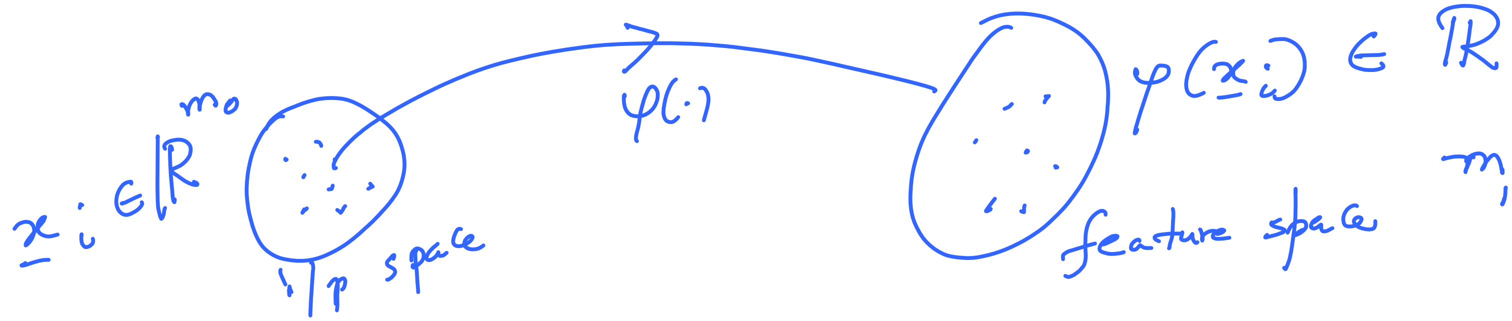
Kernel PCA

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



$m_1 > m$

Let $\varphi : \mathbb{R}^{m_0} \longrightarrow \mathbb{R}^{m_1}$

(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 $\{\varphi(\underline{x}_i)\}_{i=1}^{N}$ in the feature space

We can compute a correlation matrix

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

As with the ordinary PCA, ensure

$$\frac{1}{N}\sum_{i=1}^{N} \varphi(\underline{x}_i) = 0$$

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

We can proceed by solving

$$\tilde{R}\,\tilde{q} = \tilde{\lambda}\,\tilde{q} \qquad (1)$$

where $\tilde{\lambda}$: eigen value of $\tilde{R}$

$\tilde{q}$: Corr. eigen vector of $\tilde{R}$

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

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

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

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

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

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

Premultiply I b.s by $\varphi^T(\underline{x}_k)$

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

Rewrite III in matrix form

$$K^2 \underline{\alpha} = N \tilde{\lambda} K \underline{\alpha} \, ; \quad K := \left[K(\underline{x}_i, \underline{x}_j) \right]_{i,j=1}^{N}$$

$$K \underline{\alpha} = N \tilde{\lambda} \underline{\alpha} \quad \text{(Simplified eigen value eqn)} \qquad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}$$

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

$$\lambda_j = N \tilde{\lambda}_j \qquad j = 1, \ldots, N$$

jth eigen value of the correlation matrix $\tilde{R}$

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

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

where λ is the eigen value over the simplified eigen value eqn.

The vector α is to normalized.

