal{F}_{t-1}] \\ &\leq \exp(\lambda S_{t-1} + 4\lambda B^4 \eta_t^2) \exp(32\lambda^2 B^4 \eta_t^2), \end{aligned}$$

where the last estimate follows from Chernoff's inequality.

Thus, for any $\delta, \lambda > 0$,

$$\mathbb{E}[\exp(\lambda S_t) | \mathcal{F}_{t-1}] \leq \exp(\lambda S_{t-1} + 4\lambda B^4 \eta_t^2 + 32\lambda^2 B^4 \eta_t^2)$$

Consequently, for any $\lambda > 0$, the process

$$\{V_t\}_{t \geq 0} := \left\{ \exp\left[\lambda S_t + (4\lambda B^4 + 32\lambda^2 B^4) \cdot \sum_{k \geq t+1} \eta_k^2 \right] \right\}_{t \geq 0}$$

forms a non-negative super martingale. Here we define $V_0 := \lambda S_0$.

Moreover, for any $t \geq 1$,

$$\begin{aligned} \sum_{k \geq t+1} \eta_k^2 &\leq \sum_{k \geq 1} \eta_k^2 = \sum_{k=1}^{\infty} \frac{4}{\rho^2(k+L)^2} \\ &\leq \frac{4}{\rho^2} \int_{L+1}^{\infty} \frac{1}{x^2} dx = \frac{4}{\rho^2(L+1)}. \end{aligned}$$

Thus, Ville's inequality yields

$$\begin{aligned}\mathbb{P}\left(\sup_{t \geq 0} S_t \geq 1/2\right) &\leq \inf_{\lambda > 0} \left\{ \mathbb{P}\left(\sup_{t \geq 0} V_t \geq \frac{\lambda}{2} - \frac{16\lambda B^4(1+8\lambda)}{\rho^2(L+1)}\right) \right\} \\ &\leq \inf_{\lambda > 0} \left\{ \exp\left(-\frac{\lambda}{2} + \frac{16\lambda B^4(1+8\lambda)}{\rho^2(L+1)}\right) \mathbb{E}V_0 \right\} \\ &= \inf_{\lambda > 0} \left\{ \exp\left(-\frac{\lambda}{2} + \frac{32\lambda B^4(1+8\lambda)}{\rho^2(L+1)}\right) \cdot \mathbb{E}(\exp(\lambda S_0)) \right\}.\end{aligned}$$

To bound the last term, note that

$$\mathbb{E}(\exp(\lambda S_0)) \leq e^{\lambda/4} (1 - \delta^3) + e^\lambda \cdot \delta^3.$$

Consequently,

$$\mathbb{P}\left(\sup_{t \geq 0} S_t \geq 1/2\right) \leq \inf_{\lambda > 0} \left\{ \exp\left(\frac{32\lambda B^4(1+8\lambda)}{\rho^2(L+1)}\right) \cdot [e^{-\lambda/4} (1 - \delta^3) + e^{\lambda/2} \cdot \delta^3] \right\}.$$

By setting

$$\lambda = \frac{4}{3} \log\left(\frac{1 - \delta^3}{\delta^3}\right) \leq 4 \log(\delta^{-1}),$$

we get

$$\begin{aligned}\mathbb{P}\left(\sup_{t \geq 0} S_t \geq 1/2\right) &\leq \exp\left(128 \frac{B^4 \log(\delta^{-1})^2}{\rho^2(L+1)}\right) \cdot [2\delta \cdot (1 - \delta^3)^{2/3}] \\ &\leq 2e\delta\end{aligned}$$

whenever

$$L+1 \geq \frac{128B^4 \log(\delta^{-1})^2}{\rho^2}.$$

Thus,

$$\begin{aligned}\mathbb{P}(\forall t \geq 0 : S_t = S_t^*) &\geq 1 - \mathbb{P}\left(\sup_{t \geq 0} S_t \geq 1/2\right) \\ &\geq 1 - 2e\delta.\end{aligned}$$

The proof is completed by combining the expression above and (58). \square

11 Proof of Theorem 4

In what follows, S_t stands for L_t^{RM} for simplicity. Let us split the proofs in a few steps.

Step 1: X_n converges to 0 almost surely. Recall that for $S_t = (X_t - \theta)^2$, Proposition 5 gives the recursion

$$S_t \leq (1 - 2R\eta_t)S_{t-1} + Q_t + 2\eta_t^2 \left[\left(A \sqrt{S_{t-1}} + B \right)^2 + R_1^2 \right].$$

By taking the conditional expectation over \mathcal{F}_{t-1} and note that $\mathbb{E}(Q_t | \mathcal{F}_{t-1}) = 0$, we have

$$\begin{aligned}\mathbb{E}(S_t | \mathcal{F}_{t-1}) &\leq (1 - 2R\eta_t)S_{t-1} + 2\eta_t^2 \left(A^2 S_{t-1} + 2AB \sqrt{S_{t-1}} + B^2 + R_1^2 \right) \\ &\leq (1 - 2R\eta_t)S_{t-1} + 2\eta_t^2 \left(A^2 S_{t-1} + A^2 + B^2 S_{t-1} + R_1^2 \right) \\ &= S_{t-1} \left(1 + 2\eta_t^2(A^2 + B^2) \right) - 2R\eta_t S_{t-1} + 2\eta_t^2 \left(A^2 + R_1^2 \right)\end{aligned}$$

By applying the Robbins-Siegmund lemma in the form of (6), with

$$a_t = 2\eta_t^2(A^2 + B^2), \quad b_t = 2\eta_t^2(A^2 + R_1^2), \quad c_t = 2R\eta_t S_{t-1},$$

and noting that

$$\begin{aligned} \sum_{t \geq 1} a_t &\leq 2(A^2 + B^2)L_2^2 \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty, \\ \sum_{t \geq 1} b_t &\leq 2(A^2 + R_1^2)L_2^2 \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty, \end{aligned}$$

we conclude that S_t converges almost surely to a non-negative random variable S , and

$$\sum_{t \geq 1} c_t = \sum_{t=1}^{\infty} \eta_t S_{t-1} < \infty$$

almost surely.

If $P(S > \varepsilon) > \varepsilon$ for some $\varepsilon > 0$, then

$$\sum_{t=1}^{\infty} \eta_t S_{t-1} \geq L_1 \sum_{t=1}^{\infty} \frac{S_{t-1}}{t} \geq \frac{L_1 \varepsilon}{2} \cdot \limsup_{k \rightarrow \infty} \sum_{i=k}^{\infty} \frac{1}{i} = \infty$$

on the event $\{S > \varepsilon\}$. Thus,

$$\mathbb{P}\left(\sum_{t=1}^{\infty} \eta_t S_{t-1} = \infty\right) \geq \mathbb{P}(S > 0) > \varepsilon,$$

which is a contradiction. Therefore, X_t converges almost surely to θ .

Step 2: A moderate-deviation bound. For simplicity, write $Z_k = Y(X_k) - M(X_k)$. Note that $\{Z_k; k \geq 1\}$ are i.i.d. with mean 0 and unit variance. Moreover, for all $n \geq 1$, we have

$$X_{2^{n+1}} = X_{2^n+1} - \sum_{k=2^n+1}^{2^{n+1}} \eta_k M(X_k) - \sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k. \quad (59)$$

We will show that

$$\mathbb{P}\left(\sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \geq M \sqrt{\frac{\log \log 2^n}{2^n}}\right) \geq \frac{1}{n} \quad (60)$$

for all n sufficiently large, where

$$M := \sqrt{L_1/2}.$$

We will make use of the following moderate deviation result from [9] (Proposition 4.5 in the same paper for more details)

Lemma 6 ([9]). *Let $\{\xi_i : 1 \leq i \leq n\}$ be independent random variables satisfying $\mathbb{E}\xi_i = 0$ and $\mathbb{E} e^{t_n |\xi_i|} < \infty$ for some $t_n > 0$ and all $1 \leq i \leq n$. Assume furthermore that*

$$\sum_{i=1}^n \mathbb{E} \xi_i^2 = 1. \quad (61)$$

Define

$$W = \sum_{i=1}^n \xi_i, \quad \gamma = \sum_{i=1}^n \mathbb{E}[|\xi_i|^3 e^{x|\xi_i|}].$$

Then, for $0 \leq x \leq t_n$,

$$\frac{\mathbb{P}(W \geq x)}{1 - \Phi(x)} = 1 + O(1)(1 + x^3)\gamma e^{4x^3\gamma},$$

where Φ is the CDF of a standard Gaussian and the $O(1)$ term is universal.

