

# Online Learning for Sparse PCA in High Dimensions: Exact Dynamics and Phase Transitions

Chuang Wang and Yue M. Lu

John A. Paulson School of Engineering and Applied Sciences  
Harvard University

**Abstract**—We study the dynamics of an online algorithm for learning a sparse leading eigenvector from samples generated from a spiked covariance model. This algorithm combines the classical Oja’s method for online PCA with an element-wise nonlinearity at each iteration to promote sparsity. In the high-dimensional limit, the joint empirical measure of the underlying sparse eigenvector and its estimate provided by the algorithm is shown to converge weakly to a deterministic, measure-valued process. This scaling limit is characterized as the unique solution of a nonlinear PDE, and it provides exact information regarding the asymptotic performance of the algorithm. For example, performance metrics such as the cosine similarity and the misclassification rate in sparse support recovery can be obtained by examining the limiting dynamics. A steady-state analysis of the nonlinear PDE also reveals an interesting phase transition phenomenon. Although our analysis is asymptotic in nature, numerical simulations show that the theoretical predictions are accurate for moderate signal dimensions.

## I. INTRODUCTION

Consider the spiked covariance model [1], where we are given a sequence of  $p$ -dimensional sample vectors  $\mathbf{y}_1, \mathbf{y}_2, \dots$  that are distributed according to

$$\mathbf{y}_k = \sqrt{\frac{\omega}{p}} c_k \boldsymbol{\xi} + \mathbf{a}_k. \quad (1)$$

Here,  $\boldsymbol{\xi}$  is an unknown vector in  $\mathbb{R}^p$ ,  $c_k \sim \mathcal{N}(0, 1)$ ,  $\mathbf{a}_k \sim \mathcal{N}(0, \mathbf{I}_p)$ , and  $\omega$  is a positive quantity specifying the signal-to-noise ratio (SNR);  $(c_i, \mathbf{a}_i)$  and  $(c_j, \mathbf{a}_j)$  are independent for  $i \neq j$ . In this paper, we analyze the exact dynamics of an online (incremental) algorithm for estimating  $\boldsymbol{\xi}$  in the high-dimensional ( $p \rightarrow \infty$ ) limit.

The model in (1) arises in the theoretical study of principal component analysis (PCA), an important statistical tool in exploratory data analysis, visualization and dimension reduction. A standard method to estimate  $\boldsymbol{\xi}$  is to compute the leading eigenvector of the sample covariance matrix  $\boldsymbol{\Sigma} = \frac{1}{n} \sum_{k=1}^n \mathbf{y}_k \mathbf{y}_k^T$ . For fixed  $p$  and when the number of samples  $n$  tends to infinity, the eigenvector is a consistent estimator of  $\boldsymbol{\xi}$  (up to a normalization constant.) However, in the regime where  $p$  and  $n$  are both large and comparable in size, the estimate given by the eigenvector is no longer consistent [2], [3].

To address this issue, a flurry of work—under the name of sparse PCA—has exploited the sparsity structure of  $\boldsymbol{\xi}$  (see, e.g.,

[3]–[5].) In addition to potentially improving the estimate of  $\boldsymbol{\xi}$ , sparse PCA generates a more parsimonious and interpretable representation, using a small subset of feature variables to explain the original data.

The natural formulation of sparse PCA leads to nonconvex optimization problems [3]–[5]. Convex relaxations via semidefinite programming (SDP) [6], [7] are possible, but the computational and storage cost of SDP may become prohibitive when the dimensionality is high. Many efficient algorithms have been proposed to solve sparse PCA, in both offline [3], [8]–[11] and online [12]–[15] settings. In the latter case, which is the setting we study in this paper, sample vectors  $\{\mathbf{y}_k\}$  arrive sequentially in an infinite stream; as soon as a new sample vector (or a small batch of them) has arrived, an online algorithm computes an instantaneous update to its estimate of  $\boldsymbol{\xi}$ . Since they only keep and operate on small sets of current samples, online algorithms are memory and computationally efficient. Moreover, as they provide estimates *on-the-fly*, online algorithms are well-suited to dynamic scenarios where the principal component vectors can be time-varying.

