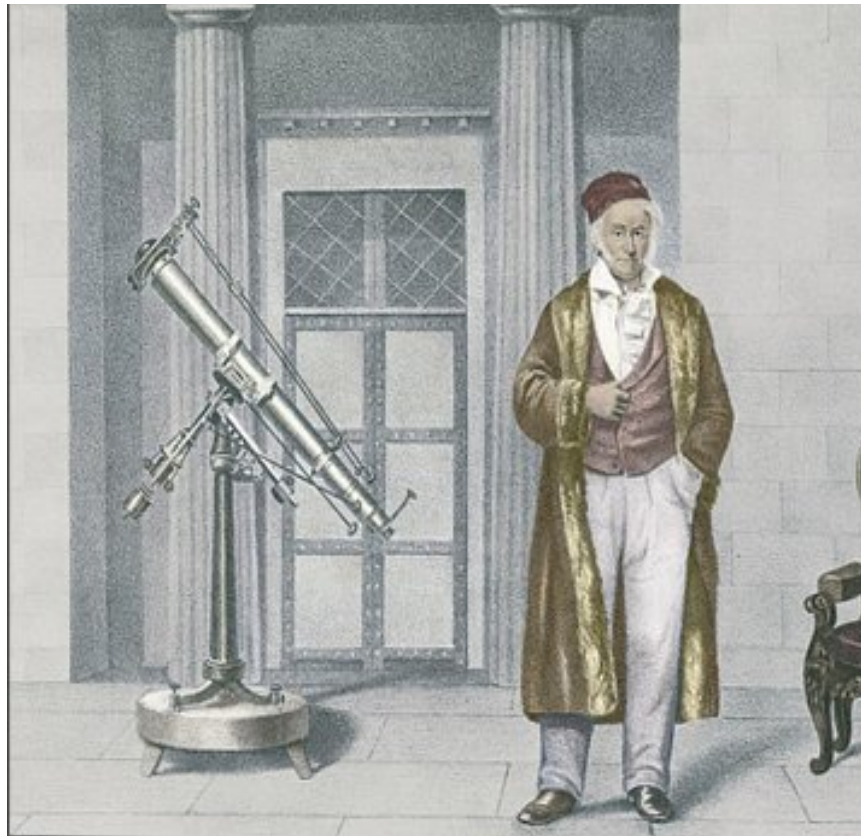
2.160 Identification, Estimation, and Learning

Part 1 Regression

Lecture 3

Recursive Least Squares with Forgetting Factor

H. Harry Asada
Department of Mechanical Engineering
MIT



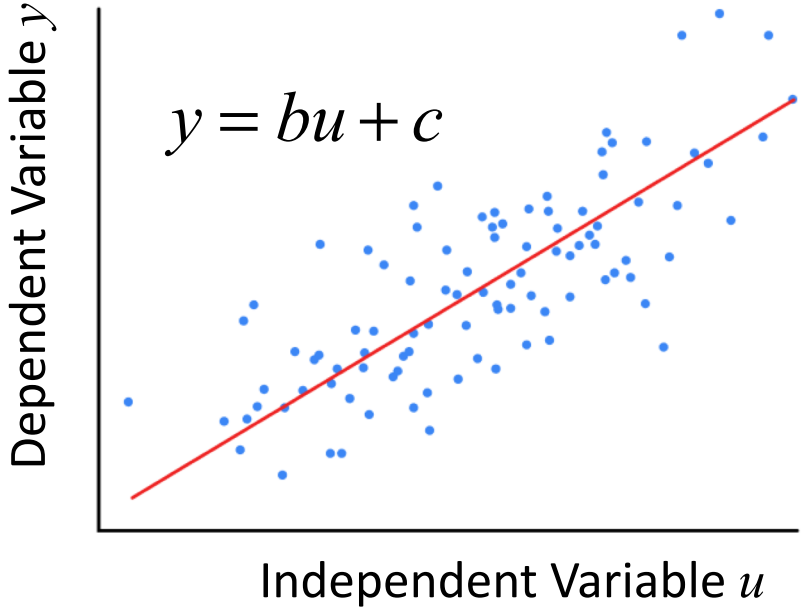
Carl Friedrich Gauss with his telescope

Recap

Linear Regression

$$y = b_1 u_1 + b_2 u_2 + \cdots + b_m u_m = \theta^T \varphi$$

Parameters $\theta = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \Re^{m\times 1}$ Regressor $\varphi = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix} \in \Re^{m\times 1}$



Example 1. Finite Impulse Response (FIR) Model

$$y(t) = g_1 u(t-1) + g_2 u(t-2) + \cdots + g_m u(t-m)$$

$$\theta = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \Re^{m\times 1} \qquad \varphi = \begin{pmatrix} u(t-1) \\ \vdots \\ u(t-m) \end{pmatrix} \in \Re^{m\times 1}$$

Example 2. Nonlinear function

$$y(t) = b_1 x_1 + b_2 x_1^3 + b_3 x_1 x_2 + b_4 e^{-3x_1}$$

Parameters are linearly involved, if we use the following regressor:

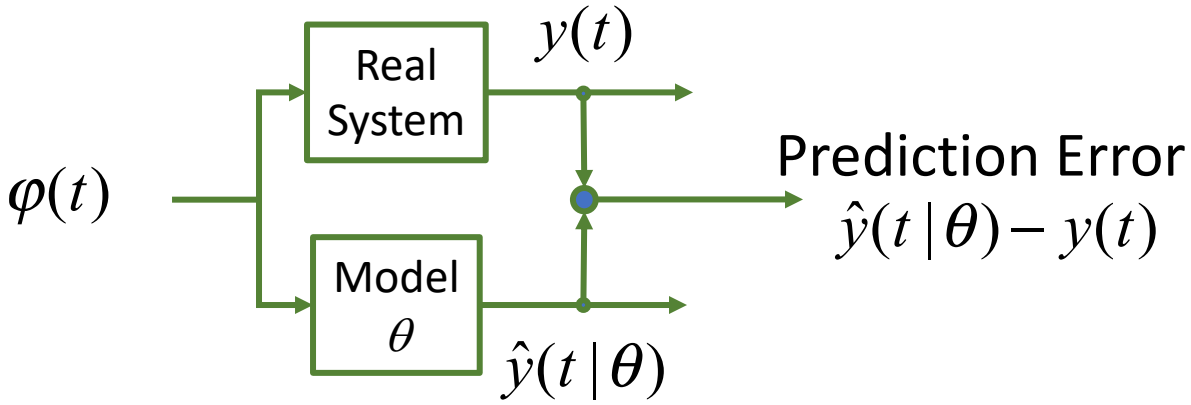
$$\varphi = \begin{pmatrix} x_1 \\ x_1^3 \\ x_1 x_2 \\ e^{-3x_1} \end{pmatrix} \in \Re^{4\times 1} \qquad \theta = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} \in \Re^{4\times 1}$$

Recap

Least Squares Estimate (LSE)

For finding parameters from data, $y(t)$ and $\varphi(t)$

Prediction – Error Formalism



Mean Squared Error:

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}(\hat{y}(t|\theta) - y(t))^2$$

Least Squares Estimate (LSE) provides the parameter vector that minimizes the above Mean Squared Error:

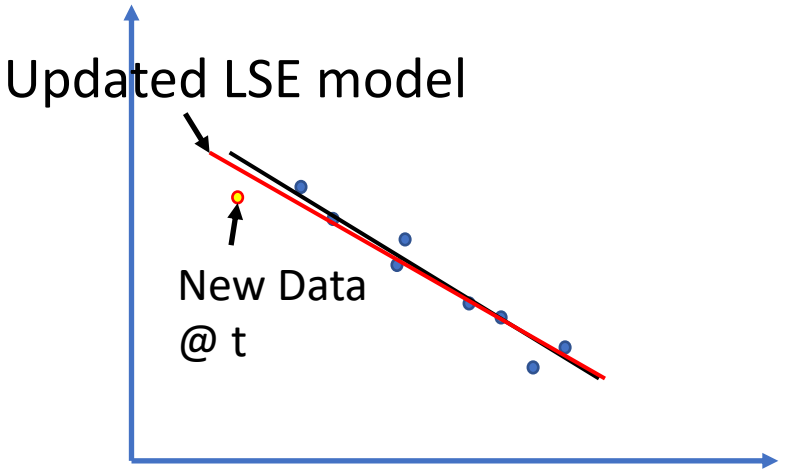
$$\hat{\theta}^{LS} = \arg\min_{\theta} V_N(\theta)$$

Problem:

