A General Framework for Analyzing Stochastic Dynamics in Learning Algorithms

Chi-Ning Chou* Mien Brabeeba Wang† Tiancheng Yu‡

June 12, 2020

Abstract

We present a general framework for analyzing high-probability bounds for stochastic dynamics in learning algorithms. Our framework composes standard techniques such as a stopping time, a martingale concentration and a closed-from solution to give a streamlined three-step recipe with a general and flexible principle to implement it. To demonstrate the power and the flexibility of our framework, we apply the framework on three very different learning problems: stochastic gradient descent for strongly convex functions, streaming principal component analysis and linear bandit with stochastic gradient descent updates. We improve the state of the art bounds on all three dynamics.

1 Introduction

Iterative methods are widely used in machine learning and stochastic optimization where they naturally induce *stochastic processes* of objective functions. For example, when an optimization algorithm uses stochastic gradient descent (SGD) updates, the loss function forms a stochastic process. Therefore, to study the performance of a learning algorithm, it usually suffices to understand the *behavior* of the corresponding stochastic process.

There have been many successes in providing theoretical guarantees for various learning algorithms. However, as the learning algorithms nowadays become increasingly complicated, it is more and more challenging to perform clean and tight theoretical analysis. Moreover, due to the lack of general principles for analysis, the existing theoretical studies are usually tailored to specific learning dynamics and hence are unlikely to extend to other problems.

Specifically, a main difficulty of the analysis lies in the non-linearity from an iterative method which convolutes the underlying drifting process with the noise from the stochasticity.[1] This creates a circular entanglement: the improvement of the drifting term requires bounds for the noise term while the latter also depends on the historical bounds of the former. Therefore, to enable a tight analysis, it is crucial to have good control of the local moment information about the process at each time step to bound the noise. However, previous analysis usually either only looks at the expectation and loses important local information (*e.g.* [CLSH19, QDC19, YWH19, ZSJ+19]) or directly applies concentration analysis to bound the past process step-by-step which can cause overcomplication (*e.g.,* [AZL17, AZLS19, RR19, ZCZG18]).

We propose an analysis framework with an attempt to address these challenges. In response to the lack of a general principle, the framework provides a *streamlined recipe* to implement the analysis. Moreover, the

*School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Supported by NSF awards CCF 1565264 and CNS 1618026. Email: chiningchou@g.harvard.edu.

†MIT CSAIL, Cambridge, Massachusetts, USA. Supported by NSF Awards CCF-1810758, CCF-0939370, CCF-1461559 and Akamai Presidential Fellowship. Email: brabeeba@mit.edu.

‡MIT LIDS, Cambridge, Massachusetts, USA. Supported by NSF BIGDATA grant 1741341. Email: yutc@mit.edu.

[1] For example, consider $x_t = x_{t-1}^2 + n_t$ where n_t is the noise. Although n_t does not explicitly convolute with x_t, after unfolding the recursion, we observe $x_t = x_{t-2}^4 + 2n_{t-1}x_{t-2}^2 + \cdots$ and $n_{t-1}x_{t-2}^2$ is entangled.

framework is *flexible* to analyze distinct dynamics as well as to provide rooms to accommodate specialized techniques. To disassemble the tangled dynamic, the recipe provides a three-step solution. In step one, the framework uses a closed-form solution from a linearization and a stopping time technique to tightly track the local moment information of the noise term without dealing with the entanglement from the drifting process. In step two, the framework disentangles the dynamic in a local interval by simultaneously bounding the noise and showing the local improvement of the stochastic process. Specifically, we use a martingale concentration on the stopped noise and summarize the improvement analysis into two easily verifiable conditions. In step three, the framework uses a flexible interval analysis to connect the local improvement back to the global improvement. We leave the details in Section 2.

Examples of implementing the framework. To show the power and the flexibility of our framework, we consider three different learning dynamics with increasing difficulties. We improve the state-of-the-art bound and match the information-theoretic lower bound for all three problems up to logarithmic factors. In the following, ϵ and δ stands for the error and the failure probability.

We start with a well-known textbook learning algorithm: *stochastic gradient descent (SGD)* algorithm for strongly convex functions. We first give an expository proof that matches the state-of-the-art $O(\epsilon^{-1}(\log \delta^{-1} + \log\log \epsilon^{-1}))$ convergence rate [HK14, RSS12]. To show the flexibility of the framework, we use a more optimized interval analysis to achieve $O(\epsilon^{-1}(\log \delta^{-1} + \log\log\log \epsilon^{-1}))$ convergence rate. See Section 3 for details.

Next, we move on to a classic dynamic with a non-convex structure: the *streaming principle component analysis (PCA)* problem. We provide a simple proof for the local convergence setting and get $O(\epsilon^{-1}(\log \delta^{-1} + \log\log \epsilon^{-1}))$ convergence rate while the previous state-of-the-art analysis [AZL17] gets $O(\epsilon^{-1}(\log \delta^{-1} + \log^3 \epsilon^{-1}))$. See Section 4 for details.[2]

Finally, we consider a problem with active dynamic where the updates are adaptive and dependant on the whole history: solving stochastic linear bandit with SGD update [JBNW17, KPM15]. This problem is not only useful for designing scalable bandit algorithms but also serves as an important intermediate step towards analyzing the Q-learning algorithm in linear parameterized Markov decision processes (MDPs) [JAZBJ18, JYWJ19]. We analyze an algorithm where previous technique cannot analyze and improve the state-of-the-art regret from $O(n(T\log^2(T)\log(T/\delta))^{1/2})$ in [JBNW17] to $O(n(T\log^2(T)\log(1/\delta))^{1/2})$. See Section 5 for details.

Related work. We focus on analyzing stochastic processes in learning algorithms which broadly appear in theoretical machine learning [Moi18, Set09, SB18], optimization theory [BBV04, Haz19, SSBD14], statistical learning theory [HTF09], etc. There have been many beautiful results providing theoretical analysis on a wide range of important learning problems, *e.g.,* principal component analysis [AZL17], non-negative matrix factorization [AGKM16, LS99, Vav10], topic models [AGH$^+$13, AGM12], matrix completion [Har14, JNS13], tensor decomposition [AGH$^+$14, GHJY15], neural networks [AZLS18, DZPS18, JGH18], etc. However, the lack of a unifying framework often makes the progress in analyzing frontier learning dynamics slow and sub-optimal. This paper attempts to propose a general framework for the future studies in new and complex learning dynamics.

The technical ingredients in the framework are standard, simple, and inspired by both the stochastic approximation theory [KY97] and a recent analysis for streaming PCA [CW19]. For example, the stopping time technique or the martingale concentration have been widely applied in many other analysis [AZL17, RSS12]. We emphasize that it is the *composition* of these tools that makes our framework powerful and flexible, and the main contribution of this paper is to propose a general and streamlined recipe that future analysis of new and complex learning algorithms can easily adopt.

[2] There are other standard guarantees such as global convergence, gap-free convergence, exponential convergence etc. [AZL17] obtains $\log^5$ in their bound in the global gap-free setting. Using their analysis techniques on the local convergence would get $\log^3$ in the bound.

2 General Framework

In the theoretical analysis of a learning algorithm, one usually identifies an *objective function* to evaluate how well the algorithm performs. When using an iterative method, the dynamic of the objective function can be characterized by a stochastic process. We use $\{X_t\}$ to denote the stochastic process of interest and the goal[3] is to find a function $T(\epsilon, \delta)$ such that for every $\epsilon, \delta > 0$, we have

$$\Pr[X_{T(\epsilon,\delta)} > \epsilon] < \delta \,.$$

For example, we would hope to prove $T(\epsilon, \delta) = \tilde{O}(\epsilon^{-1} \log \delta^{-1})$ for SGD and PCA.[4]

Prologue: Continuous analysis. It is often not obvious on how to analyze a discrete stochastic process directly. A general principle inspired by stochastic approximation theory [KY97] is to first understand the behaviors of the continuous analog, which is the limiting process by taking the learning rate to 0. The guidance from the continuous dynamic can usually be very insightful and point to a good way to analyze the discrete stochastic process. See Appendix B for more discussion.

2.1 Step 1: Linearization and moment analysis

Guided by the continuous analysis, we investigate the local behavior of $\{X_t\}$ by approximating it with a well-studied dynamic. Specifically, in this paper we focus on *linearizing* $\{X_t\}$ as follows.

$$X_t \le H_t \cdot X_{t-1} + N_t \,, \ \forall t \in \mathbb{N}$$

where $H_t > 0$ is a multiplicative factor and N_t is a minor term depending on both X_{t-1} and the stochasticity at the t^{th} step. Next, by unfolding the recursion[5], we have

$$X_t \le D_t \cdot (X_{T_0} + M_t) \ , \ \forall\, 0 \le T_0 < t \in \mathbb{N} \tag{2.1}$$

where $D_t = \prod_{t'=T_0+1}^{t} H_{t'}$ is the *drifting factor* and $M_t = \sum_{t'=T_0+1}^{t} D_{t'}^{-1} \cdot N_{t'}$. Note that $\{M_t\}$ is adaptive: it does not depend on the future. Intuitively, $D_t \cdot X_{T_0}$ is the *drifting term* that dominates the dynamic and $D_t \cdot M_t$ is the *minor term* coming from the linearization and discretization.

Observe that if M_t is small compared to X_{T_0}, then the drifting term will govern the dynamic. To achieve tighter analysis, we use a *stopping time* technique to keep track of the *local information on where* $\{X_t\}$ *is*. Hence, using the moment information of the stopped process of $\{M_t\}$, we are able to apply martingale concentration inequality and show that the (stopped) minor term is dominated by the drifting term. We call the collection of such moment bounds a *moment profile* for $\{X_t\}$.

Definition 2.2 (Moment profile)**.** *Let $\{X_t\}$ be a stochastic process described in Equation 2.1 for some $\{M_t\}$. Let $\Lambda > 0$, τ be the stopping time for the event $\{X_t \ge \Lambda\}$, $\{M_{t\wedge\tau}\}$ be the stopped process of $\{M_t\}$, and let $\{\mathcal{F}_t\}$ be a filtration of $\{X_t\}$.[6] The functions (B, μ, σ^2) form a moment profile for $\{X_t\}$ and Λ if the following hold for every $t \ge 1$.*

- *(Bounded difference)* $|M_{t\wedge\tau} - M_{(t-1)\wedge\tau}| \le B(t, \Lambda)$ *almost surely.*
- *(Conditional expectation)* $\mathbb{E}[M_{t\wedge\tau} - M_{(t-1)\wedge\tau} \mid \mathcal{F}_{t-1}] \le \mu(t, \Lambda)$.
- *(Conditional variance)* $\mathsf{Var}[M_{t\wedge\tau} - M_{(t-1)\wedge\tau} \mid \mathcal{F}_{t-1}] \le \sigma^2(t, \Lambda)$.

If we started from $T_0 \in \mathbb{N}$ instead, then the moment profile is denoted as $(B_{T_0}, \mu_{T_0}, \sigma^2_{T_0})$.

[3] See Definition A.8 for other common notions of high-probability convergence this framework can achieve.

[4] Another way to present convergence result is to find a function $\epsilon(T, \delta)$ such that $\Pr[X_T > \epsilon(T, \delta)] < \delta$ for all $T \in \mathbb{N}$ and $\delta > 0$. For example, we would hope to prove $\epsilon(T, \delta) = \tilde{O}((T \log \delta^{-1})^{1/2})$ for the linear bandit problem. For simplicity, in Section 2 we focus on the guarantee of finding $T(\epsilon, \delta)$.

[5] This is also known as the *ODE trick* in [CW19].

[6] See Appendix A for background in martingale and stopping time.

Treating Λ as a free parameter *isolates* the moment calculation from the potentially complicated global dynamic. Also, the choice of stopping time τ could be more general and provides additional room for specialized techniques. See Appendix A for backgrounds in the related math tools. Finally, note that the moment profile is for the stopped process of the minor term. In step two, we will *pull-out* the stopping time to recover the concentration on the original minor term.

User manual. It is convenient to think of the stopped martingale difference $M_{t\wedge\tau} - M_{(t-1)\wedge\tau}$ simply as $\mathbf{1}_{\{X_{t-1}<\Lambda\}} \cdot (M_t - M_{t-1})$. Namely, to calculate the moment profile, one express the three moment quantities of $M_t - M_{t-1}$ as a function of Λ conditioning on the event $\{X_{t-1} < \Lambda\}$.

2.2 Step 2: Improvement analysis

In this step, given $A_0, A_1 > 0$, the goal is to find $T_1 \in \mathbb{N}$ as small as possible such that given $X_0 \le A_0$, we have $X_{T_1} \le A_1$ with high probability.[7] We call this an improvement analysis. To show the improvement of $\{X_t\}$, according to Equation 2.1 it suffices to show that the minor term M_t is small. However, directly invoking martingale concentration inequality with a moment profile only ensures the *stopped process* of the minor term being small. Nevertheless, we can extend to the original minor term if the parameters satisfied certain *pull-out conditions* [CW19]. Meanwhile, we also need to make sure the improvement from drifting term D_t is as expected. The following proposition crystallizes these intuitions into inequalities that can be easily verified.

Proposition 2.3 (Improvement analysis)**.** *Let $\{X_t\}$ be a stochastic process described in Equation 2.1 for some $\{D_t\}$ and $\{M_t\}$. Let $\Lambda > 0$, τ be the stopping time for the event $\{X_t > \Lambda\}$, and (B, μ, σ^2) be a moment profile for $\{X_t\}$ and Λ. For every $T_1 \in \mathbb{N}$ and $\delta' > 0$, let*

$$\Delta = \Delta\left(B, \mu, \sigma^2, \Lambda, T_1, \delta'\right)$$

be the deviation from a concentration inequality[8] such that $\Pr\left[\max_{1\le t\le T_1} M_{t\wedge\tau} > \Delta\right] < \delta'$.
If we have

- *(Improvement condition) $D_{T_1} \cdot (A_0 + \Delta) \le A_1$ and*
- *(Pull-out condition) $D_t \cdot (A_0 + \Delta) \le \Lambda$ for every $1 \le t \le T_1$.*

Then we have

$$\Pr\left[\exists t \in [T_1],\ X_t > D_t \cdot (A_0 + \Delta) \mid X_0 \le A_0\right] < \delta'.$$

In particular, the above implies $\Pr[X_{T_1} > A_1 \mid X_0 \le A_0] < \delta'$. Also, the proposition can naturally extend to starting from $T_0 \in \mathbb{N}$ instead of 0.

User manual. Given a moment profile, use your favorite martingale concentration inequality and calculate Δ as a function of Λ. We will instantiate the improvement analysis with various settings of the free parameters $T_0, T_1, A_0, A_1, \Lambda$ and verify the two inequalities in the next step.

2.3 Step 3: Interval analysis

Directly applying the improvement analysis usually would not give a tight convergence rate because the magnitude of the moment quantities could vary a lot at different stages of a stochastic process. Thus, one has to utilize the *local information on where the process is* to achieve tighter analysis.

To approach optimal convergence rate, we systematically execute an *interval analysis* by tracking the stochastic process $\{X_t\}$ locally. We design a sequence $a_0, a_1, \ldots, a_\ell > 0$ such that $a_\ell = \epsilon$ and aim to design a

[7] Note that in the linear bandit problem, the goal would be to find $T_1 > T_0$ as large as possible.

[8] For example, $\Delta = \sqrt{2\sum_{t'=T_0+1}^{T_1} B_{T_0}^2(t', \Lambda)\log(1/\delta')} + \sum_{t'=T_0+1}^{T_1} \mu_{T_0}(t', \Lambda)$ if we used Azuma's inequality. See Appendix B and [CL06] for more examples.

sequence $0 = t_0 < t_1 < \cdots < t_\ell$ such that for each i if the process started from $X_{t_{i-1}} \le a_{i-1}$, then $X_{t_i} \le a_i$ with high probability according to Proposition 2.3. Intuitively, the value of X_t, as well as the three moment quantities, are expected not to change by too much within an interval $[t_{i-1}, t_i]$. Thus, one can hope to get a nearly tight characterization of the dynamic through the improvement analysis.

In general, the above interval analysis can be very *flexible* because it is easy to try different speeds of convergence, different settings of the learning rate, etc., while the cumbersome moment calculation and concentration analysis had been *isolated* in the previous step. Also, there is a *principle way* to design sequences that satisfy both the improvement condition and the pull-out condition as elaborated in Appendix B. Since the final analysis differs across different problems, it would be more illuminating to see concrete examples in the following sections.

User manual. The above high-level idea can be implemented by designing length ℓ (or $\ell+1$) sequences $\{t_i\}$, $\{a_i\}$, $\{\Lambda_i\}$, $\{\delta_i\}$ with $a_\ell = \epsilon$ and invoke Proposition 2.3 with $(T_0 = t_{i-1}, T_1 = t_i, A_0 = a_{i-1}, A_1 = a_i, \delta' = \delta_i)$ for each $i = 1, 2, \ldots, \ell$. As a consequence, this amounts to checking the two conditions in Proposition 2.3 by verifying these sequences satisfying certain inequalities and then we have $\Pr[X_{t_i} > a_i \,|\, X_{t_{i-1}} \le a_{i-1}] < \delta_i$. As for the boundary case, we pick a_0 properly such that $\Pr[X_{t_0} > a_0] < \delta_0$. Note that by the chain rule for conditional expectation and let $T(\epsilon, \delta) = t_\ell$, we achieve the goal $\Pr[X_{T(\epsilon,\delta)} > \epsilon] < \sum_i \delta_i = \delta$.

3 SGD for Strongly Convex Functions

We pick stochastic gradient descent (SGD) for strongly convex functions as the first example because it is one of the simplest and most well-known methods in learning problems. To let the reader have more familiarity with the framework, we provide an expository proof that matches the state-of-the-art analysis on $1/t$ learning rate. To demonstrate the flexibility of the framework, we further improve the bound via a more optimized interval analysis.

We briefly set up the problem and defer the details to Appendix C. The goal is to minimize a λ-strongly convex function F with a G-bounded gradient oracle using the following update rule

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \hat{\mathbf{g}}_t$$

where η_t is the learning rate, $\hat{\mathbf{g}}_t$ is the response from the gradient oracle with $\mathbb{E}[\hat{\mathbf{g}}_t]$ being the gradient of $\mathbf{w}_{t-1}$. Also, $\|\hat{\mathbf{g}}_t\| \le G$ almost surely. Since F is λ-strongly convex, there exists a unique $\mathbf{w}^*$ that attains the minimum value of F. The potential function is naturally defined as

$$X_t := \|\mathbf{w}_t - \mathbf{w}^*\|^2 . \tag{3.1}$$

Previously, the best high-probability bound is $O(G^2(\log(1/\delta) + \log\log(G/\epsilon\lambda))\epsilon^{-1}\lambda^{-2})$ where ϵ is the error and δ is the failure probability [HK14, RSS12]. We apply our framework and prove the following theorem that first matches previous bounds using learning rate $\eta_t = 1/(\lambda t)$ and then improve the bound to $O(G^2(\log(1/\delta)+\log\log\log(G/\epsilon\lambda))\epsilon^{-1}\lambda^{-2})$ using a different learning rate. Note that this partially answers the open question of [HK14] on whether the high-probability bound can be improved.

