On Accuracy of Delay and Extrapolate Method in Online Misalignment Estimation

Yu Wu, Shirin Jalali, Carl Nuzman

***Abstract*—Many computationally efficient solutions for online system identification, including the recently-proposed incremental maximum likelihood (IML) algorithm, tune their learning rates based on online estimates of the misalignment error, defined as the ℓ_2-norm of the difference between the true system parameters vector $\mathbf{w}^* \in \mathbb{R}^L$ and its current estimate $\mathbf{w}_t \in \mathbb{R}^L$. One classic approach for estimating the misalignment error is the so-called "delay and extrapolate" algorithm proposed in 1982 by Yamamoto et al. as a heuristic method. In this paper, we rigorously analyze the accuracy and effectiveness of this approach when applied to the least mean squares (LMS) algorithm. For general stationary memoryless sources and for stationary Gaussian sources with memory, we show that under mild conditions, the delay and extrapolate method can provably provide an accurate estimate of the misalignment error. We support and illustrate our theoretical analysis through simulation.**

I. Introduction

Online estimation of a linear systems's impulse response, also known as adaptive filtering, is a well-studied decades-old problem with a wide range of applications in digital signal processing and machine learning. A key application of such algorithms is in two-way audio communication systems. (Refer to Fig. 1.) In such systems, audio echo cancellation (AEC) algorithms are employed to cancel the undesired echo signal generated by the loud-speakers and received by the microphones. Well known AEC algorithms include least mean squares (LMS) [1], normalized LMS (NLMS), proportionate NLMS (PNLMS) [2], affine projection algorithm (APA) [3] and recursive least squares (RLS) [4]. In addition, recent advances in computer vision and natural language processing have motivated researchers to explore the application of deep learning in this domain as well. This has led to solutions such as fusion of RNN and NLMS [5], CNN for AEC [6] and attention-based neural network for AEC [7].

While classic algorithms, such as NLMS and APA, are versatile and computationally-efficient, they do not fully address the requirements of modern high-definition multi-channel audio systems, such as very high accuracy, fast convergencebib (or low delay), and robustness. On the other hand, the RLS algorithm has both fast convergence and high accuracy and the deep learning based algorithms can handle the cases where the system contains non-linear components. However, the high computational complexity and memory requirements of both RLS and deep learning based methods make them impractical in most applications. These shortcomings has generated a new wave of interest in developing efficient high-performance online learning algorithms.

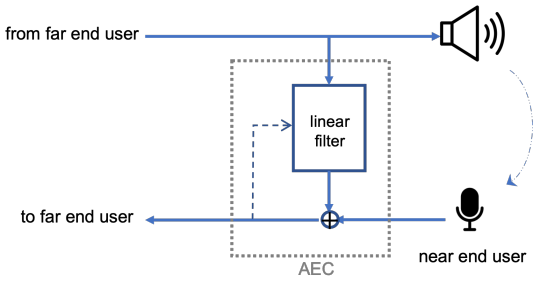


Fig. 1. AEC system model

When used with fixed parameters, such as fixed learning rate, classical adaptive filtering algorithms show a trade-off between convergence speed and accuracy. A key approach to address this issue and design fast-converging algorithms that are also accurate is to adaptively tune parameters such as the learning rate. Examples include variable step size NLMS [8]–[12], variable step size APA [13], [14], and, recently, IML and OBML [15], the last of which has optimal convergence properties for some input distributions. In many of these algorithms, the optimal choice of the learning rate is a function of the time-dependent misalignment error, which is the distance between the true unknown parameters vector $\mathbf{w}^* \in \mathbb{R}^L$ and its estimate at time t, $\mathbf{w}_t$. Since this error is unknown, to be practically useful these algorithms must be coupled with algorithms for estimating the misalignment error. One potential misalignment estimation approach is the "delay and extrapolate" algorithm, introduced in [16] in 1980s. In this approach, additional (non-causal) delay is added artificially to the control loop so that ideal filter coefficients associated with the delay are known to be zero. The error in estmating these known coefficients can be calcualted and then extrapolated to an estimate of the overall error on all filter coefficients.

The intuition behind the delay and extrapolate method is that "adaptive algorithms spread the filter misalignment evenly over all coefficients " [17]. However, the assumption that the error is uniformly spread across the coefficients is a heuristic assumption based on empirical observations and is not theoretically verified. The goal of this paper is to provide a solid theoretical foundation for the use of the delay-and-extrapolate method by analyzing the validity of this assumption under different source signal distributions. We focus on the LMS algorithm and theoretically characterize the steady-state behavior of the squared error matrix $\mathbf{U}_t \in \mathbb{R}^{L\times L}$

defined as $\mathbf{U}_t = \mathrm{E}[(\mathbf{w}_t - \mathbf{w}^*)(\mathbf{w}_t - \mathbf{w}^*)^T]$. Prior work [18], [19] provides an approximate analysis of $\mathbf{U}_t$. In our work, we derive exact expressions for $\mathbf{U}_t$, both for general memoryless sources and Gaussian sources with memory.

The organization of the paper is as follows. Section II reviews the system model and also the basics of the delay and extrapolate algorithm. Section III presents our main theoretical results on the squared error matrix $\mathbf{U}_t$, for two different types of source. Section IV explores the implications of the results of Section III for the delay and extrapolate algorithm. Section V presents our simulation results. Section VI concludes the paper. The proofs of the theoretical results are presented in the appendices.

II. Background

A. System model and the LMS algorithm

Consider a linear system described as follows. Let $\mathbf{x}_t \in \mathbb{R}^L$ and y_t denote the input and output of our system at time t, respectively. We assume that our linear time-invariant system's response is described by $\mathbf{w}^* \in \mathbb{R}^L$, such that

$$y_t = \mathbf{x}_t^T \mathbf{w}^* + z_t, \tag{1}$$

where where z_t denotes the additive noise in the system, which is typically modeled as independently and identically distributed (i.i.d.) as $\mathcal{N}(0, \sigma_z^2)$, and independent of $\mathbf{x}_t$.

We consider two models for the input:

- *sequence model:* $\mathbf{x}_t$ is a window of the most recent samples
$$\mathbf{x}_t = [x_t, \ldots, x_{t-L+1}]^T.$$
of a zero-mean, wide-sense stationary process x_t with $E[x_t x_{t+\tau}] = r(\tau)$.
- *block model:* the $\mathbf{x}_t$ are i.i.d with mean zero and covariance $\mathbf{R} \in \mathbb{R}^{L \times L}$

In the sequence model, the covariance matrix satisfies $R_{ij} = r(i-j)$, and (1) represents the convolution of the input sequence with finite impulse response $\mathbf{w}^*$. In the context of AEC (Fig. 1), x_t and y_t denote the loudspeaker and microphone signals at time t, and $\mathbf{w}^*$ is the echo impulse response.

While the sequence model aligns best with many applications, the block model is more amenable to analysis. Our theoretical results will be for the block model, although we empirically have observed that the predictions of the block model carry over well to the sequence model.

In online system identification (or learning), our goal is to adaptively estimate the unknown parameters $\mathbf{w}^*$ as the training samples $(\mathbf{x}_t, y_t)$ arrive. To achieve this goal, we start from an initial estimate $\mathbf{w}_0$, and at every time step $t+1$, we update our current estimate $\mathbf{w}_t$ to $\mathbf{w}_{t+1}$ as a function of $\mathbf{w}_t$ and our past few observations $\{(\mathbf{x}_{t'}, y_{t'})\}_{t'=t-P+1}^{t}$, where P is a small integer. A classic approach in this domain is the least mean squares (LMS) algorithm which works as follows

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \mu\big(y_t - \mathbf{x}_t^T \mathbf{w}_t\big)\mathbf{x}_t, \tag{2}$$

$\mu \in \mathbb{R}^+$ denotes the learning rate. Note that the LMS algorithm is closely connected to the stochastic gradient descent (SGD) method.