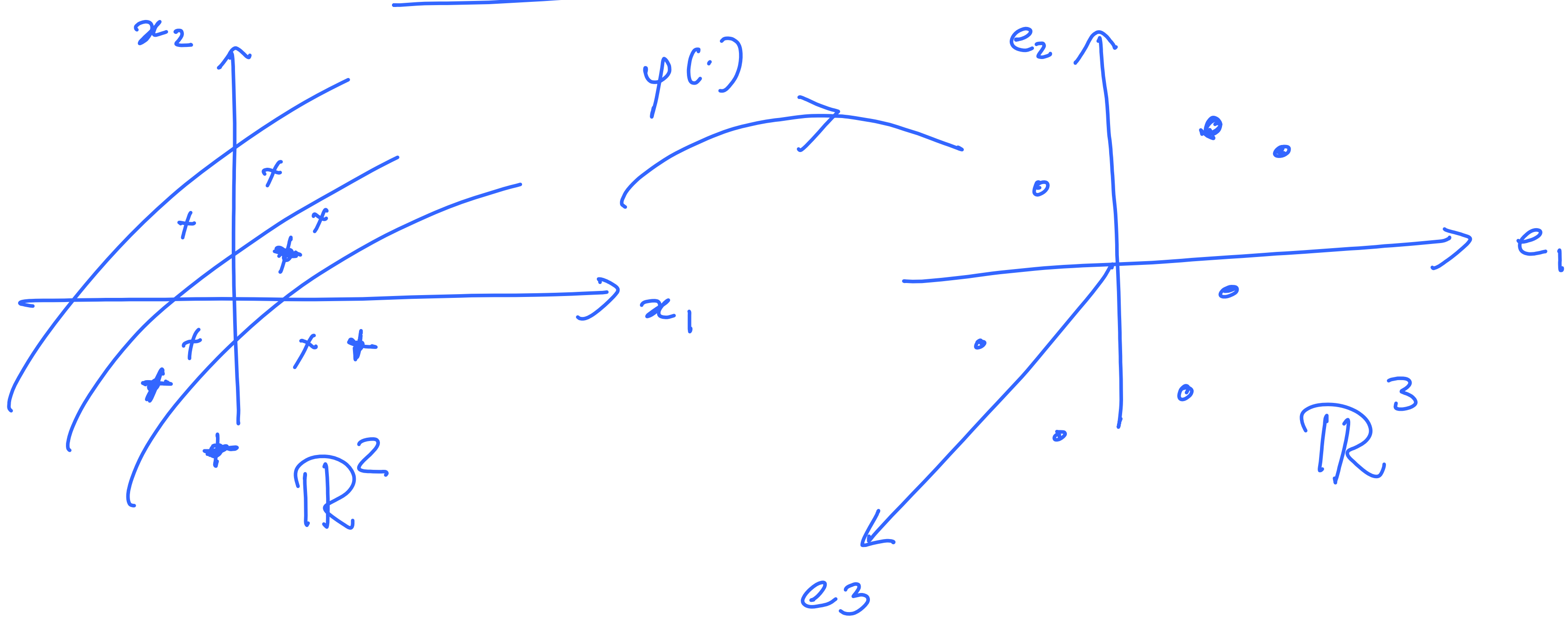
This requires eigen vector $\tilde{q}$ of the corr. matrix $\tilde{R}$ to be normalized to unity

$$\tilde{q}_k^T \tilde{q}_k = 1 \quad \forall\ k = 1, \dots, N$$

Verify :

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

Summary of the kernel PCA



Algo

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

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

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

2) Solve the eigen value problem

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

eigen vector

eigen value

3) Normalize the eigen vectors

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

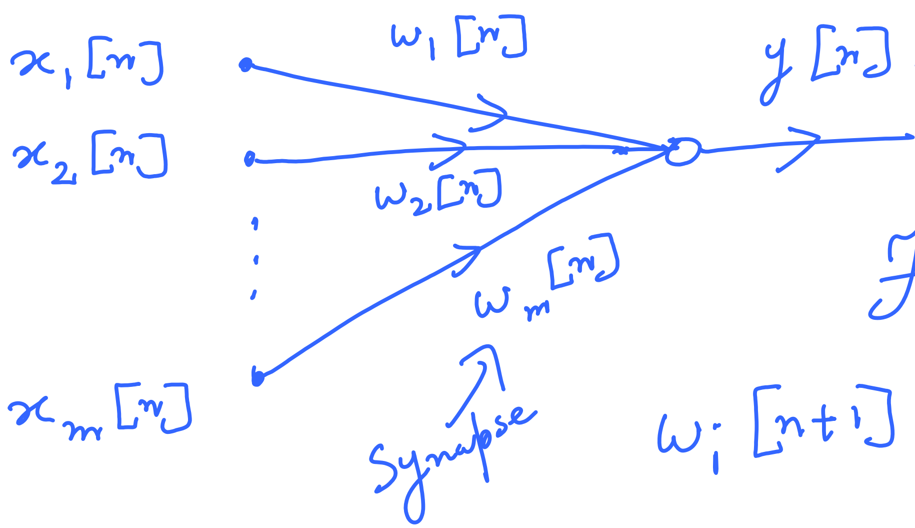
4) For extraction of principal components of a test vector $\underline{x}$

