

Non-asymptotic Error Bounds For Constant Stepsize Stochastic Approximation For Tracking Mobile Agents

Bhumesh Kumar, Vivek Borkar, Akhil Shetty *

March 4, 2019

Abstract

This work revisits the constant stepsize stochastic approximation algorithm for tracking a slowly moving target and obtains a bound for the tracking error that is valid for the entire time axis, using the Alekseev non-linear variation of constants formula. It is the first non-asymptotic bound for the entire time axis in the sense that it is not based on the vanishing stepsize limit and associated limit theorems unlike prior works, and captures clearly the dependence on problem parameters and the dimension.

Keywords: Stochastic Approximation; Constant Stepsize; Non-asymptotic bound; Alekseev’s Formula; Martingale Concentration Inequalities; Perturbation Analysis; Non-stationary Optimization

1 Introduction

1.1 Background

Robbins and Monro proposed in [39] a stochastic iterative scheme

$$x_{n+1} = x_n + a_n [h(x_n) + M_{n+1}], \quad n \geq 0, \quad (1)$$

*VB is and BK, AS were with the Department of Electrical Engineering, IIT Bombay, Powai, Mumbai, Maharashtra 400076, India. BK is now with the Department of Electrical and Computer Engineering, University of Wisconsin at Madison, Madison, WI 53706, USA. AS is now with the Department of Electrical Engineering and Computer Science, University of California at Berkeley, Cory Hall, Hearst Avenue, Berkeley, CA 94720, USA. Email: bkumar@wisc.edu, borkar.vs@gmail.com, shetty.akhil@berkeley.edu. Work of VB was supported in part by a J. C. Bose Fellowship, CEFIPRA grant No. IFC/DST-Inria-2016-01/448 “Machine Learning for Network Analytics” and a grant for ‘*Approximation for high dimensional optimization and control problems*’ from the Department of Science and Technology, Government of India.

for finding the zero(s) of a function $h(\cdot)$ given its noisy evaluations, with M_{n+1} being the measurement noise. By a clever choice of the stepsize sequence $\{a_n\}$, viz., those satisfying

$$\sum_n a_n = \infty, \quad \sum_n a_n^2 < \infty, \quad (2)$$

they were able to show almost sure (a.s.) convergence of the scheme to a zero of h under reasonable hypotheses. The scheme has since been a cornerstone of not only statistical computation, but also in a variety of engineering applications ranging from signal processing, adaptive control, to more recently, machine learning. See [8, 15] for some recent pedagogical accounts of stochastic approximation. What makes it so popular is its typically low per iterate memory and computational requirement and ability to ‘average out’ the noise, which makes it ideal for adaptive estimation/learning scenarios. A later viewpoint [17], [34] views (1) as a noisy discretization of the ordinary differential equation (ODE for short)

$$\dot{x}(t) = h(x(t)) \quad (3)$$

with decreasing stepsize and argues that the errors due to discretization and noise are asymptotically negligible under (2), so that it has the same asymptotic behaviour as (3). See [8, 10] for a fuller development of this approach.

The clean theory under (2) notwithstanding, there has also been an interest and necessity to consider constant stepsize $a_n \equiv a > 0$. The strong convergence claims under (2) can no longer be expected¹, e.g., for the simple case of $\{M_n\}$ being i.i.d. zero mean, the best one can hope for is convergence to a stationary distribution. What one can still expect is a high probability concentration around the desired target, viz., zero(s) of h , if the stepsize a is small [30, 31]. This is acceptable and in fact unavoidable in the important application area of tracking a slowly moving target or measuring a slowly exciting signal [4, 23], and other instances of learning in a slowly varying environment. This is because with decreasing stepsize, the algorithmic time scale, dictated by the decreasing stepsize, eventually becomes slower than the timescale on which the target is moving and thereby loses its tracking ability. The alternative of either frequent resets or adaptive loop gain is often not desirable because of the additional logic it requires, particularly when the algorithm is hard-wired [8, 44], and one settles for a judiciously chosen constant stepsize. Such schemes are a part of traditional signal processing and neural network algorithms [18, 19, 21, 28, 35, 41] and often show up in important applications such as quasi-stationary experimentation for meteorology [14], slowly exciting physical wave measurement [13], and more recently in online learning and non-stationary optimization [45, 51]. However, the focus in online learning is cumulative regret bounds instead of all time bounds.

These developments have motivated analysis of constant stepsize schemes [10, 12, 25, 29, 30, 37, 38, 41] in the form of various limit theorems, (non) asymptotic

¹barring some very special cases, e.g., when the right hand side of (1) is contractive uniformly w.r.t. the noise variable.

analysis, law of iterated logarithm etc., but a convenient bound valid for all time, a useful metric for tracking applications in a slowly varying environment, seems to be a topic of relatively recent interest [51]. Our objective here is to provide precisely one such bound.

1.2 Comparison with Prior Art

As already mentioned, one of the main motivation of constant stepsize stochastic approximation has been their ability to track slowly moving environments. Not surprisingly, much of the early work has come from signal processing and control, most notably in adaptive filtering, and this continues to be its primary application domain. Some representative works are [4], [20], [22], [23], [24], [41], [51], etc.

Much of this work concerns tracking in specific models and the proposed schemes usually have a very specific structure, e.g., linear. From purely theoretical angle, analyses appear in [12], [25], [30], [37], [38] among others. The emphasis of the latter is towards analyzing convergence properties in the small stepsize limit and the associated functional central limit theorem for fluctuations around the deterministic o.d.e. limit, except in case of [25], which establishes a law of iterated logarithms, and [38], which obtains confidence bounds for a specific choice of adaptive stepsizes and stopping rule. The latter is a non-asymptotic result as the title suggests, but in a different sense than us.

In the context of tracking, the functional central limit theorem characterizing a Gauss-Markov process as a limit in law of suitably scaled fluctuations is also used for suggesting performance metrics for tracking application, see, e.g., [5].

More recently constant stepsize stochastic gradient and its variants have elicited interest in machine learning literature due to the possibility of using them in conjunction with iterate averaging to get better speed than decreasing stepsizes, see, e.g, [3]. The pros and cons of these have been discussed, e.g., in [32]. This motivation, however, is not relevant for tracking because iterate averaging is also a stochastic approximation with decreasing stepsizes ($a_n = 1/(n + 1)$ to be precise) and decreasing stepsizes is simply not an option here because the iterates will eventually become slower than the slowly varying signal and lose their tracking ability.

Another strand of work analyzes tracking in the specific context of tracking the solution of an optimization problem when its parameters drift slowly [45]. Tracking problems have also been studied in the literature as regime switching stochastic approximations when the evolution is modulated by a Markov chain on a time scale equal to or faster than that of the algorithm. This situation has

been analysed through mean squared error bounds [46, 47] and is close in spirit to ours.

1.3 Our contributions

Our main result is Theorem 4.1. The highlights of this result are as follows.

1. Our set-up is applicable to a very general scenario that includes unbounded correlated noise without any explicit evolution model, no explicit strong convexity or linearity assumptions regarding the dynamics being tracked, and so on, rendering it a more general framework than in prior work.
2. We provide a bound valid for the entire time axis, not only for a finite time interval as in, e.g., ‘sample complexity’ bounds, or purely asymptotic as in, e.g., cumulative regret bounds or asymptotic error bounds. That is, it holds uniformly for all n , $0 \leq n < \infty$, not only for $n \leq$ some N or in the $n \rightarrow \infty$ limit, which need not reflect the finite time behavior. This is particularly relevant here because we are considering the problem of continuously tracking a *time-varying, in particular, non-stationary* target. Furthermore, this is achieved under a very general noise model, viz., martingale differences which allow dependence across times, requiring only uncorrelatedness. Their conditional distributions given the past are required to satisfy an exponential moment bound that is satisfied by most standard distributions such as exponential, gaussian, their mixtures, etc., except the heavy tailed ones.
3. This bound is *non-asymptotic*, i.e., it is derived for the actual constant stepsize $a > 0$ and not from an idealized limiting scenario based on a limit theorem for fluctuations in the $a \downarrow 0$ limit, as is often the case in prior studies. To the best of our knowledge, ours is the first result to achieve this. Also, our derivation of the bound allows us to keep track of its dependence on problem parameters, dimension, etc. if needed.
4. We bound the *exact* error which is given by the Alekseev formula, there is no approximation at this stage. Furthermore, we analyze this error keeping the slow movement of the target being tracked in tact, without treating it as essentially static as, e.g., in [5].

As for potential avenues for improvement, we have the following observations:

1. It appears unlikely that the bounds that use Lipschitz constants etc., can be improved much, if at all. The moment bounds on martingale differences use state of the art martingale concentration inequalities and could

improve if better inequalities become available. It may be noted that we assume exponential tails for the distributions of martingale differences. Stronger inequalities such as McDiarmid’s inequality may be used under stronger hypotheses such as uniformly bounded martingale differences, see, e.g., [6]. On a different note, if we allow heavy tailed noise, one would get weaker claims using the corresponding, naturally weaker, concentration inequalities. Rather limited results are available here, see e.g., [26] and its application to stochastic approximation in [2].

