Non-asymptotic error bounds for constant stepsize stochastic approximation for tracking mobile agents

Bhumesh Kumar\textsuperscript{1,2} \cdot Vivek Borkar\textsuperscript{1} \cdot Akhil Shetty\textsuperscript{1,3}

Received: 21 February 2018 / Accepted: 5 October 2019 / Published online: 25 October 2019
© Springer-Verlag London Ltd., part of Springer Nature 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 nonlinear 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 \cdot Constant stepsize \cdot Non-asymptotic bound \cdot Alekseev’s formula \cdot Martingale concentration inequalities \cdot Perturbation analysis \cdot Non-stationary optimization

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.

Bhumesh Kumar
bkumar@wisc.edu

Vivek Borkar
borkar.vs@gmail.com

Akhil Shetty
shetty.akhil@berkeley.edu

1 Department of Electrical Engineering, IIT Bombay, Mumbai, Maharashtra 400076, India
2 Department of Electrical and Computer Engineering, University of Wisconsin at Madison, Madison, WI 53706, USA
3 Department of Electrical Engineering and Computer Science, University of California at Berkeley, Cory Hall, Hearst Avenue, Berkeley, CA 94720, USA
1 Introduction

1.1 Background

Robbins and Monro proposed in [36] a stochastic iterative scheme

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

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

$$\sum_n a_n = \infty, \ \sum_n a_n^2 < \infty,$$

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 [8,14] 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 [15,31] views (1) as a noisy discretization of the ordinary differential equation (ODE for short)

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

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 behavior as (3). See [8,9] 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 expected, 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 [27,28]. This is acceptable and in fact unavoidable in the important application area of tracking a slowly moving target or measuring a slowly exciting signal [4,21] and other instances of learning in a slowly varying environment. This is because with decreasing stepsize, the algorithmic timescale, dictated by the decreasing stepsize, eventually becomes slower than the timescale on which the target is moving and therefore 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 hardwired [8,41], and one settles for a

\[1\text{ Barring some very special cases, e.g., when the right-hand side of (1) is contractive uniformly w.r.t. the noise variable.}\]
judiciously chosen constant stepsize. Such schemes are a part of traditional signal processing and neural network algorithms [16,17,19,25,32,38] and often show up in important applications such as quasi-stationary experimentation for meteorology [13], slowly exciting physical wave measurement [12], and more recently in online learning and non-stationary optimization [42,48]. However, the focus in online learning is cumulative regret bounds instead of all time bounds.

These developments have motivated analysis of constant stepsize schemes [9,11,23,26,27,34,35,38] in the form of various limit theorems, (non)asymptotic 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 [48]. 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 [4,18,20–22,38,48].

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 [11,23,27,34,35] among others. The emphasis of the latter is toward 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 [23], which establishes a law of iterated logarithms, and [35], 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 [29]. 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 [42]. 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 timescale equal to or faster than that of the algorithm. This situation has been analyzed through mean squared error bounds [43,44] and is close in spirit to ours.
1.3 Our contributions

Our main result is Theorem 41. The highlights of this result are as follows.

1. Our setup 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 \( n \leq \text{some} \; N \) or in the \( n \to \infty \) 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 and their mixtures, 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]. The Alekseev formula was used in [40] to obtain all time concentration bounds for decreasing stepsize stochastic approximation. The present work differs from ibid. in significant ways, specifically the following.

   (a) The core of the analysis of [40] is in the small neighborhood of a stable equilibrium of the limiting ODE whereby the stability and related properties of the linearized ODE at this point can be leveraged. This methodology is not available in the tracking problem where the full complexity of stability issues for time-dependent linear systems has to be contended with. This calls for our use of Solo’s stability criterion, discussed further later on.

   (b) The interpolated iterates on [40] are easily cast as a perturbation of the limiting ODE, which is not the case here. The entire calculation in Sect. 2.3 leading to Eqs. (24)–(27) is necessary precisely for this reason. Note in particular that the right-hand side of (24) involves \( \tilde{h} \) and not \( h \) as simple analogy with [40] might lead one to believe.

   (c) The estimates of individual error terms do go along the lines of their counterparts in [40] in spirit, but with extra complications due to coupling with the slowly varying component.
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., [24] and its application to stochastic approximation in [2].

2. One potential spot for improvement is in the use of the assumption (∗) below, which entails a stability condition for a linearized dynamics which is time-dependent. Such conditions are available only under constraints on the timescale 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 [39], is the most general available to our knowledge. (Another class of sufficient conditions available is based on the existence of Lyapunov 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 nonlinear 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 behavior, developed through a sequence of lemmas, leading to the main result in Sect. 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.