Put

$$\sigma_n^2 := \text{Var} \left(\sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \right) = \sum_{k=2^n+1}^{2^{n+1}} \eta_k^2.$$

It is easy to check that

$$L_1 \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{k^2} \leq \sigma_n^2 \leq L_2 \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{k^2}.$$

Thus,

$$\begin{aligned} \sigma_n^2 &\geq L_1 \int_{2^n+1}^{2^{n+1}} \frac{1}{x^2} = L_1 \left(\frac{1}{2^n+1} - \frac{1}{2^{n+1}} \right) \\ &= L_1 \cdot \frac{2^{n+1} - 2^n - 1}{2^{n+1}(2^n+1)} \\ &= L_1 \cdot \frac{2^n - 1}{2^{n+1}(2^n+1)}, \end{aligned}$$

and

$$\sigma_n^2 \leq L_2 \int_{2^n}^{2^{n+1}} \frac{1}{x^2} = L_2 \cdot \left(\frac{1}{2^n} - \frac{1}{2^{n+1}} \right) = \frac{L_2}{2^n}.$$

Consequently, for all $n \geq 10$,

$$\frac{L_1}{2^{n+1}} \leq \sigma_n^2 \leq \frac{L_2}{2^n}.$$

We now employ Lemma 6 to get

$$\begin{aligned} &\frac{\mathbb{P} \left(\sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \geq M \sqrt{\frac{\log \log 2^n}{2^n}} \right)}{1 - \Phi \left(\frac{M \sqrt{\log \log 2^n}}{\sigma_n 2^{n/2}} \right)} \\ &= 1 + O(1) \left(1 + x_n^3 \right) \gamma_n \exp \left(4x_n^3 \gamma_n \right) \end{aligned}$$

where

$$\begin{aligned} x_n &:= \frac{M}{\sigma_n \cdot 2^{n/2}} \cdot \sqrt{\log \log 2^n}; \\ \gamma_n &:= \sum_{k=2^n+1}^{2^{n+1}} \frac{\eta_k^3}{\sigma_n^3} \cdot \mathbb{E} \left(|Z_k|^3 \cdot \exp \left(\frac{x_n \eta_k Z_k}{\sigma_n} \right) \right). \end{aligned}$$

It is easy to see that

$$x_n \leq \frac{M \sqrt{2}}{\sqrt{L_1}} \sqrt{\log(n + \log 2)}.$$

Let us estimate the size of γ_n . Observe that for all $k \in [2^n + 1, 2^{n+1}]$, we have

$$\eta_k \leq \frac{L_2}{k} \leq \frac{L_2}{2^n + 1}.$$

Thus, for all n sufficiently large,

$$\begin{aligned}
\gamma_n &\leq \sigma_n^{-3} \cdot \max_{2^n+1 \leq k \leq 2^{n+1}} \left\{ \mathbb{E} \left(|Z_k|^3 \cdot \exp \left(\frac{x_n \eta_k Z_k}{\sigma_n} \right) \right) \right\} \cdot \sum_{k=2^n+1}^{2^{n+1}} \frac{L_2^3}{k^3} \\
&\leq \frac{2^{3/2(n+1)}}{L_1^{3/2}} \cdot \max_{2^n+1 \leq k \leq 2^{n+1}} \left\{ R_1^3 \cdot \underbrace{\exp \left(\frac{M \sqrt{2}}{\sqrt{L_1}} \sqrt{\log(n + \log 2)} \cdot \frac{L_2 R_1}{2^n + 1} \cdot \sqrt{\frac{2^{n+1}}{L_1}} \right)}_{\leq 1 \text{ when } n \text{ is large}} \right\} \\
&\times L_2^3 \cdot \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{k^3} \\
&\leq \frac{2^{3/2(n+1)} L_2^3}{L_1^{3/2}} \cdot (R_1^3 e) \cdot \int_{2^n}^{2^{n+1}} \frac{1}{x^3} dx = \frac{3 L_2^3 R_1^3 e}{2 L_1^{3/2}} \cdot \frac{2^{3/2(n+1)}}{2^{2n}} = \frac{3 \sqrt{2} L_2^3 R_1^3 e}{L_1^{3/2}} \cdot 2^{-n/2}.
\end{aligned}$$

Therefore, for all n large enough, we have

$$\frac{\mathbb{P} \left(\sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \geq M \sqrt{\frac{\log \log 2^n}{2^n}} \right)}{1 - \Phi \left(\frac{M \sqrt{\log \log 2^n}}{\sigma_n 2^{n/2}} \right)} \geq 1/2.$$

By using the tail bound $1 - \Phi(x) \geq (2\pi)^{-1/2} \cdot (2x)^{-1} \cdot e^{-x^2/2}$ (which holds for all x sufficiently large), we obtain

$$\begin{aligned}
\mathbb{P} \left(\sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \geq M \sqrt{\frac{\log \log 2^n}{2^n}} \right) &\geq \frac{1}{4\sqrt{2\pi}} \cdot \frac{\sigma_n \cdot 2^{n/2}}{M \sqrt{\log \log 2^n}} \cdot \exp \left(-\frac{M^2 \log \log 2^n}{2\sigma_n^2 2^n} \right) \\
&\geq \frac{\sqrt{L_1}}{8\sqrt{\pi}} \cdot \frac{1}{M \sqrt{\log(n + \log 2)}} \cdot (n + \log 2)^{-\frac{M^2}{L_1}} \\
&= \frac{\sqrt{L_1}}{8\sqrt{\pi}} \cdot \frac{1}{M \sqrt{\log(n + \log 2)}} \cdot (n + \log 2)^{-1/2} \\
&\geq \frac{1}{n}
\end{aligned}$$

for all n sufficiently large where we have used the fact that $M = \sqrt{L_1/2}$ in the third display.

Step 3: Wrap-up. From (60), we conclude that

$$\sum_{n=1}^{\infty} \mathbb{P} \left(\sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \geq M \sqrt{\frac{\log \log 2^n}{2^n}} \right) \geq \liminf_{k \rightarrow \infty} \left(\sum_{i=k}^{\infty} \frac{1}{i} \right) = \infty.$$

Note that the events $\left\{ \sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \geq M \sqrt{\frac{\log \log 2^n}{2^n}} \right\}$ are independent, so Borel–Cantelli lemma yields

$$\sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \geq \sqrt{\frac{L_1}{2}} \cdot \sqrt{\frac{\log \log 2^n}{2^n}} \tag{62}$$

eventually always with probability one.

Recall L from (37). Define the event

$$A = \left\{ \limsup_{t \rightarrow \infty} \left(\sqrt{\frac{t}{\log \log t}} \cdot |X_t - \theta| \right) \leq L \right\}.$$

If $\mathbb{P}(A) = 0$, then the proof is completed. Assume $\mathbb{P}(A) > 0$. From Step 1, we know that X_t converges almost surely to 0, which leads to

$$|M(X_k)| = |M(X_k) - M(\theta)| \leq 2M'(\theta) \cdot |X_k - \theta|$$

for all k sufficiently large, with probability one.

Now, on the event A , we have

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \left\{ \sqrt{\frac{2^n}{\log \log 2^n}} \cdot |X_{2^{n+1}} - X_{2^{n+1}}| \right\} \\ & \leq \limsup_{n \rightarrow \infty} \left\{ \sqrt{\frac{2^n}{\log \log 2^n}} \cdot |X_{2^{n+1}} - \theta| + \sqrt{\frac{2^n}{\log \log 2^n}} \cdot |X_{2^{n+1}} - \theta| \right\} \\ & \leq L \cdot \limsup_{n \rightarrow \infty} \left\{ \sqrt{\frac{2^n}{\log \log 2^n}} \cdot \left(\sqrt{\frac{\log \log (2^n + 1)}{2^n + 1}} + \sqrt{\frac{\log \log (2^{n+1})}{2^{n+1}}} \right) \right\} \\ & \leq L \cdot \left(1 + \frac{1}{\sqrt{2}} \right) < 2L. \end{aligned}$$

