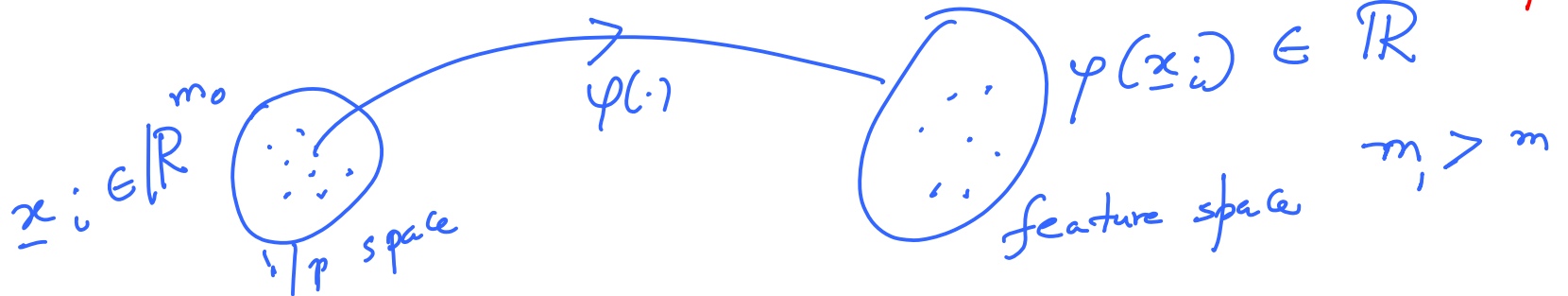


Kernel PCA

Most of PCA involve computations in the i/p space or data space.

We can also do a PCA in the feature space which is non-linearly related to the i/p space

Do a PCA in the feature space



Let $\varphi : \mathbb{R}^{m_0} \rightarrow \mathbb{R}^{m_1}$
(non linear)

$\varphi(\underline{x}_j)$ denotes the image of \underline{x}_j induced in the feature space by the non-linear map $\varphi(\cdot)$

Given $\{\underline{x}_i\}_{i=1}^N$, we compute $\{\varphi(\underline{x}_i)\}_{i=1}^N$

We can compute a correlation matrix in the feature space

$$\tilde{R} = \frac{1}{N} \sum_{i=1}^N \varphi(\underline{x}_i) \varphi^T(\underline{x}_i) \quad \text{--- (A)}$$

As with the ordinary PCA, ensure

$$\frac{1}{N} \sum_{i=1}^N \varphi(x_i) = 0$$

(Remove the bias priority to computing \tilde{R})

We can proceed by solving

$$\tilde{R} \tilde{q} = \tilde{\lambda} \tilde{q}$$

where

$\tilde{\lambda}$: eigen value of \tilde{R}
 \tilde{q} : corr. eigen vector of \tilde{R}

①

For $\tilde{\lambda} \neq 0$, satisfying (1)

\exists a corresponding set of coeffs $\{\alpha_j\}_{j=1}^N$

$$\tilde{q} = \sum_{j=1}^N \alpha_j \varphi(\underline{x}_j) \quad \text{--- (2)}$$

Plug (A), (2) in (1)

$$\sum_{i=1}^N \sum_{j=1}^N \alpha_j \varphi(\underline{x}_i) K(\underline{x}_i, \underline{x}_j) = N \tilde{\lambda} \sum_{j=1}^N \alpha_j \varphi(\underline{x}_j) \quad \text{(I)}$$
$$K(\underline{x}_i, \underline{x}_j) = \varphi^T(\underline{x}_i) \varphi(\underline{x}_j)$$

Pre multiply $\textcircled{\text{I}}$ b.s by $\varphi^T(\underline{x}_k)$

$$\sum_{i=1}^N \sum_{j=1}^N \alpha_j K(\underline{x}_k, \underline{x}_i) K(\underline{x}_i, \underline{x}_j) = N \tilde{\lambda} \sum_{j=1}^N \alpha_j K(\underline{x}_k, \underline{x}_j) \quad \textcircled{\text{II}}$$