\[ x_n = \text{Iterate at time } n \] (2.1)
\[ h(\cdot, \cdot) = \text{Driving vector field of the tracking scheme (2.1)} \]
\[ a = \text{Stepsize (2.1)} \]
\[ M_{n+1} = \text{Martingale difference noise (2.1)} \]
\[ \varepsilon_n = \text{Additive error (2.1)} \]
\[ \varepsilon^* = \text{Bound on } \|\varepsilon_{n+1}\| \] (2.1)
\[ y(\cdot) = \text{Slowly varying signal to be tracked (2.1)} \]
\[ \epsilon = \text{Small}(\ll 1) = \text{number controlling rate of } y(\cdot) \] (2.1)
\( γ(⋅) = \text{Vector field driving } y(⋅) \) (2.1)
\[
C^* = \max \left( \sup_n \mathbb{E}\left[ \|x_n\|^2 \right]^{1/2}, \sup_n \mathbb{E}\left[ \|x_n\|^4 \right]^{1/4} \right) \) (2.1)
\[
C_γ = \sup_{t \geq 0} \|y(t)\\) (2.1)
\[
C_M, \delta = \text{Constants featuring in the bound for } \|M_{n+1}\| \) (2.1)
\[
Φ = \text{Transition matrix of linear system } (2.2)
\[
z(⋅) = \text{Slowly varying equilibrium for the algorithm } (2.3)
\[
d = \text{Dimension of } x_n \text{ and } z(⋅) \) (2.1/2.3)
\[
∇ = \text{Gradient operator } (2.3)
\[
C_Φ, β = \text{Constants featuring in the exponential bound for } Φ \) (2.3)
\[
L_f = \text{Lipschitz constant for Lipschitz function } f \) (generic)
\[
G_f = \text{Constant of linear growth for function } f, \text{ i.e., } |f(x)| \leq C_f (1 + \|x\|) \) (generic)
\[
B_f = \sup_x |f(x)| \text{ for a bounded function } f \) (generic)
\[
K_γ = \max_{\|y\| \leq C_γ} \|y(y)\\) (3.2)
\[
O(⋅) = \text{Big O notation } (3.3)
\[
μ = 1/β \) (3.3)
\[
K_1 = L_h^2 (1 + C_h + C_γ) + \epsilon^* \) (3.3)
\[
K_2 = C_Φ L_h^2 \) (3.3)
\[
K_3 = K_1 + L_γ a \epsilon \) (3.3)
\[
K_4 = \max \left( 2C_M/δ^2, C_M^2/δ^2 \right) \) (3.3)
\[
K_5 = K \Phi^2 L_D/β \) (3.4)
\[
K_6 = C_Φ^3 L_γ L_D \epsilon \) (3.4)
\[
K_7 = \max \left( 24C_M δ^4, 4C_M^2 δ^4 \right) \) (3.4)
\[
K_8 = 2\sqrt{6C_M C_h^2/δ^2} \) (3.4)
\[
K_9 = C_M γ_1 d^{1.5} / δ \) (3.5)

2 Preliminaries

2.1 The tracking problem

We consider a constant stepsize stochastic approximation algorithm given by the \( d \)-dimensional iteration

\[
x_{n+1} = x_n + a \left[ h(x_n, y_n) + M_{n+1} + ε_{n+1} \right], \quad n \geq 0,
\]

\( \Box \) Springer
for tracking a slowly varying signal governed by
\[ \dot{y}(t) = a \epsilon \gamma(y(t)), \]
(5)
with \(0 < a < 1, 0 < \epsilon \ll 1\). Also, \(y_n := y(n), n \geq 0\), the trajectory of (5) sampled at unit\textsuperscript{2} 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 \(\epsilon_{n+1}\) 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 \(\epsilon\) ensures a separation of timescale 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,
- \(C^* := \max \left( \sup_n \mathbb{E}\left[ \|x_n\|^2 \right]^{1/2}, \sup_n \mathbb{E}\left[ \|x_n\|^4 \right]^{1/4} \right) < \infty\). (See [9] 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 \(\epsilon^* > 0\) such that
\[ \|\epsilon_{n+1}\| \leq \epsilon^*, \quad \forall n \geq 0, \]
(6)
- \(M_n\) is a martingale difference sequence w.r.t. the increasing \(\sigma\)-fields
\[ \mathcal{F}_n := \sigma(x_m, M_m, \epsilon_m, m \leq n), n \geq 0, \]
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
\[ P(\|M_{n+1}\| > u | \mathcal{F}_n) \leq c_1(x_n) e^{-c_2(x_n)u}, \quad n \geq 0, \]
(7)
for all \(u \geq v\) for a fixed, sufficiently large \(v > 0\) (i.e., a sub-exponential tail), with
\[ \sup_n E\left[ c_1(x_n) \right] < \infty. \]
(8)
In particular, (7), (8) together imply that there exist \(\delta, C_M > 0\) such that
\[ E\left[ e^{\delta\|M_{n+1}\|} \right] \leq C_M, \quad n \geq 0. \]
(9)
\textsuperscript{2} Without loss of generality.
Using the Taylor expansion of the exponential function, we get

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

(10)

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

\[
E[\|M_{n+1}\|^m] \leq \frac{C_M m!}{\delta^m}.
\]  

(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 ‘nonlinear variation of constants’ formula.

### 2.2 Alekseev’s formula

Consider the ODE

\[
\dot{w}(t) = f(t, w(t)), \ t \geq 0,
\]

and its perturbed version,

\[
\dot{u}(t) = f(t, u(t)) + g(t, u(t)), \ t \geq 0,
\]

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 nonlinear 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 \(t \geq t_0\),

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

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

\[
\dot{\phi}(t) = \frac{\partial f}{\partial w}(t, w(t, s, w_0))\phi(t), \ t \geq s, \tag{13}\]

with \(\Phi(s, s, w_0) = I_d\), the d-dimensional identity matrix. That is, it is the unique solution to the matrix linear differential equation

\[\hat{\phi} \text{ Springer}\]
\[
\dot{\Phi}(t, s, w_0) = \frac{\partial f}{\partial w}(t, w(t, s, w_0))\Phi(t, s, w_0)
\]

with the aforementioned initial condition at \( t = s \). Equation (12) is the Alekseev nonlinear variation of constants formula [1] (see also Lemma 3, [10]).

