Steady State and Transient MSE Analysis of Convexly Constrained Mixture Methods

Mehmet A. Donmez and Suleyman S. Kozat

***Abstract*—We investigate convexly constrained mixture methods to adaptively combine outputs of two adaptive filters running in parallel to model a desired unknown system. We compare several algorithms with respect to their mean-square error in the steady state, when the underlying unknown system is nonstationary with a random walk model. We demonstrate that these algorithms are universal such that they achieve the performance of the best constituent filter in the steady state if certain algorithmic parameters are chosen properly. We also demonstrate that certain mixtures converge to the optimal convex combination filter such that their steady-state performances can be better than the best constituent filter. We also perform the transient analysis of these updates in the mean and mean-square error sense. Furthermore, we show that the investigated convexly constrained algorithms update certain auxiliary variables through sigmoid nonlinearity, hence, in this sense, related.**

***Index Terms*— Adaptive filtering, combination methods, convex mixtures, steady-state analysis, transient analysis.**

I. Introduction

In this correspondence, we first investigate and compare four well-known convexly constrained adaptive mixture methods to combine outputs of two adaptive filters [1]–[4] with respect to their mean-square error (MSE) in the steady state. We then perform the transient analysis of these convexly constrained updates in the mean and the MSE senses. In this widely studied framework, we have two adaptive filters that work in parallel in order to model an unknown system [1]. The outputs of these algorithms are then combined using another adaptive method in order to improve the overall performance [1]. The first adaptive algorithm [1] uses a stochastic gradient update on the convexly constrained mixture parameter to minimize the final estimation error. The second algorithm is based on the exponentiated gradient (EG) algorithm [2], [5]. The EG algorithm has extensive roots in sequential learning theory [6], [7] and minimizes an approximate final estimation error while penalizing the distance between the new and the old mixture parameters. The third [3] and the fourth algorithms [4] use specific performance-based updates on the mixture parameters as further detailed in Section III. Although we specifically concentrate on the combination of two filters for presentation clarity, our results can be readily extended to mixtures having more than two filters [8].

Mixture approaches are shown to improve the steady-state and transient performance over the constituent filters under certain scenarios [1], [3], [9], [10]. The steady-state analysis of convexly constrained, affinely constrained and unconstrained mixtures are carried out in [1], [9], [10], respectively. Specifically, the adaptive convex mixture of [1] is shown to be universal with respect to the constituent filters such that this algorithm achieves the excess MSE (EMSE) performance of the best constituent filter and, in certain cases, even outperforms both [1]. Furthermore, the convexly constrained mixture methods are extensively studied in sequential learning theory under the mixture of experts framework [2], [3] and shown to be "universal" in a strong and deterministic sense such that they asymptotically achieve the performance of the best algorithm in the mixture for any bounded but arbitrary real-valued sequence. However, the results in [2], [3], [6], and [7] hold for deterministic and bounded sequences. The boundedness assumption is not correct, as an example, for Gaussian random sequences.

In this correspondence, we show that if we use the EG algorithm to update the mixing parameter, the resulting combination filter is universal with respect to the constituent filters such that the combination filter performs, at least, as well as the best constituent filter in the steady state. Specifically, we show that the EMSE of the combination filter is as small as the best of the constituent filters and, in some cases, smaller than EMSEs of the component filters in the steady state. We also show that the mixture parameter under the EG update converges to the optimum convex combination parameter that minimizes the EMSE. Note that the EG algorithm is shown to converge faster and has better tracking performance than the stochastic gradient algorithms for sparse impulse responses in certain situations [2], [5], [11]. Hence, the EG algorithm can be preferred over the stochastic gradient based algorithms for mixtures having more than two filters and when the combination favors only a few of the constituent filters. We point out that although the MSE of the EG algorithm is studied using Euler discretization in [11] under certain assumptions for uncorrelated input regressors, our framework and the analysis are significantly different since we use the EG algorithm to combine outputs of adaptive filters, which are nonlinearly coupled, such that the assumptions of [11] do not hold. The third algorithm that we investigate is based on a certain performance-based mixture of the constituent filters [3]. We analyze the steady-state behavior of [3] and show that with a proper selection of the forgetting factor, the combination filter is universal such that it performs as well as the best constituent filter in the steady state. Although the algorithm of [3] is also shown to be universal in a strong deterministic sense [3], however, we show that the mixture parameter does not converge to the optimum convex combination parameter under our assumptions (which is also supported by our experiments). The fourth algorithm that we investigate was studied in [4] and combines filters based on their performances within a time window. We demonstrate that if the mixture parameter in [4] is selected using a sufficiently large time window, the combination filter can achieve the performance of the best constituent filter in the steady state. For all algorithms, we also perform the transient analysis in the mean and the MSE senses.

In Section II, we first briefly describe the mixture framework for the combination of two adaptive filters running in parallel with the error quantities and performance measures. In Section III, we investigate four mixture methods in detail and provide MSEs along with the converged mixture weights in the steady state. In Section IV, we provide a transient analysis of the corresponding algorithms. We illustrate the introduced results through simulations under the setup of [1] in Section V. Our results accurately describe the behavior of these algorithms both in the steady state and during convergence in the studied setup. The correspondence concludes with certain remarks.

Manuscript received August 12, 2011; revised November 15, 2011 and February 07, 2012; accepted February 12, 2012. Date of publication March 08, 2012; date of current version May 11, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Maciej Niedzwiecki. This work is supported in part by IBM Faculty Award and Outstanding Young Scientist Award Program, Turkish Academy of Sciences.

The authors are with the Electrical Engineering Department, Koc University, 06660 Istanbul, Turkey (e-mail: skozat@ku.edu.tr; medonmez@ku.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2012.2189110

II. Problem Description

In this framework, we have two adaptive algorithms that run in parallel to model a desired signal $d(t)$. The desired signal $d(t)$ is given by $d(t) = \boldsymbol{w}_o^T(t)\boldsymbol{u}(t) + n(t)$, where $\boldsymbol{w}_o(t) \in \mathbb{R}^p$ is the desired system vector that varies according to a random walk model [12], i.e., $\boldsymbol{w}_o(t+1) = \boldsymbol{w}_o(t) + \boldsymbol{q}(t)$, where $\boldsymbol{q}(t)$ is a zero mean, i.i.d. random vector with covariance matrix $\boldsymbol{Q} = E[\boldsymbol{q}(t)\boldsymbol{q}^T(t)]$, $\boldsymbol{u}(t) \in \mathbb{R}^p$ is the input regressor with zero mean and correlation matrix $\boldsymbol{R} = E[\boldsymbol{u}(t)\boldsymbol{u}^T(t)]$ and the observation noise $n(t)$ is i.i.d. with zero mean and variance $E[n^2(t)] = \sigma_n^2$. The cross correlation vector between the desired signal and the input regressor is $\boldsymbol{p}(t) = E[d(t)\boldsymbol{u}(t)]$. To model the desired signal $d(t)$, we have two parallel running constituent filters each producing estimates $\hat{d}_1(t) = \boldsymbol{w}_1^T(t)\boldsymbol{u}(t)$ and $\hat{d}_2(t) = \boldsymbol{w}_2^T(t)\boldsymbol{u}(t)$ using the weight vectors $\boldsymbol{w}_1(t), \boldsymbol{w}_2(t)$ respectively. For each constituent filter, we define the estimation error, the *a priori* error and the *a posteriori* error as $e_i(t) \triangleq d(t) - \hat{d}_i(t) = d(t) - \boldsymbol{w}_i^T(t)\boldsymbol{u}(t)$, $e_{a,i}(t) \triangleq [\boldsymbol{w}_o(t) - \boldsymbol{w}_i(t)]^T\boldsymbol{u}(t)$ and $e_{p,i}(t) \triangleq [\boldsymbol{w}_o(t) - \boldsymbol{w}_i(t+1)]^T\boldsymbol{u}(t)$, respectively. For each filter, we also define MSE as $J_i(t) \triangleq E[e_i^2(t)]$ and excess MSE as $J_{\mathrm{ex},i}(t) \triangleq E[e_{a,i}^2(t)]$, with limiting values $J_i \triangleq \lim_{t\to\infty} J_i(t)$, $J_{\mathrm{ex},i} \triangleq \lim_{t\to\infty} J_{\mathrm{ex},i}(t)$ (if the limits exist). We also define the cross correlation between the *a priori* errors as $J_{\mathrm{ex},12}(t) \triangleq E[e_{a,1}(t)e_{a,2}(t)]$ with limiting value $J_{\mathrm{ex},12} \triangleq \lim_{t\to\infty} J_{\mathrm{ex},12}(t)$. We also define $\Delta J_i(t) = J_{\mathrm{ex},i}(t) - J_{\mathrm{ex},12}(t)$ for $i = 1, 2$ with the limiting values $\Delta J_i = J_{\mathrm{ex},i} - J_{\mathrm{ex},12}$ [1].