The misalignment error is defined as

$$\mathbf{e}_t = \mathbf{w}_t - \mathbf{w}^* \tag{3}$$

The resulting prediction error at the filter output is denoted

$$\xi_t = \mathrm{E}\Big[\big(y_t - \mathbf{x}_t^T \mathbf{w}_t\big)^2\Big] = \mathrm{E}\Big[\big(z_t - \mathbf{x}_t^T \mathbf{e}_t\big)^2\Big] \tag{4}$$

The misadjustment is a measure of the impact of imperfect filter coefficients on prediciton error, relative to the optimal error achieved with $\mathbf{e}_t = 0$, namely $\xi_{\min} = \mathrm{E}[z_t^2] = \sigma_z^2$. The misadjustment is defined as

$$M_t = \frac{\xi_t - \xi_{\min}}{\xi_{\min}}. \tag{5}$$

The ultimate goal of the adaptive filtering is to drive misadjustment toward zero.

B. Delay-and-extrapolate estimator

As mentioned previously, numerous algorithms have been proposed that control the learning rate and other parameters dynamically to optimize tradeoffs between convergence speed and accuracy. Many such methods ideally require knowledge of instantaneous misalignment error defined as $\|\mathbf{w}_t - \mathbf{w}^*\|^2$. While it is not practical to compute $\|\mathbf{w}_t - \mathbf{w}^*\|^2$ directly, one approach to estimate the misalignment in practice is the delay-and-extrapolate technique. We define an extended input vector $\tilde{\mathbf{x}}_t = [\mathbf{x}_{t,1}\, \mathbf{x}_{t,2}]$ of length $\tilde{L} = L_1 + L_2$. Here $\mathbf{x}_{t,2} = \mathbf{x}_t$ is the input vector previously defined, and $\mathbf{x}_{t,1}$ is a non-causal extension, given by the future L_1 samples

$$\mathbf{x}_{t,1} = [x_{t+L_1}, \ldots, x_{t+1}]^T.$$

in the sequence model, and by the last L_1 samples of the future block $\mathbf{x}_{t+1}$ in the block model. "Future" inputs are made available in practice by delaying the output y_t for at least L_1 samples before processing the data at time t. Using these definitions, (1) can be expressed $y_t = \tilde{\mathbf{x}}_t^T \tilde{\mathbf{w}}^* + z_t$ where $\tilde{\mathbf{w}}^* = [\mathbf{0}_{L_1}\, \mathbf{w}^*]$. Defining extensions $\tilde{\mathbf{w}}_t$ and $\tilde{\mathbf{e}}_t$, we see that $\mathbf{e}_{t,1} = \mathbf{w}_{t,1} - \mathbf{0} = \mathbf{w}_{t,1}$ is known, and $\mathbf{e}_{t,2} = \mathbf{e}_t$ is unknown. Assuming that these two parts of the error vector have the same average energy, we can estimate the misalignment by extrapolation as

$$\|\mathbf{w}_t - \mathbf{w}^*\|^2 = \|\mathbf{e}_{t,2}\|^2 \approx \frac{L_2}{L_1}\|\mathbf{e}_{t,1}\|^2 = \frac{L_2}{L_1}\|\mathbf{w}_{t,1}\|^2. \tag{6}$$

This heuristic estimate is based on the empirical observation that the components of the error vector $\tilde{\mathbf{e}}_t$ tend to have homogeneous variance. To quantify the accuracy of this estimate, we first need to analyze the second-order statistics of the error vector.

III. Mean squared error analysis of filter coefficients

In this section we focus on the error vector $\mathbf{e}_t$ and characterize the convergence behaviour of its second order statistics.

Our analysis applies equally to the original model (1) or its non-causal extension; for simplicity, in this section we write $L, \mathbf{e}_t$ and so on instead of $\tilde{L}, \tilde{\mathbf{e}}_t$. We focus on the LMS algorithm with a fixed learning rate $\mu > 0$ and throughout this section assume that the input vectors $\mathbf{x}_t$ are from the block model. The analysis makes small corrections to similar, well-known results in the LMS literature (see e.g. [18]). As these differences are not central to our purpose here, discussion is postponed to Appendix E.

Subtracting $\mathbf{w}^*$ from both sides of (2), and using (1) and (3) we have

$$\mathbf{e}_{t+1} = \big(I - \mu \mathbf{x}_t \mathbf{x}_t^T\big)\mathbf{e}_t + \mu \mathbf{x}_t z_t. \tag{7}$$

Then $\mathbf{e}_t$, being a function of $\{\mathbf{x}_s\}$ and $\{z_s\}$ for $s < t$, is independent of $\mathbf{x}_t$ and z_t. Taking the expected value of both sides of (7), the expected error satisfies

$$\mathrm{E}[\mathbf{e}_{t+1}] = (I - \mu \mathbf{R})\,\mathrm{E}[\mathbf{e}_t] = (I - \mu \mathbf{R})^{t+1}\,\mathrm{E}[\mathbf{e}_0]. \tag{8}$$

Therefore, if $\|I - \mu \mathbf{R}\|_2 < 1$, then $\mathrm{E}[\mathbf{e}_t] \to 0$, as $t \to \infty$. Convergence in mean is achieved for $0 < \mu < 2/\lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{R}$.

To understand the properties of the delay-and-extrapolate method, we also need to understand the mean-square behavior of the filter error, that is, the matrix $\mathbf{U}_t \in \mathbb{R}^{L \times L}$ defined as

$$\mathbf{U}_t \triangleq \mathrm{E}[\mathbf{e}_t \mathbf{e}_t^T].$$

Analyzing $\mathbf{U}_t$ also gives insight into prediction error and misalignment, as due to our independence assumptions together with (4) and (5) we have

$$\xi_t = \sigma_z^2 + \mathrm{E}\Big[\big(\mathbf{x}_t^T \mathbf{e}_t\big)^2\Big] = \sigma_z^2 + \mathrm{Tr}(\mathbf{R}\mathbf{U}_t) \tag{9}$$

and

$$M_t = \frac{\mathrm{Tr}(\mathbf{R}\mathbf{U}_t)}{\sigma_z^2}. \tag{10}$$

Our first result characterizes the dynamics $\mathbf{U}_t$, and its asymptotic behavior as t grows without bound, for non-Gaussian $\mathbf{x}_t$ with uncorrelated components.

Theorem 1. *Assume that the entries of $\mathbf{x}_t$ are i.i.d. zero mean such that i) $\mathrm{E}[\mathbf{x}_t \mathbf{x}_t^T] = \sigma_x^2 I_L$ and ii) $\mathrm{var}(x_{t,i}^2) = c\sigma_x^4 < \infty$, for all $i = 1, \ldots, L$. Then,*

$$\begin{aligned}\mathbf{U}_{t+1} =& \big(1 - 2\mu\sigma_x^2 + 2\mu^2\sigma_x^4\big)\mathbf{U}_t \\ &+ \mu^2\sigma_x^4\left((c-2)\,\mathrm{diag}(\mathbf{U}_t) + \left(\mathrm{Tr}\,(\mathbf{U}_t) + \frac{\sigma_z^2}{\sigma_x^2}\right) I_L\right)\end{aligned} \tag{11}$$

and moreover, if $0 < \mu < 2/((L+c)\sigma_x^2)$ then $\mathbf{U}_t$ converges to the finite limit

$$\mathbf{U}_\infty := \lim_{t\to\infty} \mathbf{U}_t = \left(\frac{\mu\sigma_z^2}{2 - (L+c)\mu\sigma_x^2}\right) I_L. \tag{12}$$

The proof of Theorem 1 is presented in Appendix A. From our perspective, the main significance of this result is that it shows that the diagonal entries of $\mathbf{U}_\infty$ are all identical, supporting the accuracy of the estimator (6). In the case of $\mathbf{x}_t$ is Gaussian input, we have $\mathrm{var}(x_{t,i}^2) = 2\sigma_x^4$, and hence $c = 2$. Binary input gives $c = 0$, while heavier tailed distributions can give $c > 2$. Applying Theorem 1 to (10), we obtain a characterization of the dynamics of the misadjustment.