Given a set of data

$$\varphi^{(1)}, y^{(1)}$$
$$\varphi^{(2)}, y^{(2)}$$
$$\vdots$$
$$\varphi^{(N)}, y^{(N)}$$

Find parameter vector θ



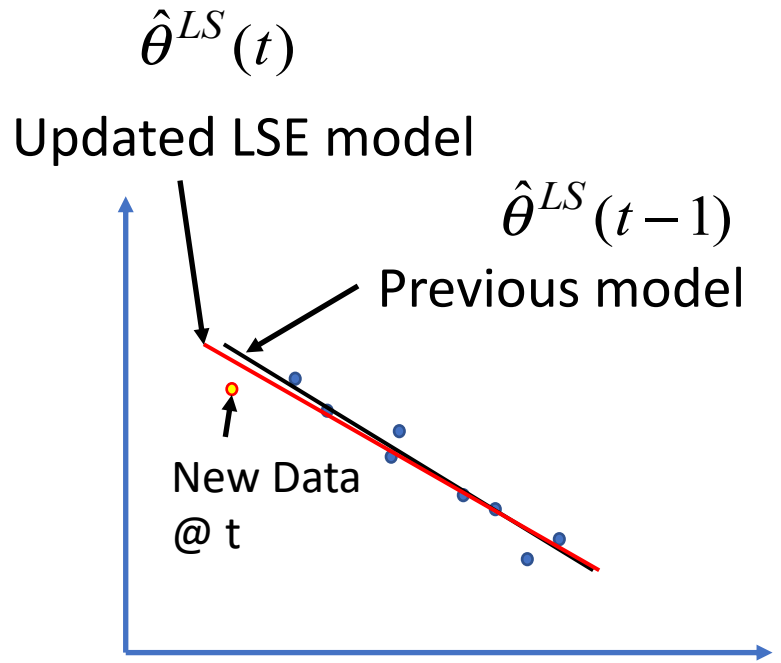
Recap

The Recursive Least Squares Algorithm

$$\hat{\theta}^{LS}(t) = \hat{\theta}^{LS}(t-1) + K_t[\, y(t) - \hat{y}(t \mid \hat{\theta}^{LS}(t-1))]$$

where $$K_t = \frac{P_{t-1}\varphi(t)}{1+\varphi^T(t)P_{t-1}\varphi(t)} \qquad t = 1,2,\cdots$$

$$P_t = P_{t-1} - \frac{P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}}{1+\varphi^T(t)P_{t-1}\varphi(t)}$$



Initial conditions:

$\hat{\theta}^{LS}(0) = \hat{\theta}_0 =$ Arbitrary, e.g. $\hat{\theta}_0 = 0$

$P_0 =$ Positive Definite Matrix, e.g. $P_0 = I$ Identity Matrix

Carl Friedrich Gauss discovered the Recursive Least Squares Algorithm in 1821.

2.3 Physical Meaning of Matrix P

$$\hat{\theta}^{LS}(t) = \hat{\theta}^{LS}(t-1) + K_t[\, y(t) - \hat{y}(t \mid \hat{\theta}^{LS}(t-1))]$$

$$P_t = P_{t-1} - \frac{P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}}{1+\varphi^T(t)P_{t-1}\varphi(t)} \quad (14) \qquad K_t = \frac{P_{t-1}\varphi(t)}{1+\varphi^T(t)P_{t-1}\varphi(t)}$$

The optimal gain K_t can be written as $\boxed{K_t = P_t\varphi(t)}$

Proof

Post-multiply $\varphi(t)$ to (14) and then multiply $1+\varphi^T(t)P_{t-1}\varphi(t)$

$$(1+\varphi^T(t)P_{t-1}\varphi(t))P_t\varphi(t) = (1+\varphi^T(t)P_{t-1}\varphi(t))P_{t-1}\varphi(t) - P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}\varphi(t)$$
$$= P_{t-1}\varphi(t)$$

$$\therefore \quad P_t\varphi(t) = \frac{P_{t-1}\varphi(t)}{1+\varphi^T(t)P_{t-1}\varphi(t)} = K_t$$ This agrees with the optimal gain K_t

Error Correction

$$\Delta\theta = \hat{\theta}^{LS}(t) - \hat{\theta}^{LS}(t-1) = K_t[\, y(t) - \hat{y}(t \mid \hat{\theta}^{LS}(t-1))]$$

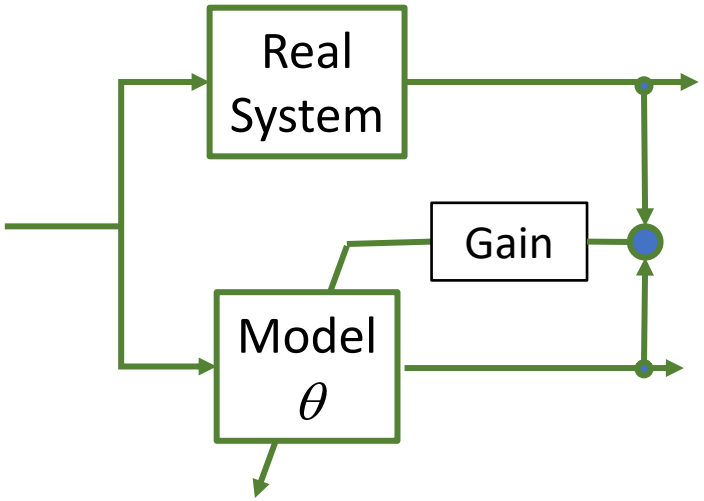
$$\therefore \Delta\theta = P_t \cdot \varphi(t) \cdot e(t)$$

(Parameter Estimation Correction)

= P_t x (Regressor) x (Prediction Error)

We can find a similar expression repeatedly in

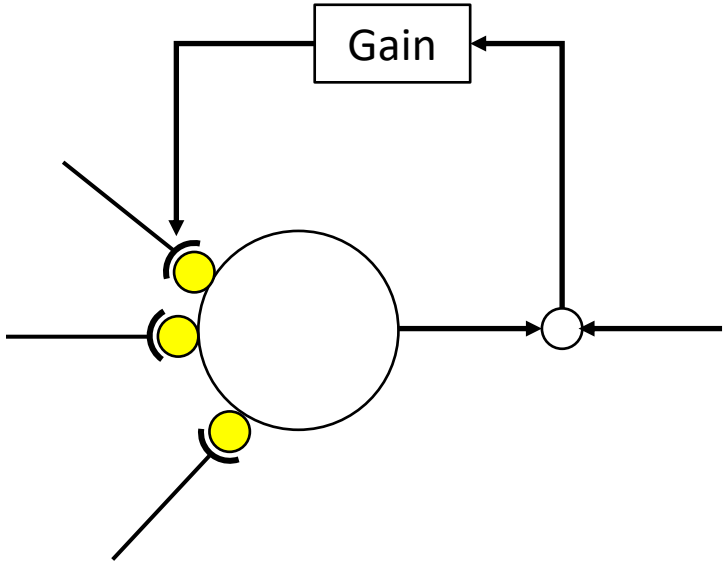
- Kalman filter
- Neuron model
- Error Backpropagation in multi-layer neural nets



Physical Sense of P_t?

$$P_t = \left[\sum_{i=1}^{t} \varphi(i)\varphi^T(i)\right]^{-1}$$

Or $$P_t^{-1} = \sum_{i=1}^{t} \varphi(i)\varphi^T(i)$$



Plot regressor data in the m-dimensional vector space.

$$\varphi(1),\varphi(2),\cdots\cdots,\varphi(t)$$

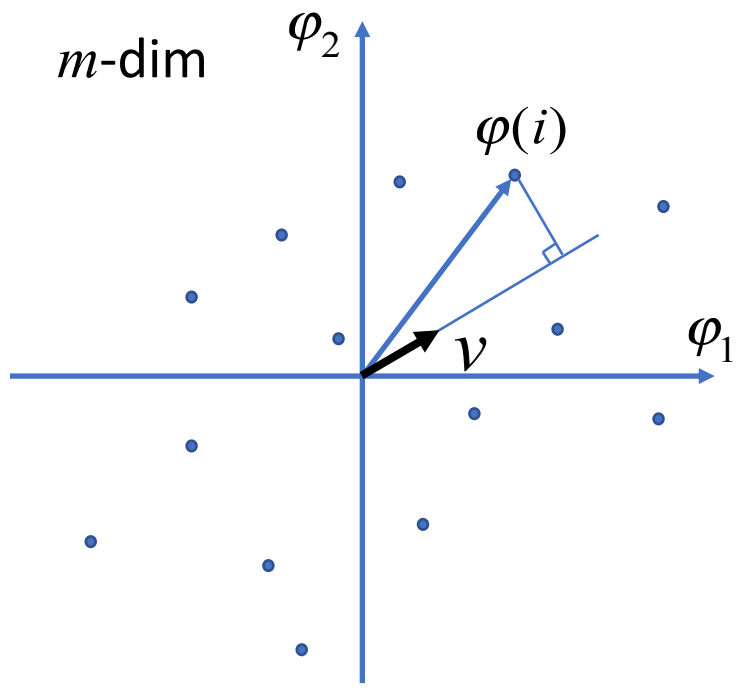
To characterize how the regressor data (input) are distributed, consider a unit vector v and quantify the total squared strength of $\varphi(1),\varphi(2),\cdots\cdots,\varphi(t)$ in the direction of vector v.