Moreover, from (59), we have

$$\begin{aligned} & \sqrt{\frac{2^n}{\log \log 2^n}} \cdot \left| \sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \right| \\ & = \sqrt{\frac{2^n}{\log \log 2^n}} \cdot \left| X_{2^{n+1}} - X_{2^n+1} + \sum_{k=2^n+1}^{2^{n+1}} \eta_k M(X_k) \right| \\ & \leq \sqrt{\frac{2^n}{\log \log 2^n}} \cdot |X_{2^{n+1}} - X_{2^n+1}| + \sqrt{\frac{2^n}{\log \log 2^n}} \cdot \left| \sum_{k=2^n+1}^{2^{n+1}} \eta_k M(X_k) \right|, \end{aligned}$$

which leads to

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \left\{ \sqrt{\frac{2^n}{\log \log 2^n}} \cdot \left| \sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \right| \right\} \\ & \leq \limsup_{n \rightarrow \infty} \left\{ \sqrt{\frac{2^n}{\log \log 2^n}} \cdot |X_{2^{n+1}} - X_{2^n+1}| \right\} \\ & + \limsup_{n \rightarrow \infty} \left\{ \sqrt{\frac{2^n}{\log \log 2^n}} \cdot 2L_2 M'(\theta) \cdot 2L \cdot \sum_{k=2^n+1}^{2^{n+1}} \sqrt{\frac{\log \log k}{k}} \cdot \frac{1}{k} \right\}. \end{aligned}$$

On the event A , by noting that the function $x \rightarrow \log \log(x)/x$ is decreasing, we get

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \left\{ \sqrt{\frac{2^n}{\log \log 2^n}} \cdot \left| \sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \right| \right\} \\ & \leq 2L + 4L L_2 M'(\theta) \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{k} \\ & \leq 2L \left(1 + 2L_2 M'(\theta) \log 2 \right) = \frac{\sqrt{L_1}}{2}. \end{aligned}$$

Therefore, on the event A ,

$$\sqrt{\frac{2^n}{\log \log 2^n}} \cdot \sum_{k=2^n+1}^{2^{n+1}} \eta_k Z_k \leq \frac{\sqrt{L_1}}{2}$$

eventually always, which is a contradiction to (62). The proof is completed. \square

A Extension to the sub-Gaussian coefficients settings

A.1 Extension of Theorem 1

In this section, we extend Theorem 1 to the settings where the coefficients A_i, B_i in (11) and (12) are allowed to be random, positive random variables. The only coefficient that has to be deterministic is C_1 in (10). It turns out that in such settings, the conclusion of Theorem 1 is still valid under a sub-Gaussian moment condition on these coefficients.

Let us first restate (11) and (12) as

$$|\mathbb{E}(U_t|\mathcal{F}_{t-1})| \leq \mathbb{E}(C_2^{(t)}|\mathcal{F}_{t-1}) \cdot \eta_t^2 + \sum_{i=1}^m \mathbb{E}(\mathcal{A}_i^{(t)}|\mathcal{F}_{t-1}) \cdot \eta_t^{1+a_i} L_{t-1}^{b_i}; \quad (63)$$

$$|U_t| \leq C_3^{(t)} \eta_t \sqrt{L_{t-1}} + \sum_{i=1}^m \mathcal{B}_i^{(t)} \cdot \eta_t^{1/2+c_i} \cdot L_{t-1}^{d_i}. \quad (64)$$

where $\{C_2^{(t)}; C_3^{(t)}; \mathcal{A}_i^{(t)}; \mathcal{B}_i^{(t)}; 1 \leq i \leq m; t \geq 1\}$ are positive random variables.

Assumption 1. *There exists $\sigma > 0$ such that*

(i).

$$\max \left\{ \mathbb{E} \left[\exp \left(\frac{|C_3^{(t)}|^2}{\sigma^2} \right) \middle| \mathcal{F}_{t-1} \right]; \left[\exp \left(\frac{|\mathcal{B}_1^{(t)}|^2}{\sigma^2} \right) \middle| \mathcal{F}_{t-1} \right]; \dots; \left[\exp \left(\frac{|\mathcal{B}_m^{(t)}|^2}{\sigma^2} \right) \middle| \mathcal{F}_{t-1} \right] \right\} \leq 2 \quad (65)$$

almost surely.

(ii). We have

$$\max_{i \geq 1} \{ \mathbb{E}(C_2^{(t)}|\mathcal{F}_{t-1}); \mathbb{E}(\mathcal{A}_1^{(t)}|\mathcal{F}_{t-1}); \dots; \mathbb{E}(\mathcal{A}_m^{(t)}|\mathcal{F}_{t-1}) \} \leq \sigma \quad (66)$$

almost surely.

Condition (65) states that the time-dependent coefficients in (64) are conditionally sub-Gaussian with respect to the sigma-fields \mathcal{F}_{t-1} . It is further required that the conditional sub-Gaussian parameter is uniformly bounded above by a deterministic constant σ . In other words, we assume a uniform bound on the conditional 2-Orlicz norm of these coefficients; see (67) below for a precise definition. At present, it is unclear whether the results extend to the sub-exponential setting. However, in the homogeneous setting, such as in Proposition 1 or Theorem 2, an extension is possible when the coefficient corresponding to η_t^2 is sub-exponential with the proof being similar to that of Theorem 1A below.

Unfortunately, these conditions are not satisfied by Oja's algorithm, since the corresponding recursion (see (33) with $U_t \equiv Q_t$) takes the form

$$|U_t| \lesssim \|X_t\|^2 \eta_t \sqrt{L_{t-1}^{Oja}} + \|X_t\|^2 \cdot (\text{eigen gap}) \cdot \eta_t (L_{t-1}^{Oja})^2 + \eta_t^2 \|X_t\|^4.$$

The three coefficients in the recursive bound above are $\{\|X_t\|^2, \|X_t\|^2, \|X_t\|^4\}$. The last coefficient, $\|X_t\|^4$, follows a sub-Weibull distribution, even when the data are Gaussian.

Condition (66) asserts that the time-dependent coefficients in (63) are almost surely bounded by a deterministic constant σ . As noted in Section 2, this is a natural extension of (12), and it is satisfied in most examples considered in the present paper. In fact, it is often the case that the coefficients $C_2^{(t)}$ and $\{\mathcal{A}_i^{(t)} : 1 \leq i \leq m\}$ are functionals of the collection $\{C_3^{(t)}, \mathcal{B}_1^{(t)}, \mathcal{B}_2^{(t)}, \dots, \mathcal{B}_m^{(t)}\}$, and therefore remain bounded.

Under Assumption 1, an analogue of Theorem 1 holds, which we state below:

Theorem 1A. *Assume (63), (64), and suppose that Assumption 1 holds for some $\sigma > 0$. Assume additionally that*

$$\min_{1 \leq i \leq m} \{(a_i + b_i) \wedge (c_i + d_i)\} > 1.$$

Then, for any $\delta \in (0, e^{-2})$, there exist constants $M, r > 0$ independent of t , and a step-size sequence $\{\eta_t\}_{t \geq 1}$, such that

$$\mathbb{P}\left(\forall t \geq 0 : L_t \leq M \cdot \frac{\log(\delta^{-1}) + \log \log(t+10)}{t+10}\right) \geq 1 - 2\delta,$$

provided that

$$\mathbb{P}(L_0 \leq r) \geq 1 - \delta.$$

Moreover, the step sizes satisfy $\eta_t \asymp 1/t$ as $t \rightarrow \infty$.

Simnilar to Theorem 1, for small δ , M scales like $O(1/\log(\delta^{-1}))$ and r scales like $O(1/\text{polylog}(\delta^{-1}))$, with respect to σ and the constants $\{a_i, b_i, c_i, d_i ; 1 \leq i \leq m\}$. The proof of Theorem 1A is similar to that of Theorem 1, albeit some changes in the maximal inequality.

A.2 Extension of Proposition 1 and Theorem 2

A.3 Some technical results

We collect some technical results and facts that are needed in this subsection. For $p \geq 1$, define the p -Orlicz norm

$$\|X\|_{\Psi_p} := \inf \left\{ t > 0 : \exp\left(\frac{|X|^p}{t^p}\right) \leq 2 \right\}. \quad (67)$$

It is well-known that $\|\cdot\|_{\Psi_p}$ is a proper norm for all $p \geq 1$. The following lemma is adapted from [53] to have explicit constants.

Lemma 7. *For $p \in \{1; 2\}$, we have*

$$\sup_X \frac{\|X - \mathbb{E}X\|_{\Psi_p}}{\|X\|_{\Psi_p}} \leq 2$$

where the supremum is taken over all random variables X that is not equal to zero with probability one.

Proof of Lemma 7. Consider the case $p = 1$. Take $\lambda > 0$ and note that

$$\begin{aligned} \mathbb{E}[\exp(\lambda|X - \mathbb{E}X|)] &\leq \mathbb{E}(\exp(\lambda|X| + \lambda|\mathbb{E}X|)) \\ &\leq \exp(\lambda|\mathbb{E}X|) \cdot \mathbb{E}(\exp(\lambda|X|)) \\ &\leq [\mathbb{E}(e^{\lambda|X|})]^2 \leq \mathbb{E}(e^{2\lambda|X|}). \end{aligned}$$

where the last line follows from Jensen's inequality.