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

Hebbian based max. eigen filter

Consider a simple neuronal model

I/p process is stationary!



$$y[n] = \sum_{i=1}^{m} w_i[n]\, x_i[n]$$

From Hebb's learning rule

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

learning rate

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

(Normalize $\underline{w}[n+1]$ L_2 norm sense)

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

Let us explore the stability analysis of this update

Consider the denominator

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

≈ 0 under the assumption $\eta <<$

$\Downarrow$ Approximation

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

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

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

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

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

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

Geometric Series

Ignore h.o.t in η

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

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

where $\underline{w}[n] = \begin{bmatrix} w_1[n] & \cdots & w_m[n] \end{bmatrix}^T$

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

$$\therefore \; w_i[n+1] = \left(w_i[n] + \eta y[n] x_i[n]\right)\left(1 - \eta y^2[n]\right)$$

$$= w_i[n] - \eta w_i[n] y^2[n] + \eta y[n] x_i[n] - \eta^2 y^3[n] x_i[n]$$

$\eta^2 \ll$ ignore

$$w_i[n+1] = w_i[n] + \eta\, y[n] \underbrace{\left(x_i[n] - w_i[n]\, y[n] \right)}_{x_i'[n]}$$

The above simplification is upto a first power of η

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

$$x_i'[n] = x_i[n] - \underbrace{y[n]\, w_i[n]}$$

forgetting term with $-$ve feed back

$$\Rightarrow \quad w_i[n+1] = \underbrace{w_i[n] + \eta\, y[n]\, x_i'[n]}$$

Rewritten Hebbian that stabilizes the update on weights

Let us do a matrix formulation

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

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

o/p

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

$$\therefore \underline{w}[n+1] = \underline{w}[n] + \eta\, y[n]\left(\underline{x}[n] - y[n]\,\underline{w}[n]\right)$$

updated wt. learning rate i/p vector

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

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

The above is a non-linear stochastic difference equation!

Asymptotic Stability Theorem

Consider the general stochastic approx. algo

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

$n = 0, 1, \ldots$ 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

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

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

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 $\bar{h}(\underline{w}) = \lim_{n \to \infty} E[h(\underline{w}, \underline{X})]$ exists $\forall \underline{w}$

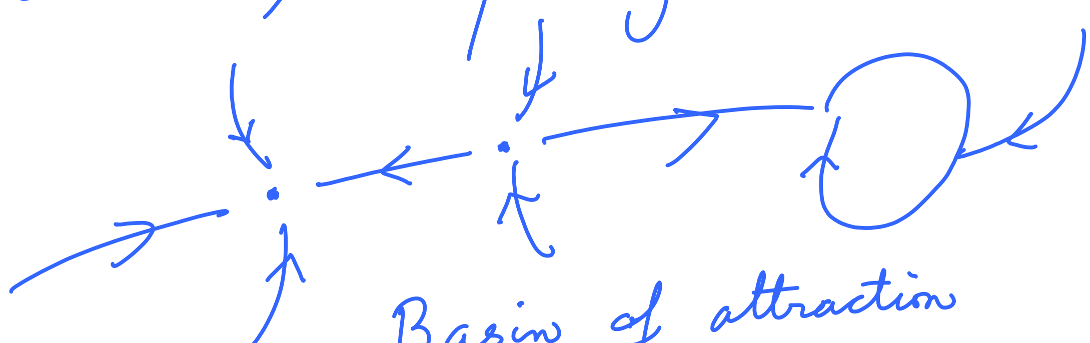
Expectation is over the p.d.f of random data vector $\underline{X}$

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

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

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



Basin of attraction

(Each attractor has a surrounding distinct region of its own $B(\underline{q})$ called basin of attraction)

The asymptotic stability theorem states that

$$\lim_{n \to \infty} \underline{w}[n] = \underline{q}, \quad \text{(infinitely often prob. 1)}$$

No idea how many iterations are needed!