In this paper, we analyze an online sparse PCA algorithm that combines the classical Oja’s method [16] with an element-wise nonlinearity (e.g., soft-thresholding) at each iteration to promote sparsity (see Section II for the exact form.) Specifically, let  $\mathbf{x}_k$  be the estimate of  $\boldsymbol{\xi}$  given by the algorithm upon receiving the  $k$ th sample; let  $x_k^i$  and  $\xi^i$  denote the  $i$ th component of each vector. Also, define the joint *empirical* measure of  $\mathbf{x}_k$  and  $\boldsymbol{\xi}$  as

$$\mu_k^p(x, \boldsymbol{\xi}) \stackrel{\text{def}}{=} \frac{1}{p} \sum_{i=1}^p \delta(x - x_k^i, \boldsymbol{\xi} - \boldsymbol{\xi}^i). \quad (2)$$

Note that  $\mu_k^p(x, \boldsymbol{\xi})$  is a random element in  $\mathcal{M}(\mathbb{R}^2)$ , the space of probability measures on  $\mathbb{R}^2$ . As the main result of this work, we show that, as  $p \rightarrow \infty$  and with suitable time-rescaling, the sequence of empirical measures  $\{\mu_k^p(x, \boldsymbol{\xi})\}_p$  converges weakly to a deterministic measure-valued process  $\mu_t(x, \boldsymbol{\xi})$ . Moreover, this limiting measure  $\mu_t(x, \boldsymbol{\xi})$  is the unique solution of a nonlinear partial differential equation (PDE.)

The deterministic *scaling limit* as specified by the PDE and its solution provides a wealth of information regarding the performance of the online sparse PCA algorithm. For example, the limiting value of the cosine similarity

$$Q_k^p \stackrel{\text{def}}{=} \frac{\mathbf{x}_k^T \boldsymbol{\xi}}{\|\mathbf{x}_k\| \|\boldsymbol{\xi}\|} \quad (3)$$

This work was supported in part by the NSF under grant CCF-1319140 and by ARO under grant W911NF-16-1-0265.

at any step  $k$  can be easily obtained by computing the expectation  $\mathbf{E}(x\xi)$  with respect to the limiting measure  $\mu_t(x, \xi)$ . More involved questions, such as the misclassification rate in sparse support recovery, can also be answered by examining  $\mu_t(x, \xi)$ . Finally, studying the PDE in its steady-state leads to an exact characterization of the long-time behavior of the online sparse PCA algorithms. This steady-state analysis also uncovers a phase transition phenomenon: the performance of the algorithm can exhibit markedly different behaviors depending on the parameter settings and SNR values.

The rest of the paper is organized as follows. In Section II, we give the details of the online sparse PCA algorithm that we analyze in this work. The scaling limit of the algorithm is presented in Section III. As a special case, we study in Section III-B the classical Oja's method and derive an analytical expression characterizing the limiting cosine similarity between its estimates and  $\xi$ . Finally, a steady-state analysis and an associated phase transition phenomenon are discussed in Section IV.

## II. ONLINE ALGORITHM FOR SPARSE PCA

We consider the online setting, where sample vectors  $\{\mathbf{y}_k\}$  arrive sequentially. We assume that the samples are generated by the spiked covariance model in (1) with a single leading eigenvector  $\xi$ . We further assume that each element of  $\xi$  is an i.i.d. sample drawn from a mixture distribution

$$\pi(\xi) = (1 - \rho)\delta(\xi) + \rho u(\xi), \quad (4)$$

where  $\rho \in (0, 1]$  is a parameter controlling the sparsity level, and  $u(\xi)$  is a density function such that  $\int \xi^2 u(\xi) d\xi = 1/\rho$ . The preceding requirement makes sure that  $\|\xi\|/\sqrt{p} \rightarrow 1$  as the dimension  $p \rightarrow \infty$ . An example of (4) is the standard Bernoulli-Gaussian distribution. By choosing

$$u(\xi) = [\delta(\xi - 1/\sqrt{\rho}) + \delta(\xi + 1/\sqrt{\rho})]/2,$$

the distribution in (4) can also describe the sparse signal model considered in [7].

In this work, we analyze a simple recursive algorithm for estimating  $\xi$  from the stream of samples  $\{\mathbf{y}_k\}$ . The algorithm starts from some initial estimate  $\mathbf{x}_0$ . Upon receiving the  $k$ th data sample  $\mathbf{y}_k$ , it updates its estimate as follows:

$$\begin{aligned} \tilde{\mathbf{x}}_k &= \mathbf{x}_{k-1} + (\tau/p) \mathbf{y}_k \mathbf{y}_k^T \mathbf{x}_{k-1} \\ \mathbf{x}_k &= \sqrt{p} \eta(\tilde{\mathbf{x}}_k) / \|\eta(\tilde{\mathbf{x}}_k)\|. \end{aligned} \quad (5)$$

Here,  $\tau > 0$  is the step size, and  $\eta(\cdot)$  is an element-wise nonlinear mapping taking the form

$$\eta(x) = x - \frac{1}{p} \phi(x), \quad (6)$$

for some piecewise smooth function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ . Clearly, the method is online (incremental): it processes one sample at a time. Once a sample has been processed, it will be discarded and never used again.

