SEARCHING THE PERFORMANCE SURFACE

Wiener -Hoff solution is analytic. Due to the special shape of the performance surface, we can think of alternative ways to find the minimum.

Search: random and CONSISTENT.

Since we know a lot about the surface, consistent. We will cover GRADIENT TECHNIQUES.

Concentrate on 2:

- NEWTON'S method: changes are always in the direction of <u>the minimum of the performance surface</u> for quadratic surfaces.
- STEEPEST DESCENT: changes are in the direction of the <u>negative gradient of the performance surface.</u> (it points to the minimum only along the principal axis of the ellipse).

All components of the weight vector change at the same time.

BASIC IDEA OF GRADIENT SEARCH

(One weigth space)

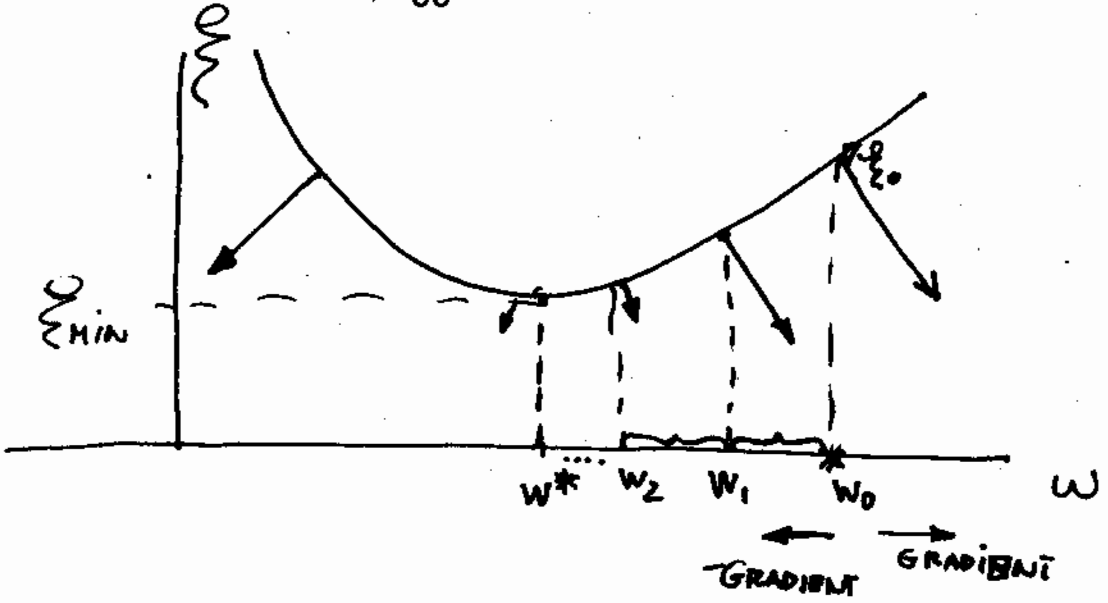
Consider the performance surface (a parabola)

$$\xi = \xi_{min} + V^{T} R V$$

.R has only one element r_{00}

$$\xi = \xi_{min} + r_{00}\,(w - w^{*})^{2} = \xi_{min} + \lambda\,(w - w^{*})^{2}$$

For the univariate case, $r_{00} = \lambda$



The problem is to find w^*

Start with an initial guess w_0 (notice that now the subscript means iteration number).

Using gradient search, one measures the gradient at w_0, and move in the surface by a quantity proportional to the NEGATIVE of the gradient.

At this new point w_1, the process is repeated until w^* is reached.

The process of measuring the gradient is called the gradient estimate, and we will assume that this value is available.

Why does this process work?

Gradient of the error points to the direction of increasing error.- Moving in the negative direction by a small amount, the move should decrease the error.

Problem is “the small amount” (may overshoot w^*).