2. One potential spot for improvement is in the use of the assumption (\dagger) below, which entails a stability condition for a linearized dynamics which is time-dependent. Such conditions are available only under constraints on the time scale separation between fast dynamics (of the algorithm) and the slow one (of the target). This, as argued later, is unavoidable, because there will be no tracking otherwise. This fact necessitates such a condition or something close to it. The one we have used, due to Solo [42], is the most general available to our knowledge. (Another class of sufficient conditions available is based on existence of Liapunov functions and not explicit like Solo’s.)

1.4 Organization

We begin by describing the problem formulation in the next section. This is followed by the Alekseev formula as a non-linear generalization of the variation of constants formula, and a key exponential stability assumption. A useful set of sufficient conditions for this assumption are recalled. Section 3 details the error analysis characterizing the tracking behaviour, developed through a sequence of lemmas and leading to the main result in section 4. Section 5 concludes with some discussion. An appendix recalls a martingale concentration inequality used in the main text.

1.5 Symbols and Notation

The section number where the notation first appears is given in parentheses.

$$x_n = \text{Iterate at time } n \quad (2.1)$$

$$h(\cdot, \cdot) = \text{driving vector field of the tracking scheme} \quad (2.1)$$

$$a = \text{Step-size} \quad (2.1)$$

$$M_{n+1} = \text{Martingale difference noise} \quad (2.1)$$

$$\varepsilon_{n+1} = \text{Additive error} \quad (2.1)$$

$$\varepsilon^* = \text{Bound on } \|\varepsilon_{n+1}\| \quad (2.1)$$

$y(\cdot)$ = Slowly varying signal to be tracked (2.1)

ϵ = Small ($\ll 1$) number controlling rate of $y(\cdot)$ (2.1)

$\gamma(\cdot)$ = Vector field driving $y(\cdot)$ (2.1)

$$C^* = \max \left(\sup_n \mathbb{E}[\|x_n\|^2]^{1/2}, \sup_n \mathbb{E}[\|x_n\|^4]^{1/4} \right) \quad (2.1)$$

$$C_\gamma = \sup_{t \geq 0} \|y(t)\| \quad (2.1)$$

C_M, δ = Constants featuring in the bound for $\|M_{n+1}\|$ (2.1)

Φ = Transition matrix of linear system (2.2)

$z(\cdot)$ = Slowly varying equilibrium for the algorithm (2.3)

d = Dimension of x_n and $z(\cdot)$ (2.1/2.3)

∇ = Gradient operator (2.3)

C_Φ, β = Constants featuring in the exponential bound for Φ (2.3)

L_f = Lipschitz constant for Lipschitz function f (generic)

G_f = constant of linear growth for function f , i.e., $|f(x)| \leq C_f(1 + \|x\|)$. (generic)

$B_f = \sup_x |f(x)|$ for a bounded function f (generic)

$$K_\gamma = \max_{\{y\} \leq C_\gamma} \|\gamma(y)\| \quad (3.2)$$

$O(\cdot)$ = Big O notation (3.3)

$$\mu = 1/\beta \quad (3.3)$$

$$K_1 = L_{\bar{h}}(1 + C_h + C_\gamma) + \epsilon^* \quad (3.3)$$

$$K_2 = C_\Phi L_{\bar{h}} \quad (3.3)$$

$$K_3 = K_1 + L_\gamma a \epsilon \quad (3.3)$$

$$K_4 = \max(2C_M/\delta^2, C_M^2/\delta^2) \quad (3.3)$$

$$K_5 = KC_\Phi^3 L_D/\beta \quad (3.4)$$

$$K_6 = C_\Phi^3 L_\gamma L_D \epsilon \quad (3.4)$$

$$K_7 = \max(24C_M\delta^4, 4C_M^2\delta^4) \quad (3.4)$$

$$K_8 = 2\sqrt{6C_M C_h^2}/\delta^2 \quad (3.4)$$

$$K_9 = \frac{C_M \gamma_1 d^{1.5}}{\delta} \quad (3.5)$$

2 Preliminaries

2.1 The tracking problem

We consider a constant step size stochastic approximation algorithm given by the d -dimensional iteration

$$x_{n+1} = x_n + a[h(x_n, y_n) + M_{n+1} + \varepsilon_{n+1}], \quad n \geq 0, \quad (4)$$

for tracking a slowly varying signal governed by

$$\dot{y}(t) = a\epsilon\gamma(y(t)), \quad (5)$$

with $0 < a < 1$, $0 < \epsilon \ll 1$. Also, $y_n := y(n), n \geq 0$, the trajectory of (5) sampled at unit² time intervals coincident with the clock of the above iteration, with slight abuse of notation. We assume that $y(t), t \geq 0$, remains in a bounded set. The term ε_{n+1} represents an added bounded component attributed to possible numerical errors (e.g., error in gradient estimation in case of stochastic gradient algorithms [8]). We assume the following:

- The smallness condition on ϵ ensures a separation of time scale between the two evolutions (4) and (5), in particular (5) has to be ‘sufficiently slow’ in a sense to be made precise later.
- $h : (x, y) \mapsto h(x, y)$ is twice continuously differentiable in x with the first and second partial derivatives in x bounded uniformly in y in a compact set, and Lipschitz in y . A common example is where $h(x, y) = -(x - y)$ corresponding to least mean square criterion for tracking in the above context with x , resp. y standing for the states of the tracking scheme and the target resp.
- $\gamma(\cdot)$ is Lipschitz continuous,
- $C^* := \max\left(\sup_n \mathbb{E}[\|x_n\|^2]^{1/2}, \sup_n \mathbb{E}[\|x_n\|^4]^{1/4}\right) < \infty$. (See [10] for sufficient conditions for uniform boundedness of second moments. Analogous conditions can be given for fourth moments.)
- $C_\gamma := \sup_{t \geq 0} \|y(t)\| < \infty$,
- there exists a constant $\varepsilon^* > 0$ such that

$$\|\varepsilon_{n+1}\| \leq \varepsilon^*, \quad \forall n \geq 0, \quad (6)$$

- M_n is a martingale difference sequence w.r.t. the increasing σ -fields

$$\mathcal{F}_n := \sigma(x_m, M_m, \varepsilon_m, m \leq n), \quad n \geq 0,$$

²without loss of generality

and satisfies: there exist continuous functions $c_1, c_2 : \mathcal{R}^d \rightarrow (0, \infty)$ with c_2 being bounded away from 0, such that

$$P(\|M_{n+1}\| > u | \mathcal{F}_n) \leq c_1(x_n) e^{-c_2(x_n)u}, \quad n \geq 0, \quad (7)$$

for all $u \geq v$ for a fixed, sufficiently large $v > 0$ (i.e., a sub-exponential tail) with

$$\sup_n E[c_1(x_n)] < \infty. \quad (8)$$

In particular, (7), (8) together imply that there exist $\delta, C_M > 0$ such that

$$E[e^{\delta\|M_{n+1}\|}] \leq C_M, \quad n \geq 0. \quad (9)$$

Using the Taylor expansion of the exponential function, we get

$$\sum_{m=0}^{\infty} \frac{\delta^m E[\|M_{n+1}\|^m]}{m!} \leq C_M, \quad n \geq 0. \quad (10)$$

As each term in the above summation is positive, we can conclude that for all $n, m \geq 0$,

$$E[\|M_{n+1}\|^m] \leq \frac{C_M m!}{\delta^m}. \quad (11)$$

We shall be interested in $m = 2, 4$.

These bounds will play an important role in our error analysis. We next state a formula due to Alekseev [1] that captures the difference between the trajectory of a system and its (regular) perturbation, and may be viewed as a ‘non-linear variation of constants’ formula.

2.2 Alekseev’s formula

Consider the ODE

$$\dot{w}(t) = f(t, w(t)), \quad t \geq 0,$$

and its perturbed version,

$$\dot{u}(t) = f(t, u(t)) + g(t, u(t)), \quad t \geq 0,$$

where $f, g : \mathcal{R} \times \mathcal{R}^d \mapsto \mathcal{R}^d$, with:

- $f(t, x)$ is measurable in t and continuously differentiable in x with bounded derivatives uniformly w.r.t. t , and,
- $g(t, x)$ is measurable in t and Lipschitz in x uniformly w.r.t. t .

Let $w(t, t_0, u_0)$ and $u(t, t_0, u_0)$ denote respectively the solutions to the above non-linear systems for $t \geq t_0$, satisfying $w(t_0, t_0, u_0) = u(t_0, t_0, u_0) = u_0$. Then for $t \geq t_0$,

$$u(t, t_0, u_0) = w(t, t_0, u_0) + \int_{t_0}^t \Phi(t, s, u(s, t_0, u_0))g(s, u(s, t_0, u_0))ds, \quad (12)$$

where $\Phi(t, s, w_0)$ for any $w_0 \in \mathcal{R}^d$ is the fundamental matrix of the linearized system

$$\dot{\phi}(t) = \frac{\partial f}{\partial w}(t, w(t, s, w_0))\phi(t), \quad t \geq s, \quad (13)$$

with $\Phi(s, s, w_0) = \mathcal{I}_d$, the d -dimensional identity matrix. That is, it is the unique solution to the matrix linear differential equation

$$\dot{\Phi}(t, s, w_0) = \frac{\partial f}{\partial w}(t, w(t, s, w_0))\Phi(t, s, w_0)$$

with the aforementioned initial condition at $t = s$. The equation (12) is the Alekseev nonlinear variation of constants formula [1] (see also Lemma 3, [11]).