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

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

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

### 2.3 Perturbation analysis

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

\[
\dot{x}(t) = h(x(t), y),
\]

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 \( \lambda \) is twice continuously differentiable with bounded first and second derivatives. (Typically, this can be verified by using the implicit function theorem.) In particular,

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

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

\[
\dot{z}(t) = \epsilon a \nabla \lambda(y(t))y(y(t))
\]

\[
= ah(\lambda(y(t)), y(t)) + \epsilon a \nabla \lambda(y(t))y(y(t))
\]

\[
= ah(z(t), y(t)) + \epsilon a \nabla \lambda(y(t))y(y(t)) = a\tilde{h}(z(t), y(t))
\]

for

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

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:

\[
x_{n+1} = x_n + a[h(x_n, y_n) + M_{n+1} + \epsilon_{n+1}]
\]

\(\square\) Springer
\begin{align}
  &= x_n + a \left[ \tilde{h}(x_n, y_n) - \epsilon \nabla \lambda(y_n) \gamma(y_n) + M_{n+1} + \varepsilon_{n+1} \right] \quad \text{(17)} \\
  &= x_n + a \left[ \tilde{h}(x_n, y_n) + \kappa_n(y_n) \right], \quad \text{(18)}
\end{align}

where

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

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 \( t_n \equiv na \ \forall n \),

\begin{equation}
  \bar{x}(t) = \bar{x}(t_n) + \frac{t - t_n}{a} \left[ \bar{x}(t_{n+1}) - \bar{x}(t_n) \right], \quad t \in [t_n, t_{n+1}]. \quad \text{(20)}
\end{equation}

Then from (18), we get

\begin{align}
  \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} a M_{k+1} + \sum_{k=0}^{n} a \varepsilon_{k+1} \\
  &= \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 \text{(21)}
\end{align}

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

\begin{align}
  \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{align}

Thus,

\begin{align}
  \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{align}

Using (20),

\begin{equation}
  \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 \text{(23)}
\end{equation}
Define

\[ \Xi(t) = \xi_1(t) + \xi_2(t) + \xi_3(t) + \xi_4(t). \]

Consider the coupled systems

\begin{align*}
\dot{z}(t) &= \tilde{h}(z(t), y(t)), \\
\dot{y}(t) &= \epsilon a \gamma(y(t)),
\end{align*}

(24)

and

\begin{align*}
\dot{\bar{x}}(t) &= \tilde{h}({\bar{x}}(t), y(t)) + \Xi(t), \\
\dot{y}(t) &= \epsilon a \gamma(y(t)).
\end{align*}

(25)

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 \( \lambda \). Then the linearization or ‘equation of variation’ of (24) is

\[ \dot{r}(t) = D(z(t), y(t))r(t). \]

(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

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

(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

\[ \varrho_n = \Phi(t_n, t_0; \bar{x}(t_0), y_0)(\bar{x}(t_0) - z(t_0)) \]

(30)

\begin{align*}
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 \text{(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 \text{(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 \text{(33)}
\end{align*}
\begin{align}
D_n &= \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s), y(s)) \epsilon_{k+1} ds, \tag{34} \\
E_n &= \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s), y(s)) \times \epsilon \nabla \lambda(y(t_k)) y'(y(t_k)) ds. \tag{35}
\end{align}

Then
\begin{align}
\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 \\
&= z(t_n) + A_n + (B_n - C_n) + C_n + D_n - E_n + \varrho_n. \tag{36}
\end{align}

Therefore,
\begin{align}
\| \bar{x}(t_n) - z(t_n) \| &\leq \| A_n \| + \| B_n - C_n \| + \| C_n \| + \| D_n \| + \| E_n \| + \| \varrho_n \|. 
\end{align}

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

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):

\((\dagger)\) There exists a \(\beta > 0\) such that \(\forall \ t > s \geq 0\) and \(x_0, y_0,\)
\begin{align}
\| \Phi(t, s; x_0, y_0) \| &\leq C \Phi e^{-\beta (t-s)}.
\end{align}

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 timescale 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

\[ \dot{x}(t) = A x(t) + B y(t), \]

where \(A\) and \(B\) are time-varying matrices.
\[ \dot{r}(t) = A(t)r(t). \] 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, [37]. 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 [39], which we recall below as a sufficient condition for \( \dagger \). (There have also been some extensions thereof to nonlinear systems; see, e.g., [33].)

Consider the following time-varying linear dynamical system:

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

(39)

and assume the following for this perturbed system:

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

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

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

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

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 \( T > 0 \),

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

4. There exists \( \delta > 0 \) such that

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

**Theorem 21** (Stability test for deterministic perturbations using eigenvalue-based characterization [39]) *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 $\bar{A}, b, \bar{a}$ are as defined in (A1)-(A4) and $\beta$ is small enough so that:

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

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 Lyapunov functions, e.g., [46,47], 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),

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

Proof  We have

$$||D_n|| = ||\sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \tilde{x}(s), y(s))\epsilon_{k+1} ds||$$

$$\leq \epsilon^* \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} ||\Phi(t_n, s; \tilde{x}(s), y(s))|| ds$$  

$$\leq C_\Phi \epsilon^* \int_{t_0}^{t_n} e^{-\beta(t_n-s)} ds$$

$$\leq \frac{C_\Phi \epsilon^*}{\beta},$$

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

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

$\square$
3.2 Bound on $E_n$

**Lemma 2** For $E_n$ defined in (35),

$$
\mathbb{E}\left[||E_n||^2\right]^{1/2} \leq \frac{K_{\gamma} L_{\lambda} C_{\Phi} \epsilon}{\beta},
$$

(44)

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

**Proof** We have,

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

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

(45)

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

(46)

Using (†),

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

$$
\leq \frac{K_{\gamma} L_{\lambda} C_{\Phi} \epsilon}{\beta}.
$$

(47)

(48)

Hence,

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

(49)

3.3 Bound on $A_n$

The next lemma is a variant of Lemma 5.5 of [40].