Corollary 1. *Under the same conditions as Theorem 1,*

$$M_{t+1} = (1 - 2\mu\sigma_x^2 + \mu^2\sigma_x^4(L+c))M_t + \mu^2\sigma_x^4 L. \tag{13}$$

and

$$M_\infty = \lim_{t\to\infty} M_t = \left(\frac{\mu\sigma_x^2 L}{2 - (L+c)\mu\sigma_x^2}\right). \tag{14}$$

Comparing (13) and (14) shows a trade-off between convergence speed and stationary error. Fastest convergence is achieved with $\mu^{-1} = (L+c)\sigma_x^2$ but yields $M_\infty = L/(L+c)$. Smaller learning rates have slower convergence but smaller asymptotic error; methods such as IML [15] use estimates of misalignment to adapt the learning rate over time to optimize this tradeoff.

Next we extend Theorem 1 to the case where the input vectors are i.i.d. zero-mean Gaussian vectors with a general covariance matrix $\mathbf{R}$. To set the stage, for a square matrix $\mathbf{R}$ and scalars μ and c, define the function

$$\phi_c(\mu, \mathbf{R}) = \mathrm{Tr}\Big(\mu\mathbf{R}(2I - c\mu\mathbf{R})^{-1}\Big) \tag{15}$$

wherever the matrix inverse exists.

Theorem 2. *Assume that the vectors $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ are i.i.d. Gaussian with $\mathbf{R} \in \mathbb{S}_{++}^L$. Then*

$$\begin{aligned}\mathbf{U}_{t+1} =& \mathbf{U}_t - \mu\mathbf{R}\mathbf{U}_t - \mu\mathbf{U}_t\mathbf{R} \\ &+ \mu^2\big[\sigma_z^2\mathbf{R} + 2\mathbf{R}\mathbf{U}_t\mathbf{R} + \mathrm{Tr}(\mathbf{R}\mathbf{U}_t)\mathbf{R}\big].\end{aligned} \tag{16}$$

Let μ^* be the smallest positive solution to $\phi_2(\mu^*, \mathbf{R}) = 1$. For $0 < \mu < \mu^*$, $\mathbf{U}_t$ converges to the finite limit

$$\mathbf{U}_\infty = \lim_{t\to\infty} \mathbf{U}_t = \frac{\mu\sigma_z^2}{1 - \phi_2(\mu, \mathbf{R})}(2I_L - 2\mu R)^{-1} \tag{17}$$

In contrast to Theorem 1, Theorem 2 shows that the diagonal entries of $\mathbf{U}_\infty$ are not exactly identical when $\mathbf{x}_t$ has an arbitrary covariance matrix. However, if the diagonal values of $\mathbf{R}$ are equal, the diagonal entries of $\mathbf{U}_\infty$ are approximately equal when the learning rate μ is sufficiently small, since in that case $(2I_L - 2\mu\mathbf{R})^{-1} \approx \frac{1}{2}(I_L + \mu\mathbf{R})$. The implications for delay-and-extrapolate estimation are considered further in the next section.

Applying Theorem 2 to (10), we obtain a bound on the dynamics of the misadjustment error, and its asymptotic limit. Let $\lambda_{\max}$ and $\lambda\min$, where $\lambda_{\max} \geq \lambda_{\min} > 0$, denote the largest and smallest eigenvalues of $\mathbf{R} \in \mathbb{S}_{++}^L$, respectively.

Corollary 2. *Under the same conditions as Theorem 2,*

$$\begin{aligned}M_{t+1} \leq& \big(1 - 2\mu\lambda_{\min} + \mu^2\big(2\lambda_{\max}^2 + \mathrm{Tr}(\mathbf{R}^2)\big)\big)M_t \\ &+ \mu^2\,\mathrm{Tr}(\mathbf{R}^2) \\ \leq& \big(1 - 2\mu\lambda_{\min} + \mu^2\lambda_{\max}^2(L+2)\big)M_t + \mu^2 L\lambda_{\max}^2\end{aligned} \tag{18}$$

and

$$M_\infty = \lim_{t\to\infty} M_t = \frac{\phi_2(\mu, \mathbf{R})}{1 - \phi_2(\mu, \mathbf{R})} \tag{19}$$

Proofs of Theorem 2 and Corollary 2 are presented in Appendices B and C, respectively. We again observe a trade-off between stationary estimation error and convergence rate. While the bound (18) is not tight in general, the limit (19) is exact. As suggested by the bound, convergence speed is slow when $\lambda_{\min} \ll \lambda_{\max}$.

To get further insights into the function $\phi_c(\mu, \mathbf{R})$ and the critical learning rate μ^*, let $\lambda_1 \geq \ldots \geq \lambda_L > 0$ denote the eigenvalues of $\mathbf{R}$. Then, $\phi_c(\mu, \mathbf{R})$ can be expressed as

$$\phi_c(\mu, \mathbf{R}) = \mu \operatorname{Tr}(\mathbf{R}(2I - \mu c \mathbf{R})^{-1}) = \sum_i \frac{\mu\lambda_i}{2 - \mu c \lambda_i}. \tag{20}$$

Thus $\phi_c(\mu, \mathbf{R})/\mu \to \operatorname{Tr}(\mathbf{R})/2$, as $\mu \to 0$. Since

$$\frac{\mu\lambda_i}{2} \leq \frac{\mu\lambda_i}{2 - \mu c \lambda_i} \leq \frac{\mu\lambda_i}{2 - \mu c \operatorname{Tr}(\mathbf{R})} \tag{21}$$

it follows that

$$\frac{\mu \operatorname{Tr}(\mathbf{R})}{2} \leq \phi_c(\mu, \mathbf{R}) \leq \frac{\mu \operatorname{Tr}(\mathbf{R})}{2 - \mu c \operatorname{Tr}(\mathbf{R})}. \tag{22}$$

As μ^* is the smallest positive number with $\phi_c(\mu^*, \mathbf{R}) = 1$. we have

$$\frac{2}{(c+1)\operatorname{Tr}(\mathbf{R})} \leq \mu^* \leq \frac{2}{\operatorname{Tr}(\mathbf{R})} \tag{23}$$

The critical learning rate μ^* can thus be easily bounded based on $\operatorname{Tr}(\mathbf{R})$, which is simply $\operatorname{Tr}(\mathbf{R}) = L\sigma_x^2$ when $\mathbf{x}_t$ comes from a stationary process.

For small enough μ, $\mathbf{U}_\infty$ can be approximated as

$$\mathbf{U}_\infty \approx \frac{\mu\sigma_z^2}{2 - \mu \operatorname{Tr}(\mathbf{R})}(I + \mu\mathbf{R}). \tag{24}$$

As a crosscheck, for the uncorrelated case with $\mathbf{R} = \sigma_x^2 I$ and $\phi_2(\mu, \mathbf{R}) = L\mu\sigma_x^2/(2 - 2\mu\sigma_x^2)$, we observe that (17) is consistent with (12) of Theorem 1. In this case $\lambda_{\max} = \lambda_{\min}$ and

$$\mu^* = \frac{2}{(L+2)\sigma_x^2}. \tag{25}$$

IV. Accuracy of delay and extrapolate

In this section, we use our results from the previous part to characterize the accuracy of the delay and extrapolate estimator (6).