Theorem 3.2 (Convergence of SGD for strongly convex functions)**.** *Consider the above setting, for every $\epsilon, \delta > 0$, we have $\|\mathbf{w}_T - \mathbf{w}^*\|^2 \le \epsilon$ with probability at least $1 - \delta$*

- *if $\eta_t = 1/(\lambda t)$ and for some $T = O(\frac{G^2(\log(1/\delta)+\log\log(G/\epsilon\lambda))}{\epsilon\lambda^2})$;*
- *if η_t is chosen properly and for some $T = O(\frac{G^2(\log(1/\delta)+\log\log\log(G/\epsilon\lambda))}{\epsilon\lambda^2})$.*

Remark. *We implement a more optimized interval analysis to prove the second item and actually provides a stronger convergence guarantee for the first item. See Appendix C for details.*

3.1 Step 1: Linearization and moment analysis

For every $t \in \mathbb{N}$, let us first linearize $X_t := \|\mathbf{w}_t - \mathbf{w}^*\|^2$ in the standard way as follows.

$$X_t \le (1 - 2\eta_t\lambda)X_{t-1} + N_t$$

where the explicit formula of N_t and the derivation of the linearization can be found in Appendix C. Next, the following lemma provides a moment profile for $\{X_t\}$.

Lemma 3.3 (Moment profile for SGD)**.** *Consider the setting in Theorem 3.2 with learning rate $\eta_t = 1/(\lambda t)$. For every $100 < T_0 + 1 \le t$ and $\Lambda > 0$, the following functions $(B_{T_0}, \mu_{T_0}, \sigma^2_{T_0})$ form a moment profile for $\{X_t\}$, Λ, and T_0.*

$$B_{T_0}(t,\Lambda) = \frac{13G^2 t}{\lambda^2 T_0^2}, \mu_{T_0}(t,\Lambda) = \frac{2G^2}{\lambda^2 T_0^2}, \text{ and } \sigma^2_{T_0}(t,\Lambda) = \frac{G^2t^2}{\lambda^2 T_0^4}\left(73\Lambda + \frac{3G^2}{\lambda^2 t^2}\right).$$

3.2 Step 2: Improvement analysis

We apply Proposition 2.3 using the moment profile of $\{X_t\}$ in Lemma 3.3 and get the following.

Lemma 3.4 (Improvement analysis for SGD)**.** *Let $\{X_t\}$ be the stochastic process described in Equation 3.1, $100 \le T_0 < T_1 \in \mathbb{N}$, $\Lambda > 0$, and $(B_{T_0}, \mu_{T_0}, \sigma^2_{T_0})$ be a moment profile for $\{X_t\}$, T_0, and Λ from Lemma 3.3. Then for every $A_0 > A_1 > 0$ and $\delta' > 0$, let*

$$\Delta = \frac{G^2 T_1 \log\frac{1}{\delta'}}{\lambda^2 T_0^2}\left(\sqrt{\frac{300T_1\Lambda\lambda^2}{G^2\log\frac{1}{\delta'}}} + 60\right).$$

Suppose that we have $A_0 + \Delta \le \Lambda$ and $\frac{T_0^2}{T_1^2}\cdot\Lambda \le A_1$. Then, $\Pr[X_{T_1} > A_1 \mid X_{T_0} \le A_0] < \delta'$.

3.3 Step 3: Interval analysis

In this step, we aim to design two sequences $\{t_i\}$ and $\{a_i\}$ of length $\ell + 1$. At the i^{th} interval, we apply Lemma 3.4 with $T_0 = t_{i-1}$, $T_1 = t_i$, $A_0 = a_{i-1}$, $A_1 = a_i$, and $\delta' = \delta/\ell$. By solving the Δ_i in $a_{i-1} + \Delta_i \le \Lambda_i$ and plugging it into $\frac{t_{i-1}^2}{t_i^2}\cdot\Lambda_i \le a_i$, it suffices to set the interval satisfying

$$a_i \ge \frac{t_{i-1}^2}{t_i^2}a_{i-1} + \sqrt{\frac{300a_{i-1}G^2\log\frac{1}{\delta'}}{\lambda^2 t_i}} + \frac{500G^2\log\frac{1}{\delta'}}{\lambda^2 t_i}.$$

Observe that since we would like to set a_i as small as possible, the third term suggests that we should set $a_i = O(\frac{G^2\log(1/\delta')}{\lambda^2 t_i})$. To be precise, let $\ell = \lceil \log\frac{a_0}{\epsilon}\rceil$, $t_0 = 0, t_1 = 100, a_0 = \frac{5000G^2\log(1/\delta')}{\lambda^2 t_1}$,

$$t_i = 2t_{i-1}, a_{i-1} = \frac{5000G^2\log\frac{1}{\delta'}}{\lambda^2 t_{i-1}}, \text{ and } \Lambda_i = 2a_{i-1},\ \forall i = 2,3,\ldots,\ell.$$

Let $T = t_\ell$. Observed that due to the choice of the parameters we have $a_\ell \le \epsilon$. Now, for each $i = 1,2,\ldots,\ell$, we invoke Lemma 3.4 with $A_0 = a_{i-1}$, $A_1 = a_i$, $T_0 = t_{i-1}$, and $T_1 = t_i$. It is easy to verify $a_{i-1} + \Delta_i \le \Lambda_i$ and $\frac{t_{i-1}^2}{t_i^2}\cdot\Lambda_i \le a_i$. Hence, $\Pr[X_T > \epsilon \,|\, X_{t_1} \le a_1] < \delta$. Also, by the strong convexity, we have $X_{t_1} \le 4G^2/\lambda^2 \le a_1$. Therefore, we have $\Pr[X_T > \epsilon] < \delta$.

4 Local Convergence for Streaming k-PCA

Next, we increase the difficulty of the problem by considering a classic non-convex optimization problem in machine learning: the streaming/online principal component analysis (PCA). Let $\mathcal{D}$ be a distribution over the unit sphere in $\mathbb{R}^n$ and $\Sigma = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$ be its covariance matrix. Given a sequence of i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$ from $\mathcal{D}$, the goal of streaming k-PCA is to output a vector that is close to the top k eigenspace of Σ using $O(nk)$ space where $k \in [n]$. Denote $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ as the eigenvalues of Σ. We analyze the classic Oja's algorithm [Oja82] with the following update rule.

$$W_t = (1 - \eta_t \mathbf{x}_t \mathbf{x}_t^\top) W_{t-1}\,,\ \forall t \geq 1$$

where $W_0 \in \mathbb{R}^{n\times k}$ is the initialization matrix and η_t is the learning rate at time t. To measure how well W_t converges to the top k eigenspace, it is standard to use the following objective function [AZL17].

$$X_t = \|Z^\top W_t (V^\top W_t)^{-1}\|_F^2 \tag{4.1}$$

where V (resp. Z) is an orthogonal basis for the Σ's eigenspace corresponds to eigenvalues $\lambda_1, \ldots, \lambda_k$ (resp. $\lambda_{k+1}, \ldots, \lambda_n$). The goal is to show that X_t converges to 0. See Appendix D for details.

There is a line of works [AZL17, DSOR15, JJK+16, LWLZ18, Sha16] studying the above dynamic. For simplicity, we focus on the *local convergence* setting and without loss of generality assume $X_0 \leq 1$. We use our framework to improve the state-of-the-art convergence rate from $O(\lambda \log^3(\delta^{-1}, \epsilon^{-1}, \lambda^{-1}, \mathsf{gap}^{-1})\epsilon^{-1}\mathsf{gap}^{-2})$ in [AZL17] to $O(\lambda \log(\log(\epsilon^{-1})\delta^{-1})\epsilon^{-1}\mathsf{gap}^{-2})$ where $\lambda = \lambda_1 + \cdots + \lambda_k$ and $\mathsf{gap} = \lambda_k - \lambda_{k+1}$.

Theorem 4.2 (Local convergence for streaming k-PCA)**.** *Consider the above setting, then for every $\epsilon > 0$ and $\delta \in (0,1)$, there exists a setting of learning rate and $T = O(\frac{\lambda \log(\log(1/\epsilon)/\delta)}{\epsilon \mathsf{gap}^2})$ such that*

$$\Pr\left[\|Z^\top W_T (V^\top W_T)^{-1}\|_F^2 > \epsilon\right] < \delta\,.$$

Remark. *There are many other guarantees for streaming PCA such as global convergence [AZL17, Sha16], gap-free convergence [AZL17], and exponential convergence [Tan19]. Since the focus of this paper is introducing the framework, we leave the analysis for these settings as future work.*

4.1 Step 1: Linearization and moment analysis

For every $t \in \mathbb{N}$, let us first linearize $X_t := \|Z^\top W_t (V^\top W_t)^{-1}\|_F^2$ as follows.

$$X_t \leq (1 - 2\eta_t \mathsf{gap}) X_{t-1} + N_t$$

where the explicit formula of N_t and the derivation of the linearization can be found in Appendix D. The following lemma provides a moment profile for $\{X_t\}$.

Lemma 4.3 (Moment profile for k-PCA)**.** *Consider the setting in Theorem 4.2 with learning rate $\eta_t = \eta = \gamma/(2\mathsf{gap})$ for some $\gamma \in (0,1)$. For every $1 \leq T_0 + 1 \leq t \in \mathbb{N}$ and $0 < \Lambda \leq 1$, the following functions $(B_{T_0}, \mu_{T_0}, \sigma_{T_0}^2)$ form a moment profile for $\{X_t\}$, Λ, and T_0.*

$$B_{T_0}(t,\Lambda) = \frac{40\eta}{(1-\gamma)^{t-T_0}}\,, \mu_{T_0}(t,\Lambda) = \frac{56\eta^2\lambda}{(1-\gamma)^{t-T_0}}\ ,\ \text{and}\ \sigma_{T_0}^2(t,\Lambda) = \frac{\eta^2\lambda\left(1136\Lambda + 512\eta^2\right)}{(1-\gamma)^{2(t-T_0)}}\,.$$

4.2 Step 2: Improvement analysis

We apply Proposition 2.3 using the moment profile of $\{X_t\}$ in Lemma 4.3 and get the following.

Lemma 4.4 (Improvement analysis for k-PCA)**.** *Let $\{X_t\}$ be the stochastic process described in Equation D.1. For every $0 < A_1 < A_0$, $0 < \delta', \Lambda \leq 1$, and $1 \leq T_0 < T_1 \in \mathbb{N}$, let*

$$\Delta = \frac{\gamma\lambda \log\frac{1}{\delta'}}{\mathsf{gap}^2(1-\gamma)^{T_1-T_0}}\left(\sqrt{\frac{568\mathsf{gap}^2\Lambda}{\gamma\lambda\log\frac{1}{\delta'}}} + \sqrt{\frac{128\mathsf{gap}}{\gamma\lambda\log\frac{1}{\delta'}}} + 94\right)$$

Suppose that $A_0 + \Delta \leq \Lambda$ and $(1-\gamma)^{T_1-T_0}\Lambda \leq A_1$, then $\Pr[X_{T_1} > A_1 \mid X_{T_0} \leq A_0] < \delta'$.

4.3 Step 3: Interval analysis

In this step, we aim to design two sequences $\{t_i\}$ and $\{a_i\}$ of length $\ell+1$. At the i^{th} interval, we invoke Lemma 4.4 with $T_0 = t_{i-1}$, $T_1 = t_i$, $A_0 = a_{i-1}$, and $A_1 = a_i$. Specifically, let $\ell = \lceil \log(1/\epsilon) \rceil$. Let $\delta' = \delta/\ell$, $t_0 = 0, a_0 = 1$. For each $i = 1, \ldots, \ell$, let

$$a_i = 2^{-i}, \gamma_i = \frac{a_{i-1}\mathsf{gap}^2}{80000\lambda \log \frac{1}{\delta'}}, t_i = t_{i-1} + \left\lceil \frac{-2}{\log(1-\gamma_i)} \right\rceil, \ \Lambda_i = 2a_{i-1}\,.$$

Let $T = t_\ell$. Observed that due to the choice of the parameters we have $a_\ell \le \epsilon$ and $1/5 \le (1-\gamma_i)^{t_i - t_{i-1}} \le 1/4$. Now, for each $i = 1, 2, \ldots, \ell$, we invoke Lemma 4.4 with $A_0 = a_{i-1}$, $A_1 = a_i$, $T_0 = t_{i-1}$, $T_1 = t_i$, and $\delta' = \delta/\ell$. It is easy to verify that $a_{i-1} + \Delta_i \le \Lambda$ and $(1-\gamma)^{t_i - t_{i-1}} \le a_i$. By Lemma 4.4 and union bounding over ℓ intervals, we have $\Pr[X_T > \epsilon] < \delta$. By expanding the definition of T with geometric series, we have $T = O(\frac{\lambda \log(\log(1/\epsilon)/\delta)}{\epsilon \mathsf{gap}^2})$ as desired.

5 Solving Linear Bandit with SGD Updates

Finally, we move on to show that our framework is also useful in active learning where the sampling is adaptive. We consider the "model-free" approach, *i.e.,* update the estimation of the unknown parameter via SGD instead of solving the linear regression directly. The idea of using an SGD update appeared in [KPM15], but their design of upper confidence bound (UCB) is heuristic, and no regret bound is provided. [JBNW17] develops an online-to-confidence-set algorithm to achieve $O(n(T\log^2(T)\log(T/\delta))^{1/2})$ regret up to iterated log-factors. They use an online Newton step predictor as a sub-routine to get rid of the dependence on historical data. In contrast, we do not need any sub-routine and update the parameter directly. As a result, we both simplify the procedure and improve the regret bound. We briefly set up the problem as follows and defer the details to Appendix E.

In stochastic linear bandit, there is an unknown parameter $\theta_* \in B(0, L_*) \subseteq \mathbb{R}^n$ and at each time step t the agent is presented with a decision set $D_t \subseteq B(0, L) \subseteq \mathbb{R}^n$, where L_* and L are known, The agent chooses an action $x_t \in D_t$ and subsequently, the agent observe the reward $y_t = \theta_*^T x_t + \epsilon_t$ where $|\epsilon_t| \le 1$ and $\mathbb{E}[\epsilon_t | x_{1:t}, \epsilon_{1:t-1}] = 0$. We make the bounded assumption of noise term only to simplify the presentation and the sub-Gaussian case can be handled by our framework similarly. The full protocol and algorithm is described below in Algorithm 1.

Algorithm 1 LinUCB-SGD

Parameters: $\lambda > 0$, $\eta = \lambda/L^2$, $\beta_t = 288 \max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}$.
Initialize: $\theta_0 \leftarrow 0$, $V_0 \leftarrow \lambda I$
for round $t = 1, \ldots, T$ **do**
  $B_t \leftarrow \{\theta : \|\theta - \theta_{t-1}\|_{V_{t-1}} \le \sqrt{\beta_t}\}$.
  Choose $x_t = \operatorname{argmin}_{x \in D_t} \min_{\theta \in B_t} \langle x, \theta \rangle$.
  Observe the reward $y_t = \langle x_t, \theta_* \rangle + \epsilon_t$.
  $\theta_t \leftarrow \theta_{t-1} + A_t\left(y_t - \theta_{t-1}^T x_t\right) x_t$ where $A_t = \eta V_{t-1}^{-1}$.
  $V_t \leftarrow V_{t-1} + \eta x_t x_t^T$.

Comparing with the standard LinUCB approach [AYPS11, DHK08], we use an SGD approach to update the estimation θ_t of θ_* with a *matrix* learning rate A_t specified in Algorithm 1 instead of computing the linear regression directly (as in standard analysis). The challenge of the SGD case is that now we need to analyze a stochastic process as follows

$$\theta_t - \theta_* = (I - A_t x_t x_t^T)(\theta_{t-1} - \theta_*) + \epsilon_t A_t x_t\,.$$

The goal is to minimize the regret at time T, defined by $R_T = \sum_{t=1}^{T} \langle x_t^* - x_t, \theta_* \rangle$, where x_t^* is the optimal action at time t. By following the standard approach in [AYPS11], it suffices to bound the dynamic of $X_t = \|\theta_t - \theta_*\|_{V_t}^2$. Specifically we have the following main theorem.

Theorem 5.1 (Regret bound for linear bandit with SGD updates)**.** *Setting parameters as in Algorithm 1, for any $\lambda, L, L_* > 0, T \in \mathbb{N}$, we have*

$$\Pr\left[\exists t \in [T],\ X_t > 288 \max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}\right] < \delta\,.$$

In particular, with probability at least $1-\delta$, we have

$$R_T \le 34\sqrt{2nT\max\left\{L_*^2L^2, n\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}\log\left(1+\frac{T}{n}\right)}\,.$$

5.1 Step 1: Linearization and moment analysis

For every $t \in \mathbb{N}$, let us first linearize $X_t := \|\theta_t - \theta_*\|_{V_t}^2$ as follows.

$$X_t \le X_{t-1} + N_t \tag{5.2}$$

where the explicit formula of N_t and the derivation of the linearization can be found in Appendix E. The following lemma provides a moment profile for $\{X_t\}$.

Lemma 5.3 (Moment profile for linear bandit with SGD updates)**.** *Consider the setting in Theorem 5.1. For notational convenience, let $v_t = \|x_t\|_{V_{t-1}^{-1}}$. For every $1 \le T_0 + 1 \le t \in \mathbb{N}$ and $\Lambda > 0$, the following functions $(B_{T_0}, \mu_{T_0}, \sigma_{T_0}^2)$ form a moment profile for $\{X_t\}$, Λ, and T_0.*

$$B_{T_0}(t,\Lambda) = 3\eta v_t\sqrt{\Lambda} + 2\eta^2 v_t^2\,, \mu_{T_0}(t,\Lambda) = 2\eta^2 v_t^2\ ,\ \text{and}\ \sigma_{T_0}^2(t,\Lambda) = 8\eta v_t^2\Lambda + 8\eta^4 v_t^4\,.$$

5.2 Step 2: Improvement analysis

We apply Proposition 2.3 using the moment profile of $\{X_t\}$ in Lemma 5.3 and get the following.

Lemma 5.4 (Improvement analysis for linear bandit with SGD updates)**.** *Consider the setting in Theorem 5.1. For every $A_1 > A_0 > 0$, $\delta' > 0$, $\Lambda > 0$, and $1 \le T_0 < T_1 \in \mathbb{N}$, let*

$$\Delta = \sqrt{144\eta n\log\left(1+\frac{\eta T_1 L^2}{n\lambda}\right)\Lambda\log\frac{1}{\delta'}} + 16\eta n\log\left(1+\frac{\eta T_1 L^2}{n\lambda}\right)\sqrt{\log\frac{1}{\delta'}}$$

Suppose, $A_0 + \Delta \le \Lambda$ and $\Lambda \le A_1$, then $\Pr\left[\max_{T_0+1\le t\le T_1} X_t > A_1 \mid X_{T_0} \le A_0\right] < \delta'$.

5.3 Step 3: Interval analysis

In general, we design multiple intervals in this step to achieve tighter analysis. In this example, this step is surprisingly easy because it suffices to set $\ell = 1, t_0 = 0, t_1 = T, a_0 = \|\theta_*\|_{V_0}^2$ and $a_1 = \Lambda_1 = 288\max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{t_1}{n}\right)\log\frac{1}{\delta}\right\}$. Now, for $i = 1$, we invoke Lemma 5.4 with $A_0 = a_0$, $A_1 = a_1$, $T_0 = 0$, $T_1 = T$, and $\delta' = \delta$. It is easy to verify $A_0 + \Delta \le \Lambda_1$ and $\Lambda_1 \le A_1$. By Lemma 5.4, this implies that

$$\Pr\left[\exists t \in [T],\ X_t > 288 \max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}\right] < \delta.$$

Acknowledgement We thank Kai-Min Chung for useful discussion in the early stage of this work and thank Boaz Barak for helpful comments on a draft of this paper.