Take $\varepsilon > 0$ and choose $\lambda = 1/2 (\|X\|_{\Psi_1} + \varepsilon)$, we get

$$\mathbb{E} \left[\exp \left(\frac{|X - \mathbb{E}X|}{2(\|X\|_{\Psi_1} + \varepsilon)} \right) \right] \leq \mathbb{E} \left[\exp \left(\frac{|X|}{\|X\|_{\Psi_1} + \varepsilon} \right) \right] \leq 2.$$

Thus,

$$\|X - \mathbb{E}X\|_{\Psi_1} \leq 2(\|X\|_{\Psi_1} + \varepsilon).$$

By letting $\varepsilon \rightarrow 0$, we conclude the proof for the case $p = 1$.

Let us turn to the case $p = 2$. By using similar argument to the case $p = 1$, we have

$$\begin{aligned} \mathbb{E} \left[\exp \left(\lambda |X - \mathbb{E}X|^2 \right) \right] &\leq \mathbb{E} \left[\exp \left(2\lambda X^2 + 2\lambda \mathbb{E}(X^2) \right) \right] \\ &\leq \mathbb{E} \left[e^{4\lambda X^2} \right] \end{aligned}$$

for all $\lambda > 0$. By taking $\lambda = 1/4 (\|X\|_{\Psi_2}^2 + \varepsilon)$ for $\varepsilon > 0$, we obtain

$$\mathbb{E} \left[\exp \left(\frac{|X - \mathbb{E}X|^2}{4(\|X\|_{\Psi_2}^2 + \varepsilon)} \right) \right] \leq \mathbb{E} \left[\exp \left(\frac{|X|}{\|X\|_{\Psi_2}^2 + \varepsilon} \right) \right] \leq 2.$$

Thus,

$$\|X - \mathbb{E}X\|_{\Psi_2} \leq 2 \sqrt{\|X\|_{\Psi_2}^2 + \varepsilon}.$$

The proof is completed by taking $\varepsilon \rightarrow 0$. \square

Lemma 8. *Let X be a random variable such that $\|X\|_{\Psi_p} < \infty$ for $p \in \{1; 2\}$. Then,*

- If $p = 1$, then

$$\mathbb{E} [\exp(\lambda(X - \mathbb{E}X))] \leq \exp(10\lambda^2 \|X\|_{\Psi_1}^2)$$

for all $|\lambda| \leq 1/(4\|X\|_{\Psi_1})$.

- If $p = 2$, then

$$\mathbb{E} [\exp(\lambda(X - \mathbb{E}X))] \leq \exp(6\lambda^2 \|X\|_{\Psi_2}^2)$$

for all $\lambda \in \mathbb{R}$.

Proof of Lemma 8. Put $Y = X - \mathbb{E}X$. Then, by Lemma 7, we have $\|Y\|_{\Psi_p} \leq 2\|X\|_{\Psi_p}$. Take $\varepsilon > 0$ and define the rescaled version of Y

$$Y_1 := \frac{Y}{2\|X\|_{\Psi_p} + \varepsilon}.$$

It is obvious that $\|Y_1\|_{\Psi_p} < 1$ and thus, the elementary inequality $e^x \leq 1 + x + (x^2/2)e^{|x|}$ gives

$$\mathbb{E} [\exp(\lambda Y_1)] \leq 1 + \mathbb{E} \left[\frac{\lambda^2 Y_1^2}{2} \cdot e^{|\lambda Y_1|} \right]. \quad (68)$$

Consider the case $p = 1$. Observe that $x^2 \leq (16/e^2)e^{|x|/2}$ so (68) implies

$$\mathbb{E} [\exp(\lambda Y_1)] \leq 1 + (8/e^2)\lambda^2 \mathbb{E} [e^{|Y_1|/2 + \lambda|Y_1|}].$$

For all $|\lambda| \leq 1/2$, we have

$$\mathbb{E} [\exp(\lambda Y_1)] \leq 1 + (8/e^2)\lambda^2 \mathbb{E} (e^{|Y_1|}) \leq 1 + (16/e^2)\lambda^2 \leq \exp((16/e^2)\lambda^2).$$

Consequently,

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp\left(\frac{16}{e^2} (2\|X\|_{\Psi_1} + \varepsilon)^2 \lambda^2\right)$$

for all λ satisfies

$$|\lambda| \leq \frac{1}{4\|X\|_{\Psi_1} + 2\varepsilon}.$$

By letting $\varepsilon \rightarrow 0$, we obtain

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp(9\|X\|_{\Psi_1}^2 \lambda^2)$$

for all $|\lambda| \leq 1/(4\|X\|_{\Psi_1})$.

Next, consider the case $p = 2$. By using the bound $x \leq e^x$ and Cauchy-Schwartz inequality in (68), we obtain

$$\begin{aligned} \mathbb{E}[\exp(\lambda Y_1)] &\leq 1 + \frac{\lambda^2}{2} e^{\lambda^2/2} \cdot \mathbb{E}(e^{Y_1^2}) \\ &\leq 1 + \lambda^2 e^{\lambda^2/2} \\ &\leq (1 + \lambda^2) e^{\lambda^2/2} \leq e^{(3/2)\lambda^2}, \end{aligned}$$

for all $\lambda \in \mathbb{R}$.

Consequently, by letting $\varepsilon \rightarrow 0$,

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp(6\|X\|_{\Psi_1}^2 \lambda^2)$$

for all $\lambda \in \mathbb{R}$. The proof is completed. \square

We now modify Lemma 4 to account for the random coefficients setting.

Lemma 3A. *Let M_t be an adapted process with the corresponding filtration $\{\mathcal{F}_t; t \geq 0\}$ with $M_0 = 0$. Suppose X_t and Y_t are positive random variables such that*

$$\begin{aligned} |M_t - M_{t-1}| &\leq X_t, \\ \left| \mathbb{E}(M_t - M_{t-1} \mid \mathcal{F}_{t-1}) \right| &\leq Y_t. \end{aligned}$$

Assume additionally that

$$\mathbb{E}\left[\exp\left(\frac{|X_t|}{c_t}\right) \mid \mathcal{F}_{t-1}\right] \leq 2$$

almost surely for all $t \in [1, T]$. Then,

$$P\left(\exists t \in [1, T] : M_t \geq V_T^*(\delta) + \sum_{i=1}^t Y_i\right) \leq \delta$$

for all $T \geq 1$ and $\delta > 0$, where

$$\begin{aligned} V_T^*(\delta) &:= \inf_{|\lambda| < \lambda_0} \left\{ \frac{\log(\delta^{-1})}{\lambda} + 40\lambda \sum_{k=1}^T c_k^2 \right\}; \\ \lambda_0 &:= \frac{1}{8 \times \max_{1 \leq i \leq T} c_i}. \end{aligned}$$

Proof of Lemma 3A. As in the proof of Lemma 4, define $d_0 := 0$ and $d_k := M_k - M_{k-1}$. It is easy to check that

$$M_k = \sum_{i=1}^k [d_i - \mathbb{E}(d_i | \mathcal{F}_{i-1})] + \sum_{i=1}^k \mathbb{E}(d_i | \mathcal{F}_{i-1}).$$

Let $h_i := d_i - \mathbb{E}(d_i | \mathcal{F}_{i-1})$. It is easy to check that $\{h_i; 1 \leq i \leq T\}$ forms a martingale difference sequence and

$$\begin{aligned} |\mathbb{E}(d_i | \mathcal{F}_{i-1})| &\leq Y_i; \\ |d_i| &\leq X_i \end{aligned}$$

almost surely.

Conditional on \mathcal{F}_{t-1} , observe that

$$\|h_i\|_{\Psi_1} \leq 2\|X_i\|_{\Psi_1} \leq 2c_i$$

almost surely.

Thus, by Lemma 8 and the fact that $\mathbb{E}(h_i | \mathcal{F}_{i-1}) = 0$, we have

$$\mathbb{E}\left[e^{\lambda h_i} \middle| \mathcal{F}_{i-1}\right] \leq e^{40\lambda^2 c_i^2}$$

for all $|\lambda| \leq 1/(8c_i)$.

Therefore,

$$\mathbb{P}\left(\exists t \in [1, T] : M_t \geq x + \sum_{i=1}^t Y_i\right) \leq \mathbb{P}\left(\exists t \in [1, T] : \sum_{k=1}^t h_k \geq x\right)$$

for all $x > 0$.

