

# Adaptive Filtering Algorithms for Set-Valued Observations—Symmetric Measurement Approach to Unlabeled and Anonymized Data

Vikram Krishnamurthy, *Fellow IEEE*, September 1, 2022

**Abstract**—Suppose  $L$  simultaneous independent stochastic systems generate observations, where the observations from each system depend on the underlying parameter of that system. The observations are unlabeled (anonymized), in the sense that an analyst does not know which observation came from which stochastic system. How can the analyst estimate the underlying parameters of the  $L$  systems? Since the anonymized observations at each time are an unordered set of  $L$  measurements (rather than a vector), classical stochastic gradient algorithms cannot be directly used. By using symmetric polynomials, we formulate a symmetric measurement equation that maps the observation set to a unique vector. By exploiting that fact that the algebraic ring of multi-variable polynomials is a unique factorization domain over the ring of one-variable polynomials, we construct an adaptive filtering algorithm that yields a statistically consistent estimate of the underlying parameters. We analyze the asymptotic covariance of these estimates to quantify the effect of anonymization. Finally, we characterize the anonymity of the observations in terms of the error probability of the maximum a posteriori Bayesian estimator. Using Blackwell dominance of mean preserving spreads, we construct a partial ordering of the noise densities which relates the anonymity of the observations to the asymptotic covariance of the adaptive filtering algorithm.

**Keywords:** Adaptive filtering, Blackwell dominance, symmetric transformation, polynomial ring, Algebraic Liapunov equation, anonymization

## I. INTRODUCTION

The classical stochastic gradient algorithm operates on a *vector-valued* observation process that is inputted to the algorithm at each time instant. Suppose due to anonymization, the observation at each time is a *set* (i.e., the elements are unordered rather than a vector). Given these anonymized observation sets over time, how to construct a stochastic gradient algorithm to estimate the underlying parameter?

Figure 1 shows the schematic setup comprising  $L$  simultaneous independent stochastic systems indexed by  $l = 1, \dots, L$ , evolving over discrete time  $k = 1, 2, \dots$ . Each stochastic system  $l$  is parametrized by true model  $\theta_l^o \in \mathbb{R}^D$  and generates observations  $y_l(k) \in \mathbb{R}^D$  given input signal  $D \times D$  dimensional matrix  $\psi(k)$ :

$$y_l(k) = \psi(k) \theta_l^o + v_l(k), \quad l \in [L] \stackrel{\text{defn}}{=} \{1, \dots, L\} \quad (1)$$

Vikram Krishnamurthy is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, 14853, USA. Email: vikramk@cornell.edu. This research was supported by the National Science Foundation grant CCF-2112457, U.S. Army Research Office grant W911NF-21-1-0093 and US. Air Force Office of Scientific Research grant FA9550-22-1-0016.

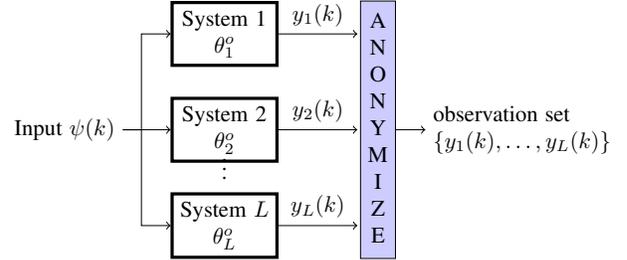


Fig. 1: Schematic setup comprising  $L$  stochastic systems. Given the sequence of anonymized observation sets  $(\{y_1(k), \dots, y_L(k)\}, k = 1, 2, \dots)$ , the aim is to estimate the underlying parameter set  $\theta^o = \{\theta_1^o, \dots, \theta_L^o\}$  of the  $L$  systems.

We assume that  $v_l(k) \in \mathbb{R}^D$  is iid random sequence with bounded second moment. We (the analyst) know (or can choose) the input signal sequence  $(\psi(k), k = 1, 2, \dots)$ . For convenience, assume that elements of  $(\psi(k), k = 1, 2, \dots)$  are zero mean iid sequences of random variables. Thus the output of the  $L$  stochastic systems at time  $k$  is the observation **matrix**

$$\mathbf{y}(k) = [y_1(k), \dots, y_L(k)]' \in \mathbb{R}^{L \times D}$$

where  $a'$  denotes transpose of matrix  $a$ .

The analyst observes at each time  $k$  the anonymized (unlabeled) observation **set**

$$y(k) = \sigma_k(\mathbf{y}(k)) = \{y_1(k), \dots, y_L(k)\} \quad (2)$$

The anonymization map  $\sigma_k$  is a permutation over the set  $\{1, 2, \dots, L\}$ . By anonymization<sup>1</sup> we mean that by transforming the matrix  $\mathbf{y}$  with ordered rows to set  $y$  with unordered rows, the index label  $l \in \{1, 2, \dots, L\}$  is hidden; that is, the observations are unlabeled. The time dependence of  $\sigma_k$  emphasizes that the permutation map operating on  $\mathbf{y}(k)$  changes at each time  $k$ .

**Aim.** The analyst only sees the anonymized observation set  $y(k)$  at each time  $k$ . Given the time sequence of observation sets  $(y(k), k = 1, 2, \dots)$ , the aim of the analyst is to estimate the underlying set of true parameters  $\theta^o = \{\theta_1^o, \dots, \theta_L^o\}$  of the  $L$  stochastic systems. Note that the analyst aim is estimate the *set*  $\theta^o$ ; due to the anonymization (unknown permutation map), in general, it is impossible to estimate which parameter belongs to which stochastic system.

<sup>1</sup>For now we use anonymization to denote masking the index label  $l$  of the stochastic process. Sec. I-C motivates this in terms of  $k$ -anonymity.

*Remarks:* (i) Another way of viewing the estimation objective is: Given noisy measurements of unknown permutations of the rows a matrix, how to estimate the elements of the matrix? Our main result is to propose a symmetric transform framework that circumvents modeling the permutations  $\sigma_k$  and is completely agnostic to the probabilistic structure of  $\sigma_k$ .

(ii) The assumption that  $\psi(k)$  is a  $D \times D$  matrix in (1) is without loss of generality. The classical LMS framework involves scalar valued observations  $o_l(k) = \psi'(k)\theta_l^o + e_l(k)$  where  $\psi(k) \in \mathbb{R}^D$  is the known regression vector, and  $e_l(k)$  is a noise process. If we stack  $D$  such scalar observations into the vector  $y_l(k)$ , then we obtain (1).

(iii) The model reflects *uncertainty associated with the origin of the measurements* (arbitrary permutation) in addition to their inaccuracy (additive noise). If we knew which observation  $m$  was associated with which stochastic system  $l$ , then we can estimate each  $\theta_l^*$  independently as the solution of the following stochastic optimization problem:  $\theta_l^* = \arg \min_{\theta} \mathbb{E}\{(y_l(k) - \psi(k)\theta_l)^2\}$ . Then the classical LMS algorithm can be applied to estimate each  $\theta_l^*$  recursively as:

$$\theta_l(k+1) = \theta_l(k) + \epsilon \psi(k)(y_l(k) - \psi(k)\theta_l(k)) \quad (3)$$

where the fixed step size  $\epsilon > 0$  is a small positive constant.

(iv) Since the ordering of the elements of the set  $\{y_1(k), \dots, y_L(k)\}$  is arbitrary, we cannot use the LMS algorithm (3). If we naively choose a random permutation of the set  $y(k)$  as the observation vector, and feed this  $L$ -dimensional observation vector into  $L$  LMS algorithms (3), then the estimates will not in general converge to  $\theta_l^o$ ,  $l = 1, \dots, L$ .

(v) Finally, the above formulation only makes sense in the stochastic case. The deterministic case is trivial. If the noise  $v_l(k) = 0$  and input matrix  $\psi(k)$  is invertible, then we need only one observation  $y$  to completely determine the parameter set  $\theta^o$ , regardless of the permutation  $\sigma_k$ .

#### *Stochastic Optimization with Anonymized Observations. Circumventing Data Association*

Broadly, there are two classes of methods for dealing with unlabeled observation model (1), (2). One class of methods is based on data association [1], [2], [3]. Data association deals with the question: How can the observations from multiple simultaneous processes be assigned to specific processes when there is uncertainty about which observation came from which process? Since the observations are anonymized wrt to the index label  $l$  of the random processes, one approach is to construct a classifier that assigns at each time  $k$  the observation  $y_l(k)$  to a specific process  $m$ . Because the number of process/observation pairs grows combinatorially with the number of processes and observations, a brute force approach to the data association problem is computationally prohibitive. Data association is studied extensively in Bayesian filtering for target tracking. In this paper we are dealing with stochastic optimization instead of Bayesian estimation, where we wish to preserve the convex structure of the problem.

The second class of methods bypasses data association, i.e., labels are no longer estimated (assigned) to the anonymized observations. This paper focuses on using symmetric transforms to bypass data association, as discussed next.

#### *A. Main Idea. Symmetric Transforms & Adaptive Filtering*

Since the assignment step in data association can destroy the convexity structure of a stochastic optimization problem, a natural question is: *Can data association be circumvented in a stochastic optimization problem?* A remarkable approach developed in the 1990s by Kamen and coworkers [4], [5] in the context of Bayesian estimation, involves using symmetric transforms. This ingenious idea circumvents data association; see also [6] and references therein. In this paper we extend this idea of symmetric transforms to stochastic optimization. Specifically, we show that the symmetric transform approach preserves convexity. Since [4] deals with Bayesian filtering for estimating the state, convexity is irrelevant. In comparison, preservation of convexity is crucial in stochastic optimization problems to ensure that the estimates of a stochastic gradient algorithm converge to the global minimum.

To explain our main ideas, suppose there are  $L = 3$  scalar-valued random processes, so each observation  $y_l(k)$  is scalar-valued. Further for simplicity assume the input signal  $\psi(k) = 1$ ; so the observations are  $y_l(k) = \theta_l^o + v_l(k)$ . Given the anonymized observation set  $y(k) = \{y_1(k), \dots, y_3(k)\}$  at each time  $k$ , how to estimate the parameters  $\theta_1^o, \theta_2^o, \theta_3^o$ ? Our main idea is to use the set  $y(k)$  to construct a pseudo-measurement vector  $z(k) \in \mathbb{R}^3$ . Suppressing the time dependency ( $k$ ) for notational convenience, we construct the pseudo-measurements  $z_1, z_2, z_3$  via a symmetric transform as follows:

$$\begin{aligned} z_1 &= S_1\{y_1, y_2, y_3\} = y_1 + y_2 + y_3 \\ z_2 &= S_2\{y_1, y_2, y_3\} = y_1 y_2 + y_1 y_3 + y_2 y_3 \\ z_3 &= S_3\{y_1, y_2, y_3\} = y_1 y_2 y_3 \end{aligned} \quad (4)$$

The key point is that the pseudo-observations  $z_l$  are symmetric in  $y_1, y_2, y_3$ . Any permutation of the elements of  $\{y_1, \dots, y_3\}$  does not affect  $z_l$ . In this way, we have circumvented the data association problem; there is no need to assign (classify) an observation to a specific process. But we have introduced a new problem: estimating  $\theta^o$  using the pseudo-observations is no longer a convex stochastic optimization problem. To estimate  $\theta^o$  we minimize the second order moments to compute:

$$\begin{aligned} \theta^* &= \arg \min_{\theta} \{\mathbb{E}\{(z_1 - (\theta_1 + \theta_2 + \theta_3))^2\} \\ &\quad + \mathbb{E}\{(z_2 - (\theta_1\theta_2 + \theta_1\theta_3 + \theta_2\theta_3))^2\} + \mathbb{E}\{(z_3 - \theta_1\theta_2\theta_3)^2\}\} \end{aligned} \quad (5)$$

Clearly the multi-linear objective (5) is non-convex in  $\theta_1, \theta_2, \theta_3$ . However, the problem is convex in the symmetric transformed variables (denoted as  $\lambda$  below), and the original variables  $\theta$  can be evaluated by inverting the symmetric transform. We formalize this as follows:

**Result 1.** *(Informal version of Theorem 1) The global minimum  $\theta^*$  of the non-convex objective (5) can be computed in three steps:*

(i) *Given the observations  $y(k)$ , compute the pseudo-observations  $z(k)$  using (4).*

(ii) *Using these pseudo-observations, estimate the pseudo parameters  $\lambda_1 = \theta_1 + \theta_2 + \theta_3$ ,  $\lambda_2 = \theta_1\theta_2 + \theta_1\theta_3 + \theta_2\theta_3$ ,  $\lambda_3 = \theta_1\theta_2\theta_3$ . Clearly (5) is a stochastic convex optimization problem in pseudo-parameters  $\lambda_1, \lambda_2, \lambda_3$ . Let  $\lambda_1^*, \lambda_2^*, \lambda_3^*$  denote the estimates.*

(iii) Finally, solve the polynomial equation  $s^3 + \lambda_1^* s^2 + \lambda_2^* s + \lambda_3^* = 0$ . Then the roots<sup>2</sup> are  $\theta^*$ . Computing the roots of a polynomial is equivalent to computing the eigenvalues of the corresponding companion matrix (Matlab command `roots`).

Put simply the above result says that while (5) is non-convex in the roots of a polynomial, it is convex in the coefficients of the polynomial! To explain Step (ii), clearly (5) is convex in the pseudo-parameters  $\lambda_1, \lambda_2, \lambda_3$ . We can straightforwardly compute the global minimum in terms of these pseudo parameters as  $\lambda_1^* = \mathbb{E}\{z_1\}$ ,  $\lambda_2^* = \mathbb{E}\{z_2\}$ ,  $\lambda_3^* = \mathbb{E}\{z_3\}$ .

To explain Step (iii) of the above result, we use a crucial property of symmetric functions. The reader can verify that the following monic polynomial in variable  $s$  satisfies

$$(s + \theta_1)(s + \theta_2)(s + \theta_3) = s^3 + \lambda_1 s^2 + \lambda_2 s + \lambda_3$$

The above equation states that a monic polynomial with pseudo-parameters  $\lambda_1, \lambda_2, \lambda_3$  as coefficients has the parameters  $\theta_1, \theta_2, \theta_3$  as roots of the polynomial. By the fundamental theorem of algebra, there is a unique invertible map between the coefficients of a monic polynomial and the set of roots of the polynomial. As a result having computed the global minimum  $\lambda^*$  of the above objective (5) (since it is convex in  $\lambda$ ), we can compute the unique parameter set  $\theta^*$ , which are the set of roots of the corresponding polynomial. Thus we have computed the global minimum  $\theta^*$  of the non-convex objective (5). To summarize Result 1 gives a constructive method to estimate the true parameter set  $\theta^o$  given anonymized observations (albeit in an extremely simplified setting).

## B. Main Results and Organization

- 1) Our first main result in Sec. II, extends the above simplistic formulation to a random input process  $\psi(k)$  rather than a constant. To achieve this, Theorem 1 exploits the homogeneous property of the symmetric transform  $S$  to construct a consistent estimator for  $\theta^o$ . In Theorem 1, we will construct a stochastic gradient algorithm that generates a sequence of estimates  $\lambda(k)$  that provably converges to  $\lambda^*$  (since the problem is convex). The roots of the corresponding polynomial converge to  $\theta^*$ .
- 2) Sec. III extends this symmetric transform approach to the case where each anonymized observation  $y_l(k)$  is a vector in  $\mathbb{R}^D$  where  $D \geq 2$  in (1). For this vector case, three issues need to be resolved:
  - a) It is not possible to use the scalar symmetric transform (4) element-wise on vector observations. Naively applying the scalar symmetric transforms element wise yields “ghost” parameters estimates that are jumbled across the various stochastic systems (see Sec.III-A.)
  - b) Since a scalar symmetric transform (or equivalently, the one variable polynomial transform) is not useful, we will use a two-variable polynomial transform inspired by [7]. However, a new issue arises. In the scalar observation case, we use the fundamental theorem of algebra to construct a unique mapping between the roots of a polynomial and the coefficients of the

polynomial. Unfortunately, in general the fundamental theorem of algebra does not extend to polynomials in two variables. The key point we will exploit below is that the ring of two-variable polynomials is a unique factorization domain over the ring of one-variable polynomials. This gives us a constructive method to extend Theorem 1 to sets of vector observations ( $D \geq 2$ ). This is the content of our main result Theorem 2.