For our univariate case,

$$\nabla = \frac{\partial \xi}{\partial w} = 2\lambda(w - w^*)$$

So , in general

$$w_{k+1} = w_k + \mu(-\nabla_k) =$$

$$= w_k - 2\lambda\mu(w_k - w^*).$$

where ∇_k is the estimate of the gradient at w_k.

Rearranging terms,

$$w_{k+1} = (1 - 2\lambda\mu)w_k + 2\mu\lambda w^*.$$

This is a linear, first order, constant coefficient difference equation. The solution can be obtained recursively,

$$w_1 = (1-2\lambda\mu)\, w_0 + 2\mu\lambda w^*$$

$$w_2 = (1-2\lambda\mu)\left[(1-2\lambda\mu)\, w_0 + 2\mu\lambda w^*\right] + 2\mu\lambda w^*$$

$$= (1-2\lambda\mu)^2 w_0 + 2\mu\lambda w^* (1+(1-2\lambda\mu))$$

$$w_k = (1-2\lambda\mu)^k w_0 + \sum_{i=1}^{k-1} (1-2\lambda\mu)^i \, 2\mu\lambda w^*$$

$$= (1-2\lambda\mu)^k w_0 + 2\mu\lambda w^* \frac{1-(1-2\mu\lambda)^k}{1-(1-2\lambda\mu)}$$

$$= (1-2\lambda\mu)^k \, (w_0 - w^*) + w^*$$

When k goes to infinity,

$$(1-2\lambda\mu)^k (w_0 - w^*) \longrightarrow 0$$

and w approaches w^*.

To meet this condition, the geometric ratio $r = 1-2\lambda\mu$

$$|1-2\lambda\mu| < 1$$

$$0 < \mu < \frac{1}{\lambda}$$

must be positive, smaller than 1/λ, the eigenvalue of R.

The rate of convergence (or the speed of convergence) is also obviously dependent on r, the geometric ratio.

when |r|<1, rate of convergence increases for smaller rs. For r=0 the minimum is reached in a single step.

For positive values of r, there is no oscilation (overdamped).

For negative values of r; overshoot the minimum (underdamped).

When $|r| \geq 1$ there will be no convergence.

THE LEARNING CURVE

We may want to know how the MSE progresses towards the minimum, for different w_k.

The MSE will converge towards the MSE_{min}.

$$\xi_k = \xi_{min} + \lambda (w_k - w^*)^2$$
$$= \xi_{min} + \lambda (w_0 - w^*)^2 (1 - 2\mu\lambda)^{2k}$$

The geometric ratio of this progression is

$$r_{mse} = r^2 = (1 - 2\mu\lambda)^2$$

r can never be negative, so the progression approaches assymptotically MSE_{min}.

The sequence of MSE_k is called the learning curve.

NEWTON'S ALGORITHM

Is a special case of the gradient search when r=0.

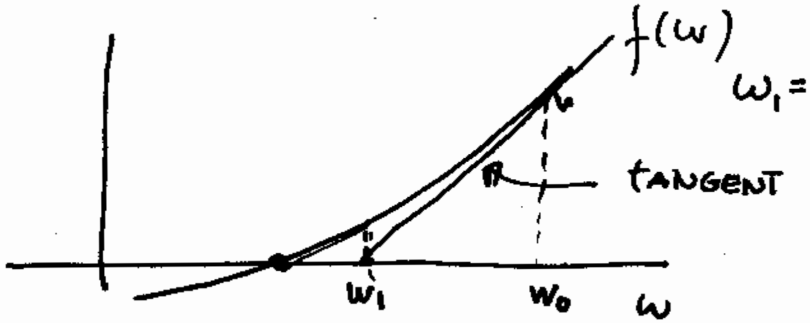
$$r=0 \Longrightarrow 1-2\lambda\mu=0 \Longrightarrow \mu=\frac{1}{2\lambda}$$

Note that to implement the Newton method one NEEDS to know the EIGENVALUES of R. So more info is required, but for QUADRATIC surfaces the minimum is reached in ONE step.