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

$$
\int_{t_k}^{t_{k+1}} e^{-\beta(t_n-s)} ||\bar{x}(s) - \bar{x}(t_k)|| ds
\leq [K_1 + G_{\tilde{h}} ||\bar{x}(t_k)|| + ||M_{k+1}||]e^{-\beta(t_n-t_{k+1})} a^2.
$$
Proof Using (20) and (1), for $s \in [t_k, t_{k+1}]$,

$$||\bar{x}(s) - \bar{x}(t_k)|| = \frac{s - t_k}{\alpha}||\bar{x}(t_{k+1}) - \bar{x}(t_k)||$$

$$= (s - t_k)[||\tilde{h}(\bar{x}(t_k), y(t_k)) + M_{k+1} + \varepsilon_{k+1}||]$$

$$\leq (s - t_k)[||\tilde{h}(\bar{x}(t_k), y(t_k))|| + ||M_{k+1}|| + ||\varepsilon_{k+1}||]$$

$$\leq (s - t_k)[G_{\tilde{h}}(1 + ||\bar{x}(t_k)|| + ||y(t_k)||) + ||M_{k+1}|| + \varepsilon^*] \quad (50)$$

$$\leq (s - t_k)[G_{\tilde{h}}(1 + C_{\gamma}) + G_{\tilde{h}}||\bar{x}(t_k)|| + ||M_{k+1}|| + \varepsilon^*]$$

$$\leq (s - t_k)[K_1 + G_{\tilde{h}}||\bar{x}(t_k)|| + ||M_{k+1}||] \quad (51)$$

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,

$$\int_{t_k}^{t_{k+1}} e^{-\beta(t_n - s)}||\bar{x}(s) - \bar{x}(t_k)||ds$$

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

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

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

Proof We have

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

$$\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} ||\Phi(t_n, s; x(s), y(s))||$$

$$\times ||[\tilde{h}(\bar{x}(t_k), y(t_k)) - \tilde{h}(\bar{x}(s), y(s))]||ds$$

$$\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} C_{\phi}e^{-\beta(t_n - s)}L_{\tilde{h}} \times \left(||\bar{x}(t_k) - \bar{x}(s)|| + ||y(t_k) - y(s)||\right) \quad (52)$$

$$\leq C_{\phi}L_{\tilde{h}} \sum_{k=0}^{n-1} \left[K_1 + G_{\tilde{h}}||\bar{x}(t_k)|| + ||M_{k+1}||e^{-\beta(t_n - t_{k+1})}a^2ight.$$  

$$+ K_{\gamma}ae \int_{t_k}^{t_{k+1}} (s - t_k)e^{-\beta(t_n - s)}ds \right] \quad (53)$$
\[ \leq C \phi L \sum_{k=0}^{n-1} \left[ [K_1 + G \| \bar{x}(t_k) \| + \| M_{k+1} \|] e^{-\beta (t_n - t_k) a^2} \\
+ K_\gamma a \epsilon e^{-\beta (t_n - t_k) a^2} \right] \]
\[ = C \phi L \sum_{k=0}^{n-1} [K_1 + G \| \bar{x}(t_k) \| + K_\gamma a \epsilon + \| M_{k+1} \|] e^{-\beta (t_n - t_k) a^2} \]
\[ \leq a C \phi L \left( (K_1 + K_\gamma a \epsilon) \mu \\
+ \sum_{k=0}^{n-1} (G \| \bar{x}(t_k) \| + \| M_{k+1} \|) a e^{-\beta (t_n - t_k) a^2} \right), \quad (54) \]

where \( \mu := \frac{1}{\beta} \). Equation (53) follows from Lemma 3. Denote the terms \( C \phi L \), \( K_1 + K_\gamma a \epsilon, \sum_{k=0}^{n-1} \| M_{k+1} \| a e^{-\beta (t_n - t_k) a^2} \) and \( \sum_{k=0}^{n-1} \| \bar{x}(t_k) \| e^{-\beta (t_n - t_k) a^2} \) by \( K_2, K_3, F_n \) and \( \tilde{F}_n \). Note that \( K_3 \) is \( O(1) \). Then

\[ \| A_n \| \leq a K_2 (K_3 \mu + F_n + G \tilde{F}_n), \]
\[ \mathbb{E}[\| A_n \|^2]^{1/2} \leq a K_2 \left( K_3 \mu + \mathbb{E}[F_n^2]^{1/2} + G \tilde{F}_n^{1/2} \right). \]