The outputs of the constituent filters are then combined using another adaptive layer to produce the final estimate of the desired signal as $\hat{d}(t) = \lambda(t)\hat{d}_1(t) + [1-\lambda(t)]\hat{d}_2(t)$, where $\lambda(t)$ is the mixing parameter constrained to be in [0, 1]. If $\boldsymbol{y}(t) \triangleq [\hat{d}_1(t)\ \hat{d}_2(t)]^T$ and $\boldsymbol{w}(t) \triangleq [\lambda(t)\ 1-\lambda(t)]^T$, then we have $\hat{d}(t) = \boldsymbol{w}^T(t)\boldsymbol{y}(t)$. The final estimation error is given as $e(t) = d(t) - \hat{d}(t)$. In this correspondence, we investigate four methods to train the combination weight $\lambda(t)$. Assuming convergence, the optimal mean combination weights in terms of minimizing the MSE under convex constraint are given by [1]

$$\boldsymbol{w}_{\mathrm{o},c} \triangleq \begin{cases} [1 \quad 0]^T : & J_{\mathrm{ex},1} \le J_{\mathrm{ex},12} \le J_{\mathrm{ex},2} \\ [0 \quad 1]^T : & J_{\mathrm{ex},2} \le J_{\mathrm{ex},12} \le J_{\mathrm{ex},1} \\ [\frac{\Delta J_2}{\Delta J_1+\Delta J_2} \quad \frac{\Delta J_1}{\Delta J_1+\Delta J_2}]^T : & J_{\mathrm{ex},12} < J_{\mathrm{ex},i},\ i=1,2 \end{cases} \quad (1)$$

in the steady state.

III. Steady-State Performances of Convexly Constrained Mixtures

In this section, we analyze four methods to train the mixture parameter $\lambda(t)$. The *a priori* error of the combination filter is $e_a(t) = \lambda(t)e_{a,1}(t) + (1-\lambda(t))e_{a,2}(t)$. If $J_{\mathrm{ex}}(t) \triangleq E[e_a^2(t)]$, then we get

$$J_{\mathrm{ex}}(t) = E\Big[\lambda^2(t)e_{a,1}^2(t) + (1-\lambda(t))^2 e_{a,2}^2(t) + 2\lambda(t)(1-\lambda(t))e_{a,1}(t)e_{a,2}(t)\Big]$$

and $J_{\mathrm{ex}} \triangleq \lim_{t\to\infty} J_{\mathrm{ex}}(t)$. Without loss of generality, we assume that $J_{\mathrm{ex},1} < J_{\mathrm{ex},2}$ in the following. Hence, for each algorithm, we have two separate cases depending on the relative value of $J_{\mathrm{ex},12}$, i.e., $J_{\mathrm{ex},1} \le J_{\mathrm{ex},12} < J_{\mathrm{ex},2}$ or $J_{\mathrm{ex},12} < J_{\mathrm{ex},1} < J_{\mathrm{ex},2}$, to investigate the steady-state behavior.

A. Algorithm 1

For the convexly constrained algorithm from [1], the mixture parameter is given by

$$\lambda_\alpha(t) = \frac{1}{1+\exp[-\alpha(t)]}$$

where $\alpha(t)$ is trained using a stochastic gradient update to minimize the final prediction error as

$$\begin{aligned}\alpha(t+1) &= \alpha(t) - \frac{\mu_\alpha}{2}\nabla_{\alpha(t)}e^2(t) \\ &= \alpha(t) + \mu_\alpha e(t)\left[\hat{d}_1(t) - \hat{d}_2(t)\right]\lambda_\alpha(t)[1-\lambda_\alpha(t)]. \end{aligned} \quad (2)$$

For (2), we have [1]

$$J_{\mathrm{ex}} = \begin{cases} J_{\mathrm{ex},1} : & J_{\mathrm{ex},1} \le J_{\mathrm{ex},12} < J_{\mathrm{ex},2} \\ J_{\mathrm{ex},12} + \frac{\Delta J_1 \Delta J_2}{\Delta J_1 + \Delta J_2} : & J_{\mathrm{ex},12} < J_{\mathrm{ex},1} < J_{\mathrm{ex},2} \end{cases}$$

where $J_{\mathrm{ex},12} + \frac{\Delta J_1 \Delta J_2}{\Delta J_1+\Delta J_2} < J_{\mathrm{ex},1}$. Furthermore, if $\boldsymbol{w}_\alpha(t) \triangleq [\lambda_\alpha(t)\ 1-\lambda_\alpha(t)]^T$, then we have [1]

$$\lim_{t\to\infty} E[\boldsymbol{w}_\alpha(t)] = \begin{cases} [1\ 0]^T : & J_{\mathrm{ex},1} \le J_{\mathrm{ex},12} < J_{\mathrm{ex},2} \\ [\frac{\Delta J_2}{\Delta J_1+\Delta J_2} \quad \frac{\Delta J_1}{\Delta J_1+\Delta J_2}]^T : & J_{\mathrm{ex},12} < J_{\mathrm{ex},1} < J_{\mathrm{ex},2}. \end{cases}$$

Hence, in the steady state, the mixture performs as well as the best component filter and, in some cases, outperforms both. Moreover, the combination weight vector $\boldsymbol{w}_\alpha(t)$ converges to the optimal weight vector $\boldsymbol{w}_{\mathrm{o},c}$ under the convex constraint.

B. Algorithm 2

The second convexly constrained update is based on the EG algorithm [2]. The EG algorithm has extensive roots in competitive online learning theory and has been used in different signal processing problems such as in echo cancellation [5], [11]. Here, we use the EG algorithm to train the mixture weights, where the combination weight is updated as [2], [5]

$$\begin{aligned}\lambda_\rho(t+1) &= \arg\min_{\lambda\in[0,1]}\Bigg\{\mathrm{d}(\boldsymbol{w}, \boldsymbol{w}_\rho(t)) + \frac{\mu_\rho}{2} \\ &\quad \times\Bigg[e^2(t) + \frac{\partial(d(t) - \boldsymbol{w}^T\boldsymbol{y}(t))^2}{\partial\lambda}\Big|_{\lambda=\lambda_\rho(t)} \\ &\quad \times(\lambda - \lambda_\rho(t))\Bigg]\Bigg\} \quad (3) \\ &= \lambda_\rho(t)\exp\left[\mu_\rho e(t)\hat{d}_1(t)\right]\lambda_\rho(t)\exp\left[\mu_\rho e(t)\hat{d}_1(t)\right] \\ &\quad + [1-\lambda_\rho(t)]\exp\left[\mu_\rho e(t)\hat{d}_2(t)\right] \quad (4)\end{aligned}$$

where $\mathrm{d}(\boldsymbol{w}, \boldsymbol{w}_\rho(t)) = \lambda\ln\left(\frac{\lambda}{\lambda_\rho(t)}\right) + (1-\lambda)\ln\left(\frac{1-\lambda}{1-\lambda_\rho(t)}\right)$ is the Kullback–Leibler distance between the old and new weights, the second term on the right hand side of (3) is the first order Taylor's approximation of $(d(t) - \boldsymbol{w}^T\boldsymbol{y}(t))^2$ around $\lambda = \lambda_\rho(t)$, measuring the "fit" of the new weight to the data, $\boldsymbol{w} = [\lambda\ 1-\lambda]^T$, $\boldsymbol{w}_\rho(t) \triangleq [\lambda_\rho(t)\ 1-\lambda_\rho(t)]^T$ and $e(t) = d(t) - \boldsymbol{w}_\rho^T(t)\boldsymbol{y}(t)$. After recognizing $e^{-\rho(t)} = \frac{1-\lambda_\rho(t)}{\lambda_\rho(t)}$ and some algebra, the update on $\lambda_\rho(t)$ in (4) can be written as

$$\lambda_\rho(t) = \frac{1}{1+\exp[-\rho(t)]} \quad (5)$$

with

$$\begin{aligned}\rho(t+1)&=\rho(t)+\mu_\rho e(t)\left(\hat{d}_1(t)-\hat{d}_2(t)\right)\\&=\rho(t)+\mu_\rho\left\{\lambda_\rho(t)e_{a,1}(t)+[1-\lambda_\rho(t)]\,e_{a,2}(t)+n(t)\right\}\\&\quad\times\left[e_{a,2}(t)-e_{a,1}(t)\right].\end{aligned}\tag{6}$$

We note that the update in (6) is similar to the update in (2) without the extra $[\lambda(t)(1-\lambda(t))]$ multiplier in (2). In [1], it is pointed out that the update in (2) may slow down when $\lambda(t)$ is too close to 0 or 1 due to $[\lambda(t)(1-\lambda(t))]$. As a possible remedy to this problem, $\lambda(t)$ is restricted to an interval excluding 0 and 1 [1]. Note that this problem is not present in (6).