To bound the last probability, note that the process

$$S_t := \exp\left(\lambda \sum_{k=1}^t h_k - 40\lambda^2 \sum_{k=1}^t c_k^2\right)$$

forms a non-negative supermartingale, which leads to

$$\begin{aligned} \mathbb{P}\left(\exists t \in [1, T] : \sum_{k=1}^t h_k \geq x\right) &\leq \mathbb{P}\left(\exists t \in [1, T] : \lambda \sum_{k=1}^t h_k - \lambda^2 \sum_{k=1}^t c_k^2 \geq \lambda x - \lambda^2 \sum_{k=1}^t c_k^2\right) \\ &\leq \exp\left(-\lambda x + 40\lambda^2 \sum_{k=1}^T c_k^2\right) \end{aligned}$$

where the last inequality follows from Ville's inequality.

Thus,

$$\mathbb{P}\left(\exists t \in [1, T] : \sum_{k=1}^t h_k \geq \frac{x}{\lambda} + 40\lambda \sum_{k=1}^T c_k^2\right) \leq e^{-x},$$

for all $|\lambda| < 1 / (8 \times \max\{c_1; c_2; \dots; c_T\})$.

The proof is completed by setting $x = \log(\delta^{-1})$. □

Remark 1. If one assumes X_t is conditionally sub-Gaussian in Lemma 3A, the variance term $V_T^*(\delta)$ is the same as in Lemma 4, up to a constant factor, that is

$$V_T^{SG}(\delta) \asymp \sqrt{\log(\delta^{-1}) \sum_{k=1}^T c_k^2}.$$

This corresponds to the sub-Gaussian settings in Theorem 1A.

A.4 Proof of Theorem 1A

It suffices to show a variant of Lemma 5 under conditions (63) and (64). Once such a result is proven, the stitching argument can be kept unchanged. An analog of Lemma 5 under (63) and (64) is

Lemma 4A. *Let X_0, X_1, \dots, X_n be a sequence of random variables that satisfy (10) with constant step sizes η . Suppose the noise process in (10) satisfies (63) and (64) with the corresponding parameter σ . Then, there exists a number $D_{\sigma, C_1, m}$ that depends only on σ, C_1 and m such that with*

$$D_{\delta}^*(r) := \frac{D_{\sigma, C_1, m}}{(1 - C_1 \eta)^n} \cdot \left(\sqrt{\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\eta r} + \sum_{i=1}^m \eta^{c_i} r^{d_i} \right) + \eta + \sum_{i=1}^m \eta^{a_i} r^{b_i} \right),$$

we have

$$\mathbb{P}(\exists t \in [1, n] : L_t \geq (1 - C_1 \eta)^t (\varepsilon_0 + D_{\delta}^*(r)), A) \leq \delta$$

whenever $\varepsilon_0 + D_{\delta}^*(r) \leq r$ and $A \subset \{L_0 \leq \varepsilon_0\}$ such that $\mathbb{P}(A) > 0$.

Proof of Lemma 4A. The only step we need to adapt is the concentration inequality for the stopped process $\{U_{\tau \wedge t}^* ; 1 \leq t \leq n\}$. Recall the notation from the proof of Lemma 5. In this case, we still have Thus,

$$\begin{aligned} |U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^*| &\leq u_t; \\ \left| \mathbb{E}(U_{\tau \wedge t}^* - U_{\tau \wedge (t-1)}^* | \mathcal{F}_{t-1}) \right| &\leq v_t \end{aligned}$$

where

$$\begin{aligned} u_t &:= (1 - C_1 \eta)^{-t} \cdot \left[\mathbb{E}(C_2^{(t)} | \mathcal{F}_{t-1}) \eta^2 + \sum_{k=1}^m \mathbb{E}(\mathcal{A}_i^{(t)} | \mathcal{F}_{t-1}) \cdot \eta^{1+a_k} r^{b_k} \right]; \\ v_t &:= (1 - C_1 \eta)^{-t} \cdot \left(C_3^{(t)} \eta \sqrt{r} + \sum_{k=1}^m \mathcal{B}_k^{(t)} \eta^{1/2+c_k} r^{d_k} \right). \end{aligned}$$

By (64), we have

$$|u_t| \leq (1 - C_1 \eta)^{-t} \sigma \cdot \left(\eta^2 + \sum_{k=1}^m \eta^{1+a_k} r^{b_k} \right).$$

Furthermore, conditionally on \mathcal{F}_{t-1} ,

$$\|v_t\|_{\Psi_2} \leq (1 - C_1 \eta)^{-t} \sigma \left(\eta \sqrt{r} + \sum_{k=1}^m \eta^{1/2+c_k} r^{d_k} \right)$$

almost surely. Put

$$c_t := (1 - C_1 \eta)^{-t} \sigma \left(\eta \sqrt{r} + \sum_{k=1}^m \eta^{1/2+c_k} r^{d_k} \right).$$

Therefore, by using Remark 1, we have

$$\mathbb{P}(\exists t \in [1, n] : U_{\tau \wedge t}^* \geq D_{\delta}^*(r) | A) \leq \frac{\delta}{\mathbb{P}(A)}$$

where

$$D_\delta^*(r) \stackrel{\sigma, C_1, m}{\asymp} (1 - C_1 \eta)^{-n} \left(\sqrt{\log(\delta^{-1}) \cdot \left(\sum_{t=1}^n c_t^2 \right)} + \eta^2 + \sum_{k=1}^m \eta^{1+a_k} r^{b_k} \right)$$

$$\stackrel{\sigma, C_1, m}{\asymp} (1 - C_1 \eta)^{-n} \left(\sqrt{\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\eta r} + \sum_{i=1}^m \eta^{c_i} r^{d_i} \right) + \eta + \sum_{i=1}^m \eta^{a_i} r^{b_i} \right).$$

The rest of the proof can be kept unchanged. \square

Note that the form of $D_\delta^*(r)$ in Lemma 4A is the same as $D_\delta(r)$ in Lemma 5, up to a constant factor. Therefore, the argument in Section 7.2 carries over and this concludes Theorem 1A. \square

B Technical proofs

Proof of Lemma 1

Rewrite (8) in the form of (6), it is easy to check that $a_t = 0$, $b_t = \beta_t$ and $c_t = -\eta_t X_t$. Since $\sum_t \beta_t < \infty$ almost surely, the Robbins-Siegmund's lemma can be applied to obtain

$$X_t \xrightarrow{\text{a.s.}} X \text{ and } \sum_{t=1}^{\infty} \eta_t X_t < \infty \text{ a.s.}$$

It suffices to show that $X \equiv 0$ a.s. To see this, suppose there exists $\delta > 0$ such that $\mathbb{P}(X > \delta) > \delta$. Then, one can find a set A with probability at least $\delta/2$ such that

$$\liminf_{t \rightarrow \infty} X_t > \delta/2.$$

Therefore, the series $\sum_{t=1}^{\infty} \eta_t X_t$ diverges on A , which contradicts the condition $\sum_{t=1}^{\infty} \eta_t X_t < \infty$ almost surely. The proof is completed. \square

Proof of Proposition 4

In the proof below, S_t stands for L_t^{Oja} . Let us start with the error bound for (Kra-Oja). For simplicity, we will simply write \mathbf{v}_t instead of $\hat{\mathbf{v}}_t$. Define

$$y_t := \mathbf{X}_t^\top \mathbf{v}_{t-1},$$

$$\mathbf{z}_t := y_t \left[\mathbf{X}_t - \frac{y_t}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbf{v}_{t-1} \right].$$

Note that in the update rule of (Kra-Oja), \mathbf{z}_t is orthogonal to \mathbf{v}_{t-1} . To see this, write

$$\begin{aligned} & \left\langle \mathbf{v}_{t-1}, \eta_t y_t \left[\mathbf{X}_t - \frac{y_t}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbf{v}_{t-1} \right] \right\rangle \\ &= \eta_t y_t \left\langle \mathbf{v}_{t-1}, \mathbf{X}_t - \frac{y_t}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbf{v}_{t-1} \right\rangle \\ &= \eta_t y_t \left[\langle \mathbf{v}_{t-1}, \mathbf{X}_t \rangle - y_t \right] = 0. \end{aligned}$$

From the orthogonality between \mathbf{z}_t and \mathbf{v}_{t-1} , one has

$$\begin{aligned} \|\mathbf{v}_t\|^2 &= \|\mathbf{v}_{t-1}\|^2 + \eta_t^2 y_t^2 \left\| \mathbf{X}_t - \frac{y_t}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbf{v}_{t-1} \right\|^2 \\ &= \|\mathbf{v}_{t-1}\|^2 \left(1 + \eta_t^2 \cdot \frac{\langle \mathbf{v}_{t-1}, \mathbf{X}_t \rangle^2}{\|\mathbf{v}_{t-1}\|^2} \cdot \left\| \mathbf{X}_t - \frac{y_t}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbf{v}_{t-1} \right\|^2 \right). \end{aligned}$$