- c) The final issue is that of homogeneity of the symmetric transform. In the scalar case, the homogeneity property is crucial in the proof of Theorem 1. We construct a suitable multidimensional generalization for the vector case in order to prove Theorem 2.
- 3) *Asymptotic Covariance of Adaptive Filtering Algorithm:* Sec. III-D analyzes the convergence and asymptotic covariance of the adaptive filtering algorithm (28). In the stochastic approximation literature [8], [9], the asymptotic rate of convergence is specified in terms of the asymptotic covariance of the estimates. We study the asymptotic efficiency of the proposed adaptive filtering algorithm. Specifically we address the question: *How much larger is the asymptotic covariance due to use of the symmetric transform to circumvent anonymization, compared to the classical LMS algorithm when there is no anonymization?*
- 4) *Mixture Model for Noisy Matrix Permutations:* We can assign a probability law to the permutation process  $\sigma$  in the anonymized observation model (1), (2) as follows:

$$y(k) = \sigma(x(k)) \begin{matrix} \theta^o & \psi(k) \\ L \times L & L \times D \end{matrix} + \sigma(x(k)) \begin{matrix} v(k) \\ D \times D \\ L \times D \end{matrix} \quad (6)$$

Here  $\sigma(x(k))$  denotes a randomly chosen  $L \times L$  permutation matrix that evolves according to some random process  $x$ . So (6) is a probabilistic mixture model. The matrix valued observations  $y(k)$  are random permutations of the rows of matrix  $\theta^o \psi(k)$  corrupted by noise. Given these observations, the aim is to estimate the matrix  $\theta^o$ . Note that there are  $L!$  possible permutation matrices  $\sigma$ . In the context of mixture models, Section IV and Appendix A present two results:

- (i) *Mean-preserving Blackwell dominance and Anonymity of permutation process:* Section IV uses the error probability of the Bayesian posterior estimate of the random permutation state  $x(k)$  in (6) as a measure of anonymity. This is in line with [10] where anonymity is studied in the context of mutual information and error probabilities. We will then use Blackwell dominance and a novel result in mean preserving spreads to relate this anonymity to the covariance of our proposed adaptive filtering algorithms.
- (ii) *Recursive Maximum likelihood estimation of  $\theta^o$*  In Appendix A, we discuss a recursive maximum likelihood estimation (MLE) algorithm for the parameters  $\theta^o$ . This requires knowing the density of  $v$  and the mixture probabilities (of course these can be estimated, but given the  $L!$  state space dimension, this becomes intractable). A more serious issue is that the likelihood is not necessarily concave in  $\theta$ . In comparison, our symmetric function approach yields a convex stochastic optimization problem.

<sup>2</sup>Strictly speaking  $\theta_1, \theta_2, \theta_3$  are factors. The root is the negative of a factor.

### C. Applications of Anonymized Observation Model

We classify applications of the anonymized observation model (1), (2) into two types: (i) Due to sensing limitations, the sensor provides noisy measurements from multiple processes, and there is uncertainty as to which measurement came from which process and (ii) examples where the identities of the processes generating the measurements are purposefully hidden to preserve anonymity.

1. *Sensing/Tracking Multiple Processes with Unlabeled Observations*: The classical observation model comprises a sensor (e.g. radar) that generates noisy measurements where, due to sensing limitations, there is uncertainty in the origin of the measurements. The observations are unlabeled and not assigned to a specific target process [1]. In this context, estimating the underlying parameter  $\theta^o$  of the target processes is identical to our estimation objective. As mentioned earlier, data association is widely studied in Bayesian estimation for target tracking. In this paper we focus on stochastic optimization with anonymized observations. For example, to estimate the underlying parameters, or more generally adaptively optimize a stochastic system comprising  $L$  parallel process.

2. *Adaptive Estimation with  $k$ -Anonymity and  $l$ -diversity*: We now discuss examples where the labels (identities) of the  $L$  processes are purposefully hidden. Anonymization of trajectories arises in several applications including health care where wearable monitors generate time series of data uniquely matched to an individual, and connected vehicles, where location traces are recorded over time.

The concept of  $k$ -anonymity<sup>3</sup> (we will call this  $L$ -anonymity since we use  $k$  for time) was proposed by [11]. It guarantees that there are at least  $L$  identical records in a data set that are indistinguishable. In our formulation, due to the anonymization step (2), the identities (indexes)  $l$  of the  $L$  processes are indistinguishable. More generally, in the model (1), (2), the identity  $l$  of each target itself can be a categorical vector  $[l_1, \dots, l_N]$ . For example if each process models GPS data trajectories of individuals, the categorical data  $\psi_l(k)$  records discrete-valued variables such as individuals identity, specific locations visited, etc. To ensure  $L$ -anonymity, these categorical vectors are all allocated a single vector, thereby maintaining anonymity of the categorical data. Thus the analyst only sees the anonymized observation set  $y(k)$ .

Note that  $L$ -anonymity hides identity  $l$  but discloses attribute information, namely the noisy observation set  $y(k)$ . To enhance  $L$ -anonymity, the attributes in  $L$ -anonymized data are often  $M$ -diversified<sup>4</sup> [12]: each equivalence class is constructed so that there are at least  $M$  distinct parameters. In our notation, if at least  $M$  processes have distinct parameter vectors  $\theta_l$ ,  $l = 1, \dots, M$ , then  $M$ -diversity of the attribute data is achieved.

<sup>3</sup>Data anonymity is mainly studied under two categories:  $k$ -anonymity and differential privacy. Differential privacy methods typically add noise to trajectory data providing a provable privacy guarantee for the data set. Even though we consider Laplacian noise for  $v$  in the numerical studies and this can be motivated in terms of differential privacy; we will not discuss differential privacy in this paper.

<sup>4</sup>The terminology used in the literature is “ $l$ -diversified”; but we use  $l$  for the index of the target process.

In our formulation, the input signal matrices  $\psi(k)$  are the same for all  $L$  processes. Thus the input matrices also preserve  $L$ -anonymity. If the analyst could specify a different input signal  $\psi_l$  to each system  $l$ , then the analyst can straightforwardly estimate  $\theta_l^o$  for each target process  $l$ , thereby breaking anonymity; see Remark 6 after Theorem 1 below.

3. *Product Sentiment given Anonymized Ratings*: Reputation agencies such as Yelp post anonymized ratings or products. Market analysts aim to estimate the true sentiment of the group of users given these anonymized ratings [13].

4. *Evaluating Effectiveness of Teaching Strategy given Anonymized Responses*: A teacher instructs  $L$  students with input signal  $\psi(k)$ . Each student  $l$  has prior knowledge  $\theta_l^o$ , and responds to the teaching input with answer  $y_l(k)$ . The identity  $l$  of the student is hidden from the teacher. Based on these anonymous responses, the teacher aims to estimate the students prior knowledge  $\theta^o$ . See also [14] for other examples. Anonymized trials are also used in evaluating the effectiveness of drugs vs placebo.

## II. ADAPTIVE FILTERING WITH SCALAR ANONYMIZED OBSERVATIONS

For ease of exposition, we first discuss the problem of estimating the true parameter  $\theta^o$  when the observation  $y_l(k)$  of each process  $l$  is a scalar; so  $D = 1$  in (1) and  $\psi(k)$  is a scalar. Since there are  $L$  independent scalar processes in (1), the parameters generating these  $L$  processes is  $\theta^o = \{\theta_1^o, \dots, \theta_L^o\}$ .

Given the anonymized observation set  $y(k) = \{y_1(k), \dots, y_L(k)\}$  at each time  $k$ , our main idea is to construct a pseudo-measurement vector  $z(k) \in \mathbb{R}^L$ . Suppressing the time dependency ( $k$ ) for notational convenience, we construct the  $L$  pseudo-measurements  $z_l, l \in [L]$  via a symmetric transform<sup>5</sup> [15] as follows:

$$z = S\{y\} \iff z_l = S_l\{y_1, \dots, y_L\} \stackrel{\text{defn}}{=} \sum_{i_1 < i_2 < \dots < i_l} y_{i_1} y_{i_2} \dots y_{i_l}, \quad l \in [L] \quad (7)$$

Recall our notation  $[L] = \{1, \dots, L\}$ . It is easily shown using the classical Vieta’s formulas [16], that the pseudo-measurements  $z_l, l \in [L]$  in (7) are the coefficients of the following  $L$ -order polynomial in variable  $s$ :

$$S\{y\}(s) \stackrel{\text{defn}}{=} \prod_{l=1}^L (s + y_l) = s^L + \sum_{l=1}^L z_l s^{L-l} \quad (8)$$

As an example, consider  $L = 3$  independent scalar processes. Then the pseudo-observations using (7) are given by (4). The reader can verify that the pseudo-observations  $z_1, z_2, z_3$  are the coefficients of the polynomial  $(s + y_1)(s + y_2)(s + y_3)$ .

Note that each  $z_l$  is permutation invariant: any permutation of the elements of  $\{y_1, \dots, y_L\}$  does not affect  $z_l$ . That is why our notation above involves the set  $\{y_1, y_2, \dots, y_L\}$ .

<sup>5</sup>By symmetric transform  $S_l$ , we mean  $S_l\{y_1, \dots, y_L\} = S_l\{P \cdot \{y_1, \dots, y_L\}\}$  for any permutation  $P$  of  $\{y_1, \dots, y_L\}$ . Thus while the elements  $\{y_1, \dots, y_L\}$  are arbitrarily ordered, the value of  $S_l\{\cdot\}$  is unique. Eq (8) gives a systematic construction of such symmetric transforms that is uniquely invertible, see (14).

*Remark:* It is easily verified from (7) that the symmetric transforms  $S_l$  is homogeneous of degree  $l$ : for any  $c \in \mathbb{R}$ ,

$$S_l\{c\theta_1, \dots, c\theta_L\} = c^l S_l\{\theta_1, \dots, \theta_L\}, \quad l \in [L] \quad (9)$$

#### A. Symmetric Transform and Estimation Objective

Given the set valued sequence of anonymized observations,  $y(1), y(2), \dots, y(k), \dots$  generated by (1), our aim is to estimate the true parameter set  $\theta^o = \{\theta_1^o, \dots, \theta_L^o\}$ . To do so, we first construct the pseudo measurement vectors  $z(1), z(2), \dots, z(k)$  via (7). Denoting  $\theta = \{\theta_1, \dots, \theta_L\}$ , our objective is to estimate the set  $\theta^* = \{\theta_1^*, \dots, \theta_L^*\}$  that minimizes:

$$\theta^* = \arg \min_{\theta} \sum_{l \in [L]} \mathbb{E} |z_l - S_l\{\psi \theta_1, \psi \theta_2, \dots, \psi \theta_L\}|^2 \quad (10)$$

where  $z_l = S_l\{\psi \theta_1^o + v_1, \dots, \psi \theta_L^o + v_L\}$

Recall the symmetric transform  $S_l$  is defined in (7). Finally, define the symmetric transforms on the model parameters as

$$\lambda = S\{\theta\} \iff \lambda_l = S_l\{\theta_1, \dots, \theta_L\}, \quad l \in [L]. \quad (11)$$

Note that  $\lambda = [\lambda_1, \dots, \lambda_L]'$  is an  $L$ -dimension vector whereas  $\theta$  is a set with  $L$  (unordered) elements.

From (10), we see that  $\theta^*$  is a second order method of moments estimate of  $\theta^o$  wrt pseudo observations. Importantly, this estimate is independent of the anonymization map  $\sigma$ .

#### B. Main Result. Consistent Estimator for $\theta^o$ .

We are now ready to state our main result, namely an adaptive filtering algorithm to estimate  $\theta^o$  given anonymized scalar observations. The result says that while objective (10) is non-convex in  $\theta$ , we can reformulate it as a convex optimization problem in terms of  $\lambda$  defined in (11). The intuition is that the objective (10) is non-convex in the roots of the polynomial (namely,  $\theta$ ), but is convex in the *coefficients* of the polynomial (namely,  $\lambda$ ); and by the fundamental theorem of algebra there is a one-to-one map from the coefficients  $\lambda$  to the roots  $\theta$ . Therefore, by mapping observations to pseudo observations (coefficients of the symmetric polynomial), we can construct a globally optimal estimate of (10).

**Theorem 1.** *Consider the sequence of anonymized observation sets  $(y(k), k \geq 1)$  generated by (1) and (2), where  $\psi(k)$  is a known iid scalar sequence. Then*

- 1) *The objective (10) can be expressed as  $L$  decoupled convex optimization problems in terms of  $\lambda$  defined in (11):*

$$\min_{\lambda_l} \mathbb{E} |z_l - \psi^l \lambda_l|^2 \quad \text{where} \quad (12)$$

$$z_l(k) = (\psi(k))^l \lambda_l^o + w_l(k)$$

*The process  $w(k)$  is defined explicitly in (59) below.*

- 2) *The global minimizer  $\theta^*$  of objective (10) is consistent in the sense that  $\theta^* = \theta^o$ .*
- 3) *With pseudo observations  $z(k) = S\{y(k)\}$  defined in (7), consider the following bank of  $L$  decoupled adaptive filtering algorithms operating on  $z(k)$ : Choose  $\lambda(0) \in \mathbb{R}^L$ . Then for  $l \in [L]$ , update as*

$$\lambda_l(k+1) = \lambda_l(k) + \epsilon \psi^l(k) (z_l(k) - \psi^l(k) \lambda_l(k))$$

$$\theta(k+1) = \Re \epsilon (S^{-1}(\lambda(k+1))) \quad (13)$$

*Here  $S^{-1}$  is defined in (14) and  $\Re$  denotes the real part of the complex vector. The estimates  $\theta(k)$  converge in probability and mean square to  $\theta^*$  (see Theorem 3).*

*Discussion:* 1. Theorem 1 gives a tractable and consistent method for estimating the parameter set  $\theta^o$  of the  $L$  stochastic systems given set valued anonymized observations  $y(1), y(2), \dots$ . We emphasize that since the observations  $y(k)$  are set-valued, the ordering of the elements of  $\theta^o$  cannot be recovered; Statement 1 of the theorem asserts that the set-valued estimate  $\theta^*$  converges to  $\theta^o$ . Statement 2 of the theorem gives an adaptive filtering algorithm (13) that operates on the pseudo observation vector  $z(k)$ . Applying the transform  $S^{-1}$  to the estimates  $\lambda(k)$  generated by (13) yields estimates  $\theta(k)$  that converge to the global minimum  $\theta^*$ . Since by assumption  $\theta^o \in \mathbb{R}^L$ , the second step of (13) chooses the real part of the possibly complex valued roots.

2. An important property of the symmetric operator  $S$  is that it is uniquely invertible since any  $L$ -th degree polynomial has a unique set of at most  $L$  roots. Indeed, given  $\lambda = S\{\theta\}$ ,  $\theta = S^{-1}(\lambda)$  are the unique set of roots  $\{\theta_1, \dots, \theta_L\}$  of the polynomial with coefficients  $\lambda_l, l \in [L]$ , that is,

$$\theta = S^{-1}(\lambda) \iff s^L + \sum_{l=1}^L \lambda_l s^{l-1} = \prod_{l=1}^L (s + \theta_l) \quad (14)$$

Note that  $S^{-1}(\cdot)$  maps the vector  $\lambda$  to unique set  $\theta$ . Recall that  $S\{\cdot\}$  maps set  $\theta$  to unique vector  $\lambda$ . Computing the roots of a polynomial is equivalent to computing the eigenvalues of the companion matrix e.g., Matlab command `roots`.

3. Typically the roots of a polynomial can be a sensitive function of the coefficients. However, this does not affect algorithm (13) since it operates on the coefficients only. The roots are *not* fed back iteratively into algorithm (13). In Section III-D and Theorem 6 below, we will quantify this sensitivity in terms of the asymptotic covariance of algorithm (13).

4. The adaptive filtering algorithm (13) uses a constant step size; hence it converges weakly (in distribution) to the true parameter  $\theta^o$  [9]. Since we assumed  $\theta^o$  is a constant, weak convergence is equivalent to convergence in probability. Later we will analyze the tracking capabilities of the algorithm when  $\theta^o$  evolves in time according to a hyper-parameter.

5. A stochastic gradient algorithm operating directly on objective (10) is

$$\theta(k+1) = \theta(k) - \epsilon \nabla_{\theta} \sum_{l \in [L]} |z_l(k) - S_l\{\psi(k) \theta_1(k), \dots, \psi(k) \theta_L(k)\}|^2 \quad (15)$$

We show via numerical examples in Sec.V that objective (10) has local minima and stochastic gradient algorithm (15) can get stuck at these local minima. In comparison, the formulation involving pseudo-measurements yields a convex (quadratic) objective and algorithm (13) provably converges to the global minimum. There is also another problem with (15). If the initial condition  $\theta(0)$  is chosen with equal elements, then since the gradient  $\nabla_{\theta}$  is symmetric (wrt  $y$  and  $\theta$ ), all the elements of the estimate  $\theta(k)$  have equal elements at each time  $k$ , regardless of the choice of  $\theta^o$ , and so algorithm (15) will not converge to  $\theta^o$ .