Steady-State Behavior of $\lambda_\rho(t)$: The derivations follow as in [1]. Here, we first obtain an expression for the adaptation parameter in the steady state. If $\bar{\lambda}_\rho(t)\triangleq E[\lambda_\rho(t)]$, then, as $t\to\infty$, we get

$$E\left[\rho(t+1)\right]=E\left[\rho(t)\right]+\mu_\rho\left(1-\bar{\lambda}_\rho(t)\right)\Delta J_2(t)-\mu_\rho\bar{\lambda}_\rho(t)\Delta J_1(t)\tag{7}$$

after some algebra, where we assume that $\lambda_\rho(t)$ and $e_{a,i}(t)$ are independent in the steady state for $i=1,2$ [1]. Furthermore, under the assumption of zero variance for $\lambda_\rho(t)$ as $t\to\infty$ [1], we get

$$J_{\rm ex}=\bar{\lambda}_\rho^2J_{{\rm ex},1}+(1-\bar{\lambda}_\rho)^2J_{{\rm ex},2}+2\bar{\lambda}_\rho(1-\bar{\lambda}_\rho)J_{{\rm ex},12}\tag{8}$$

where $\bar{\lambda}_\rho\triangleq\lim_{t-\infty}E[\lambda_\rho(t)]$. Depending on variances and cross correlation of the *a priori* errors, we have two cases:

1) $J_{{\rm ex},1}\le J_{{\rm ex},12}<J_{{\rm ex},2}$: Here, we have $\Delta J_1\le 0$ and $\Delta J_2>0$ so that the term $(1-\bar{\lambda}_\rho(t))\Delta J_2(t)-\bar{\lambda}_\rho(t)\Delta J_1(t)$ is positive since $1>\bar{\lambda}_\rho(t)>0$ for all t. Then, we get $E[\rho(t)]\to\infty$ as $t\to\infty$. This implies that $\rho(t)\to\infty$ and $\lambda_\rho(t)\to 1$ almost surely as $t\to\infty$ so that $J_{\rm ex}=J_{{\rm ex},1}$. That is, in this case, the combination performs as well as the best component filter. In addition, since we have $\lim_{t-\infty}E[\boldsymbol{w}_\rho(t)]=[1\;0]^T$, we conclude that the combination vector $\boldsymbol{w}_\rho(t)$ converges to the optimum weight vector $\boldsymbol{w}_{\rm o,c}$ under the convex constraint.
2) $J_{{\rm ex},12}<J_{{\rm ex},1}<J_{{\rm ex},2}$: We have $\Delta J_i>0$, $i=1,2$. As $t\to\infty$, a stationary point of (7) may be characterized by

$$\left(1-\bar{\lambda}_\rho(t)\right)\Delta J_2(t)=\bar{\lambda}_\rho(t)\Delta J_1(t)$$

so that $\bar{\lambda}_\rho=\frac{\Delta J_2}{\Delta J_1+\Delta J_2}$. If we substitute $\bar{\lambda}_\rho$ in (8), then we get $J_{\rm ex}=J_{{\rm ex},12}+\frac{\Delta J_1\Delta J_2}{\Delta J_1+\Delta J_2}$, after some algebra. Using $0<\frac{\Delta J_i}{\Delta J_1+\Delta J_2}<1$ yields $J_{\rm ex}<\min\{J_{{\rm ex},1},J_{{\rm ex},2}\}$. Thus, the combination filter outperforms both of the constituent filters. In addition, since we have

$$\lim_{t\to\infty}E\left[\boldsymbol{w}_\rho(t)\right]=\left[\frac{\Delta J_2}{\Delta J_1+\Delta J_2}\quad\frac{\Delta J_1}{\Delta J_1+\Delta J_2}\right]^T$$

the combination weight $\boldsymbol{w}_\rho(t)$ converges to the optimal weight vector $\boldsymbol{w}_{\rm o,c}$ under the convex constraint.

Hence, the combination filter is universal with respect to the constituent filters and its weight vector converges to its optimal value.

C. Algorithm 3

The third update uses a performance-based mixture of the component filters and has deep roots in computational learning theory [6], [7]. Here, the combination weights are selected as certain functions of the accumulated loss of each constituent filter as (9), shown at the bottom of the page, where $0<a\le 1$. After some algebra, the same update on $\lambda_\epsilon(t)$ can be written as

$$\lambda_\epsilon(t)=\frac{1}{1+\exp\left[-\epsilon(t)\right]}\tag{10}$$

with

$$\epsilon(t+1)=a\epsilon(t)+\mu_\epsilon\left(e_{a,2}(t)-e_{a,1}(t)\right)\left(e_1(t)+e_2(t)\right).\tag{11}$$

Steady-State Behavior of $\lambda_\epsilon(t)$: To obtain an expression for the adaptation parameter in the steady state, we use

$$\begin{aligned}E[\epsilon(t+1)]&=aE[\epsilon(t)]+\mu_\epsilon E[(e_{a,2}(t)-e_{a,1}(t))(e_1(t)+e_2(t))]\\&=aE[\epsilon(t)]+\mu_\epsilon\left(J_{{\rm ex},2}(t)-J_{{\rm ex},1}(t)\right)\end{aligned}\tag{12}$$

where we assume that $e_{a,i}(t)$ and $n(t)$ are independent for $i=1,2$ [12]. Along with the configuration of EMSEs, we need to consider also $0<a<1$ and $a=1$ cases separately.

a) $0<a<1$: For convergence of (12), if $d(t)\triangleq\sum_{i=0}^t a^{t-i}(J_{{\rm ex},2}(i)-J_{{\rm ex},1}(i))-\sum_{i=0}^t a^{t-i}(J_{{\rm ex},2}-J_{{\rm ex},1})$, then we recognize that $d(t+1)=ad(t)+b(t+1)-b$, so $|d(t+1)|\le a|d(t)|+|(J_{{\rm ex},2}(t)-J_{{\rm ex},1}(t))-(J_{{\rm ex},2}-J_{{\rm ex},1})|$ by the triangular inequality where $b(t)\triangleq J_{{\rm ex},2}(t)-J_{{\rm ex},1}(t)$ and $b\triangleq J_{{\rm ex},2}-J_{{\rm ex},1}$. Then, it can be easily shown that $d(t)\to 0$ as $t\to\infty$ so that

$$\lim_{t\to\infty}E\left[\epsilon(t)\right]=\frac{\mu_\epsilon(J_{{\rm ex},2}-J_{{\rm ex},1})}{1-a}.\tag{13}$$

The final EMSE of the combination filter is $J_{\rm ex}=\bar{\lambda}_\epsilon^2J_{{\rm ex},1}+(1-\bar{\lambda}_\epsilon)^2J_{{\rm ex},2}+2(1-\bar{\lambda}_\epsilon)\bar{\lambda}_\epsilon J_{{\rm ex},12}$, under the assumption of zero variance for $\lambda_\epsilon(t)$ as $t\to\infty$ [1] for any given a where $\bar{\lambda}_\epsilon\triangleq\lim_{t-\infty}E[\lambda_\epsilon(t)]$. Note that (13) does not depend on $J_{{\rm ex},12}$. Depending on the variances and the cross-EMSE of the *a priori* errors, there are two subcases:

a.1) $J_{{\rm ex},1}\le J_{{\rm ex},12}<J_{{\rm ex},2}$: Under this configuration, the optimal combination parameter λ in (1) is equal to 1 and the EMSE of the optimal combination filter is $J_{{\rm ex},1}$. Hence, for the combination filter to achieve the performance of the best constituent filter, we need to have $\lambda_\epsilon=1$, i.e., $E[\epsilon(t)]\to\infty$ as $t\to\infty$, which is true if and only if $a=1$. For any a, the

$$\lambda_\epsilon(t)=\frac{\exp\left\{-\mu_\epsilon\sum_{i=1}^{t-1}\left[a^{(t-1-i)}e_1^2(i)\right]\right\}}{\exp\left\{-\mu_\epsilon\sum_{i=1}^{t-1}\left[a^{(t-1-i)}e_1^2(i)\right]\right\}+\exp\left\{-\mu_\epsilon\sum_{i=1}^{t-1}\left[a^{(t-1-i)}e_2^2(i)\right]\right\}},\tag{9}$$

difference between the EMSEs of the combination filter and the best constituent filter is

$$f(\bar{\lambda}_\epsilon) \triangleq J_{\rm ex} - J_{\rm ex,1} = (1-\bar{\lambda}_\epsilon)\left[(1+\bar{\lambda}_\epsilon)(J_{\rm ex,12}-J_{\rm ex,1}) + (1-\bar{\lambda}_\epsilon)(J_{\rm ex,2}-J_{\rm ex,12})\right] \geq 0 \quad (14)$$

where the equality is reached if and only if $a = 1$ so that the update (11) does not achieve the performance of the best constituent filter if $a \neq 1$.

a.2) $J_{\rm ex,12} < J_{\rm ex,1} < J_{\rm ex,2}$: Here, the difference between the EMSEs of the combination filter and the best constituent filter is, i.e., $f(\bar{\lambda}_\epsilon)$ in (14), a convex function of $\bar{\lambda}_\epsilon$ with roots $\frac{\Delta J_2 - \Delta J_1}{\Delta J_2 + \Delta J_1}$ and 1. Hence, for $\bar{\lambda}_\epsilon \in \left(\frac{\Delta J_2-\Delta J_1}{\Delta J_2+\Delta J_1}, 1\right)$, $f(\cdot)$ is negative, i.e., $J_{\rm ex} < J_{\rm ex,1}$. We note that $\bar{\lambda}_\epsilon \in \left(\frac{\Delta J_2-\Delta J_1}{\Delta J_2+\Delta J_1}, 1\right)$ if and only if $a \in \left(1+\mu_\epsilon \frac{\Delta J_2-\Delta J_1}{\ln\left(\frac{2\Delta J_1}{\Delta J_2-\Delta J_1}\right)}, 1\right)$ assuming that $\mu_\epsilon \frac{\Delta J_2-\Delta J_1}{\ln\left(\frac{2\Delta J_1}{\Delta J_2-\Delta J_1}\right)} < 0$. Then, the combination filter outperforms the constituent filters for any $a \in \left(1+\mu_\epsilon \frac{\Delta J_2-\Delta J_1}{\ln\left(\frac{2\Delta J_1}{\Delta J_2-\Delta J_1}\right)}, 1\right)$.