The update steps in (5) as well as the expression in (6) need some explanations. First, we note that, without the nonlinear mapping (*i.e.*, by setting  $\eta(x) = x$ ), the recursions in (5) are

exactly the original Oja's method [16] for online PCA. The nonlinearity (6) in  $\eta(\cdot)$  is introduced to promote sparsity of the estimates. To see this, we consider an optimization formulation for sparse PCA in the *offline* setting:

$$\hat{\mathbf{x}} = \arg \min_{\|\mathbf{x}\|=\sqrt{p}} \frac{-\mathbf{x}^T \Sigma \mathbf{x}}{2} + \sum_{i=1}^p \Phi(x^i), \quad (7)$$

where  $\Sigma$  is the population (or sample) covariance matrix, and  $\Phi(\cdot)$  is an element-wise penalty function that favors sparse solutions. For example,  $\Phi(x) = \lambda|x|$  for lasso-type penalizations; or we can choose  $\Phi(x) = \lambda_1 x^2 + \lambda_2 |x|$  for the elastic net [5]. To solve (7), we use a proximal gradient method [17] followed by a projection onto the sphere of radius  $\sqrt{p}$ :

$$\begin{aligned} \tilde{\mathbf{x}}_k &= \mathbf{x}_{k-1} + (\tau/p) \Sigma \mathbf{x}_{k-1} \\ \mathbf{x}_k &= \sqrt{p} \operatorname{prox}_{\tau\Phi/p}(\tilde{\mathbf{x}}_k) \left\| \operatorname{prox}_{\tau\Phi/p}(\tilde{\mathbf{x}}_k) \right\|^{-1}, \end{aligned}$$

where  $\operatorname{prox}_{\tau\Phi/p}$  denotes the proximal operator of the function  $\tau\Phi(x)/p$ . Replacing the covariance matrix  $\Sigma$  by its instantaneous (and noisy) version  $\mathbf{y}_k \mathbf{y}_k^T$  and using the approximation  $\operatorname{prox}_{\tau\Phi/p}(x) \approx x - \tau(\frac{\partial}{\partial x} \Phi)/p$  (see, *e.g.*, [17, p. 138] for a justification of this approximation which holds for large  $p$ ), we reach our algorithm in (5) as well as the form given in (6).

*Example 1:* Consider a lasso-type penalization in (7) where  $\Phi(x) = \frac{\beta}{\tau}|x|$  for some  $\beta > 0$ . The associated proximal operator is the standard soft-thresholding function with parameter  $\beta/p$ , which can be approximated, for large  $p$ , as

$$\operatorname{prox}_{\tau\Phi/p}(x) \approx x - \frac{\beta \operatorname{sgn}(x)}{p}.$$

This corresponds to choosing  $\phi(x) = \beta \operatorname{sgn}(x)$  in (6). In what follows, we refer to this particular variant of the algorithm as Oja's algorithm with iterative soft thresholding (OIST for short.)

## III. DYNAMICS IN HIGH DIMENSIONS: SCALING LIMITS

In what follows, we analyze the dynamics of the online sparse PCA algorithm in (5) in the large  $p$  limit. The central object in our analysis is the empirical measure  $\mu_k^p(x, \xi)$  as defined in (2). Here, the subscript  $k$  indicates the iteration step, and the superscript  $p$  makes explicit the dependence of the measure on the dimension  $p$ .

The measure  $\mu_k^p$  contains a great deal of information about the algorithm. For example, using the notation

$$\langle f, \mu_k^p \rangle \stackrel{\text{def}}{=} \frac{1}{p} \sum_{i \leq p} f(x_k^i, \xi^i),$$

for a test function  $f(x, \xi)$ , we can write the cosine similarity defined in (3) as  $Q_k^p = \langle x\xi, \mu_k^p \rangle / \sqrt{\langle \xi^2, \mu_k^p \rangle}$ . Similarly, more involved quantities such as the misclassification rate in sparse support recovery can also be written in terms of  $\mu_k^p$ .