Re write $\textcircled{\text{II}}$ in matrix form

$$K^2 \underline{\alpha} = N \tilde{\lambda} K \underline{\alpha}; \quad K := \left[K(\underline{x}_i, \underline{x}_j) \right]_{i,j=1}^N$$
$$K \underline{\alpha} = N \tilde{\lambda} \underline{\alpha} \quad (\text{Simplified eigen value eqn}) \quad \underline{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}$$

Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$ denote the eigen values of the kernel matrix K

$$\lambda_j = N \tilde{\lambda}_j \quad j = 1, \dots, N$$

j^{th} eigen value of the correlation matrix \tilde{R}

$$K \alpha = \lambda \alpha$$

$(\lambda = N \tilde{\lambda})$

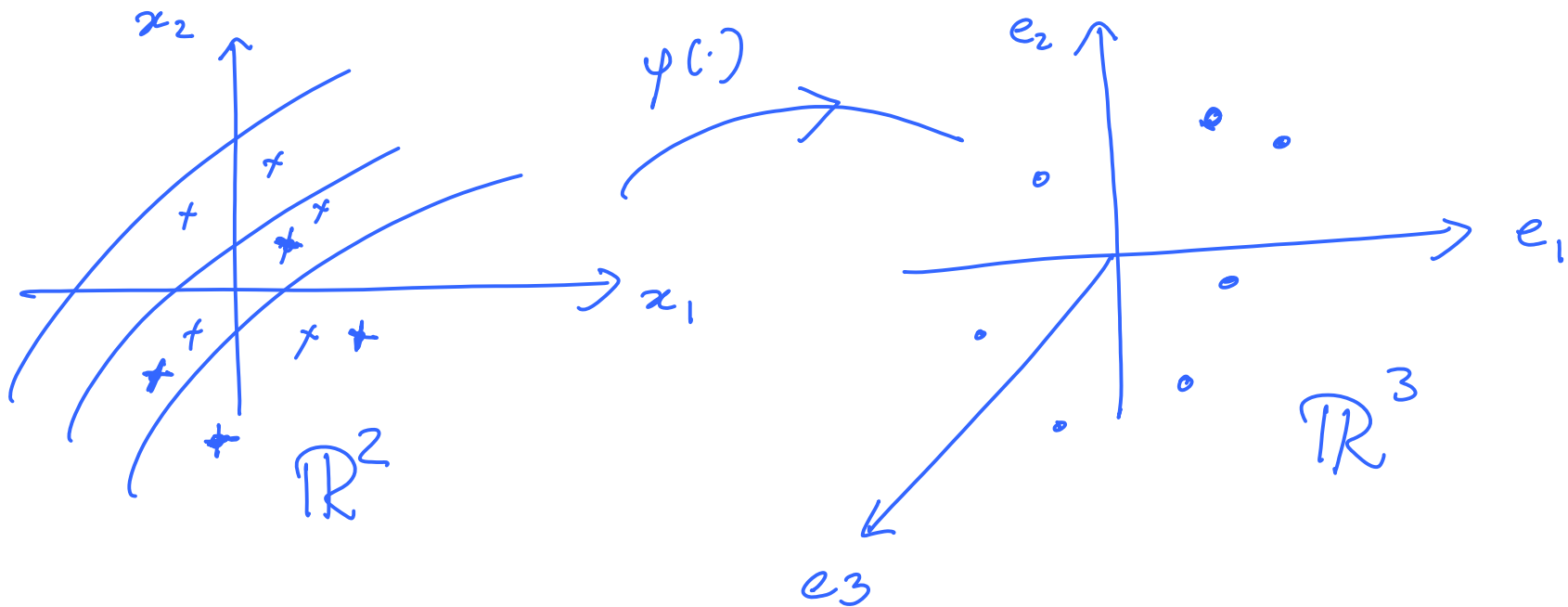
where $\tilde{\lambda}$ is the eigen value over the simplified eigen value eqn.

The vector α is to be normalized.