b) $a = 1$: We have $E[\epsilon(t+1)] = E[\epsilon(t)] + K(t)$, where $K(t) \triangleq \mu_\epsilon(J_{\rm ex,2}(t) - J_{\rm ex,1}(t))$ converges to a positive constant since $J_{\rm ex,1} < J_{\rm ex,2}$ so that $E[\epsilon(t)] \to \infty$ as $t \to \infty$. This implies that $\epsilon(t) \to \infty$ and $\lambda_\epsilon(t) \to 1$ almost surely as $t \to \infty$ so that $J_{\rm ex} = J_{\rm ex,1}$. Thus, the combination filter performs as well as the best component filter. The final combination weight vector is $\lim_{t\to\infty} E[\boldsymbol{w}_\epsilon(t)] = [1\ 0]^T$.

Hence, $a = 1$ is a necessary condition for the combination filter to achieve the performance of the best constituent filter. Note that when $a \neq 1$, the combination filter may outperform the constituent filters in specific configurations of EMSEs. However, if the cross EMSE is $J_{\rm ex,12} > J_{\rm ex,1}$ and $a \neq 1$, then the combination performs worse than the best constituent filter. Hence, unlike [1], the algorithm of [3] achieves (but not outperforms) the best constituent filter when $a = 1$ and if $a \neq 1$, then the algorithm may outperform or perform worse than the best constituent filter depending on the configuration of EMSEs. Moreover, the weight vector convergence does not appear.

D. Algorithm 4

The fourth update we investigate is studied in [4]. Here, the combination weight is given by

$$\lambda_\gamma(t) = \frac{\left[\sum_{n=0}^{M-1} e_1^2(t-n)\right]^{-\frac{M}{2}}}{\left[\sum_{n=0}^{M-1} e_1^2(t-n)\right]^{-\frac{M}{2}} + \left[\sum_{n=0}^{M-1} e_2^2(t-n)\right]^{-\frac{M}{2}}} \quad (15)$$

where M is the time window to evaluate the performance-based weighting. The same update on $\lambda_\gamma(t)$ can be written as

$$\lambda_\gamma(t) = \frac{1}{1+\exp[-\gamma(t)]}$$

where

$$\gamma(t) \triangleq \frac{M}{2} \ln\left[\frac{\sum_{n=0}^{M-1} e_2^2(t-n)}{\sum_{n=0}^{M-1} e_1^2(t-n)}\right]. \quad (16)$$

Steady-State Behavior of $\lambda_\gamma(t)$: To get the steady-state behavior, we use

$$E[\lambda_\gamma(t)] \approx m \frac{1}{1+E\left[\frac{\sum_{n=0}^{M-1} e_2^2(t-n)}{\sum_{n=0}^{M-1} e_1^2(t-n)}\right]^{-\frac{M}{2}}} \approx \frac{1}{1+\left[\frac{\sum_{n=0}^{M-1} E[e_2^2(t-n)]}{\sum_{n=0}^{M-1} E[e_1^2(t-n)]}\right]^{-\frac{M}{2}}}. \quad (17)$$

We emphasize that although the approximations in (17) are strong especially for small M, we observe a close agreement with our simulations for relatively large M, e.g., $M > 30$. Since as $t \to \infty$, $E[e_i^2(t)] \to J_{ex} i + \sigma_n^2$ for $i = 1, 2$, we get

$$\bar{\lambda}_\gamma \triangleq \lim_{t\to\infty} E[\lambda_\gamma(t)] = \frac{1}{1+\left[\frac{J_{\rm ex,2}+\sigma_n^2}{J_{\rm ex,1}+\sigma_n^2}\right]^{-\frac{M}{2}}} \quad (18)$$

and the final EMSE of the combination filter is $J_{\rm ex} = \bar{\lambda}_\gamma^2 J_{\rm ex,1} + (1-\bar{\lambda}_\gamma)^2 J_{\rm ex,2} + 2(1-\bar{\lambda}_\gamma)\bar{\lambda}_\gamma J_{\rm ex,12}$ for any given M under the assumption of zero variance for $\lambda_\gamma(t)$ in the steady state [1]. Depending on M, we have two cases:

a) $M \to \infty$: Since we have $\frac{(J_{\rm ex,2}+\sigma_n^2)}{(J_{\rm ex,1}+\sigma_n^2)} > 1$, we get $\lim_{t\to\infty} E[\lambda_\gamma(t)] = 1$. Hence,

$$J_{\rm ex} = J_{\rm ex,1}. \quad (19)$$

Thus, the combination filter performs as well as the best constituent filter. The final combination weight vector is $\lim_{t\to\infty} E[\boldsymbol{w}_\gamma(t)] = [1\ 0]^T$.

b) $M < \infty$: Depending on the *a priori* errors and the cross-EMSE between the component filters, there are two subcases:

b.1) $J_{\rm ex,1} \leq J_{\rm ex,12} < J_{\rm ex,2}$: In this case, the optimal combination parameter λ in (1) is 1 and the EMSE of the optimal combination filter is $J_{\rm ex,1}$. The combination filter achieves the performance of the best constituent filter if $E[\gamma(t)] \to \infty$ as $t \to \infty$ if and only if $M \to \infty$. The difference between the EMSEs of the combination filter and the best constituent filter is

$$J_{\rm ex} - J_{\rm ex,1} = (1-\bar{\lambda}_\gamma)\left[(1+\bar{\lambda}_\gamma)(J_{\rm ex,12}-J_{\rm ex,1}) + (1-\bar{\lambda}_\gamma)(J_{\rm ex,2}-J_{\rm ex,12})\right] \geq 0$$

where the equality is reached if and only if $M \to \infty$ so that the algorithm does not achieve the performance of the best constituent filter if $M < \infty$.

b.2) $J_{\rm ex,12} < J_{\rm ex,1} < J_{\rm ex,2}$: In this case, the difference between the EMSEs of the combination filter and the best constituent filter, i.e., $f(\bar{\lambda}_\gamma)$ in (14), is negative for $M \in \left(2\frac{\log(2\Delta J_1)-\log(\Delta J_2-\Delta J_1)}{\log\left(J_{\rm ex,1}+\sigma_n^2\right)-\log\left(J_{\rm ex,2}+\sigma_n^2\right)}, \infty\right)$ so that the combination filter outperforms the constituent filters, i.e., $J_{\rm ex} < J_{\rm ex,1}$.

Hence, $M \to \infty$ is a necessary condition for the combination filter to perform as well as the best constituent filter. The combination filter using the update rule (15) with $M < \infty$ may outperform the constituent filters in certain configurations of the EMSEs. However, if the cross EMSE is sufficiently large, then the combination filter performs worse than the best component filter when $M < \infty$. Hence, unlike [1], update (15) achieves (but not outperforms) the best constituent filter when $M \to \infty$ and if $M < \infty$, then the algorithm may outperform

or perform worse than the best constituent filter depending on the configuration of EMSEs. Moreover, it does not offer the desirable weight vector convergence.

IV. Transient Analysis of the Convexly Constrained Mixtures

In this section, we perform mean and mean-square convergence analysis of the studied algorithms. The derivations follow [13]. We use the following assumptions [13]:

- A1) $n(t)$ is independent of $\boldsymbol{u}(t)$.
- A2) $\rho(t), \epsilon(t), \gamma(t)$ vary slowly enough so that $E[e^k_{a,i}(t)e^l_{a,j}(t)h(t)|h(t)] = E[e^k_{a,i}(t)e^l_{a,j}(t)]h(t)$, where $h(t) \in \{\rho(t), \epsilon(t), \gamma(t)\}$, $i, j = 1, 2$ and $k, l = 0, \ldots, 4$, $k + l \leq 4$.
- A3) $e_{a,1}(t)$ and $e_{a,2}(t)$ are jointly Gaussian and zero mean, implying [13] $E[e^4_{a,i}(t)] = 3J^2_{\mathrm{ex},i}(t)$, $i = 1, 2$,, $E[e^k_{a,1}(t)e^l_{a,2}(t)] = 0$, $k + l = 3$, $E[e^k_{a,1}(t)e^l_{a,2}(t)] = 3J_{\mathrm{ex},1}(t)J_{\mathrm{ex},12}(t)$, $k = 3$, $l = 1$, $E[e^k_{a,1}(t)e^l_{a,2}(t)] = 3J_{\mathrm{ex},12}(t)J_{\mathrm{ex},2}(t)$, $k = 1$, $l = 3$, $E[e^k_{a,1}(t)e^l_{a,2}(t)] = 2J^2_{\mathrm{ex},12}(t) + J_{\mathrm{ex},1}(t)J_{\mathrm{ex},2}(t)$, $k = l = 2$.