6. *Anonymization of input signal  $\psi(k)$* : We assumed that the input signal matrices  $\psi(k)$  are the same for all  $L$  processes. If the analyst can specify a different input signal  $\psi_l$  to each system  $l$ , then the analyst can estimate  $\theta_l^o$  for each target process  $l$  via classical least squares, thereby breaking anonymity as follows: Minimizing  $\mathbb{E}\{\sum_{l \in [L]} y_l - \psi_l \theta_l\}^2 = \mathbb{E}\{z_1 - \sum_{l \in [L]} \psi_l \theta_l\}^2$  wrt  $\theta_l$  yields the classical least squares estimator. Thus the analyst only needs the pseudo observations  $z_1(k) = \sum_l y_l(k)$  to estimate  $\theta_l^o$  and thereby break anonymity.

In our formulation, since the regression input signals  $\psi_l$  are identical, minimizing  $\mathbb{E}\{z_1 - \psi \sum_{l \in [L]} \theta_l\}^2$  only estimates the sum of parameters, namely  $\sum_l \theta_l^o$ ; the individual parameters are not identifiable. This is why we require pseudo-observations  $z_1, \dots, z_L$  to estimate the elements  $\theta_l^o, l \in [L]$ .

### III. ADAPTIVE FILTERING GIVEN VECTOR ANONYMIZED OBSERVATIONS

We now consider the case  $D \geq 2$ , namely, for each process  $l \in [L]$ , the observation  $y_l(k)$  in (1) is a  $D$ -dimensional vector. We observe the (unordered) set  $y(k) = \{y_1(k), \dots, y_L(k)\}$  at each time  $k$ . That is, we do not know which observation vector  $y_l(k)$  came from which process  $l$ . Given the anonymized observation set (2), the aim is to estimate  $\theta^o \in \mathbb{R}^{L \times D}$ .

*Remark.* For each observation vector  $y_l \in \mathbb{R}^D$ , let  $y_{l,i}$  denote the  $i$ -th component. Note that the elements of each vector  $y_l$  are ordered, namely  $y_l = [y_{l1}, \dots, y_{lD}]'$ , but the first index  $l$  (identity of process) is anonymized yielding the observation set  $y = \{y_1, \dots, y_L\}$ .

As mentioned in Sec.I, for this vector case, three issues need to be resolved: First, naively applying the scalar symmetric transforms element wise yields “ghost” parameter estimates that are jumbled across the various stochastic systems. (We discuss this in more detail below.) Second, we need a systematic way to encode the observation vectors via a symmetric transform that is invertible. We will use a two-variable polynomial transform. However, a new issue arises; in general the fundamental theorem of algebra, namely that an  $L$ -th degree polynomial has up to  $L$  complex valued roots, does not extend to polynomials in two variables. We will construct an invertible map for two-variable polynomials. This gives us a constructive method to extend Theorem 1 to vector observations  $D \geq 2$ . The final issue is that of homogeneity of the symmetric transform. Recall in the scalar case, the homogeneity property (9) was crucial in the proof of Theorem 1. We need to generalize this to the vector case. The main result (Theorem 2 below) addresses these three issues.

#### A. Symmetric Transform for Vector Observations

This section constructs the symmetric transform  $S$  for vector observations. The construction involves a polynomial in two variables,  $s$  and  $t$ . It is convenient to first define the symmetric transform for an arbitrary set  $\alpha = \{\alpha_1, \dots, \alpha_L\}$  where  $\alpha_l \in \mathbb{R}^D$ . The symmetric transform is defined as

$$S\{\alpha\}(s, t) = \prod_{l=1}^L (s + \sum_{i=1}^D \alpha_{l,i} t^{i-1}) = s^L + \sum_{l=1}^L \sum_{m=1}^{M_l} S_{l,m}\{\alpha\} s^{L-1} t^{m-1} \quad (16)$$

where  $M_l \stackrel{\text{defn}}{=} (L-l)(D-1) + D$

So the symmetric transform is the array of polynomial coefficients  $S_{l,m}\{\alpha\}$  of the above two variable polynomial. We write this notationally as

$$S\{\alpha\} = [S_{l,m}\{\alpha\}, m = 1, \dots, M_l, l \in [L]]$$

When  $D = 1$ , we see that the symmetric transform (16) specializes to (7).

Another equivalent way of expressing the above symmetric transform involves convolutions: The  $M_l$  dimensional vector  $S_l\{\alpha\} = [S_{l1}\{\alpha\}, \dots, S_{lM_l}\{\alpha\}]'$  satisfies

$$S_l\{\alpha\} = \sum_{i_1 < i_2 < \dots < i_l} \alpha_{i_1} \otimes \alpha_{i_2} \otimes \dots \otimes \alpha_{i_l}, \quad l \in [L] \quad (17)$$

where  $\otimes$  denotes convolution. Eq. (17) serves as a constructive computational method to compute the symmetric transform of a set  $\alpha$ .

With the above definition of the symmetric transform, consider the observation set  $y(k) = \{y_1(k), \dots, y_L(k)\}$  at each time  $k$ . We define the pseudo-observations as

$$z(k) = S\{y(k)\} \quad (18)$$

**Example.** To illustrate the polynomial  $S\{y\}(s, t)$ , consider  $L = 2$  independent processes each of dimension  $D = 2$ . Then with  $y_1 = [y_{11}, y_{12}]'$ ,  $y_2 = [y_{21}, y_{22}]'$ , the symmetric polynomial (16) in variables  $s, t$  is

$$S\{y\}(s, t) = (s + y_{11} + y_{12}t)(s + y_{21} + y_{22}t) \quad (19)$$

Then the pseudo observations  $z_{lm}$  specified by the RHS of (16) are the coefficients of this polynomial, namely

$$\begin{aligned} z_{11} &= y_{11} y_{21}, \quad z_{12} = y_{11} y_{22} + y_{12} y_{21}, \quad z_{13} = y_{12} y_{22}, \\ z_{21} &= y_{11} + y_{21}, \quad z_{22} = y_{12} + y_{22} \end{aligned} \quad (20)$$

In the convolution notation (17), the pseudo-observations are

$$z_1 = [z_{11}, z_{12}, z_{13}]' = y_1 \otimes y_2, \quad z_2 = [z_{21}, z_{22}]' = y_1 + y_2$$

We see from this example that the pseudo-observations (20) generated by the vector symmetric transform (16) is a superset of the scalar symmetric transforms applied to each component of the vector observation. Specifically pseudo-observations for the first elements of  $y_1$  and  $y_2$ , namely  $y_{11}, y_{2,1}$  are  $z_{11}, z_{21}$ . Similarly pseudo-observations on the second elements of  $y_1$  and  $y_2$ , namely  $y_{12}, y_{22}$  are  $z_{13}, z_{22}$ . But  $z_{12}$  in (20) is the extra pseudo-observation that cannot be obtained by simply constructing symmetric transforms of each individual element. In Sec.III-A below, we will discuss the importance of the above vector symmetric transform compared to a naive application of scalar symmetric transform element-wise.

*Why a naive element-wise symmetric transform is not useful:* Instead of the vector symmetric transform defined in (16), why not perform the scalar symmetric transform on each of the  $D$  components separately? To make this more precise, let us define the *naive* vector symmetric transform which uses the scalar symmetric transform  $S_l, l \in [L]$  in (8) as follows:

$$\bar{z}_{lj} = \bar{S}_{l,j}\{y\} = S_l\{y_{1,j}, \dots, y_{L,j}\}, \quad j \in \{1, \dots, D\} \quad (21)$$

This is simply the scalar symmetric transform  $S\{y_{1,j}, \dots, y_{L,j}\}$  applied separately to each component  $j = 1, \dots, D$ .

In analogy to (10), we can define the estimation objective in terms of the naive vector transform as

$$\bar{\theta}^* = \arg \min_{\theta} \sum_{l \in [L]} \sum_{j=1}^D \mathbb{E} |\bar{z}_{lj} - \bar{S}_{l,j} \{\psi\theta_1, \psi\theta_2, \dots, \psi\theta_L\}|^2 \quad (22)$$

The naive symmetric transform  $\bar{S}$  in (21), (22) loses ordering information of the vector elements; for example given two processes ( $L = 2$ ) each of dimension  $D = 2$ ,  $\bar{S}$  does not distinguish between observation set  $\{[y_{1,1}, y_{1,2}], [y_{2,1}, y_{2,2}]\}$  and the observation set  $\{[y_{1,1}, y_{2,2}], [y_{2,1}, y_{1,2}]\}$ . It follows that  $\bar{\theta}^*$  in (22) is not a consistent estimator for  $\theta^o$ ; see remark following proof of Statement 4 of Theorem 2. Specifically, if the true parameters are  $\theta^o = \{[\theta_{11}^o, \theta_{12}^o], [\theta_{21}^o, \theta_{22}^o]\}$ , then the estimates can converge to the parameters of the “ghost processes”  $\{[\theta_{11}^o, \theta_{22}^o], [\theta_{21}^o, \theta_{12}^o]\}$ . That is, the parameter estimates get jumbled between the stochastic systems. Such “ghost” target estimates are common in data association in target tracking, and we will demonstrate a similar phenomenon in numerical examples of Sec.V when using the naive symmetric transform on anonymized vector observations.

In comparison, the vector symmetric transform (16) systematically encodes the observations with no information loss. For example in the  $D = 2, L = 2$  case, the extra pseudo-observation  $z_{12}$  in (19) allows to distinguish between these observation sets. (See also Appendix A for the example  $D = 3, L = 3$ .) To summarize, the vector symmetric transform is fundamentally different to the scalar symmetric transform. We will use the vector symmetric transform as a consistent estimator for  $\theta^o$  below.

### B. Main Result. Consistent Estimator for $\theta^o$

We first formalize our estimation objective based on the anonymized observations. Then we present the main result.

Denoting  $\theta = \{\theta_1, \dots, \theta_L\}$ , our objective is to estimate the set  $\theta^* = \{\theta_1^*, \dots, \theta_L^*\}$  that minimizes the following expected cost (where  $\|\cdot\|_F$  denotes the Frobenius norm): Compute

$$\theta^* = \arg \min_{\theta} \mathbb{E} \|S\{y_1, \dots, y_L\} - S\{\psi\theta_1, \psi\theta_2, \dots, \psi\theta_L\}\|_F^2 \quad (23)$$

Recall that  $\theta_l \in \mathbb{R}^D$  for each  $l \in [L]$ . For notational convenience we use  $\{\psi\theta\}$  to denote the set  $\{\psi\theta_1, \psi\theta_2, \dots, \psi\theta_L\}$ . Also  $y = \{y_1, \dots, y_L\}$  is the (anonymized) observation set.

*Remark.* As in the scalar case, we note that  $\theta^*$  in (23) is a second order method of moments estimate of  $\theta^o$  wrt pseudo observations, independent of anonymization map  $\sigma$ .

We are now ready to state our main result, namely an adaptive filtering algorithm to estimate  $\theta^o$  given the anonymized observation vectors. As in the scalar case, the main idea is that we have a convex optimization problem in the symmetric transform variables (denoted as  $\lambda$  below), and the variables  $\theta$  can be evaluated by inverting the symmetric transform.

**Theorem 2.** Consider the sequence of anonymized observation sets,  $(y(k), k \geq 1)$  generated by (1), (2), where  $\psi(k), k \geq 1$  is a known iid sequence of  $D \times D$  matrices. Then

- 1) The symmetric transform polynomial  $S\{y\}(s, t)$  in (16) can be decomposed into signal and noise polynomials as

$$S\{y\}(s, t) = S\{\psi\theta^o\}(s, t) + w(s, t) \quad (24)$$

where  $w(s, t)$  is a noise polynomial whose coefficients are zero mean. (We define  $w(s, t)$  in (63) below.)

- 2) The symmetric transform  $S$  has the following homogeneity property: With  $\lambda_{l,m} \stackrel{\text{defn}}{=} S_{l,m}\{\theta\}$ , then

$$S_{l,m}\{\psi\theta\} = \sum_{n \in M_l} \lambda_{l,n} S_{l,n}\{\psi^{l,m}\}, \quad l \in [L] \quad (25)$$

Here for  $\lambda_{m,n} = \sum_{i_1 \leq i_2 \leq \dots \leq i_l} \theta_{1,i_1} \theta_{2,i_2} \dots \theta_{l,i_l}$ , we construct  $\psi^{l,m}$  as the following  $D \times l$  matrix of elements from input matrix  $\psi$ :

$$\psi^{l,m} \stackrel{\text{defn}}{=} \begin{bmatrix} \psi_{i_1,1} & \psi_{i_2,1} & \dots & \psi_{i_l,1} \\ \psi_{i_1,2} & \psi_{i_2,2} & \dots & \psi_{i_l,2} \\ \vdots & \vdots & \dots & \vdots \\ \psi_{i_1,D} & \psi_{i_2,D} & \dots & \psi_{i_l,D} \end{bmatrix} \quad (26)$$

- 3) With pseudo observations  $z = S\{y\}$  defined in (7) and  $\psi^{l,m}$  defined in (26), the objective (23) can be expressed as  $L$  decoupled convex optimization problems:

$$\begin{aligned} [\lambda_{1,1}^*, \dots, \lambda_{l,M_l}^*] &= \arg \min_{\lambda_{1,1}, \dots, \lambda_{l,M_l}} \sum_{m \in M_l} \mathbb{E} |z_{lm}| \\ &\quad - \sum_{n \in M_l} \lambda_{l,n} S_{l,n}\{\psi^{l,m}\}|^2 \quad l \in [L] \quad (27) \\ \theta^* &= S^{-1}(\lambda^*) \end{aligned}$$

- 4) The global minimizer  $\theta^*$  of objective (23) is consistent in the sense that  $\theta^* = \theta^o$ .
- 5) With pseudo-observations  $z(k) = S\{y(k)\}$  computed by (17), consider the following  $L$  decoupled adaptive filtering algorithms operating on quadratic objective (27): Choose initial condition  $\lambda_{l,m}(0) \in \mathbb{R}$  arbitrarily. Update each element of  $\lambda_{l,m}$ ,  $m \in M_l, l \in [L]$  as

$$\begin{aligned} \lambda_{l,m}(k+1) &= \lambda_{l,m}(k) + \epsilon S_{l,m}(\psi^{lm}(k)) \\ &\quad \times \sum_{m \in M_l} (z_{l,m}(k) - \sum_{n \in M_l} \lambda_{l,n}(k) S_{l,n}\{\psi^{lm}(k)\}), \quad (28) \\ \theta(k+1) &= \Re\epsilon(S^{-1}(\lambda(k+1))) \end{aligned}$$

Here  $\epsilon > 0$  is the algorithm step size,  $S^{-1}$  is evaluated via (31), (32),  $\psi^{l,m}$  is constructed in (26), and  $S_{l,n}\{\cdot\}$  is computed in (17). Then the estimates  $\theta(k)$  converge in probability and mean square to  $\theta^*$  (see Theorem 3).

### C. Discussion of Theorem 2

Despite the complex notation, the important takeaway from (25) is that  $S_{l,m}\{\psi\theta\}$  is a linear function of  $\lambda_{l,n} = S_{l,m}\{\theta\}$ . Therefore the objective (23) becomes a convex (quadratic) optimization problem (27). Thus similar to the scalar case in Theorem 1, we have converted a non-convex problem in the roots of a two-dimensional polynomial to a convex problem in the coefficients of the polynomial. Since the map between the set of roots and vector of coefficients roots is uniquely invertible, the optimization objectives (23) and (27) are equivalent.

*Homogeneity of Symmetric Transform:* The fundamental theorem of symmetric functions states that any symmetric polynomial can be expressed as a polynomial in terms of elementary symmetric functions [17, Theorem 4.3.7]. However, Theorem 2 exploits the linear map  $\psi\theta$  to obtain the specific result (25), namely  $S_{l,m}\{\psi\theta\} = \sum_{n \in M_l} S_{l,n}\{\theta\} S_{l,n}\{\psi^{l,m}\}$ . This qualifies as a vector version of the homogeneity property (9) in the scalar case. The scale factor is  $S_{l,n}\{\psi^{l,m}\}$ .

As a simple example of evaluating the matrix  $\psi^{l,m}$  in (26), suppose  $L = 3, D = 3$ . Then since  $\lambda_{11} = \theta_{11}\theta_{21}\theta_{31}$ , it follows from (26) and (16) that

$$\psi^{11} = \begin{bmatrix} \psi_{11} & \psi_{11} & \psi_{11} \\ \psi_{12} & \psi_{12} & \psi_{12} \\ \psi_{13} & \psi_{13} & \psi_{13} \end{bmatrix}, \quad S_{11}\{\psi^{11}\} = \psi_{11}^3, \quad (29)$$

$S_{12}\{\psi^{11}\} = 3\psi_{11}^2\psi_{12}$ . Also, since  $\lambda_{12} = \theta_{11}\theta_{21}\theta_{32} + \theta_{11}\theta_{22}\theta_{31} + \theta_{12}\theta_{21}\theta_{31}$ , it follows from that

$$\psi^{12} = \begin{bmatrix} \psi_{11} & \psi_{11} & \psi_{21} \\ \psi_{12} & \psi_{12} & \psi_{22} \\ \psi_{13} & \psi_{13} & \psi_{23} \end{bmatrix}, \quad S_{11}\{\psi^{12}\} = \psi_{11}^2\psi_{21}, \quad (30)$$

$$S_{12}\{\psi^{21}\} = \psi_{11}^2\psi_{12} + \psi_{11}\psi_{12}\psi_{21} + \psi_{12}\psi_{11}\psi_{21}$$

*S is uniquely invertible:* The fundamental theorem of algebra, namely that an  $L$ -th degree polynomial has up to  $L$  complex valued roots, does not, in general, extend to polynomials in two variables. However, the above special construction which encodes the observations as coefficients of powers of  $t$ , ensures that  $S$  is a uniquely invertible transform between the set of observations and matrix of polynomial coefficients. This is because the ring  $F(s, t)$  of two-variable polynomials is a unique factorization domain over the ring  $F(s)$  of one-variable polynomials [16, Theorem 2.25].