### A. The Main Convergence Result

To establish the scaling limit of  $\mu_k^p$ , we first embed the discrete-time sequence in continuous-time by defining

$$\mu_t^p \stackrel{\text{def}}{=} \mu_{\lfloor pt \rfloor}^p,$$

where  $\lfloor \cdot \rfloor$  is the floor function. Similarly, we can define  $Q_t^p$  as the continuous-time rescaled version of  $Q_k^p$ . Note that this type of time embedding and rescaling is standard in studying the convergence of stochastic processes [18]. (Some technicalities before we move on: since the empirical measure is random,  $\mu_t^p$  is a piecewise-constant càdlàg process taking values in  $\mathcal{M}(\mathbb{R}^2)$ , the space of probability measures on  $\mathbb{R}^2$ . In short,  $\mu_t^p$  is a random element in  $D(\mathbb{R}^+, \mathcal{M}(\mathbb{R}^2))$ , for which the notion of weak convergence is well-defined. See, e.g., [19].)

*Theorem 1:* Suppose that  $\mu_0^p$ , the empirical measure at time  $k = 0$ , converges (weakly) to a deterministic measure  $\mu_0 \in \mathcal{M}(\mathbb{R}^2)$  and that  $Q_0 = \langle x\xi, \mu_0 \rangle \neq 0$ . Then, as  $p \rightarrow \infty$ , the measure-valued stochastic process  $\mu_t^p$  converges weakly to a deterministic process  $\mu_t$ , characterized as the unique solution to the following nonlinear PDE (given in the weak form): for any positive, bounded and  $C^3$  test function  $f(x, \xi)$ ,

$$\begin{aligned} \langle f, \mu_t \rangle &= \langle f, \mu_0 \rangle + \int_0^t \left\langle \Gamma(x, \xi, Q_s, R_s) \frac{\partial}{\partial x} f, \mu_s \right\rangle ds \\ &+ \frac{\tau^2}{2} \int_0^t (1 + \omega Q_s^2) \left\langle \frac{\partial^2}{\partial x^2} f, \mu_s \right\rangle ds, \end{aligned} \quad (8)$$

where

$$Q_t = \iint x\xi d\mu_t, \quad R_t \stackrel{\text{def}}{=} \iint x\phi(x) d\mu_t; \quad (9)$$

with  $\phi(x)$  being the function introduced in (6), and

$$\Gamma(x, \xi, Q, R) \stackrel{\text{def}}{=} \tau\omega Q\xi - \phi(x) - x \left[ \tau\omega Q^2 - R + \frac{\tau^2}{2}(1 + \omega Q^2) \right]. \quad (10)$$

*Remark 1:* The deterministic measure-valued process  $\mu_t(x, \xi)$  characterizes the exact dynamics of the online sparse PCA algorithm in (5) in the high-dimensional limit. The nonlinear PDE (8) specifies the time evolution of  $\mu_t(x, \xi)$ . Note that (8) is presented in the weak form. If the strong, density valued solution exists, then it must satisfy

$$\begin{aligned} \frac{\partial}{\partial t} P_t(x|\xi) &= - \frac{\partial}{\partial x} \left[ \Gamma(x, \xi, Q_t, R_t) P_t(x|\xi) \right] \\ &+ \frac{\tau^2(1 + \omega Q_t^2)}{2} \frac{\partial^2}{\partial x^2} P_t(x|\xi), \end{aligned} \quad (11)$$

where we use  $P_t(x|\xi)$  to denote the conditional density of  $x$  given  $\xi$  at time  $t$ . The joint density can then be computed as  $P_t(x, \xi) = P_t(x|\xi)\pi(\xi)$ , where  $\pi(\xi)$  is the marginal density defined in (4).

*Remark 2:* For each  $\xi$ , the PDE (11) resembles a Fokker-Planck equation [20] describing the time-evolution of the probability density associated with a particle undergoing a drift-diffusion process in one spatial dimension. There is, however, one important distinction: the PDEs associated with different values of  $\xi$  are *coupled* via the quantities  $Q_t$  and  $R_t$ , which