This requires eigen vector \tilde{q}_k of the corr. matrix \tilde{R}
to be normalized to unity
 $\tilde{q}_k^T \tilde{q}_k = 1 \quad \forall k = 1, \dots, N$

Verify : $\alpha_k^T \alpha_k = \frac{1}{\lambda_k}$

Summary of the kernel PCA



1) Given Algo training samples $\{\underline{x}_i\}_{i=1}^N$, compute the $N \times N$ Kernel matrix

$$K := [K(\underline{x}_i, \underline{x}_j)]$$

$$K(\underline{x}_i, \underline{x}_j) = \varphi^T(\underline{x}_i) \varphi(\underline{x}_j)$$

Solve the eigen value problem

$$K \underline{\alpha} = \lambda \underline{\alpha}$$

eigen vector
eigen value

3)

Normalize the eigen vectors

$$\alpha_k^T \alpha_k = \frac{1}{\lambda_k}$$

4)

For extraction of principal components of a

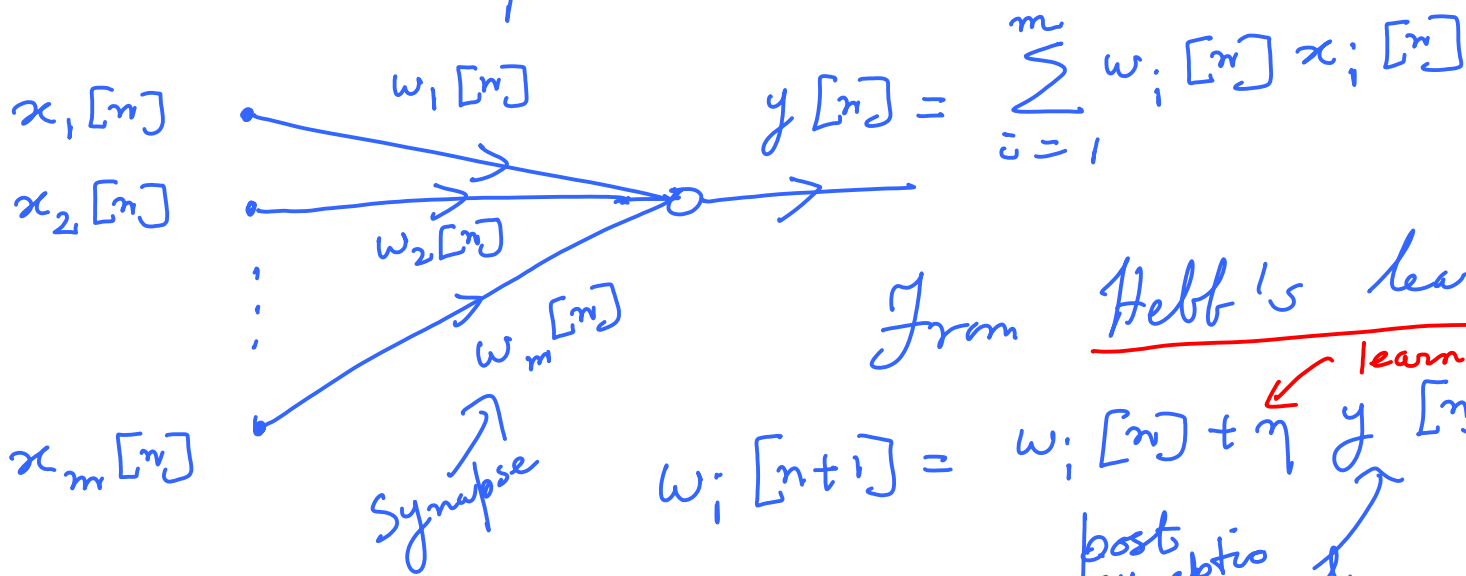
test vector \underline{x}

$$a_k = \frac{\tilde{q}_k^T \varphi(\underline{x})}{\lambda_k} = \langle \tilde{q}_k, \varphi(\underline{x}) \rangle$$

Hebbian based max. eigen filter

Consider a simple neuronal model

I/p process
is
stationary!