Normally Netwon's method is presented differently, because it evolved from polynomial root finding.

$$f(\omega)=0$$

Start with w=w0, calculate f'(w0) and calculate w1 as



$$\omega_1=\omega_0-\frac{f(\omega_0)}{f'(\omega_0)}$$

or in general

$$\omega_{k+1} = \omega_k - \frac{f(\omega_k)}{f'(\omega_k)}$$

Now for discrete problems, must compute the derivative by differences. The backward difference is normally used

$$f'(\omega_k) = \frac{f(\omega_k) - f(\omega_{k-1})}{\omega_k - \omega_{k-1}}$$

So

$$\omega_{k+1} = \omega_k - \frac{f(\omega_k)\ (\omega_k - \omega_{k-1})}{f(\omega_k) - f(\omega_{k-1})}$$

To apply this theory to the search of the minimum of the performance surface, note that at $\omega = \omega^*$

$$\nabla = 0$$

So the equation becomes $\xi'(\omega) = 0 \Longrightarrow f(\omega) = \xi'(\omega)$

and one needs to substitute f(w) by

and the expression becomes

$$\omega_{k+1} = \omega_k - \frac{\xi'(\omega_k)}{\xi''(\omega_k)} =$$

Note that we need to estimate both $\xi'(\omega_k)$ AND $\xi''(\omega_k)$

For quadratic surfaces this search is very fast. For nonquadratic, the search may not converge.

$$\xi = \xi_{min} + \lambda(\omega-\omega^*)^2$$

$$\frac{\partial \xi}{\partial w_k} = 2\lambda\,(w_k - w^*) \quad ; \quad \frac{\partial \xi^2}{\partial w_k^2} = 2\lambda$$

So $$\omega_{k+1} = \omega_k - \frac{2\lambda(\omega_k - \omega^*)}{2\lambda} = \underline{\omega^*}$$

EXTENSION TO MULTIDIMENSIONAL CASES

In multi-D, the gradient vector is

$$\nabla = 2\bar{R}\,\vec{w} - 2\bar{P}$$

Left multiplying by 1/2 $R^{-1}$and substituing $W^{*}=R^{-1}P$,

$$\tfrac{1}{2} R^{-1}\nabla = W - 2R^{-1}P = W - W^{*}$$

So $\quad W^{*} = W - \tfrac{1}{2} R^{-1}\nabla$

This gives the recursive equation

$$\bar{W}_{k+1} = \bar{W}_k - \tfrac{1}{2}\,\underline{\bar{R}}^{-1}\bar{\nabla}_k$$

and is the Newton's method for L dimensions.

$$\xi = E\left[d^2(k)\right] - 2\underline{P}^T\underline{W} + \underline{W}^T R\,\underline{W}$$

For quadratic surfaces, we obtain the minimum in one step (k=1). However we can use other values of μ (other than 1/2).

$$W_{k+1} = W_k - \mu R^{-1} \nabla_k$$

$$0 < \mu < 1$$

Any value between 0 and 1 will make the geometric projection ot converge.

Calculating the gradient $\frac{\partial \xi}{\partial w} = 2RV = 2R(w - w^*)$

we get $\omega_{k+1} = \omega_k - 2\mu(\omega_k - \omega^*) =$

$$w_{k+1} = (1-2\mu)\,\omega_k + 2\mu\omega^*$$

So the iterative solution has the same form as before.

$$\bar{\omega}_k = \bar{\omega}^* + (1-2\mu)^k(\bar{\omega}_0 - \bar{w}^*)$$

STEEPEST DESCENT

Weights are adjusted in the direction of the gradient. The method is easier to implement and of more consistent performance.