Now,
\[ \mathbb{E}[\tilde{F}_n^2]^{1/2} = \sum_{k=0}^{n-1} \left( a \mathbb{E}[\| \bar{x}(t_k) \|^2]^{1/2} e^{-\beta (t_n - t_k) a^2} \right) \]
\[ \leq C^* \mu \quad (55) \]

and
\[ \mathbb{E}[F_n^2]^{1/2} = \sum_{k=0}^{n-1} \mathbb{E}[\| M_{k+1} \|^2]^{1/2} a e^{-\beta (n-k) a} \]
\[ \leq \sum_{i=0}^{n-1} \sqrt{2C M} \delta a e^{-\beta (n-k) a} \]
\[ \leq K_4 \mu, \quad (56) \]

where \( K_4 = \frac{\sqrt{2C M}}{\delta} \). Therefore,
\[ \mathbb{E}[\| A_n \|^2]^{1/2} \leq a K_2 \left( K_3 \mu + K_4 \mu + G \tilde{F}_n C^* a \mu \right). \]

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

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

Therefore,

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

$$\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} ||\Phi(t_n, s; \bar{x}(s), y(s)) - \Phi(t_n, s; \bar{x}(t_k), y(t_k))||$$

$$\times ||M_{k+1}|| ds.$$  

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

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

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

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

and

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

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

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

\[ \dot{\Phi}_j(t, s; \bar{x}(t_k), y(t_k)) = D(z(t, s; \bar{x}(s), y(s)), y(t, s; y(s)))\Phi_j(t, s; \bar{x}(t_k), y(t_k)) \]

\[ + \left[ 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))) \right] \times \Phi_j(t, s; \bar{x}(t_k), y(t_k)). \]  

(62)

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

\[ \Phi_j(t_n, s; \bar{x}(t_k), y(t_k)) - \Phi_j(t_n, s; \bar{x}(s), y(s)) \]

\[ = \int_s^{t_n} \Phi(t_n, t; \bar{x}(t), y(t)) \]

\[ \times \left[ 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))) \right] \]

\[ \times \Phi_j(t, s; \bar{x}(t_k), y(t_k)) dt. \]  

(63)

Combining equation (63) for all columns, we get

\[ \Phi(t_n, s; \bar{x}(t_k), y(t_k)) - \Phi(t_n, s; \bar{x}(s), y(s)) \]

\[ = \int_s^{t_n} \Phi(t_n, t; \bar{x}(s), y(s)) \]

\[ \times \left[ 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))) \right] \]

\[ \times \Phi(t, s; \bar{x}(t_k), y(t_k)) dt. \]

Therefore,

\[ ||B_n - C_n|| \leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} \left| \left| \Phi(t_n, t; \bar{x}(s), y(s)) \right| \right| \]

\[ \times \left| \left| 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))) \right| \right| \]

\[ \times ||\Phi(t, s; \bar{x}(t_k), y(t_k))|| \times ||M_{k+1}|| dt \, ds \]

\[ \leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_0^2 L_D \times e^{-\beta(t_n-s)} \times e^{-\beta(t-s)} \times \]

\[ \left[ ||z(t, s; \bar{x}(s), y(s)) - z(t, s; \bar{x}(t_k), y(t_k))|| 

+ ||y(s) - y(t_k)|| \right] \times ||M_{k+1}|| dt \, ds \]  

(64)

In fact, the classical variation of constants formula for linear systems which it generalizes.
\[
\leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_\phi^2 L_D e^{-\beta(t_n-s)} \times e^{-\beta(t-s)} \times \left( \left[ ||\bar{x}(s) - \bar{x}(t_k)|| + ||y(s) - y(t_k)|| \right] C_\phi e^{-\beta(t-s)} + ||y(s) - y(t_k)|| \right) \times ||M_{k+1}|| \, dt \, ds,
\]

where (64) follows from (\dagger) and the Lipschitz property of \( D(\cdot, \cdot) \), while (65) follows from (14) and (\dagger). We split the analysis into two terms as follows:

\[
G_n := \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_\phi^3 L_D e^{-\beta(t_n-s)} ||\bar{x}(s) - \bar{x}(t_k)|| \times e^{-2\beta(t-s)} \times ||M_{k+1}|| \, dt \, ds
\]

\[
= \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_\phi^3 L_D e^{-\beta(t_n-s)} ||\bar{x}(s) - \bar{x}(t_k)|| \times \left( 1 - e^{-2\beta(t_n-s)} \right) \times ||M_{k+1}|| \, ds
\]

\[
\leq K_5 \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} e^{-\beta(t_n-s)} \times ||\bar{x}(s) - \bar{x}(t_k)|| \times \left( 1 - e^{-2\beta(t_n-s)} \right) \times ||M_{k+1}|| \, ds
\]

\[
\leq K_5 \sum_{k=0}^{n-1} a^2 e^{-\beta(t_n-t_k)} \left[ K_1 ||M_{k+1}|| + ||M_{k+1}||^2 + G_h ||\bar{x}(t_k)|| \times ||M_{k+1}|| \right]
\]