A. Transient Analysis of Algorithm 2

The update (6) can be written as

$$\begin{aligned} e\rho(t+1) = \rho(t) + \mu_\rho \Big[& -\lambda_\rho(t)e^2_{a,1}(t) + (1-\lambda_\rho(t))\, e^2_{a,2}(t) \\ & + (2\lambda_\rho(t)-1)e_{a,1}(t)e_{a,2}(t) \\ & + n(t)\,(e_{a,2}(t) - e_{a,1}(t))\Big]. \end{aligned} \tag{20}$$

The first order Taylor's approximation of $\lambda_\rho(\rho(t)) \triangleq \frac{1}{(1+\exp(-\rho(t)))}$ around $\bar{\rho}(t) \triangleq E[\rho(t)]$ is given by

$$\begin{aligned} \lambda_\rho\left(\rho(t)\right) &\approx \lambda_\rho\left(\bar{\rho}(t)\right) + \frac{d\lambda_\rho}{d\rho(t)}\left(\bar{\rho}(t)\right)\left(\rho(t) - \bar{\rho}(t)\right) \\ &= \bar{\lambda}_\rho(t) + \bar{\lambda}_\rho(t)\left(1-\bar{\lambda}_\rho(t)\right)\left(\rho(t) - \bar{\rho}(t)\right) \end{aligned} \tag{21}$$

where $\bar{\lambda}_\rho(t) \triangleq \lambda_\rho(\bar{\rho}(t))$. Using (21) in (20) yields

$$\begin{aligned} \rho(t+1) = \rho(t) + \mu_\rho \Big[& -\left(\bar{\lambda}_\rho(t) + \bar{\lambda}_\rho(t)\left(1-\bar{\lambda}_\rho(t)\right)\left(\rho(t)-\bar{\rho}(t)\right)\right)e^2_{a,1}(t) \\ & + \left(1-\bar{\lambda}_\rho(t) - \bar{\lambda}_\rho(t)\left(1-\bar{\lambda}_\rho(t)\right)\left(\rho(t)-\bar{\rho}(t)\right)\right)e^2_{a,2}(t) \\ & + \Big(2\bar{\lambda}_\rho(t) + 2\bar{\lambda}_\rho(t)\left(1-\bar{\lambda}_\rho(t)\right) \\ & \quad \times \left(\rho(t)-\bar{\rho}(t)\right) - 1\Big)\, e_{a,1}(t)e_{a,2}(t) \\ & + n(t)\left(e_{a,2}(t) - e_{a,1}(t)\right)\Big]. \end{aligned} \tag{22}$$

Taking the expectation of (22) and using A1), A2) yields

$$\begin{aligned} \bar{\rho}(t+1) = \bar{\rho}(t) + \mu_\rho \Big[& -\bar{\lambda}_\rho(t)J_{\mathrm{ex},1}(t) + \left(1-\bar{\lambda}_\rho(t)\right)J_{\mathrm{ex},2}(t) \\ & + \left(2\bar{\lambda}_\rho(t) - 1\right)J_{\mathrm{ex},12}(t)\Big]. \end{aligned} \tag{23}$$

Moreover, by using (21) in $e_a(t) = \lambda_\rho(t)e_{a,1}(t) + (1-\lambda_\rho(t))e_{a,2}(t)$, we get

$$\begin{aligned} e_a(t) = \left(\bar{\lambda}_\rho(t) + \bar{\lambda}_\rho(t)\left(1-\bar{\lambda}_\rho(t)\right)\left(\rho(t)-\bar{\rho}(t)\right)\right)\left(e_{a,1}(t)\right. \\ \left. - e_{a,2}(t)\right) + e_{a,2}(t) \end{aligned} \tag{24}$$

which yields $E[e_a(t)] = 0$ using A1) and A2). We next find the EMSE of the combination filter by squaring (24) and taking the expectation, yielding

$$\begin{aligned} E\left[e^2_a(t)\right] = & \left[\bar{\lambda}^2_\rho(t) + \sigma^2_\rho(t)\bar{\lambda}^2_\rho(t)\left(1-\bar{\lambda}_\rho(t)\right)^2\right] \\ & \times \left[J_{\mathrm{ex},1}(t) + J_{\mathrm{ex},2}(t) - 2J_{\mathrm{ex},12}(t)\right] \\ & + 2\bar{\lambda}_\rho(t)\left(J_{\mathrm{ex},12}(t) - J_{\mathrm{ex},2}(t)\right) + J_{\mathrm{ex},2}(t) \end{aligned} \tag{25}$$

where $\sigma^2_\rho(t) \triangleq E[(\rho(t) - \bar{\rho}(t))^2]$ with A1) and A2). To evaluate (25), we need have $\sigma^2_\rho(t)$. To obtain a recursion for $\sigma^2_\rho(t)$, we square (22), take the expected value and subtract the square of (23), yielding, using A1), A2), and A3), after straightforward algebra,

$$\sigma^2_\rho(t+1) = \left(1 + 2\mu_\rho G_1(t) + \mu^2_\rho G_2(t)\right)\sigma^2_\rho(t) + \mu^2_\rho F(t) \tag{26}$$

where, omitting t,

$$\begin{aligned} F = & 2(1-\bar{\lambda}_\rho)^2 J^2_{\mathrm{ex},2} + (2\bar{\lambda}_\rho - 1)^2\left[J^2_{\mathrm{ex},12} + J_{\mathrm{ex},1}J_{\mathrm{ex},2}\right] \\ & + 2\bar{\lambda}^2_\rho J^2_{\mathrm{ex},1} + 4(2\bar{\lambda}_\rho - 1)(1-\bar{\lambda}_\rho)J_{\mathrm{ex},12}J_{\mathrm{ex},2} \\ & - 4\bar{\lambda}_\rho(1-\bar{\lambda}_\rho)J^2_{\mathrm{ex},12} - 4\bar{\lambda}_\rho(2\bar{\lambda}_\rho - 1)J_{\mathrm{ex},1}J_{\mathrm{ex},12} \\ & + (J_{\mathrm{ex},1} + J_{\mathrm{ex},2} - 2J_{\mathrm{ex},12})\sigma^2_n \end{aligned} \tag{27}$$

$$G_1 = -\bar{\lambda}_\rho(1-\bar{\lambda}_\rho)[J_{\mathrm{ex},1} + J_{\mathrm{ex},2} - 2J_{\mathrm{ex},12}] \tag{28}$$

$$\begin{aligned} G_2 = & 3\bar{\lambda}^2_\rho(1-\bar{\lambda}_\rho)^2\Big[J^2_{\mathrm{ex},1} + 2(J_{\mathrm{ex},1}J_{\mathrm{ex},2} + 2J^2_{\mathrm{ex},12}) \\ & - 4J_{\mathrm{ex},1}J_{\mathrm{ex},12} - 4J_{\mathrm{ex},12}J_{\mathrm{ex},2} + J^2_{\mathrm{ex},2}\Big]. \end{aligned} \tag{29}$$

Here, we analyze the bias/variance relation of Algorithm 2. From (23), when the step size is large, the combination filter could better track the constituent filters. However, a larger step size may cause $\sigma^2_\rho(t)$ to be large so that the EMSE of the combination filter (25) may become unstable during the initial iterations. Note that from (26) when $\bar{\lambda}_\rho \triangleq \lim_{t\to\infty}\bar{\lambda}_\rho(t) = 0$ or 1, $\sigma^2_\rho(t)$ is unbounded as $t \to \infty$ since $G_1 \triangleq \lim_{t\to\infty} G_1(t) = 0$, $G_2 \triangleq \lim_{t\to\infty} G_2(t) = 0$ and $F \triangleq \lim_{t\to\infty} F(t) > 0$. However, in our simulations, we observe that $\bar{\lambda}^2_\rho(t)(1-\bar{\lambda}_\rho(t))^2$ converges to 0 faster than $\sigma^2_\rho(t)$ goes to infinity so that the term $\lim_{t\to\infty}\sigma^2_\rho(t)\bar{\lambda}^2_\rho(t)(1-\bar{\lambda}_\rho(t))^2 = 0$ in (25). Hence, the effect of the variance of the combination parameter on the EMSE of the combination filter diminishes in the steady state when $\bar{\lambda}_\rho = 0$ or 1 so that the EMSE of the combination filter converges to the EMSE of the best constituent filter in the mean and the MSE senses. When $\bar{\lambda}_\rho = \frac{\Delta J_2}{\Delta J_1 + \Delta J_2}$, we observe from (26) that $\sigma^2_\rho(t)$ converges when $|1 + 2\,\mu_\rho G_1(t) + \mu^2_\rho G_2(t)| < 1$ for all t, i.e., $-2 < 2\,\mu_\rho G_1(t) + \mu^2_\rho G_2(t) < 0$ and under this condition

$$\sigma^2_\rho \triangleq \lim_{t\to\infty}\sigma^2_\rho(t) = -\frac{\mu_\rho F}{2G_1 + \mu_\rho G_2}.$$

We observe from (27) that $F > 0$ and from (28) and (29) that $2G_1 + \mu_\rho G_2 < 0$ when $-2 < 2\,\mu_\rho G_1(t) + \mu^2_\rho G_2(t) < 0$ for all t so that $\sigma^2_\rho > 0$ and the term $\sigma^2_\rho(t)\bar{\lambda}^2_\rho(t)(1-\bar{\lambda}_\rho(t))^2$ in (25) converges. Hence, from (25), there is a bias term $\sigma^2_\rho\bar{\lambda}^2_\rho(1-\bar{\lambda}_\rho)^2(\Delta J_1 + \Delta J_2)$ in the EMSE of the combination filter in the steady state which introduces a bias/variance tradeoff as in the stochastic gradient algorithms [12], e.g., the tradeoff between the bias and the step size of LMS algorithm. Since all the terms in (27), (28), and (29) can be calculated (recursively), this concludes the transient analysis of Algorithm 2.