References

[AGH+13] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288, 2013.

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[AGKM16] Sanjeev Arora, Rong Ge, Ravi Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization—provably. *SIAM Journal on Computing*, 45(4):1582–1611, 2016.

[AGM12] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE, 2012.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[AZL17] 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.

[AZLS18] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 6673–6685, 2019.

[BBV04] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CL06] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.

[CLSH19] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[CW19] Chi-Ning Chou and Mien Brabeeba Wang. ODE-inspired analysis for the biological version of oja's rule in solving streaming pca. *arXiv preprint arXiv:1911.02363*, 2019.

[DHK08] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[DSOR15] Christopher De Sa, Kunle Olukotun, and Christopher Ré. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2332–2341. JMLR.org, 2015.

[DZPS18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[Har14] Moritz Hardt. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 651–660. IEEE, 2014.

[Haz19] Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

[HK14] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[JAZBJ18] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[JBNW17] Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, pages 99–109, 2017.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[JJK+16] 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.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

[JYWJ19] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

[KPM15] Nathaniel Korda, LA Prashanth, and Rémi Munos. Fast gradient descent for drifting least squares regression, with application to bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[KY97] Harold J. Kushner and G. George Yin. *Stochastic Approximation Algorithms and Applications*. Springer New York, 1997.

[LG16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016.

[LS99] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[LWLZ18] Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167(1):75–97, 2018.

[Moi18] Ankur Moitra. *Algorithmic aspects of machine learning*. Cambridge University Press, 2018.

[Oja82] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[QDC19] Xingye Qiao, Jiexin Duan, and Guang Cheng. Rates of convergence for large-scale nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 10768–10779, 2019.

[RR19] Dominic Richards and Patrick Rebeschini. Optimal statistical rates for decentralised non-parametric regression with linear speed-up. In *Advances in Neural Information Processing Systems*, pages 1214–1225, 2019.

[RSS12] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Set09] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[Sha16] Ohad Shamir. Convergence of stochastic gradient descent for pca. In *International Conference on Machine Learning*, pages 257–265, 2016.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Tan19] Cheng Tang. Exponentially convergent stochastic k-pca without variance reduction. In *Advances in Neural Information Processing Systems*, pages 12393–12404, 2019.

[Vav10] Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.

[YWH19] Yue Yu, Jiaxiang Wu, and Longbo Huang. Double quantization for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 4440–4451, 2019.

[ZCZG18] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

[ZSJ+19] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11127–11135, 2019.

Appendix

- Appendix A provides sufficient tools and backgrounds for the techniques used in this paper.
- Appendix B provides an in-depth discussion on the framework.
- Appendix C provides the details on the example of SGD for strongly convex functions.
- Appendix D provides the details on the example of local convergence of k-PCA.
- Appendix E provides the details on the example of linear bandit with SGD updates.

A Tools and Preliminaries

A.1 Martingale, stopped process, and concentration inequality

Stochastic process is a central tool in this paper. In this subsection, we will introduce preliminary mathematical background related to stochastic process. Interested readers can find detailed exposition in standard text such as [LG16]. Let us start with the definition on adapted random process.

Definition A.1 (Adapted random process)**.** *Let $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$ be a sequence of random variables and $\{\mathcal{F}_t\}_{t\in\mathbb{N}_{\geq 0}}$ be a filtration. We say $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$ is an adapted random process with respect to $\{\mathcal{F}_t\}_{t\in\mathbb{N}_{\geq 0}}$ if for each $t\in\mathbb{N}_{\geq 0}$, the σ-algebra generated by $X_0, X_1, \ldots, X_t$ is contained in $\mathcal{F}_t$.*

In most of the situation, we use $\mathcal{F}_t$ to denote the *natural filtration* of $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$, namely, $\mathcal{F}_t$ is defined as the σ-algebra generated by $X_0, X_1, \ldots, X_t$. One of the most common adapted processes is the martingale.

Definition A.2 (Martingale)**.** *Let $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ be a sequence of random variables and let $\{\mathcal{F}_t\}_{\mathbb{N}}$ be its natural filtration. We say $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ is a martingale if for each $t\in\mathbb{N}$, $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = M_t$.*

Note that for any adapted random process $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$, one can always turn it into a martingale by defining $M_0 = X_0$ and for any $t\in\mathbb{N}$, let $M_t = X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$. When the bounded difference and moments of a martingale can be bounded almost surely, we can apply any of the martingale concentration inequality in [CL06]. Furthermore, all the martingale inequality in [CL06] can be strengthened into the maximal form to avoid union bound by using Doob's maximal inequality [LG16] instead of Markov's inequality. For example, a small modification of the classical Freedman's inequality gives the following lemma.

Lemma A.3 (A martingale concentration inequality)**.** *Let $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ be an adapted stochastic process. Let $T_0 < T \in \mathbb{N}_{\geq 0}$ and $B, \sigma_t, \mu_t \geq 0$ be some constants for all $t\in[T]$. Suppose for each $t = T_0+1, 2, \ldots, T$, $M_t - M_{t-1} \leq B$ almost surely, $\mathsf{Var}[M_t \mid \mathcal{F}_{t-1}] \leq \sigma_t^2$, and $\mathbb{E}[M_t - M_{t-1} \mid \mathcal{F}_{t-1}] \leq \mu_t$, then for every $\delta\in(0,1)$ we have*

$$\Pr\left[\exists T_0+1\leq t\leq T,\ M_t - M_0 \geq 2\max\left\{\sqrt{\sum_{t'=T_0+1}^{T_1}\sigma_{t'}\log\frac{1}{\delta}}, 2B\log\frac{1}{\delta}\right\} + \sum_{t'=T_0+1}^{T_1}\mu_{t'}\right] < \delta\,.$$

However, in general, it is difficult to obtain a good bound on the bounded difference and the moments of a stochastic process. One powerful technique to deal with this issue is using the *stopping time* defined as follows.

Definition A.4 (Stopping time)**.** *Let $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$ be an adapted process associated with filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}_{\geq 0}}$. An integer-valued random variable τ is a stopping time for $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$ if for all $t\in\mathbb{N}$, $\{\tau = t\}\in\mathcal{F}_t$.*

Let $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ be an adapted process, the most common stopping time for $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ is of the following form. For any $a\in\mathbb{R}$, let

$$\tau := \min_{M_t > a} t\,.$$

Namely, τ is the first time when the martingale becomes at least a. For convenience, in the rest of the paper, we would define stopping time of this form by saying "τ is the stopping time for the event $\{M_t > a\}$".

Given an adapted process $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ and a stopping time τ, it is then natural to consider the corresponding *stopped process* $\{M_{t\wedge\tau}\}_{t\in\mathbb{N}_{\geq 0}}$ where $t\wedge\tau = \min\{t,\tau\}$ is also a random variable. An useful and powerful fact here is that the stopped process of a martingale is also a martingale. Also, notice that we have

$$M_{t\wedge\tau} - M_{(t-1)\wedge\tau} = \mathbf{1}_{\tau\geq t}(M_t - M_{t-1})\,.$$

Therefore, by considering a stopping time for the event $\{M_t > \Lambda\}$, we form a small bounded difference and moments for the stopped process in the moment profile (see Definition 2.2) and therefore can obtain a concentration on the stopped process. For example, if we apply Lemma A.3 on the stopped process, we have

$$\Pr\left[\exists 1\leq t\leq T,\ M_{t\wedge\tau} - M_0 \geq 2\max\left\{\sqrt{\sum_{t'=1}^{T_1}\sigma_{t'}\log\frac{1}{\delta}}, 2B\log\frac{1}{\delta}\right\} + \sum_{t'=1}^{T_1}\mu_{t'}\right] < \delta\,.$$

A.2 The pull-out lemma

In the end of the previous subsection, we see how to apply concentration inequality on the stopped process. To extend the concentration to the original process, the pull-out lemma [CW19] provides a sufficient condition to remove the stopping time without paying any extra factor. Here we restate the lemma and give a whole proof for the completeness of presentation.

Lemma A.5 (The Pull-out lemma [CW19])**.** *Let $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ be an adapted stochastic process and τ be a stopping time. Let $\{M_t^*\}_{t\in\mathbb{N}_{\geq 0}}$ be the maximal process of $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ where $M_t^* = \max_{1\leq t'\leq t} M_t$. For any $t\in\mathbb{N}$, $\Delta\in\mathbb{R}$, and $\delta\in(0,1)$, suppose*

1. *$\Pr[M_{t\wedge\tau}^* > \Delta] < \delta$ and*
2. *For any $1\leq t' < t$, $\Pr[\tau\geq t'+1 \mid M_{t'}^*\leq\Delta] = 1$.*

Then, we have

$$\Pr[M_t^* > \Delta] < \delta\,.$$

For the completeness of the presentation, we provide a proof for the pull-out lemma as follows.

Proof of Lemma A.5. The main idea is to use an auxiliary stopping time ξ for the event $\{M_{t\wedge\tau}^* > \Delta\}$. First, we would like to show that if τ stopped before time t, then ξ must stop no later than τ too. For the sake of contradiction, assume $\tau < t$ but $\xi > \tau$. Then we know from the definition of ξ that $M_\tau^* \leq \Delta$ because $\xi > \tau$. From the second condition of the lemma statement, we have $\tau \geq \tau + 1$ which is a contradiction. Thus, we have $\Pr[\tau < t,\ \xi > \tau] = 0$.

Next, consider the following decomposition of the error probability $\Pr[M_t^* > \Delta]$.

$$\begin{aligned}\Pr[M_t^* > \Delta] &= \Pr[M_t^* > \Delta,\ \tau\geq t] + \Pr[M_t^* > \Delta,\ \tau < t,\ \xi\leq\tau]\\ &\quad + \Pr[M_t^* > \Delta,\ \tau < t,\ \xi > \tau]\\ &= \Pr[M_t^* > \Delta,\ \tau\geq t] + \Pr[M_t^* > \Delta,\ \tau < t,\ \xi\leq\tau]\end{aligned}$$

where $\Pr[M_t^*,\ \tau < t,\ \xi > \tau] = 0$ as explained in the previous paragraph. Now, observe that when $\tau\geq t$, we have $t = t\wedge\tau$. Also, if $\xi\leq\tau < t$ then $M_t^*, M_{t\wedge\xi}^* > \Delta$ according to the definition of ξ. Namely, we can turn the process into stopped process in the above equation as follows.

$$\begin{aligned}&= \Pr[M_{t\wedge\tau}^* > \Delta,\ \tau\geq t] + \Pr[M_{t\wedge\tau}^* > \Delta,\ \tau < t,\ \xi\leq\tau]\\ &\leq \Pr[M_{t\wedge\tau}^* > \Delta] < \delta\end{aligned}$$

where the last inequality is due to the first condition in the lemma statement. Thus, we have $\Pr[M_t^* > \Delta] < \delta$ as desired. ∎

A.3 Matrix norms and inequalities

As many common potential functions are defined as the norm of certain matrix, here we provide some common matrix norms and inequalities that will be useful.

Definition A.6 (Matrix norms)**.** *Let $A \in \mathbb{R}^{n\times m}$.*

- *The Frobenius norm of A is defined as*

$$\|A\|_F := \sqrt{\mathsf{tr}(A^\top A)}\,.$$

- *The operator norm of A is defined as*

$$\|A\| := \sup_{\mathbf{x}\in\mathbb{R}^n\setminus\{0^n\}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}\,.$$

- *The Schatten p norm of A for some $p \geq 1$ is defined as*

$$\|A\|_p := \mathsf{tr}(|A|^p)^{1/p}$$

  where $|A| := \sqrt{A^\top A}$.

- *The matrix inner product of $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ is defined as*

$$\langle \mathbf{x}, \mathbf{y}\rangle_A := \mathbf{x}^\top A\mathbf{y}\,.$$

  When A is a square matrix, the A-norm of $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_A := \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle_A}\,.$$

Note that $\|A\|_F = \|A\|_2$ and $\|A\|_p^p = \sum_{i=1}^{n\wedge m} \sigma_i(A)^p$ where $\sigma_1(A) \geq \cdots \geq \sigma_n(A) \geq 0$ are the singular values of A. Therefore, we have $\|AA^\top\|_1 = \|A\|_F^2$ and $\|A\|_\infty = \|A\|$.

Lemma A.7 (Matrix inequalities)**.** *Let $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times k}$.*

- $\|AB\|_F \leq \|A\|\|B\|_F$.
- $\|A\| \leq \|A\|_F \leq \sqrt{m}\|A\|$.
- *(Matrix Cauchy-Schwarz inequality):* $|\mathsf{tr}(AB)| \leq \|A\|_F\|B\|_F$.
- *(Matrix Hölder inequality):* $|\mathsf{tr}(AB)| \leq \|A\|_p\|B\|_q$ *for every* $p, q \geq 1$ *such that* $1/p + 1/q = 1$.
- *(Matrix AM-GM inequality):* $|2\mathsf{tr}(AB)| \leq \mathsf{tr}(A^\top A) + \mathsf{tr}(B^\top B) = \|A\|_F^2 + \|B\|_F^2$.

A.4 Convergence guarantees

Here we define four common notions of high-probability convergence with increasing generality.

Definition A.8 (Convergence guarantees)**.** *Let $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$ be a stochastic process with learning rate $\{\eta_t\}_{t\in\mathbb{N}_{\geq 0}}$. There are four types of high-probability bounds of interest as follows. Let $\delta > 0$ be the failure probability.*

- *(Weak non-uniform bound) There exists a function $T(\epsilon, \delta)$ such that for any $\epsilon > 0$, there exists a learning rate $\{\eta_t\}_{t\in\mathbb{N}_{\geq 0}}$ such that we have $\Pr[X_{T(\epsilon,\delta)} > \epsilon] < \delta$.*
- *(Strong non-uniform bound) There exists a learning rate $\{\eta_t\}_{t\in\mathbb{N}_{\geq 0}}$ and a function $T(\epsilon, \delta)$ such that for any $\epsilon > 0$, we have $\Pr[X_{T(\epsilon,\delta)} > \epsilon] < \delta$.*

- *(Weak uniform bound) There exists a learning rate* $\{\eta_t\}_{t\in\mathbb{N}_{\geq 0}}$ *and a function* $\epsilon(t,T,\delta)$ *such that for any* $T\in\mathbb{N}$ *large enough, we have* $\Pr[\exists t\in[T],\ X_t>\epsilon(t,T,\delta)]<\delta$.
- *(Strong uniform bound) There exists a learning rate* $\{\eta_t\}_{t\in\mathbb{N}_{\geq 0}}$ *and a function* $\epsilon(t,\delta)$ *such that for any* $T\in\mathbb{N}$ *large enough, we have* $\Pr[\exists t\in[T],\ X_t>\epsilon(t,\delta)]<\delta$.

For example, we obtain weak non-uniform bound in Section C.3.2, obtain strong uniform bound in Section C.3.1 and Section D.3, and obtain weak uniform bound in Section E.3.

B Details on the framework

In this section, we elaborate on the framework and provide complete proofs for the lemmas and propositions in the main article.

B.1 Why non-linearity in the dynamic creates entanglement?

In this subsection, we talk about how non-linearity in the stochastic updates can create the entanglement between the noise and the process. Consider the following general update rule

$$X_t = f(X_{t-1}, N_t).$$

By doing Taylor expansion on f, we have

$$X_t = a_0(N_t) + a_1(N_t)X_{t-1} + a_2(N_t)X_{t-1}^2 + \cdots.$$

For f to be nonlinear, either for $i\geq 1$, $a_i(N_t)$ is not a constant or for $i>1$, $a_i(N_t)\neq 0$. If $a_i(N_t)$ is not constant, then the update rule is entangled already by looking at $a_i(N_t)X_{t-1}^i$. So it suffices to assume that only $a_0(N_t)$ depends on N_t. Now if $i>1$ and $a_i(N_t)\neq 0$, we can unfold the expression one more step and get

$$a_iX_{t-1}^i = a_i(a_0(N_{t-1}) + a_1(N_{t-1})X_{t-2} + a_2(N_{t-1})X_{t-2}^2 + \cdots)^i = ia_ia_0(N_{t-1})(a_iX_{t-2}^i)^i + \cdots.$$

We can see that $a_0(N_{t-1})$ and X_{t-2} are entangled together.

B.2 Continuous analysis

In this subsection, we talk about how continuous analysis can serve as a guide on how to analyze the discrete dynamic. This subsection is not needed for the use of the framework, so the reader is welcome to skip it for the first time reading through this paper. The continuous analysis helps us in three ways:

1. Determine the intrinsic behaviors of the dynamic.
2. Give us a guide on how to analyze the dynamic.
3. Help us to write down a closed-form solution of the dynamic that is adapted.

Determine the intrinsic behaviors of the dynamic. One important thing to do is to understand the intrinsic behaviors of the dynamic first. For example, if at the continuous limit, the system is chaotic, it is probably useless to analyze further since we can't hope the discrete dynamic to be any better than chaos. Therefore, one natural way to study the stochastic system is to consider its continuous limit and then study the corresponding random dynamical system to characterize different fixed points, saddle points, limit cycles, etc. Only by fully understanding the continuous counterpart, we can cope with the intrinsic behaviors of the dynamic in the discrete setting. This leads us to our second point.

Give us a guide on how to analyze the dynamic. By looking at the continuous system as a random dynamical system, we can obtain a strategy to analyze the dynamic. For example, even without the noise, it is not obvious how to analyze

$$X_t = X_{t-1} + \eta X_{t-1}(1 - X_{t-1})\,.$$

However, at the continuous limit, this gives us

$$dX_t = X_{t-1}(1 - X_{t-1})dt$$

which has a stable fixed point at 1 and an unstable fixed point at 0. This suggests that around 0, we should linearize at 0

$$X_t = (1 + \eta(1 - X_{t-1}))X_{t-1}\,.$$

On the other hand, when the dynamic gets closer to 1, we should linearize at 1,

$$X_t - 1 = (1 - \eta X_{t-1})(X_{t-1} - 1)\,.$$

We recommend [CW19] for a detailed discussion on related ideas.

Help us to write down a closed-form solution of the dynamic that is adapted. In a stochastic differential equation, we solve the dynamic by writing it down as a stochastic integral. This suggests that to solve a stochastic difference equation, we should write it as a linear combination of the noise. This is how Equation 2.1 appears. Recall that given a linear dynamic

$$X_t = H_t \cdot X_{t-1} + N_t\,,$$

Equation 2.1 gives us

$$X_t = \prod_{i=1}^{t} H_t \left(X_0 + \sum_{i=1}^{t} \frac{N_i}{\prod_{j=1}^{i} H_j} \right)\,.$$

Comparing with the continuous counterpart, we have

$$\frac{dX(t)}{dt} = H(t)X(t) + N(t)$$

and

$$X(t) = e^{H(t)} \left(X(0) + \int_0^t e^{-H(s)} N(s) ds \right)$$

which we can see the correspondence easily. Furthermore, some readers might argue that Equation 2.1 is simply an unrolling recursion to write down a closed-form solution, but we claim that it is a very special closed-form solution in which the summation of the noise term is adapted.