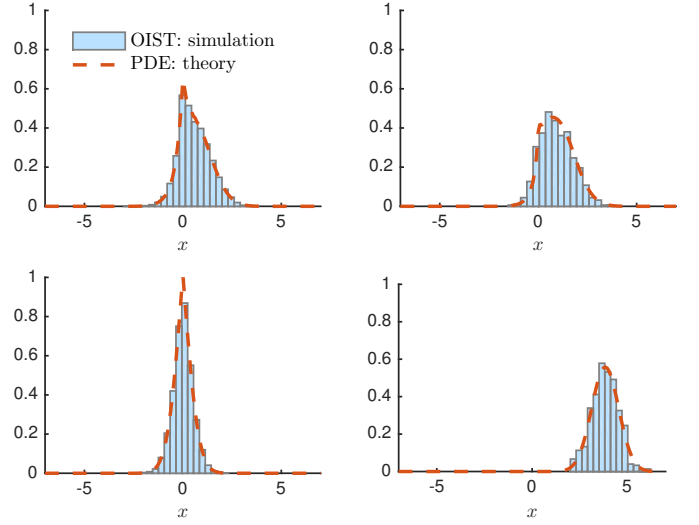


Fig. 1. Theory v.s. simulations. The figures show comparisons between the limiting conditional densities  $P_t(x|\xi)$  as predicted by the PDE (11) and the empirical densities obtained from Monte Carlo simulations. Top row:  $t = 1$ ; bottom row:  $t = 15$ ; left column:  $\xi = 0$ ; and right column:  $\xi = 1/\sqrt{\rho}$ . See Example 2 for details of the experiment.

themselves depend on the current densities  $P_t(x|\xi)$ . To see this, we rewrite (9) as

$$Q_t = \mathbf{E}_\xi \left( \xi \int x P_t(x|\xi) dx \right) \quad (12)$$

$$R_t = \mathbf{E}_\xi \left( \int x\phi(x) P_t(x|\xi) dx \right), \quad (13)$$

where  $\mathbf{E}_\xi$  denotes the expectation with respect to the variable  $\xi$  drawn from the prior distribution  $\pi(\xi)$ .

*Proposition 1:* Under the same assumptions of Theorem 1, the stochastic process  $Q_t^p \stackrel{\text{def}}{=} Q_{\lfloor tp \rfloor}^p$  converges weakly, as  $p \rightarrow \infty$ , to the deterministic function  $Q_t$  defined in (9).

*Remark 3:* We note that  $Q_t^p$  describes the time-evolution of the cosine similarity (3) between the estimate given by the algorithm and the unknown vector  $\xi$ . This result shows that the dynamics of  $Q_t^p$  converges to a deterministic curve  $Q_t$ , which can be computed from the limiting measure  $\mu_t$ .

*Example 2:* The proofs of Theorem 1 and Proposition 1 will be presented elsewhere. Here, we verify the accuracy of the theoretical predictions made in them via numerical simulations. In our experiment, we generate a vector  $\xi$  whose components are i.i.d. and drawn from a marginal distribution  $\pi(\xi) = (1 - \rho)\delta(\xi) + \rho\delta(\xi - 1/\sqrt{\rho})$ . The sparsity level is set to  $\rho = 0.05$ . Starting from a random initial estimate  $x_0$  with i.i.d. entries drawn from a normal distribution  $\mathcal{N}(\frac{1}{\sqrt{2}}, \frac{1}{2})$ , we use the OIST version of the online sparse PCA algorithm (see Example 1) to estimate  $\xi$ . The dimension is set to  $p = 10,000$ , and the other parameters are  $\tau = 0.5$ ,  $\beta = 0.27$ , and  $\omega = 1$ .

In Figure 1, we compare the predicted limiting conditional densities  $P_t(x|\xi = 0)$  and  $P_t(x|\xi = 1/\sqrt{\rho})$  against the empirical densities observed in the simulations, at two different times ( $t = 1$  and  $t = 15$ .) The PDE in (11) is solved numerically. We can see from the figure that the limiting densities given by the theory provide accurate predictions for the simulation

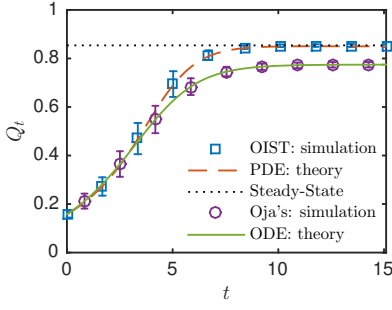


Fig. 2. The comparison between the analytical predictions of the cosine similarity  $Q_t$  and Monte Carlo simulations. For OIST, the theoretical curve is computed by using (12); for Oja's method, we use the closed-form formula in (14). The theoretical predictions are plotted as dashed and solid lines, whereas the average values of 120 Monte Carlo simulations are plotted as squares and circles. The error bars show confidence intervals of  $\pm 2$  standard deviations. The black dotted line indicates the theoretical prediction of the steady-state given by the solution of the fixed-point equations in (19).

results. In Figure 2, we verify the limiting form of the cosine overlap  $Q_t$  as given in (12). For simulations, we average over 120 independent instances of OIST, and plot the mean values and confidence intervals ( $\pm 2$  standard deviations.) Again, we can see that the asymptotic results match with simulation data very well. Also shown in the figure are results for the standard Oja's method, for which we can obtain a *closed-form* analytical formula for  $Q_t$ . This is the focus of the following subsection.