*Evaluating  $S^{-1}$ :* Given the observations  $y$ , the transform  $S\{y\}$  computes the pseudo-observations via convolution (17). We now discuss how to compute  $\theta = S^{-1}(\lambda)$  given  $\lambda$ . This computation is required in (27) to compute  $\theta^*$  and also in the adaptive filtering algorithm (28) below.

As in the scalar case (8), given  $\lambda_{1,1}, \dots, \lambda_{L,1}$ , we first compute  $\theta_{1,1}, \dots, \theta_{L,1}$  by solving for the roots of the polynomial:

$$\prod_{l=1}^L (s + \theta_{l,1}) = s^L + \sum_{l=1}^L \lambda_{l,1} s^{L-l} \quad (31)$$

Next, solve for the remaining elements of  $\theta_{l,m}$  iteratively over  $m = 2, 3, \dots, D$ . For each  $m \geq 2$ , given  $\lambda_{1,m}, \dots, \lambda_{L,m}$  and  $\{\theta_{1,n}, \dots, \theta_{L,n}\}, n = 1, \dots, m-1$ , we solve the following linear system of equations<sup>6</sup> for  $\theta_{1,m}, \dots, \theta_{L,m}$ :

$$\begin{aligned} S_{1,m}\{\theta_{1,m}, \dots, \theta_{L,m}\} &= \lambda_{1,m} \\ S_{2,m}\{\theta_{1,m}, \dots, \theta_{L,m}\} &= \lambda_{2,m} \\ &\vdots \\ S_{L,m}\{\theta_{1,m}, \dots, \theta_{L,m}\} &= \lambda_{L,m} \end{aligned} \quad (32)$$

By the property of elementary symmetric polynomials, the linear system (32) has full rank.

<sup>6</sup>It follows from the definition that  $S_{i,m}$  is linear in  $\theta_{1,m}, \dots, \theta_{L,m}$  with linear coefficients specified by  $\{\theta_{1,n}, \dots, \theta_{L,n}\}, n = 1, \dots, m-1$

To summarize, computing  $S^{-1}$  for the vector case requires solving a single polynomial equation (as in the scalar case) and then  $D-1$  additional linear algebraic equations.

#### D. Convergence of Adaptive Filtering Algorithm and Asymptotic Efficiency

This section analyzes the convergence and asymptotic covariance of the adaptive filtering algorithm (28). The convergence is typically studied via two approaches: mean square convergence and weak convergence (since  $\theta^o$  is assumed to be a constant, weak convergence to  $\theta^o$  is equivalent to convergence in probability). We refer to the comprehensive books [18], [9], [8] for details. Below we state the main convergence result (which follows directly from these references). More importantly, we then discuss the asymptotic efficiency of the adaptive filtering algorithm (28). Specifically we address the question: *How much larger is the asymptotic covariance with the symmetric transform and anonymized observations, compared to the classical LMS algorithm with no anonymization?*

The algorithm (28) can be represented abstractly as

$$\lambda(k+1) = \lambda(k) + \epsilon \Psi(k) (z(k) - \Psi(k)\lambda(k)) \quad (33)$$

where  $\Psi(k)$  is the block diagonal matrix  $\text{diag}(S_{l,m}, l \in [L], m \in M_l)$ .

Let  $\mathcal{F}_k$  be the  $\sigma$ -algebra generated by  $\{\Psi(n), v(n), n < k, \lambda(n), n \leq k\}$ , and denote the conditional expectation with respect to  $\mathcal{F}_k$  by  $\mathbb{E}_k$ . We assume the following conditions:

(A) The signal  $\{\Psi(k), v(k)\}$  is independent of  $\{\lambda(k)\}$ . Either  $\{\Psi(k), v(k)\}$  is a sequence of bounded signals such that there is a symmetric and positive definite matrix  $Q$  such that  $\mathbb{E}\Psi(k)\Psi'(k) = Q$

$$\left| \sum_{n=k}^{\infty} \mathbb{E}_k[\Psi(n)\Psi'(n) - Q] \right| \leq K, \quad \left| \sum_{n=k}^{\infty} \mathbb{E}_k \Psi(n)e(n) \right| \leq K, \quad (34)$$

or  $\{\Psi(k), v(k)\}$  is a sequence of martingale difference signals satisfying  $\mathbb{E}|\Psi(k)|^{4+\Delta} < \infty$  and  $\mathbb{E}|\Psi(k)v(k)|^{2+\Delta} < \infty$  for some  $\Delta > 0$ .

Assumption A includes correlated mixing processes [19, p.345]. and where the remote past and distant future are asymptotically independent. The boundedness is a mild restriction, for example, one may consider truncated processes. Practical implementations of stochastic gradient algorithms use a projection: when the estimates are outside a bounded set  $H$ , they are projected back to the constrained set  $H$ . [9] has extensively discusses such projection algorithms. For unbounded signals, (A) allows for martingale difference sequences.

**Theorem 3** ([9]). *Consider the adaptive filtering algorithm (33). Assume (A). Then*

- 1) (Mean Squared convergence). *For sufficiently large  $k$ , the estimates  $\lambda(k)$  from adaptive filtering algorithm (28) have mean square error  $\mathbb{E}\{\|\lambda(k) - \lambda^o\|^2\} = O(\epsilon)$ .*
- 2) (Convergence in probability)  *$\lim_{\epsilon \downarrow 0} P(\sup_{t \leq T} |\lambda^\epsilon(t) - \lambda^o| > \eta) = 0$  as  $T \rightarrow \infty$  for all  $\eta > 0$ . Here  $\lambda^\epsilon(t) = \lambda(k)$ ,  $t \in [\epsilon k, (\epsilon+1)k)$  denotes the continuous-time interpolated process constructed from  $\lambda(k)$ .*

3) (Asymptotic Normality). As  $k \rightarrow \infty$ , for small  $\epsilon$ , the estimates  $\lambda(k)$  from algorithm (28) satisfy the central limit theorem (where  $\xrightarrow{D}$  denotes convergence in distribution)

$$\epsilon^{-1/2} (\lambda(k) - \lambda^\circ) \xrightarrow{D} \mathbf{N}(0, \Sigma) \quad (35)$$

Here the asymptotic covariance  $\Sigma$  satisfies the algebraic Lyapunov equation

$$Q \Sigma + \Sigma Q = R \quad (36)$$

4) (Asymptotic Covariance of Estimates). Therefore, the estimates  $\theta(k) = S^{-1}(\lambda(k))$  satisfy

$$\begin{aligned} \epsilon^{-1/2} (\theta(k) - \theta^\circ) &\xrightarrow{D} \mathbf{N}(0, \bar{\Sigma}), \\ \bar{\Sigma} &= (\nabla S^{-1}(\lambda^\circ))' \Sigma \nabla S^{-1}(\lambda^\circ). \end{aligned} \quad (37)$$

*Remarks.* (i) Statements 1,2 and 3 of the above result are well known [9]. The expression for  $\bar{\Sigma}$  in (37) follows from the ‘‘delta-method’’ for asymptotic normality [20]. The delta-method requires that  $S^{-1}$  is continuously differentiable. This holds since the solutions of a polynomial equation are continuously differentiable in the coefficients of the polynomial.

(ii) Recall  $\theta(k) = S^{-1}(\lambda(k))$  is a set (and not a vector). So we interpret (37) after ordering the elements in some specific way. In the scalar case, we can impose that the elements are ascending ordered, namely,  $\theta_1 \leq \theta_2 \leq \dots \leq \theta_L$ . For the vector case,  $\theta$  can be ordered such that the first elements of the parameter vector of the  $L$  processes are in ascending order,  $\theta_{11} \leq \theta_{21} \leq \dots \leq \theta_{L1}$ .

(iii) In the stochastic approximation literature [8], [9], the asymptotic rate of convergence is specified in terms of the asymptotic covariance of the estimates, namely  $\Sigma$  in (35) and  $\bar{\Sigma}$  in (37). Since we want to quantify the asymptotic convergence rate, we will focus on evaluating  $\Sigma$  and  $\bar{\Sigma}$ .

### Loss in Efficiency due to Anonymity

We now evaluate the asymptotic covariance matrices  $\Sigma$  in (35) and  $\bar{\Sigma}$  in (37) to quantify the asymptotic rate of convergence of adaptive filtering algorithm (28). To obtain a tractable closed form expression, we consider the scalar observation case  $D = 1$ . So  $\Sigma$  and  $\bar{\Sigma}$  are  $L \times L$  covariance matrices. (Recall there are  $L$  anonymized processes.)

We assume that the zero mean noise process  $v(k)$  is iid across the  $L$  processes with  $\text{Var}\{v_l(k)\} = \sigma_l^2$ . Also we choose the regression input matrix as  $\psi(k) \sim \mathbf{N}(0, I_{L \times L})$ . Using (13) it follows that for  $l \in [L]$ ,

$$Q = \text{diag}[\text{Cov}(\psi^{2l})], \quad \text{Cov}(\psi^{2l}) = (2l-1)(2l-3) \dots 1 \quad (38)$$

Next define  $R_l = \text{Cov}[\psi^l(k)(z_l(k) - \psi^l(k)\lambda_l)]$  evaluated at  $\lambda_l^\circ$ . We have

$$R_l = \text{Cov}(\psi^l(\psi^l(\lambda_l^\circ - \lambda_l) + w_l))|_{\lambda=\lambda^\circ} = \text{Cov}(\psi^l w_l) \quad (39)$$

These can be evaluated using the expression for  $w$  in (59).

Finally from Theorem 6 in Appendix, the sensitivity of the  $l$ -th polynomial root  $\theta_l$  wrt  $m$ -th coefficient  $\lambda_m$  is

$$\nabla S^{-1}(\lambda) = \left[ \frac{d\theta_l}{d\lambda_m} \right], \quad \text{where} \quad \frac{d\theta_l}{d\lambda_m} = \frac{(-1)^{m+1} (-\theta_l)^{L-m}}{\frac{dS\{\theta\}(-\theta)}{d\theta}|_{\theta=\theta_l}} \quad (40)$$

The above formula assumes that the polynomial does not have repeated roots; otherwise the sensitivity is infinite since  $dS(-\theta)/d\theta = 0$  at a repeated root.

With the above characterization of  $Q, R, \nabla S^{-1}(\lambda)$ , we now evaluate  $\Sigma$  and  $\bar{\Sigma}$  explicitly for  $L = 2$ .

**Lemma 1.** Consider the anonymized model (1), (2) with  $D = 1, L = 2$ . Assume the zero mean noise process  $v(k)$  is iid across the  $L$  processes with  $\text{Var}\{v_l(k)\} = \sigma^2$ , and  $\psi(k) \sim \mathbf{N}(0, I_{L \times L})$ . Then the asymptotic covariance  $\bar{\Sigma}$  (see (37)) of the estimates  $\theta(k)$  generated by algorithm (13) satisfies

$$\text{Tr}(\bar{\Sigma}) = \frac{6\sigma^2(\theta_1^2 + \theta_2^2) + \sigma^4}{(\theta_1 - \theta_2)^2} \quad (41)$$

*Remark.* From (41),  $\inf \text{Tr}(\bar{\Sigma}) = 3\sigma^4$  when  $\theta_1 = -\theta_2 \rightarrow \infty$ . In comparison, for the classical LMS algorithm when the observations are not anonymized, the asymptotic covariance for  $D = 2$  is  $\text{Tr}(\text{Cov}(\text{LMS})) = \sigma^2$ . So for  $D = 2$ , at best, the adaptive filtering algorithm (13) with anonymized observations is 3 times less efficient than the classical LMS.

*Proof.* From (38),  $Q = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$ . Also (39) yields  $R = \text{Cov} \begin{bmatrix} \psi(v_1 + v_2) \\ \psi^2(\psi v_2 \theta_1 + \psi v_1 \theta_2 + v_1 v_2) \end{bmatrix} = \text{diag}(2\sigma^2, 15\sigma^2(\theta_1^2 + \theta_2^2) + 3\sigma^4)$ . Finally (40) yields  $\nabla S^{-1}(\lambda) = \begin{bmatrix} \frac{\theta_1}{\theta_1 - \theta_2} & \frac{\theta_2}{\theta_2 - \theta_1} \\ \frac{\theta_1}{\theta_2 - \theta_1} & \frac{\theta_2}{\theta_1 - \theta_2} \end{bmatrix}$ . Then evaluating  $\Sigma = \frac{1}{2} Q^{-1} R$ , and  $\bar{\Sigma}$  using (37) yields (41).  $\square$

To summarize, Lemma 1 shows that for  $D = 2$ , at best, adaptive filtering with anonymized data has three times the asymptotic variance compared to the classical LMS algorithm.

### E. Analysis for Tracking a Markov hyper-parameter

So far we assumed that the true parameter  $\theta^\circ$  was constant. An important property of a constant step size adaptive filtering algorithm (28) is the ability to track a time evolving true parameter. Suppose the true parameter  $\theta^\circ(k)$  evolves according to a slow Markov chain with unknown transition matrix. How well does the adaptive filtering algorithm track (estimate)  $\theta^\circ(k)$ ? Our aim is to quantify the mean squared tracking error.

(B) Suppose that exists a small parameter  $\mu > 0$  and  $\theta^\circ(k)$  is a discrete-time Markov chain, whose state space is

$$M_l = \{a_1, \dots, a_m\}, \quad a_i \in \mathbb{R}^{L \times D}, \quad i = 1, \dots, m, \quad (42)$$

and whose transition probability matrix  $P^\mu = I + \mu Q$ , where  $I$  is an  $\mathbb{R}^{m \times m}$  identity matrix and  $Q = (q_{ij}) \in \mathbb{R}^{m \times m}$  is an irreducible generator (i.e.,  $Q$  satisfies  $q_{ij} \geq 0$  for  $i \neq j$  and  $\sum_{j=1}^m q_{ij} = 0$  for each  $i = 1, \dots, m$ ) of a continuous-time Markov chain.

The time evolving parameter  $\theta^\circ(k)$  is called a hyperparameter. Although the dynamics of the hyperparameter  $\theta^\circ(k)$  are used in our analysis below, the implementation of the adaptive filtering algorithm (13), does not use this information.

Define the tracking error of the adaptive filtering algorithm (28) as  $\tilde{\lambda}(k) \stackrel{\text{defn}}{=} \lambda(k) - \lambda^\circ(k)$ . The aim is to determine bounds on the tracking error  $\tilde{\lambda}(k)$  and therefore  $\theta(k)$ .

**Theorem 4.** Under (A), (B), for sufficiently large  $k$ ,

$$\mathbb{E}|\tilde{\lambda}(k)|^2 = O(\epsilon + \mu + \mu^2/\epsilon) \quad (43)$$

Therefore, choosing  $\mu = O(\epsilon)$ , the mean squared-tracking error is  $\mathbb{E}|\tilde{\lambda}(k)|^2 = O(\epsilon)$  and so  $\mathbb{E}|\hat{\theta}(k)|^2 = O(\epsilon)$

The proof follows from [21]. The theorem implies that even if the hyperparameter  $\theta^o$  evolves on the same time scale (speed) as the adaptive filtering algorithm, the algorithm can track the hyperparameter with mean squared error  $O(\epsilon)$ .

#### IV. MIXTURE MODEL FOR ANONYMIZATION

This section uses a Bayesian interpretation of the anonymity map  $\sigma$  in (2) to present a performance analysis of the adaptive filtering algorithm (28). Thus far we have assumed nothing about the permutation (anonymization) process  $\sigma$  in (2). The symmetric transform based algorithms proposed in Sections II and III are oblivious to any assumptions on  $\sigma$ . Below we formulate a probabilistic model for the permutation process  $\sigma$ . Based on this probabilistic model, we address two questions:

- 1) *How do noisy observations of the permutation process affect anonymity of the identity of the target processes?* We will consider the expected error probability of the maximum posterior estimate of the permutation process as a measure of the anonymity of the permutation process. This is in line with [10] where the error probability of an estimator (and also mutual information) is used as a measure of anonymity.
- 2) *How does anonymity of process  $\sigma$  in terms of Bayesian error probabilities relate to the asymptotic covariance of the adaptive filtering algorithm (28)?* Our main result below (Theorem 5) shows that if the observation likelihood of noise process one Blackwell dominates that of noise process two, then anonymity of the permutation process two is higher than that of one; and also the asymptotic covariance of the parameter estimates of the adaptive filtering algorithm (28) is higher.