Define $\Delta = \lim_{t\to\infty} \Delta_t = \lim_{t\to\infty} \|\mathbf{e}_{t,2}\|_2^2$ and its estimate $\hat{\Delta} = \lim_{t\to\infty} \hat{\Delta}_t = \lim_{t\to\infty} \frac{L_2}{L_1}\|\mathbf{e}_{t,1}\|_2^2$. For uncorrelated inputs, Theorem 1 shows that the diagonal entries of $\mathbf{U}_\infty$ are all equal, which implies that the expected values of Δ and $\hat{\Delta}$ are equal ($\mathrm{E}[\hat{\Delta}] = \mathrm{E}[\Delta]$). For Gaussian sources on the other hand, Theorem 2 shows that, if μ is small enough and the covariance matrix $\mathbf{R}$ has a constant diagonal (which holds for instance for all stationary sources), then $\mathrm{E}[\hat{\Delta}] \approx \mathrm{E}[\Delta]$.

We are also interested in the variability of the estimate. Define the relative estimation error as

$$\upsilon = \frac{\hat{\Delta} - \Delta}{\Delta} = \frac{\frac{1}{L_1}\|\mathbf{e}_{t,1}\|^2}{\frac{1}{L_2}\|\mathbf{e}_{t,2}\|^2} - 1. \tag{26}$$

When μ is small enough to achieve small misadjustment, the error vector $\mathbf{e}_t$ for large t can be thought of as the sum of many small, independent increments (see (3)). Thus if $\mathbf{x}_t$ is not heavy-tailed, we can expect $\mathbf{e}_t$ for large t to be approximately Gaussian. In the case of uncorrelated inputs, Theorem 1 shows that $\tilde{\mathbf{e}}_t$ would then be approximately distributed as $\mathcal{N}(\mathbf{0}, \nu^2 I_{\tilde{L}})$ with some variance ν^2. In this case, $\mathbf{e}_{t,1}$ and $\mathbf{e}_{t,2}$, asymptotically, converge to independent Gaussian random vectors with i.i.d. Gaussian entries, and Δ and $\hat{\Delta}$ converge to two independent chi-square random variables with L_2 and L_1 degrees of freedom, respectively. The ratio of two independent chi-square random variables is known to have a F distribution [20]. More precisely, if $U \sim \chi^2_{n_1}$ and $V \sim \chi^2_{n_2}$, where U and V are independent, then $X = \frac{\frac{1}{n_1}U}{\frac{1}{n_2}V}$ follows an F distribution F_{n_1,n_2}, with

$$\mathrm{E}[X] = \frac{n_2}{n_2 - 2},$$

and

$$\operatorname{Var}(X) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2-2)^2(n_2-4)}$$

Thus

$$\mathrm{E}[\upsilon] = \frac{L_2}{L_2 - 2} - 1 = \frac{2}{L_2 - 2} \tag{27}$$

and

$$\operatorname{Var}(\upsilon) = \frac{2L_2^2}{(L_2-2)^2(L_2-4)}\left(1 + \frac{L_2 - 2}{L_1}\right). \tag{28}$$

For large filter length $L_2 \gg L_1$, the variance of υ is approximately equal to $2/L_1$. In general, increasing L_1 lowers the variance, at the cost of added delay in the system, and (28) can be used to choose the smallest delay consistent with a target estimation accuracy.

For general Gaussian sources, Theorem 2 shows that $\mathbf{U}_t$ converges to $\mathbf{U}_\infty$ which is no longer proportional to identity in general, and therefore Δ and $\hat{\Delta}$ are no longer independent and no longer chi-square random variables. However, for small μ, $\mathbf{U}_\infty \sim (I - \mu\mathbf{R})^{-1} \approx I + \mu\mathbf{R}$ is approximately proportional to identity if $\mathbf{R}$ is Toeplitz, and $\mathbf{e}_{\infty,1}$ and $\mathbf{e}_{\infty,2}$ are approximately uncorrelated, so that (27) and (28) should still hold approximately.

Moreover, when $\mathbf{R}$ is Toeplitz, as occurs in the sequential model, then $\mathbf{R}$ is also centrosymmetric, meaning that $\mathbf{R}_{n,m} = \mathbf{R}_{L+1-n,L+1-m}$, then $\mathbf{U}_\infty$ is also centrosymmetric by the closure of this property under matrix inversion. The delay and extrapolate estimate is then accurate on expectation, by symmetry, in the special case of $L_1 = L_2$.

V. Simulation results

In this section, we conduct numerical simulations to verify and illustrate properties of the delay-and-extrapolate estimator.

A. Estimator variance in steady state

We first performed numerical experiments to confirm and illustrate the effect of the delay parameter L_1 in controlling the variance of the misalignment estimator in steady state. We fixed $L_2 = 20$, so that (28) predicts $\text{Var}(\upsilon) = \frac{25}{162}(1 + \frac{18}{L_1})$. This predicted variance is plotted in Fig. 2 as a function of L_1, together with empirical estimates of this variance obtained by simulation with uncorrelated and correlated inputs. For uncorrelated signals, we ran 5000 trials, running the LMS algorithm for 20,000 iterations in each trial, using $\mu = \frac{2}{3\,\text{Tr}\,\mathbf{R}} = \frac{2}{3(L_1+L_2)}$ for 20,000 iterations of the LMS algorithm. In the correlated case, we reduced the learning rate to $\mu = \frac{1}{6(L_1+L_2)}$ and to ensure convergence, increased the number of LMS iterations to 250,000. After convergence, we calculate the empirical variance of relative error over all trials when L_1 is in the interval $[10, 30]$. As expected, the analysis and simulation results agree very well for the uncorrelated case, and approximately in the correlated case, demonstrating the usefulness of (28) in designing the system delay L_1.

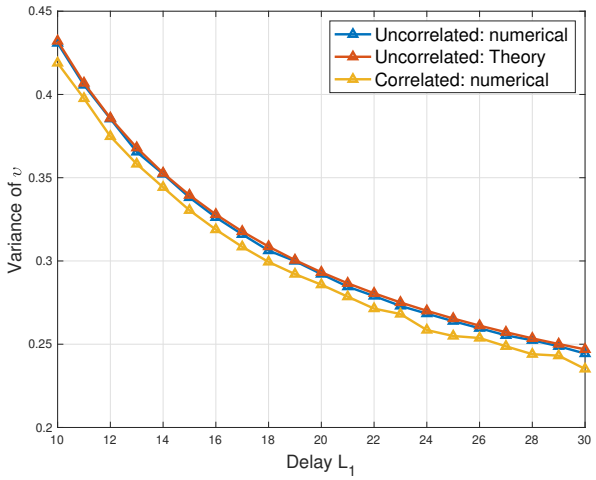


Fig. 2. Numerical and theoretical variance of relative error with smaller μ for correlated source

B. Convergence time

The previous section demonstrates the accuracy of the delay-and-extrapolate estimator after convergence to steady-state. In practice, it is important to keep in mind that accuracy is not guaranteed before convergence. To illustrate the convergence properties of the estimator, we show in Figure 3 the evolution of the true misalignment $\|\mathbf{e}_{t,2}\|^2$ and estimated misalignment $(L_2/L_1)\|\mathbf{e}_{t,1}\|^2$ as a function of time index t, averaged over 5000 trials. In each trial, the filter $\mathbf{w}^*$ is a Gaussian vector with norm $10\sqrt{5}$, the initial filter estimate is zero, and $L_1 = 10$, $L_2 = 20$, $\sigma_x^2 = \sigma_z^2 = 1$, and $\mu = 0.0044$. In the uncorrelated case, we have $c = 2$, and in the correlated case, we have $\rho = 0.95$.