### B. The Nonsparse Case: Oja's Method

As mentioned earlier, the classical Oja's method [16] can be viewed as a special case of the algorithm in (5). It corresponds to setting  $\phi(x) = 0$  in (6), *i.e.*, the algorithm does not apply the nonlinear mapping  $\eta(x)$ . In this case, the limiting PDE (11) can be converted to a linear Fokker-Planck equation for the Ornstein-Uhlenbeck process, for which analytical solutions exist. For brevity, we omit discussions of this analytical solution of the PDE. Instead, we show a related result regarding the cosine similarity  $Q_t$ , which is an important figure of merit for the algorithm.

*Proposition 2:* For Oja's method, assume that we start the algorithm with a nonzero cosine similarity, *i.e.*,  $Q_0^p \rightarrow Q_0 \neq 0$  as  $p \rightarrow \infty$ . Then the dynamics of the cosine similarity  $Q_t^p \rightarrow Q_t$ , where  $Q_t$  is given by

$$Q_t^2 = \begin{cases} \alpha_2 \left[ \alpha_1 + \left( \frac{\alpha_2}{Q_0^2} - \alpha_1 \right) e^{-2\alpha_2 t} \right]^{-1} & \text{if } \alpha_2 \neq 0 \\ \left( 2\alpha_1 t + Q_0^{-2} \right)^{-1} & \text{if } \alpha_2 = 0. \end{cases} \quad (14)$$

Here,  $\alpha_1 = \tau\omega(1 + \frac{\tau}{2})$  and  $\alpha_2 = \tau(\omega - \frac{\tau}{2})$ .

*Proof (sketch):* We substitute  $f(x, \xi) = x\xi$  into the weak form (8) of the limiting PDE. The left-hand side is then exactly  $Q_t$ . Using the facts that  $\mathbf{E}_\xi \xi^2 = 1$ ,  $\phi(t) = 0$ , and after some manipulations, we can simplify the right-hand side of (8) and get  $Q_t = Q_0 + \int_0^t (-\alpha_1 Q_s^3 + \alpha_2 Q_s) ds$ . Solving this ordinary differential equation leads to (14). ■

In the long-time limit, we have

$$\lim_{t \rightarrow \infty} Q_t^2 = \max \left\{ 0, \frac{\omega - \frac{\tau}{2}}{\omega(1 + \frac{\tau}{2})} \right\}. \quad (15)$$

This result indicates that for any finite step size  $\tau > 0$ , Oja's method for online PCA cannot reach perfect estimation (*i.e.*,  $Q_\infty = 1$ ) even with infinite number of samples. Moreover, the formula also points out a simple phase transition phenomenon: when  $\tau > 2\omega$ , the estimates obtained by the algorithm will be uncorrelated with  $\xi$ .

## IV. STEADY STATE ANALYSIS AND PHASE TRANSITIONS

In this section, we study the long-time limit of OIST for sparse PCA. This steady-state analysis reveals an interesting phase transition phenomenon associated with OIST, which we also briefly discuss.

In the long-time limit, upon reaching the steady-state, the left-hand side of (11) becomes 0. It follows that the steady-state density functions satisfy the equation

$$\frac{\tau^2(1 + \omega Q^2)}{2} \frac{\partial}{\partial x} P(x | \xi) = \Gamma(x, \xi, Q, R) P(x | \xi), \quad (16)$$

where  $P(x | \xi)$ ,  $Q$ ,  $R$  are the steady-state versions of  $P_t(x | \xi)$ ,  $Q_t$  and  $R_t$ , respectively. Solving (16) and expanding  $\Gamma$  according to its definition in (10), we find the steady-state conditional density in the form of a Boltzmann distribution:

$$P(x | \xi) = \frac{1}{Z_\xi} \exp \left( - \frac{h(Q, R)x^2 + \Phi(x) - \tau\omega Q\xi x}{g(Q)} \right), \quad (17)$$

where  $Z_\xi$  is the partition function,

$$\begin{aligned} g(Q) &= \tau^2(1 + \omega Q^2)/2 \\ h(Q, R) &= (\tau\omega Q^2 - R + g(Q))/2 \end{aligned} \quad (18)$$

and  $\Phi(x)$  is an antiderivative of  $\phi(x)$ . Note that  $\Phi(x)$  can be any such antiderivative, since any constant added to  $\Phi(x)$  will be absorbed into the normalization constant  $Z_\xi$ .