The generalization of Alekseev's nonlinear variation of constants for differing initial conditions [7] is given by

$$u(t, t_0, u_0) = w(t, t_0, w_0) + \Phi(t, t_0, u_0)(u_0 - w_0) + \int_{t_0}^t \Phi(t, s, u(s, t_0, u_0))g(s, u(s, t_0, u_0))ds \quad (14)$$

where the additional additive term captures the contribution due to differing initial conditions. This term will decay exponentially under our assumption (\dagger) below.

2.3 Perturbation analysis

In view of the ODE approach described earlier, we consider the candidate ODE

$$\dot{x}(t) = h(x(t), y) \quad (15)$$

where we have treated the y component as frozen at a fixed value in view of its slow evolution (recall that $\epsilon \ll 1$). We assume that this ODE has a globally stable equilibrium $\lambda(y)$ where λ is twice continuously differentiable with bounded first and second derivatives. (Typically, this can be verified by using the implicit function theorem.) In particular,

$$h(\lambda(y), y) = 0 \quad \forall y \implies h(\lambda(y(t)), y(t)) = 0 \quad \forall t \geq 0.$$

Define $z(t) = \lambda(y(t)), t \geq 0$. Then

$$\dot{z}(t) = \epsilon a \nabla \lambda(y(t)) \gamma(y(t))$$

$$\begin{aligned}
&= ah(\lambda(y(t)), y(t)) + \epsilon a \nabla \lambda(y(t)) \gamma(y(t)) \\
&= ah(z(t), y(t)) + \epsilon a \nabla \lambda(y(t)) \gamma(y(t)) = a\tilde{h}(z(t), y(t))
\end{aligned}$$

for

$$\tilde{h}(z, y) := h(z, y) + \epsilon \nabla \lambda(y) \gamma(y).$$

The corresponding Euler scheme would be

$$z_{n+1} = z_n + a\tilde{h}(z_n, y_n).$$

The tracking algorithm (4) can therefore be equivalently written as:

$$x_{n+1} = x_n + a[h(x_n, y_n) + M_{n+1} + \varepsilon_{n+1}] \quad (16)$$

$$= x_n + a[\tilde{h}(x_n, y_n) - \epsilon \nabla \lambda(y_n) \gamma(y_n) + M_{n+1} + \varepsilon_{n+1}] \quad (17)$$

$$= x_n + a[\tilde{h}(x_n, y_n) + \kappa_n(y_n)], \quad (18)$$

where,

$$\kappa_n(y_n) = -\epsilon \nabla \lambda(y_n) \gamma(y_n) + M_{n+1} + \varepsilon_{n+1}. \quad (19)$$

Let $\bar{x}(t)$ be the linearly interpolated trajectory of the stochastic approximation iterates such that $\bar{x}(t_k) = x_k$. That is, for $t_n \equiv na \forall n$,

$$\bar{x}(t) = \bar{x}(t_n) + \frac{t - t_n}{a} [\bar{x}(t_{n+1}) - \bar{x}(t_n)], \quad t \in [t_n, t_{n+1}]. \quad (20)$$

Then from (18), we get

$$\begin{aligned}
\bar{x}(t_{n+1}) &= \bar{x}(t_0) + \sum_{k=0}^n a\tilde{h}(\bar{x}(t_k), y(t_k)) - \sum_{k=0}^n a\epsilon \nabla \lambda(y(t_k)) \gamma(y(t_k)) \\
&\quad + \sum_{k=0}^n aM_{k+1} + \sum_{k=0}^n a\varepsilon_{k+1} \quad (21)
\end{aligned}$$

$$\begin{aligned}
&= \bar{x}(t_0) + \sum_{k=0}^n \int_{t_k}^{t_{k+1}} \tilde{h}(\bar{x}(t_k), y(t_k)) ds - \sum_{k=0}^n \int_{t_k}^{t_{k+1}} \epsilon \nabla \lambda(y(t_k)) \gamma(y(t_k)) ds \\
&\quad + \sum_{k=0}^n \int_{t_k}^{t_{k+1}} M_{k+1} ds + \sum_{k=0}^n \int_{t_k}^{t_{k+1}} \varepsilon_{k+1} ds. \quad (22)
\end{aligned}$$

For $k \geq 0$ and $s \in [t_k, t_{k+1}]$, define perturbation terms:

$$\begin{aligned}
\zeta_1(s) &:= \tilde{h}(\bar{x}(t_k), y(t_k)) - \tilde{h}(\bar{x}(s), y(s)), \\
\zeta_2(s) &:= M_{k+1}, \\
\zeta_3(s) &:= \varepsilon_{k+1}, \\
\zeta_4(s) &:= -\epsilon \nabla \lambda(y(t_k)) \gamma(y(t_k)).
\end{aligned}$$

Thus

$$\begin{aligned}\bar{x}(t_{n+1}) &= \bar{x}(t_0) + \int_{t_0}^{t_{n+1}} \tilde{h}(\bar{x}(s), y(s)) ds \\ &\quad + \int_{t_0}^{t_{n+1}} \left(\zeta_1(s) + \zeta_2(s) + \zeta_3(s) + \zeta_4(s) \right) ds.\end{aligned}$$

Using (20),

$$\bar{x}(t) = \bar{x}(t_0) + \int_{t_0}^t \tilde{h}(\bar{x}(s), y(s)) ds + \int_{t_0}^t \left(\zeta_1(s) + \zeta_2(s) + \zeta_3(s) + \zeta_4(s) \right) ds. \quad (23)$$

Define

$$\Xi(t) = \zeta_1(t) + \zeta_2(t) + \zeta_3(t) + \zeta_4(t).$$

Consider the coupled systems

$$\dot{z}(t) = \tilde{h}(z(t), y(t)), \quad (24)$$

$$\dot{y}(t) = \epsilon a \gamma(y(t)), \quad (25)$$

and

$$\dot{\bar{x}}(t) = \tilde{h}(\bar{x}(t), y(t)) + \Xi(t), \quad (26)$$

$$\dot{y}(t) = \epsilon a \gamma(y(t)). \quad (27)$$

The ODE (26) can be seen as a perturbation of the (24), with the perturbation term being $\Xi(t)$.

Let $D(\cdot, \cdot) \in \mathbb{R}^{d \times d}$ denote the Jacobian matrix of h (and therefore of \tilde{h}) in the first argument, and $\Gamma(\cdot) \in \mathbb{R}^{d \times d}$ the Jacobian matrix of λ . Then the linearization or ‘equation of variation’ of (24) is

$$\dot{r}(t) = D(z(t), y(t))r(t). \quad (28)$$

For $t \geq s \geq 0$ and $x, y \in \mathbb{R}^d$, let $\Phi(t, s; x_0, y_0)$ denote the fundamental matrix for the time varying linear system (28), i.e., the solution to the matrix-valued differential equation

$$\dot{\Phi}(t, s; x_0, y_0) = D(z(t), y(t))\Phi(t, s; x_0, y_0), \quad t \geq s, \quad (29)$$

with initial condition $\Phi(s, s; x_0, y_0) = I$. Then by Alekseev’s formula,

$$\bar{x}(t) = z(t) + \Phi(t, t_0; \bar{x}(t_0), y_0)(\bar{x}(t_0) - z(t_0)) + \int_{t_0}^t \Phi(t, s; \bar{x}(s), y(s))\Xi(s) ds.$$

Define

$$\varrho_n = \Phi(t_n, t_0; \bar{x}(t_0), y_0)(\bar{x}(t_0) - z(t_0)) \quad (30)$$

$$A_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s), y(s)) [\tilde{h}(\bar{x}(t_k), y(t_k)) - \tilde{h}(\bar{x}(s), y(s))] ds, \quad (31)$$

$$B_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s), y(s)) M_{k+1} ds, \quad (32)$$

$$C_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(t_k), y(t_k)) M_{k+1} ds, \quad (33)$$

$$D_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s), y(s)) \varepsilon_{k+1} ds, \quad (34)$$

$$E_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s), y(s)) \times \varepsilon \nabla \lambda(y(t_k)) \gamma(y(t_k)) ds. \quad (35)$$