From Hebb's learning rule

$$w_i[n+1] = w_i[n] + \eta y[n] x_i[n]$$

learning rate η $y[n]$ $x_i[n]$
 post synaptic signal presynaptic signal

In the original form of Hebb's rule, the rule is unbounded and physically inadmissible

Oja (1982) did a normalization to the update

$$w_i[n] + \eta y[n] x_i[n]$$

(Normalize)
 $\frac{w[n+1]}{\|w[n+1]\|_{L_2 \text{ norm}}}$
Sense

$$w_i[n+1] = \frac{w_i[n] + \eta y[n] x_i[n]}{\left(\sum_{i=1}^m (w_i[n] + \eta y[n] x_i[n])^2 \right)^{\frac{1}{2}}}$$

Let us explore the stability analysis of this update

Consider the denominator

$$\sqrt{\sum_{i=1}^m w_i^2[n] + \eta^2 y^2[n] x_i^2[n] + 2\eta w_i[n] y[n] x_i[n]}$$

≈ 0 under the assumption $\eta \ll 1$

Approximation

$$\sqrt{\sum_{i=1}^m w_i^2[n] + 2\eta y[n] \underbrace{\sum_{i=1}^m w_i[n] x_i[n]}_{y[n]}}$$

$$\text{Let } b = \sum_{i=1}^n w_i^2[n]$$

$$\sqrt{b + 2\eta y^2[n]}$$

$$\sqrt{b \left(1 + \frac{2\eta y^2[n]}{b} \right)}$$

$$\approx \sqrt{b} \left(1 + \frac{\eta y^2[n]}{b} \right)$$

$$\sqrt{1+2x} \approx 1 + \frac{1}{2} \cdot 2x$$

$$\sqrt{b} \left(1 + \frac{\eta y^2[n]}{b} \right)$$

Geometric Series

\approx

$$\frac{1}{\sqrt{b}} \left(1 - \frac{\eta y^2[n]}{b} + O(\eta^2) \right)$$

Ignore h.o.t in η

At each time step n , weight vector i.e., $\|\underline{w}[n]\| = 1 \forall n$ is normalized

where $\underline{w}[n] = [w_1[n] \dots w_m[n]]^T$

The denominator simplifies to $(1 - \eta y^2[n])$

$$\begin{aligned} \therefore w_i[n+1] &= \frac{(w_i[n] + \eta y[n] x_i[n]) (1 - \eta y^2[n])}{1 - \eta^2 y^3[n] x_i[n]} \\ &= w_i[n] - \eta w_i[n] y^2[n] + \eta y[n] x_i[n] \end{aligned}$$

$\eta^2 \ll 1$
ignore

$$w_i[n+1] = w_i[n] + \eta y[n] \left(x_i[n] - w_i[n] y[n] \right)$$

The above simplification is upto a $x_i'[n]$
first power of η

One can think of Oja's approximation as a modification when i/p changes to the

form

$$x_i' [n] = x_i [n] - \underbrace{y [n] w_i [n]}_{\text{forgetting term with -ve feed back}}$$

\Rightarrow

$$w_i [n+1] = \underbrace{w_i [n] + \eta y [n] x_i' [n]}_{\text{Rewritten Hebbians that stabilizes the updates on weights}}$$

Let us do a matrix formulation

$$\underline{x}[n] = [x_1[n] \ x_2[n] \ \dots \ x_m[n]]^T$$

$$\underline{w}[n] = [w_1[n] \ \dots \ w_m[n]]^T$$

o/p ∇

$$y[n] = \underline{x}^T[n] \underline{w}[n] \quad \text{or} \quad \underline{w}^T[n] \underline{x}[n]$$

$$\therefore \underline{w}[n+1] = \underbrace{\underline{w}[n]}_{\text{updated wt.}} + \underbrace{\eta}_{\text{learning rate}} y[n] \left(\underbrace{\underline{x}[n]}_{\text{i/p vector}} - y[n] \underline{w}[n] \right)$$

Plugging $y[n]$ in terms of $\underline{w}[n]$ and $\underline{x}[n]$

$$\underline{w}[n+1] = \underline{w}[n] + \eta \left(\begin{array}{c} \underline{w}^T[n] \underline{x}[n] \underline{x}[n] \\ - \underbrace{\underline{w}^T[n] \underline{x}[n] \underline{x}^T[n] \underline{w}[n]}_{y^2[n]} \underline{w}[n] \end{array} \right)$$

The above is a non-linear stochastic
difference equation!