Notice that $\sum_{i=1}^{t} \prod_{j=1}^{i} H_j^{-1} N_i$ in Equation 2.1 does not depend on the future event. This is naturally true because it is the discrete counterpart of a stochastic integral. However, if we write the closed-form solution as

$$X_t = \prod_{i=1}^{t} H_t X_0 + \sum_{i=1}^{t} \prod_{j=i+1}^{t} H_j N_i\,,$$

the process $\sum_{i=1}^{t} \prod_{j=i+1}^{t} H_j N_i$ is no longer adapted and hence we are not able to apply martingale concentration technique. In the situation where we need to write down closed form solution not from a linear approximation, translating how the continuous counterpart writes down the stochastic integral will help us to write down a closed form solution with adapted noise.

B.3 Linearization and moment analysis

In this section, we provide the proof of *unfolding the recursion, a.k.a.,* the ODE trick in [CW19].

Lemma B.1. *Let $\{X_t\}_{t\geq \mathbb{N}_{\geq 0}}$, $\{N_t\}_{t\in\mathbb{N}}$, and $\{H_t\}_{t\in\mathbb{N}}$ be sequences of random variables with the following dynamic*

$$X_t \leq H_t \cdot X_{t-1} + N_t \tag{B.2}$$

for all $t \in \mathbb{N}$. Then for all $T_0, t \in \mathbb{N}_{\geq 0}$ such that $T_0 < t$, we have

$$X_t \leq D_t \cdot (X_{T_0} + M_t)$$

where $D_t = \prod_{t'=T_0+1}^{t} H_{t'}$ and $M_t = \sum_{t'=T_0+1}^{t} D_{t'}^{-1} \cdot N_{t'}$.

Proof of Lemma B.1. For each $T_0 + 1 \leq t' \leq t$, dividing Equation B.2 with $D_{t'}$ on both sides, we have

$$\frac{X_{t'}}{D_{t'}} = \frac{X_{t'-1}}{D_{t'-1}} + \frac{N_{t'}}{D_{t'}}\,.$$

We get the desiring expression by telescoping the above equation from $t' = T_0 + 1$ to t. □

B.4 Improvement analysis

Let us restate the main proposition of the improvement analysis as follows.

Proposition 2.3 (Improvement analysis)**.** *Let $\{X_t\}$ be a stochastic process described in Equation 2.1 for some $\{D_t\}$ and $\{M_t\}$. Let $\Lambda > 0$, τ be the stopping time for the event $\{X_t > \Lambda\}$, and (B, μ, σ^2) be a moment profile for $\{X_t\}$ and Λ. For every $T_1 \in \mathbb{N}$ and $\delta' > 0$, let*

$$\Delta = \Delta\left(B, \mu, \sigma^2, \Lambda, T_1, \delta'\right)$$

be the deviation from a concentration inequality[9] such that $\Pr\left[\max_{1\leq t\leq T_1} M_{t\wedge\tau} > \Delta\right] < \delta'$.
If we have

- *(Improvement condition) $D_{T_1} \cdot (A_0 + \Delta) \leq A_1$ and*
- *(Pull-out condition) $D_t \cdot (A_0 + \Delta) \leq \Lambda$ for every $1 \leq t \leq T_1$.*

Then we have

$$\Pr\left[\exists t \in [T_1],\ X_t > D_t \cdot (A_0 + \Delta) \mid X_0 \leq A_0\right] < \delta'\,.$$

In particular, the above implies $\Pr[X_{T_1} > A_1 \mid X_0 \leq A_0] < \delta'$. Also, the proposition can naturally extend to starting from $T_0 \in \mathbb{N}$ instead of 0.

Proof of Proposition 2.3. We would like to apply the pull-out lemma, *i.e.,* Lemma A.5, on $\{M_{t\wedge\tau}\}_{t\in\mathbb{N}}$ and thus have to verify the following two conditions. First, note that $\{M_{t\wedge\tau}\}_{t\in\mathbb{N}}$ forms a martingale. Thus, due to the martingale concentration inequality and the choice of Δ, we have

$$\Pr\left[\max_{1\leq t\leq T_1} M_{t\wedge\tau} > \Delta\right] < \delta'$$

and thus we satisfy the first condition of the pull-out lemma. Next, for every $1 \leq t \leq T_1$, suppose $\max_{1\leq t'\leq t} M_{t'} \leq \Delta$, then we have the following from the recursion formula (see Equation 2.1).

$$X_t = D_t \cdot (X_0 + M_t) \leq D_t \cdot (A_0 + \Delta) \leq \Lambda$$

[9] For example, $\Delta = \sqrt{2\sum_{t'=T_0+1}^{T_1} B_{T_0}^2(t', \Lambda)\log(1/\delta')} + \sum_{t'=T_0+1}^{T_1} \mu_{T_0}(t', \Lambda)$ if we used Azuma's inequality. See Appendix B and [CL06] for more examples.

where the last inequality is from the pull-out condition in the proposition statement. Thus, by the choice of τ, we have

$$\Pr\left[\tau > t \;\middle|\; \max_{1\le t'\le t} M_{t'} \le \Delta\right] = 1\,.$$

The above two satisfy the second condition of the pull-out lemma as desired. As a result, by the pull-out lemma (see Lemma A.5), we have pulled out the stopping time as follows.

$$\Pr\left[\max_{1\le t\le T_1} M_t > \Delta\right] < \delta'\,.$$

Finally, by the recursion formula, we have

$$\Pr\left[\exists t \in [T_1],\ X_t > D_t \cdot (A_0 + \Delta) \mid X_0 \le A_0\right] < \delta'\,.$$

In particular, by combining with the improvement condition in the proposition statement, we have

$$\Pr[X_{T_1} > A_1 \mid X_0 \le A_0] < \delta'$$

as desired. □

B.5 Interval analysis

In the improvement analysis, *i.e.,* Proposition 2.3, we get the improvement guarantee $\Pr[X_{T_1} > A_1 \mid X_{T_0} \le A_0] < \delta'$ as long as the two conditions are satisfied. Recall that we start from X_0 and want to see how small X_T could be with high probability. As the first try, we can invoke Proposition 2.3 by setting $T_0 = 0$, $T_1 = T$, and $A_0 = X_0$ then see what is the smallest A_1 we can get. Nevertheless, such analysis in general would not be tight because it does not use the local information.

To fully leverage the improvement analysis, we perform an interval analysis by designing sequences $\{a_i\}$, $\{t_i\}$, $\{\delta_i\}$ of length ℓ (or $\ell + 1$) such that we invoke Proposition 2.3 by setting $T_0 = t_{i-1}$, $T_1 = t_i$, $A_0 = a_{i-1}$, $A_1 = a_i$, and $\delta' = \delta_i$ for each $i = 1, 2, \ldots, \ell$. Namely, in the i^{th} interval, we would like to show $\Pr[X_{t_i} > A_i \mid X_{t_{i-1}} \le A_{i-1}] < \delta_i$ and thus by union bound we would have $\Pr[X_{t_\ell} > A_\ell \mid X_{t_0} \le A_0] < \sum_i \delta_i$.

Note that with the general recipe as above, in principle one can get the tightest bound by solving the following optimization problem.

$$\begin{aligned}
\underset{\ell\in\mathbb{N},\{t_i\},\{a_i\},\{\delta_i\},\{\Lambda_i\}}{\text{minimize}} \quad & a_\ell \\
\text{subject to} \quad & a_i \ge \prod_{t=t_{i-1}+1}^{t_i} H_t \cdot (a_{i-1} + \Delta_i)\,,\ \forall i = 1, 2, \ldots, \ell \\
& a_{i-1} + \Delta_i \le \Lambda_i\,,\ \forall i = 1, 2, \ldots, \ell \\
& \Pr[X_0 > a_0] < \delta_0 \\
& \sum_i \delta_i \le \delta \\
& 0 = t_0 < t_1 < \cdots < t_\ell = T\,.
\end{aligned}$$

However, in general the above optimization problem might be complicated to solve optimally by hands. Thus, we provide some common ways to set the intervals as a principle to implement interval analysis.

How to set $\{a_i\}$. For simplicity, let us focus on the setting where the goal is moving from X_0 to ϵ where $X_0 \gg \epsilon > 0$. Namely, $a_0 = X_0$ and $a_\ell = \epsilon$. The principle here can be easily adapted to other settings. We provide three common ways of setting $\{a_i\}$: the *greedy improvement*, the *multiplicative improvement*, and the *polynomial improvement*. See Table 1 for a summary.

Type	$\{a_i\}$	# intervals	Example
Greedy	Pick a_i as small as possible	Problem-dependent	Section E.3
Multiplicative	$a_i = \frac{a_{i-1}}{2} = 2^{-i} \cdot X_0$	$\lceil \log \frac{X_0}{\epsilon} \rceil$	Section C.3.1 & Section D.3
Polynomial	$a_i = \frac{\epsilon}{4} \cdot \left(\frac{4a_{i-1}}{\epsilon}\right)^{\frac{3}{4}} = \frac{\epsilon}{4} \cdot (\frac{4X_0}{\epsilon})^{(\frac{3}{4})^i}$	$\left\lceil \frac{\log \log \frac{4X_0}{\epsilon}}{\log \frac{4}{3}} \right\rceil$	Section C.3.2

Table 1: Three common ways of setting $\{a_i\}$ in an interval analysis. We specify how to pick a_i according to a_{i-1} in the second column and calculate the number of intervals in the third column. Note that the constants used here are arbitrarily chosen and can be further optimized during the implementation. For concrete examples, see the fourth column.

How to set $\{\delta_i\}$. In general, a handy way to set $\{\delta_i\}$ is setting $\delta_i = \delta/(2i^2)$. First, note that $\sum_i \delta_i \le \delta$ as desired. Second, in high probability bound, we usually get $\log \frac{1}{\delta'}$ dependency from the martingale concentration. In such case, we have $\log \frac{1}{\delta_i} = \log \frac{1}{\delta} + 2 \log i + 1$ where $\log \frac{1}{\delta}$ is essential from concentration and $\log i$ is the cost of union bound. See Section C.3 and Section D.3 for concrete examples.

C Details on SGD for Strongly Convex Functions

Let F be a convex function over some convex domain $\mathcal{W}$ equipped with norm $\|\cdot\|$. There is a G-bounded stochastic gradient oracle where on input $\mathbf{w}$, it returns a random $\hat{\mathbf{g}}$ such that $\mathbb{E}[\hat{\mathbf{g}}]$ is the gradient of F at $\mathbf{w}$ and $\|\hat{\mathbf{g}}\|^2 \le G^2$ almost surely for some constant $G > 0$. The goal is to minimize F over $\mathcal{W}$ with the help of this stochastic gradient oracle. The stochastic gradient descent (SGD) algorithm maintains a vector $\mathbf{w}_t$ at time t as follows.

Algorithm 2 SGD for strongly convex function

Input: Time parameter $T \in \mathbb{N}$ and step size parameters $\{\eta_t\}_{t \in \mathbb{N}}$.

1: Initialize $\mathbf{w}_0 \in \mathcal{W}$.
2: **for** $t = 1, \dots, T$ **do**
3: Query the stochastic gradient oracle of F at $\mathbf{w}_{t-1}$ and get $\hat{\mathbf{g}}_t$.
4: Let $\mathbf{w}_t = \mathsf{Proj}_{\mathcal{W}}(\mathbf{w}_{t-1} - \eta_t \hat{\mathbf{g}}_t)$ where $\mathsf{Proj}_{\mathcal{W}}$ is an orthogonal projection operator for $\mathcal{W}$.

Output: $\mathbf{w}_T$.

We say F is λ-strongly convex for some $\lambda > 0$ if for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ and any subgradient $\mathbf{g}$ at $\mathbf{w}$,

$$F(\mathbf{w}') \ge F(\mathbf{w}) + \mathbf{g}^\top (\mathbf{w}' - \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}' - \mathbf{w}\|^2 .$$

Due to the strong convexity, there exists a unique minimizer $\mathbf{w}^* \in \mathcal{W}$ of F. Thus, the potential function is defined as $X_t = \|\mathbf{w}_t - \mathbf{w}^*\|^2$. Note that due to the strong convexity again, we have

$$G\|\mathbf{w} - \mathbf{w}^*\| \ge \|\mathbf{g}\| \cdot \|\mathbf{w} - \mathbf{w}^*\| \ge \langle \mathbf{g}, \mathbf{w} - \mathbf{w}^* \rangle \ge \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}^*\|^2$$

where $\mathbf{g}$ is the gradient of $\mathbf{w}$. By dividing $\frac{2}{\lambda}\|\mathbf{w} - \mathbf{w}^*\|$ and square both side, we have for every $t \in \mathbb{N}$

$$X_t \le \frac{4G^2}{\lambda^2}$$

almost surely. The goal is to prove the following theorem, which is a formal version of Theorem 3.2.

Theorem C.1 (Convergence of SGD for strongly convex functions)**.** *Let F be a λ-strongly convex function over some convex domain. An SGD algorithm with gradient bounded by $G > 0$ almost surely has the following convergence rate. For every $\delta \in (0,1)$,*

- *if $\eta_t = 1/(\lambda t)$, we have the following strong uniform convergence*

$$\Pr\left[\exists t \in \mathbb{N},\ X_t > \frac{10000G^2\left(\log\frac{1}{\delta} + 2\log\log(t+2)\right)}{\lambda^2 t}\right] < \delta\,;$$

- *there exists a deterministic η_t such that for every $\epsilon \in (0,1)$, there exists*

$$T \le \frac{14500G^2\left(\frac{1}{\delta} + 2\log\log\log\frac{G}{\lambda\epsilon} + 10\right)}{\epsilon\lambda^2}$$

such that

$$\Pr\left[X_T > \epsilon\right] < \delta\,.$$

Comparison. For the learning rate $\eta_t = 1/(\lambda t)$, we match the state-of-the-art $O(\frac{G^2(\log\delta^{-1}+\log\log t)}{\lambda^2 t})$ convergence rate. Specifically, [HK14] shows the strong non-uniform convergence and [RSS12] shows the weak uniform convergence while Theorem C.1 gives the strong uniform convergence (see Definition A.8 for the definitions of these convergence guarantees). In addition, Theorem C.1 further gives the first convergence rate that breaks the $\log\log t$ barrier to $O(\frac{G^2(\log\delta^{-1}+\log\log\log t)}{\lambda^2 t})$ and partially answers the open problem in [HK14]. Note that information-theoretically the convergence rate is at least $\Omega(\frac{G^2\log\delta^{-1}}{\lambda^2 t})$.

Section structure. In the rest of this section, we provide the linearization and moment analysis in Section C.1, the improvement analysis in Section C.2, and the interval analysis in Section C.3.

C.1 Linearization and moment analysis

Let us start with the linearization for X_t. For every $0 \le T_0 < T_1$, we rewrite the dynamic as follows.

$$\begin{aligned}
X_t &= \|\mathbf{w}_t - \mathbf{w}^*\|^2 = \|\mathsf{Proj}_{\mathcal{W}}(\mathbf{w}_{t-1} - \eta_t\hat{\mathbf{g}}_t) - \mathbf{w}^*\|^2 \le \|\mathbf{w}_{t-1} - \eta_t\hat{\mathbf{g}}_t - \mathbf{w}^*\|^2 \\
&= \|\mathbf{w}_{t-1} - \mathbf{w}^*\|^2 - 2\eta_t\hat{\mathbf{g}}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}^*) + \eta_t^2\|\hat{\mathbf{g}}_t\|^2 \\
&= (1 - 2\eta_t\lambda)X_{t-1} + 2\eta_t\left(\lambda X_{t-1} - \hat{\mathbf{g}}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}^*)\right) + \eta_t^2\|\hat{\mathbf{g}}_t\|^2\,.
\end{aligned}$$

Let $M_t = \sum_{t'=T_0+1}^{t} \frac{2\eta_{t'}\left(\lambda X_{t'-1} - \hat{\mathbf{g}}_{t'}^\top(\mathbf{w}_{t'-1} - \mathbf{w}^*)\right) + \eta_{t'}^2\|\hat{\mathbf{g}}_{t'}\|^2}{\prod_{t''=T_0+1}^{t'}(1-2\eta_{t''}\lambda)}$ for all $T_0 + 1 \le t \le T_1$. We can unfold the recursion and get the following.

$$X_t = \prod_{t'=T_0+1}^{t}(1 - 2\eta_{t'}\lambda)\left(X_{T_0} + M_t\right)\,.$$

Lemma C.2 (Moment profile for SGD)**.** *Consider the setting in Theorem C.1. Let $\Lambda > 0$ and τ is the stopping time for the event $\{X_t > \Lambda\}$. For every $T_0 + 1 \le t \le T$, the following following functions $(B_{T_0}, \mu_{T_0}, \sigma_{T_0}^2)$ form a moment profile for $\{X_t\}$, Λ, and T_0.*

- *(Bounded difference) $|M_{t\wedge\tau} - M_{(t-1)\wedge\tau}| \le B(t,\Lambda) = \eta_t \cdot \frac{12G/\lambda + \eta_t G^2}{\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda)}$ almost surely.*
- *(Conditional expectation) $\mathbb{E}\left[M_{t\wedge\tau} - M_{(t-1)\wedge\tau} \mid \mathcal{F}_{t-1}\right] \le \mu(t,\Lambda) = \eta_t^2 \cdot \frac{G^2}{\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda)}$.*
- *(Conditional variance) $\mathsf{Var}\left[M_{t\wedge\tau} - M_{(t-1)\wedge\tau} \mid \mathcal{F}_{t-1}\right] \le \sigma^2(t,\Lambda) = \eta_t^2 \cdot \frac{72G^2\Lambda + 2\eta_t^2 G^4}{\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda)^2}$.*

Specifically,

- *if* $\eta_t = 1/(\lambda t)$ *and* $T_0 \geq 73$*, we can pick*

$$B_{T_0}(t,\Lambda) = \frac{13G^2 t}{\lambda^2 T_0^2}, \mu_{T_0}(t,\Lambda) = \frac{2G^2}{\lambda^2 T_0^2}, \text{ and } \sigma^2_{T_0}(t,\Lambda) = \frac{G^2 t^2}{\lambda^2 T_0^4}\left(73\Lambda + \frac{3G^2}{\lambda^2 t^2}\right).$$

- *if* $\eta_t = \eta = \gamma/(2\lambda)$ *for some* $\gamma > 0$*,*

$$B_{T_0}(t,\Lambda) = \frac{G^2\gamma \cdot \left(6+\frac{\gamma}{4}\right)}{\lambda^2(1-\gamma)^{t-T_0}}, \mu_{T_0}(t,\Lambda) = \frac{G^2\gamma^2}{4\lambda^2(1-\gamma)^{t-T_0}}, \text{ and } \sigma^2_{T_0}(t,\Lambda) = \frac{G^2\gamma^2 \cdot \left(18\Lambda + \frac{G^2\gamma^2}{8\lambda^2}\right)}{\lambda^2(1-\gamma)^{2(t-T_0)}}.$$

Proof of Lemma C.2. For bounded difference, we have

$$|M_{t\wedge\tau} - M_{(t-1)\wedge\tau}| = \left|\mathbf{1}_{\tau\geq t}\frac{2\eta_t\left(\lambda X_{t-1} - \hat{\mathbf{g}}_t^\top(\mathbf{w}_{t-1}-\mathbf{w}^*)\right) + \eta_t^2\|\hat{\mathbf{g}}\|^2}{\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda)}\right|.$$