Then

$$\begin{aligned} \bar{x}(t_n) &= z(t_n) + \int_{t_0}^{t_n} \Phi(t_n, s; \bar{x}(s), y(s)) \zeta_1(s) ds + \int_{t_0}^{t_n} \Phi(t_n, s; \bar{x}(s), y(s)) \zeta_2(s) ds \\ &\quad + \int_{t_0}^{t_n} \Phi(t_n, s; \bar{x}(s), y(s)) \zeta_3(s) ds + \int_{t_0}^{t_n} \Phi(t_n, s; \bar{x}(s), y(s)) \zeta_4(s) ds + \varrho_n \end{aligned} \quad (36)$$

$$= z(t_n) + A_n + (B_n - C_n) + C_n + D_n - E_n + \varrho_n. \quad (37)$$

Therefore

$$\|\bar{x}(t_n) - z(t_n)\| \leq \|A_n\| + \|B_n - C_n\| + \|C_n\| + \|D_n\| + \|E_n\| + \|\varrho_n\|.$$

Also

$$\begin{aligned} \mathbb{E}[\|\bar{x}(t_n) - z(t_n)\|^2]^{1/2} &\leq \mathbb{E}[\|A_n\|^2]^{1/2} + \mathbb{E}[\|E_n\|^2]^{1/2} + \mathbb{E}[\|C_n\|^2]^{1/2} + \\ &\quad \mathbb{E}[\|D_n\|^2]^{1/2} + \mathbb{E}[\|B_n - C_n\|^2]^{1/2} + \mathbb{E}[\|\varrho_n\|^2]^{1/2}. \end{aligned} \quad (38)$$

We shall individually bound the above error terms in the next section under the important assumption of *exponential stability* of the equation of variation (28):

(†) There exists a $\beta > 0$ such that $\forall t > s \geq 0$ and x_0, y_0 ,

$$\|\Phi(t, s; x_0, y_0)\| \leq C_{\Phi} e^{-\beta(t-s)}.$$

This seemingly restrictive assumption requires some discussion, we argue in particular that some such assumption is *essential* if one is to obtain bounds valid for all time.

To begin, since the idea is to have the parametrized o.d.e. (15), which is a surrogate for the original iteration, track its unique asymptotically stable

equilibrium parametrized by y as the parameter $y \approx y(t)$ changes slowly, it is essential that its rate of approach to the equilibrium, dictated by the spectrum of its linearized drift at this equilibrium, should be much faster than the rate of change of the parameter. This already makes it clear that there will be a requirement of minimum time scale separation for tracking to work at all.

A stronger motivation comes from the fact that the tracking error, given exactly by the Alekseev formula, depends on the linearization of the o.d.e. itself around its ideal trajectory $z(\cdot)$, which is a time-varying linear differential equation of the type $\dot{r}(t) = A(t)r(t)$. It is well known in control theory that this can be unstable even if the matrix $A(t)$ is stable for each t , see, e.g., Example 8.1, p. 131, [40]. Stability is guaranteed to hold only in the special case of $A(t)$ varying slowly with time. The most general result in this direction is that of [42], which we recall below as a sufficient condition for (†). (There have also been some extensions thereof to nonlinear systems, see, e.g., [36].)

Consider the following time varying linear dynamical system:

$$\dot{x}(t) = [A(t) + P(t)]x(t) \quad (39)$$

and assume the following for this perturbed system:

1. There exists $\bar{A} > 0$ such that

$$\limsup_{T \uparrow \infty} \frac{1}{T} \int_{t_0}^{t_0+T} \|A(s)\| ds \leq \bar{A} \quad \forall t_0.$$

2. There exists $\gamma \in (0, 1], b > 0$ and $\beta > 0$ sufficiently small in the sense made precise in the theorem below, such that

$$\sum_{t=n_0}^{n_0+n} \|A(t_2 + (t-1)T) - A(t_1 + (t-1)T)\| \leq Tb + T^\gamma(n+1)\beta \quad \forall n, n_0,$$

whenever $|t_2 - t_1| \leq T$.

3. Let $\alpha(t)$ be the real part of the eigenvalue of $A(t)$ whose real part is the largest in absolute value. Then there exists $\bar{\alpha} < 0$ such that, for any $T > 0$,

$$\limsup_{N \uparrow \infty} \frac{1}{N} \sum_{n=n_0}^{n_0+N} \alpha(s + nT) \leq \bar{\alpha} \quad \forall s, n_0.$$

4. There exists $\delta > 0$ such that

$$\limsup_{T \uparrow \infty} \int_{t_0}^{t_0+T} \|P(s)\| ds \leq \delta, \quad \forall t_0.$$

Theorem 2.1 (Stability test for deterministic perturbations using eigenvalue based characterization [42]). *If the previously mentioned assumptions $(A_1) - (A_4)$ hold, the system $\dot{x}(t) = (A(t) + P(t))x(t)$ is exponentially stable provided we chose $\epsilon, \delta > 0$ small enough so that*

$$\bar{\alpha} + \epsilon < 0,$$

and

$$\bar{\alpha} + \epsilon + M_\epsilon \delta < 0,$$

with $M_\epsilon = 3\left(\frac{2(\bar{A}+b)}{\epsilon} + 1\right)^{p-1}/2$, where $\bar{A}, b, \bar{\alpha}$ are as defined in $(A_1)-(A_4)$ and β is small enough so that:

$$\bar{\alpha} + \epsilon + M_\epsilon \delta + 2(\ln M_\epsilon)^{\gamma/(\gamma+1)}[\beta(M_\epsilon + \epsilon/(\bar{A} + b))]^{1/(\gamma+1)} < 0.$$

The correspondence of the foregoing with our framework is given by $A(\cdot) \leftrightarrow D(\cdot, \cdot)$, $P(\cdot) \leftrightarrow \Xi(\cdot)$.

We note here that there are also some sufficient conditions for stability of time-varying linear systems in terms of Liapunov functions, e.g., [49], [50], but they appear not so easy to verify.

3 Error bounds

Here we obtain the error bounds through a sequence of lemmas.

3.1 Bound on D_n

Lemma 1. *For D_n defined in (34),*

$$\mathbb{E}[\|D_n\|^2]^{1/2} \leq \frac{C_\Phi \epsilon^*}{\beta} \quad (40)$$

Proof. We have

$$\begin{aligned} \|D_n\| &= \left\| \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s), y(s)) \varepsilon_{k+1} ds \right\| \\ &\leq \epsilon^* \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \|\Phi(t_n, s; \bar{x}(s), y(s))\| ds \end{aligned} \quad (41)$$

$$\begin{aligned} &\leq C_\Phi \epsilon^* \int_{t_0}^{t_n} e^{-\beta(t_n-s)} ds \\ &\leq \frac{C_\Phi \epsilon^*}{\beta}, \end{aligned} \quad (42)$$

where (41) and (42) follow from (6) and (†) respectively. Therefore, for all $n \geq 0$, we have

$$\mathbb{E}[\|D_n\|^2]^{1/2} \leq \frac{C_\Phi \epsilon^*}{\beta}. \quad (43)$$

■

3.2 Bound on E_n

Lemma 2. For E_n defined in (35),

$$\mathbb{E}[\|E_n\|^2]^{1/2} \leq \frac{K_\gamma L_\lambda C_\Phi \epsilon}{\beta} \quad (44)$$

where $K_\gamma := \max_{\|y\| \leq C_\gamma} \|\gamma(y)\|$.

Proof. We have,

$$\begin{aligned} \|E_n\| &= \left\| \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; x(s), y(s)) \times \epsilon \nabla \lambda(y(t_k)) \gamma(y(t_k)) ds \right\| \\ &\leq \epsilon \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \|\Phi(t_n, s; x(s), y(s))\| \times \|\nabla \lambda(y(t_k))\| \times \|\gamma(y(t_k))\| ds, \end{aligned} \quad (45)$$

$$\leq \epsilon K_\gamma L_\lambda \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \|\Phi(t_n, s; x(s), y(s))\| ds. \quad (46)$$

Using (†),

$$\|E_n\| \leq \epsilon K_\gamma L_\lambda C_\Phi \int_{t_0}^{t_n} e^{-\beta(t_n-s)} ds \quad (47)$$

$$\leq \frac{K_\gamma L_\lambda C_\Phi \epsilon}{\beta}. \quad (48)$$

Hence

$$\mathbb{E}[\|E_n\|^2]^{1/2} \leq \frac{K_\gamma L_\lambda C_\Phi \epsilon}{\beta}. \quad (49)$$

■

3.3 Bound on A_n

The next lemma is a variant of Lemma 5.5 of [43].