In both cases, the estimator starts out as an underestimate of the true misalignment, since $\mathbf{e}_{0,1} = 0$ while $\mathbf{e}_{0,2}$ is large. The estimator is only guaranteed to be accurate once U_t approximates U_∞.

In the uncorrelated case, (13) can be rewritten as

$$M_t - M_\infty = \alpha^t (M_0 - M_\infty) \tag{29}$$

where $\alpha = (1 - 2\mu\sigma_x^2 + \mu^2\sigma_x^4(L + c)) \approx 0.9917$. Solving for the time when $M_\tau = 2M_\infty$, given that $M_0 = 500$ and $M_\infty = 0.067$, we get convergence at $\tau \approx 1000$, after which the estimate is accurate on average.

In the correlated case, the dynamics are more complex and depend on various eigenvalues of $\mathbf{R}$. In the proof of Theorem 2, (D.17) suggests that various components of U_t decay with parameters $\alpha_i \approx 1 - 2\mu\lambda_i$, if μ is small. In our case, $\lambda_1 \approx 14.7$ and $\lambda_{L_2} \approx 0.026$ yielding corresponding convergence times ranging from $\tau_1 \approx 64$ to $\tau_{L_2} \approx 38,000$. As a rule of thumb, to estimate the number of iterations required before the estimator becomes reliable, one could use the average eigenvalue $\bar{\lambda} = \text{Tr}\,\mathbf{R}/L$ and assume geometric decay with $\alpha = 1 - 2\bar{\lambda}$.

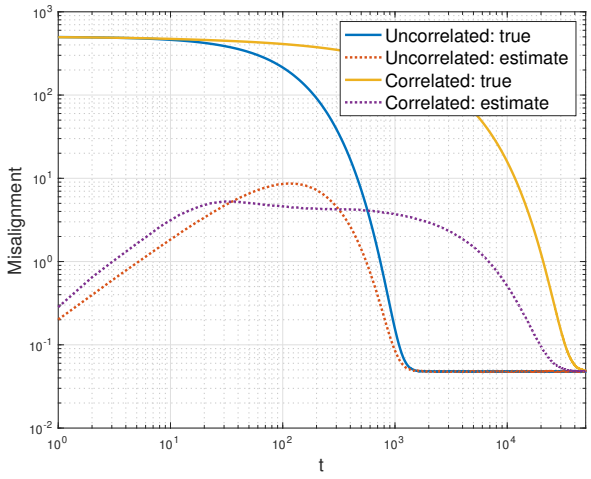


Fig. 3. Dynamics of true and estimated misalignment, averaged over 5000 instances

VI. Conclusion

In this paper, we theoretically analyze the properties of filter error variance matrix and corresponding convergence condition linked to learning rate for both uncorrelated and correlated input signal in adaptive filtering system. We further evaluated the Delay and Extrapolate method and proved that under certain assumption, this method is useful for the estimation of misalignment. The results have been verified by the simulated experiments. This paper lays a solid foundation for future studies on filter mean squared error and related variable learning rate adaptive algorithms.

REFERENCES

[1] Bernard Widrow and Marcian E Hoff, "Adaptive switching circuits," Tech. Rep., Stanford Univ Ca Stanford Electronics Labs, 1960.

[2] Constantin Paleologu, Jacob Benesty, and Silviu Ciochină, "An improved proportionate nlms algorithm based on the l 0 norm," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 309–312.

[3] Kazuhiko Ozeki and Tetsuo Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electronics and Communications in Japan (Part I: Communications)*, vol. 67, no. 5, pp. 19–27, 1984.

[4] S Haykin, "Adaptive filter theory. pearson education india," in *27th Annual International Conference of the Engineering in Medicine and Biology Society*. IEEE Press, 2008, pp. 1212–1215.

[5] Lu Ma, Hua Huang, Pei Zhao, and Tengrong Su, "Acoustic echo cancellation by combining adaptive digital filter and recurrent neural network," *arXiv preprint arXiv:2005.09237*, 2020.

[6] Hongsheng Chen, Teng Xiang, Kai Chen, and Jing Lu, "Nonlinear residual echo suppression based on multi-stream conv-tasnet," *arXiv preprint arXiv:2005.07631*, 2020.

[7] Jung-Hee Kim and Joon-Hyuk Chang, "Attention wave-u-net for acoustic echo cancellation.," in *Interspeech*, 2020, pp. 3969–3973.

[8] Ahmed I Sulyman and Azzedine Zerguine, "Convergence and steady-state analysis of a variable step-size nlms algorithm," *Signal Processing*, vol. 83, no. 6, pp. 1255–1273, 2003.

[9] Hyun-Chool Shin, Ali H Sayed, and Woo-Jin Song, "Variable step-size nlms and affine projection algorithms," *IEEE signal processing letters*, vol. 11, no. 2, pp. 132–135, 2004.

[10] Jacob Benesty, Hernan Rey, Leonardo Rey Vega, and Sara Tressens, "A nonparametric vss nlms algorithm," *IEEE Signal Processing Letters*, vol. 13, no. 10, pp. 581–584, 2006.

[11] Poogyeon Park, Moonsoo Chang, and Namwoong Kong, "Scheduled-stepsize nlms algorithm," *IEEE Signal Processing Letters*, vol. 16, no. 12, pp. 1055–1058, 2009.

[12] Hsu-Chang Huang and Junghsi Lee, "A new variable step-size nlms algorithm and its performance analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 2055–2060, 2011.

[13] Felix Albu, Constantin Paleologu, and Jacob Benesty, "A variable step size evolutionary affine projection algorithm," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 429–432.

[14] Alberto Gonzalez, Miguel Ferrer, Maria de Diego, and Gema Piñero, "An affine projection algorithm with variable step size and projection order," *Digital Signal Processing*, vol. 22, no. 4, pp. 586–592, 2012.

[15] Shirin Jalali, Carl Nuzman, and Yue Sun, "Incremental maximum likelihood estimation for efficient adaptive filtering," *arXiv preprint arXiv:2209.01594*, 2022.

[16] Seiichi Yamamato and Seishi Kitayama, "An adaptive echo canceller with variable step gain method," *IEICE Trans. (1976-1990)*, vol. 65, no. 1, pp. 1–8, 1982.

[17] Constantin Paleologu, Silviu Ciochină, Jacob Benesty, and Steven L Grant, "An overview on optimized nlms algorithms for acoustic echo cancellation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–19, 2015.

[18] S Jaggi and AB Martinez, "Upper and lower bounds of the misadjustment in the lms algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 164–166, 1990.

[19] Aurelio Uncini, *Fundamentals of adaptive signal processing*, Springer, 2015.

[20] Richard J Larsen and Morris L Marx, *Introduction to Mathematical Statistics and Its Applications: Pearson New International Edition PDF EBook*, Pearson Higher Ed, 2013.

[21] Leon Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biometrika*, vol. 12, no. 1/2, pp. 134–139, 1918.