By Cauchy-Schwartz's inequality, it is easy to check that

$$\begin{aligned} \frac{\langle \mathbf{v}_{t-1}, \mathbf{X}_t \rangle^2}{\|\mathbf{v}_{t-1}\|^2} &\leq \|\mathbf{X}_t\|^2 \leq B^2, \\ \left\| \mathbf{X}_t - \frac{y_t}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbf{v}_{t-1} \right\|^2 &\leq 2\|\mathbf{X}_t\|^2 + \frac{2y_t^2}{\|\mathbf{v}_{t-1}\|^2} \\ &\leq 2B^2 + 2\frac{\langle \mathbf{v}_{t-1}, \mathbf{X}_t \rangle^2}{\|\mathbf{v}_{t-1}\|^2} \leq 4B^2. \end{aligned}$$

Thus,

$$\|\mathbf{v}_{t-1}\|^2 \leq \|\mathbf{v}_t\|^2 \leq \|\mathbf{v}_t\|^2 (1 + 4B^4\eta_t^2)$$

and

$$\|\mathbf{z}_t\|^2 \leq 4B^4\|\mathbf{v}_{t-1}\|^2$$

for all $t \geq 1$.

Recall that due to the observation in (32), we can assume the principal eigenvector is \mathbf{e}_1 . Write

$$\begin{aligned} S_t &= \frac{\|\mathbf{v}_t\|^2 - v_{t,1}^2}{\|\mathbf{v}_t\|^2} \\ &\leq \frac{\|\mathbf{v}_t\|^2 - v_{t,1}^2}{\|\mathbf{v}_{t-1}\|^2} \\ &\leq \frac{\|\mathbf{v}_{t-1}\|^2 + \eta_t^2\|\mathbf{z}_t\|^2 - (v_{t-1,1}^2 + 2\eta_t v_{t-1,1} z_{t,1} + \eta_t^2 z_{t,1}^2)}{\|\mathbf{v}_{t-1}\|^2} \\ &= \frac{\|\mathbf{v}_{t-1}\|^2 - v_{t-1,1}^2}{\|\mathbf{v}_{t-1}\|^2} + \eta_t^2 \cdot \frac{\|\mathbf{z}_t\|^2 - z_{t,1}^2}{\|\mathbf{v}_{t-1}\|^2} - \frac{2\eta_t v_{t-1,1} z_{t,1}}{\|\mathbf{v}_{t-1}\|^2} \\ &\leq S_{t-1} + 4B^4\eta_t^2 - \frac{2\eta_t v_{t-1,1} z_{t,1}}{\|\mathbf{v}_{t-1}\|^2}. \end{aligned}$$

Now define

$$Q_t := -\frac{2\eta_t v_{t-1,1} z_{t,1}}{\|\mathbf{v}_{t-1}\|^2} + \mathbb{E}\left(\frac{2\eta_t v_{t-1,1} z_{t,1}}{\|\mathbf{v}_{t-1}\|^2} \middle| \mathcal{F}_{t-1}\right).$$

By using the orthogonality between \mathbf{z}_t and \mathbf{v}_{t-1} , we have

$$\begin{aligned} |v_{t-1,1} z_{t,1}| &= \left| \sum_{k=2}^p v_{t-1,k} z_{t,k} \right| \\ &\leq \|\mathbf{z}_t\| \cdot \sqrt{\|\mathbf{v}_{t-1}\|^2 - v_{t-1,1}^2} \\ &= \|\mathbf{z}_t\| \cdot \|\mathbf{v}_{t-1}\| \cdot \sqrt{S_t}. \end{aligned}$$

Thus,

$$\begin{aligned} |Q_t| &\leq 2\eta_t \sqrt{S_{t-1}} \cdot \frac{\|\mathbf{z}_t\|}{\|\mathbf{v}_{t-1}\|} \\ &\leq 2\eta_t \sqrt{S_{t-1}} \cdot \frac{|\langle \mathbf{v}_{t-1}, \mathbf{X}_t \rangle|}{\|\mathbf{v}_{t-1}\|} \cdot \left\| \mathbf{X}_t - \frac{y_t}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbf{v}_{t-1} \right\| \\ &\leq 2\eta_t \sqrt{S_{t-1}} \cdot B \cdot 2B = 8B^2\eta_t \sqrt{S_{t-1}}. \end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{E}\left(\frac{2\eta_t v_{t-1,1} z_{t,1}}{\|\mathbf{v}_{t-1}\|^2} \middle| \mathcal{F}_{t-1}\right) &= \frac{2\eta_t v_{t-1,1}}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbb{E}(z_{t,1} \middle| \mathcal{F}_{t-1}) \\
&= \frac{2\eta_t v_{t-1,1}}{\|\mathbf{v}_{t-1}\|^2} \cdot \mathbb{E}\left[y_t \left(X_{t,1} - \frac{y_t v_{t-1,1}}{\|\mathbf{v}_{t-1}\|^2}\right) \middle| \mathcal{F}_{t-1}\right] \\
&= \frac{2\eta_t v_{t-1,1}}{\|\mathbf{v}_{t-1}\|^2} \cdot \left[\lambda_1 v_{t-1,1} - \frac{v_{t-1,1}}{\|\mathbf{v}_{t-1}\|^2} \mathbb{E}(y_t^2 \middle| \mathcal{F}_{t-1})\right] \\
&= \frac{2\eta_t \lambda_1 v_{t-1,1}^2}{\|\mathbf{v}_{t-1}\|^2} - \frac{2\eta_t v_{t-1,1}^2}{\|\mathbf{v}_{t-1}\|^4} \cdot \sum_{i=1}^p \lambda_i v_{t-1,i}^2 \\
&= 2\eta_t \lambda_1 (1 - S_{t-1}) - 2\eta_t (1 - S_{t-1}) \left[\lambda_1 \frac{v_{t-1,1}^2}{\|\mathbf{v}_{t-1}\|^2} + \sum_{i=2}^p \frac{\lambda_i v_{t-1,i}^2}{\|\mathbf{v}_{t-1}\|^2} \right] \\
&\geq 2\eta_t \lambda_1 (1 - S_{t-1}) - 2\eta_t \lambda_1 (1 - S_{t-1})^2 - 2\eta_t \lambda_2 S_{t-1} (1 - S_{t-1}) \\
&= 2\eta_t \lambda_1 (1 - S_{t-1}) S_{t-1} - 2\eta_t \lambda_2 S_{t-1} (1 - S_{t-1}) \\
&= 2\eta_t (\lambda_1 - \lambda_2) S_{t-1} (1 - S_{t-1}).
\end{aligned}$$

Consequently,

$$\begin{aligned}
S_t &\leq S_{t-1} + 4B^4 \eta_t^2 + Q_t - 2\eta_t (\lambda_1 - \lambda_2) S_{t-1} (1 - S_{t-1}) \\
&= S_{t-1} [1 - 2\eta_t (\lambda_1 - \lambda_2)] + 2\eta_t (\lambda_1 - \lambda_2) S_{t-1}^2 + Q_t + 4B^4 \eta_t^2.
\end{aligned}$$

This completes the proof of the error bound for (Kra-Oja). Let us consider (Oja). For this update rule, it is easy to check that

$$\begin{aligned}
\|\mathbf{v}_t\|^2 &= \|\mathbf{v}_{t-1}\|^2 + 2\eta_t y_t^2 + \eta_t^2 \|\mathbf{X}_t \mathbf{X}_t^\top \mathbf{v}_{t-1}\|^2 \\
&= \|\mathbf{v}_{t-1}\|^2 \left(1 + \frac{2\eta_t y_t^2}{\|\mathbf{v}_{t-1}\|^2} + \eta_t^2 \frac{\|\mathbf{X}_t \mathbf{X}_t^\top \mathbf{v}_{t-1}\|^2}{\|\mathbf{v}_{t-1}\|^2}\right) \\
&\leq \|\mathbf{v}_{t-1}\|^2 \cdot \left(1 + \frac{2\eta_t y_t^2}{\|\mathbf{v}_{t-1}\|^2} + \eta_t^2 B^4\right).
\end{aligned}$$