By Cauchy-Schwarz inequality and the fact that $\|\hat{\mathbf{g}}_t\| \leq G$, we have

$$\leq \mathbf{1}_{\tau\geq t}\eta_t \cdot \frac{2\lambda X_{t-1} + 2G\sqrt{X_{t-1}} + \eta_t G^2}{\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda)}.$$

Because $X_t \leq \frac{4G^2}{\lambda^2}$, we have

$$\leq \frac{12G^2/\lambda + \eta_t G^2}{\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda)}. \tag{C.3}$$

For the conditional expectation, notice that $\mathbb{E}[\hat{\mathbf{g}}_t^\top(\mathbf{w}_{t-1}-\mathbf{w}^*) \mid \mathcal{F}_{t-1}] \geq \lambda X_{t-1}$ due to strong convexity and $\|\hat{\mathbf{g}}_t\| \leq G$ due to G-boundedness, we have

$$\mathbb{E}\left[M_{t\wedge\tau} - M_{(t-1)\wedge\tau} \mid \mathcal{F}_{t-1}\right] \leq \eta_{t'}^2 \cdot \frac{G^2}{\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda)}.$$

For the conditional variance, we have by Equation C.3 and the definition of τ,

$$\mathsf{Var}\left[M_{t\wedge\tau} - M_{(t-1)\wedge\tau} \mid \mathcal{F}_{t-1}\right] \leq \mathbf{1}_{\tau\geq t}\cdot(M_t - M_{t-1})^2 \leq \eta_t^2 \cdot \frac{72G^2\Lambda + 2\eta_t^2 G^4}{\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda)^2}.$$

Notice that for $T_0 \geq 73$

$$\prod_{t'=T_0+1}^{t}(1-2\eta_{t'}\lambda) = \frac{T_0(T_0-1)}{t(t-1)} \geq \frac{72T_0^2}{73t^2}. \tag{C.4}$$

Now plug the specific learning rate back, we obtain the bound in the lemma statement. ∎

C.2 Improvement analysis

Lemma C.5. *Let* $\{X_t\}$ *be the stochastic process described in Equation 3.1,* $1 \leq T_0 < T_1 \in \mathbb{N}$*,* $\Lambda > 0$*, and* $(B_{T_0}, \mu_{T_0}, \sigma^2_{T_0})$ *be a moment profile for* $\{X_t\}$*,* T_0*, and* Λ *from Lemma C.2.*

- *If* $\eta_t = 1/\lambda t$*, then for every* $A_0 > A_1 > 0$*,* $\delta' > 0$*,* $\Lambda > 0$*, and* $73 \leq T_0 < T_1 \in \mathbb{N}$*, let*

$$\Delta = \frac{G^2 T_1 \log\frac{1}{\delta'}}{\lambda^2 T_0^2}\left(\sqrt{\frac{292 T_1 \Lambda\lambda^2}{G^2\log\frac{1}{\delta'}}} + 58\right).$$

Suppose that we have $A_0 + \Delta < \Lambda$ *and* $\frac{T_0(T_0-1)}{T_1(T_1-1)}\Lambda < A_1$*. Then,*

$$\Pr\left[\exists T_0+1 \le t \le T_1\,,\ X_t > \frac{T_0(T_0-1)}{t(t-1)}\cdot\Lambda \;\middle|\; X_{T_0} < A_0\right] < \delta'.$$

In particular, the above implies $\Pr[X_{T_1} > A_1 \mid X_{T_0} \le A_0] \le \delta'$*.*

- *If* $\eta_t = \eta = \gamma/(2\lambda)$ *for some* $\gamma > 0$*, then for every* $A_0 > A_1 > 0$*,* $\delta' > 0$*,* $\Lambda > 0$*, and* $1 \le T_0 < T_1 \in \mathbb{N}$*, let*

$$\Delta = \frac{G^2\gamma\log\frac{1}{\delta'}}{\lambda^2(1-\gamma)^{T_1-T_0}}\left(\sqrt{\frac{72\Lambda\lambda^2}{G^2\gamma\log\frac{1}{\delta'}}}+27\right).$$

Suppose that we have $A_0 + \Delta < \Lambda$ *and* $(1-\gamma)^{T_1-T_0}\Lambda < A_1$*. Then,*

$$\Pr\left[\exists T_0+1 \le t \le T_1\,,\ X_t > (1-\gamma)^{t-T_0}\Lambda \;\middle|\; X_{T_0} < A_0\right] < \delta'\,.$$

In particular, the above implies $\Pr[X_{T_1} > A_1 \mid X_{T_0} \le A_0] \le \delta'$*.*

Proof of Lemma C.5.

- ($\eta_t = 1/\lambda t$) Given the moment profile for $\{X_t\}$ in Lemma C.2, we apply Proposition 2.3 and get the following deviation using Lemma A.3.

$$\begin{aligned}
&2\max\left\{\sqrt{\sum_{t'=T_0+1}^{T_1}\sigma_{T_0}^2(t',\Lambda)\log\frac{1}{\delta'}}, 2\max_{T_0+1\le t'\le T_1} B_{T_0}(t',\Lambda)\log\frac{1}{\delta'}\right\} + \sum_{t'=T_0+1}^{T_1}\mu_{T_0}(t',\Lambda)\\
&= \max\left\{2\sqrt{\sum_{t'=T_0+1}^{T_1}\frac{G^2t'^2}{\lambda^2T_0^4}\left(73\Lambda+\frac{3G^2}{\lambda^2t'^2}\right)\log\frac{1}{\delta'}}, 52\frac{G^2T_1}{\lambda^2T_0^2}\log\frac{1}{\delta'}\right\} + \sum_{t'=T_0+1}^{T_1}\frac{2G^2}{\lambda^2T_0^2}\\
&\le \sqrt{\frac{292G^2T_1^3\Lambda\log\frac{1}{\delta'}}{\lambda^2T_0^4}} + \frac{G^2\sqrt{12\log\frac{1}{\delta'}}}{\lambda^2T_0^2} + 52\frac{G^2T_1}{\lambda^2T_0^2}\log\frac{1}{\delta'} + \frac{2G^2T_1}{\lambda^2T_0^2}\\
&\le \frac{G^2T_1\log\frac{1}{\delta'}}{\lambda^2T_0^2}\left(\sqrt{\frac{292T_1\Lambda\lambda^2}{G^2\log\frac{1}{\delta'}}}+58\right) = \Delta\,.
\end{aligned}$$

  Next, by Proposition 2.3 we get the desiring improvement inequalities.

- ($\eta_t = \eta = \gamma/(2\lambda)$) Due to the choice of learning rate, we have

$$\sum_{t=T_0+1}^{T_1}\prod_{t'=T_0+1}^{t}\frac{1}{(1-2\eta_{t'}\lambda)} = \frac{(1-\gamma)^{T_0-T_1}-1}{\gamma}\,.$$

  Given the moment profile for $\{X_t\}$ in Lemma C.2, we apply Proposition 2.3 and get the following deviation

using Lemma A.3.

$$
\begin{aligned}
&2\max\left\{\sqrt{\sum_{t'=T_0+1}^{T_1}\sigma_{T_0}^2(t',\Lambda)\log\frac{1}{\delta'}},\,2\max_{T_0+1\le t'\le T_1}B_{T_0}(t',\Lambda)\log\frac{1}{\delta'}\right\}+\sum_{t'=T_0+1}^{T_1}\mu_{T_0}(t',\Lambda)\\
&=\max\left\{2\sqrt{\sum_{t'=T_0+1}^{T_1}\frac{G^2\gamma^2\left(18\Lambda+\frac{G^2\gamma^2}{8\lambda^2}\right)\log\frac{1}{\delta'}}{\lambda^2(1-\gamma)^{2(t'-T_0)}}},\,\frac{25G^2\gamma\log\frac{1}{\delta'}}{\lambda^2(1-\gamma)^{T_1-T_0}}\right\}+\sum_{t'=T_0+1}^{T_1}\frac{G^2\gamma^2}{4\lambda^2(1-\gamma)^{t'-T_0}}\\
&\le\sqrt{\frac{72G^2\gamma\Lambda\log\frac{1}{\delta'}}{\lambda^2(1-\gamma)^{2(T_1-T_0)}}}+\frac{G^2\gamma\sqrt{\gamma\log\frac{1}{\delta'}}}{\lambda^2(1-\gamma)^{T_1-T_0}}+\frac{25G^2\gamma\log\frac{1}{\delta'}}{\lambda^2(1-\gamma)^{T_1-T_0}}+\frac{G^2\gamma}{4\lambda^2(1-\gamma)^{T_1-T_0}}\\
&\le\frac{G^2\gamma\log\frac{1}{\delta'}}{\lambda^2(1-\gamma)^{T_1-T_0}}\left(\sqrt{\frac{72\Lambda\lambda^2}{G^2\gamma\log\frac{1}{\delta'}}}+27\right)=\Delta\,.
\end{aligned}
$$

Next, by Proposition 2.3 we get the desiring improvement inequalities.

□

C.3 Interval analysis

C.3.1 Learning rate $\eta_t=\frac{1}{\lambda t}$

As a warm up, here we first improve the state of art *weak* uniform convergence bound for SGD with $\eta_t=\frac{1}{\lambda t}$ to *strong* uniform convergence bound. See Definition A.8 for the formal definitions of the convergence guarantees.

Proof of item 1 in Theorem C.1 when $\eta_t=1/\lambda t$. Let $\ell=\infty$ and let

$$t_0=0,\ t_1=73,\ t_i=2t_{i-1},\ \forall i\ge 2\,.$$

For $i\ge 1$, let

$$\delta_i=\frac{\delta}{2i^2},\ a_0=\frac{5000G^2\log\frac{1}{\delta}}{\lambda^2 t_1},\ a_i=\frac{5000G^2\log\frac{1}{\delta_i}}{\lambda^2 t_i},\text{ and }\Lambda_i=2a_{i-1}\,.$$

Now, for each $i\in\mathbb{N}$, we invoke Lemma C.5 with $A_0=a_{i-1}$, $A_1=a_i$, $T_0=t_{i-1}$, $T_1=t_i$, and $\delta'=\delta_i$. Let us verify the two conditions. First, we verify the pull out condition as follows.

$$
\begin{aligned}
a_{i-1}+\Delta_i&=a_{i-1}+\frac{G^2t_i\log\frac{1}{\delta_i}}{\lambda^2t_{i-1}^2}\left(\sqrt{\frac{292t_{i-1}\Lambda_i\lambda^2}{G^2\log\frac{1}{\delta_i}}}+58\right)\\
&=a_{i-1}+\frac{2\sqrt{292\cdot5000\cdot4}a_{i-1}}{5000}+\frac{58a_{i-1}}{5000}<2a_{i-1}<\Lambda_i\,.
\end{aligned}
$$

Next, we verify the improvement condition $\prod_{t=t_{i-1}+1}^{t_i}(1-2\eta_t\lambda)(a_{i-1}+\Delta_i)<a_i$ as follows.

$$\frac{t_{i-1}(t_{i-1}-1)}{t_i(t_i-1)}\cdot 2a_{i-1}<\frac{2a_{i-1}}{4}=\frac{a_i}{2}\,.$$

As $\Lambda_i\le\frac{10000G^2(\log(1/\delta)+\log\log 2t^2)}{\lambda^2 t}$ for every $t_{i-1}+1\le t\le t_i$ and $\log\log 2t^2\le 2\log\log(t+2)$, by Lemma C.5, this implies that

$$\Pr\left[\exists t_{i-1}+1\le t\le t_i\,,\ X_t>\frac{10000G^2\left(\log\frac{1}{\delta}+2\log\log(t+2)\right)}{\lambda^2 t}\,\middle|\,X_{t_{i-1}}\le a_{i-1}\right]<\delta_i\,.$$

Also, since $X_t \le \frac{4G^2}{\lambda^2}$ for all $t \le t_1$ and $\sum_{i=1}^{\infty} \delta_i \le \delta$, by union bound we have

$$\Pr\left[\exists t \in \mathbb{N},\ X_t > \frac{10000G^2\left(\log\frac{1}{\delta} + 2\log\log(t+2)\right)}{\lambda^2 t}\right] < \delta\,.$$

□

C.3.2 Learning rate $\eta_t = \gamma/(2\lambda)$

The goal of this subsubsection is to prove the second item of Theorem C.1. In order to improve to the $\log\log\log$ region, we need to improve polynomially at each interval. This is specified in the following lemma.

Lemma C.6 (Moving from a to $a\left(\frac{\epsilon}{4a}\right)^{\frac{1}{4}}$ when $a \ge \epsilon$)**.** *Let $\{X_t\}$ be the stochastic process described in Equation 3.1 with learning rate $\eta_t = \eta = \gamma/(2\lambda)$ and $\delta' > 0$. Let $a \ge \epsilon, T_0, T_1 \in \mathbb{N}, \gamma \in (0,1)$. If*

$$T_1 - T_0 = \left\lceil \frac{\log(\left(\frac{\epsilon}{4a}\right)^{\frac{1}{4}}/2)}{\log(1-\gamma)} \right\rceil, \gamma = \frac{\lambda^2\epsilon^{\frac{3}{4}}a^{\frac{1}{4}}\log\frac{4a}{\epsilon}}{4608G^2\log\frac{1}{\delta'}}$$

then we have

$$\Pr\left[\exists T_0+1 \le t \le T_1,\ X_t > 2(1-\gamma)^t a,\ X_{T_1} > a\left(\frac{\epsilon}{4a}\right)^{\frac{1}{4}} \,\middle|\, X_{T_0} \le a\right] < \delta'.$$

Proof of Lemma C.6. Let $\Lambda = 2a$ and invoke Lemma C.5. First, we verify the pullout condition as follows.

$$\begin{aligned}
&a + \frac{G^2\gamma\log\frac{1}{\delta'}}{\lambda^2(1-\gamma)^{T_1-T_0}}\left(\sqrt{\frac{72\Lambda\lambda^2}{G^2\gamma\log\frac{1}{\delta'}}} + 27\right) \\
&= a + \sqrt{\frac{72\cdot 2a\cdot\epsilon^{\frac{3}{4}}a^{\frac{1}{4}}\log\frac{4a}{\epsilon}}{4608\cdot(1-\gamma)^{2(T_1-T_0)}}} + \frac{27\cdot\epsilon^{\frac{3}{4}}a^{\frac{1}{4}}\log\frac{4a}{\epsilon}}{4608\cdot(1-\gamma)^{T_1-T_0}}
\end{aligned}$$

Due to the choice of γ, T_1, T_0, we have $2\cdot(4a)^{\frac{1}{4}}\epsilon^{-\frac{1}{4}} \le (1-\gamma)^{-(T_1-T_0)} \le 4\cdot(4a)^{\frac{1}{4}}\epsilon^{-\frac{1}{4}}$.

$$\le a + \sqrt{\frac{(4a)^{\frac{1}{2}}\cdot 72\cdot 2a\cdot\epsilon^{\frac{3}{4}}a^{\frac{1}{4}}\log\frac{4a}{\epsilon}}{4608\cdot\epsilon^{\frac{1}{2}}}} + \frac{(4a)^{\frac{1}{4}}\cdot 27\cdot\epsilon^{\frac{3}{4}}a^{\frac{1}{4}}\log\frac{4a}{\epsilon}}{4608\cdot\epsilon^{\frac{1}{4}}}$$

Because $a \ge \epsilon$, we have $\frac{1}{4}\log\frac{4a}{\epsilon} \le (\frac{4a}{\epsilon})^{\frac{1}{4}}$ and $\frac{1}{2}\log\frac{4a}{\epsilon} \le (\frac{4a}{\epsilon})^{\frac{1}{2}}$ so the inequality becomes

$$\le a + \sqrt{\frac{a^2}{2}} + \frac{a}{20} < 2a = \Lambda\,.$$

Next, we verify the improvement condition as follows.

$$(1-\gamma)^{T_1-T_0}\Lambda \le \frac{\left(\frac{\epsilon}{4a}\right)^{\frac{1}{4}}}{2}2a = a\left(\frac{\epsilon}{4a}\right)^{\frac{1}{4}}\,.$$

Therefore by Lemma C.5, we have

$$\Pr\left[\exists T_0+1 \le t \le T_1\,,\ X_t > 2(1-\gamma)^t a,\ X_{T_1} > a\left(\frac{\epsilon}{4a}\right)^{\frac{1}{4}} \,\middle|\, X_{T_0} \le a\right] < \delta'\,.$$

□

Now we describe how to set the intervals properly in each region using the improvement analysis instantiated in Lemma C.6. Let $\epsilon, \delta \in (0,1)$, $a_0 = \frac{4G^2}{\lambda^2}$ and $t_0 = 0$. For the i^{th} interval,

$$a_i = a_{i-1}\left(\frac{\epsilon}{4a_{i-1}}\right)^{\frac{1}{4}}, \ \gamma_i = \frac{\lambda^2\epsilon^{\frac{3}{4}}a_{i-1}^{\frac{1}{4}}\log\frac{4a_{i-1}}{\epsilon}}{4608G^2\log\frac{2i^2}{\delta}}, \ t_i = t_{i-1} + \left\lceil \frac{\log\left(\left(\frac{\epsilon}{4a_{i-1}}\right)^{\frac{1}{4}}/2\right)}{\log(1-\gamma_i)} \right\rceil .$$

Now we set the learning rate of the algorithm. Given t, there is an i such that $t_{i-1} < t \le t_i$. Let $\eta_t = \gamma_i/2\lambda$. Now we are ready to prove the main theorem.