APPENDIX A
PROOF OF THEOREM 1

Recall that $\mathbf{U}_t \triangleq \mathrm{E}[\mathbf{e}_t\mathbf{e}_t^T]$ and define $\hat{\mathbf{U}}_t \triangleq \mathbf{e}_t\mathbf{e}_t^T$ and $\hat{\mathbf{R}}_t \triangleq \mathbf{x}_t\mathbf{x}_t^T$. Then, applying (7) given i) the independence of $\mathbf{x}_t$ and z_t, ii) $\{z_t\}_{t=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_z^2)$ and iii) $\mathrm{E}[\mathbf{x}_t\mathbf{x}_t^T] = \sigma_x^2 I_L$, we have

$$\begin{aligned}\mathbf{U}_{t+1} &= \mathrm{E}[\mathbf{e}_{t+1}\mathbf{e}_{t+1}^T] \\ &= \mathrm{E}[(I-\mu\hat{\mathbf{R}}_t)\mathbf{e}_t\mathbf{e}_t^T(I-\mu\hat{\mathbf{R}}_t)] + \mu^2\sigma_x^2\sigma_z^2 I_L \\ &= (1-2\mu\sigma_x^2)\mathbf{U}_t + \mu^2\sigma_z^2\sigma_x^2 I_L + \mu^2\,\mathrm{E}[\hat{\mathbf{R}}_t\hat{\mathbf{U}}_t\hat{\mathbf{R}}_t]. \end{aligned} \tag{A.1}$$

To simplify the third term, let $N_t \triangleq \mathrm{E}[\hat{\mathbf{R}}_t\hat{\mathbf{U}}_t\hat{\mathbf{R}}_t]$. Note that

$$\begin{aligned}\mathbf{N}_t &= \mathrm{E}[\mathbf{x}_t\mathbf{x}_t^T\hat{\mathbf{U}}_t\mathbf{x}_t\mathbf{x}_t^T] \\ &= \mathrm{E}[\mathbf{x}_t(\mathbf{x}_t^T\hat{\mathbf{U}}_t\mathbf{x}_t)\mathbf{x}_t^T] \\ &= \mathrm{E}[\mathbf{x}_t(\sum_{k=1}^{L}\sum_{l=1}^{L}(\mathbf{x}_t)_k(\hat{\mathbf{U}}_t)_{k,l}(\mathbf{x}_t)_l)\mathbf{x}_t^T].| \end{aligned} \tag{A.2}$$

Therefore,

$$(\mathbf{N}_t)_{i,j} = \mathrm{E}[(\mathbf{x}_t)_i(\sum_{k=1}^{L}\sum_{l=1}^{L}(\mathbf{x}_t)_k(\hat{\mathbf{U}}_t)_{k,l}(\mathbf{x}_t)_l)(\mathbf{x}_t)_j].$$

To simplify the notation, we temporarily drop the subscript t, and let x_i, e_j, w_k denote the i-th, j-th, k-th entries of $\mathbf{x}_t$, $\mathbf{e}_t$ and $\mathbf{w}_t$, respectively. Then

$$(\mathbf{N}_t)_{ij} = \mathrm{E}[x_i(\sum_{k=1}^{L}\sum_{l=1}^{L}x_k e_k e_l x_l)x_j]. \tag{A.3}$$

For $i=j$,

$$\begin{aligned}(\mathbf{N}_t)_{i,i} &= \mathrm{E}[x_i^2\sum_{k=1}^{L}\sum_{l=1}^{L}x_k x_l e_k e_l] \\ &= 0 + \mathrm{E}[x_i^4 e_i^2] + \sum_{k\neq i}^{L}\mathrm{E}[x_i^2 x_k^2 e_k^2] \qquad \text{(A.4)} \\ &= c\sigma_x^4\,\mathrm{E}[e_i^2] + \sigma_x^4\sum_{k=1}^{L}\mathrm{E}[e_k^2], \qquad \text{(A.5)}\end{aligned}$$

where the last line follows because $\mathrm{E}[x_i^4] = (c+1)\sigma_x^4$. For $i \neq j$,

$$\begin{aligned}(\mathbf{N}_t)_{i,j} &= 0 + \mathrm{E}[x_i^2x_j^2e_ie_j] + \mathrm{E}[x_i^2x_j^2e_je_i] \qquad \text{(A.6)} \\ &= \sigma_x^4\,\mathrm{E}[e_ie_j] + \sigma_x^4\,\mathrm{E}[e_je_i] = 2\sigma_x^4\,\mathrm{E}[e_ie_j]. \qquad \text{(A.7)}\end{aligned}$$

Therefore, in summary,

$$\mathbf{N}_t = \sigma_x^4(2\mathbf{U}_t + (c-2)\,\mathrm{diag}(\mathbf{U}_t) + \mathrm{Tr}\,(\mathbf{U}_t)I_L). \tag{A.8}$$

Combining (A.1) and (A.8) yields (11). Applying Lemma 1 , proved in Appendix D, with $\mathbf{R} = \sigma_x^2 I_L$, yields (12). In applying the lemma, we use $\phi_c(\mu,\sigma_x^2 I_L) = \mu L\sigma_x^2/\big(2-\mu c\sigma_x^2\big)$ and hence $\mu^* = 2/\big((L+c)\sigma_x^2\big)$.

APPENDIX B
PROOF OF THEOREM 2

Since $\mathbf{x}_t$ and $\mathbf{e}_t$ are independent, following the steps used in analyzing (A.1), we have

$$\begin{aligned}\mathbf{U}_{t+1} =& \mathbf{U}_t - \mu\mathbf{R}\mathbf{U}_t - \mu\mathbf{U}_t\mathbf{R} \\ &+ \mu^2\sigma_z^2\mathbf{R} + \mu^2\,\mathrm{E}[\hat{\mathbf{R}}_t\mathbf{e}_t\mathbf{e}_t^T\hat{\mathbf{R}}_t^T]. \end{aligned} \tag{B.9}$$

Here, as before, $\hat{\mathbf{R}}_t = \mathbf{x}_t\mathbf{x}_t^T$. Define $\mathbf{N}_t \triangleq \mathrm{E}[\hat{\mathbf{R}}_t\mathbf{e}_t\mathbf{e}_t^T\hat{\mathbf{R}}_t^T]$. To simplify the proof, similar to the proof of Theorem 1, we temporarily drop the subscript t. Then,

$$(\mathbf{N}_t)_{ij} = \sum_{k,l}(\mathbf{U}_t)_{k,l}\,\mathrm{E}[x_ix_kx_lx_j]. \tag{B.10}$$

Since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_L,\mathbf{R})$, employing Isserlis' Theorem [21], it follows that

$$\mathrm{E}[x_ix_kx_lx_j] = \mathbf{R}_{ik}\mathbf{R}_{lj} + \mathbf{R}_{il}\mathbf{R}_{kj} + \mathbf{R}_{ij}\mathbf{R}_{kl}.$$

Using the symmetry of $\mathbf{U}_t$, we have

$$(\mathbf{N}_t)_{i,j} = 2(\mathbf{R}\mathbf{U}_t\mathbf{R})_{ij} + \mathbf{R}_{ij}\sum_{k,l}(\mathbf{U}_t)_{kl}\mathbf{R}_{kl}. \tag{B.11}$$

Further using symmetry of $\mathbf{U}_t$, we have

$$\begin{aligned}\sum_{k,l}(\mathbf{U}_t)_{kl}\mathbf{R}_{kl} &= \sum_k\sum_l\mathbf{R}_{kl}(\mathbf{U}_t)_{lk} \\ &= \sum_k(\mathbf{R}\mathbf{U}_t)_{kk} = \mathrm{Tr}(\mathbf{R}\mathbf{U}_t)\end{aligned}$$

and so

$$\mathbf{N}_t = 2\mathbf{R}\mathbf{U}_t\mathbf{R} + \mathrm{Tr}(\mathbf{R}\mathbf{U})\mathbf{R} \tag{B.12}$$

which yields the desired expression (16). Applying Lemma 1, proved in Appendix D, with $c=2$, yields (17).