It is important to emphasize that (17) is only an *implicit* definition of the steady-state distribution. This is because the expression relies on two constants  $Q$  and  $R$ , whose values are determined by the self-consistent equations (12) and (13) (with  $t \rightarrow \infty$ ) involving  $P(x | \xi)$ .

In what follows, we focus on OIST as discussed in Example 1. Here,  $\phi(x) = \beta \text{sgn}(x)$ , and thus we can set  $\Phi(x) = \beta|x|$ . It follows that the exponent in (17) is a piecewise quadratic polynomial. This convenient form allows us to further simplify the right-hand sides of (12) and (13). After some manipulations (which are omitted here), we can obtain the following fixed-point equations for determining  $Q$  and  $R$ :

$$\begin{aligned} Q &= \sqrt{\frac{g(Q)}{h(Q, R)}} \mathbf{E}_\xi \left( \xi \frac{z_+ f(z_+) - z_- f(z_-)}{f(z_+) + f(z_-)} \right), \\ R &= \beta \sqrt{\frac{g(Q)}{h(Q, R)}} \mathbf{E}_\xi \left( \frac{\frac{2}{\pi} - z_+ f(z_+) - z_- f(z_-)}{f(z_+) + f(z_-)} \right), \end{aligned} \quad (19)$$

where  $g(Q)$ ,  $h(Q, R)$  are the functions defined in (18),  $z_\pm = (g(Q)h(Q, R))^{-\frac{1}{2}} (\beta \pm \tau\omega\xi Q)/2$ , and  $f(\cdot)$  is the scaled complementary error function defined as  $f(x) = \frac{2}{\pi} e^{x^2} \int_x^\infty e^{-z^2} dz$ .

One can check that  $\{Q^0 = 0, R^0 = \frac{\tau^2}{2}\}$  is always a solution to the fixed-point equations (19). We call any such solution with

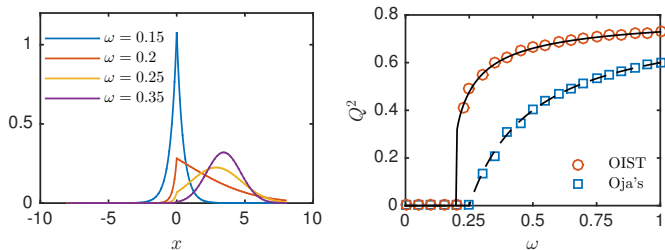


Fig. 3. Steady-state distributions and phase transitions. Left-hand side: The steady-state densities  $P(x|\xi = 1/\sqrt{\rho})$  at different SNR values. Right-hand side: Theoretical predictions of the steady-state cosine overlap  $Q$  as a function of the SNR parameter  $\omega$ . Black solid line: theoretical prediction for OIST; red dots: simulation results; black dashed line: theoretical prediction for Oja's method; blue squares: simulation results.

$Q = 0$  an *uninformative* solution, since it corresponds to a final estimate  $\mathbf{x}$  that is uncorrelated with  $\xi$ . It is also revealing to examine the corresponding steady-state distributions. Substituting  $Q^0, R^0$  into (17), we find that, for any  $\xi$ , the conditional density is of the form

$$P(x|\xi) = \frac{\beta}{\tau^2} e^{-\frac{2\beta}{\tau^2}|x|}. \quad (20)$$

Since  $P(x|\xi)$  does not depend on  $\xi$ , the variables  $x$  and  $\xi$  are independent; thus, the estimate provided by the algorithm contains no information about  $\xi$  in the long-time limit.

In the low SNR regime, such uninformative fixed-points are the only solutions to (19). The situation improves when we increase the SNR parameter  $\omega$ . At a certain critical value  $\omega_c$ , a nontrivial fixed point  $\{Q^*, R^*\}$  with  $Q^* \neq 0$  emerges. This corresponds to the case when the estimate  $\mathbf{x}$  becomes informative. We will present more detailed analysis of this phase transition phenomenon elsewhere. In what follows, we illustrate it using a numerical example.