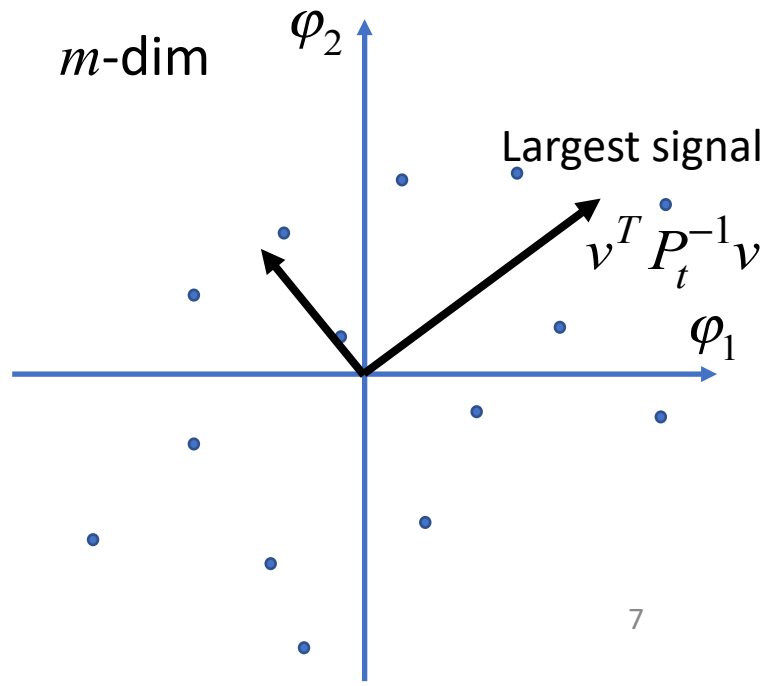
$$\sum_{i=1}^{t}(\varphi^T(i)\cdot v)^2=\sum_{i=1}^{t}v^T\varphi(i)\varphi^T(i)\cdot v$$

$$=v^T\sum_{i=1}^{t}\varphi(i)\varphi^T(i)\cdot v=v^T P_t^{-1}v$$ Quadratic form of P_t^{-1}

Find the max. (min.) of the total squared signal strength.
→ Find the direction of unit vector v that maximizes the quadratic function $v^T P_t^{-1}v$ subject to $|v|^2=1$.

This is an eigenvalue problem.





Quick Math Review: Eigenvalues

The maximization problem can be solved by using Lagrange multiplier λ.
We first define an extended cost functional including the constraint:

$$L = v^T \sum_{i=1}^{t} \varphi\varphi^T v - \lambda(v^T v - 1)$$

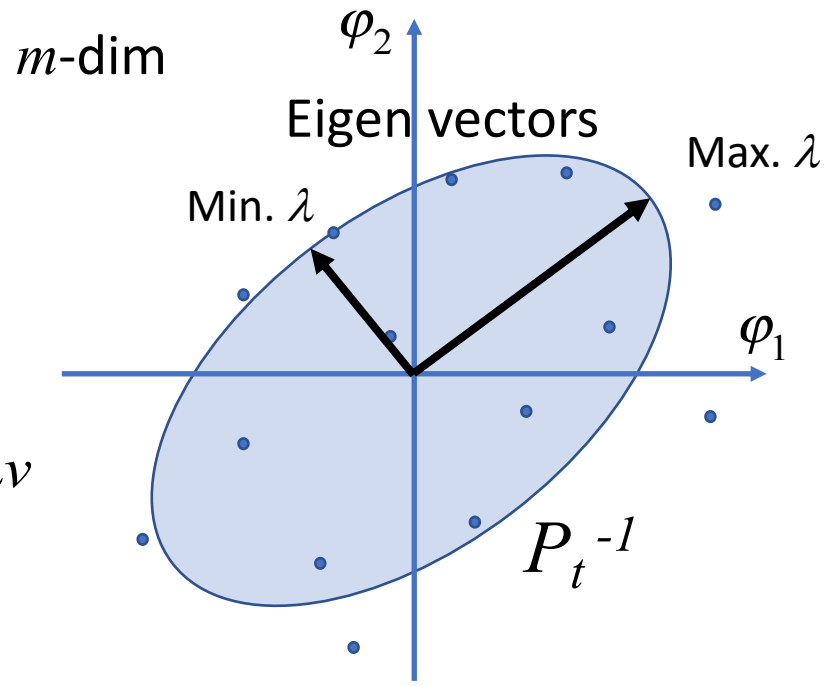
The max. direction of unit vector v is:

$$v = \arg\max_{v} L(v, \lambda)$$

Necessary conditions:

$$\frac{\partial L}{\partial v} = 0 \qquad 2\sum_{i=1}^{t} \varphi\varphi^T v - 2\lambda v = 0 \qquad \therefore \sum_{i=1}^{t} \varphi\varphi^T v = \lambda v$$

This implies that v is an eigenvector of matrix $\sum_{i=1}^{t} \varphi(i)\varphi^T(i)$



Math Review

The Lagrange Multiplier Method

Maximize (minimize) $f(x,y)$

Subject to $g(x,y)=0$

Differentiable

At a local maximum point A, the gradient of $f(x,y)$ is aligned with the gradient of $g(x,y)$

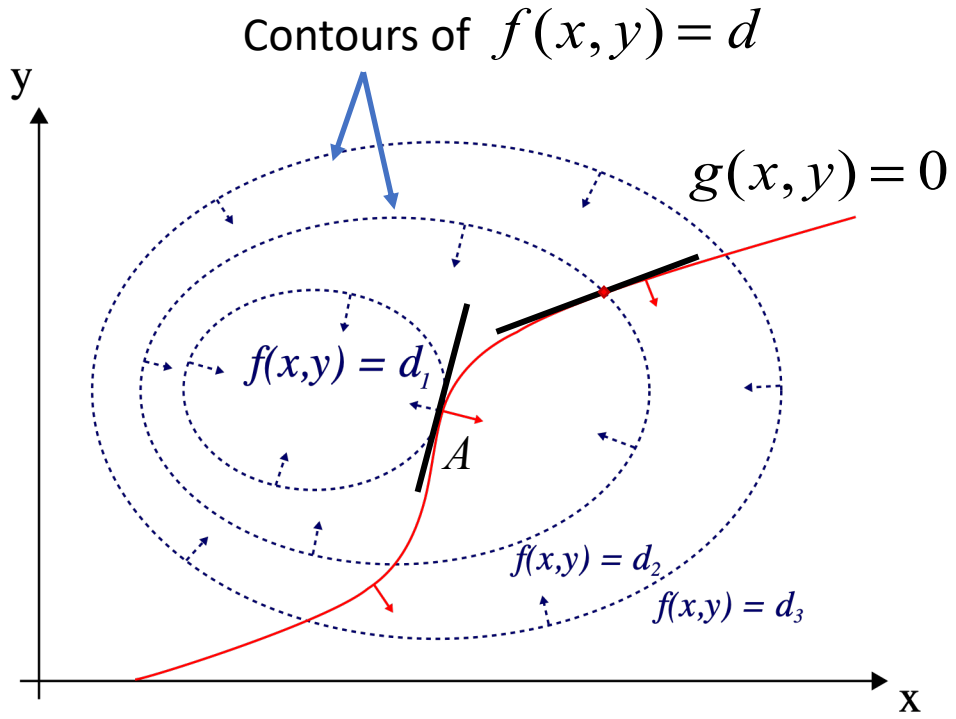
$$\begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \lambda \begin{pmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{pmatrix}$$

or $\nabla_{x,y} f(x,y) - \lambda \cdot \nabla_{x,y} g(x,y) = 0$

for some λ.