APPENDIX C
PROOF OF COROLLARY 2

The limit expression (19) follows directly from (5) and the definition of $\phi_c(\mu,\mathbf{R})$.

It remains to prove the bound (18). Let $\mathbf{R} = V\mathbf{\Lambda}V^T$ be the eigensystem of $\mathbf{R}$, and define $\mathbf{B}_t = V^T\mathbf{U}_tV$. Using the similarity invariance of the trace operator, we have

$$\mathrm{Tr}(\mathbf{R}^n\mathbf{U}_t) = \mathrm{Tr}(\mathbf{\Lambda}^n\mathbf{B}_t) \leq \lambda_{\max}^{n-1}\,\mathrm{Tr}(\mathbf{\Lambda}\mathbf{B}_t) = \lambda_{\max}^{n-1}\,\mathrm{Tr}(\mathbf{R}\mathbf{U}_t).$$

Similarly, $\mathrm{Tr}(\mathbf{R}^n\mathbf{U}_t) \geq \lambda_{\min}^{n-1}\,\mathrm{Tr}(\mathbf{R}\mathbf{U}_t)$.

Multiplying (16) by $\mathbf{R}$ and taking the trace operator yields

$$\begin{aligned}\mathrm{Tr}(\mathbf{R}\mathbf{U}_{t+1}) =& \,\mathrm{Tr}(\mathbf{R}\mathbf{U}_t) - 2\mu\,\mathrm{Tr}(\mathbf{R}^2\mathbf{U}_t) + \mu^2\big[\sigma_z^2\,\mathrm{Tr}(\mathbf{R}^2) \\ &+2\,\mathrm{Tr}(\mathbf{R}^3\mathbf{U}_t) + \mathrm{Tr}(\mathbf{R}\mathbf{U}_t)\,\mathrm{Tr}(\mathbf{R}^2)\big] \\ \leq& \big(1-2\mu\lambda_{\min}+2\mu^2\lambda_{\max}^2+\mu^2\,\mathrm{Tr}(\mathbf{R}^2)\big)\,\mathrm{Tr}(\mathbf{R}U_t) \\ &+\mu^2\sigma_z^2\,\mathrm{Tr}(\mathbf{R}^2) \end{aligned} \tag{C.13}$$

The result then follows from (5).

APPENDIX D
MATRIX SYSTEM CONVERGENCE LEMMA

Lemma 1. *Let $\mathbf{R}$ be a positive-definite matrix of dimension L with eigensystem $\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, and choose $c \geq 0$. Let μ^* be the smallest positive solution to $\phi_c(\mu^*, \mathbf{R}) = 1$. If $L + c \geq 2$ and $0 < \mu < \mu^*$, or if $L = 1$ and $0 < \mu < \mu^*/2$, then the matrix dynamical system defined by*

$$\begin{aligned}\mathbf{U}_{t+1} =& \mathbf{U}_t - \mu\mathbf{R}\mathbf{U}_t - \mu\mathbf{U}_t\mathbf{R} \\ &+ \mu^2\big[\sigma_z^2\mathbf{R} + 2\mathbf{R}\mathbf{U}_t\mathbf{R} + \mathrm{Tr}(\mathbf{R}\mathbf{U}_t)\mathbf{R} \\ &+ (c-2)\mathbf{R}\mathbf{V}\,\mathrm{diag}(\mathbf{V}^T\mathbf{U}_t\mathbf{V})\mathbf{V}^T\mathbf{R}\big]. \end{aligned} \tag{D.14}$$

converges to

$$\mathbf{U}_\infty = \lim_{t\to\infty}\mathbf{U}_t = \frac{\mu\sigma_z^2}{1-\phi_c(\mu,R)}(2I_L - \mu cR)^{-1} \tag{D.15}$$

Proof:

Applying the change of variables $\mathbf{B}_t = \mathbf{V}^T\mathbf{U}_t\mathbf{V}$ and using the orthogonality of $\mathbf{V}$, (D.14) becomes

$$\begin{aligned}\mathbf{B}_{t+1} =& \mathbf{B}_t - \mu\mathbf{\Lambda}\mathbf{B}_t - \mu\mathbf{B}_t\mathbf{\Lambda} \\ &+ \mu^2\big[\sigma_z^2\mathbf{\Lambda} + 2\mathbf{\Lambda}\mathbf{B}_t\mathbf{\Lambda} + \mathrm{Tr}(\mathbf{\Lambda}\mathbf{B}_t)\mathbf{\Lambda} \\ &+ (c-2)\mathbf{\Lambda}\,\mathrm{diag}(\mathbf{B}_t)\mathbf{\Lambda}\big]. \end{aligned} \tag{D.16}$$

Let $\mathbf{b}_t \in \mathbb{R}^L$ and $\boldsymbol{\lambda} \in \mathbb{R}^L$ be vectors formed from the diagonal elements of $\mathbf{B}_t$ and $\mathbf{\Lambda}$ respectively. These diagonal elements satisfy the dynamical system

$$\mathbf{b}_{t+1} = \Big(I_L - 2\mu\mathbf{\Lambda} + \mu^2 c\mathbf{\Lambda}^2 + \mu^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)\mathbf{b}_t + \mu^2\sigma_z^2\boldsymbol{\lambda}. \tag{D.17}$$

not involving the off-diagonal elements of B_t. If the matrix $\mathbf{\Theta} := 2\mu\mathbf{\Lambda} - \mu^2 c\mathbf{\Lambda}^2 - \mu^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T$ is non-singular, and if the matrix $I_L - \Theta$ has norm less than 1, then the system (D.17) will converge to the limit

$$\mathbf{b}_\infty = \Big(2\mu\mathbf{\Lambda} - \mu^2 c\mathbf{\Lambda}^2 - \mu^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\mu^2\sigma_z^2\boldsymbol{\lambda}. \tag{D.18}$$

Both of the conditions will hold if all of the eigenvalues of $\mathbf{\Theta}$ lie in $(0, 2)$.

To upper bound the eigenvalues of $\mathbf{\Theta}$, we can take $c = 0$. By the Gershgorin circle theorem, the maximum eigenvalue of $\mathbf{\Theta}$ is upper bounded by

$$\begin{aligned}\lambda_{\max}(\mathbf{\Theta}) \leq& \max_i(2\mu\lambda_i - \mu^2\lambda_i^2 + \mu^2\lambda_i\sum_{j\neq i}\lambda_j) \\ \leq& \max_i(\mu\lambda_i(2 - \mu\lambda_i + \mu(\mathrm{Tr}\,\mathbf{R} - \lambda_i))) \\ \leq& \max_i(\mu\lambda_i(2 + \mu\,\mathrm{Tr}\,\mathbf{R} - 2\mu\lambda_i)) \\ \leq& \frac{1}{8}(2 + \mu\,\mathrm{Tr}\,\mathbf{R})^2 \end{aligned} \tag{D.19}$$

where the last line is obtained by maximizing the previous line over all real λ_i. Since $\mu < \mu^* \leq 2/\,\mathrm{Tr}\,\mathbf{R}$, via (23), the max eigenvalue of $\mathbf{\Theta}$ is less than 2.