Lemma 3. For a suitable constant $K_1 > 0$,

$$\begin{aligned} &\int_{t_k}^{t_{k+1}} e^{-\beta(t_n-s)} \|\bar{x}(s) - \bar{x}(t_k)\| ds \\ &\leq [K_1 + G_{\bar{h}}] \|\bar{x}(t_k)\| + \|M_{k+1}\| e^{-\beta(t_n-t_{k+1})} a^2. \end{aligned}$$

Proof. Using (20) and (1), for $s \in [t_k, t_{k+1}]$,

$$\begin{aligned}
\|\bar{x}(s) - \bar{x}(t_k)\| &= \frac{s - t_k}{a} \|\bar{x}(t_{k+1}) - \bar{x}(t_k)\| \\
&= (s - t_k) \|\tilde{h}(\bar{x}(t_k), y(t_k)) + M_{k+1} + \varepsilon_{k+1}\| \\
&\leq (s - t_k) [\|\tilde{h}(\bar{x}(t_k), y(t_k))\| + \|M_{k+1}\| + \|\varepsilon_{k+1}\|] \\
&\leq (s - t_k) [G_{\tilde{h}}(1 + \|\bar{x}(t_k)\| + \|y(t_k)\|) + \|M_{k+1}\| + \varepsilon^*] \quad (50) \\
&\leq (s - t_k) [G_{\tilde{h}}(1 + C_\gamma) + G_{\tilde{h}}\|\bar{x}(t_k)\| + \|M_{k+1}\| + \varepsilon^*] \\
&\leq (s - t_k) [K_1 + G_{\tilde{h}}\|\bar{x}(t_k)\| + \|M_{k+1}\|], \quad (51)
\end{aligned}$$

where $K_1 = G_{\tilde{h}}(1 + C_\gamma) + \varepsilon^*$. Also,

$$\int_{t_k}^{t_{k+1}} (s - t_k) e^{-\beta(t_n - s)} ds \leq e^{-\beta(t_n - t_{k+1})} a^2.$$

Therefore

$$\begin{aligned}
\int_{t_k}^{t_{k+1}} e^{-\beta(t_n - s)} \|\bar{x}(s) - \bar{x}(t_k)\| ds &\leq \\
&[K_1 + G_{\tilde{h}}\|\bar{x}(t_k)\| + \|M_{k+1}\|] e^{-\beta(t_n - t_{k+1})} a^2.
\end{aligned}$$

■

Lemma 4. For A_n as defined in (31),

$$\mathbb{E}[\|A_n\|^2]^{1/2} = O(a).$$

Proof. We have

$$\begin{aligned}
\|A_n\| &= \left\| \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; x(s), y(s)) \times [\tilde{h}(\bar{x}(t_k), y(t_k)) - \tilde{h}(\bar{x}(s), y(s))] ds \right\| \\
&\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \|\Phi(t_n, s; x(s), y(s))\| \\
&\quad \times \|\tilde{h}(\bar{x}(t_k), y(t_k)) - \tilde{h}(\bar{x}(s), y(s))\| ds \\
&\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} C_\Phi e^{-\beta(t_n - s)} L_{\tilde{h}} \times [\|\bar{x}(t_k) - \bar{x}(s)\| + \|y(t_k) - y(s)\|] \quad (52) \\
&\leq C_\Phi L_{\tilde{h}} \sum_{k=0}^{n-1} \left[[K_1 + G_{\tilde{h}}\|\bar{x}(t_k)\| + \|M_{k+1}\|] e^{-\beta(t_n - t_{k+1})} a^2 \right. \\
&\quad \left. + K_\gamma a \int_{t_k}^{t_{k+1}} (s - t_k) e^{-\beta(t_n - s)} ds \right] \quad (53) \\
&\leq C_\Phi L_{\tilde{h}} \sum_{k=0}^{n-1} [K_1 + G_{\tilde{h}}\|\bar{x}(t_k)\| + \|M_{k+1}\|] e^{-\beta(t_n - t_{k+1})} a^2
\end{aligned}$$

$$\begin{aligned}
& + K_\gamma a \epsilon e^{-\beta(t_n - t_{k+1})} a^2 \Big] \\
& = C_\Phi L_{\tilde{h}} \sum_{k=0}^{n-1} [K_1 + G_{\tilde{h}} \|\bar{x}(t_k)\| + K_\gamma a \epsilon + \|M_{k+1}\|] \times e^{-\beta(t_n - t_{k+1})} a^2 \\
& \leq a C_\Phi L_{\tilde{h}} \left((K_1 + K_\gamma a \epsilon) \mu + \sum_{k=0}^{n-1} (G_{\tilde{h}} \|\bar{x}(t_k)\| + \|M_{k+1}\|) a e^{-\beta(t_n - t_{k+1})} \right),
\end{aligned} \tag{54}$$

where $\mu := \frac{1}{\beta}$. Equation (53) follows from Lemma 3. Denote the terms $C_\Phi L_{\tilde{h}}$, $K_1 + K_\gamma a \epsilon$, $\sum_{k=0}^{n-1} \|M_{k+1}\| a e^{-\beta(t_n - t_{k+1})}$ and $\sum_{k=0}^{n-1} a \|\bar{x}(t_k)\| e^{-\beta(t_n - t_{k+1})}$ by K_2 , K_3 , F_n and \tilde{F}_n . Note that K_3 is $O(1)$. Then

$$\begin{aligned}
\|A_n\| & \leq a K_2 (K_3 \mu + F_n + G_{\tilde{h}} \tilde{F}_n), \\
\mathbb{E}[\|A_n\|^2]^{1/2} & \leq a K_2 \left(K_3 \mu + \mathbb{E}[F_n^2]^{1/2} + G_{\tilde{h}} \mathbb{E}[\tilde{F}_n^2]^{1/2} \right).
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}[\tilde{F}_n^2]^{1/2} & = \sum_{k=0}^{n-1} \left(a \mathbb{E}[\|\bar{x}(t_k)\|^2]^{1/2} e^{-\beta(t_n - t_{k+1})} \right) \\
& \leq C^* \mu
\end{aligned} \tag{55}$$

and

$$\begin{aligned}
\mathbb{E}[F_n^2]^{1/2} & = \sum_{k=0}^{n-1} \mathbb{E}[\|M_{k+1}\|^2]^{1/2} a e^{-\beta(n-k)a} \\
& \leq \sum_{i=0}^{n-1} \frac{\sqrt{2C_M}}{\delta} a e^{-\beta(n-k)a} \\
& \leq K_4 \mu
\end{aligned} \tag{56}$$

where $K_4 = \frac{\sqrt{2C_M}}{\delta}$. Therefore

$$\mathbb{E}[\|A_n\|^2]^{1/2} \leq a K_2 \left(K_3 \mu + K_4 \mu + G_{\tilde{h}} C^* a \mu \right).$$

Hence $\mathbb{E}[\|A_n\|^2]^{1/2} = O(a)$. ■

3.4 Bound on $B_n - C_n$

Lemma 5. For B_n and C_n defined in (32) and (33)

$$\mathbb{E}[\|B_n - C_n\|^2]^{1/2} = O(a).$$

Proof. From (32) and (33) we have

$$B_n - C_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} [\Phi(t_n, s; \bar{x}(s), y(s)) - \Phi(t_n, s; \bar{x}(t_k), y(t_k))] M_{k+1} ds$$

Therefore

$$\begin{aligned} \|B_n - C_n\| &= \left\| \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} [\Phi(t_n, s; \bar{x}(s), y(s)) - \Phi(t_n, s; \bar{x}(t_k), y(t_k))] M_{k+1} ds \right\| \\ &\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \|\Phi(t_n, s; \bar{x}(s), y(s)) - \Phi(t_n, s; \bar{x}(t_k), y(t_k))\| \\ &\quad \times \|M_{k+1}\| ds. \end{aligned}$$

From (28), we know that $\Phi(t, s; \bar{x}(s), y(s))$ and $\Phi(t, s; \bar{x}(t_k), y(t_k))$ are fundamental matrices for the linear systems given by: for $t \geq s$,

$$\dot{\chi}(t, s; \bar{x}(s), y(s)) = D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s))) \chi(t, s; \bar{x}(s), y(s)), \quad (57)$$

and

$$\dot{\tilde{\chi}}(t, s; \bar{x}(t_k), y(t_k)) = D(z(t, s; \bar{x}(t_k), y(t_k)), y(t, s; y(t_k))) \tilde{\chi}(t, s; \bar{x}(t_k), y(t_k)). \quad (58)$$

So $\Phi(t, s; \bar{x}(s), y(s))$ and $\Phi(t, s; \bar{x}(t_k), y(t_k))$ satisfy the following matrix valued differential equations

$$\dot{\Phi}(t, s; \bar{x}(s), y(s)) = D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s))) \Phi(t, s; \bar{x}(s), y(s)), \quad (59)$$

and

$$\dot{\Phi}(t, s; \bar{x}(t_k), y(t_k)) = D(z(t, s; \bar{x}(t_k), y(t_k)), y(t, s; y(t_k))) \Phi(t, s; \bar{x}(t_k), y(t_k)). \quad (60)$$

For each column indexed by j , the differential equations (59) and (60) can be equivalently written as

$$\dot{\Phi}_j(t, s; \bar{x}(s), y(s)) = D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s))) \Phi_j(t, s; \bar{x}(s), y(s)), \quad (61)$$

and

$$\begin{aligned} \dot{\Phi}_j(t, s; \bar{x}(t_k), y(t_k)) &= D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s))) \Phi_j(t, s; \bar{x}(t_k), y(t_k)) + \\ &\quad [D(z(t, s; \bar{x}(t_k), y(t_k)), y(t, s; y(t_k))) \\ &\quad - D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s)))] \\ &\quad \times \Phi_j(t, s; \bar{x}(t_k), y(t_k)). \end{aligned} \quad (62)$$