B. Transient Analysis of Algorithm 3

The update rule for $\epsilon(t)$ can be written as

$$\epsilon(t+1)=a\epsilon(t)+\mu_\epsilon\left[e_{a,2}^2(t)-e_{a,1}^2(t)+2n(t)(e_{a,2}(t)-e_{a,1}(t))\right] \tag{30}$$

yielding

$$\bar{\epsilon}(t+1)=a\bar{\epsilon}(t)+\mu_\epsilon\left[J_{\mathrm{ex},2}(t)-J_{\mathrm{ex},1}(t)\right] \tag{31}$$

with A1). We next use the first order Taylor's approximation of $\lambda_\epsilon(t)$ around the expected value $\bar{\epsilon}(t)\triangleq E[\epsilon(t)]$ as $\lambda_\epsilon(t)\approx\bar{\lambda}_\epsilon(t)+\bar{\lambda}_\epsilon(t)(1-\bar{\lambda}_\epsilon(t))(\epsilon(t)-\bar{\epsilon}(t))$, where $\bar{\lambda}_\epsilon(t)\triangleq\lambda(\bar{\epsilon}(t))$. Applying this to $e_a(t)=\lambda_\epsilon(t)e_{a,1}(t)+(1-\lambda_\epsilon(t))e_{a,2}(t)$ yields

$$e_a(t)=\left[\bar{\lambda}_\epsilon(t)+\bar{\lambda}_\epsilon(t)\left(1-\bar{\lambda}_\epsilon(t)\left(\epsilon(t)-\bar{\epsilon}(t)\right)\right)\right]\left[e_{a,1}(t)-e_{a,2}(t)\right]+e_{a,2}(t) \tag{32}$$

and $E[e_a(t)]=0$ with A1) and A2). We obtain EMSE of the combination filter by squaring (32) and taking the expectation as

$$\begin{aligned}E\left[e_a^2(t)\right]=&\left[\bar{\lambda}_\epsilon^2(t)+\sigma_\epsilon^2(t)\bar{\lambda}_\epsilon^2(t)\left(1-\bar{\lambda}_\epsilon(t)\right)^2\right]\\&\times\left[J_{\mathrm{ex},1}(t)+J_{\mathrm{ex},2}(t)-2J_{\mathrm{ex},12}(t)\right]\\&+2\bar{\lambda}_\epsilon(t)\left(J_{\mathrm{ex},12}(t)-J_{\mathrm{ex},2}(t)\right)+J_{\mathrm{ex},2}(t)\end{aligned} \tag{33}$$

where $\sigma_\epsilon^2(t)\triangleq E[(\epsilon(t)-\bar{\epsilon}(t))^2]$ is the variance of $\epsilon(t)$ using A1) and A2). To obtain a recursion for $\sigma_\epsilon^2(t)$, we square (30), take expectation and then subtract the square of (31). This yields, using A1), A2), and A3),

$$\begin{aligned}\sigma_\epsilon^2(t+1)=a^2\sigma_\epsilon^2(t)+2\mu_\epsilon^2\big[&J_{\mathrm{ex},1}^2(t)+J_{\mathrm{ex},2}^2(t)-2J_{\mathrm{ex},12}^2(t)\\&+2\sigma_n^2\left(J_{\mathrm{ex},1}(t)+J_{\mathrm{ex},2}(t)-2J_{\mathrm{ex},12}(t)\right)\big].\end{aligned} \tag{34}$$

Here, we analyze the bias/variance relation of Algorithm 3. From (31), the combination filter could better track the constituent filters when the step size is large. However, a larger step size may cause $\sigma_\epsilon^2(t)$ to be large so that the EMSE of the combination filter (33) may become unstable during the initial iterations. When $0<a<1$, we have $\bar{\epsilon}\triangleq\lim_{t\to\infty}\bar{\epsilon}(t)=\frac{\mu_\epsilon(J_{\mathrm{ex},2}-J_{\mathrm{ex},1})}{1-a}$ and $\bar{\lambda}_\epsilon\triangleq\lim_{t\to\infty}\bar{\lambda}_\epsilon(t)$. From (34), $\sigma_\epsilon^2(t)$ converges and

$$\begin{aligned}\sigma_\epsilon^2&\triangleq\lim_{t\to\infty}\sigma_\epsilon^2(t)\\&=\frac{2\mu_\epsilon^2\left[J_{\mathrm{ex},1}^2+J_{\mathrm{ex},2}^2-2J_{\mathrm{ex},12}^2+2\sigma_n^2(J_{\mathrm{ex},1}+J_{\mathrm{ex},2}-2J_{\mathrm{ex},12})\right]}{1-a^2}.\end{aligned}$$

Hence, the term $\sigma_\epsilon^2(t)\bar{\lambda}_\epsilon^2(t)(1-\bar{\lambda}_\epsilon(t))^2$ in (33) converges. Note that from (33) this term introduces a bias in the EMSE of the combination filter in the steady state. When $a=1$, it follows from (31) that $\bar{\lambda}_\epsilon=0$ or 1. From (34), $\sigma_\epsilon^2(t)$ is unbounded as $t\to\infty$. However, in our simulations, we observe that $\bar{\lambda}_\epsilon^2(t)(1-\bar{\lambda}_\epsilon(t))^2$ converges to 0 faster than $\sigma_\epsilon^2(t)$ goes to infinity so that the term $\lim_{t\to\infty}\sigma_\epsilon^2(t)\bar{\lambda}_\epsilon^2(t)(1-\bar{\lambda}_\epsilon(t))^2=0$ in (33). Hence, the effect of the variance of the combination parameter on the EMSE of the combination filter diminishes in the steady state when $a=1$ so that the EMSE of the combination filter converges to the EMSE of the best constituent filter in the mean and the MSE senses. This concludes the transient analysis of Algorithm 3.

C. Transient Analysis of Algorithm 4

Taking expectation of (16) yields

$$\begin{aligned}\bar{\gamma}(t)&=\frac{M}{2}E\left[\ln\left(\frac{\sum_{n=0}^{M-1}e_2^2(t-n)}{\sum_{n=0}^{M-1}e_1^2(t-n)}\right)\right]\\&\approx\frac{M}{2}\ln\left(\frac{\sum_{n=0}^{M-1}J_{\mathrm{ex},2}(t-n)}{\sum_{n=0}^{M-1}J_{\mathrm{ex},1}^2(t-n)}\right).\end{aligned}$$

If we use the first order Taylor's approximation of $\lambda_\gamma(t)$ around the expected value $\bar{\gamma}(t)\triangleq E[\gamma(t)]$, then we get

$$\lambda_\gamma(t)\approx\bar{\lambda}_\gamma(t)+\bar{\lambda}_\gamma(t)\left(1-\bar{\lambda}_\gamma(t)\right)\left(\gamma(t)-\bar{\gamma}(t)\right) \tag{35}$$

where $\bar{\lambda}_\gamma(t)\triangleq\lambda(\bar{\gamma}(t))$. Using (35) in $e_a(t)$ yields

$$\begin{aligned}e_a(t)=&\left[\bar{\lambda}_\gamma(t)+\bar{\lambda}_\gamma(t)\left(1-\bar{\lambda}_\gamma(t)\left(\gamma(t)-\bar{\gamma}(t)\right)\right)\right]\\&\times\left[e_{a,1}(t)-e_{a,2}(t)\right]+e_{a,2}(t)\end{aligned} \tag{36}$$

and $E[e_a(t)]=0$ under A1) and A2). To get the EMSE of the combination filter, we first use the first-order Taylor's approximation of $\lambda_\gamma^2(t)$ around the expected value $\bar{\gamma}(t)\triangleq E[\gamma(t)]$ to get

$$\lambda_\gamma^2(t)\triangleq\bar{\lambda}_\gamma^2(t)+2\bar{\lambda}_\gamma^2(t)(1-\bar{\lambda}_\gamma(t))(\gamma(t)-\bar{\gamma}(t)). \tag{37}$$

Using (35) and (37) in $e_a^2(t)$ and taking expectation yields

$$\begin{aligned}E\left[e_a^2(t)\right]=&\bar{\lambda}_\gamma^2(t)J_{\mathrm{ex},1}(t)+\left(1-\bar{\lambda}_\gamma^2(t)\right)^2J_{\mathrm{ex},2}(t)\\&+2\bar{\lambda}_\gamma^2(t)\left(1-\bar{\lambda}_\gamma^2(t)\right)J_{\mathrm{ex},12}(t).\end{aligned} \tag{38}$$

This concludes the transient analysis of Algorithm 4.