(66)

where \( K_5 \) denotes \( C_\phi^3 L_D/2\beta \) and (66) follows from Lemma 3,

\[
H_n := \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_\phi^2 L_D e^{-\beta(t_n-s)} ||y(s) - y(t_k)|| e^{-2\beta(t-s)} ||M_{k+1}|| \, dt \, ds
\]

\[
+ \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \int_s^{t_n} C_\phi^2 L_D e^{-\beta(t_n-s)} ||y(s) - y(t_k)|| e^{-\beta(t-s)} ||M_{k+1}|| \, dt \, ds
\]

\[
= \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} C_\phi^3 L_D e^{-\beta(t_n-s)} ||y(s) - y(t_k)|| \left[ \frac{1}{2\beta} \left( 1 - e^{-2\beta(t_n-s)} \right) \right] ||M_{k+1}|| \, ds
\]

\[
+ \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} C_\phi^3 L_D e^{-\beta(t_n-s)} ||y(s) - y(t_k)|| \left[ \frac{1}{\beta} \left( 1 - e^{-\beta(t_n-s)} \right) \right] ||M_{k+1}|| \, ds
\]
\[ \leq \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \frac{C_\Phi^2 (2 + C_\Phi) L D}{2 \beta} e^{-\beta(t_n - s)} K \| a e(s - t_k) \| M_{k+1} \| ds \]
\[ \leq K_6 \sum_{k=0}^{n-1} e^{-\beta(t_n - t_k) + 1} a^3 \| M_{k+1} \|, \quad (67) \]

where (67) follows from (5) and
\[ K_6 := (2 + C_\Phi) C_\Phi^2 K \| L D \| e/2 \beta. \]

Further define \( G_{1,n}, \) \( G_{2,n} \) and \( G_{3,n} \) as follows
\[
G_{1,n} = \sum_{k=0}^{n-1} a e^{-\beta(t_n - t_k + 1)} \| M_{k+1} \|,
\]
\[
G_{2,n} = \sum_{k=0}^{n-1} a e^{-\beta(t_n - t_k + 1)} \| M_{k+1} \|^2,
\]
\[
G_{3,n} = \sum_{k=0}^{n-1} a e^{-\beta(t_n - t_k + 1)} \| \bar{x}(t_k) \| \| M_{k+1} \|.
\]

Then
\[
\| B_n - C_n \| \leq G_n + H_n
\]
\[
\leq K_5 a (K_1 G_{n,1} + G_{n,2} + G_h G_{n,3}) + (K_6 a^2 G_{n,1}).
\]

Therefore,
\[
\mathbb{E}[\| B_n - C_n \|^2]^{1/2} \leq K_5 a (K_1 \mathbb{E}[G_{n,1}^2]^{1/2} + \mathbb{E}[G_{n,2}^2]^{1/2} + G_h \mathbb{E}[G_{n,3}^2]^{1/2})
\]
\[
+ a^2 K_6 \mathbb{E}[G_{n,1}^2]^{1/2}.
\]

We now bound each of the terms in the previous expression. Using a calculation similar to the one used for (56), we have
\[
\mathbb{E}[G_{n,1}^2]^{1/2} \leq K_4 \mu, \quad (68)
\]
\[
\mathbb{E}[G_{n,2}^2]^{1/2} = \sum_{k=0}^{n-1} \mathbb{E}[\| M_{k+1} \|^4]^{1/2} a e^{-\beta(n-k)a}
\]
\[
\leq \sum_{k=0}^{n-1} \frac{\sqrt{24 CM}}{\delta^2} a e^{-\beta(n-k)a}
\]
\[
\leq K_7 \left( \sum_{k=0}^{n-1} a e^{-\beta(n-k)a} \right)
\]
\[
\leq K_7 \mu, \quad (69)
\]
where $K^7 = \sqrt{24CM/\delta^2}$,

\[
\mathbb{E}[G_{n,3}^2]^{1/2} = \mathbb{E}\left[\sum_{k=0}^{n-1} ae^{-\beta(t_n-t_{k+1})} E\left[||\bar{x}(t_k)||^2||M_{k+1}||^2\right]\right]^{1/2}
\leq \sum_{k=0}^{n-1} ae^{-\beta(t_n-t_{k+1})} \left(\mathbb{E}[||\bar{x}(t_k)||^4]\mathbb{E}[||M_{k+1}||^4]\right)^{1/4}
\leq \sum_{k=0}^{n-1} ae^{-\beta(t_n-t_{k+1})} C^* \left(\mathbb{E}[||M_{k+1}||^4]\right)^{1/2}
\leq \sum_{k=0}^{n-1} ae^{-\beta(t_n-t_{k+1})} C^* \left(\frac{(24CM)^{1/4}}{\delta}\right)
\leq K^8 \mu, \quad (70)
\]

where $K^8 = \frac{C^*(24CM)^{1/4}}{\delta}$.

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

\[
\mathbb{E}[||B_n - C_n||^2]^{1/2} \leq K_5 a\left(K_1 K_4 \mu + K_7 \mu + G_{\tilde{h}} K_8 \mu\right) + a^2 K_6 K_8 \mu = O(a).
\] \(\square\)

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 51 in Appendix, with

\[
\alpha_{k,n} = \int_{t_k}^{t_{k+1}} \Phi(t_n, s, \bar{x}(t_k), y(t_k)) ds,
\gamma_1 = \frac{C\beta}{\gamma}, \quad \gamma_2 = 1, \quad \beta_n = a,
\]

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

\[
E[||C_n||^2] = \int_0^{\infty} P(||C_n||^2 \geq s) ds
= \int_0^{\infty} P(||C_n|| \geq s) ds.
\]
Using the martingale concentration inequality provided in 51 in Appendix, we have

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

(72)

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

\[
\int_0^{K_9} 2d^2 \exp\left(-\frac{cs}{d^3a}\right) ds = \frac{2d^5a}{c} \left( 1 - \exp\left(-\frac{cK_9}{d^3a}\right) \right)
\leq \left( \frac{2d^5}{c} \right)a,
\]