Proof of item 2 in Theorem C.1 when using η_t specified above. By Lemma C.6 and union bound, we have

$$\Pr[\exists i \in \mathbb{N}, \ X_{t_i} > a_i] < \sum_{i=1}^{\infty} \frac{\delta}{2i^2} < \delta .$$

Now it remains to determine at which interval i, $a_i \le \epsilon$ and the length of t_i. We can rewrite the recursive relation as

$$a_i = a_{i-1}^{\frac{3}{4}}\left(\frac{\epsilon}{4}\right)^{\frac{1}{4}} .$$

By multiplying $\frac{4}{\epsilon}$ at both side, we have

$$a_i\left(\frac{\epsilon}{4}\right)^{-1} = \left(a_{i-1}\left(\frac{\epsilon}{4}\right)^{-1}\right)^{\frac{3}{4}} .$$

This implies that

$$a_i\left(\frac{\epsilon}{4}\right)^{-1} = \left(a_0\left(\frac{\epsilon}{4}\right)^{-1}\right)^{(\frac{3}{4})^i} = \left(\frac{16G^2}{\epsilon\lambda^2}\right)^{(\frac{3}{4})^i} . \tag{C.7}$$

Therefore when $\ell = \left\lceil \frac{\log(\log\frac{16G^2}{\epsilon\lambda^2}/2)}{\log\frac{4}{3}} \right\rceil$, we have $4 \ge \left(\frac{16G^2}{\epsilon\lambda^2}\right)^{(\frac{3}{4})^\ell} \ge 4^{3/4}$ and thus

$$a_\ell = \left(\frac{16G^2}{\epsilon\lambda^2}\right)^{(\frac{3}{4})^\ell} \cdot \frac{\epsilon}{4} \le 4 \cdot \frac{\epsilon}{4} = \epsilon .$$

Let $T = t_\ell$. The algorithm will output X_T. Finally, we aim to show $T = O(\frac{G^2(\log\frac{1}{\delta} + \log\log\log\frac{G}{\epsilon\lambda})}{\epsilon\lambda^2})$.

$$T = \sum_{i=1}^{\ell} \left\lceil \frac{\log\left(\left(\frac{\epsilon}{4a_{i-1}}\right)^{\frac{1}{4}}/2\right)}{\log(1-\gamma_i)} \right\rceil \le \sum_{i=1}^{\ell} 1 + \frac{2\log\frac{4a_{i-1}}{\epsilon}}{4\gamma_i} \le \sum_{i=1}^{\ell} 1 + \frac{4608G^2\log\frac{2i^2}{\delta}}{2\lambda^2\epsilon^{\frac{3}{4}}a_{i-1}^{\frac{1}{4}}}$$

$$\le \frac{3260G^2\log\frac{2\ell^2}{\delta}}{\epsilon\lambda^2} \sum_{i=1}^{\ell} \left(\frac{\epsilon}{4a_{i-1}}\right)^{\frac{1}{4}} .$$

By Equation C.7, we have

$$\le \frac{3260G^2\log\frac{2\ell^2}{\delta}}{\epsilon\lambda^2} \sum_{i=1}^{\ell} \left(\frac{1}{(16G^2/\epsilon\lambda^2)^{\frac{1}{4}}}\right)^{(\frac{3}{4})^{i-1}} .$$

By the choice of ℓ, we have $(16G^2/\epsilon\lambda^2)^{\frac{1}{4}(3/4)^\ell} \geq 4^{\frac{1}{4}\cdot\frac{3}{4}} \geq 1.29$ and thus

$$\left(\frac{16G^2}{\epsilon\lambda^2}\right)^{\frac{1}{4}(3/4)^i} = \left(\frac{16G^2}{\epsilon\lambda^2}\right)^{\frac{1}{4}(3/4)^\ell(4/3)^{\ell-i}} \geq 1.29^{(4/3)^{\ell-i}}$$

for all $0 \leq i \leq \ell$. That is, by flipping the order of the summation, we have

$$T \leq \frac{3260G^2 \log\frac{2\ell^2}{\delta}}{\epsilon\lambda^2} \sum_{i=1}^{\ell} \left(\frac{1}{1.29}\right)^{(\frac{4}{3})^{i-1}}.$$

Since for $i > 7$, $(4/3)^{i-1} > i - 1$, we have

$$\begin{aligned} &\leq \frac{3260G^2 \log\frac{2\ell^2}{\delta}}{\epsilon\lambda^2} \left(\sum_{i=1}^{7} \left(\frac{1}{1.29}\right)^{(\frac{4}{3})^{i-1}} + \sum_{i=8}^{\ell} \left(\frac{1}{1.29}\right)^{i-1}\right) \\ &\leq \frac{14500G^2 \log\frac{2\ell^2}{\delta}}{\epsilon\lambda^2}. \end{aligned}$$

We also have

$$\log(2\ell^2) = 1 + 2\log\left\lceil \frac{\log(\log\frac{16G^2}{\epsilon\lambda^2}/2)}{\log\frac{4}{3}} \right\rceil \leq 10 + 2\log\log\log\frac{G}{\epsilon\lambda}.$$

So in total we have

$$T \leq \frac{14500G^2\left(\log\frac{1}{\delta} + 2\log\log\log\frac{G}{\lambda\epsilon} + 10\right)}{\epsilon\lambda^2}.$$

□

D Details on k-PCA

Let $\mathcal{D}$ be a distribution over the unit sphere in $\mathbb{R}^n$ and $\Sigma = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$ be its covariance matrix. Given a sequence of i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$ from $\mathcal{D}$, the goal of streaming k-PCA is to output a k dimensional subspace that is close to the top k eigenspace of A using $O(nk)$ space a given $1 \leq k \leq n$.

We analyze the following Oja's algorithm [Oja82] which maintains a matrix $W_t \in \mathbb{R}^{n\times k}$ at time t.

Algorithm 3 Oja's algorithm for streaming k-PCA

Input: Time parameter $T \in \mathbb{N}$, learning rate $\{\eta_t\}_{t\in\mathbb{N}}$, initial matrix $W_0 \in \mathbb{R}^{n\times k}$, and sequence of $\mathbf{x}_1, \ldots, \mathbf{x}_T \sim \mathcal{D}$.

1: **for** $t = 1, \ldots, T$ **do**

2: Let $W_t = (1 - \eta_t \mathbf{x}_t\mathbf{x}_t^\top)W_{t-1}$.

Output: $QR(W_t)$, an orthonormal basis of the column space of W_t.

To measure how well W_t converges to the top k eigenspace, it is standard to use the following objective function.

$$X_t = \|Z^\top QR(W_t)\|_F^2 = \|Z^\top W_t(V^\top W_t)^{-1}\|_F^2 \tag{D.1}$$

where $QR(\cdot)$ stands for the QR decomposition and V (resp. Z) is an orthogonal basis for the eigenspace corresponds to eigenvalues $\lambda_1, \ldots, \lambda_k$ (resp. $\lambda_{k+1}, \ldots, \lambda_n$). Denote $\Sigma = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$, $\Sigma_{\leq k} = V\mathsf{diag}(\lambda_1, \ldots, \lambda_k)V^\top$, and $\Sigma_{>k} = Z\mathsf{diag}(\lambda_{k+1}, \ldots, \lambda_n)Z^T$. The goal is to show that X_t converges to 0 efficiently.

There are two common convergence guarantees for k-PCA. The *local convergence* which starts from a good initialization such that $X_0 \leq 1$ and the *global convergence* where W_0 is randomly chosen. On the other

hand, it is also common to consider the following two eigengap settings: the *gap-dependent setting* which assumes $\mathsf{gap} = \lambda_k - \lambda_{k+1} > 0$ and the *gap-free setting* where the goal is showing that W_t is close to the top eigenspace corresponds to eigenvalue $\geq \lambda_k - \rho$ for some parameter $\rho > 0$. In this paper, since the goal is to demonstrate the power of the proposed framework, we focus on the simplest non-trivial setting: the local convergence for gap-dependent k-PCA. Specifically, we apply the framework and prove the following theorem.

Theorem D.2 (Local convergence for gap-dependent streaming k-PCA)**.** *Let $\mathcal{D}$ be a distribution over the unit sphere in $\mathbb{R}^n$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be the eigenvalues of its covariance matrix, and let W_t be the output of the Oja's algorithm at time $t \in \mathbb{N}$. Let $\lambda = \lambda_1 + \cdots + \lambda_k$ and $\mathsf{gap} = \lambda_k - \lambda_{k+1}$. For every $\delta \in (0,1)$, there exists a setting of the learning rate such that we have the following strong uniform convergence.*

$$\Pr\left[\exists t \in \mathbb{N},\ \|Z^\top W_t (V^\top W_t)^{-1}\|_F^2 > \frac{30000\lambda(\log\frac{1}{\delta} + 2\log\log(t+2))}{\mathsf{gap}^2 t}\right] < \delta\,.$$

Comparison. The convergence rate of the Oja's algorithm is a well-studied problem [AZL17, DSOR15, Sha16]. For the gap-dependent local convergence, the previous state-of-the-art analysis [AZL17] gives $O(\frac{\lambda(\log\delta^{-1}+\log^3 t)}{\mathsf{gap}^2 t})$ convergence rate while in Theorem D.2 we improve to $O(\frac{\lambda(\log\delta^{-1}+\log\log t)}{\mathsf{gap}^2 t})$. Note that there is an information-theoretic lower bound $\Omega(\frac{k\lambda_k}{\mathsf{gap}^2 t})$. See [AZL17] for more comparisons with other previous works as well as other settings.

Section structure. In the rest of this section, we provide the linearization and moment analysis in Section D.1, the improvement analysis in Section D.2, and the interval analysis in Section D.3.

D.1 Moment analysis

Here we provide the details on the moment analysis. The linearization is summarized in Lemma D.3 and the moment profile is given in Lemma D.8. The proof of this step looks relatively lengthy because the objective function X_t has an inverse term $(V^\top W_t)^{-1}$. Conceptually, the proofs are straightforward by properly rearranging the terms and applying matrix inequalities (see Lemma A.7).

Lemma D.3 (Linearization for k-PCA)**.** *For any $t \in \mathbb{N}$, we have*

$$X_t \leq (1 - 2\eta_t \mathsf{gap}) X_{t-1} + N_t$$

with

$$\begin{aligned} N_t = {} & 2\eta_t \cdot \left(-\mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t) + \mathbb{E}[\mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t)] + \mathsf{tr}(Y_{t-1}^\top C_t) - \mathbb{E}[\mathsf{tr}(Y_{t-1}^\top C_t)]\right) \\ & + \frac{2\eta_t^2 a_t}{1+\eta_t a_t} \cdot \left(\mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t) - \mathsf{tr}(Y_{t-1}^\top C_t)\right) + \frac{2\eta_t^2}{(1+\eta_t a_t)^2} \cdot \left(\|Y_{t-1} B_t\|_F^2 + \|C_t\|_F^2\right) \end{aligned}$$

where

$$\begin{aligned} Y_t &= Z^\top W_t (V^\top W_t)^{-1}\,, & a_t &= \mathbf{x}_t^\top W_{t-1} (V^\top W_{t-1})^{-1} V^\top \mathbf{x}_t\,, \\ B_t &= V^\top \mathbf{x}_t \mathbf{x}_t^\top W_{t-1} (V^\top W_{t-1})^{-1}\,, & C_t &= Z^\top \mathbf{x}_t \mathbf{x}_t^\top W_{t-1} (V^\top W_{t-1})^{-1}\,. \end{aligned}$$

Proof of Lemma D.3. First by Sherman-Morisson formula, we have

$$(V^\top W_t)^{-1} = (V^\top W_{t-1})^{-1} - \frac{\eta_t}{1+\eta_t a_t} (V^\top W_{t-1})^{-1} B_t\,.$$

Now we have

$$\begin{aligned} Y_t &= Z^\top W_{t-1} (V^\top W_t)^{-1} + \eta_t Z^\top \mathbf{x}_t \mathbf{x}_t^\top W_{t-1} (V^\top W_t)^{-1} \\ &= Y_{t-1} - \frac{\eta_t}{1+\eta_t a_t} Y_{t-1} B_t + \eta_t C_t - \frac{\eta_t^2}{1+\eta_t a_t} C_t B_t\,. \end{aligned}$$

Because $C_t B_t = a_t C_t$, we have

$$\begin{aligned} &= Y_{t-1} - \frac{\eta_t}{1+\eta_t a_t} Y_{t-1} B_t + \left(\eta_t - \frac{\eta_t^2 a_t}{1+\eta_t a_t}\right) C_t \\ &= Y_{t-1} - \frac{\eta_t}{1+\eta_t a_t} Y_{t-1} B_t + \frac{\eta_t}{1+\eta_t a_t} C_t \,. \end{aligned} \tag{D.4}$$

Next, expand the square of the Frobenius norm of Equation D.4 as follows.

$$\begin{aligned} \|Y_t\|_F^2 = \mathsf{tr}(Y_t^\top Y_t) &= \mathsf{tr}(Y_{t-1}^\top Y_{t-1}) - \frac{2\eta_t}{1+\eta_t a_t} \cdot \big(\mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t) + \mathsf{tr}(Y_{t-1}^\top C_t)\big) \\ &+ \frac{\eta_t^2}{(1+\eta_t a_t)^2} \cdot \big(\|Y_{t-1} B_t\|_F^2 - 2\mathsf{tr}(Y_{t-1}^\top C_t B_t) + \|C_t\|_F^2\big) \,. \end{aligned}$$

By AM-GM inequality for matrix (see Lemma A.7), we have

$$\begin{aligned} &\le \mathsf{tr}(Y_{t-1} Y_{t-1}^\top) - \frac{2\eta_t}{1+\eta_t a_t} \cdot \big(\mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t) + \mathsf{tr}(Y_{t-1}^\top C_t)\big) \\ &+ \frac{2\eta_t^2}{(1+\eta_t a_t)^2} \cdot \big(\|Y_{t-1} B_t\|_F^2 + \|C_t\|_F^2\big) \,. \end{aligned}$$

Plus and minus $2\eta_t \cdot \big(\mathsf{tr}(C_t^\top Y_{t-1}) - \mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t)\big)$, we have

$$\begin{aligned} &\le \mathsf{tr}(Y_{t-1} Y_{t-1}^\top) - 2\eta_t \cdot \big(\mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t) - \mathsf{tr}(Y_{t-1}^\top C_t)\big) \\ &+ \frac{2\eta_t^2 a_t}{1+\eta_t a_t} \cdot \big(\mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t) - \mathsf{tr}(C_t^\top Y_{t-1})\big) + \frac{2\eta_t^2}{(1+\eta_t a_t)^2} \cdot \big(\|Y_{t-1} B_t\|_F^2 + \|C_t\|_F^2\big) \,. \end{aligned} \tag{D.5}$$

Finally, let us calculate the expectation of the second term as follows. Recall that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] = \Sigma$ is the covariance matrix.

$$\begin{aligned} \mathbb{E}[\mathsf{tr}(Y_{t-1}^\top Y_{t-1} B_t) | \mathcal{F}_{t-1}] &= \mathsf{tr}(Y_{t-1}^\top Y_{t-1} V^\top \Sigma W_{t-1} (V^\top W_{t-1})^{-1}) \\ &= \mathsf{tr}(Y_{t-1}^\top Y_{t-1} \Sigma_{\le k} V^\top W_{t-1} (V^\top W_{t-1})^{-1}) \\ &= \mathsf{tr}(Y_{t-1}^\top Y_{t-1} \Sigma_{\le k}) \ge \lambda_k \mathsf{tr}(Y_{t-1}^\top Y_{t-1}) \end{aligned} \tag{D.6}$$

and

$$\begin{aligned} \mathbb{E}[\mathsf{tr}(Y_{t-1}^\top C_t) | \mathcal{F}_{t-1}] &= \mathsf{tr}(Y_{t-1}^\top Z^\top \Sigma W_{t-1} (V^\top W_{t-1})^{-1}) \\ &= \mathsf{tr}(Y_{t-1}^\top \Sigma_{>k} Z^\top W_{t-1} (V^\top W_{t-1})^{-1}) \\ &= \mathsf{tr}(Y_{t-1}^\top \Sigma_{>k} Y_{t-1}) \le \lambda_{k+1} \mathsf{tr}(Y_{t-1}^\top Y_{t-1}) \,. \end{aligned} \tag{D.7}$$

Combining Equation D.5, Equation D.6, and Equation D.7, as $X_t = \|Y_t\|_F^2$, we have

$$X_t \le (1 - 2\eta_t \mathsf{gap}) \cdot X_{t-1} + N_t$$

as desired. □

Given the linearization in Lemma D.3, we let $M_t = \sum_{t'=T_0+1} \prod_{t''=T_0+1}^{t'} (1 - 2\eta_{t'} \mathsf{gap})^{-1} N_{t'}$ so that $X_t = \prod_{t'=T_0+1}^{t} (1 - 2\eta_{t'} \mathsf{gap}) \cdot (X_{T_0} + M_t)$. The following lemma provides the moment profile for $\{X_t\}$ under this linearization.

Lemma D.8 (Moment profile for k-PCA)**.** *Consider the setting in Theorem D.2 with learning rate $\eta_t = \eta = \gamma/(2\mathsf{gap})$ for some $\gamma > 0$. Let $0 < \Lambda \le 1$ and τ is the stopping time for the event $\{X_t > \Lambda\}$. For every $T_0 + 1 \le t \le T$, the following following functions $(B_{T_0}, \mu_{T_0}, \sigma_{T_0}^2)$ form a moment profile for $\{X_t\}$, Λ, and T_0.*

- *(Bounded difference)* $|M_{t\wedge\tau} - M_{(t-1)\wedge\tau}| \le B_{T_0}(t,\Lambda) = \eta_t \cdot \frac{40}{(1-\gamma)^{t-T_0}}$.
- *(Conditional expectation)* $\left|\mathbb{E}\left[M_{t\wedge\tau} - M_{(t-1)\wedge\tau} \,|\, \mathcal{F}_{t-1}\right]\right| \le \mu_{T_0}(t,\Lambda) = \eta_t^2\lambda \cdot \frac{56}{(1-\gamma)^{t-T_0}}$.
- *(Conditional variance)* $\left|\mathsf{Var}\left[M_{t\wedge\tau} - M_{(t-1)\wedge\tau} \,|\, \mathcal{F}_{t-1}\right]\right| \le \sigma_{T_0}^2(t,\Lambda) = \eta_t^2\lambda \cdot \frac{1136\Lambda + 512\eta_t^2}{(1-\gamma)^{2(t-T_0)}}$.

Proof of Lemma D.8. Let us start with three useful facts we are going to heavily apply throughout the proof. First, $I = VV^\top + ZZ^\top$ because V and Z form an orthonormal eigenbasis for the covariance matrix. Second, the following show that $\mathbf{1}_{\tau\ge t}a_t$ is small almost surely.

$$\begin{aligned}|\mathbf{1}_{\tau\ge t}a_t| &= |\mathbf{1}_{\tau\ge t}\mathbf{x}_t^\top(VV^\top + ZZ^\top)W_{t-1}(V^\top W_{t-1})^{-1}V^\top\mathbf{x}_t| \\ &\le \|V^\top\mathbf{x}_t\|_2^2 + |\mathbf{1}_{\tau\ge t}\mathbf{x}_t^\top ZY_{t-1}V^\top\mathbf{x}_t|\,.\end{aligned}$$

Because $\|\mathbf{x}_t\|_2 = 1$, we have

$$\le 1 + \|\mathbf{1}_{\tau\ge t}Y_{t-1}\|_F \le 2\,.$$

Third, let $A_t = x_t^\top W_{t-1}(V^\top W_{t-1})^{-1}$. We have

$$\begin{aligned}\|\mathbf{1}_{\tau\ge t}x_t^T A_{t-1}\|_2 &= \|\mathbf{1}_{\tau\ge t}x_t^T(VV^T + ZZ^T)A_{t-1}\|_2 \\ &\le \|x_t^T V\|_2 + \|\mathbf{1}_{\tau\ge t}x_t^T ZX_{t-1}\|_2 \\ &\le 1 + \|\mathbf{1}_{\tau\ge t}X_{t-1}\|_F < 2\,.\end{aligned} \tag{D.9}$$

Thus, as $\eta_t \le 1/4$, we have

$$\mathbf{1}_{\tau\ge t}\frac{2\eta_t^2 a_t}{1+\eta_t a_t} \le 8\eta_t^2 \quad\text{and}\quad \mathbf{1}_{\tau\ge t}\frac{2\eta_t^2}{(1+\eta_t a_t)^2} \le 8\eta_t^2\,.$$

- (Bounded difference) First, observe that $\|\mathbf{1}_{\tau\ge t}B_t\|_F$ and $\|\mathbf{1}_{\tau\ge t}C_t\|_F$ are small almost surely. Concretely, by Cauchy-Schwarz inequality and the fact that B_t, C_t are rank 1 matrix, we have

  $$\begin{aligned}\|\mathbf{1}_{\tau\ge t}B_t\|_F = \|\mathbf{1}_{\tau\ge t}B_t\| = \|V^\top x_t\|_2\|x_t^\top A_{t-1}\|_2 < 2\,,\\ \|\mathbf{1}_{\tau\ge t}C_t\|_F = \|\mathbf{1}_{\tau\ge t}C_t\| = \|Z^\top x_t\|_2\|x_t^\top A_{t-1}\|_2 < 2\,.\end{aligned}$$