Treating (62) as a perturbation of (61) and applying Alexseev's formula (12)³ to each column of $\Phi(\bullet, \bullet; \bullet, \bullet)$, we have

$$\begin{aligned}
& \Phi_j(t_n, s; \bar{x}(t_k), y(t_k)) - \Phi_j(t_n, s; \bar{x}(s), y(s)) \\
&= \int_s^{t_n} \Phi(t_n, t; \bar{x}(t), y(t)) \\
&\quad \times [D(z(t, s; \bar{x}(t_k), y(t_k)), y(t, s; y(t_k))) - D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s)))] \\
&\quad \times \Phi_j(t, s; \bar{x}(t_k), y(t_k)) dt \tag{63}
\end{aligned}$$

Combining the equations (63) for all columns, we get

$$\begin{aligned}
& \Phi(t_n, s; \bar{x}(t_k), y(t_k)) - \Phi(t_n, s; \bar{x}(s), y(s)) \\
&= \int_s^{t_n} \Phi(t_n, t; \bar{x}(s), y(s)) \\
&\quad \times [D(z(t, s; \bar{x}(t_k), y(t_k)), y(t, s; y(t_k))) - D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s)))] \\
&\quad \times \Phi(t, s; \bar{x}(t_k), y(t_k)) dt
\end{aligned}$$

Therefore

$$\begin{aligned}
\|B_n - C_n\| &\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} \|\Phi(t_n, t; \bar{x}(s), y(s))\| \\
&\quad \times \| [D(z(t, s; \bar{x}(t_k), y(t_k)), y(t, s; y(t_k))) \\
&\quad - D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s)))] \| \\
&\quad \times \|\Phi(t, s; \bar{x}(t_k), y(t_k))\| \times \|M_{k+1}\| dt ds \\
&\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_{\Phi}^2 L_D \times e^{-\beta(t_n-s)} \times e^{-\beta(t-s)} \times \\
&\quad \left[\|z(t, s; \bar{x}(s), y(s)) - z(t, s; \bar{x}(t_k), y(t_k))\| \right. \\
&\quad \left. + \|y(s) - y(t_k)\| \right] \times \|M_{k+1}\| dt ds \tag{64}
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_{\Phi}^2 L_D \times e^{-\beta(t_n-s)} \times e^{-\beta(t-s)} \times \\
&\quad \left[[\|\bar{x}(s) - \bar{x}(t_k)\| + \|y(s) - y(t_k)\|] C_{\Phi} e^{-\beta(t-s)} \right. \\
&\quad \left. + \|y(s) - y(t_k)\| \right] \times \|M_{k+1}\| dt ds, \tag{65}
\end{aligned}$$

where (64) follows from (†) and the Lipschitz property of $D(\cdot, \cdot)$ while (65)

³in fact, the classical variation of constants formula for linear systems which it generalizes

follows from (14) and (†). We split the analysis into two terms as follows:

$$\begin{aligned}
G_n &:= \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_{\Phi}^3 L_D e^{-\beta(t_n-s)} \|\bar{x}(s) - \bar{x}(t_k)\| \times e^{-2\beta(t-s)} \times \|M_{k+1}\| dt ds \\
&= \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} C_{\Phi}^3 L_D e^{-\beta(t_n-s)} \|\bar{x}(s) - \bar{x}(t_k)\| \times \frac{1 - e^{-2\beta(t_n-s)}}{2\beta} \times \|M_{k+1}\| ds \\
&= K_5 \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} e^{-\beta(t_n-s)} \times \|\bar{x}(s) - \bar{x}(t_k)\| \times (1 - e^{-2\beta(t_n-s)}) \times \|M_{k+1}\| ds \\
&\leq K_5 \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} e^{-\beta(t_n-s)} \|\bar{x}(s) - \bar{x}(t_k)\| \times \|M_{k+1}\| ds \\
&\leq K_5 \sum_{k=0}^{n-1} a^2 e^{-\beta(t_n-t_{k+1})} [K_1 \|M_{k+1}\| + \|M_{k+1}\|^2 + G_{\bar{h}} \|\bar{x}(t_k)\| \|M_{k+1}\|]
\end{aligned} \tag{66}$$

where K_5 denotes $C_{\Phi}^3 L_D / 2\beta$ and (66) follows from Lemma 3,

$$\begin{aligned}
H_n &:= \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_{\Phi}^3 L_D \times e^{-\beta(t_n-s)} \|y(s) - y(t_k)\| e^{-2\beta(t-s)} \|M_{k+1}\| dt ds \\
&\quad + \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_{\Phi}^2 L_D \times e^{-\beta(t_n-s)} \|y(s) - y(t_k)\| e^{-\beta(t-s)} \|M_{k+1}\| dt ds \\
&= \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} C_{\Phi}^3 L_D e^{-\beta(t_n-s)} \|y(s) - y(t_k)\| \left[\frac{1}{2\beta} (1 - e^{-2\beta(t_n-s)}) \right] \|M_{k+1}\| ds \\
&\quad + \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} C_{\Phi}^2 L_D e^{-\beta(t_n-s)} \|y(s) - y(t_k)\| \left[\frac{1}{\beta} (1 - e^{-\beta(t_n-s)}) \right] \|M_{k+1}\| ds \\
&\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \frac{C_{\Phi}^2 (\frac{1}{2} + C_{\Phi}) L_D}{\beta} e^{-\beta(t_n-s)} K_{\gamma} a \epsilon (s - t_k) \|M_{k+1}\| ds \\
&\leq K_6 \sum_{k=0}^{n-1} e^{-\beta(t_n-t_{k+1})} a^3 \|M_{k+1}\|
\end{aligned} \tag{67}$$

where (67) follows from (5) and $K_6 := (\frac{1}{2} + C_{\Phi}) C_{\Phi}^2 K_{\gamma} L_D \epsilon / \beta$. Further define $G_{1,n}$, $G_{2,n}$ and $G_{3,n}$ as follows

$$\begin{aligned}
G_{1,n} &= \sum_{k=0}^{n-1} a e^{-\beta(t_n-t_{k+1})} \|M_{k+1}\|, \\
G_{2,n} &= \sum_{k=0}^{n-1} a e^{-\beta(t_n-t_{k+1})} \|M_{k+1}\|^2,
\end{aligned}$$

$$G_{3,n} = \sum_{k=0}^{n-1} a e^{-\beta(t_n - t_{k+1})} \|\bar{x}(t_k)\| \|M_{k+1}\|.$$

Then

$$\begin{aligned} \|B_n - C_n\| &\leq G_n + H_n \\ &\leq K_5 a (K_1 G_{n,1} + G_{n,2} + G_{\bar{h}} G_{n,3}) + (K_6 a^2 G_{n,1}). \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{E}[\|B_n - C_n\|^2]^{1/2} &\leq K_5 a (K_1 \mathbb{E}[G_{n,1}^2]^{1/2} + \mathbb{E}[G_{n,2}^2]^{1/2} + G_{\bar{h}} \mathbb{E}[G_{n,3}^2]^{1/2}) + \\ &\quad a^2 K_6 \mathbb{E}[G_{n,1}^2]^{1/2} \end{aligned}$$

We now bound each of the terms in the previous expression. Using a calculation similar to the one used for (56), we have

$$\begin{aligned} \mathbb{E}[G_{n,1}^2]^{1/2} &\leq K_4 \mu, \tag{68} \\ \mathbb{E}[G_{n,2}^2]^{1/2} &= \sum_{k=0}^{n-1} E[\|M_{k+1}\|^4]^{1/2} a e^{-\beta(n-k)a} \\ &\leq \sum_{k=0}^{n-1} \frac{\sqrt{24C_M}}{\delta^2} a e^{-\beta(n-k)a} \\ &\leq K_7 \left(\sum_{k=0}^{n-1} a e^{-\beta(n-k)a} \right) \\ &\leq K_7 \mu, \tag{69} \end{aligned}$$

where $K_7 = \sqrt{24C_M}/\delta^2$,

$$\begin{aligned} \mathbb{E}[G_{n,3}^2]^{1/2} &= \mathbb{E} \left[\sum_{k=0}^{n-1} a e^{-\beta(t_n - t_{k+1})} E \left[\|\bar{x}(t_k)\|^2 \|M_{k+1}\|^2 \right]^{1/2} \right] \\ &\leq \sum_{k=0}^{n-1} a e^{-\beta(t_n - t_{k+1})} \left(\mathbb{E}[\|\bar{x}(t_k)\|^4] \mathbb{E}[\|M_{k+1}\|^4] \right)^{1/4} \\ &\leq \sum_{k=0}^{n-1} a e^{-\beta(t_n - t_{k+1})} C^* \left(\mathbb{E}[\|M_{k+1}\|^4] \right)^{1/2} \\ &\leq \sum_{k=0}^{n-1} a e^{-\beta(t_n - t_{k+1})} C^* \left(\frac{(24C_M)^{1/4}}{\delta} \right) \\ &\leq K_8 \mu, \tag{70} \end{aligned}$$

where $K_8 = \frac{C^* (24C_M)^{1/4}}{\delta}$.