(73)

and

\[
\int_{K_9}^{\infty} 2d^2 \exp\left(-\frac{c\sqrt{s}}{d^{3/2}a}\right) ds = \frac{4d^5a}{c^2} \exp\left(-\frac{c\sqrt{K_9}}{d^{3/2}a}\right) \left( a + \frac{c\sqrt{K_9}}{d^{3/2}} \right)
\leq \frac{4d^5a}{c^2} \left( \frac{ad^{3/2}}{c\sqrt{K_9}} \right) \left( a + \frac{c\sqrt{K_9}}{d^{3/2}} \right)
= O(\max\{a^3d^{6.5}, a^2d^5\}),
\]

(74)

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

\[
\mathbb{E}[||C_n||^2] \leq \left( \frac{2d^5}{c} \right)a + O(\max\{a^3d^{6.5}, a^2d^5\})
= O(\max\{a^3d^{6.5}, ad^5\})
\]

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

\( \Box \)

### 4 Main result

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

**Theorem 41** The mean square deviation of tracked iterates from a non-stationary trajectory satisfies:

\[
\mathbb{E}\left[ ||x_n - \lambda(y(n))||^2 \right]^{1/2} \leq \frac{C_{\phi} \epsilon^*}{\beta} + \frac{K_{\gamma} L_{\phi} C_{\phi} \epsilon}{\beta}
+ O(\max\{a^{1.5}d^{3.25}, a^{0.5}d^{2.5}\})
+ C_{\phi} e^{-\beta(t_n - t_0)} ||x_0 - \lambda(y(0))||.
\]

(76)
Proof Using (38), (†) and Lemmas 1-6, we get

\[
\mathbb{E}[||\bar{x}(t_n) - z(t_n)||^2]^{1/2} \leq \frac{C_\phi \epsilon^*}{\beta} + \frac{K_y L_\lambda C_\phi \epsilon}{\beta} + O(a)
\]
\[+ O(\max\{a^{1.5}d^{3.25}, a^{0.5}d^{2.5}\})
\]
\[+ C_\phi e^{-\beta(t_n-t_0)}||\bar{x}(t_0) - z(t_0)||
\]
\[= \frac{C_\phi \epsilon^*}{\beta} + \frac{K_y L_\lambda C_\phi \epsilon}{\beta}
\]
\[+ O(\max\{a^{1.5}d^{3.25}, a^{0.5}d^{2.5}\})
\]
\[+ C_\phi e^{-\beta(t_n-t_0)}||\bar{x}(t_0) - z(t_0)||.
\]

The claim follows. \(\square\)

Remark 1. The \(O(\cdot)\) notation is used above to isolate the dependence on the step-size \(a\). 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 \(\epsilon^*\) and \(\epsilon\) is natural to expect, these being contributions from bounded additive error component \(\epsilon_n\) 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 stepsize stochastic approximation algorithm for tracking a slowly varying dynamical system and obtained a non-asymptotic bound valid for all time, with dependence on stepsize and dimension explicitly given. The latter in particular provides insight into stepsize 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 [45], which is much more complex.

Appendix: A martingale concentration inequality

We state here the martingale concentration inequality we have used, from [40], which in turn is a slight adaptation of the results of [30].

Theorem 51 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 \(\delta, C > 0\)
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 $\beta_n$ is some positive sequence. Then for $\eta > 0$, there exists some constant $c > 0$ depending on $\delta$, $C$, $\gamma_1$, $\gamma_2$ such that
$$P(||S_n|| > \eta) \leq \begin{cases} 2d^2 e^{-\frac{c\eta^2}{\delta^3 \beta_n}} & \text{if } \eta \in \left(0, \frac{C\gamma_1 d^{1.5}}{\delta}\right], \\ 2d^2 e^{-\frac{c\eta}{\delta^{1.5} \beta_n}} & \text{otherwise} \end{cases}$$