Define Lagrange Function:

$$L(x,y,\lambda) = f(x,y) - \lambda \cdot g(x,y)$$



The necessary condition for (x,y) is a maximum (minimum) point of $f(x,y)$ subject to $g(x,y)=0$ is :

$$\frac{\partial L}{\partial x} = 0$$

$$\frac{\partial L}{\partial y} = 0$$

$$\frac{\partial L}{\partial \lambda} = 0 \quad \rightarrow \quad g(x,y)=0$$

Quick Math Review: Eigenvalues

The maximization problem can be solved by using Lagrange multiplier λ.

We first define an extended cost functional including the constraint:

$f(x,y)$ → $g(x,y)=0$

$$L = v^T \sum_{i=1}^{t} \varphi\varphi^T v - \lambda(v^T v - 1)$$

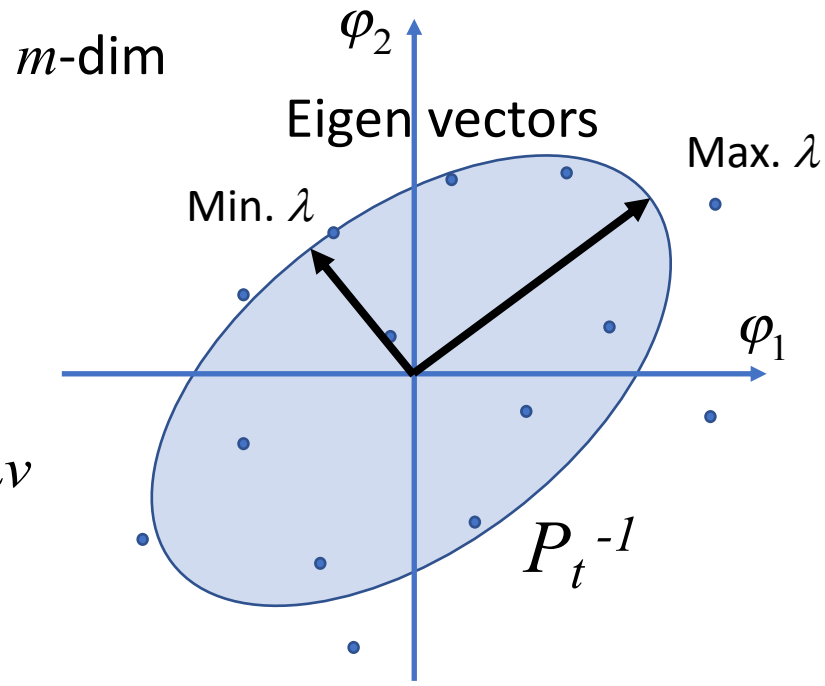
The max. direction of unit vector v is:

$$v = \arg\max_{v} L(v,\lambda)$$

Necessary conditions:

$$\frac{\partial L}{\partial v} = 0 \qquad 2\sum_{i=1}^{t} \varphi\varphi^T v - 2\lambda v = 0 \qquad \therefore \sum_{i=1}^{t} \varphi\varphi^T v = \lambda v$$

This implies that v is an eigenvector of matrix $\sum_{i=1}^{t} \varphi(i)\varphi^T(i)$



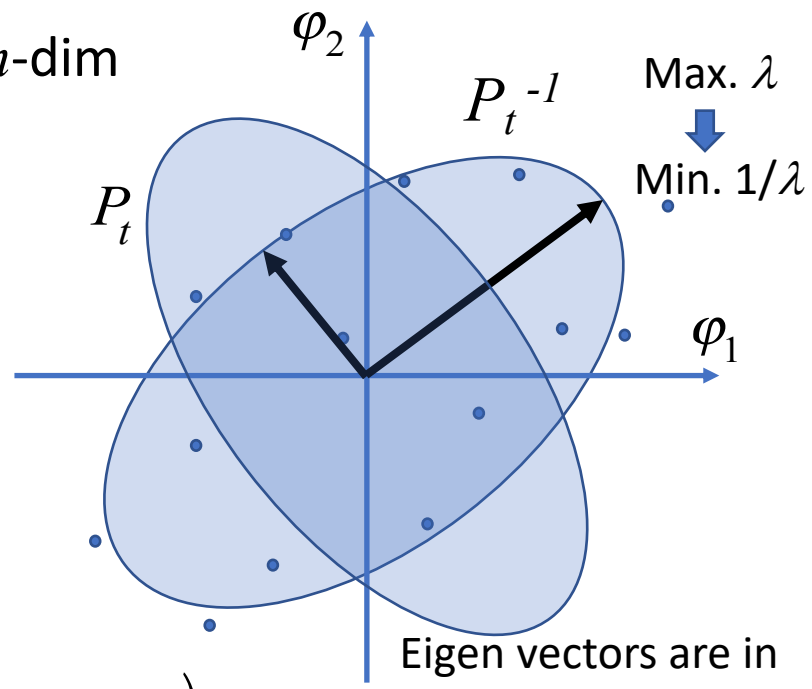
Error Correction

$$\Delta\theta = \hat{\theta}^{LS}(t) - \hat{\theta}^{LS}(t-1) = K_t[\, y(t) - \hat{y}(t \mid \hat{\theta}^{LS}(t-1))]$$

$$\therefore \Delta\theta = P_t \cdot \varphi(t) \cdot e(t)$$

Assume $\sum_{i=1}^{t} \varphi(i)\varphi^T(i)$ is non-singular.

$$P_t^{-1} = \sum_{i=1}^{t} \varphi(i)\varphi^T(i) = T \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix} T^T \qquad T = \begin{pmatrix} v_1 & \cdots & v_m \end{pmatrix}$$



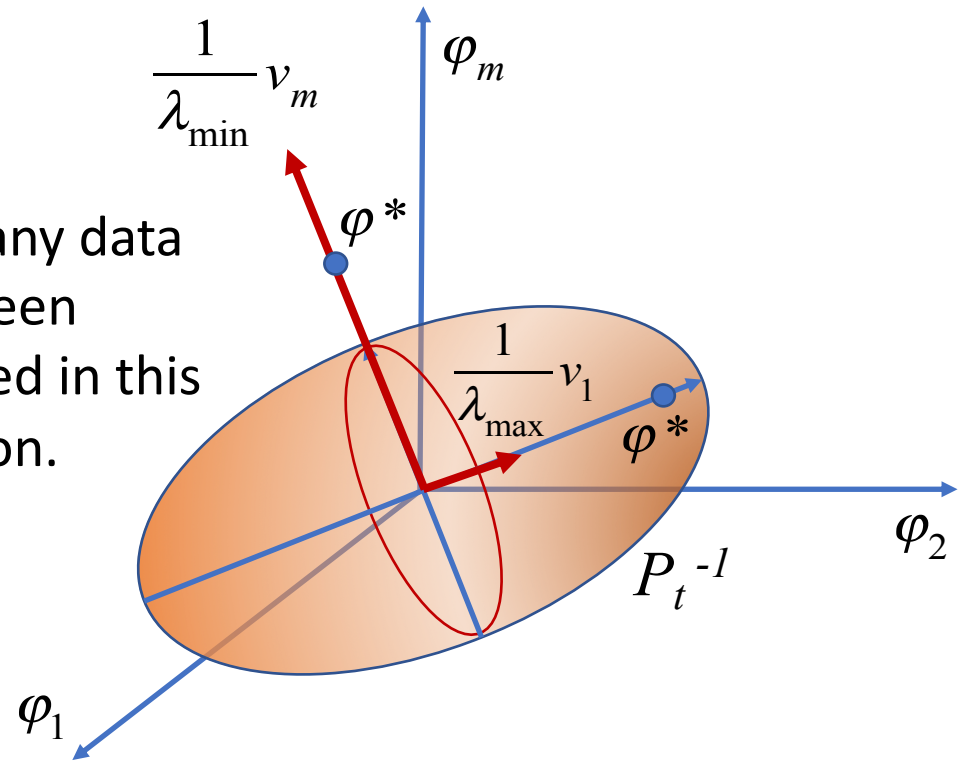
Eigen vectors are in the same directions. Eigenvalues become reciprocal.

$$\left(\sum_{i=1}^{t} \varphi(i)\varphi^T(i) \right)^{-1} = (T^T)^{-1} \begin{pmatrix} \frac{1}{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_m} \end{pmatrix} T^{-1} = T \begin{pmatrix} \frac{1}{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_m} \end{pmatrix} T^T$$

$$T^T = T^{-1}$$

Orthonormal

Error Correction $\Delta\theta = P_t \cdot \varphi(t) \cdot e(t)$



Not many data have been obtained in this direction.

Not strong signal strength in this direction:
Not familiar.