From a probabilistic point of view, the anonymized observation model (1), (2) can be constructed as the following random permutation mixture model of the rows of matrix  $\theta^o$ :

$$\underset{L \times D}{y(k)} = \underset{L \times L}{\sigma(x(k))} \underset{L \times D}{\mathbf{y}(k)} = \underset{L \times L}{\sigma(x(k))} \underset{L \times D}{\theta^o} \underset{D \times D}{\psi(k)} + \underset{L \times D}{\sigma(x(k))} \underset{L \times D}{v(k)} \quad (44)$$

Here  $\sigma(x(k))$  denotes a randomly chosen  $L \times L$  permutation matrix that evolves according to a random process

$$x \in \mathcal{X}, \quad \mathcal{X} \subseteq \{1, 2, \dots, X\} \text{ where } X = L!$$

since there are  $L!$  possible permutations. Also  $y(k) = [y_1(k), \dots, y_L(k)]'$  where each  $y_l(k) \in \mathbb{R}^D$ . Recall that  $\psi(k)$  is a known input (regression) matrix and  $v(k) = [v_1(k), \dots, v_L(k)]'$  is a  $L \times D$  matrix valued noise process whose elements are zero mean. As previously, we assume for simplicity that  $v_l(k)$  and  $v_m(k)$ ,  $l \neq m$  are iid vectors in  $\mathbb{R}^D$ .

##### A. Anonymity of Permutation Process $x$ and asymptotic covariance of adaptive filtering algorithm

This section characterizes the anonymity of the permutation process  $x$  in terms of average error probability of the *maximum a posteriori* (MAP) state estimate. Our assumptions are:

- 1) The permutation process  $x$  is iid with known probabilities  $\pi(i) \stackrel{\text{defn}}{=} P(x(k) = i)$ .
- 2) The regression matrix  $\psi(k) = I$ . From a Bayesian point of view, this is without loss of generality since  $\psi(k)$  is known and invertible. So we can post-multiply (44) by  $\psi^{-1}(k)$  to obtain an equivalent observation process.

Given the observation model (44), define the  $L \times D$ -variate observation likelihood given state  $x(k) = i$  as

$$B_{iy} = p(y(k) = y | x(k) = i) \propto p_v(y - q_i), \quad (45)$$

where  $q_i \stackrel{\text{defn}}{=} \sigma(i) \theta^o \in \mathbb{R}^{L \times D}$

Here  $p_v$  denotes the  $L \times D$ -variate density of noise process  $v$ . Since the  $L$  noise processes are independent, with  $y_l = y' e_l$  where  $e_l \in \mathbb{R}^L$  is the unit vector with 1 in the  $l$ -th position,

$$B_{iy} = \prod_{l=1}^L B_{iy_l}, B_{iy_l} = \prod_{m=1}^D B_{i,y_l,m}, B_{i,y_l,m} = p_{v_{lm}}(y_{lm} - \theta_{lm}^o) \quad (46)$$

The anonymity of the  $x$  depends on the prior  $\pi$  of the permutation process  $x$  and the observation likelihood  $B$ .

*Perfect Anonymity.* If all  $X$  permutations are equi-probable, i.e.,  $\pi(i) = 1/X$ , then clearly  $P(x(k) = i | y(k)) = 1/X$ . So the probability of error of the maximum a posteriori estimate  $\hat{x}_k$  is  $P(\hat{x}_k \neq x(k)) = (X-1)/X$  which is the largest possible value. So for discrete uniform prior on the permutation process, perfect anonymity of the identities of the  $L$  processes holds (even with no measurement noise).

*Zero Anonymity.* If  $\pi(x) = 1$  for some state  $x = i^*$ , then the error probability is zero and there is no anonymity.

*Anonymity of Permutation Process  $x$  wrt observation likelihood:* In the rest of this section, we analyze the anonymity of a Bayesian estimator of the permutation process  $x$  in terms of the observation likelihood  $B$ , or equivalently, the noise  $v(k)$ . We start with Bayes formula for the posterior of permutation state  $x(k)$  given observation  $y(k)$ . Define the diagonal matrix  $B_y = \text{diag}[B_{1y}, \dots, B_{Xy}]$ . Then given the prior  $\pi$  and observation  $y(k)$ , the posterior  $\pi(k) = [\pi_1(k), \dots, \pi_X(k)]'$  where  $\pi_i(k) = p(x(k) = i | y(k))$  is given by Bayes formula:

$$\pi(k) = T(\pi, y(k)) \stackrel{\text{defn}}{=} \frac{B_{y(k)} \pi}{\sigma(\pi, y(k))}, \text{ where } \sigma(\pi, y) = \mathbf{1}' B_{y(k)} \pi \quad (47)$$

Finally, given the posterior computed by (47), define the maximum a posteriori (MAP) permutation state estimate as

$$\hat{x}(k) = \arg \max_i \pi_i(k)$$

**Lemma 2.** *The expected error probability of the MAP state estimate is (where  $\mathcal{Y}$  below denotes the observation space)*

$$P_e(\pi; B) = \mathbb{E}_y \{P(x(k) \neq \hat{x}(k) | y)\} = 1 - \int_{\mathcal{Y}} \max_i e'_i B_y \pi dy$$

where  $e_l \in \mathbb{R}^X$  is the unit vector with 1 in the  $i$ -th position.

We normalize the expected error probability by defining the anonymity of permutation process  $x$  as

$$\mathcal{A}(\pi, B) = P_e(\pi; B) \frac{X}{X-1} \in [0, 1] \quad (48)$$

So the anonymity  $\mathcal{A} = 0$  when  $P_e(\pi; B) = 0$ , and  $\mathcal{A} = 1$  when  $P_e(\pi; B) = \frac{X-1}{X}$ .

## B. Blackwell Dominance and Main Result

We now use a novel result involving Blackwell dominance of mean preserving spreads to relate the anonymity to the covariance of adaptive filtering algorithm (28).

**Definition 1** (Blackwell ordering of stochastic kernels). *We say that likelihood  $B$  Blackwell dominates likelihood  $\bar{B}$ , i.e.,  $B \geq_B \bar{B}$  if  $\bar{B} = BM$  where  $M$  is a stochastic kernel. That is,  $\int_{\mathcal{Y}} M_{\bar{y},y} dy = 1$  and  $M_{\bar{y},y} \geq 0$ .*

Intuitively  $\bar{B}$  is noisier than  $B$ . Thus observation  $y$  with conditional distribution specified by  $B$  is said to be more informative than (Blackwell dominates) observation  $\bar{y}$  with conditional distribution  $\bar{B}$ , see [2] for several applications. When  $y$  belongs to a finite set, it is well known [23] that  $B \geq_B \bar{B}$  implies that  $\bar{B}$  has smaller Shannon capacity than  $B$ .

*Main Result:* First we list the main assumptions:

- (A1)  $B \geq_B \bar{B}$
  - (A2)  $\int_{\mathcal{Y}} B_{iy} y dy = q_i$  and  $\int_{\mathcal{Y}} \bar{B}_{iy} y dy = q_i$  (zero mean noise)
- Recall  $q_i$  is defined in (45).

Since the observations of the  $L$  processes are independent, Blackwell dominance of the  $l$  individual likelihoods  $B_{iy_l} \geq_B \bar{B}_{iy_l}$ ,  $l \in [L]$  is sufficient for (A1). The mean preserving spread assumption (A2) on  $B$  and  $\bar{B}$  implies that the observation noise is zero mean. This is a classical assumption for the convergence of the stochastic gradient algorithm (28).

We are now ready to state the main result. Theorem 5 shows that Blackwell ordering of observation likelihoods yields an ordering for error probabilities (anonymity) and also a partial ordering on the asymptotic covariance matrices of the adaptive filtering algorithm (28). So the more the anonymity of the permutation process, the higher the asymptotic covariance of the adaptive filtering algorithm (28). To the best of our knowledge, this result is new.

**Theorem 5.** *Consider observations  $y(k)$  generated by (44).*

- 1)  $\text{Cov}_B(y) \preceq \text{Cov}_{\bar{B}}(y)$  implies  $\text{Cov}_B(S\{y\}) \preceq \text{Cov}_{\bar{B}}(S\{y\})$  for the symmetric transform  $S$ .
- 2) Assume (A1). Then the average error probabilities satisfy  $P_e(\pi; B) \leq P_e(\pi; \bar{B})$ , and therefore the anonymity satisfies  $\mathcal{A}(\pi, B) \leq \mathcal{A}(\pi, \bar{B})$ .
- 3) Assume (A1), (A2). Then  $\text{Cov}_B(y) \preceq \text{Cov}_{\bar{B}}(y)$ . Therefore, the asymptotic covariance of  $\lambda(k)$  in (35) of the adaptive filtering algorithm satisfies  $\Sigma(B) \leq \Sigma(\bar{B})$ . Also the asymptotic covariance of  $\theta(k)$  in (37) satisfies  $\bar{\Sigma}(B) \leq \bar{\Sigma}(\bar{B})$ .

The proof in the appendix uses mean-preserving convex dominance from Blackwell's classic paper [5]. Note that Theorem 5 does not require the noise to be Gaussian; for example, the noise can be finite valued random variables.

To summarize, we have linked anonymity of the observations (error probability of the Bayesian MAP estimate) to the asymptotic covariance (convergence rate) of the adaptive filtering algorithm (28).

## V. NUMERICAL EXAMPLES

*Example 1: Symmetric Transform for Scalar case  $D = 1$ :*

The aim of this example is to show that objective (10) has local

minima wrt  $\theta$ ; and therefore the classical stochastic gradient algorithm (15) gets stuck in a local minimum. In comparison, the objective (12) in terms of pseudo-measurements is convex (quadratic) wrt  $\lambda$  and therefore the adaptive filtering algorithm (13) converges to the global minimum  $\theta^*$ .

We consider  $L = 3$  independent scalar processes ( $D = 1$ ) with anonymized observations generated as in (2). The true model that generates the observations is  $\theta^o = [-2, 5, 8]'$ . The regression signal  $\psi(k) \sim \mathbf{N}(0, \sigma^2)$  where  $\sigma = 1$ . The noise error  $v(k) \sim \mathbf{N}(0, \sigma_v^2)$  where  $\sigma_v = 10^{-2}$ .

We ran the adaptive filtering algorithm (13) on a sample path of  $2 \times 10^5$  anonymized observations generated by the above model with step size  $\epsilon = 10^{-4}$ . For initial condition  $\theta(0) = [1, 2, 3]'$ , Figure 2a shows that the estimates generated by Algorithm (13) converges to  $\theta^o$ . As can be seen from Figure 2a, the sample path of the estimates initially are coalesced, and then split. This is because the estimates  $\theta_1(k)$  and  $\theta_2(k)$  are initially complex conjugates; since we plot the real parts, the estimates of  $\theta_1(k)$  and  $\theta_2(k)$  are identical.

We also ran the classical stochastic gradient algorithm (15) on the anonymized observations. Recall this algorithm minimizes (10) directly. The step size chosen was  $\epsilon = 10^{-7}$  (larger step sizes led to instability). For initial condition  $\theta(0) = [1, 2, 3]'$ , Figure 2(b) shows that the estimates converge to a local stationary point  $[-2.02, 6.12, 6.45]'$  which is not  $\theta^o$ . On the other hand for initial condition  $\theta(0) = [3, 6, 9]'$ , we found that the estimates converged to  $\theta^o$ . This provides numerical verification that objective (10) is non-convex. Besides the non-convex objective, another problem with the algorithm (15) is that if we choose  $\theta(0) = [c, c, c]$  for any  $c \in \mathbb{R}$ , then all elements of  $\theta(k)$  are identical, regardless of  $\theta^o$ .

There are two takeaways from this numerical example. First, despite the anonymization, one can still consistently estimate the true parameter set  $\theta^o$ . Second, the objective (10) is non-convex in  $\theta$  but convex - so a classical stochastic gradient algorithm can get stuck in a local minimum. But since the objective is convex in the polynomial coefficients  $\lambda$ , which are constructed as pseudo-observations via the symmetric transform, algorithm (13) converges to the global minimum.

*Example 2: Recursive Maximum Likelihood vs Symmetric Transform:* The recursive EM algorithm ((55) (REM) in Appendix) requires knowledge of the noise distribution and probabilities of permutation process  $x$ . When these are known, REM performs extremely well. But in the *mis-specified case*, where the assumed noise distribution is different to the actual distribution, REM can yield a significant bias in the estimates.

We simulated anonymized observations (1), (2) for  $D = 1$ ,  $L = 2$  with zero mean iid Laplacian noise  $v$  with standard deviation 2. The true parameter is  $\theta^o = [4, 5]$  for  $k \leq 3 \times 10^5$  time points and then changes to  $[1, 3]$ . We ran REM (55) assuming unit variance Gaussian noise. The step size  $\epsilon = 5 \times 10^{-5}$  and initial estimate  $\theta(0) = [1, 2]'$ . Figure 3a shows that the algorithm yields a significant bias in the estimate for  $\theta^o$ ; the estimates  $\theta(k)$  converge to  $[3.5590, 5.4559]'$  for the first  $3 \times 10^5$  points and then to  $[0.7405, 3.2658]'$ .

We then computed the pseudo-observations (7) using the scalar symmetric transform (11) and ran the adaptive filtering algorithm (13) with step size  $\epsilon = 2 \times 10^{-5}$  and initial condition

$\theta(0) = [1, 2]'$ . Figure 3b displays the sample path estimates  $\theta(k)$ . We see empirically that the convergence of adaptive filtering algorithm is slower than the recursive EM, but the estimates converge to the true parameter  $\theta^\circ$  (with no bias).

*Example 3: Symmetric Transform for Vector case.*  $D = 2$ ,  $L = 2$ : We consider  $L = 2$  independent processes each of dimension  $D = 2$  with anonymized observations generated by (2). The true models that generate the observations for the two independent processes via (1) are  $\theta_1^\circ = [-2, 6]'$ ,  $\theta_2^\circ = [4, 5]'$ . The  $2 \times 2$  input regression matrix in (1) was chosen with iid elements  $\psi_{ij}(k) \sim \mathcal{N}(0, 1)$ . The 2-dimensional noise error vector  $v(k)$  has iid elements  $\mathcal{N}(0, \sigma_v^2)$  where  $\sigma_v = 10^{-1}$ .

Given the anonymized observations, we constructed the pseudo-observations using the vector symmetric transform (16). We ran the adaptive filtering algorithm (28) with step size  $\epsilon = 10^{-4}$  on these pseudo-observations. Figure 4a shows that the estimates converge to the true model set  $\theta^\circ$ .

Next we constructed the naive pseudo observations from the anonymized observations by using the naive transform  $\bar{S}$  (21). We then ran the adaptive filtering algorithm (28) with step size  $\epsilon = 10^{-5}$  on these naive pseudo-observations. We see from Figure 4b that the estimates converge to  $\{[-2, 5]', [4, 6]'\}$  instead of the true model set  $\{[-2, 6]', [4, 5]'\}$ . So naively applying the scalar symmetric transform element-wise can result in estimates that swap the elements of  $\theta^\circ$ . In comparison, the vector symmetric transform together with algorithm (28) yield consistent estimates of  $\theta^\circ$ .

*Example 4: Mid-sized Example:* In Appendix G (supplementary document) we consider the case  $L = 4$  and  $D = 10$ . We show that algorithm (28) successfully estimates the parameters. In comparison, the naive symmetric transform loses ordering information resulting in ghost process estimates.

## VI. CONCLUSIONS

We proposed a symmetric transform based adaptive filtering algorithm for parameter estimation when the observations are a set (unordered) rather than a vector. Such observation sets arise due to uncertainty in sensing or deliberate anonymization of data. By exploiting the uniqueness of factorization over polynomial rings, Theorems 1 and 2 showed that the adaptive filtering algorithms converge to the true parameter (global minimum). Lemma 1 characterized the loss in efficiency due to anonymization by evaluating the asymptotic covariance of the algorithm via the algebraic Liapunov equation. Theorem 4 characterized the mean squared error when the underlying true parameter evolves over time according to an unknown Markov chain. Finally Theorem 5 related the asymptotic covariance (convergence rate) of the adaptive filtering algorithm to a Bayesian interpretation of anonymity of the observations via mean preserving Blackwell dominance.