References

1. Alekseev VM (1961) An estimate for the perturbations of the solutions of ordinary differential equations. Westnik Moskov Unn Ser 1:28–36
2. Anantharam V, Borkar VS (2012) Stochastic approximation with long range dependent and heavy tailed noise. Queueing Syst 71(1–2):221–242
3. Bach F, Moulines E (2013) Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In: Advances in neural information processing systems, pp 773–781
4. Benveniste A (1987) Design of adaptive algorithms for the tracking of time-varying systems. Int J Adapt Control Signal Process 19(1):3–29
5. Benveniste A, Ruget G (1982) A measure of the tracking capability of recursive stochastic algorithms with constant gains. IEEE Trans Autom Control 47(3):639–649
6. Borkar VS (2002) On trapping probability of stochastic approximation. Comb Probab Comput 11(1):11–20
7. Borkar AV, Borkar VS, Sinha A (2019) Aerial monitoring of slow moving convoys using elliptical orbits. Eur J Control 46(2):90–102
8. Borkar VS (2008) Stochastic approximation: a dynamical systems viewpoint. Hindustan Publishing Agency and Cambridge University Press, New Delhi and Cambridge
9. Borkar VS, Meyn SP (2000) The ODE method for convergence of stochastic approximation and reinforcement learning. SIAM J Control Optim 38(2):447–469
10. Brauer F (1966) Perturbations of non-linear systems of differential equations. J Math Anal Appl 14(2):198–206
11. Bucklew JA, Kurtz TG (1993) Weak convergence and local stability properties of fixed step size recursive algorithms. IEEE Trans Inf Theory 39(3):966–978
12. Burke WL (1971) Gravitational radiation damping of slowly moving systems calculated using matched asymptotic expansions. J Math Phys 12(3):401–418
13. Chappell CF (1986) Quasi-stationary convective events. In: Ray PS (ed) Mesoscale meteorology and forecasting. Springer, New York, pp 289–310
14. Han-Fu Chen (2006) Stochastic approximation and its applications. Springer, New York
15. Derevitskii DP, Fradkov AL (1974) Two models analyzing the dynamics of adaptation algorithms. Autom Remote Control 35(1):59–67
16. Diamantaras KI, Kung SY (1996) Principal component neural networks: theory and applications. Wiley, New York
17. Eweda E (1994) Comparison of RLS, LMS, and sign algorithms for tracking randomly time-varying channels. IEEE Trans Signal Process 42(11):2937–2944
18. Farden D (1981) Tracking properties of adaptive signal processing algorithms. IEEE Trans Acoust Speech Signal Process 29(3):439–446
19. Finnoff W (1993) Diffusion approximations for the constant learning rate backpropagation algorithm and resistance to local minima. In: Advances in neural information processing systems, pp. 459–466
20. Guo L, Ljung L (1995) Exponential stability of general tracking algorithms. IEEE Trans Autom Control 40(8):1376–1387
21. Guo L, Ljung L (1995) Performance analysis of general tracking algorithms. IEEE Trans Autom Control 40(8):1388–1402
22. Guo L, Ljung L, Wang G-J (1997) Necessary and sufficient conditions for stability of LMS. IEEE Trans Autom Control 42(6):761–770
23. Joslin JA, Heunis AJ (2000) Law of the iterated logarithm for a constant-gain linear stochastic gradient algorithm. SIAM J Control Optim 39(2):533–570
24. Joulin A (2007) On maximal inequalities for stable stochastic integrals. Potential Anal 26(1):57–78
25. Kuan CM, Hornik K (1991) Convergence of learning algorithms with constant learning rates. IEEE Trans Neural Netw 2(5):484–489
26. Kushner HJ, Hai H (1981) Averaging methods for the asymptotic analysis of learning and adaptive systems, with small adjustment rate. SIAM J Control Optim 19(5):635–650
27. Kushner HJ, Huang H (1981) Asymptotic properties of stochastic approximations with constant coefficients. SIAM J Control Optim 19(1):87–105
28. Kushner HJ, Yin GG (1997) Stochastic approximation algorithms and applications. Springer, New York
29. Laxminarayanan C, Szepesvari C (2018) Linear stochastic approximation: how far does constant stepsize and iterate averaging go? In: Proceedings of the 21st international conference on artificial intelligence and statistics (AISTATS)
30. Liu Q, Watbled F (2009) Exponential inequalities for martingales and asymptotic properties for the free energy of directed polymers in a random environment. Stoch Process Their Appl 119(10):3101–3132
31. Ljung L (1977) Analysis of recursive stochastic algorithms. IEEE Trans Autom Control 22(4):551–575
32. Ng SC, Leung SH, Luk A (1999) Fast convergent generalized back-propagation algorithm with constant learning rate. Neural Process Lett 9(1):13–23
33. Peuteman J, Aeyels D (2002) Exponential stability of slowly time-varying non-linear systems. Math Control Signals Syst 15:202–228
34. Pflug GCh (1986) Stochastic minimization with constant step-size: asymptotic laws. SIAM J Control Optim 24(4):655–666
35. Pflug GCh (1990) Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size. Monatshefte für Mathematik 110(3):297–314
36. Robbins H, Monro J (1951) A stochastic approximation method. Ann Math Stat 22(3):400–407
37. Rugh WJ (1993) Linear System Theory. Prentice Hall, Englewood Cliffs, NJ
38. Sharma R, Sethares WA, Bucklew JA (1996) Asymptotic analysis of stochastic gradient-based adaptive filtering algorithms with general cost functions. IEEE Trans Signal Process 44(9):2186–2194
39. Victor, Solo (1994) On the stability of slowly time-varying linear systems, Mathematics of Control, Signals, and Systems (MCSS), 7.4, pp. 331–350
40. Thoppe G, Borkar VS (2019) A concentration bound for stochastic approximation via Alekseev’s formula. Stoch Syst 9(1):1–26
41. Utkin V, Guldner J, Shi J (2017) Sliding mode control in electro-mechanical systems. CRC Press, Boca Raton
42. Wilson C, Veeravalli V, Nedic A (2018) Adaptive sequential stochastic optimization. IEEE Trans Autom Control 64(2):496–509
43. Yin G, Ion C, Krishnamurthy V (2009) How does a stochastic optimization/approximation algorithm adapt to a randomly evolving optimum/root with jump Markov sample paths. Math Program 120(1):67–99
44. Yin G, Krishnamurthy V, Ion C (2004) Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization. SIAM J Optim 14(4):1187–1215
45. Zerihun T, Ladde GS (2013) Fundamental properties of solutions of nonlinear stochastic differential equations and Method of variation of parameters. Dyn Syst Appl 22:433–458
46. Zhou B (2016) On asymptotic stability of linear time-varying systems. Automatica 68:266–276
47. Zhou B (2017) Stability analysis of non-linear time-varying systems by Lyapunov functions with indefinite derivatives. IET Control Theory Appl 9:1434–1442
48. Zhu J, Spall JC (2016) Tracking capability of stochastic gradient algorithm with constant gain. In: 55th IEEE Conference on decision and control (CDC), Las Vegas, pp 4522–4527

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.