Write

$$\begin{aligned}
S_t &= 1 - \frac{v_{t,1}^2}{\|\mathbf{v}_t\|^2} \\
&\leq 1 - \left(1 + \frac{2\eta_t y_t^2}{\|\mathbf{v}_{t-1}\|^2} + \eta_t^2 B^4\right)^{-1} \cdot \frac{v_{t-1,1}^2 + 2\eta_t y_t v_{t-1,1} X_{t,1}}{\|\mathbf{v}_{t-1}\|^2} \\
&\leq 1 - \left(1 - \frac{2\eta_t y_t^2}{\|\mathbf{v}_{t-1}\|^2} - \eta_t^2 B^4\right) \cdot \frac{v_{t-1,1}^2 + 2\eta_t y_t v_{t-1,1} X_{t,1}}{\|\mathbf{v}_{t-1}\|^2} \\
&= 1 - \frac{v_{t-1,1}^2 + 2\eta_t y_t v_{t-1,1} X_{t,1}}{\|\mathbf{v}_{t-1}\|^2} + 2 \frac{\eta_t y_t^2 v_{t-1,1}^2}{\|\mathbf{v}_{t-1}\|^4} \\
&\quad + 4\eta_t^2 \cdot \frac{y_t^4}{\|\mathbf{v}_{t-1}\|^4} + B^4 \eta_t^2 \frac{v_{t-1,1}^2 + 2\eta_t y_t^2}{\|\mathbf{v}_{t-1}\|^2} \\
&\leq S_{t-1} - 2\eta_t \frac{v_{t-1,1}}{\|\mathbf{v}_{t-1}\|^2} \cdot \underbrace{y_t \left(X_{t,1} - \frac{v_{t-1,1}}{\|\mathbf{v}_{t-1}\|^2}\right)}_{z_{t,1}} + \eta_t^2 (4B^4 + B^4(1 + 2B^2)).
\end{aligned}$$

The middle term in the display above is exactly the same as in Q_t for (Kra-Oja). One can use the same bound to proceed. The proof is completed. \square

Proof of Proposition 2

In the proof below, for simplicity, we use X_t to denote L_t^{SGD} . Since $\mathbf{x}_* \in \mathcal{X}$, it is a fixed point of $\Pi_{\mathcal{X}}$. Thus, we can write

$$\begin{aligned} X_t &= \|\mathbf{x}_t - \mathbf{x}_*\|^2 \\ &= \|\Pi_{\mathcal{X}}(\mathbf{x}_{t-1} - \eta_t g(\mathbf{x}_{t-1}, \xi_t)) - \Pi_{\mathcal{X}}(\mathbf{x}_*)\|^2 \\ &\leq \|\mathbf{x}_{t-1} - \eta_t g(\mathbf{x}_{t-1}, \xi_t) - \mathbf{x}_*\|^2 \\ &= \|\mathbf{x}_{t-1} - \mathbf{x}_*\|^2 - 2\eta_t \langle g(\mathbf{x}_{t-1}, \xi_t), \mathbf{x}_{t-1} - \mathbf{x}_* \rangle \\ &\quad + \eta_t^2 \|g(\mathbf{x}_{t-1}, \xi_t)\|^2 \\ &\leq X_{t-1} - 2\eta_t \langle g(\mathbf{x}_{t-1}, \xi_t), \mathbf{x}_{t-1} - \mathbf{x}_* \rangle + B^2 \eta_t^2. \end{aligned}$$

For the middle term in the display above, write

$$\begin{aligned} &2\eta_t \langle g(\mathbf{x}_{t-1}, \xi_t), \mathbf{x}_{t-1} - \mathbf{x}_* \rangle \\ &= 2\eta_t \langle \nabla F(\mathbf{x}_{t-1}), \mathbf{x}_{t-1} - \mathbf{x}_* \rangle + 2\eta_t \underbrace{\langle g(\mathbf{x}_{t-1}, \xi_t) - \nabla F(\mathbf{x}_{t-1}), \mathbf{x}_{t-1} - \mathbf{x}_* \rangle}_{Y_t} \\ &\geq 2\eta_t \left(F(\mathbf{x}_{t-1}) - F(\mathbf{x}_*) + \frac{\lambda}{2} \|\mathbf{x}_{t-1} - \mathbf{x}_*\|^2 \right) + 2\eta_t Y_t \\ &\geq 2\lambda\eta_t \|\mathbf{x}_{t-1} - \mathbf{x}_*\|^2 + 2\eta_t Y_t \end{aligned}$$

where we have used (λ -SC) in the last line. Thus,

$$X_t \leq (1 - 2\lambda\eta_t) X_{t-1} + 2\eta_t Y_t + B^2 \eta_t^2$$

which gives (22). To finish the proof, note that

$$\begin{aligned} \mathbb{E}(Y_t | \mathcal{F}_{t-1}) &= 0, \\ |Y_t| &\leq \|g - \nabla F\|_{\infty} \cdot \|\mathbf{x}_{t-1} - \mathbf{x}_*\| \\ &\leq B \sqrt{X_{t-1}}. \end{aligned}$$

The proof is completed. \square

Proof of Proposition 3

By using the (μ -smooth) and (P-L), we have

$$\begin{aligned} F(\mathbf{x}_t) - F(\mathbf{x}^*) &\leq F(\mathbf{x}_{t-1}) - F(\mathbf{x}^*) - \eta_t \langle \nabla F(\mathbf{x}_{t-1}), g(\mathbf{x}_{t-1}, \xi_t) \rangle + \frac{\mu\eta_t^2}{2} \|g(\mathbf{x}_{t-1}, \xi_t)\|^2 \\ &\leq F(\mathbf{x}_{t-1}) - F(\mathbf{x}^*) - \eta_t \|\nabla F(\mathbf{x}_{t-1})\|^2 + \eta_t \langle \nabla F(\mathbf{x}_{t-1}), \nabla F(\mathbf{x}_{t-1}) - g(\mathbf{x}_{t-1}, \xi_t) \rangle \\ &\quad + 2\mu B^2 \eta_t^2 \\ &\leq F(\mathbf{x}_{t-1}) - F(\mathbf{x}^*) - \tau\eta_t (F(\mathbf{x}_{t-1}) - F(\mathbf{x}^*)) + \eta_t Y_t + 2\mu B^2 \eta_t^2 \\ &\leq (1 - \tau\eta_t) (F(\mathbf{x}_{t-1}) - F(\mathbf{x}^*)) + \eta_t Y_t + 2\mu B^2 \eta_t^2, \end{aligned}$$

where

$$Y_t := \langle \nabla F(\mathbf{x}_{t-1}), \nabla F(\mathbf{x}_{t-1}) - g(\mathbf{x}_{t-1}, \xi_t) \rangle.$$

It is easy to check that

$$\begin{aligned} \mathbb{E}(Y_t | \mathcal{F}_{t-1}) &= 0; \\ |Y_t| &\leq B \|\nabla F(\mathbf{x}_{t-1})\| \\ &\leq B \sqrt{\mu} \cdot \sqrt{F(\mathbf{x}_{t-1}) - F(\mathbf{x}^*)} \end{aligned}$$

where the last line follows from the smoothness of F . The proof is completed. \square

Proof of Proposition 5

In the proof below, we use S_t to denotes L_t^{RM} for simplicity. Write

$$\begin{aligned} S_t &= (X_t - X_{t-1} + X_{t-1} - \theta)^2 \\ &= (X_{t-1} - \theta)^2 + 2(X_{t-1} - \theta)(X_t - X_{t-1}) + (X_t - X_{t-1})^2 \\ &= S_{t-1} - 2\eta_t(X_{t-1} - \theta)Y(X_{t-1}) + \eta_t^2 Y(X_{t-1})^2 \\ &= S_{t-1} - 2\eta_t(X_{t-1} - \theta)M(X_{t-1}) - 2\eta_t(X_{t-1} - \theta)(Y(X_{t-1}) - M(X_{t-1})) \\ &\quad + \eta_t^2 Y(X_{t-1})^2 \\ &= S_{t-1} - 2\eta_t \cdot S_{t-1} \cdot \frac{M(X_{t-1}) - M(\theta)}{X_{t-1} - \theta} - 2\eta_t(X_{t-1} - \theta)\xi_{X_{t-1}} + \eta_t^2 Y(X_{t-1})^2. \end{aligned}$$

Note that

$$\frac{M(X_{t-1}) - M(\theta)}{X_{t-1} - \theta} = M'((1-u)X_{t-1} + u\theta) \geq R$$

almost surely, where $u \in (0, 1)$ is random and depends on X_{t-1} .