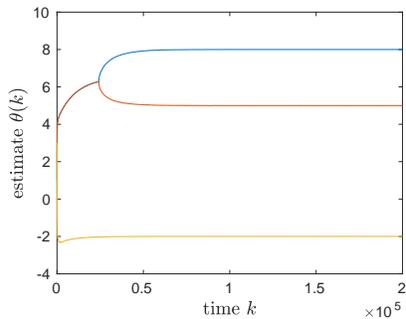
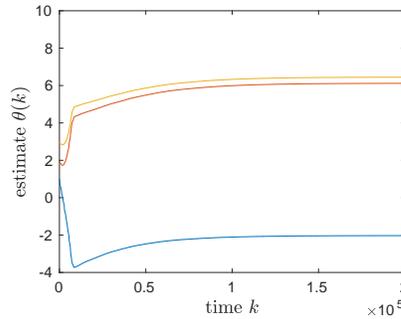
The tools used in this paper, namely symmetric transforms to circumvent data association, polynomial rings to characterize the attraction points of an adaptive filtering (stochastic gradient) algorithm, and Blackwell dominance to relate a Bayesian interpretation of anonymity to the convergence rate of the adaptive filtering algorithm, can be extended to other formulations. In future work, it is worth addressing distributed

methods for learning with unlabeled data, for example, [3] proposes powerful distributed methods. Also the effect of quantizing the anonymized data can be studied using [26].

**Supplementary Document.** The supplementary document contains all proofs, additional simulation examples and description of a recursive maximum likelihood estimator for  $\theta^\circ$ .

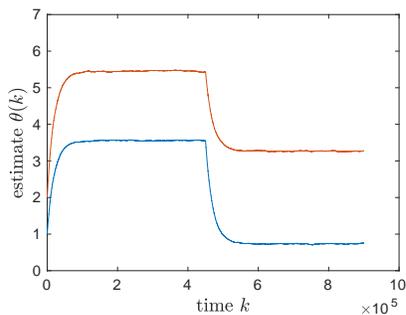
## REFERENCES

- [1] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation*. New York: John Wiley, 2008.
- [2] S. Marano, V. Matta, and P. Willett, "Some approaches to quantization for distributed estimation with data association," *IEEE transactions on signal processing*, vol. 53, no. 3, pp. 885–895, 2005.
- [3] Y. Bar-Shalom, T. E. Fortmann, and P. G. Cable, "Tracking and data association," 1990.
- [4] E. W. Kamen, "Multiple target tracking based on symmetric measurement equations," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, pp. 371–374, 1992.
- [5] E. W. Kamen and C. R. Sastry, "Multiple target tracking using products of position measurements," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 29, no. 2, pp. 476–493, 1993.
- [6] U. D. Hanebeck, M. Baum, and P. Willett, "Symmetrizing measurement equations for association-free multi-target tracking via point set distances," in *Signal Processing, Sensor/Information Fusion, and Target Recognition XXVI*, vol. 10200. SPIE, 2017, pp. 48–61.
- [7] D. J. Muder and S. D. O'Neil, "Multidimensional sme filter for multitarget tracking," in *Signal and Data Processing of Small Targets 1993*, vol. 1954. SPIE, 1993, pp. 587–599.
- [8] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, ser. Applications of Mathematics. Springer-Verlag, 1990, vol. 22.
- [9] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Recursive Algorithms and Applications*, 2nd ed. Springer-Verlag, 2003.
- [10] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching anonymized and obfuscated time series to users' profiles," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 724–741, 2019.
- [11] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.
- [12] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.
- [13] T. J. Lampoltshammer, L. Thurnay, G. Eibl *et al.*, "Impact of anonymization on sentiment analysis of twitter postings," in *Data Science—Analytics and Applications*. Springer, 2019, pp. 41–48.
- [14] I. Lourenço, R. Mattila, C. R. Rojas, and B. Wahlberg, "Cooperative system identification via correlative learning," *IFAC-PapersOnLine*, vol. 54, no. 7, pp. 19–24, 2021.
- [15] I. G. Macdonald, *Symmetric functions and Hall polynomials*. Oxford university press, 1998.
- [16] N. Jacobson, *Basic algebra I*. Courier Corporation, 2012.
- [17] B. Sagan, *The symmetric group: representations, combinatorial algorithms, and symmetric functions*. Springer Science & Business Media, 2001, vol. 203.
- [18] A. Sayed, *Adaptive Filters*. Wiley, 2008.
- [19] S. N. Ethier and T. G. Kurtz, *Markov Processes—Characterization and Convergence*. Wiley, 1986.
- [20] A. W. Van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 2000, vol. 3.
- [21] G. Yin and V. Krishnamurthy, "Least mean square algorithms with Markov regime switching limit," *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 577–593, May 2005.
- [22] V. Krishnamurthy, *Partially Observed Markov Decision Processes. From Filtering to Controlled Sensing*. Cambridge University Press, 2016.
- [23] C. E. Shannon, "A note on a partial ordering for communication channels," *Information and control*, vol. 1, no. 4, pp. 390–397, 1958.
- [24] D. Blackwell, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, pp. 265–272, 1953.
- [25] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—part ii: Polynomial escape from saddle-points," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1257–1270, 2021.
- [26] L. Wang and G. Yin, "Asymptotically efficient parameter estimation using quantized output observations," *Automatica*, vol. 43, no. 7, pp. 1178–1191, 2007.

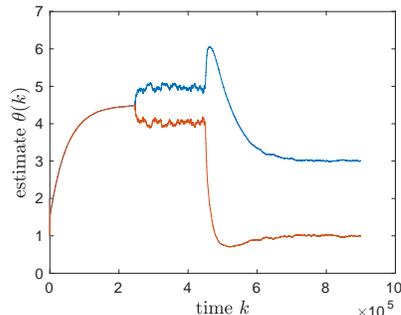
(a) Algorithm (13) converges to global optimum  $\theta^o$ .

(b) Classical stochastic gradient algorithm (15) gets stuck in local minimum.

Fig. 2: Anonymized estimation problem in Example 1 of Sec V. The initial condition is  $\theta(0) = [1, 2, 3]'$  and the true parameter is  $\theta^o = [-2, 5, 8]'$ . Fig.2a shows that the parameter estimates generated by Algorithm (13) converge to  $\theta^o$ . Fig.2b shows that the parameter estimates generated by stochastic gradient algorithm (15) operating on (10) do not converge to  $\theta^o$ .



(a) Recursive EM Algorithm (55).



(b) Adaptive Filtering algorithm (13).

Fig. 3: Recursive Expectation Maximization algorithm vs Symmetric Transform based Adaptive Filtering algorithm. Both algorithms operate on anonymized observations (1), (2) corrupted by Laplacian noise. The true parameter is  $\theta^o = [4, 5]'$ . The recursive EM shows a significant bias in the mis-specified case; in comparison the symmetric transform based algorithm converges to the true parameter value but the convergence is slower. The parameters are specified in Example 2.

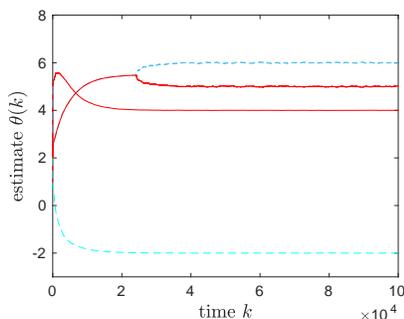
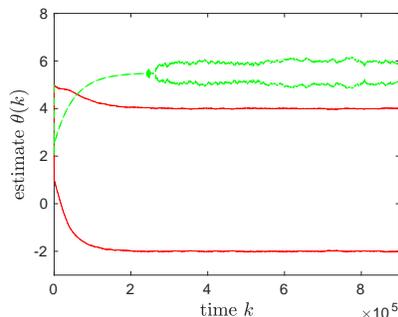
(a) Algorithm (28) operating on vector symmetric transforms converges to global optimum  $\theta^o$ .(b) Algorithm (28) operating on pseudo-observations generated by the naive symmetric transform (21) converges to the ghost process parameters  $\{[-2, 5]', [4, 6]'\}$  instead of the true model set  $\{[-2, 6]', [4, 5]'\}$ .

Fig. 4: Anonymized estimation problem in Example 3 of Sec. V.

## APPENDIX

**Supplementary Document**

Adaptive Filtering Algorithms for Set-Valued  
Observations–Symmetric Measurement Approach to  
Unlabeled and Anonymized Data  
by Vikram Krishnamurthy

**Abstract**—This supplementary document contains:

- 1) **Description of the maximum likelihood estimator of the parameter  $\theta^\circ$  via a recursive Expectation Maximization algorithm in Sec. A.**
- 2) **Proofs of the theorems stated in the main paper.**
- 3) **A medium sized simulation example illustrating the adaptive filtering algorithm and ghost processes in Sec.G**
- 4) **Example of the symmetric transform  $S$  for  $D = 3, L = 3$  in Sec.H.**

### A. Maximum Likelihood Estimation

This section discusses maximum likelihood (ML) estimation of  $\theta^\circ$  given observations generated by (1), (2). The results of this section are not new - they are used to benchmark the symmetric transform based algorithms derived in the paper.

To give some context, we mentioned in the Introduction that given an observation set  $y$  (instead of a vector), feeding it in an arbitrary order into a bank of LMS algorithms will not converge to  $\theta^\circ$  in general. A more sophisticated approach is to order the elements of the observation set at each time based on an estimate of the permutation map  $\sigma_k$ . We can interpret the recursive MLE algorithm below as computing the posterior of  $\sigma_k$  and then feeding it into a stochastic gradient algorithm.

Before proceeding it is worthwhile to summarize the disadvantages of the MLE approach of this section:

- 1) The density function of the noise process  $v$  in (1) and the probability law of the random process  $x$  in (44) need to be known. For example if  $x$  was an iid process, the in principle one can recursively estimate the probabilities of  $x$ . However if  $x$  is an arbitrary non-stationary process, then the MLE approach is not useful.
- 2) The state space dimension of  $x$  is  $L!$ , i.e., factorial in the number of processes  $L$ . In comparison, for the symmetric function approach, the number of coefficients of the symmetric transform polynomial is  $O(L^2)$ , see (16).
- 3) The likelihood is not necessarily concave in  $\theta$  and so computing the global maximum of the likelihood can be intractable. However, when  $v$  in (1) is Gaussian, then (44), (49), imply that the likelihood is concave in  $\theta$ .
- 4) Why not use the MLE approach together with the symmetric transform? This is not tractable since after applying the symmetric transform, the noise distribution has complicated form (63) that is not amenable to MLE.

We assume that the permutation process  $x$  in (44) is an  $L!$  Markov chain with known transition matrix

$$P(x(k+1) = j | x(k) = i) = P_{ij}, \quad i, j \in \mathcal{X} \quad (49)$$

Then (44) is a Hidden Markov model (HMM) or dynamic mixture model. Notice that the matrix valued observations  $y(k)$  are generated as random (Markovian) permutations of the rows of matrix  $\theta^\circ \psi(k)$  corrupted by noise. Given these observations, the aim is to estimate the matrix  $\theta^\circ$ .

In this section, our aim is to compute the MLE for  $\theta^\circ$ . Given  $N$  data points, the MLE is defined as  $\hat{\theta} = \arg \sup_{\theta \in \Theta} \log p(y(1), \dots, y(N); \theta)$ . We assume that  $\Theta$  is a compact subset of  $\mathbb{R}^{L \times D}$  and so the MLE is

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log p_\theta(y(1:N)), \quad (50)$$

where  $y(1:N) \stackrel{\text{defn}}{=} (y(1), \dots, y(N))$

Under quite general conditions the MLE  $\hat{\theta}$  of a HMM is strongly consistent (converges w.p.1 to  $\theta^\circ$ ) and efficient (achieves the Cramer-Rao lower bound), see [1].

*Remark.* With suitable abuse of notation, note that  $y(k)$  in (44) is a matrix, whereas  $y(k)$  in (2) is a set. In the probabilistic setting that we now consider, this distinction is irrelevant. For example, we could have denoted the anonymization operation (2) as choosing amongst the permutation matrices with equal probability  $1/L!$ . In the symmetric transform formulation in previous sections, we did not impose assumptions on how the elements of the observation set are permuted; the algorithm (28) was agnostic to the order of the elements in the set  $y(k)$ . In comparison, in this section we postulate that the Markov process  $x$  permutes the observations.

*Expectation Maximization (EM) Algorithm:* The process  $x$  is the latent (unobserved) data that permutes the observations from the  $L$  processes yielding the matrix  $y(k)$  in (44). The Expectation Maximization (EM) algorithm is a convenient numerical method for computing the MLE when there is latent data. Starting with an initial estimate  $\theta^0$ , the EM algorithm iteratively generates a sequence of estimates  $\theta^i$ , where each iteration  $i = 1, 2, \dots$  comprises two steps:

*Step 1. Expectation step:* Compute the auxiliary likelihood

$$Q(\theta, \theta^i) \stackrel{\text{defn}}{=} \mathbb{E}\{\log p_\theta(y(1:N), x(1:N) | y(1:N), \theta^i)\} \quad (51)$$

where  $y(1:N) = (y(1), \dots, y(N))$  and  $x(1:N) = (x(1), \dots, x(N))$ . In our case, from (1), (44), (49), imply

$$Q(\theta, \theta^i) = \sum_{k=1}^N \sum_{i=1}^X \pi_i(k|N) \log p_v(y(k) - \sigma(i) \psi(k) \theta) \quad (52)$$

The smoothed probabilities  $\pi_i(k|N)$  are computed using a forward backward algorithm [2]; we omit details here.

*Step 2. Maximization step:* Compute  $\theta^{i+1} = \arg \max_\theta Q(\theta, \theta^i)$ .

Under mild continuity conditions of  $Q(\theta, \theta^i)$  wrt  $\theta$ , it is well known [4] that the EM algorithm climbs the likelihood surface and converges to a local stationary point  $\theta^*$  of the log likelihood  $\log p(y(1), \dots, y(N); \theta)$ .

*Recursive EM Algorithm for Anonymized Observations – IID Permutations:* We are interested in sequential (on-line) estimation that generates a sequence of estimates  $\theta(k)$  over time  $k$ . So we formulate a recursive (on-line) EM algorithm. In the numerical examples presented in Sec. V, we will consider the case where permuting process  $x$  is iid with  $\pi(i) \stackrel{\text{defn}}{=} P(x(k) = i)$ , rather than a more general Markov chain. (Recursive EM algorithms can also be developed for HMMs, but the convergence proof is more technical.)

Since  $x$  and  $y$  are iid processes, assuming  $|\mathbb{E}_{\theta^\circ}\{\mathbb{E}\{\log p_\theta(y(k), x(k))|y(k), \bar{\theta}\}\}| < \infty$ , it follows from Kolmogorov's strong law of large numbers that

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} Q(\theta, \bar{\theta}) &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \mathbb{E}\{\log p_\theta(y(k), x(k))|y(k), \bar{\theta}\} \\ &= \mathbb{E}_{\theta^\circ}\{\mathbb{E}\{\log p_\theta(y, x)|y, \bar{\theta}\}\} \quad \text{w.p.1} \end{aligned} \quad (53)$$

The recursive EM algorithm is a stochastic gradient ascent algorithm that operates on the above objective:

$$\theta(k+1) = \theta(k) + \epsilon \nabla_\theta \mathbb{E}\{\log p_\theta(y, x)|y, \theta(k)\}|_{\theta=\theta(k)} \quad (54)$$

where  $\epsilon > 0$  is a constant step size. Then starting with initial estimate  $\theta(0)$ , the recursive EM algorithm generates estimates  $\theta(k)$ ,  $k = 1, 2, \dots$ , as follows:

$$\begin{aligned} \theta(k+1) &= \theta(k) + \epsilon \sum_{i \in \mathcal{X}} \pi_i(k) \nabla_\theta [\log p_v(y(k) - \sigma(i) \psi(k) \theta(k))] \\ \pi_i(k) &\propto \pi_0(i) p_v(y(k) - \sigma(i) \psi(k) \theta(k)) \end{aligned} \quad (55)$$

So (55) uses a weighted combination of the posterior probability of all possible permutations to scale the gradient of the auxiliary likelihood  $Q$ ; and these scaled gradients are used in the stochastic gradient ascent algorithm.