Asymptotic Stability Theorem

Consider the general stochastic approx. algo

$$\underline{w}[n+1] = \underline{w}[n] + \eta(n) h(\underline{w}(n), \underline{x}(n))$$

$n = 0, 1, \dots$
time steps

η is a sequence of +ve scalars

$h(\cdot)$ is a deterministic function with some regularity conditions

The following conditions must be satisfied

1) The sequence $\eta[n]$ is a decreasing seq. of +ve real nos

$$\lim_{n \rightarrow \infty} \eta[n] = 0$$

$$\sum_{n=1}^{\infty} \eta^p[n] < \infty$$

$$p > 1$$

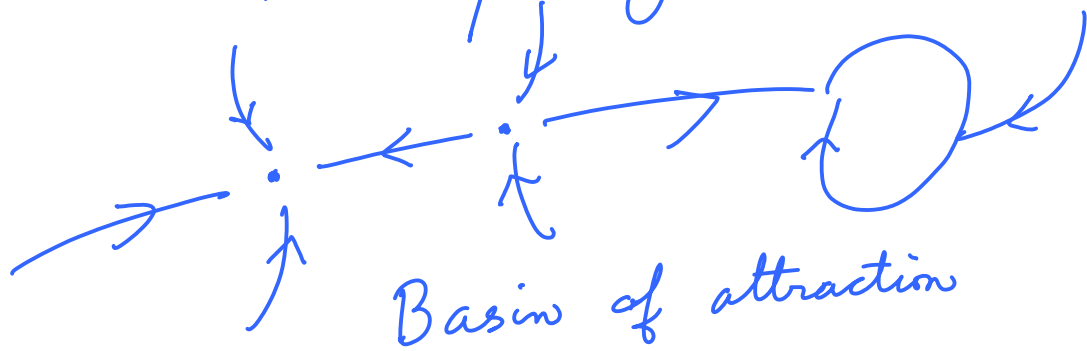
Controls the convergence rate

- 2) The sequence of parameter vectors \underline{w} is bounded with probability '1'
- 3) The update function $h(\underline{w}, \underline{x})$ is continuously differentiable w.r.t. \underline{w} and \underline{x} and the derivatives are bounded in time.
- 4) The limit $\bar{h}(\underline{w}) = \lim_{n \rightarrow \infty} E[h(\underline{w}, \underline{x})]$ exists $\forall \underline{w}$
Expectation is over the p.d.f of random data vector \underline{x}

5) There is a locally asymptotically stable
(in the Lyapunov sense) solution to the
ordinary differential eqn (ODE)

$$\frac{d \underline{w}(t)}{dt} = \underline{h}(w(t)) \text{ ————— } \textcircled{A}$$

6) Let \underline{q} be the soln to (A) with the basin of attraction $B(\underline{q})$.
 The parameter vector $\underline{w}(n)$ enters a compact subset A of the basin of attraction, infinitely often, with prob. 1



(Each attractor has a surrounding distinct region of its own BCD called basin of attraction)

The asymptotic stability theorem states that

$$\lim_{n \rightarrow \infty} w[n] = -\rho \quad \left(\begin{array}{l} \text{infinitely often} \\ \text{prob! 1} \end{array} \right)$$

No idea how many iterations are needed! ☹

Stability analysis of the max. eigen filter

To satisfy condition (i) of the stability theorem,

$$\begin{aligned} \text{Let } \eta[n] &= \frac{1}{n} \\ h(\underline{w}, \underline{x}) &= \frac{\underline{x}(n) y(n) - y^2(n) \underline{w}(n)}{\underline{x}(n) \underline{x}^T(n) \underline{w}(n) - [\underline{w}^T(n) \underline{x}(n) \underline{x}^T(n) \underline{w}(n)] \underline{w}(n)} \quad \textcircled{B} \end{aligned}$$

(B) Satisfies condition 3 of the Stability thm.