  Thus, by matrix Cauchy-Schwarz inequality (see Lemma A.7), we have

  $$|\mathbf{1}_{\tau\ge t}\mathsf{tr}(Y_{t-1}^\top C_t)| \le \|\mathbf{1}_{\tau\ge t}Y_{t-1}\|_F \cdot \|\mathbf{1}_{\tau\ge t}C_t\|_F \le 2\sqrt{\Lambda}\,.$$

  By matrix Holder's inequality (see Lemma A.7), the fact that $\|AA^\top\|_1 = \|A\|_F^2, \|A\|_\infty = \|A\|$ and for rank 1 matrix $\|AA^T\| = \|A\|^2 = \|A\|_F^2$, we have

  $$\begin{aligned}&|\mathbf{1}_{\tau\ge t}\mathsf{tr}(Y_{t-1}^\top Y_{t-1}B_t)| \le \mathbf{1}_{\tau\ge t}\|Y_{t-1}^\top Y_{t-1}\|_1\|B_t\|_\infty = \mathbf{1}_{\tau\ge t}\|Y_{t-1}\|_F^2\|B_t\|_F \le 2\Lambda\,, \text{ and}\\ &\|\mathbf{1}_{\tau\ge t}Y_{t-1}B_t\|_F^2 \le \mathbf{1}_{\tau\ge t}\|Y_{t-1}Y_{t-1}^\top\|_1\|B_tB_t^\top\|_\infty \le \mathbf{1}_{\tau\ge t}\|Y_{t-1}\|_F^2\|B_t\|_F^2 \le 4\Lambda\,.\end{aligned}$$

  By triangle inequality and $\eta_t \le 1/4$, we have

  $$|M_{t\wedge\tau} - M_{(t-1)\wedge\tau}| = \frac{|\mathbf{1}_{\tau\ge t}N_t|}{(1-\gamma)^{t-T_0}} \le \eta_t \cdot \frac{20\Lambda + 12\sqrt{\Lambda} + 32\eta_t}{(1-\gamma)^{t-T_0}} \le \cdot\frac{40\eta_t}{(1-\gamma)^{t-T_0}}\,.$$

- (Conditional expectation) By reusing the inequalities in the calculation of bounded difference, we have

$$\begin{aligned}&\left|\mathbf{1}_{\tau\geq t}\mathbb{E}\left[\frac{2\eta_t^2 a_t}{1+\eta_t a_t}\cdot\mathsf{tr}(Y_{t-1}^\top Y_{t-1}B_t)+\frac{2\eta_t^2}{(1+\eta_t a_t)^2}\cdot(\|Y_{t-1}B_t\|_F^2+\|C_t\|_F^2)\,|\,\mathcal{F}_{t-1}\right]\right|\\&\leq\eta_t^2\Lambda\cdot\left(4\mathbb{E}\left[\|\mathbf{1}_{\tau\geq t}a_tB_t\|_F\,|\,\mathcal{F}_{t-1}\right]+8\mathbb{E}\left[\|\mathbf{1}_{\tau\geq t}B_t\|_F^2\,|\,\mathcal{F}_{t-1}\right]+8\mathbb{E}\left[\|\mathbf{1}_{\tau\geq t}C_t\|_F^2\,|\,\mathcal{F}_{t-1}\right]\right).\end{aligned}$$

  Note that $a_t^2=\mathsf{tr}(B_t^\top B_t)$ and thus $|a_t|=\|B_t\|_F$. Namely, $\|a_tB_t\|_F=\|B_t\|_F^2$. Also, by matrix inequalities (see Lemma A.7), we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{1}_{\tau\geq t}B_t\|_F^2\,|\,\mathcal{F}_{t-1}]&\leq\mathbb{E}[\mathbf{1}_{\tau\geq t}\|x_t^\top A_{t-1}\|_F^2\|V^\top x_t\|_F^2\,|\,\mathcal{F}_{t-1}]\\&\leq 2\mathbb{E}[\|V^\top x_t\|_F^2\,|\,\mathcal{F}_{t-1}]=2\lambda\end{aligned}$$

  and

$$\begin{aligned}\mathbb{E}[\|\mathbf{1}_{\tau\geq t}C_t\|_F^2|\mathcal{F}_{t-1}]&\leq\mathbb{E}[\mathbf{1}_{\tau\geq t}\|x_tA_{t-1}\|_F^2\|Z^\top x\|_F^2|\mathcal{F}_{t-1}]\\&\leq\mathbb{E}[\mathbf{1}_{\tau\geq t}\|x_tA_{t-1}\|_F^2|\mathcal{F}_{t-1}]\\&=\mathbf{1}_{\tau\geq t}Tr(A_{t-1}^T\Sigma A_{t-1})\\&=Tr(A_{t-1}^T\Sigma_{\leq k}A_{t-1})+\mathbf{1}_{\tau\geq t}Tr(A_{t-1}^T\Sigma_{>k}A_{t-1})\\&=\lambda+\mathbf{1}_{\tau\geq t}\lambda_{k+1}X_{t-1}\leq 2\lambda\,.\end{aligned}$$

  As for the $\mathsf{tr}(Y_{t-1}^\top C_t)$ term, use the identity $I=VV^\top+ZZ^\top$, we have

$$\begin{aligned}|\mathsf{tr}(Y_{t-1}^\top C_t)|&=|\mathsf{tr}(Y_{t-1}^\top Z^\top\mathbf{x}_t\mathbf{x}_t^\top W_{t-1}(V^\top W_{t-1})^{-1})|\\&=|\mathsf{tr}(Y_{t-1}^\top Z^\top\mathbf{x}_t\mathbf{x}_t^\top(VV^\top+ZZ^\top)W_{t-1}(V^\top W_{t-1})^{-1})|\\&\leq|\mathsf{tr}(Y_{t-1}^\top Z^\top\mathbf{x}_t\mathbf{x}_t^\top V)|+|\mathsf{tr}(Y_{t-1}^\top Z^\top\mathbf{x}_t\mathbf{x}_t^\top ZY_{t-1})|\,.\end{aligned}$$

  By the matrix AM-GM inequality, we have

$$\leq\frac{1}{2}\|V^\top\mathbf{x}_t\|_2^2+\frac{3}{2}\mathsf{tr}(Y_{t-1}^\top Z^\top\mathbf{x}_t\mathbf{x}_t^\top ZY_{t-1})\,.$$

  Thus,

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\tau\geq t}\left|\frac{2\eta_t^2a_t}{1+\eta_ta_t}\mathsf{tr}(Y_{t-1}^\top C_t)\right|\,|\,\mathcal{F}_{t-1}\right]&\leq\mathbf{1}_{\tau\geq t}4\eta_t^2\cdot\mathbb{E}\left[\|V^\top\mathbf{x}_t\|_2^2+3\mathsf{tr}(Y_{t-1}^\top Z^\top\mathbf{x}_t\mathbf{x}_t^\top ZY_{t-1})\,|\,\mathcal{F}_{t-1}\right]\\&\leq\mathbf{1}_{\tau\geq t}4\eta_t^2\cdot(\lambda+3\lambda_{k+1}X_{t-1})\\&\leq 4\eta_t^2\cdot(\lambda+3\lambda_{k+1}\Lambda)\,.\end{aligned}$$

  To sum up, we have

$$\mathbb{E}[|M_{t\wedge\tau}-M_{(t-1)\wedge\tau}|\,|\,\mathcal{F}_{t-1}]=\frac{\mathbb{E}[|\mathbf{1}_{\tau\geq t}N_t|\,|\,\mathcal{F}_{t-1}]}{(1-\gamma)^{t-T_0}}\leq\eta_t^2\lambda\cdot\frac{52\Lambda+4}{(1-\gamma)^{t-T_0}}\leq\eta_t^2\lambda\cdot\frac{56}{(1-\gamma)^{t-T_0}}\,.$$

- (Conditional variance) Let us start with a rough estimation as follows.

$$\begin{aligned}\mathsf{Var}[|\mathbf{1}_{\tau\geq t}N_t|\,|\,\mathcal{F}_{t-1}]\leq\eta_t^2\mathbf{1}_{\tau\geq t}\cdot\mathbb{E}\left[216\mathsf{tr}(Y_{t-1}^\top Y_{t-1}B_t)^2+216\mathsf{tr}(Y_{t-1}^\top C_t)^2\,|\,\mathcal{F}_{t-1}\right]\\+\eta_t^2\mathbf{1}_{\tau\geq t}\cdot\mathbb{E}\left[96\eta_t^2\|Y_{t-1}B_t\|_F^2+256\eta_t^4\|C_t\|_F^2\,|\,\mathcal{F}_{t-1}\right].\end{aligned}$$

  By reusing the previous calculation, we have

$$\begin{aligned}&\mathbb{E}[\mathbf{1}_{\tau\geq t}\mathsf{tr}(Y_{t-1}^\top Y_{t-1}B_t)^2\,|\,\mathcal{F}_{t-1}]\leq\mathbb{E}[\mathbf{1}_{\tau\geq t}\|B_t\|_F^2\|Y_{t-1}\|_F^4\,|\,\mathcal{F}_{t-1}]\leq 2\lambda\Lambda^2\,,\\&\mathbb{E}[\mathbf{1}_{\tau\geq t}\mathsf{tr}(Y_{t-1}^\top C_t)^2\,|\,\mathcal{F}_{t-1}]\leq\mathbb{E}[\mathbf{1}_{\tau\geq t}\|Y_{t-1}\|_F^2\|C_t\|_F^2\,|\,\mathcal{F}_{t-1}]\leq 2\lambda\Lambda\,,\\&\mathbb{E}[\mathbf{1}_{\tau\geq t}\|Y_{t-1}B_t\|_F^2\,|\,\mathcal{F}_{t-1}]\leq\mathbb{E}[\mathbf{1}_{\tau\geq t}\|Y_{t-1}\|_F^2\|B_t\|_F^2\,|\,\mathcal{F}_{t-1}]\leq 2\lambda\Lambda\,,\text{ and}\\&\mathbb{E}[\mathbf{1}_{\tau\geq t}\|C_t\|_F^2\,|\,\mathcal{F}_{t-1}]\leq 2\lambda\,.\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathsf{Var}[\|M_{t\wedge\tau} - M_{(t-1)\wedge\tau}\| \,|\, \mathcal{F}_{t-1}] &= \frac{\mathsf{Var}[\|\mathbf{1}_{\tau\geq t}N_t\| \,|\, \mathcal{F}_{t-1}]}{(1-\gamma)^{2(t-T_0)}} \leq \eta_t^2\lambda \cdot \frac{512\Lambda^2 + 624\Lambda + 512\eta_t^2}{(1-\gamma)^{2(t-T_0)}} \\ &\leq \eta_t^2\lambda \cdot \frac{1136\Lambda + 512\eta_t^2}{(1-\gamma)^{2(t-T_0)}}\,.\end{aligned}$$

□

D.2 Improvement analysis

Lemma D.10 (Improvement analysis for k-PCA)**.** *Let $\{X_t\}$ be the stochastic process described in Equation D.1. For every $A_0 > A_1 > 0$, $\delta' > 0$, $0 < \Lambda \leq 1$, and $1 \leq T_0 < T_1 \in \mathbb{N}$, let*

$$\Delta = \frac{\gamma\lambda\log\frac{1}{\delta'}}{\mathit{gap}^2(1-\gamma)^{T_1-T_0}}\left(\sqrt{\frac{568\mathit{gap}^2\Lambda}{\gamma\lambda\log\frac{1}{\delta'}}} + \sqrt{\frac{128\mathit{gap}}{\gamma\lambda\log\frac{1}{\delta'}}} + 94\right).$$

Suppose that we have $A_0 + \Delta < \Lambda$ and $(1-\gamma)^{T_1-T_0}\Lambda < A_1$. Then,

$$\Pr\left[\exists T_0+1 \leq t \leq T_1\,,\ X_t > (1-\gamma)^{t-T_0}\Lambda \;\middle|\; X_{T_0} \leq A_0\right] < \delta'\,.$$

In particular, the above implies $\Pr[X_{T_1} > A_1 \mid X_{T_0} \leq A_0] \leq \delta'$.

Proof of Lemma D.10. Given the moment profile $(B_{T_0}, \mu_{T_0}, \sigma^2_{T_0})$ in Lemma D.8, we apply Proposition 2.3 and get the following bound for the deviation.

$$\begin{aligned}&2\max\left\{\sqrt{\sum_{t=T_0+1}^{T_1}\sigma^2_{T_0}(t,\Lambda)\log\frac{1}{\delta'}}, 2\max_{T_0+1\leq t\leq T_1} B_{T_0}(t,\Lambda)\log\frac{1}{\delta'}\right\} + \sum_{t'=T_0+1}^{T_1}\mu_{T_0}(t,\Lambda) \\ =\ &\max\left\{2\sqrt{\sum_{t=T_0+1}^{T_1}\eta_t^2\lambda\frac{1136\Lambda+512\eta_t^2}{(1-\gamma)^{2(t-T_0)}}\log\frac{1}{\delta'}}, \eta_t\frac{160}{(1-\gamma)^{T_1-T_0}}\log\frac{1}{\delta'}\right\} + \sum_{t=T_0+1}^{T_1}\eta_t^2\lambda\frac{56}{(1-\gamma)^{t-T_0}}\,.\end{aligned}$$

By geometric series, we have

$$\leq 2\sqrt{\eta_t^2\lambda\frac{1136\Lambda+512\eta_t^2}{\gamma(1-\gamma)^{2(T_1-T_0)}}\log\frac{1}{\delta'}} + \eta_t\frac{160}{(1-\gamma)^{T_1-T_0}}\log\frac{1}{\delta'} + \eta_t^2\lambda\frac{56}{\gamma(1-\gamma)^{T_1-T_0}}\,.$$

Since $\eta_t = \gamma/(2\mathsf{gap}) \leq \frac{1}{4}$ and $\Lambda \leq 1$, we have

$$\begin{aligned}&\leq 2\sqrt{\gamma\lambda\frac{142\Lambda+32\gamma/\mathsf{gap}}{\mathsf{gap}^2(1-\gamma)^{2(T_1-T_0)}}\log\frac{1}{\delta'}} + \frac{80\gamma}{\mathsf{gap}(1-\gamma)^{T_1-T_0}}\log\frac{1}{\delta'} + \gamma\lambda\frac{14}{\mathsf{gap}^2(1-\gamma)^{T_1-T_0}} \\ &\leq 2\sqrt{\gamma\lambda\frac{142\Lambda}{\mathsf{gap}^2(1-\gamma)^{2(T_1-T_0)}}\log\frac{1}{\delta'}} + 2\sqrt{\frac{32\gamma^2\lambda}{\mathsf{gap}^3(1-\gamma)^{2(T_1-T_0)}}\log\frac{1}{\delta'}} + \gamma\lambda\frac{94}{\mathsf{gap}^2(1-\gamma)^{T_1-T_0}}\log\frac{1}{\delta'} \\ &\leq \frac{\gamma\lambda\log\frac{1}{\delta'}}{\mathsf{gap}^2(1-\gamma)^{T_1-T_0}}\left(\sqrt{\frac{568\mathsf{gap}^2\Lambda}{\gamma\lambda\log\frac{1}{\delta'}}} + \sqrt{\frac{128\mathsf{gap}}{\lambda\log\frac{1}{\delta'}}} + 94\right).\end{aligned}$$

Next, by Proposition 2.3 we get the desiring improvement inequalities. □

D.3 Interval analysis

Finally, we perform an interval analysis and complete the proof of Theorem D.2.

Proof of Theorem D.2. Let $\ell = \infty$ and let $t_0 = 0, a_0 = 1$. For each $i \geq 1$, let

$$\delta_i = \frac{\delta}{2i^2},\ a_i = 2^{-i},\ \gamma_i = \frac{a_{i-1}\mathsf{gap}^2}{29500\lambda \log\frac{1}{\delta_i}},\ t_i = t_{i-1} + \left\lceil \frac{-2}{\log(1-\gamma_i)} \right\rceil,\ \Lambda_i = 2a_{i-1}\,,$$

and $\eta_t = \gamma_i/2\mathsf{gap}$ for every $t_{i-1}+1 \leq t \leq t_i$. Let $T = t_\ell$. Observed that due to the choice of the parameters we have $a_\ell \leq \epsilon$ and $1/5 \leq (1-\gamma_i)^{t_i - t_{i-1}} \leq 1/4$. Now, for each $i = 1, 2, \ldots, \ell$, we invoke Lemma D.10 with $A_0 = a_{i-1}$, $A_1 = a_i$, $T_0 = t_{i-1}$, $T_1 = t_i$, and $\delta' = \delta_i$. Let us verify the two conditions. First, we verify the pull out condition as follows.

$$\begin{aligned}
a_{i-1} + \Delta_i &= a_{i-1} + \frac{\gamma_i \lambda \log\frac{1}{\delta_i}}{\mathsf{gap}^2(1-\gamma_i)^{t_i - t_{i-1}}}\left(\sqrt{\frac{568\mathsf{gap}^2\Lambda_i}{\gamma_i\lambda\log\frac{1}{\delta_i}}} + \sqrt{\frac{128\mathsf{gap}}{\lambda\log\frac{1}{\delta_i}}} + 94\right) \\
&\leq a_{i-1} + \frac{5a_{i-1}}{29500}\cdot\left(\sqrt{1126\times 29500} + \sqrt{128} + 94\right) \\
&< 2a_{i-1} < \Lambda_i\,.
\end{aligned}$$

For the improvement condition, we have

$$(1-\gamma)^{t_i - t_{i-1}} \cdot 2a_{i-1} < \frac{2a_{i-1}}{4} = a_i\,.$$

By Lemma D.10 and union bounding over the intervals, we have

$$\Pr[\exists i \in \mathbb{N},\ t_{i-1}+1 \leq t \leq t_i \text{ s.t. } X_t > \Lambda_i] < \sum_{i=1}^{\infty}\delta_i \leq \delta\,.$$

To have an explicit upper bound for the convergence rate, note that

$$\left\lceil \frac{-2}{\log(1-\gamma_i)} \right\rceil \leq 1 + \frac{2}{\gamma_i} = \frac{60000\lambda(\log\frac{1}{\delta} + \log 2i^2)\cdot 2^i}{\mathsf{gap}^2}\,.$$

So $t_i \leq \frac{60000\lambda(\log\frac{1}{\delta}+\log 2i^2)2^{i+1}}{\mathsf{gap}^2}$. Now, for every $t \in \mathbb{N}$, let $i = \left\lceil \log \frac{t\mathsf{gap}^2}{120000\lambda(\log\frac{1}{\delta}+2\log\log(t+2))} \right\rceil$. We have $t \geq t_{i-1}$. We also have $\Lambda_i = 2a_{i-1} = 2^{-i-2} \leq \frac{30000\lambda(\log\frac{1}{\delta}+2\log\log(t+2))}{\mathsf{gap}^2 t}$. Finally, the following strong uniform convergence holds.

$$\Pr\left[\exists t \in \mathbb{N},\ X_t > \frac{30000\lambda(\log\frac{1}{\delta} + 2\log\log(t+2))}{\mathsf{gap}^2 t}\right] < \delta\,.$$

□

E Details on Solving Linear Bandit with SGD Updates

In this subsection, we study linear bandit with SGD dynamic. In stochastic linear bandit, there is a true parameter $\theta_* \in B(0, L_*) \subseteq \mathbb{R}^n$ and at each time step t the agent is presented with a decision set $D_t \subseteq B(0,L) \subseteq \mathbb{R}^n$. The agent chooses an action $x_t \in D_t$ and subsequently, the agent observe the reward

$$y_t = \theta_*^\top x_t + \epsilon_t\,.$$

where $|\epsilon_t| \leq 1$ and $\mathbb{E}[\epsilon_t | x_{1:t}, \epsilon_{1:t-1}] = 0$. We make the bounded assumption of noise term only to simplify the presentation and the sub-Gaussian case can be handled by our framework similarly. The full protocol and algorithm is described in Algorithm 1.