We consider OIST at different SNR values. The other parameters in the algorithm are the same as those used in Example 2. The left-side of Figure 3 shows the limiting steady-state conditional densities  $P(x|\xi = 1/\sqrt{\rho})$  for increasing values of the SNR parameter  $\omega$ . At a low SNR value ( $\omega = 0.15$ ), we get the zero-mean (uninformative) Laplace distribution in (20). As  $\omega$  increases, the modes of the conditional densities move towards  $1/\sqrt{\rho}$ , starting to reveal information about  $\xi$ . In the right-side of Figure 3, we show the steady-state values of the cosine overlap  $Q$  as a function of  $\omega$ . A clear phase transition appears at a critical value  $\omega_c$ . The theoretical prediction (the solid line in the figure), obtained by numerically solving the fixed-point equations (19), matches very well with Monte Carlo simulations of the algorithm (shown as red dots.) Also shown in the figure are the results for Oja's method, with its theoretical prediction given by (15). Comparing OIST with Oja's method, we see that OIST has a lower phase transition threshold and that it also achieves a higher steady-state value for  $Q$ . This improvement in performance can be attributed to the fact that OIST exploits the sparsity structure of  $\xi$  via iterative thresholding.

## V. CONCLUSION

We analyzed the dynamics of an online sparse PCA algorithm in the high-dimensional limit. The joint empirical measure of

the underlying sparse eigenvector and its estimate as provided by the algorithm converges weakly to a deterministic process, characterized as the unique solution of a nonlinear PDE. This scaling limit provides exact information regarding the asymptotic performance of the algorithm. As a special case, we derived a closed-form expression for the limiting dynamics of the cosine similarity associated with Oja's method, a classical algorithm for online PCA. We also studied the steady-state of the nonlinear PDE and observed a phase transition phenomenon. The theoretical framework in this work is general. It paves the way towards understanding the dynamics of other online algorithms for various high-dimensional estimation problems. The theoretical analysis also provides insights and can lead to more principled ways of optimizing parameters in the algorithm to further improve performance.

## REFERENCES

- [1] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Stat.*, vol. 29, no. 2, pp. 295–327, Apr. 2001.
- [2] B. Nadler, "Finite sample approximation results for principal component analysis: A matrix perturbation approach," *Ann. Stat.*, vol. 36, no. 6, pp. 2791–2817, 2008.
- [3] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *J. Am. Stat. Assoc.*, vol. 104, no. 486, pp. 682–693, Jun. 2009.
- [4] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the LASSO," *J. Comp. Graph. Stat.*, vol. 12, no. 3, pp. 531–547, Sep. 2003.
- [5] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comp. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.
- [6] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Rev.*, vol. 49, no. 3, pp. 434–448, Jan. 2007.
- [7] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," *Ann. Stat.*, vol. 37, no. 5B, pp. 2877–2921, Oct. 2009.
- [8] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *J. Multivar. Anal.*, vol. 99, no. 6, pp. 1015–1034, Jul. 2008.
- [9] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, no. 517–553, 2010.
- [10] Z. Ma, "Sparse principal component analysis and iterative thresholding," *Ann. Stat.*, vol. 41, no. 2, pp. 772–801, Apr. 2013.
- [11] Y. Deshpande and A. Montanari, "Information-theoretically optimal sparse PCA," in *IEEE International Symposium on Information Theory*, 2014.
- [12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.
- [13] R. Arora, A. Cotter, K. Livescu, and N. Srebro, "Stochastic optimization for PCA and PLS," in *Proc. 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct. 2012.
- [14] A. Balsubramani, S. Dasgupta, and Y. Freund, "The fast convergence of incremental PCA," in *Adv. Neural Inf. Process. Syst.*, 2013.
- [15] W. Yang and H. Xu, "Streaming sparse principal component analysis," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 494–503.
- [16] E. Oja and J. Karhunen, "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix," *J. Math. Anal. Appl.*, vol. 106, no. 1, pp. 69–84, 1985.
- [17] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, Jan. 2014.
- [18] P. Billingsley, *Convergence of probability measures*, 2nd ed. New York: Wiley, 1999.
- [19] O. Kallenberg, *Foundations of modern probability*, 2nd ed. Springer, 2002.
- [20] H. Risken, *The Fokker-Planck equation: Methods of solution and applications*, 2nd ed. New York: Springer-Verlag, 1996.