Taking expectation w.r.t. pdf of \underline{X}

$$\begin{aligned} T_h &= \lim_{n \rightarrow \infty} E \left(\begin{array}{c} \underline{x}(n) \underline{x}^T(n) \underline{w}(n) \\ \underline{w}^T(n) \underline{x}(n) \underline{x}^T(n) \underline{w}(n) \\ \underline{w}(n) \end{array} \right) \\ &= R \underline{w}(\infty) - \left[\underline{w}^T(\infty) R \underline{w}(\infty) \right] \underline{w}(\infty) \end{aligned}$$

Correlation matrix $E(\underline{X}\underline{X}^T)$ ignore \underline{w}

$$\frac{d}{dt} \underline{w}(t) = \underline{h}(w(t)) \left(\begin{array}{l} \text{NOTE:} \\ \frac{d}{dt} w(t) \propto \Delta W(n) \\ \text{assuming a} \\ \text{sampling interval } \Delta T \end{array} \right)$$

$$= R \underline{w}(t) - \left[\underline{w}^T(t) R \underline{w}(t) \right] \underline{w}(t)$$

(C)

Let $\underline{w}(t)$ be expanded in terms of the complete set of orthonormal eigenvectors of R

(Eigen expansion)

$$\underline{w}(t) = \sum_{k=1}^m \theta_k(t) \underline{q}_k \quad \text{--- (1)}$$

time varying
projection of $\underline{w}(t)$
on \underline{q}_k

Now, $R \underline{q}_k = \lambda_k \underline{q}_k$ --- (2) eigen vector

$$\underline{q}_k^T R \underline{q}_k = \lambda_k \quad \text{--- eigen value}$$

$$\therefore \sum_{k=1}^m \frac{d}{dt} \theta_k(t) \underline{q}_k = \overline{h}(\underline{w}(t))$$

$\frac{d}{dt} \underline{w}(t)$ Using ① and ② in ③

Consider $\underbrace{\omega^T(t) R \omega(t)}_{\text{eigen expansion}}$

$$\sum_{l=1}^m \theta_l(t) \underline{q}_l^T \cdot R \cdot \sum_{k=1}^m \theta_k(t) \underline{q}_k$$

$$= \sum_{l=1}^m \sum_{k=1}^m \underbrace{\theta_l(t) \theta_k(t)}_{\text{eigen}} \underbrace{\underline{q}_l^T R \underline{q}_k}_{\lambda_k \underline{q}_k}$$

eigen värde eqn

$$= \sum_{l=1}^m \sum_{k=1}^m \chi_k \theta_l(t) \theta_k(t) \underbrace{\underline{q}_l^T \underline{q}_k}_{\delta_{l,k}}$$

$$= \sum_{l=1}^m \lambda_l \theta_l^2(t) \quad \underline{\hspace{10em}} \quad \textcircled{3}$$

Using $\textcircled{3}$ and $\textcircled{1}$

$$\left[\omega^T(t) R \omega(t) \right]_{\omega(t)} \text{ evaluates to } \sum_{l=1}^m \lambda_l \theta_l^2(t) \quad \sum_{k=1}^m \theta_k(t) - q_k$$

Thus,

$$\bar{h}(\underline{w}(t)) = \sum_{k=1}^m \lambda_k \theta_k(t) \underline{q}_k - \sum_{l=1}^m \lambda_l \theta_l^2(t)$$

$\underline{w}^T(t) R \underline{w}(t)$
 \downarrow
 $\sum_{k=1}^m \theta_k(t) \underline{q}_k$
 $\uparrow \underline{w}(t)$

Written carefully,

$$\sum_{k=1}^m \frac{d}{dt} \theta_k(t) \underline{q}_k = \frac{d}{dt} \underline{w}(t)$$

$$\bar{h}(\underline{w}(t)) = \sum_{k=1}^m \lambda_k \theta_k(t) \underline{q}_k - \sum_{l=1}^m \lambda_l \theta_l^2(t)$$

④

$\{q_k\}_{k=1}^m$ are orthonormal \Rightarrow linear independence vectors
 Eqn (4) is a linear combination of linearly independent vectors whose coeffs are governed by the differential eqn given by

$$\frac{d}{dt} \Theta_k(t) = \lambda_k \Theta_k(t) - \Theta_k(t) \sum_{l=1}^m \lambda_l \Theta_l^2(t)$$