Rely more on new data in this direction

$$P_t \varphi^* = \frac{1}{\lambda_{\min}} \varphi^* \qquad |\Delta\theta| \text{ is large.}$$

Suppose that a new sample φ^* (test point) is observed, which is on the line of v_1 associated with the largest eigenvalue of matrix $\sum_{i=1}^{t} \varphi(i)\varphi^T(i)$

In the direction of v_1, we have already obtained a lot of data; we already know well in this direction.

No need to make a large correction to $\Delta\theta$

$$|P_t \varphi^*| \propto \frac{1}{\lambda_{\max}} \qquad P_t \varphi^* = \frac{1}{\lambda_{\max}} \varphi^*$$

$|\Delta\theta|$ is small.

In summary, the correction gain $K_t = P_t \varphi(t)$ is attenuated depending on how much the system already know in each direction:

Less known → large gain
Plenty of data → small gain

Quick Math Review: Positive-Definiteness

Consider a real square matrix $A \in \Re^{m \times m}$

$x^T Ax > 0$ For all $x \in \Re^{m \times 1}$ Other than $x \neq 0$ $\Longleftrightarrow$ Matrix A is Positive-Definite.

If A is given by $A = \sum_i \varphi(i)\varphi^T(i)$, then A is real, symmetric, and

1) A is at least positive semi-definite with eigenvalues:

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m = \lambda_{\min} \geq 0$$

$$\vdots \quad \vdots \quad \vdots$$

$v_1 \quad v_2 \quad v_m$ Corresponding eigen vectors

2) A can be decomposed to

$$A = \begin{pmatrix} v_1 & \cdots & v_m \end{pmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_m^T \end{pmatrix}$$

Orthogonal Matrix

Eigen Decomposition

$$T \in \Re^{m \times m}, \quad T^T = T^{-1}$$

Check property 1)

Consider Quadratic Form:

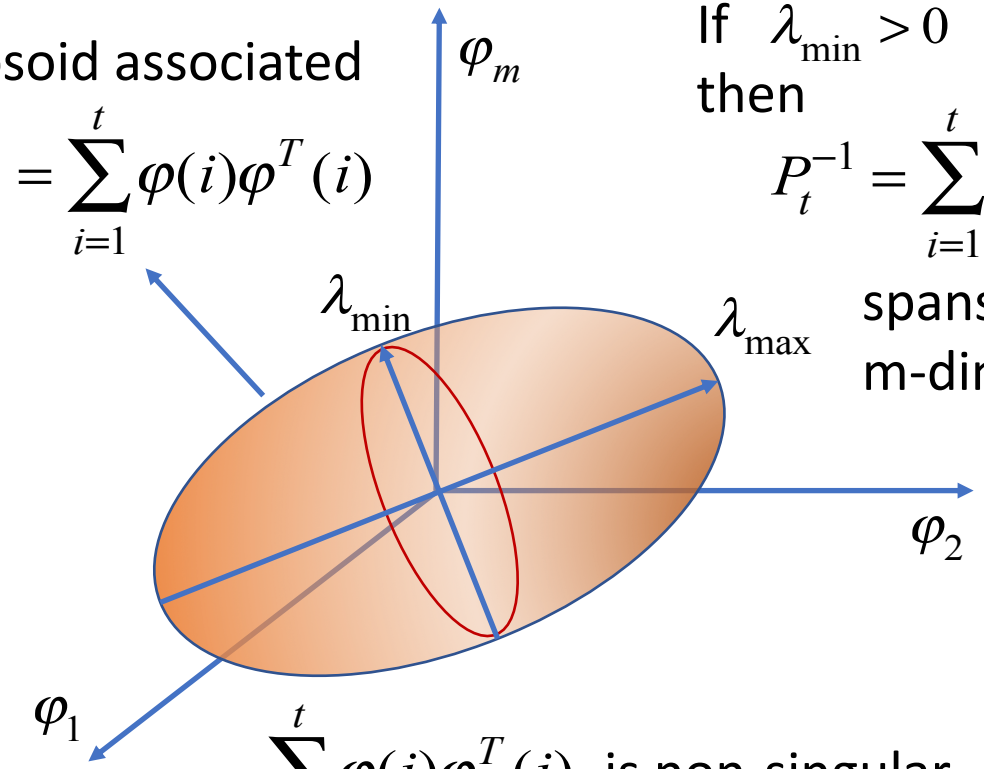
$$x^T Ax = x^T \sum \varphi\varphi^T x$$
$$= \sum (\varphi^T x)^2 \geq 0$$

This implies that matrix A is positive semi-definite.

Persistently Exciting Input / Regressor

m-dimensional regressor space

The ellipsoid associated with $P_t^{-1} = \sum_{i=1}^{t} \varphi(i)\varphi^T(i)$



If $\lambda_{\min} > 0$ then $P_t^{-1} = \sum_{i=1}^{t} \varphi(i)\varphi^T(i)$ spans the entire m-dim space.

$\sum_{i=1}^{t} \varphi(i)\varphi^T(i)$ is non-singular.

The data must cover the entire m-dimensional regressor space.

Persistently Exciting (PE)

If $v^T \sum_{i=1}^{t} \varphi(i)\varphi^T(i) \cdot v = 0$

for a particular v;

$$\sum_{i=1}^{t} (\varphi^T(i) \cdot v)^2 = 0$$ This implies

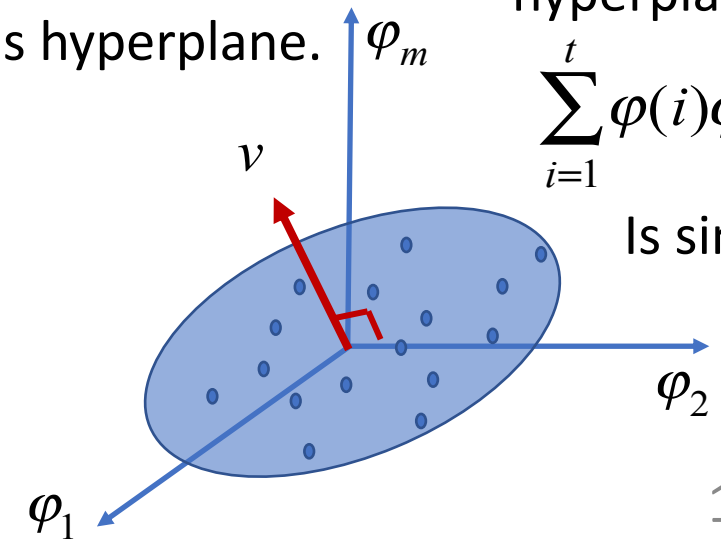
$$\varphi^T(i) \cdot v = 0, \quad 1 \le \forall i \le t$$

All the regressor points are confined within this hyperplane.

The ellipsoid collapses to the hyperplane.

$\sum_{i=1}^{t} \varphi(i)\varphi^T(i)$

Is singular.



Convergence Properties of the Recursive Formula

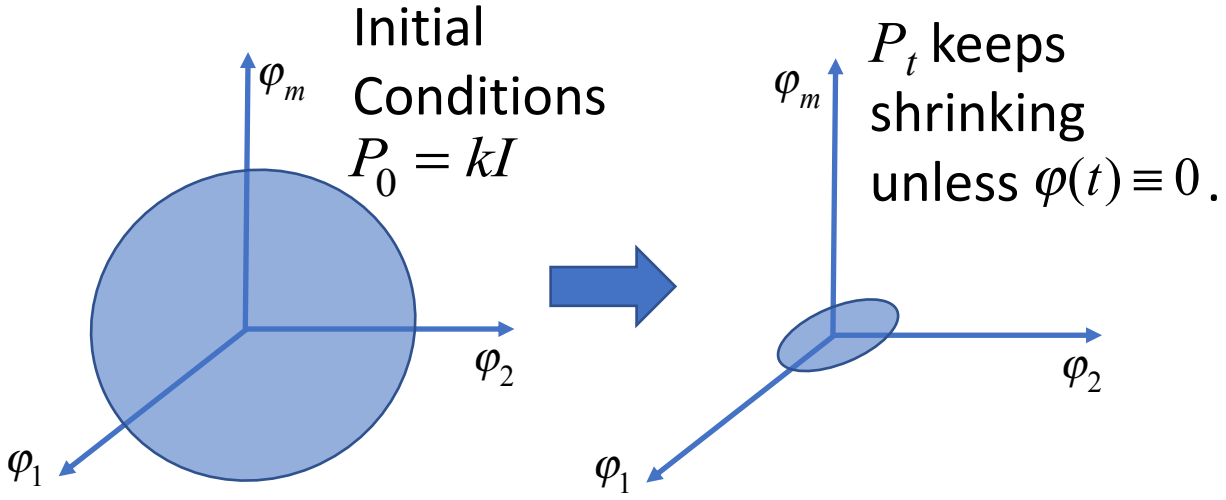
Incremental change to P_t

$$\Delta P = P_t - P_{t-1} = -\frac{P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}}{1+\varphi^T(t)P_{t-1}\varphi(t)} \leq 0$$

Negative semi-definite

Exercise: Show why it is negative semi-definite.