Using (68), (69) and (70), we have

$$\begin{aligned}\mathbb{E}[\|B_n - C_n\|^2]^{\frac{1}{2}} &\leq K_5 a (K_1 K_4 \mu + K_7 \mu + G_{\bar{h}} K_8 \mu) + a^2 K_6 K_8 \mu \\ &= O(a).\end{aligned}\tag{71}$$

■

3.5 Bound on C_n

Lemma 6. For C_n defined in (33)

$$\mathbb{E}[\|C_n\|^2]^{1/2} = O(\max\{a^{1.5} d^{3.25}, a^{0.5} d^{2.5}\}).$$

Proof. It is easy to verify that C_n satisfies the condition for the martingale concentration inequality provided in Theorem 5.1 in Appendix, with

$$\begin{aligned}\alpha_{k,n} &= \int_{t_k}^{t_{k+1}} \Phi(t_n, s, \bar{x}(t_k), y(t_k)) ds, \\ \gamma_1 &= \frac{C_{\Phi}}{\beta}, \quad \gamma_2 = 1, \quad \beta_n = a,\end{aligned}$$

for $k, n \geq 0$. Thus,

$$\begin{aligned}E[\|C_n\|^2] &= \int_0^{\infty} P(\|C_n\|^2 \geq s) ds \\ &= \int_0^{\infty} P(\|C_n\| \geq \sqrt{s}) ds\end{aligned}$$

Using the martingale concentration inequality provided in 5.1 in Appendix, we have

$$\begin{aligned}E[\|C_n\|^2] &= \int_0^{K_9} 2d^2 \exp\left(\frac{-cs}{d^3 a}\right) ds \\ &\quad + \int_{K_9}^{\infty} 2d^2 \exp\left(\frac{-c\sqrt{s}}{d^{3/2} a}\right) ds,\end{aligned}\tag{72}$$

where $K_9 = \frac{C_M \gamma_1 d^{1.5}}{\delta}$. Analysing the terms separately, we have

$$\begin{aligned}\int_0^{K_9} 2d^2 \exp\left(\frac{-cs}{d^3 a}\right) ds &= \frac{2d^5 a}{c} \left(1 - \exp\left(\frac{-cK_9}{d^3 a}\right)\right) \\ &\leq \left(\frac{2d^5}{c}\right) a,\end{aligned}\tag{73}$$

and

$$\int_{K_9}^{\infty} 2d^2 \exp\left(\frac{-c\sqrt{s}}{d^{3/2} a}\right) ds = \frac{4d^5 a}{c^2} \exp\left(\frac{-c\sqrt{K_9}}{d^{3/2} a}\right) \left(a + \frac{c\sqrt{K_9}}{d^{3/2}}\right)$$

$$\leq \frac{4d^5 a}{c^2} \left(\frac{ad^{3/2}}{c\sqrt{K_9}} \right) \left(a + \frac{c\sqrt{K_9}}{d^{3/2}} \right) \quad (74)$$

$$= O(\max\{a^3 d^{6.5}, a^2 d^5\}), \quad (75)$$

where (74) follows from the fact that $e^{-1/a} \leq a$ for $a > 0$. From (73) and (75), we have

$$\begin{aligned} \mathbb{E}[\|C_n\|^2] &\leq \left(\frac{2d^5}{c} \right) a + O(\max\{a^3 d^{6.5}, a^2 d^5\}) \\ &= O(\max\{a^3 d^{6.5}, a^2 d^5\}) \\ \therefore \mathbb{E}[\|C_n\|^2]^{1/2} &= O(\max\{a^{1.5} d^{3.25}, a^{0.5} d^{2.5}\}). \end{aligned}$$

■

4 Main result

Combining the foregoing bounds leads to our main result stated as follows.

Theorem 4.1. *The mean square deviation of tracked iterates from a non-stationary trajectory satisfies:*

$$\begin{aligned} \mathbb{E}[\|x_n - \lambda(y(n))\|^2]^{1/2} &\leq \frac{C_\Phi \varepsilon^*}{\beta} + \frac{K_\gamma L_\lambda C_\Phi \epsilon}{\beta} \\ &\quad + O(\max\{a^{1.5} d^{3.25}, a^{0.5} d^{2.5}\}) \\ &\quad + C_\Phi e^{-\beta(t_n - t_0)} \|x_0 - \lambda(y(0))\| \end{aligned} \quad (76)$$

Proof. Using (38), (†) and lemmas 1-6, we get

$$\begin{aligned} \mathbb{E}[\|\bar{x}(t_n) - z(t_n)\|^2]^{1/2} &\leq \frac{C_\Phi \varepsilon^*}{\beta} + \frac{K_\gamma L_\lambda C_\Phi \epsilon}{\beta} + O(a) \\ &\quad + O(\max\{a^{1.5} d^{3.25}, a^{0.5} d^{2.5}\}) \\ &\quad + C_\Phi e^{-\beta(t_n - t_0)} \|\bar{x}(t_0) - z(t_0)\| \\ &= \frac{C_\Phi \varepsilon^*}{\beta} + \frac{K_\gamma L_\lambda C_\Phi \epsilon}{\beta} \\ &\quad + O(\max\{a^{1.5} d^{3.25}, a^{0.5} d^{2.5}\}) \\ &\quad + C_\Phi e^{-\beta(t_n - t_0)} \|\bar{x}(t_0) - z(t_0)\|. \end{aligned}$$

The claim follows. ■

Remark: 1. The $O(\cdot)$ notation is used above to isolate the dependence on the stepsize a . The exact constants involved are available in the relevant lemmas, but are suppressed in order to improve clarity.

2. The linear complexity of the error bound in ε^* and ϵ is natural to expect, these being contributions from bounded additive error component ε_n and rate of variation of the tracking signal, respectively. The $O(\cdot)$ term is due to the martingale noise and discretization. The last term accounts for the effect of initial condition.

3. By setting $\epsilon = 0$ in (76), we can recover as a special case a bound valid for all time for a stationary target. Then $y(\cdot) \equiv y^*$, a constant, and $z(\cdot) \equiv x^* = \lambda(y^*)$, also a constant, viz., an equilibrium for the system $\dot{x}(t) = h(x(t), y^*)$.

5 Conclusion and Future Work

We analyzed a constant step-size stochastic approximation algorithm for tracking a slowly varying dynamical system and obtained a *non-asymptotic* bound *valid for all time*, with dependence on step-size and dimension explicitly given. The latter in particular provides insight into step-size selection in high dimensional regime.

A natural extension would be to the problem of tracking a stochastic dynamics. Indeed, a suitable extension of Alekseev's formula is available for this purpose [48], which is much more complex.

Appendix: A martingale concentration inequality

We state here the martingale concentration inequality we have used, from [43], which in turn is a slight adaptation of the results of [33].

Theorem 5.1. *Let $S_n = \sum_{k=1}^n \alpha_{k,n} X_k$, where X_k is a \mathbb{R}^d valued \mathcal{F}_k - adapted martingale difference sequence and $\alpha_{k,n}$ is a sequence of bounded pre-visible real valued $d \times d$ random matrices, i.e., $\alpha_{k,n} \in \mathcal{F}_{k-1}$ and there exists finite number, say $A_{k,n}$, such that $\|\alpha_{k,n}\| \leq A_{k,n}$. Suppose that for some $\delta, C > 0$*

$$\mathbb{E}[e^{\delta\|X_k\|} \mid \mathcal{F}_{k-1}] \leq C, \quad k \geq 1.$$

Further assume that there exist constants $\gamma_1, \gamma_2 > 0$, independent of n , so that $\sum_{k=1}^n A_{k,n} \leq \gamma_1$ and $\max_{1 \leq k \leq n} A_{k,n} \leq \gamma_2 \beta_n$, where β_n is some positive sequence. Then for $\eta > 0$, there exists some constant $c > 0$ depending on $\delta, C, \gamma_1, \gamma_2$ such that

$$P(\|S_n\| > \eta) \leq \begin{cases} 2d^2 e^{-\frac{c\eta^2}{d^3\beta_n}} & \text{if } \eta \in \left(0, \frac{C\gamma_1 d^{1.5}}{\delta}\right], \\ 2d^2 e^{-\frac{c\eta}{d^{1.5}\beta_n}} & \text{otherwise.} \end{cases}$$

References

- [1] Alekseev, V. M. “An estimate for the perturbations of the solutions of ordinary differential equations.” *Westnik Moskov Unn. Ser 1*, pp. 28–36, 1961.
- [2] Anantharam, V. and Borkar, V. S., “Stochastic approximation with long range dependent and heavy tailed noise.” *Queueing Systems* 71.1-2 pp. 221-242, 2012.
- [3] Bach, F. and Moulines, E., “Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$.”, *NIPS*, pp. 773-781, 2013.
- [4] Benveniste, A. “Design of adaptive algorithms for the tracking of time-varying systems.” *International Journal of adaptive control and signal processing*, 19.1 pp. 3–29, 1987.
- [5] Benveniste, A. and Ruget, G. “A measure of the tracking capability of recursive stochastic algorithms with constant gains.” *IEEE Transactions on Automatic Control*, 47.3 pp. 639–649, 1982.
- [6] Borkar, V. S., “On trapping probability of stochastic approximation.” *Combinatorics, Probability and Computing* 11.1, pp. 11-20, 2002.