To show that all eigenvalues of $\mathbf{\Theta}$ are positive, note that a matrix of the form $D - \mathbf{a}\mathbf{a}^T$ with invertible D has inverse

$$\big(D - \mathbf{a}\mathbf{a}^T\big)^{-1} = D^{-1} + \frac{D^{-1}\mathbf{a}\mathbf{a}^T D^{-1}}{1 - \mathbf{a}^T D^{-1}\mathbf{a}}. \tag{D.20}$$

as long as $\mathbf{a}^T D^{-1}\mathbf{a} \neq 1$. Taking $D = 2\mu\mathbf{\Lambda} - \mu^2 c\mathbf{\Lambda}^2$ and $\mathbf{a} = \mu\boldsymbol{\lambda}$, we see that $\mathbf{\Theta}$ is invertible unless $\mathbf{a}^T D^{-1}\mathbf{a} = 1$. We know D is invertible whenever $\mu < 2/(c\lambda_k)$ for all k. This follows from the fact that $\mu^* < 2/(c\lambda_k)$ for each k, since the individual summands in the definition of $\phi_c(\mu, \mathbf{R})$ increase from 0 to ∞ as μ increases from 0 to $2/(c\lambda_k)$. Moreover,

$$\begin{aligned}\mathbf{a}^T D^{-1}\mathbf{a} &= \sum_k \frac{\mu^2\lambda_k^2}{2\mu\lambda_k - c\mu^2\lambda_k^2} \\ &= \sum_k \frac{\mu\lambda_k}{2 - c\mu\lambda_k} = \phi_c(\mu, \mathbf{R}). \end{aligned} \tag{D.21}$$

so $\mathbf{\Theta}$ is non-singular as long as $\phi_c(\mu, \mathbf{R}) \neq 1$. Given that $\phi_c(\mu, \mathbf{R}) < 1$ on $0 < \mu < \mu^*$, the eigenvalues of $\mathbf{\Theta}$ are non-zero on this interval. Since all eigenvalues of $\mathbf{\Theta}$ are positive for sufficiently small $\mu > 0$, continuity of eigenvalues implies that all eigenvalues of $\mathbf{\Theta}$ must be positive on $0 < \mu < \mu^*$. Thus we have shown convergence of the diagonal entries of $\mathbf{B}_t$ to $\mathbf{b}_\infty$ if $0 < \mu < \mu^*$.

From (D.18), (D.20), and (D.21), the limit achieved for the diagonal values of $\mathbf{B}_t$ is

$$\begin{aligned}\mathbf{b}_\infty &= \frac{1}{1 - \mathbf{a}^T D^{-1}\mathbf{a}} D^{-1}\mathbf{a}\mu\sigma_z^2 \\ &= \frac{\mu\sigma_z^2}{1 - \phi_c(\mu, \mathbf{R})}(2I_L - \mu c\mathbf{\Lambda})^{-1}\mathbf{1}. \end{aligned} \tag{D.22}$$

The last step of the proof is to show that the off-diagonal values of $\mathbf{B}_t$ go to zero in the limit. In that case,

$$\mathbf{B}_\infty = \lim_{t\to\infty}\mathbf{B}_t = \frac{\mu\sigma_z^2}{1 - \phi_c(\mu, \mathbf{R})}(2I_L - \mu c\mathbf{\Lambda})^{-1} \tag{D.23}$$

so that $\mathbf{U}_\infty = \mathbf{V}\mathbf{B}_\infty\mathbf{V}^T$ has the desired form.

From (D.16), the off-diagonal values of $\mathbf{B}_t$ satisfy the dynamics

$$(\mathbf{B}_{t+1})_{nm} = \big(1 - \mu\lambda_n - \mu\lambda_m + 2\mu^2\lambda_n\lambda_m\big)(\mathbf{B}_t)_{nm}$$

Convergence to zero will be established by showing that for $0 < \mu < \mu^*$, the factor $\alpha_{nm} := \mu\lambda_n + \mu\lambda_m - 2\mu^2\lambda_n\lambda_m$ lies in $(0, 2)$ for each n, m.

To show $\alpha_{nm} > 0$, define $t_{n,m} = \lambda_n + \lambda_m$. Considering $t_{n,m}$ fixed and varying λ_n, we have

$$\begin{aligned}\alpha_{nm} &= \mu t_{nm} - 2\mu^2\lambda_n(t_{nm} - \lambda_n) \\ &\geq \mu t_{nm} - \frac{\mu^2 t_{nm}^2}{2} \\ &\geq \mu t_{nm}\left(1 - \frac{\mu t_{nm}}{2}\right). \end{aligned} \tag{D.24}$$

Thus $\alpha_{nm} > 0$ for $0 < \mu < 2/t_{nm}$. Since we have $\mu < \mu^* \leq 2/\,\mathrm{Tr}(\mathbf{R}) \leq 2/t_{nm}$ if follows that $\alpha_{nm} > 0$.

To show $\alpha_{nm} < 2$, it is sufficient to show $\mu(\lambda_n + \lambda_m) < 2$.

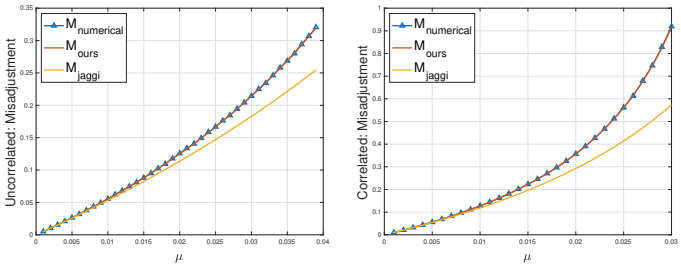
We have

$$\mu(\lambda_n + \lambda_m) < \mu^*(\lambda_n + \lambda_m) \leq \mu^* \operatorname{Tr}(\mathbf{R}) \leq 2.$$

□

APPENDIX E
EXACTNESS OF ANALYSIS

Analysis of the LMS algorithm that we are aware of in literature, well-represented by [18], make certain approximations that yield small differences compared with our results in Section III. In this section, we make a note of these differences and illustrate them numerically.



(a) Uncorrelated input signals (b) Correlated input signals

Fig. 4. Comparison to different misadjustments over step size

In the uncorrelated case, the usual approximation is

$$M_\infty^{\text{approx}} = \frac{\mu\sigma_x^2 L}{2 - (L+1)\mu\sigma_x^2}$$

while while Corollary 1 gives

$$M_\infty = \left(\frac{\mu\sigma_x^2 L}{2 - (L+c)\mu\sigma_x^2}\right).$$

The difference is the coefficient of variation of $\mathbf{x}_{t,i}^2$, $c = \operatorname{var}(x_{t,i}^2)/\sigma_x^4$. Our result shows that the region of stability gets smaller and the misadjustment gets larger for heavyier tailed input distributions. Figure 4(a) plots both M_∞ and M_∞^{approx} versus μ for the uncorrelated case, with $c = 5$, $\sigma_x^2 = 2$, and $L = 20$. We also performed empirical experiments, running the LMS algorithm for 20,000 iterations with Laplace-distributed, i.i.d inputs, measuring the final misadjustment, and averaging over 5,000 independent trials. The empirical results match our analysis, and show the effect of the c parameter.

In the correlated case, we only considered Gaussian inputs. The approximation in [18] is expressed, in our notation,

$$M_\infty^{\text{approx}} = \frac{2\phi_2(\frac{\mu}{2}, \mathbf{R})}{1 - 2\phi_2(\frac{\mu}{2}, \mathbf{R})}$$

while Corollary 2 gives

$$M_\infty = \frac{\phi_2(\mu, \mathbf{R})}{1 - \phi_2(\mu, \mathbf{R})}$$

These two values are plotted versus μ in Figure 4(b), where we see that they agree for small μ but differ as μ approaches μ^*. Experiments are again performed using 20,000 iterations of the LMS algorithm over 5,000 trials, to obtain an empirical curve, which closely matches our result. In these experiments and formulas, the covariance matrix was defined by $R_{ij} = 0.95^{-|i-j|}$, with $L = 20$ and $\sigma_z^2 = 1$.