$P_t \leq P_{t-1}$ Non-increasing



Analysis of Convergence Properties:

$$\hat{\theta}^{LS}(t) \underset{t\to\infty}{\mapsto} converge$$

Recursive Least Squares (RLS) with initial conditions:

$\hat{\theta}(0) =$ Arbitrary

$P_0 =$ Positive definite

Converges in the sense $\lim_{t\to\infty}\left|\hat{\theta}(t)-\hat{\theta}(t-1)\right| = 0$

Goodwin and Sin, Chapter 3

Caveat! This does not mean that $\hat{\theta}^{LS}(t)$ converges to the true value θ_{true}.

It depends on input (regressor) $\varphi(t)$

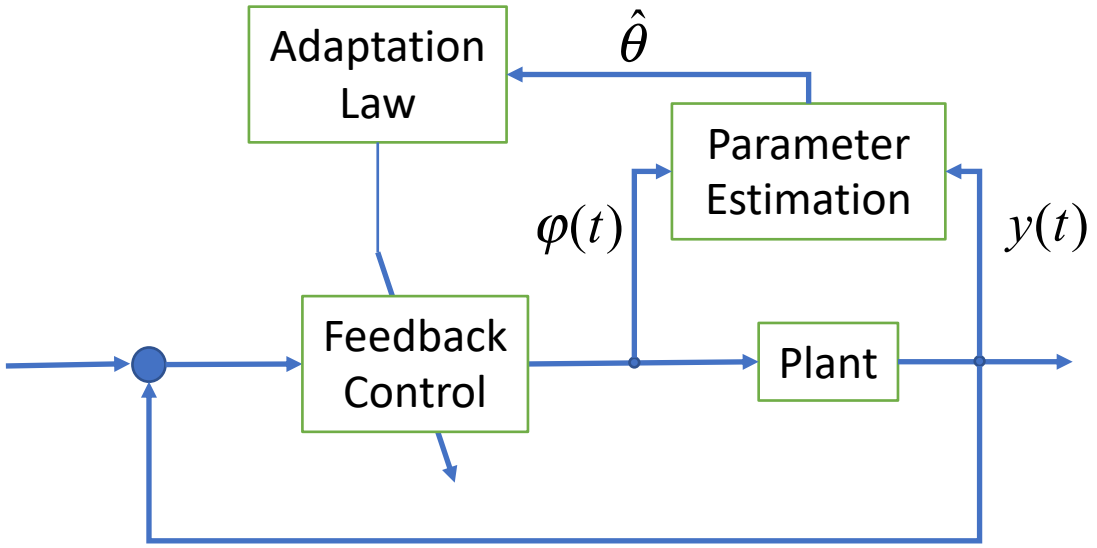
$$\hat{\theta}^{LS}(t) \underset{t\to\infty}{\mapsto} \theta_{true}$$

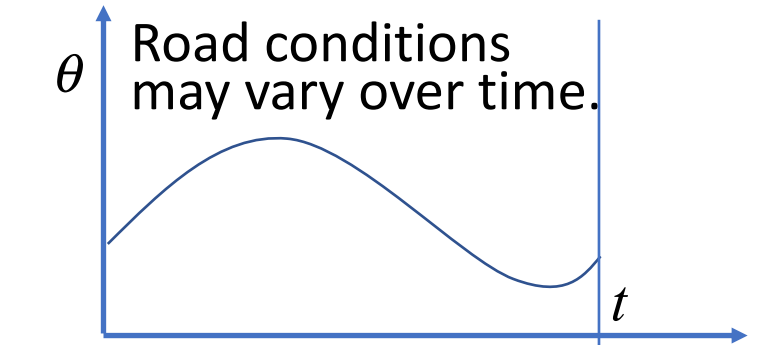
Persistent Excitation conditions required.

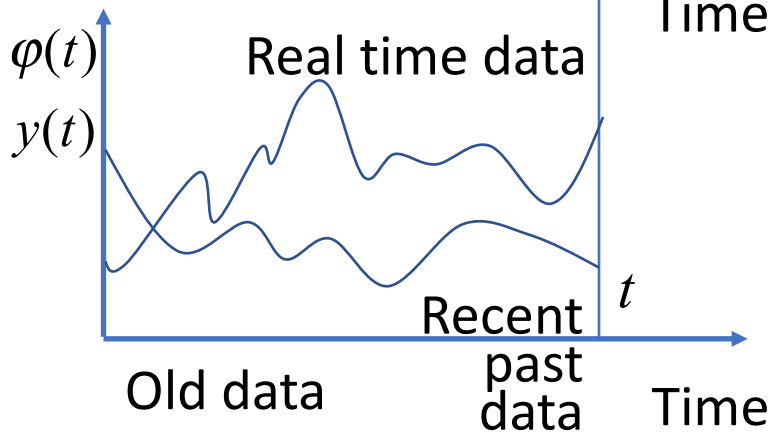
2.5 Estimation of Time-Varying Parameters

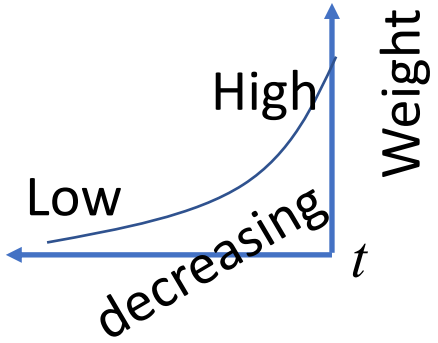
Smart traction + suspension control

Indirect Adaptive Control







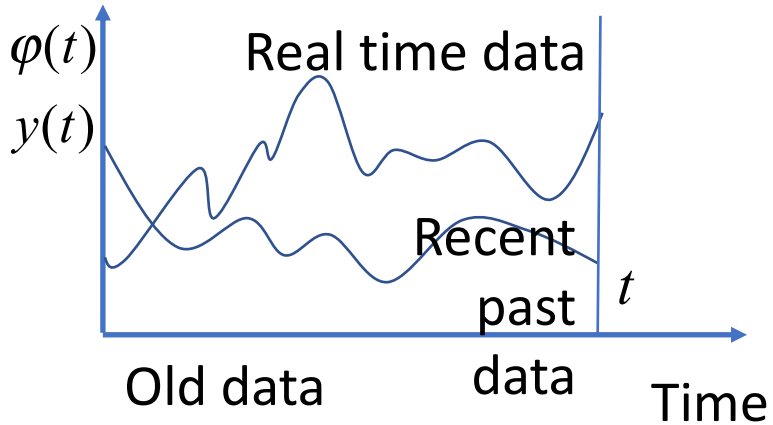


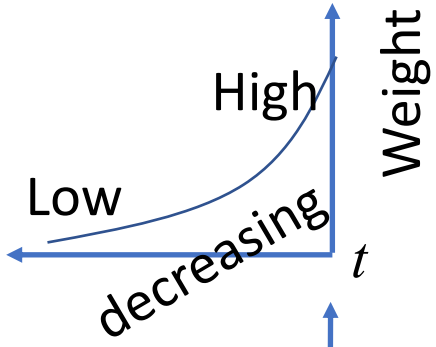
Plant conditions, such as road conditions, may vary over time. How can a control system estimate varying parameters in real time?

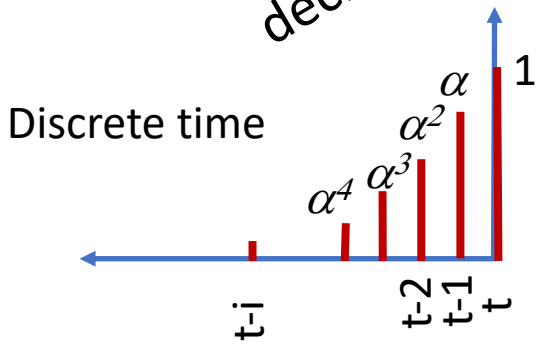
Suppose that the system keeps measuring regressors and output variables. The parameters should be estimated based on recent past data rather than older data.

This parameter estimation problem is important for the system to adapt to the varying environment conditions: Adaptive control.