Stability analysis of the max. eigen filter

To satisfy Condition (1) of the stability theorem,

Let $\eta[n] = \frac{1}{n}$

$$h(\underline{w}, \underline{x}) = \underline{x}(n)\, y(n) - y^2(n)\, \underline{w}(n)$$

$$= \underline{x}(n)\, \underline{x}^T(n)\, \underline{w}(n) - \left[\underline{w}^T(n)\, \underline{x}(n)\, \underline{x}^T(n)\, \underline{w}(n)\right] \underline{w}(n)$$

B

(B) Satisfies condition 3 of the Stability thm.

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

$$h = \lim_{n \to \infty} E\left(\underline{x}(n)\,\underline{x}^T(n)\,\underline{w}(n) - \left(\underline{w}^T(n)\,\underline{x}(n)\,\underline{x}^T(n)\,\underline{w}(n)\right)\underline{w}(n) \right)$$

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

Correlation matrix $E(xx^T)$ ignore 'n'

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

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

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,

$$R\,\underline{q}_k = \lambda_k \underline{q}_k \quad \text{--- (2)}$$

eigen vector

$$\underline{q}_k^T R\, \underline{q}_k = \lambda_k$$

eigen value

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

Using 1 and 2 in C

Consider $\underline{w}^T(t) R \underline{w}(t)$

eigen expansion

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

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

eigen value eqn

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

$$= \sum_{l=1}^{m} \lambda_l \, \theta_l^2(t) \qquad \text{—————} \quad (3)$$

Using (3) and (1)

$\left[\underline{w}^T(t) \, R \, \underline{w}(t)\right] \underline{w}(t)$ evaluates to

$$\underbrace{\sum_{l=1}^{m} \lambda_l \, \theta_l^2(t) \sum_{k=1}^{m} \theta_k(t) \, \underline{q}_k}$$

Thus,

$$\bar{h}(\underline{w}(t)) = \sum_{k=1}^{m} \lambda_k \theta_k(t)\, \underline{q}_k - \overbrace{\sum_{\ell=1}^{m} \lambda_\ell \theta_\ell^2(t)}^{\underline{w}^T(t)\, R\, \underline{w}(t)} \cdot \underbrace{\sum_{k=1}^{m} \theta_k(t)\, \underline{q}_k}_{\underline{w}(t)}$$

Written carefully,

$$\underbrace{\sum_{k=1}^{m} \frac{d}{dt} \theta_k(t)\, \underline{q}_k}_{\frac{d\underline{w}(t)}{dt}} = \overbrace{\sum_{k=1}^{m} \lambda_k \theta_k(t)\, \underline{q}_k - \sum_{\ell=1}^{m} \lambda_\ell \theta_\ell^2(t) \cdot \sum_{k=1}^{m} \theta_k(t)\, \underline{q}_k}^{\bar{h}(\underline{w}(t))} \qquad (4)$$

$\{\underline{q}_k\}_{k=1}^{m}$ are orthonormal vectors $\Rightarrow$ linear independence

Eqn (4) is a linear combination of linearly independent vectors whose coeffts are governed by the differential eqn given by

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

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

$\theta_k(t)$ are the principal modes

Case 1 : $1 < k \leq m$

For this treatment,

let $\alpha_k(t) \triangleq \dfrac{\theta_k(t)}{\theta_1(t)}$ $\quad 1 \leq k \leq m$