V. Simulations

In this section, we present performance of the combination algorithms through simulations using the setup of [1]. Here, we have two LMS filters with the same input regressor and different step sizes running in parallel as the constituent filters with updates $\boldsymbol{w}_i(t+1)=\boldsymbol{w}_i(t)+\mu_i e_i(t)\boldsymbol{u}(t)$, for $i=1,2$, where $\mu_1=0.1$ and $\mu_2=0.001$. The input regressor $\boldsymbol{u}(t)\in\mathbb{R}^7$ is zero mean and i.i.d. Gaussian with variance selected to yield $\mathrm{Tr}(\boldsymbol{R})=1$, where $\mathrm{Tr}(\cdot)$ is the trace. The underlying signal is generated as $d(t)=\boldsymbol{w}_o^T(t)\boldsymbol{u}(t)+n(t)$, where $n(t)$ is the additive i.i.d Gaussian noise with variance $\sigma_n^2=0.01$ and $\boldsymbol{w}_o(t+1)=\boldsymbol{w}_o(t)+\boldsymbol{q}(t)$. The initial value of $\boldsymbol{w}_o(t)$ is selected as $\boldsymbol{w}_o(0)=[0.90,-0.53,0.21,-0.028,0.78,0.52,-0.08]^T$ [1]. Theoretical EMSEs of the combination filters and the cross-EMSE between them are given by $J_{\mathrm{ex},i}=\frac{\mu_i\sigma_n^2\mathrm{Tr}(\boldsymbol{R})+\mu_i^{-1}\mathrm{Tr}(\boldsymbol{Q})}{2-\mu_i\mathrm{Tr}(\boldsymbol{R})}$, $J_{\mathrm{ex},12}=\frac{\mu_{12}\sigma_n^2\mathrm{Tr}(\boldsymbol{R})+2\frac{\mathrm{Tr}(\boldsymbol{Q})}{\mu_1+\mu_2}}{2-\mu_{12}\mathrm{Tr}(\boldsymbol{R})}$ under the separation assumption [1], where $\mu_{12}=\frac{2\mu_1\mu_2}{\mu_1+\mu_2}$ and theoretical $J_{\mathrm{ex},i}$ attains the minimum at $\mu_{\mathrm{opt}}=\sqrt{\frac{\mathrm{Tr}(\boldsymbol{Q})}{\sigma_n^2\mathrm{Tr}(\boldsymbol{R})}+\frac{[\mathrm{Tr}(\boldsymbol{Q})]^2}{4\sigma_n^4}}-\frac{\mathrm{Tr}(\boldsymbol{Q})}{2\sigma_n^2}$. We measure the performance using the same figure of merit as in [1]. The normalized square deviation (NSD) of the component filters and the combination filters are defined as $\mathrm{NSD}_i\triangleq\frac{J_{\mathrm{ex},i}}{J_{\mathrm{ex,opt}}}$, $\mathrm{NSD}_{\mathrm{alg2}}\triangleq\frac{J_{\mathrm{ex,alg2}}}{J_{\mathrm{ex,opt}}}$, $\mathrm{NSD}_{\mathrm{alg3}}\triangleq\frac{J_{\mathrm{ex,alg3}}}{J_{\mathrm{ex,opt}}}$, $\mathrm{NSD}_{\mathrm{alg4}}\triangleq\frac{J_{\mathrm{ex,alg4}}}{J_{\mathrm{ex,opt}}}$, where $J_{\mathrm{ex,alg}i}$ is the EMSE of the ith combination filter and $J_{\mathrm{ex,opt}}$ is the EMSE calculated using μ_{opt}.

In Fig. 1, we plot the NSDs for all algorithms as a function of $\mathrm{Tr}(\boldsymbol{Q})$, $\boldsymbol{Q}=E[\boldsymbol{q}(t)\boldsymbol{q}^T(t)]$. For these simulations, the step size in (6) is set to

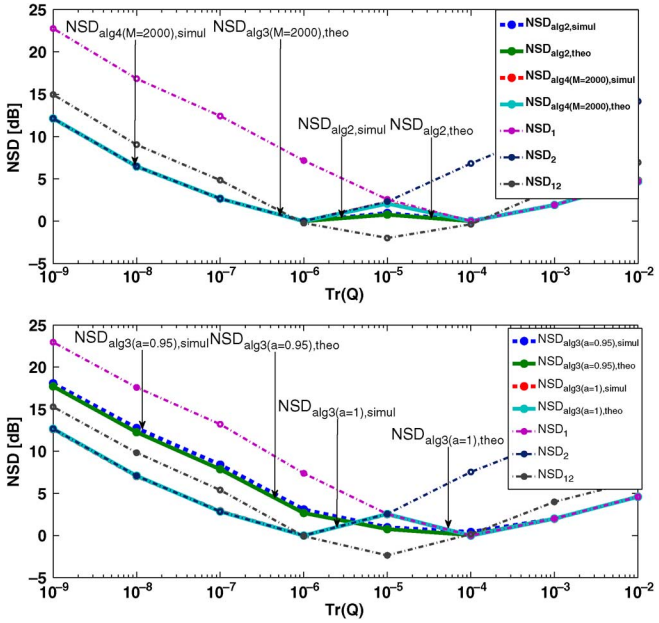

Fig. 1. Theoretical and simulated NSDs as a function of $\mathrm{Tr}(\boldsymbol{Q})$. (a) Second and third combination filters, $\mu_\rho = 30$, $M = 2000$. (b) Fourth combination filter, $\mu_\epsilon = 30$, $a = 1$, $a = 0.95$.

$\mu_\rho = 30$ and the step size in (11) is set to $\mu_\epsilon = 30$ to guarantee convergence. To test our theoretical analysis on the forgetting factor in (11), we simulate (11) using $a = 1$ and $a = 0.95$. We test the update in (15) using a time window $M = 2000$. The simulations are done over 6×10^5 samples, averaged over 20 independent trials. The final EMSEs are calculated by averaging the last 1000 samples of each iteration.

We observe in Fig. 1(a) that the combination filter using the EG update (6) is universal with respect to the combination filters and even performs better than both when $J_{\mathrm{ex},12} < \min\{J_{\mathrm{ex},1}, J_{\mathrm{ex},2}\}$ (as shown in Section III-B). The update (15) with $M = 2000$ achieves the performance of the best constituent filter since $M = 2000$ is sufficiently large to yield (19). Similarly, the update (11) is also universal when $a = 1$ such that it achieves the performance of the best constituent filter for all $\mathrm{Tr}(\boldsymbol{Q})$ in Fig. 1(b). For the update (11) with $a = 0.95$, we observe that for certain $\mathrm{Tr}(\boldsymbol{Q})$, the update performs better than both constituent filters. However, since $a \neq 1$, the update (11) performs worse than the best constituent filter as predicted in (14) and Section III-Ca2) for certain $\mathrm{Tr}(\boldsymbol{Q})$. For all algorithms, we observe that our steady-state analysis accurately describes the simulations.

For the simulations related to the transient analysis, the underlying signal is generated from a stationary model as $d(t) = \boldsymbol{w}_o^T \boldsymbol{u}(t) + n(t)$ [1], where $n(t)$ is the additive i.i.d noise with variance $\sigma_n^2 = 0.01$ and $\boldsymbol{w}_o = [0.24, -0.45, -0.35, 0.04, -0.17, 0.74, 0.14]^T$. Moreover, to test the switching performance, we abruptly change $\boldsymbol{w}_o$ to $\boldsymbol{w}_o = [0.34, 0.45, -0.41, 0.46, 0.14, -0.44, -0.24]^T$ in the middle of the simulations [1]. Here, the input regressor $\boldsymbol{u}(t) \in \mathbb{R}^7$ is zero mean i.i.d. Gaussian, where the variance of each entry is set to 1. As the constituent filters, we have two LMS filters with the same input regressor and different step sizes running in parallel with updates $\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \mu_i e_i(t)\boldsymbol{u}(t)$, for $i = 1, 2$, $\mu_1 = 0.15$, $\mu_2 = 0.002$. For the combination algorithms, we set $\mu_\rho = 1$ for Algorithm 2, $\mu_\epsilon = 1$ and $a = 0.98$ for Algorithm 3, $M = 200$ for the Algorithm 4. Results are averaged over 1000 independent trials. In Fig. 2(a), we plot the MSE curve for Algorithm 2, labeled as "$\mathrm{Alg.2_{simul}}$", the theoretical derived