Recursive Least Squares with Exponential Forgetting Factor







The original squared error: $V_t(\theta) = \frac{1}{t}\sum_{i=1}^{t}(\underbrace{y(i) - \hat{y}(i\,|\,\theta)}_{e(i)})^2$

Weighted squared error

$$J_t(\theta) = e^2(t) + \alpha e^2(t-1) + \cdots + \alpha^{t-1}e^2(1)$$

$$J_t(\theta) = \sum_{i=1}^{t}\alpha^{t-i}e^2(i)$$

$\alpha =$ Exponential Forgetting Factor; $0 < \alpha \le 1$

0.8
0.64
0.51
0.41
0.33
0.26
0.21
.
.

The Least Squares Estimate:

$$\hat{\theta}(t) = \arg\min_{\theta} J_t(\theta)$$

This least square estimate and its recursive version can be obtained in the same way as the previous case.

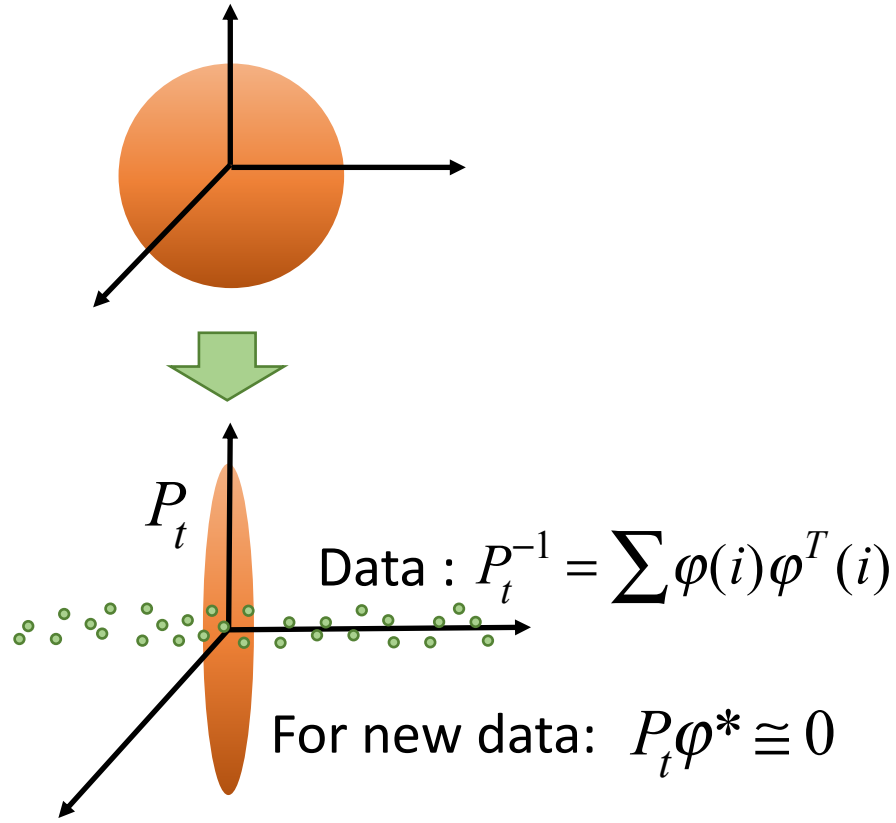
$\frac{dJ_t}{d\theta} = 0$ and apply the Matrix Inversion Lemma.

Recursive Least Squares with Exponential Forgetting Factor

$$\hat{\theta}(t) = \hat{\theta}(t-1) + K_t[\, y(t) - \hat{y}(t \mid \hat{\theta}(t-1))]$$

where $$K_t = \frac{P_{t-1}\varphi(t)}{\alpha + \varphi^T(t) P_{t-1}\varphi(t)} \qquad t = 1,2,\cdots$$

$$P_t = \frac{1}{\alpha}\left[P_{t-1} - \frac{P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}}{\alpha + \varphi^T(t) P_{t-1}\varphi(t)} \right] \qquad 0 < \alpha \le 1$$



The only difference is that 1 in the denominator is replaced by α and that the updated P matrix is divided by α.

Caveat!

When the system enters a "steady-state", producing the same $\varphi(t), y(t)$ for a long time, the total squared signal strength keeps increasing in that direction of the regressor vector $\rightarrow \quad P_t = \left(\sum \varphi(i)\varphi^T(i) \right)^{-1}$ becomes very thin.

Failure Scenarios

$$P_t = \frac{1}{\alpha}\left[P_{t-1} - \frac{P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}}{\alpha + \varphi^T(t)P_{t-1}\varphi(t)} \right]$$

When a car is running through a boring country side, almost constant data will be obtained, resulting in

$$P_{t-1}\varphi^* \cong 0.$$

However, the reciprocal of forgetting factor,

$$\frac{1}{\alpha} > 1$$

is multiplied repeatedly.

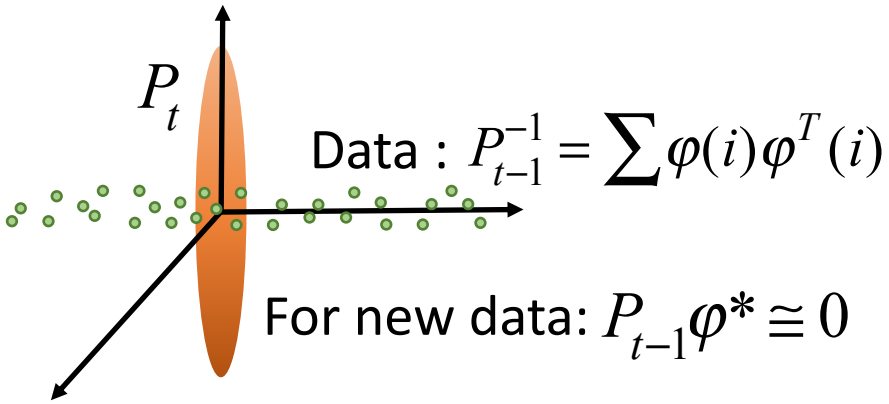
$$P_t = \frac{1}{\alpha}\left[P_{t-1} - \frac{P_{t-1}\varphi^*(t)\varphi^{*T}(t)P_{t-1}}{\alpha + \varphi^{*T}(t)P_{t-1}\varphi^*(t)} \right] \cong \frac{1}{\alpha} P_{t-1}$$

$$P_t \cong \frac{1}{\alpha^t} P_0 \quad \underset{t\to\infty}{\to} \quad \textbf{Blows up!}$$

$$1 < \alpha < 1$$

Driving along a cornfield

Boring, no change in sensor readings



Example

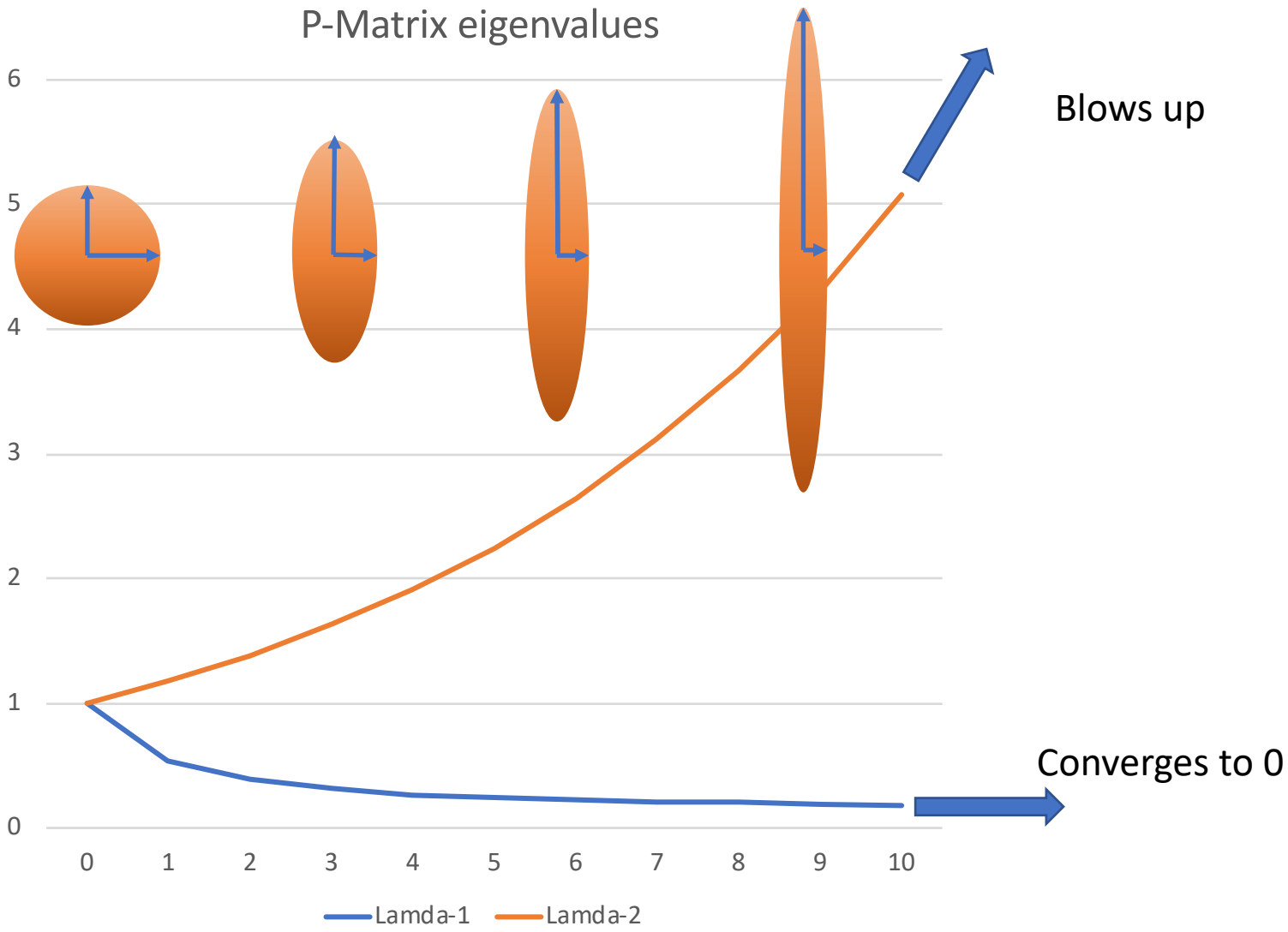
Recursive Least Squares
with Forgetting Factor