Thus,

$$S_t \leq (1 - 2R\eta_t)S_{t-1} - 2\eta_t(X_{t-1} - \theta)\xi_{X_{t-1}} + \eta_t^2 Y(X_{t-1})^2.$$

Moreover,

$$\begin{aligned} Y(X_{t-1})^2 &= (M(X_{t-1}) + \xi_{X_{t-1}})^2 \\ &\leq 2M(X_{t-1})^2 + 2R_1^2 \\ &\leq 2P^2(\sqrt{S_{t-1}}) + 2R_1^2 \end{aligned}$$

where R_1 is given in (36). The proof is completed. \square

Lemma 9. Let $\delta \in (0, e^{-2})$. Then, the function

$$f(x) := \frac{\log \log(x) + \log(\delta^{-1})}{x}$$

is decreasing on $[e, \infty]$.

Proof of Lemma 9. It is easy to check that the derivative of f is

$$f'(x) := \frac{(\log x)^{-1} - \log \log x - \log(\delta^{-1})}{x^2}.$$

On $[e, \infty)$, we have $1/\log(x) < 2 = \log(e^2) \leq \log(\delta^{-1})$ and $\log \log(x) \geq 0$. Thus, f is decreasing. \square

Lemma 10. Let L be an integer greater than 2 and suppose $\alpha = \max\{L-1, 33\}$. Defined by $t_k = \alpha^{k+1} - \alpha$. Then,

$$31.5 \times \frac{t_{k-1} + L - 1}{(\alpha^2 - 1)(t_{k-1} + \alpha^2)} \leq 1$$

for all $k \geq 1$.

Proof of Lemma 10. we need to show that

$$\begin{aligned} 31.5 \times (t_{k-1} + L - 1) &\leq (\alpha - 1)(t_{k-1} + \alpha) \\ \iff t_{k-1}(32.5 - \alpha) &\leq \alpha(\alpha - 1) - 31.5 \times (L - 1). \end{aligned}$$

The statement above is true for $k = 1$ since $\alpha \geq \max\{L-1; 32\}$ and $t_0 = 0$.

For $k \geq 2$, note that the left-hand side in the last display is negative (since $\alpha \geq 33$) while the right-hand side is positive. The proof is completed. \square

Lemma 11. Let $\Omega = \{(x, y, z) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ : x + y + z = 1\}$. Then, we have

$$\min_{\substack{x,y,z>0 \\ x+y+z=1}} \left\{ \max \left\{ \frac{1}{x}; \frac{8}{y}; \frac{16}{z^2} \right\} \right\} \leq 31.5.$$

Proof of Lemma 11. It is easy to see that

$$\min_{\substack{x,y,z>0 \\ x+y+z=1}} \left\{ \max \left\{ \frac{1}{x}; \frac{8}{y}; \frac{16}{z^2} \right\} \right\} \leq A$$

where A satisfies

$$\begin{cases} \frac{1}{x} = \frac{8}{y} = \frac{16}{z^2} = A; \\ \frac{1}{A} + \frac{8}{A} + \frac{4}{\sqrt{A}} = 1. \end{cases}$$

The second equality is a quadratic equation in A , and is equivalent to

$$A - 4\sqrt{A} - 9 = 0.$$

Solving this equation gives

$$A = (2 + \sqrt{13})^2 < 31.5.$$

□

C Bibliography

- [1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.
- [2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.
- [3] Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–492. IEEE, 2017.
- [4] Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309. PMLR, 2016.
- [5] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental pca. *Advances in neural information processing systems*, 26, 2013.
- [6] Vivek S Borkar and Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.
- [7] Selina Carter and Arun K Kuchibhotla. Statistical inference for online algorithms. *arXiv preprint arXiv:2505.17300*, 2025.
- [8] Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

- [9] Louis HY Chen, Xiao Fang, and Qi-Man Shao. From stein’s identities to moderate deviations. *The Annals of Probability*, pages 262–293, 2013.
- [10] 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.
- [11] Chi-Ning Chou and Mien Brabecba Wang. Ode-inspired analysis for the biological version of oja’s rule in solving streaming pca. In *Conference on Learning Theory*, pages 1339–1343. PMLR, 2020.
- [12] Sanjoy Dasgupta, Syamantak Kumar, Shourya Pandey, and Purnamrita Sarkar. Low precision streaming PCA. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [13] Victor H de la Pena and Michael J Klass2and Tze Leung Lai. Pseudo-maximization and self-normalized processes. *Probability Surveys*, 4:172–192, 2007.
- [14] Victor H De la Pena, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.
- [15] Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *International conference on machine learning*, pages 2332–2341. PMLR, 2015.
- [16] John C Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.
- [17] Yasong Feng, Yifan Jiang, Tianyu Wang, and Zhiliang Ying. The anytime convergence of stochastic gradient descent with momentum: From a continuous-time perspective. *arXiv preprint arXiv:2310.19598*, 2023.
- [18] EG Gladyshev. On stochastic approximation. *Theory of Probability & Its Applications*, 10(2):275–278, 1965.
- [19] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Advances in neural information processing systems*, 27, 2014.
- [20] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17, 2020.
- [21] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021.
- [22] De Huang, Jonathan Niles-Weed, Joel A. Tropp, and Rachel Ward. Matrix concentration for products. *Foundations of Computational Mathematics*, 22(6):1767–1799, 2022.
- [23] De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-pca: Efficient guarantees for oja’s algorithm, beyond rank-one updates. In *Conference on Learning Theory*, pages 2463–2498. PMLR, 2021.
- [24] Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja’s algorithm. In *Conference on learning theory*, pages 1147–1164, 2016.

- [25] Rajeeva Laxman Karandikar and Mathukumalli Vidyasagar. Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. *Journal of Optimization Theory and Applications*, pages 1–39, 2024.
- [26] Syamantak Kumar, Shourya Pandey, and Purnamrita Sarkar. Beyond sin-squared error: linear time entrywise uncertainty quantification for streaming pca. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence, UAI ’25*. JMLR.org, 2025.
- [27] Syamantak Kumar and Purnamrita Sarkar. Streaming pca for markovian data. *Advances in Neural Information Processing Systems*, 36:64650–64662, 2023.
- [28] Syamantak Kumar and Purnamrita Sarkar. Oja’s algorithm for streaming sparse pca. *arXiv preprint arXiv:2402.07240*, 2024.
- [29] Harold Kushner. Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):87–96, 2010.
- [30] Harold J Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.
- [31] Tze Leung Lai. On confidence sequences. *The Annals of Statistics*, 4(2):265–280, 1976.
- [32] Tze Leung Lai. Extended stochastic lyapunov functions and recursive algorithms in linear stochastic systems. In *Stochastic Differential Systems: Proceedings of the 4th Bad Honnef Conference, June, 20–24, 1988*, pages 206–220. Springer, 2006.
- [33] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- [34] Chris Junchi Li, Wenlong Mou, Martin Wainwright, and Michael Jordan. Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. In *Conference on Learning Theory*, pages 909–981. PMLR, 2022.
- [35] Jinlan Liu, Dongpo Xu, Yinghua Lu, Jun Kong, and Danilo P Mandic. Last-iterate convergence analysis of stochastic momentum methods for neural networks. *Neurocomputing*, 527:27–35, 2023.
- [36] Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory*, pages 2963–2983. PMLR, 2022.
- [37] Robert Lunde, Purnamrita Sarkar, and Rachel Ward. Bootstrapping the error of oja’s algorithm. *Advances in neural information processing systems*, 34:6240–6252, 2021.
- [38] Liam Madden, Emiliano Dall’Anese, and Stephen Becker. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36, 2024.
- [39] Tudor Manole and Aaditya Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. 69(7):4641–4658, 2023.
- [40] Diego Martinez-Taboada, Tomas Gonzalez, and Aaditya Ramdas. Vector-valued self-normalized concentration inequalities beyond sub-gaussianity. *arXiv preprint arXiv:2511.03606*, 2025.
- [41] Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. *Advances in neural information processing systems*, 26, 2013.

- [42] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.
- [43] Morenikeji Neri and Thomas Powell. A quantitative robbins-siegmund theorem. *The Annals of Applied Probability, to appear*, 2024.
- [44] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15:267–273, 1982.
- [45] Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.
- [46] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.
- [47] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.
- [48] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [49] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.
- [50] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971. PMLR, 2021.
- [51] Ohad Shamir. Convergence of stochastic gradient descent for pca. In *International Conference on Machine Learning*, pages 257–265. PMLR, 2016.
- [52] Ohad Shamir. Fast stochastic algorithms for svd and pca: Convergence properties and convexity. In *International Conference on Machine Learning*, pages 248–256. PMLR, 2016.
- [53] R. Vershynin. *High-dimensional probability: an introduction with applications in data science*. Cambridge University Press., 2018.
- [54] Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H Kennedy, and Aaditya Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics*, 52(6):2613–2640, 2024.
- [55] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.
- [56] Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. Time-uniform self-normalized concentration for vector-valued processes. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291, pages 5714–5715, 2025.