$k = 1, 2, \dots, m$

$\Theta_k(t)$ are the principal modes

Case 1 : $1 < k \leq m$

For this treatments,

$$\text{let } \alpha_k(t) \stackrel{\Delta}{=} \frac{\theta_k(t)}{\theta_1(t)} \quad 1 \leq k \leq m$$

$\theta_1(t) \neq 0$
w. prob 1

and $\omega(t)$ is randomly chosen

$$\frac{d\alpha_k(t)}{dt} = \underbrace{\frac{1}{\theta_1(t)} \frac{d\theta_k(t)}{dt}}_{\text{Term 1}} - \underbrace{\frac{\theta_k(t)}{\theta_1^2(t)} \frac{d\theta_1(t)}{dt}}_{\text{Term 2}}$$

$$\frac{1}{\theta_1(t)} \frac{d\theta_k(t)}{dt} = \frac{1}{\theta_1(t)} \left[\lambda_k \theta_k(t) - \theta_k(t) \sum_{l=1}^m \lambda_l \theta_l^2(t) \right]$$

$$-\frac{\theta_k(t)}{\theta_1^2(t)} \frac{d\theta_1(t)}{dt} = -\frac{\theta_k(t)}{\theta_1^2(t)} \left[\lambda_1 \theta_1(t) - \theta_1(t) \sum_{l=1}^m \lambda_l \theta_l^2(t) \right]$$

Summing the terms in $\delta(a)$ & $\delta(b)$

$$\frac{d \alpha_k(t)}{dt} = \frac{\theta_k(t)}{\theta_1(t)} (\lambda_k - \lambda_1)$$

$$= \alpha_k(t) (\lambda_k - \lambda_1)$$

$$= - (\lambda_1 - \lambda_k) \alpha_k(t)$$

Assume λ_1 is max. value for all $\lambda_k, k \neq 1$

Assuming eigen values of R are distinct

and $\lambda_1 > \lambda_2 > \dots > \lambda_m$

$(\lambda_1 - \lambda_k) \propto$ $\xrightarrow{\text{time constant}}$

$$\alpha_k(t) = e^{-at} u(t) \quad k \neq 1$$
$$\frac{d\alpha_k}{dt} = -a \alpha_k$$
$$a > 0$$
$$a = \lambda_1 - \lambda_k$$

$\Rightarrow \alpha_k(t) \xrightarrow[t \rightarrow \infty]{=} 0$ for $1 < k \leq m$

CASE 2 : $k = 1$

$$\frac{d\theta_1(t)}{dt} = \lambda_1 \theta_1(t) - \theta_1(t) \sum_{l=1}^m \lambda_l \theta_l^2(t)$$
$$= \lambda_1 \theta_1(t) - \lambda_1 \theta_1^3(t) - \theta_1(t) \sum_{l=2}^m \lambda_l \theta_l^2(t)$$

From Case 1,
of our analysis

$$\alpha_l \xrightarrow[t \rightarrow \infty]{} 0$$

————— (6)

∴ The governing equation is

$$\frac{d\theta_1(t)}{dt} = \lambda_1 \theta_1(t) (1 - \theta_1^2(t)) \quad (\text{asymptotically}) \quad \textcircled{7}$$

To analyze the stability of this system, we need a positive definite function called Lyapunov function.

(Part of non-linear dynamics)

Let \underline{s} be the state vector of an autonomous system

Let $V(t)$ be the Lyapunov function of the system

An equilibrium state $\underline{s}^{(eq)}$ of the system is automatically

stable if

$$\frac{dV(t)}{dt} < 0 \quad \text{for } \underline{s} \in \mathcal{U} - \underline{s}^{(eq)}$$

where \mathcal{U} is a small neighborhood around \underline{s}

For our problem at hand, the differential eqn has a Lyapunov function given by

$$V(t) = \left[\theta_1^2(t) - 1 \right]^2 \quad (8)$$

To validate the assertion,

$$(1) \quad \frac{dV(t)}{dt} < 0 \quad \forall t$$

(2) $V(t)$ has a minimum

(You may question this @ this stage)

Now $\frac{dV(t)}{dt} = 4\theta_1(t) \left(\theta_1^2(t) - 1 \right) \frac{d\theta_1(t)}{dt}$ — (9a)