$$P_t = \frac{1}{\alpha}\left[P_{t-1} - \frac{P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}}{\alpha + \varphi^T(t)P_{t-1}\varphi(t)} \right]$$

$$P_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\varphi(t) = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$\alpha = 0.85$$

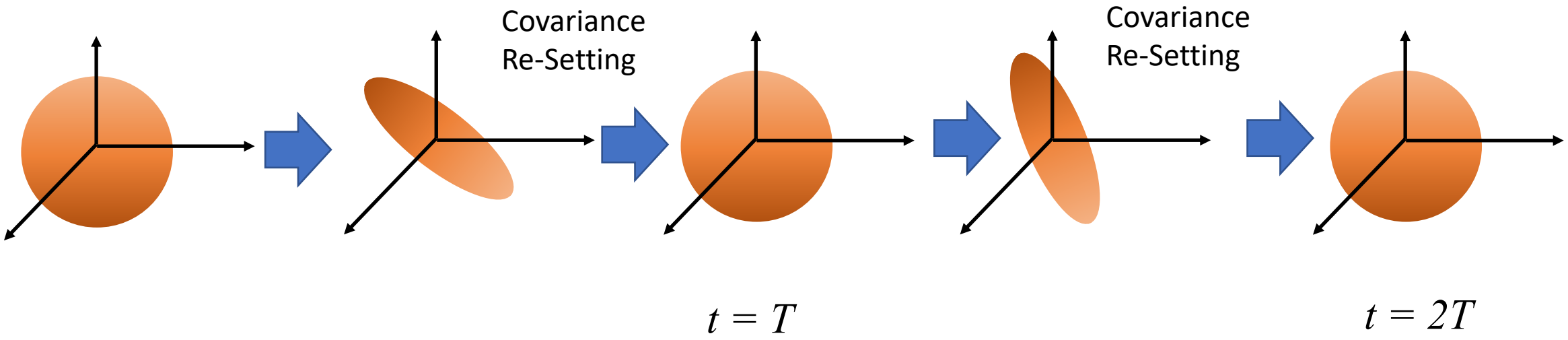


Remedy

- Covariance Re-setting
- Occasionally, P is re-set to a large positive-definite matrix, such as the identity matrix:

$$P_t = \kappa I, \quad 0 < \kappa < \infty, @t = T, 2T, 3T, ...$$

- This covariance re-setting revitalizes the RLS algorithm.



Example

Covariance Re-setting

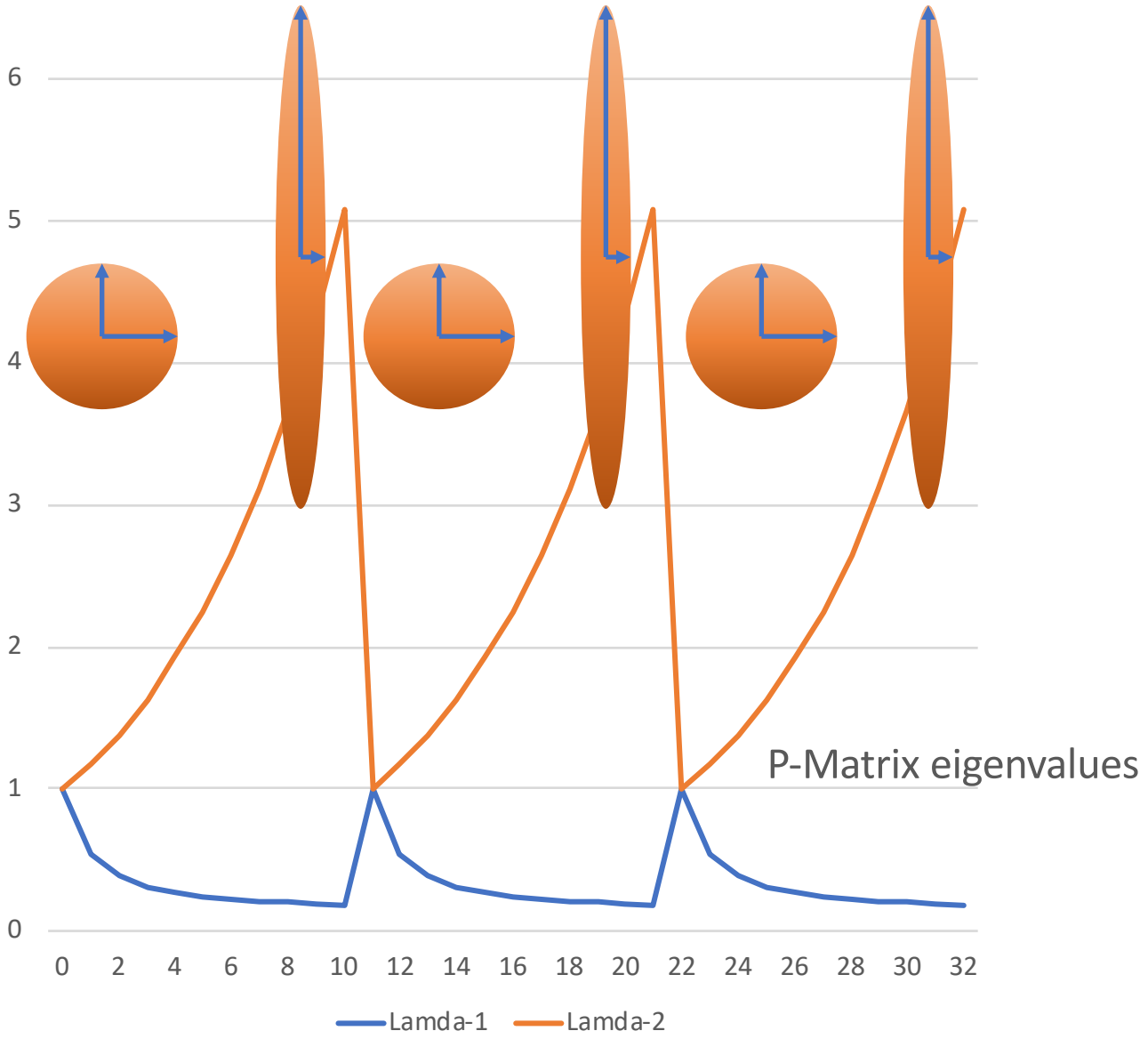
$$P_t = \frac{1}{\alpha}\left[P_{t-1} - \frac{P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}}{\alpha + \varphi^T(t)P_{t-1}\varphi(t)} \right]$$

$$T = 11$$

$$P_{kT} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad k = 0,1,2,3,.....$$

$$\varphi(t) = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$\alpha = 0.85$$



Remaining Topics

- ❑ Orthogonal Projection
- ❑ Multi-Output, Weighted Least Squares Estimate

See the Lecture Notes Chapter 2