- [7] Borkar, A.V. and Borkar, V.S. and Sinha, A. “Aerial monitoring of slow moving convoys using elliptical orbits.” *European Journal of Control*, available online, 2018.
- [8] Borkar, V. S. “Stochastic Approximation: A Dynamical Systems Viewpoint.” *Hindustan Publishing Agency, New Delhi, and Cambridge University Press, Cambridge, UK*, 2008.
- [9] Borkar, V. S. “Probability Theory: An Advanced Course.” *Springer Science & Business Media, New York*, 2012.
- [10] Borkar, V. S. and Meyn, S. P. “The ODE method for convergence of stochastic approximation and reinforcement learning.” *SIAM Journal on Control and Optimization* 38.2, pp. 447–469, 2000.
- [11] Brauer, F. “Perturbations of non-linear systems of differential equations.” *Journal of Mathematical Analysis and Applications* 14.2, pp. 198–206, 1966.
- [12] Bucklew, J. A. and Kurtz, T. G., “Weak convergence and local stability properties of fixed step size recursive algorithms.” *IEEE Transactions on Information Theory* 39(3), pp. 966-978, 1993,.
- [13] Burke, W. L., “Gravitational radiation damping of slowly moving systems calculated using matched asymptotic expansions.” *Journal of Mathematical Physics* 12.3, pp. 401–418, 1971.
- [14] Chappell, C. F. “Quasi-stationary convective events.” *Mesoscale Meteorology and Forecasting, Springer*, pp. 289-310, 1986.
- [15] Chen, Han-Fu, “Stochastic Approximation and Its Applications.” *Springer Science & Business Media* 64, 2006.
- [16] Dalal, G. and Szorenyi, B. and Thoppe, G. and Mannor, S. “Concentration Bounds for Two Time-scale Stochastic Approximation with Applications to Reinforcement Learning.” arXiv preprint *arXiv:1703.05376*, 2017.
- [17] Derevitskii, D. P. and Fradkov, A. L., “Two models analyzing the dynamics of adaptation algorithms”, *Automation and Remote Control* 35(1), pp. 59–67, 1974.
- [18] Diamantaras, K. I. and Kung, S. Y. “Principal Component Neural Networks: Theory and Applications.” *John Wiley & Sons, Inc.*, 1996.
- [19] Eweda, E. “Comparison of RLS, LMS, and sign algorithms for tracking randomly time-varying channels.” *IEEE Transactions on Signal Processing* 42.11, pp. 2937–2944, 1994.
- [20] Farden, D. “Tracking properties of adaptive signal processing algorithms.” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29.3 pp. 439-446 1981.

- [21] Finnoff, W. “Diffusion approximations for the constant learning rate back-propagation algorithm and resistance to local minima.” *Advances in Neural Information Processing Systems*, pp. 459-466, 1993.
- [22] Guo, L. and Ljung, L. “Exponential stability of general tracking algorithms.” *IEEE Transactions on Automatic Control*, 40.8 pp. 1376-1387, 1995.
- [23] Guo, L. and Ljung, L. “Performance analysis of general tracking algorithms.” *IEEE Transactions on Automatic Control* 40.8 pp. 1388–1402, 1995.
- [24] Guo, L., Ljung, L. and Wang, G.-J., “Necessary and sufficient conditions for stability of LMS.” *IEEE Transactions on Automatic Control* 42.6 pp. 76-1-770, 1997.
- [25] Joslin, J. A. and Heunis, A. J., “Law of the iterated logarithm for a constant-gain linear stochastic gradient algorithm.” *SIAM Journal on Control and Optimization* 39.2, pp. 533-570, 2000.
- [26] Joulin, A., “On maximal inequalities for stable stochastic integrals”, *Potential Analysis* 26.1, pp. 57-78, 2007.
- [27] Khalil, H. K., “Nonlinear Systems” (3rd ed.), *Prentice Hall, Englewood Cliffs, NJ*, 1996.
- [28] Kuan, C. M. and Hornik, K. “Convergence of learning algorithms with constant learning rates.” *IEEE Transactions on Neural Networks*, 2.5 pp. 484–489, 1991.
- [29] Kushner, H. J. and Hai H. “Averaging methods for the asymptotic analysis of learning and adaptive systems, with small adjustment rate.” *SIAM Journal on Control and Optimization* 19.5 pp. 635-650, 1981.
- [30] Kushner, H. J. and Huang, H. “Asymptotic properties of stochastic approximations with constant coefficients.” *SIAM Journal on Control and Optimization* 19.1 pp. 87–105, 1981.
- [31] Kushner, H. J. and Yin, G. G. “Stochastic Approximation Algorithms and Applications.” *Springer Verlag, New York*, 1997.
- [32] Laxminarayanan, C. and Szepesvari, C., “Linear stochastic approximation: how far does constant stepsize and iterate averaging go?” *Proc. 21st Intel. Conf. on Artificial Intelligence and Statistics (AISTATS)*, Lazarote, Spain, 2018.
- [33] Liu, Q., and Watbled, F. “Exponential inequalities for martingales and asymptotic properties for the free energy of directed polymers in a random environment.” *Stochastic Processes and Their Applications* 119.10 pp. 3101-3132, 2009.

- [34] Ljung, L. “Analysis of recursive stochastic algorithms”, *IEEE Transactions on Automatic Control* 22(4), pp. 551–575, 1977.
- [35] Ng, S. C. and Leung, S. H. and Luk, A. “Fast convergent generalized back-propagation algorithm with constant learning rate.” *Neural Processing Letters*, 9.1 pp. 13–23, 1999.
- [36] Peuteman, J. and Aeyels, D., “Exponential stability of slowly time-varying non-linear systems.” *Mathematics of Control, Signals and Systems* 15, pp. 202-228, 2002.
- [37] Pflug, G. Ch. “Stochastic minimization with constant step-size: asymptotic laws.” *SIAM Journal on Control and Optimization* 24.4 pp. 655–666, 1986.
- [38] Pflug, G. Ch. “Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size.” *Monatshefte für Mathematik* 110.3 pp. 297-314, 1990.
- [39] Robbins, H. and Monro, J., “A stochastic approximation method.” *The Annals of Mathematical Statistics* 22.3, pp. 400–407, 1951.
- [40] Rugh, W. J., “Linear System Theory.” *Prentice Hall, Englewood Cliffs, NJ*, 1993.
- [41] Sharma, R. and Sethares, W. A., and Bucklew, J. A., “Asymptotic analysis of stochastic gradient-based adaptive filtering algorithms with general cost functions.” *IEEE Transactions on Signal Processing*, 44.9 pp. 2186-2194, 1996.
- [42] Solo, Victor, “On the stability of slowly time-varying linear systems.” *Mathematics of Control, Signals, and Systems (MCSS)* 7.4, pp. 331–350, 1994.
- [43] Thoppe, G. and Borkar, V. S. “A concentration bound for stochastic approximation via Alekseev’s formula.” *Stochastic Systems*, to appear *arXiv:1506.08657*, 2019.
- [44] Utkin, V. and Guldner, J. and Shi, J. “Sliding Mode Control in Electro-mechanical Systems.” *CRC press*, 2017.
- [45] Wilson, C. and Veeravalli, V. and Nedic A. “Adaptive Sequential Stochastic Optimization.” arXiv preprint *arXiv:1610.01970*, 2018.
- [46] Yin , G. and Ion, C. and Krishnamurthy V. “How does a stochastic optimization/approximation algorithm adapt to a randomly evolving optimum/root with jump Markov sample paths.” *Mathematical Programming* 120.1 pp. 67-99, 2009.
- [47] Yin , G. and Krishnamurthy V. and Ion, C. “Regime switching stochastic approximation algorithms with application To adaptive discrete stochastic optimization.” *SIAM Journal on Optimization* 14.4 pp. 1187-1215, 2004.

- [48] Zerihun, T. and Ladde, G. S., “Fundamental properties of solutions of nonlinear stochastic differential equations and Method of variation of parameters.” *Dynamical Systems and Applications* 22, 2013, pp. 433-458.
- [49] Zhou, B., “On asymptotic stability of linear time-varying systems.” *Automatica* 68, pp. 266–276, 2016.
- [50] Zhou, B., “Stability analysis of non-linear time-varying systems by Lyapunov functions with indefinite derivatives.” *IET Control Theory & Applications* 9, pp. 1434–1442, 2017.
- [51] Zhu, J. and Spall, J. C. “Tracking capability of stochastic gradient algorithm with constant gain.” *55th IEEE Conf. on Decision and Control (CDC)*, Las Vegas, pp. 4522–4527, 2016.