$\theta_1(t) \neq 0$ w. prob 1 and $\underline{w}(0)$ is randomly chosen

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

$$\frac{1}{\theta_1(t)}\frac{d\theta_k(t)}{dt} = \frac{1}{\theta_1(t)}\left[\lambda_k\theta_k(t) - \theta_k(t)\sum_{l=1}^{m}\lambda_l\theta_l^2(t)\right] \quad 5(a)$$

$$-\frac{\theta_k(t)}{\theta_1^2(t)}\frac{d\theta_1(t)}{dt} = -\frac{\theta_k(t)}{\theta_1^2(t)}\left[\lambda_1\theta_1(t) - \theta_1(t)\sum_{l=1}^{m}\lambda_l\theta_l^2(t)\right] \quad 5(b)$$

Summing the terms in 5(a) & 5(b)

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

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

$$= -\overbrace{\left(\lambda_1 - \lambda_k\right)}^{+ve}\alpha_k(t)$$

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

Assuming eigen values of $\underline{R}$ are distinct

and $\lambda_1 > \lambda_2 > \cdots > \lambda_m$

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

$$(\lambda_1 - \lambda_k) \propto \frac{1}{\text{time constant}}$$

$$\Rightarrow \alpha_k(t) \xrightarrow[t \to \infty]{= +ve} 0 \quad \text{for } 1 < k \leq m$$

CASE 2 : $k = 1$

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

$$= \lambda_1 \theta_1(t) - \lambda_1 \theta_1^3(t) - \theta_1(t) \sum_{l=2}^{m} \lambda_l \theta_l^2(t) \quad \text{———} \quad (6)$$

From Case 1 of our analysis, $\alpha_l \xrightarrow[t \to \infty]{} 0$, $l \neq 1$

0

$\therefore$ The governing equation is

$$\frac{d\theta_1(t)}{dt} = x_1\, \theta_1(t)\left(1 - \theta_1^2(t)\right) \quad \text{(asymptotically)} \qquad (7)$$

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 equillibrium state $\underline{s}^{(eq)}$ of the system is automatically stable if

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

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

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

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

(You may question this @ this stage)

To validate the assertion,

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

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

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

But $$\frac{d\theta_1(t)}{dt} = \lambda_1\,\theta_1(t) - \underbrace{\theta_1(t)\sum_{l=1}^{m}\lambda_l\,\cancel{\theta_l^2(t)}}_{\text{asymptotically } 0 \text{ except } l=1}$$

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

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

$$= -\lambda_1\,\theta_1(t)\left(\theta_1^2(t) - 1\right) \qquad \text{——} \quad (9b)$$

Plug (9b) in (9a)

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

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 2^{nd} condition is also satisfied.

$$\theta_1(t) \xrightarrow[t \to \infty]{} \pm 1$$

$$\theta_k(t) \xrightarrow[t \to \infty]{} 0 \qquad 1 < k \le m \quad (\text{Case } A)$$

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

$$\underline{w}(n+1) = \underline{w}(n) + \eta\left(\underline{x}(n)\,\underline{x}^T(n)\,\underline{w}(n) - \underline{w}^T(n)\,\underline{x}(n)\,\underline{x}^T(n)\,\underline{w}(n)\cdot\underline{w}(n)\right)$$

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

2) $\theta_1(t)$ will converge to ± 1

Formally stated,

$$\underline{w}(t) \xrightarrow[t \to \infty]{} \underline{q}_1$$ 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 $\mathcal{A}$ of the set of vectors / $\lim_{n \to \infty} \underline{w}(n) = \underline{q}$, $\left(\underbrace{\text{i.o.}}_{\text{infinitely often}} \text{ with prob. } 1\right)$

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'.

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

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

$\Longrightarrow$ As iterations $n \to \infty$, $\underline{w}(n)$ will eventually be within A & will remain inside A i.o. with prob. 1

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

Rest of all the conditions in AST are met,

$\Rightarrow$ $\underline{w}(n)$ converges to $\underline{q}_1$ 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 $y(n)$; $\sigma^2(n) = E(y^2(n))$ scalar

i.e., Prove $\sigma^2(n) \xrightarrow[n \to \infty]{} \lambda_1$

2) Per Oja's update, Hebbian based filter converges to a fixed point with prob. 1.

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

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