*Remark.* Let us explain the rationale behind the recursive EM algorithm (55). First, assuming  $\mathbb{E}_{\theta^\circ}|\log p_\theta(y)| < \infty$ , it follows by Kolmogorov's strong law of large numbers that the log likelihood satisfies

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log p_\theta(y(1:N)) = \mathbb{E}_{\theta^\circ}\{\log p_\theta(y)\} \quad \text{w.p.1} \quad (56)$$

Next Fisher's identity relates the gradient of the log likelihood to that of the auxiliary likelihood  $Q$ :

$$\nabla_\theta \log p_{\bar{\theta}}(y(1:N)) = \nabla_\theta Q(\theta, \bar{\theta})|_{\theta=\bar{\theta}}$$

Thus for  $N \rightarrow \infty$ , it follows from Fisher's identity, (53) and (56) that

$$\nabla_\theta \log p_\theta(y)|_{\bar{\theta}} = \nabla_\theta \mathbb{E}\{\log p_\theta(y, x)|y, \bar{\theta}\}|_{\theta=\bar{\theta}} \quad (57)$$

The regularity conditions for (57) to hold are (i)  $L(\theta)$  is differentiable on  $\Theta$ . (ii) For any  $\bar{\theta} \in \Theta$ ,  $Q(\theta, \bar{\theta})$  is continuously differentiable on  $\Theta$ . (iii) For any  $\theta \in \Theta$ , both  $|\log p_\theta(y, x)| \leq \alpha$  and  $\|\nabla_\theta \log p_\theta(y, x)\| \leq \beta$  for all  $y$  with  $\mathbb{E}\{\alpha\} < \infty$ ,  $\mathbb{E}\{\beta\} < \infty$ . Note that (ii) and (iii) are sufficient (via the dominated convergence theorem) for  $\nabla_\theta \mathbb{E}\{\log p_\theta(y, x)|y, \bar{\theta}\} = \mathbb{E}\{\nabla_\theta \log p_\theta(y, x)|y, \bar{\theta}\}$ .

In light of (57), we see that (55) is a stochastic gradient algorithm to maximize the objective  $\mathbb{E}_{\theta^\circ}\{\log p_\theta(y)\}$  wrt  $\theta$ . Moreover, we can rewrite this objective in terms of the Kullback Liebler (KL) divergence:

$$\begin{aligned} &\arg \max_{\theta \in \Theta} \mathbb{E}_{\theta^\circ}\{\log p_\theta(y)\} \\ &= \arg \min_{\theta \in \Theta} \mathbb{E}_{\theta^\circ}\{\log p_{\theta^\circ}(y)\} - \mathbb{E}_{\theta^\circ}\{\log p_\theta(y)\} \\ &= \arg \min_{\theta \in \Theta} D(\theta^\circ || \theta) \end{aligned} \quad (58)$$

where  $D(\theta^\circ || \theta)$  is the KL divergence between  $p_{\theta^\circ}(y)$  and  $p_\theta(y)$ . To summarize, the recursive EM algorithm (55) is

a stochastic gradient algorithm to minimize the KL divergence (58).

## Proofs of Theorems

### B. Proof of Theorem 1

Starting from (8), by expanding the symmetric polynomial coefficients we have in polynomial notation

$$\begin{aligned} S\{y_1, \dots, y_L\}(s) &= S\{\psi\theta_1^\circ, \dots, \psi\theta_L^\circ\}(s) + w(s), \\ \text{where } w(s) &\stackrel{\text{defn}}{=} \sum_{\mathcal{I} \subseteq [L], \mathcal{I} \neq \emptyset} \prod_{l \in \mathcal{I}} v_l S\{\psi\theta_l^\circ, i \in [L] - \mathcal{I}\}(s) \end{aligned} \quad (59)$$

The definition of the noise polynomial  $w$  in (59) involves the summation over all non-empty subsets  $\mathcal{I}$  of  $[L]$ .

*Remark.* To illustrate formula (59), consider  $L = 2$ . First expanding out we have

$$\begin{aligned} S\{y_1, y_2\}(s) &= (s + \psi\theta_1^\circ + v_1)(s + \psi\theta_2^\circ + v_2) \\ &= (s + \psi\theta_1^\circ)(s + \psi\theta_2^\circ) + w(s) \\ w(s) &= s(v_1 + v_2) + \psi\theta_1^\circ v_2 + \psi\theta_2^\circ v_1 + v_1 v_2 \end{aligned} \quad (60)$$

Let us verify the expression for  $w(s)$  in (59): Since,  $[L] = \{1, 2\}$ , for  $\mathcal{I} = \{1\}$ ,  $S\{\psi\theta_2^\circ\}(s) = (s + \psi\theta_2^\circ)$ ; for  $\mathcal{I} = \{2\}$ ,  $S\{\psi\theta_1^\circ\}(s) = (s + \psi\theta_1^\circ)$ ; and for  $\mathcal{I} = \{1, 2\}$ ,  $S\{\emptyset\} = 1$ . Then  $w(s) = v_1 S\{\psi\theta_2^\circ\}(s) + v_2 S\{\psi\theta_1^\circ\}(s) + v_1 v_2 S\{\emptyset\}$  which yields the above expression. Note that for  $\mathcal{I} = \emptyset$ ,  $S\{\psi\theta_1^\circ, \psi\theta_2^\circ\}(s) = (s + \psi\theta_1^\circ)(s + \psi\theta_2^\circ)$ .

From (59), since  $v_l$  are zero mean mutually independent, clearly  $\mathbb{E}\{w(s)\} = 0$ ; equivalently, the vector  $w$  is zero mean, i.e.,  $\mathbb{E}\{w_l\} = 0$ ,  $l \in [L]$ . We can then express (59) component-wise by reading off the  $L$  coefficients of the polynomial:

$$\begin{aligned} z_l &= S_l\{\psi\theta_1^\circ, \dots, \psi\theta_l^\circ\} + w_l \\ &= \psi^l S_l\{\theta_1^\circ, \dots, \theta_l^\circ\} + w_l = \psi^l \lambda_l^\circ + w_l, \quad l \in [L] \end{aligned} \quad (61)$$

where the second equality follows from (9). So we can rewrite objective (10) as  $L$  decoupled convex optimization problems in terms of the variables  $\lambda$ :

$$\min_{\lambda_l} \mathbb{E}|z_l - \psi^l \lambda_l|^2 \quad \text{where } z_l(k) = (\psi(k))^l \lambda_l^\circ + w_l(k), \quad (62)$$

Notice that each of the  $L$  objectives in (62) are quadratic (convex) in  $\lambda_l$ . It is easily verified that  $\lambda_l = \lambda^\circ$ ,  $l \in [L]$  is the unique minimizer that solves (62). Since  $\lambda = \lambda^\circ$  is the unique minimizer of (62) and  $S$  is uniquely invertible (see (14)), it follows that  $\theta = S^{-1}(\lambda) = S^{-1}(\lambda^\circ) = \theta^\circ$  is the unique minimizer of (10).

Finally, because each stochastic optimization problem (12) is quadratic, the adaptive filtering algorithm (13) yields estimates  $\lambda(k)$  that converge to  $\lambda^\circ$  in probability. Therefore the unique set of roots  $\theta(k) = S^{-1}(\lambda(k))$  converges to  $\theta^\circ$  in probability. (Recall from (14) that  $\theta(k)$  are the set of roots of the polynomial with coefficients  $\lambda(k)$ .)

### C. Proof of Theorem 2

**Statement 1.** Evaluating (16) with  $y$  in (1) we obtain

$$S\{y_1, \dots, y_L\}(s, t) = S\{\psi\theta_1^o, \dots, \psi\theta_L^o\}(s, t) + w(s, t)$$

$$w(s, t) \stackrel{\text{defn}}{=} \sum_{\mathcal{I} \subseteq [L], \mathcal{I} \neq \emptyset} \prod_{l \in \mathcal{I}} \left[ \sum_{j=1}^D v_{l,j} t^{j-1} \right] S\{\psi\theta_i^o, i \in [L] - \mathcal{I}\}(s, t) \quad (63)$$

The definition of the noise polynomial  $w$  in (63) involves the summation over all non-empty subsets  $\mathcal{I}$  of  $[L]$ .

Since  $v_l$  are zero mean mutually independent, it follows from (63) that  $\mathbb{E}\{w(s, t)\} = 0$ ; equivalently, the matrix  $w$  comprising the coefficients of the polynomial  $w(s, t)$  is zero mean, i.e.,  $\mathbb{E}\{w_{l,m}\} = 0$ ,  $m \in M_l$ ,  $l \in [L]$ .

*Remark.* Since the notation in (63) is complex, we illustrate formula (63) for the case  $L = 2$ ,  $D = 2$ . Let  $\psi_i$  denote the  $i$ -th row of  $\psi$ . Evaluating (16) with  $y$  given by (1), we obtain the polynomial

$$S\{y\}(s, t) = (s + \psi'_1\theta_1^o + t\psi'_2\theta_1^o)(s + \psi'_1\theta_2^o + t\psi'_2\theta_2^o) + w(s, t)$$

$$\text{where } w(s, t) = (s + \psi'_1\theta_1^o + t\psi'_2\theta_1^o)(v_{2,1} + v_{2,2}t) + (s + \psi'_1\theta_2^o + t\psi'_2\theta_2^o)(v_{1,1} + v_{1,2}t) + (v_{1,1} + v_{1,2}t)(v_{2,1} + v_{2,2}t) \quad (64)$$

We now show that (63) gives the same expression for  $w(s, t)$  as (64). For  $[L] = \{1, 2\}$ , we evaluate each subset  $\mathcal{I}$  and the corresponding term in (63). For  $\mathcal{I} = \{1\}$ ,  $S\{\psi\theta_2^o\}(s, t) = (s + \psi'_1\theta_2^o + t\psi'_2\theta_2^o)$  and the multiplying noise term is  $v_{1,1} + v_{1,2}t$ . For  $\mathcal{I} = \{2\}$ ,  $S\{\psi\theta_1^o\}(s, t) = (s + \psi'_1\theta_1^o + t\psi'_2\theta_1^o)$  and the multiplying noise term is  $v_{2,1} + v_{2,2}t$ . Finally for  $\mathcal{I} = \{1, 2\}$ ,  $S\{\emptyset\} = 1$  and the multiplying noise term is  $(v_{1,1} + v_{1,2}t)(v_{2,1} + v_{2,2}t)$ . Adding these three terms as in (63), we obtain noise polynomial  $w(s, t)$  in (64). Note that for  $\mathcal{I} = \emptyset$ ,  $S\{\psi\theta_1^o, \psi\theta_2^o\}(s, t)$ , we obtain the signal polynomial  $(s + \psi'_1\theta_1^o + t\psi'_2\theta_1^o)(s + \psi'_1\theta_2^o + t\psi'_2\theta_2^o)$ .

**Statement 2.** To keep the notation manageable we prove the result for  $L = 3$ , The general proof is identical but the notation becomes unreadable.

Suppose  $S_{lm}\{\theta\} = \sum_{i,j,k} \theta_{1i} \theta_{2j} \theta_{3k}$  where this sum is symmetric over indices  $[i, j, k]$  in a certain set. We denote this symmetric sum as  $S_{lm}\{\theta\} = [\theta_{1i} \theta_{2j} \theta_{3k}] \circ [i, j, k]$ .

*Example.* If  $D = 3$ ,  $L = 3$ ,  $S_{12}(\theta) = \theta_{11}\theta_{21}\theta_{32} + \theta_{11}\theta_{22}\theta_{31} + \theta_{12}\theta_{21}\theta_{31} = [\theta_{1i} \theta_{2j} \theta_{3k}] \circ [1, 1, 2]$ .

Then

$$S_{lm}\{\psi\theta\} = \sum_{i,j,k} \psi'_i \theta_1 \psi'_j \theta_2 \psi'_k \theta_3 = [\psi'_i \theta_1 \psi'_j \theta_2 \psi'_k \theta_3] \circ [i, j, k]$$

We want to show that this yields the RHS of (25). The trick is to encode the above expression in terms of a polynomial in

variable  $t$ :

$$S_{lm}\{\psi\theta\}(t) = \sum_{i,j,k} \sum_{p=1}^D \psi'_{i,p} \theta_{1p} t^{p-1} \sum_{q=1}^D \psi'_{j,q} \theta_{2q} t^{q-1} \times \sum_{r=1}^D \psi'_{k,r} \theta_{3r} t^{r-1}$$

$$= [\psi_{ip} \psi_{jq} \psi_{kr} \theta_{1p} \theta_{2q} \theta_{3r}] \circ ([p, q, r] \in t^0) + ([p, q, r] \in t^1) + ([p, q, r] \in t^2) + \dots \circ [i, j, k] \quad (65)$$

where we grouped the symmetric coefficients of powers of  $t$  in the last equation above. Clearly  $S_{lm}\{\psi\theta\} = S_{lm}\{\psi\theta\}(1)$ , i.e., by setting  $t = 1$ .

Next examining (65), we see that for each  $n$ , the symmetric coefficients of  $t^{n-1}$  satisfy

$$[\psi_{ip} \psi_{jq} \psi_{kr} \theta_{1p} \theta_{2q} \theta_{3r}] \circ ([p, q, r] \in t^{n-1}) \circ [i, j, k] = S_{ln}\{\psi^{lm}\} S_{ln}\{\theta\} \quad (66)$$

$$\psi^{lm} = \begin{bmatrix} \psi_{i,1} & \psi_{j,2} & \psi_{k,3} \\ \psi_{i,2} & \psi_{j,2} & \psi_{k,3} \\ \vdots & \vdots & \vdots \\ \psi_{i,D} & \psi_{j,D} & \psi_{k,D} \end{bmatrix}$$

Thus (65), (66) with coefficients  $\lambda_{ln} = S_{ln}(\theta)$  yields

$$S_{lm}\{\psi\theta\}(t) = S_{l1}\{\psi^{l,m}\} \lambda_{l1} + S_{l2}\{\psi^{l,m}\} \lambda_{l2} t + \dots + S_{l,M_l}\{\psi^{l,m}\} \lambda_{l,M_l} t^{M_l}$$

**Statement 3.** This follows immediately by substituting (24), (25) into (23).

**Statement 4.** For notational convenience we use  $\{\psi\theta\}$  to denote the set  $\{\psi\theta_1, \psi\theta_2, \dots, \psi\theta_L\}$ . Let  $S_{l,m}$ ,  $m \in M_l$ ,  $l \in [L]$  denote the coefficients of the polynomial  $S\{\psi\theta\}$ . Using (24),  $S_{l,m}\{y\} = S_{l,m}\{\psi\theta^o + v\} = S_{l,m}\{\psi\theta^o\} + w_{l,m}$  where  $w_{l,m}$  is zero mean iid over time; so we can rewrite the estimation objective (23) as

$$\arg \min_{\theta} \mathbb{E} \left\{ \sum_{l \in [L]} \sum_{m \in M_l} \|S_{l,m}\{\psi\theta\} - S_{l,m}\{\psi\theta^o\}\|^2 \right\}$$

Clearly the above minimum is achieved by choosing  $\theta^*$  such that the polynomial coefficients satisfy

$$S_{l,m}\{\psi\theta^*\} = S_{l,m}\{\psi\theta^o\} \quad \text{w.p.1, } m \in M_l, l \in [L]$$

Next since  $S$  is uniquely invertible, applying  $S^{-1}$  to the polynomial coefficients yields the unique set of roots

$$\{\psi\theta_l^*, m \in M_l, l \in [L]\} = \{\psi\theta_l^o, m \in M_l, l \in [L]\} \quad \text{w.p.1}$$

for all random variable realizations  $\psi$ . This in turn implies  $\{\theta_l^*, m \in M_l, l \in [L]\} = \{\theta_l^o, m \in M_l, l \in [L]\}$ .

*Remark.* The reader may wonder why the above proof breaks down for the naive symmetric transform (21). We note that for the naive vector symmetric transform (21), the above proof of Statement 4 does not hold. Even though  $\tilde{S}_{l,j}\{\psi\theta^*\} = \tilde{S}_{l,j}\{\psi\theta^o\}$  for all  $l, j$ , it does not follow that  $\{\psi\theta_l^*, m \in M_l, l \in [L]\} = \{\psi\theta_l^o, m \in M_l, l \in [L]\}$  w.p.1. This is because as discussed below (22), the naive symmetric transform does not preserve the ordering of vectors.

### D. Sensitivity of Symmetric Transform Polynomial

Here we derive the expression in (40).

**Theorem 6.** Suppose  $\theta = \{\theta_1, \dots, \theta_L\}$  is the set of factors of the polynomial  $S\{\theta\}(s) = s^L + \sum_{l=1}^L \lambda_l s^{L-l}$ . That is,  $S\{\theta\}(s) = \prod_{l=1}^L (s + \theta_l)$ . Assume  $\theta \in \theta$  is a distinct (non-repeated) factor. Then

$$\frac{d\theta}{d\lambda_l} = (-1)^{l+1} \theta^{L-l} \left[ \frac{dS\{\theta\}(-\theta)}{d\theta} \right]^{-1} \quad (67)$$

**Proof:** The proof follows from the following two lemmas.

**Lemma 3.** Suppose  $\theta = \{\theta_1, \dots, \theta_L\}$  is the set of factors of the polynomial  $S\{\theta\}(s) = s^L + \sum_{l=1}^L \lambda_l s^{L-l}$ . That is,  $S\{\theta\}(s) = \prod_{l=1}^L (s + \theta_l)$ . Then  $\theta$  is also the set of roots of the polynomial  $S\{\theta\}(-s) = s^L + \sum_{l=1}^L (-1)^l \lambda_l s^{L-l}$ . That is, for any  $\theta \in \theta$ , it follows that  $S\{\theta\}(-\theta) = 0$ .