By expanding the SGD update, we have the following dynamics.

$$\theta_t - \theta_* = (I - A_t x_t x_t^\top)(\theta_{t-1} - \theta_*) + \epsilon_t A_t x_t \,.$$

The goal is to minimized the regret at time T, defined by $R_T = \sum_{t=1}^{\top}(x_t^* - x_t)^\top \theta_*$, where x_t^* is the optimal action at time t. The regret of Algorithm 1 is bounded by the following.

Theorem E.1. *Setting parameters as in Algorithm 1, with probability $1-\delta$, for any $\lambda > 0$, $L > 0$, $L_* > 0$,*

$$R_T \le 34\sqrt{2nT \max\left\{L_*^2 L^2, n\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}\log\left(1+\frac{T}{n}\right)}\,.$$

In particular, we have $R_T = O(n\sqrt{T\log^2 T\log(1/\delta)})$.

In order to obtain the above regret bound, we follow the standard approach in [AYPS11] and study the dynamic of $X_t = \|\theta_t - \theta_*\|_{V_t}^2$ using our framework. Specifically, we will obtain the following theorem.

Theorem E.2. *Setting parameters as in Algorithm 1, for any $\lambda > 0, L > 0, L_* > 0, T > 0$, we have*

$$\Pr[\exists t \in [T],\ X_T > 288\max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}] < \delta\,.$$

Comparison. Linear bandit is a well-studied problem via the standard LinUCB approach [AYPS11, DHK08]. However, the standard approaches reuses historical data and therefore it is difficult for it to be scalable. In this work, we analyze the SGD version of LinUCB. The idea of using an SGD update appeared in [KPM15], but their design of upper confidence bound (UCB) is heuristic, and no regret bound is provided. [JBNW17] develops an online-to-confidence-set algorithm to achieve $O(n(T\log^2(T)\log(T/\delta))^{1/2})$ regret up to iterated log-factors. They use an online Newton step predictor as a sub-routine to get rid of the dependence on historical data. In contrast, we do not need any sub-routine and update the parameter directly and our regret bound is $O(n(T\log^2 T\log(1/\delta))^{1/2})$ which matches the lower bound $O(nT^{1/2})$ up to logarithmic factors. As a result, we both simplify the procedure and improve the regret bound for linear bandit with SGD updates.

Section structure. In the rest of this section, we provide the linearization and moment analysis in Section E.1, the improvement analysis in Section E.2, and the interval analysis in Section E.3.

E.1 Linearization and moment analysis

We would like to apply our framework on the quantity $X_t = \|\theta_t - \theta_*\|_{V_t}^2$. We have the following lemma on linearization.

Lemma E.3 (Linearization and moment analysis for linear bandit with SGD updates)**.** *Consider the setting in Theorem E.2. Let $\eta \le \frac{\lambda}{L^2}$. For all $t \in \mathbb{N}$, we have*

$$X_t \le X_{t-1} + N_t$$

where

$$N_t = 2\eta\epsilon_t(\theta_{t-1}-\theta_*)^\top x_t - 2\eta^3\epsilon_t(\theta_{t-1}-\theta_*)^\top x_t\|x_t\|_{V_{t-1}^{-1}}^4 + 2\epsilon_t^2\eta^2\|x_t\|_{V_{t-1}^{-1}}^2\,.$$

For $\Lambda > 0$, let τ is the stopping time for the event $\{X_t > \Lambda\}$. For every $T_0 + 1 \le t \le T$, the following following functions $(B_{T_0}, \mu_{T_0}, \sigma_{T_0}^2)$ form a moment profile for $\{X_t\}$, Λ, and T_0.

- *(Bounded difference)* $|\mathbf{1}_{\tau\ge t}N_t| \le B_{T_0}(t,\Lambda) = 3\eta\|x_t\|_{V_{t-1}^{-1}}\sqrt{\Lambda} + 2\eta^2\|x_t\|_{V_{t-1}^{-1}}^2$.

- *(Conditional expectation)* $|\mathbb{E}[\mathbf{1}_{\tau\geq t}N_t \,|\, \mathcal{F}_{t-1}]| \leq \mu_{T_0}(t,\Lambda) = 2\eta^2\|x_t\|^2_{V_{t-1}^{-1}}$.
- *(Conditional variance)* $|\mathsf{Var}[\mathbf{1}_{\tau\geq t}N_t \,|\, \mathcal{F}_{t-1}]| \leq \sigma^2_{T_0}(t,\Lambda) = 18\eta^2\|x_t\|^2_{V_{t-1}^{-1}}\Lambda + 8\eta^4\|x_t\|^4_{V_{t-1}^{-1}}$.

Proof of Lemma E.3. First notice that since $\eta \leq \frac{\lambda}{L^2}$, $\eta\|x_t\|^2_{V_{t-1}^{-1}} \leq 1$. By Equation 5.2, we have

$$\|\theta_t - \theta_*\|^2_{V_t} = \|(I - A_t x_t x_t^\top)(\theta_{t-1} - \theta_*)\|^2_{V_t} + 2\langle (I - A_t x_t x_t^\top)(\theta_{t-1} - \theta_*), \epsilon_t A_t x_t\rangle_{V_t} + \epsilon_t^2\|A_t x_t\|^2_{V_t}.$$

For the first term, we have

$$\begin{aligned}
&\|(I - A_t x_t x_t^\top)(\theta_{t-1} - \theta_*)\|^2_{V_t} \\
&= \|(\theta_{t-1} - \theta_*)\|^2_{V_t} - 2\langle \theta_{t-1} - \theta_*, A_t x_t x_t^\top(\theta_{t-1} - \theta_*)\rangle_{V_t} + \|A_t x_t x_t^\top(\theta_{t-1} - \theta_*)\|^2_{V_t} \\
&= \|(\theta_{t-1} - \theta_*)\|^2_{V_{t-1}} + \eta((\theta_{t-1} - \theta_*)^\top x_t)^2 - 2\eta((\theta_{t-1} - \theta_*)^\top x_t)^2 - 2\eta^2((\theta_{t-1} - \theta_*)^\top x_t)^2\|x_t\|^2_{V_{t-1}^{-1}} \\
&+ \eta^2((\theta_{t-1} - \theta_*)^\top x_t)^2\|x_t\|^2_{V_{t-1}^{-1}} + \eta^3((\theta_{t-1} - \theta_*)^\top x_t)^2\|x_t\|^4_{V_{t-1}^{-1}} \\
&\leq \|(\theta_{t-1} - \theta_*)\|^2_{V_{t-1}} + (\eta^3\|x_t\|^4_{V_{t-1}^{-1}} - \eta^2\|x_t\|^2_{V_{t-1}^{-1}} - \eta)((\theta_{t-1} - \theta_*)^\top x_t)^2 \\
&\leq \|\theta_{t-1} - \theta_*\|^2_{V_{t-1}}.
\end{aligned}$$

For the second term we have

$$\begin{aligned}
&2\langle (I - A_t x_t x_t^\top)(\theta_{t-1} - \theta_*), \epsilon_t A_t x_t\rangle_{V_t} \\
&= 2\langle \theta_{t-1} - \theta_*, \epsilon_t A_t x_t\rangle_{V_t} - 2\langle A_t x_t x_t^\top(\theta_{t-1} - \theta_*), \epsilon_t A_t x_t\rangle_{V_t} \\
&= 2\eta\epsilon_t(\theta_{t-1} - \theta_*)^\top x_t + 2\eta^2\epsilon_t(\theta_{t-1} - \theta_*)^\top x_t \cdot \|x_t\|^2_{V_{t-1}^{-1}} - 2\eta^2\epsilon_t(\theta_{t-1} - \theta_*)^\top x_t \cdot \|x_t\|^2_{V_{t-1}^{-1}} \\
&- 2\eta^3\epsilon_t(\theta_{t-1} - \theta_*)^\top x_t \cdot \|x_t\|^4_{V_{t-1}^{-1}} \\
&= 2\eta\epsilon_t(\theta_{t-1} - \theta_*)^\top x_t - 2\eta^3\epsilon_t(\theta_{t-1} - \theta_*)^\top x_t \cdot \|x_t\|^4_{V_{t-1}^{-1}}.
\end{aligned}$$

For the third term, we have

$$\epsilon_t^2\|A_t x_t\|^2_{V_t} = \epsilon_t^2\eta^2\|x_t\|^2_{V_{t-1}^{-1}} + \epsilon_t^2\eta^3\|x_t\|^4_{V_{t-1}^{-1}} \leq 2\epsilon_t^2\eta^2\|x_t\|^2_{V_{t-1}^{-1}}.$$

So in total we have

$$X_t \leq X_{t-1} + N_t$$

where

$$N_t = 2\eta\epsilon_t(\theta_{t-1} - \theta_*)^\top x_t - 2\eta^3\epsilon_t(\theta_{t-1} - \theta_*)^\top x_t\|x_t\|^4_{V_{t-1}^{-1}} + 2\epsilon_t^2\eta^2\|x_t\|^2_{V_{t-1}^{-1}}.$$

In particular,

$$\|\theta_t - \theta_*\|^2_{V_t} \leq \|\theta_0 - \theta_*\|^2_{V_0} + \sum_{i=1}^t N_i.$$

For the bounded difference by $\eta\|x_t\|^2_{V_{t-1}^{-1}} \leq \frac{1}{2}$ and Cauchy-Schwarz inequality we have

$$|\mathbf{1}_{\tau\geq t}N_t| \leq 3\eta\|x_t\|_{V_{t-1}^{-1}}\sqrt{\Lambda} + 2\eta^2\|x_t\|^2_{V_{t-1}^{-1}}.$$

For the conditional expectation, we have

$$|\mathbb{E}[\mathbf{1}_{\tau\geq t}N_t \,|\, \mathcal{F}_{t-1}]| \leq 2\eta^2\|x_t\|^2_{V_{t-1}^{-1}}.$$

For the conditional variance, we have

$$|\mathsf{Var}[\mathbf{1}_{\tau\geq t}N_t \,|\, \mathcal{F}_{t-1}]| \leq 18\eta^2\|x_t\|^2_{V_{t-1}^{-1}}\Lambda + 8\eta^4\|x_t\|^4_{V_{t-1}^{-1}}.$$

□

E.2 Improvement analysis

Notice that we need to bound $\|x_t\|^2_{V_{t-1}^{-1}}$. So we begin with following helper lemma.

Lemma E.4. *We have*

$$\log\left(\frac{\det(V_t)}{\det(V_0)}\right) \le \sum_{i=1}^{t} \eta\|x_i\|^2_{V_{i-1}^{-1}} \le 2\log\left(\frac{\det(V_t)}{\det(V_0)}\right) \le 2n\log\left(1+\frac{\eta t L^2}{n\lambda}\right).$$

Proof of Lemma E.4. By definition of V_t,

$$\begin{aligned}\det(V_t) &= \det(V_{t-1} + \eta x_t x_t^\top)\\ &= \det(V_{t-1})\det(I + \eta V_{t-1}^{-1/2} x_t (V_{t-1}^{-1/2} x_t)^\top)\\ &= \det(V_{t-1})(1+\eta\|x_t\|^2_{V_{t-1}^{-1}}).\end{aligned}$$

Unfolding the recursion we have

$$\det(V_t) = \det(V_0)\prod_{i=1}^{t}(1+\eta\|x_i\|^2_{V_{i-1}^{-1}}).$$

We then obtain

$$\log\left(\frac{\det(V_t)}{\det(V_0)}\right) = \sum_{i=1}^{t}\log(1+\eta\|x_i\|^2_{V_{i-1}^{-1}}) \le \sum_{i=1}^{t}\eta\|x_i\|^2_{V_{i-1}^{-1}}, \text{ and}$$

$$\sum_{i=1}^{t}\eta\|x_i\|^2_{V_{i-1}^{-1}} \le \sum_{i=1}^{t} 2\log(1+\eta\|x_i\|^2_{V_{i-1}^{-1}}) = 2\log\left(\frac{\det(V_t)}{\det(V_0)}\right).$$

The last step is to bound the determinant $\det(V_t)$ by

$$\log\frac{\det(V_t)}{\det(V_0)} \overset{(i)}{\le} \log\left(\frac{\operatorname{trace}(V_t)/n}{\lambda}\right)^n \overset{(ii)}{\le} \log\left(\frac{\eta t L^2 + n\lambda}{n\lambda}\right)^n = n\log\left(1+\frac{\eta t L^2}{n\lambda}\right)$$

where (i) is by AM-GM inequality and (ii) is by the definition of V_t. □

Now we can bound the deviation induced by the noise.

Lemma E.5 (Improvement analysis for linear bandit with SGD updates)**.** *Consider the setting in Theorem E.2. For every $A_0 > A_1 > 0$, $\delta' > 0$, $\Lambda > 0$, and $1 \le T_0 < T_1 \in \mathbb{N}$, let*

$$\Delta = \sqrt{144\eta n\log\left(1+\frac{\eta T_1 L^2}{n\lambda}\right)\Lambda\log\frac{1}{\delta'}} + 16\eta n\log\left(1+\frac{\eta T_1 L^2}{n\lambda}\right)\log\frac{1}{\delta'}.$$

Suppose that we have $A_0 + \Delta < \Lambda$ and $\Lambda < A_1$. Then,

$$\Pr\left[\max_{T_0+1\le t\le T_1} X_t > \Lambda \;\middle|\; X_{T_0} \le A_0\right] < \delta'.$$

In particular, the above implies $\Pr[X_{T_1} > A_0 \mid X_{T_0} \le a] \le \delta'$.

Proof of Lemma E.5. Given the moment profile $(B_{T_0}, \mu_{T_0}, \sigma^2_{T_0})$ in Lemma E.3, we apply a martingale concentration inequality (see Theorem 6.2 in [CL06]) to obtain the deviation of $\sum_{t=T_0}^{T_1 \wedge \tau} N_t$ as follows.

$$\begin{aligned}
&\sqrt{\sum_{t=T_0+1}^{T_1} 2\sigma^2_{T_0}(t,\Lambda)\log\frac{1}{\delta'} + 2B^2_{T_0}(t,\Lambda)\log\frac{1}{\delta'}} + \sum_{t=T_0+1}^{T_1}\mu_{T_0}(t,\Lambda) \\
&\le \sqrt{\sum_{t=T_0+1}^{T_1}\left(72\eta^2\|x_t\|^2_{V_{t-1}^{-1}}\Lambda + 32\eta^4\|x\|^4_{V_{t-1}^{-1}}\right)\log\frac{1}{\delta'}} + \sum_{t=T_0+1}^{T_1} 2\eta^2\|x_t\|^2_{V_{t-1}^{-1}} \\
&\le \sqrt{\sum_{t=T_0+1}^{T_1} 72\eta^2\|x_t\|^2_{V_{t-1}^{-1}}\Lambda\log\frac{1}{\delta'}} + \sqrt{\sum_{t=T_0+1}^{T_1} 32\eta^4\|x_t\|^4_{V_{t-1}^{-1}}\log\frac{1}{\delta'}} + \sum_{t=T_0+1}^{T_1} 2\eta^2\|x_t\|^2_{V_{t-1}^{-1}} \\
&\le \sqrt{\sum_{t=T_0+1}^{T_1} 72\eta^2\|x_t\|^2_{V_{t-1}^{-1}}\Lambda\log\frac{1}{\delta'}} + \sum_{t=T_0+1}^{T_1} 6\eta^2\|x_t\|^2_{V_{t-1}^{-1}}\sqrt{\log\frac{1}{\delta'}} + \sum_{t=T_0+1}^{T_1} 2\eta^2\|x_t\|^2_{V_{t-1}^{-1}}\,.
\end{aligned}$$

By Lemma E.4, we have

$$\le \sqrt{144\eta n\log\left(1+\frac{\eta T_1 L^2}{n\lambda}\right)\Lambda\log\frac{1}{\delta'}} + 16\eta n\log\left(1+\frac{\eta T_1 L^2}{n\lambda}\right)\sqrt{\log\frac{1}{\delta'}}\,.$$

We get what we want from Proposition 2.3. □

E.3 Interval Analysis

Now we are ready to prove Theorem E.2

Proof of Theorem E.2. Let $t_0 = 0, t_1 = T, a_0 = \|\theta_*\|^2_{V_0}$ and

$$a_1 = \Lambda = 288\max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}$$

by plugging $\eta = \frac{\lambda}{L^2}$. Now, we invoke Lemma E.5 with $A_0 = a_0$, $A_1 = a_1$, $T_0 = 0$, $T_1 = T$, and $\delta' = \delta$. Let us verify the two conditions. First, we verify the pull out condition. We have

$$\begin{aligned}
a_0 + \Delta &\le L_*^2\lambda + \sqrt{144\frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\Lambda\log\frac{1}{\delta}} + 16\frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\sqrt{\log\frac{1}{\delta}} \\
&= \frac{\Lambda}{288} + \frac{\Lambda}{\sqrt{2}} + \frac{16\Lambda}{288} < \Lambda\,.
\end{aligned}$$

For the improvement condition, we have $\Lambda_1 \le a_1$ trivially. By Lemma E.5 this implies that

$$\Pr\left[\exists t \in [T],\ X_t > 288\max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}\right] < \delta\,.$$

□

Finally we prove the regret bound.

Proof of Theorem E.1. To prove the regret bound, by setting

$$\beta_t = 288\max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{t}{n}\right)\log\frac{1}{\delta}\right\}\,,$$

Theorem E.2 can guarantee with probability $1-\delta$, $\theta_* \in B_t$. As a result, defining $\tilde{\theta}_t = \arg\min_{\theta \in B_t} \min_{x \in D_t} \langle x, \theta \rangle$, $\langle x_t^*, \theta_* \rangle \leq \langle x_t, \tilde{\theta}_t \rangle$. Therefore,

$$\langle x_t^* - x_t, \theta_* \rangle \leq \langle x_t, \tilde{\theta}_t - \theta_* \rangle \leq \|x_t\|_{V_{t-1}^{-1}} \|\tilde{\theta}_t - \theta_*\|_{V_{t-1}} \leq 2\|x_t\|_{V_{t-1}^{-1}} \sqrt{\beta_t}\,.$$

Taking the sum,

$$R_T \leq \sum_{i=1}^{\top} 2\sqrt{\beta_t} \|x_i\|_{V_{i-1}^{-1}}\,.$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&= 2\sqrt{288T \max\left\{L_*^2\lambda, \frac{n\lambda}{L^2}\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}} \sqrt{\sum_{i=1}^{\top} \|x_i\|_{V_{i-1}^{-1}}^2} \\
&\overset{(i)}{\leq} 34\sqrt{T \max\left\{L_*^2 L^2, n\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\}} \sqrt{2n\log\left(1+\frac{T}{n}\right)} \\
&\leq 34\sqrt{2nT \max\left\{L_*^2 L^2, n\log\left(1+\frac{T}{n}\right)\log\frac{1}{\delta}\right\} \log\left(1+\frac{T}{n}\right)}
\end{aligned}$$

where (i) is by Lemma E.4. □