But $\frac{d\theta_1(t)}{dt} = \lambda_1 \theta_1(t) - \theta_1(t) \sum_{l=1}^m \lambda_l \theta_l^2(t)$

$\underbrace{\hspace{10em}}_{\text{asymptotically except } l=1}$

$= \lambda_1 \theta_1(t) - \lambda_1 \theta_1(t) \theta_1^2(t)$
 $= \lambda_1 \theta_1(t) (1 - \theta_1^2(t))$
 $= -\lambda_1 \theta_1(t) (\theta_1^2(t) - 1)$

Plug (9b) in (9a)

(9b)

$$\frac{dV(t)}{dt} = -4\lambda_1 \theta_1^2(t) (\theta_1^2(t) - 1)^2$$

From the +ve definite property of R i.e., Correlation matrix

since eigen value λ_1 is +ve

$$\frac{dV(t)}{dt} < 0$$

2) $V(t)$ has a minimum @ $\theta_1(t) = \pm 1$, and so the 2nd condition is also satisfied.

$$\theta_1(t) \xrightarrow{t \rightarrow \infty} \pm 1$$

$$\theta_k(t) \xrightarrow{t \rightarrow \infty} 0$$

$$1 < k \leq m \quad (\text{Case A})$$

Note this carefully
Case 1 of the analysis

2 conclusions can be drawn

1) The only principal mode of the stochastic approx. algo. described in

$$\underline{w}(n+1) = \underline{w}(n) + \eta \begin{pmatrix} \underline{x}(n) \underline{x}^T(n) \underline{w}(n) \\ \underline{w}^T(n) \underline{x}(n) \underline{x}^T(n) \underline{w}(n) \\ \underline{w}(n) \end{pmatrix}$$

is $\theta_1(t)$. All other modes die to zero.

2) $\theta_1(t)$ will converge to ± 1

Formally stated,

$\underline{w}(t) \xrightarrow[t \rightarrow \infty]{} \underline{q}_1$ where \underline{q}_1 is the
normalized eigen vector associated with the largest
eigen value λ_1 of the correlation matrix R .

According to Condition (6) of the asymptotic stability theorem (AST), \exists a subset A of the set of vectors / $\lim_{n \rightarrow \infty} \underline{w}(n) = \underline{q}$, (i.o. with prob. 1)
 infinitely often

To satisfy condition (2) of the AST, we hard limit the entries of $\underline{w}(n)$ so that their magnitudes remain below a threshold 'a'.

$$\| \underline{w}(n) \| = \max_j |w_j(n)| \leq a$$

Let A be compact subset of \mathbb{R}^m defined by
the set of vectors whose norm $\leq a$.

Sanger (1989) showed that

If $\|\underline{w}^{(n)}\| \leq a$ and the constant is sufficiently
large, then $\|\underline{w}^{(n+1)}\| < \|\underline{w}^{(n)}\|$ with prob. 1

\Rightarrow As iterations $n \rightarrow \infty$, $\underline{w}^{(n)}$ will eventually be
within A & will remain inside A i.o. with

Basin of attraction $B(\underline{q}_1)$ includes all vectors with
norm bounded $\Rightarrow A \in B(\underline{q}_1)$

Rest of all the conditions in A ST are met,

\Rightarrow $w(n)$ Converges to q , with prob. 1
and has λ_1 as the associated eigen value

\therefore A single neuron under normalized Hebbian
update, aligns to the principal eigen vector
of the correlation matrix

Summary of the Hebbian-based eigen filter

1) For stationary inputs $\underline{x}(n)$, a single neuron extracts the principal eigen component of the Correlation matrix R.

Verify : λ_1 is related to the variance of the
o/p $y(n)$; $\sigma^2(n) = E(y^2(n))$
i.e., Prove $\sigma^2(n) \xrightarrow{n \rightarrow \infty} \lambda_1$ ← scalar

2) Per Oja's update, Hebbian based filter converges to a fixed point with

prob. 1.

$$(a) \quad \lim_{n \rightarrow \infty} \sigma^2(n) = \lambda_1$$

$$(b) \quad \lim_{n \rightarrow \infty} \underline{w}(n) = \underline{q}_1 \quad ;$$