**Lemma 4.** Suppose  $P_\beta(\theta) \stackrel{\text{defn}}{=} \theta^L + \sum_{m=0}^{L-1} \beta_m \theta^m = 0$ , i.e.,  $\theta$  is a root of the polynomial  $P_\beta(\theta)$ . Assume  $\theta$  is a distinct (non-repeated) root. Then

$$\frac{d\theta}{d\beta_m} = -\theta^m \left[ \frac{dP_\beta}{d\theta} \right]^{-1} \quad (68)$$

We now use Lemma 4 with Lemma 3 to obtain an expression for  $d\theta/d\lambda_l$ . Note that  $S\{\theta\}(-\theta) = P_\beta(\theta)$  by choosing  $\beta_m = (-1)^{L-m} \lambda_{L-m}$  for  $m = 0, \dots, L-1$ . Then from Lemma 4,

$$\frac{d\theta}{d\lambda_{L-m}} = -(-1)^{L-m} \theta^m \left[ \frac{dP_\beta}{d\theta} \right]^{-1} \quad (69)$$

$$\begin{aligned} \frac{dP_\beta}{d\theta} &= (L-1)\theta^{L-1} + \sum_{m=0}^{L-1} m \beta_m \theta^{m-1} \\ &= (L-1)\theta^{L-1} + \sum_{m=0}^{L-1} m (-1)^{L-m} \lambda_{L-m} \theta^{m-1} \text{ (Lemma 3)} \\ &= (L-1)\theta^{L-1} + \sum_{l=1}^L (L-l) (-1)^l \lambda_l \theta^{L-l-1} \end{aligned}$$

choosing  $l = L - m$

$$= \frac{dS\{\theta\}(-\theta)}{d\theta}$$

Therefore plugging  $l = L - m$  in (69), we obtain (67).

### E. Proof of Lemma 2

The MAP estimate is correct when the event  $I(\hat{x}(k) = x(k))$  occurs. Denoting  $i^* = \arg \max_i \pi_i$ , the conditional probability that the MAP estimate is correct given observation  $y(k)$  is

$$\mathbb{E}\{I(x(k) = i^*)|y(k)\} = \sum_{i=1}^X \pi_i(k) I(i^* = i) = \max_i \pi_i(k)$$

The error event is  $1 - I(x(k) = \hat{x}(k))$ . Therefore the error probability of the MAP estimate is  $1 - \max_i \pi_i(k) = 1 -$

$\max_i e_i' T(\pi, y(k))$ . Finally, the expected error probability of the MAP estimate over all possible realizations of  $y$  is

$$P_e(\pi; B) = \int_{\mathcal{Y}} (1 - \max_i e_i' T(\pi, y)) \sigma(\pi, y) = 1 - \int_{\mathcal{Y}} \max_i e_i' B_y \pi \quad (70)$$

Here  $\int_{\mathcal{Y}}$  denotes integration wrt Lebesgue measure when  $y \in \mathbb{R}^{L \times D}$  or counting measure when  $y$  is a subset of integers.

### F. Proof of Theorem 5

**Statement 1.** Since  $y_l$  are independent wrt  $l$ ,  $\text{Cov}_B(y) \preceq \text{Cov}_{\bar{B}}(y)$  implies  $\text{Var}_B(y_l) \leq \text{Var}_{\bar{B}}(y_l)$ . Next  $\text{Cov}(S\{y\})$  is a  $L \times L$  diagonal matrix with  $l$  element  $\sum_{i_1 < i_2 < \dots < i_l} \text{Var } y_{i_1} \text{Var } y_{i_2} \dots \text{Var } y_{i_l}$ . This together with  $\text{Var}_B(y_l) \leq \text{Var}_{\bar{B}}(y_l)$  implies  $\text{Cov}_B(S\{y\}) \preceq \text{Cov}_{\bar{B}}(S\{y\})$ .

**Statement 2.** The proof below exploits that facts that  $P_e(\pi, B) = 1 - \sum_y \max_i B_y \pi$  is concave in  $\pi$ , and that (A1), namely,  $B \geq_B \bar{B}$  holds. In particular, (A1) implies the following factorization of Bayes formula:

$$T(\pi, \bar{y}; \bar{B}) = \int_{\mathcal{Y}} T(\pi, y; B) \frac{\sigma(\pi, y; B)}{\sigma(\pi, \bar{y}; \bar{B})} M_{y, \bar{y}}$$

where  $\sigma(\pi, \bar{y}; \bar{B}) = \sum_y \sigma(\pi, y; B) M_{y, \bar{y}}$ . So  $\frac{\sigma(\pi, y; B)}{\sigma(\pi, \bar{y}; \bar{B})} M_{y, \bar{y}}$  qualifies as a measure wrt  $y$ . Since  $P_e(T(\pi, \bar{y}; \bar{B})) = 1 - \max_i T(\pi, \bar{y}; \bar{B})$  is concave in  $\pi$ , it follows using Jensen's inequality that

$$\begin{aligned} P_e(T(\pi, \bar{y}; \bar{B})) &= P_e\left(\int_{\mathcal{Y}} T(\pi, y; B) \frac{\sigma(\pi, y; B)}{\sigma(\pi, \bar{y}; \bar{B})} M_{y, \bar{y}}\right) \\ &\geq \int_{\mathcal{Y}} P_e(T(\pi, y; B)) \frac{\sigma(\pi, y; B)}{\sigma(\pi, \bar{y}; \bar{B})} M_{y, \bar{y}} \end{aligned}$$

So cross multiplying by  $\sigma(\pi, \bar{y}; \bar{B})$  and integrating wrt  $y$  implies

$$\int_{\mathcal{Y}} P_e(T(\pi, \bar{y}; \bar{B})) \sigma(\pi, \bar{y}; \bar{B}) \geq \int_{\mathcal{Y}} P_e(T(\pi, y; B)) \sigma(\pi, y; B)$$

which in turn implies  $P_e(\pi; \bar{B}) \geq P_e(\pi; B)$ .

**Statement 3.** Blackwell's classic paper [5, Theorem 3] shows that (A1) and (A2) imply that  $\sum_{y_{lm}} B_{iy_{lm}} y_{lm}^2 \leq \sum_{y_{lm}} \bar{B}_{iy_{lm}} y_{lm}^2$ , i.e.,  $\text{Cov}_B(y) \leq \text{Cov}_{\bar{B}}(y)$ . Here we give the proof in more transparent notation. Below we omit the  $l, m$  subscripts. Using Blackwell dominance (A1), it follows that

$$\int_{\mathcal{Y}} y \bar{B}_{iy} = \int_{\mathcal{Y}} y \int_{\bar{\mathcal{Y}}} B_{i, \bar{y}} M_{\bar{y}, y} = \int_{\bar{\mathcal{Y}}} B_{i, \bar{y}} \int_{\mathcal{Y}} y M_{\bar{y}, y}$$

by Fubini's theorem assuming  $\int_{\mathcal{Y}} |y| \bar{B}_{iy} < \infty$ . The mean preserving assumption (A2) implies that the above expression equals  $\int_{\bar{\mathcal{Y}}} \bar{y} B_{i, \bar{y}}$ . Therefore Blackwell dominance (A1) and mean preserving spread (A2) imply that the kernel  $M$  satisfies

$$\int_{\mathcal{Y}} y M_{\bar{y}, y} = \bar{y} \quad (71)$$

Next for any convex function  $\phi$ , applying Blackwell dominance, it follows that

$$\begin{aligned} \int_{\mathcal{Y}} \phi(y) \bar{B}_{i,y} &= \int_{\bar{\mathcal{Y}}} B_{i\bar{y}} \int_{\mathcal{Y}} \phi(y) M_{\bar{y},y} \\ &\geq \int_{\bar{\mathcal{Y}}} B_{i\bar{y}} \phi\left(\int_{\mathcal{Y}} y M_{\bar{y},y}\right) \quad (\text{by Jensen's inequality}) \\ &= \int_{\bar{\mathcal{Y}}} B_{i\bar{y}} \phi(\bar{y}) \quad (\text{by (71)}) \end{aligned} \quad (72)$$

So choosing  $\phi(y) = y^2$ , it follows that  $\text{Cov}_B(y) \leq \text{Cov}_{\bar{B}}(y)$ .

Therefore, from Statement 1, we have,  $\text{Cov}_B(S\{y\}) \leq \text{Cov}_{\bar{B}}(S\{y\})$ , or equivalently,  $R(B) \preceq R(\bar{B})$  (wrt positive definite ordering).

Next, it can be shown [6] by differentiation that the solution of the algebraic Lyapunov equation (36) satisfies

$$\Sigma(B) = \int_0^\infty \exp(sQ) R(B) \exp(sQ) ds$$

So clearly if  $R(B) \preceq R(\bar{B})$ , then  $\Sigma(B) \preceq \Sigma(\bar{B})$ . Finally, since  $\bar{\Sigma} = \nabla S^{-1} \Sigma \nabla S^{-1}$ , it follows that  $\bar{\Sigma}(B) \preceq \bar{\Sigma}(\bar{B})$ .

### G. Simulation Example. Noisy Matrix Permutation

The aim of this section is to provide a medium-sized numerical example of estimating  $\theta^\circ$  with the vector symmetric transform and adaptive filtering algorithm (28). We also show that applying the naive symmetric transform (21) element wise (as opposed to the vector symmetric transform) loses order information.

We consider the case  $L = 4$  and  $D = 10$ . The true parameter is

$$\theta^\circ = \begin{bmatrix} 1 & 3 & 4 & 5 & 7 & 9 & 10 & 11 & 12 & 13 \\ 2 & 4 & 5 & 10 & 8 & 7 & 1 & 8 & 9 & 10 \\ 3 & 1 & 2 & 7 & 6 & 5 & 4 & 5 & 7 & 9 \\ 6 & 12 & 18 & 24 & 36 & 43 & 50 & 10 & 1 & 3 \end{bmatrix}$$

The regression matrix  $\psi(k)$  was chosen as  $I_{3 \times 3}$ . The anonymized  $D$ -dimension observation vectors were generated according to (1), (2).

The pseudo observation vectors are constructed at each time  $k$  using (17) as

$$\begin{aligned} z_1 &= y_1 + y_2 + y_3 + y_4, \\ z_2 &= y_1 \otimes y_2 + y_1 \otimes y_3 + y_1 \otimes y_4 + y_2 \otimes y_3 + y_2 \text{Cov } y_4 \\ &\quad + y_3 \otimes y_4, \\ z_3 &= y_1 \otimes y_2 \otimes y_3 + y_1 \otimes y_2 \otimes y_4 + y_1 \otimes y_3 \otimes y_4 \\ &\quad + y_2 \otimes y_3 \otimes y_4, \\ z_4 &= y_1 \otimes y_2 \otimes y_3 \otimes y_4 \end{aligned} \quad (73)$$

where  $\otimes$  denotes the convolution operator and each  $y_i(k) \in \mathbb{R}^D$ ,  $i = 1, \dots, 4$ .

We ran 100 independent trials of the adaptive filtering algorithm (28) on 100 independent pseudo observation sequences. We computed the relative error of the average estimate  $\theta^{\text{avg}}(k)$  over the 100 trials at time  $k = 50,000$ :

$$|\theta_{ij}^{\text{avg}}(k) - \theta_{ij}^\circ| / \theta_{ij}^\circ \leq 7 \times 10^{-4}$$

Thus algorithm (28), based on the vector symmetric transform, successfully estimates the parameters.

Next, we ran adaptive filtering algorithm using the naive symmetric transform (21). We see from the estimate  $\theta(k)$  at  $k = 50,000$  below, that all order information is lost (the boxes indicate the nearest estimates to the first row of  $\theta^\circ$ ):

$$\begin{bmatrix} \boxed{1.0052} & 1.0053 & 1.9923 & \boxed{5.0129} & 6.0105 & 5.0033 \\ 2.0041 & \boxed{2.9971} & \boxed{4.0048} & 7.0028 & \boxed{6.9913} & 7.0086 \\ 3.0023 & 4.0016 & 4.9988 & 9.9934 & 8.0095 & \boxed{9.0024} \\ 5.9997 & 12.0000 & 17.9965 & 24.0002 & 36.0066 & 43.0028 \\ & 0.9939 & 4.9955 & 1.0032 & 2.9986 & \\ & 3.9995 & 7.9970 & 6.9999 & 8.9985 & \\ \boxed{9.9936} & 9.9942 & 8.9969 & 10.0001 & & \\ 50.0024 & \boxed{10.9959} & \boxed{12.0031} & \boxed{12.9960} & & \end{bmatrix}$$

We found in numerical examples that when rows of  $\theta^\circ$  are different from each other, the naive transform is able to estimate the order; but when the elements of two rows are close, then the estimate can switch rows resulting in a ghost estimate.

### H. Symmetric Transform for $D = 3, L = 3$ .

This final section of the supplementary document gives a complete evaluation of the symmetric transform  $S$  for  $D = 3, L = 3$  to illustrative (17) in the paper.

Recall  $\lambda_{l,m}$  is the coefficient of  $s^{l-1}t^{m-1}$  in the polynomial  $S\{\theta\}(s, t)$ :

$$\begin{array}{l} \lambda_{11} = [1, 1, 1] \\ \lambda_{12} = [1, 1, 2] \\ \lambda_{13} = [1, 1, 3] + [1, 2, 2] \\ \lambda_{14} = [1, 2, 3] + [2, 2, 2] \\ \lambda_{15} = [1, 3, 3] + [2, 2, 3] \\ \lambda_{16} = [2, 3, 3] \\ \lambda_{17} = [3, 3, 3] \end{array} \quad \left| \begin{array}{l} \lambda_{21} = [0, 1, 1] \\ \lambda_{22} = [0, 1, 2] \\ \lambda_{23} = [0, 1, 3] + [0, 2, 2] \\ \lambda_{24} = [0, 2, 3] \\ \lambda_{25} = [0, 3, 3] \end{array} \right.$$

$$\lambda_{31} = [0, 0, 1], \lambda_{32} = [0, 0, 2], \lambda_{33} = [0, 0, 3].$$

To explain the compact notation above:  $[a, b, c] = \sum_{\sigma\{a,b,c\}} \theta_{1a}\theta_{2b}\theta_{3c} = \theta_{1a}\theta_{2b}\theta_{3c} + \theta_{1a}\theta_{2c}\theta_{3b} + \theta_{1b}\theta_{2a}\theta_{3c} + \theta_{1b}\theta_{2c}\theta_{3a} + \theta_{1c}\theta_{2a}\theta_{3b} + \theta_{1c}\theta_{2b}\theta_{3a}$  and we set  $\theta_{i0} = 1$  for all  $i$ .

So  $\lambda_{11} = [1, 1, 1] = \theta_{11}\theta_{21}\theta_{31}$  since  $\sigma\{1, 1, 1\} = \{1, 1, 1\}$ . Also,  $\lambda_{12} = [1, 1, 2]$  is constructed by taking all permutations of  $[1, 1, 2]$ ; so  $\lambda_{12} = \theta_{11}\theta_{21}\theta_{32} + \theta_{11}\theta_{22}\theta_{31} + \theta_{12}\theta_{21}\theta_{31}$ . Similarly,  $\lambda_{23} = [0, 1, 3] + [0, 2, 2] = \theta_{11}\theta_{23} + \theta_{21}\theta_{33} + \theta_{11}\theta_{33} + \theta_{12}\theta_{22} + \theta_{12}\theta_{32} + \theta_{22}\theta_{32}$  since  $\theta_{i0} = 1$  by convention.

In the convolution notation of (17) we can express  $\lambda_1 \in \mathbb{R}^7, \lambda_2 \in \mathbb{R}^5, \lambda_3 \in \mathbb{R}^3$  as:

$$\lambda_1 = \theta_1 \otimes \theta_2 \otimes \theta_3, \lambda_2 = \theta_1 \otimes \theta_2 + \theta_1 \otimes \theta_3 + \theta_2 \otimes \theta_3, \lambda_3 = \theta_1 + \theta_2 + \theta_3$$

### REFERENCES

- [1] O. Cappe, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models*. Springer-Verlag, 2005.
- [2] V. Krishnamurthy, *Partially Observed Markov Decision Processes. From Filtering to Controlled Sensing*. Cambridge University Press, 2016.
- [3] V. Krishnamurthy, "Quickest Detection POMDPs with Social Learning: Interaction of local and global decision makers." *IEEE Transactions on Information Theory*, vol. 58, pp. 5563–5587, 2012.
- [4] C. F. J. Wu, "On the convergence properties of the EM algorithm;" *Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.

- [5] D. Blackwell, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, pp. 265–272, 1953.
- [6] B. D. O. Anderson and J. B. Moore, *Optimal filtering*. Englewood Cliffs, New Jersey: Prentice Hall, 1979.