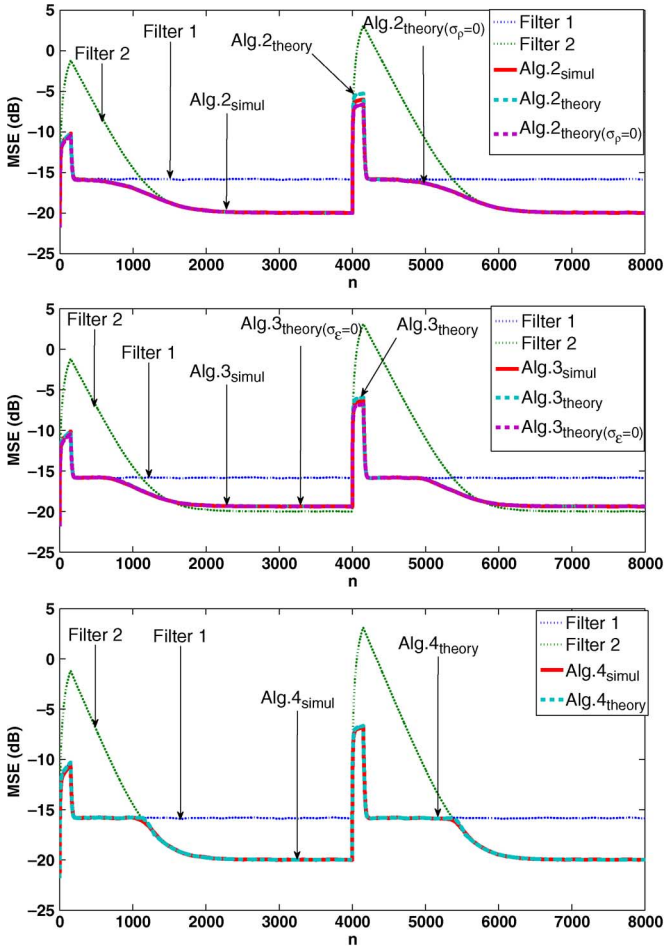

Fig. 2. MSE curves for all algorithms. Labels are described in the text. (a) Algorithm 2, $\mu_\rho = 1$. (b) Algorithm 3, $\mu_\epsilon = 1$ and $a = 0.98$. (c) Algorithm 4, $M = 200$.

MSE curve using (25), labeled as "$\mathrm{Alg.2_{theory}}$". In Fig. 2(a), we also plot the theoretical derived MSE curve, where we set $\sigma_\rho(t) = 0$, labeled as "$\mathrm{Alg.2_{theory(\sigma_\rho=0)}}$", as suggested in [13]. We observe that our analysis closely describes the transient behavior of Algorithm 2 in these simulations. We repeat the same experiment for Algorithm 3 and display the results in Fig. 2(b). We use the same labeling as in Fig. 2(a), however, use (33) to calculate the theoretical curves. We point out that since $a = 0.98$, as predicted from the steady-state analysis, the mixture does not converge to the best constituent filter as seen in Fig. 2(b) (unlike Algorithm 2 in Fig. 2(a)). The same simulations are performed for Algorithm 4 as shown in Fig. 2(c), however, we used (38) to calculate the theoretical curve. We again observe that our transient analysis closely describes the behavior of Algorithm 3 and 4. We observe that $M = 200$ is sufficiently large for these simulations that the mixture converges to the best constituent filter.

Here, we investigate the tradeoff between the transient and steady-state behaviors for the combination algorithms as follows. In this setup, the desired signal is generated as $d(t) = \boldsymbol{w}_o^T \boldsymbol{u}(t) + n(t)$, where $n(t)$ is the additive i.i.d noise with variance $\sigma_n^2 = 0.01$, $\boldsymbol{w}_o = [0.25, -0.47, -0.37, 0.04, -0.18, 0.78, 0.14]^T$ and the input regressor $\boldsymbol{u}(t) \in \mathbb{R}^7$ is zero mean i.i.d. Gaussian, where the variance of each entry is set to 1. As the input filters, there are two LMS filters running in parallel to model $d(t)$ with the same input regressor and the step sizes $\mu_1 = 0.15$, $\mu_2 = 0.002$, respectively. We first fix the

TABLE I
COMPARISON OF COMBINATION ALGORITHMS

μ_α	μ_ρ	Parameters of Alg. 3	Parameter of Alg. 4	n (iteration index)
$\mu_\alpha = 0.5$	$\mu_\rho = 0.5$	a=0.999, $\mu_\epsilon = 0.5$	M=2000	$n_1 = 1864$, $n_2 = 1092$, $n_3 = 3055$, $n_4 = 2595$
$\mu_\alpha = 5$	$\mu_\rho = 5$	a=0.999, $\mu_\epsilon = 5$	M=1700	$n_1 = 1063$, $n_2 = 1009$, $n_3 = 3181$, $n_4 = 2336$
$\mu_\alpha = 100$	$\mu_\rho = 100$	a=0.999, $\mu_\epsilon = 100$	M=1400	$n_1 = 1063$,$n_2 = 1011$, $n_3 = 3193$, $n_4 = 2088$

step size of Algorithm 1, i.e., $\mu_\alpha = 0.5$, and generate the theoretical $\mathrm{MSE}(n)$ curve versus iteration index n. Then, we determine the value of n where $\mathrm{MSE}(n)$ is 3 dB above the minimum MSE and label it n_1. We adjust the step size of Algorithm 2 μ_ρ, the step size μ_ϵ and the forgetting factor a of Algorithm 3 and the time window M of Algorithm 4 such that the final MSE of each algorithm is equal to the final MSE of Algorithm 1. Then, the theoretical $\mathrm{MSE}(n)$ curve versus iteration index n for each algorithm is generated and the values of n where $\mathrm{MSE}(n)$ is 3 dB above the minimum MSE are determined and labeled by n_2, n_3, n_4, respectively. The performance of the combination algorithm with the smallest n_i is the best for this example. We repeat this process for different selections of μ_α including $\mu_\alpha = 5$ and $\mu_\alpha = 100$ and summarize the results in Table I. We observe that in these simulations Algorithm 2 provide a better converge tradeoff.

VI. CONCLUSION

We investigated and compared four convexly constraint adaptive mixture methods to adaptively combine outputs of constituent filters that work in parallel on system modeling. We derive the corresponding MSEs and the converged mixture weights in the steady state under nonstationary random walk model. We also performed the transient analysis in the mean and MSE sense for all algorithms. We observe that these convex mixture methods are universal such that they achieve the performance of the best constituent filter in the steady state. We observe that the EG update (6) under the mixture of experts framework can also outperform the best constituent filter under certain configuration of the EMSEs of the constituent filters (similar to the algorithm from [1]). We also demonstrate that the MSE in the steady state of the algorithms from [3] and [4] heavily depends on the corresponding algorithmic parameters, i.e., the forgetting factor in [3] and the window length in [4]. We observe that our derivations accurately describes the behavior of all algorithms under the setup of [1]. Our main contributions are as follows: 1) we show that the algorithm from [2] is universal and its combined weight vector converges to the optimal convex mixture; 2) we demonstrate that the algorithm from [3] is only universal if the memory constant is unitary (no decay is allowed if universality is required), but the weight vector does not convergence to the optimal convex mixture; and 3) we show that the algorithm from [4] is always universal (but not better than the best filter) only for very long windows, however, does not offer the desirable weight vector convergence.

REFERENCES

[1] J. Arenas-Garcia, A. R. Figueiras-Vidal, and A. H. Sayed, "Mean-square performance of a convex combination of two adaptive filters," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 1078–1090, 2006.

[2] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *J. Inf. Comput.*, vol. 132, pp. 1–64, 1997.

[3] A. C. Singer and M. Feder, "Universal linear prediction by model order weighting," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2685–2699, 1999.

[4] M. Niedzwiecki, "Identification of nonstationary stochastic systems using parallel estimation schemes," *IEEE Trans. Autom. Control*, vol. 35, no. 3, pp. 329–334, 1990.

[5] J. Benesty and Y. Huang, "The LMS, PNLMS, and exponentiated gradient algorithms," in *Proc. EUSIPCO*, 2004, vol. 1, pp. 721–724.

[6] V. Vovk, "A game of prediction with expert advice," *J. Comput. Syst. Sci.*, vol. 56, pp. 153–173, 1998.

[7] N. Cesa-Bianchi and G. Lugosi, "On prediction of individual sequences relative to a set of experts," in *Proc. IEEE Int. Symp. Inf. Theory*, Massachusetts, 1998, pp. 16–21.

[8] J. Arenas-Garcia, V. Gomez-Verdejo, and A. R. Figueiras-Vidal, "New algorithms for improved adaptive convex combination of LMS transversal filters," *IEEE Trans. Instrum. Meas.*, vol. 54, pp. 2239–2249, Dec. 2005.

[9] N. J. Bershad, J. C. M. Bermudez, and J. Tourneret, "An affine combination of two lms adaptive filtersransient mean-square analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1853–1864, 2008.

[10] S. S. Kozat, A. T. Erdogan, A. C. Singer, and A. H. Sayed, "Steady state MSE performance analysis of mixture approaches to adaptive filtering," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4050–4063, Aug. 2010.

[11] S. I. Hill and R. C. Williamson, "Convergence of exponentiated gradient algorithms," *IEEE Trans. Signal Process.*, vol. 49, no. 6, pp. 1208–1215, 2001.

[12] A. H. Sayed, *Fundamentals of Adaptive Filtering*. New York: Wiley, 2003.

[13] R. Candido, M. T. M. Silva, R. Candido, and V. H. Nascimento, "Transient and steady-state analysis of the affine combination of two adaptive filters," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4064–4078, 2010.