Multi-D extension is immediate

$$\bar{W}_{k+1} = \bar{W}_k - \mu \bar{\nabla}_k$$

Now calculating gradient $\nabla_k = 2\bar{R}(\bar{w}_k - \bar{w}^*)$

$$\bar{W}_{k+1} = (I - 2\mu R)\,\bar{w}_k + 2\mu \bar{R}\bar{W}^*$$

Since R is not diagonal W_{K+1} will depend on W_k and on all other coefficients through R. In the principal coordinate system,

$$\bar{V}_{k+1} = \bar{V}_k - 2\mu R\bar{V}_k = (I - 2\mu\bar{R})\,\bar{V}_k$$

$$\bar{Q}\bar{V}'_{k+1} = (I - 2\mu R)\,Q V'_k$$

$$= (I - 2\mu\Lambda)\,V'_k$$

As Λ is diagonal, we can talk about individual solutions along the principal axes

$$V'_{k+1} = (I - 2\mu\Lambda)^{k} V'_0$$

This can be written <u>separately</u>

$$v'_i(k+1) = (1 - 2\mu\lambda_i)\, v'_i(k) \quad 0 \le i \le L-1$$

For convergence, <u>each</u> equation must satisfy

$$|1 - 2\mu\lambda_i| < 1 \Longrightarrow 0 < \mu < \frac{1}{\lambda_i}$$

So, to have convergence in the N dimensional system, it is necessary and sufficient that

$$0 < \mu < \frac{1}{\lambda_{max}}$$

(in practice for smooth convergence, μ should be 2 orders of magnitude smaller than λ_{max})

In each axis there is a straight geometric progression for the sequence of weights v_i.

$$\left\{ \begin{array}{ll} \bar{w}_k = w^* + (I - 2\mu R)^k (\bar{w}_0 - \bar{w}^*) & \text{STEEPEST} \\ w_k = w^* (1 - 2\mu)^k (w_0 - w^*) & \text{NEWTON} \end{array} \right.$$

LEARNING CURVES FOR MULTI-D.

NEWTON

$$\xi = \xi_{MIN} + V^T R V = \xi_{MIN} + (1 - 2\mu)^k V_0^T R (1 - 2\mu)^k V_0$$

$$= \xi_{MIN} + (1 - 2\mu)^{2k} V_0^T R V_0$$

There is only ONE time constant (exp. decay).

$$r_{MSE} = r^2 = (1 - 2\mu)^2$$

STEEPEST DESCENT

$$\xi = \xi_{MIN} + V^T R V = \xi_{MIN} + \left[(I - 2\mu \Lambda)^k V_0' \right]^T \Lambda \left[(I - 2\mu \Lambda)^k V \right.$$

$$\xi_K = \xi_{MIN} + V_0'^T (I - 2\mu \Lambda)^{2k} V_0'$$

$$\xi_K = \xi_{MIN} + \sum_{n=0}^{L} v_{0n}'^2 \lambda_n (1 - 2\mu \lambda_n)^{2k}$$

For this case the learning curve is a SUM of geometric progressions of the form

$$r_n^2 = (1 - 2\mu \lambda_n)^2$$

Define the ADAPTIVE TIME CONSTANT as the time it takes for the error to decrease to 1/e (~37%) of its initial value.

$$1 \longrightarrow$$

$$\exp\left(-\frac{1}{\tau}\right) = r = 1 - \frac{1}{\tau} + \frac{1}{2!\tau^2} + \dots$$

$$\simeq 1 - \frac{1}{\tau} = 1 - 2\mu\lambda_i$$

$$\boxed{\tau_i \simeq \frac{1}{2\mu\lambda_i}}$$

So τ_i only depends on the eigenvalue and it is independent of P.

Each MODE will converge at different rates depending on μ and λ_i .

The SYSTEM BEHAVIOR WILL BE CONTROLLED BY THE SLOWEST MODE. Call it τ

$$\tau = \underset{i}{\mathrm{MAX}} \left\{ \frac{1}{2\mu\lambda_i} \right\} = \frac{1}{2\mu \underset{i}{\mathrm{MIN}} \{\lambda_i\}}$$

Therefore, the smallest eigenvalue CONTROLS the adaptation time (i.e. the algorithm can not converge faster than the slowest mode).

μ FOR MAXIMUM CONVERGENCE.

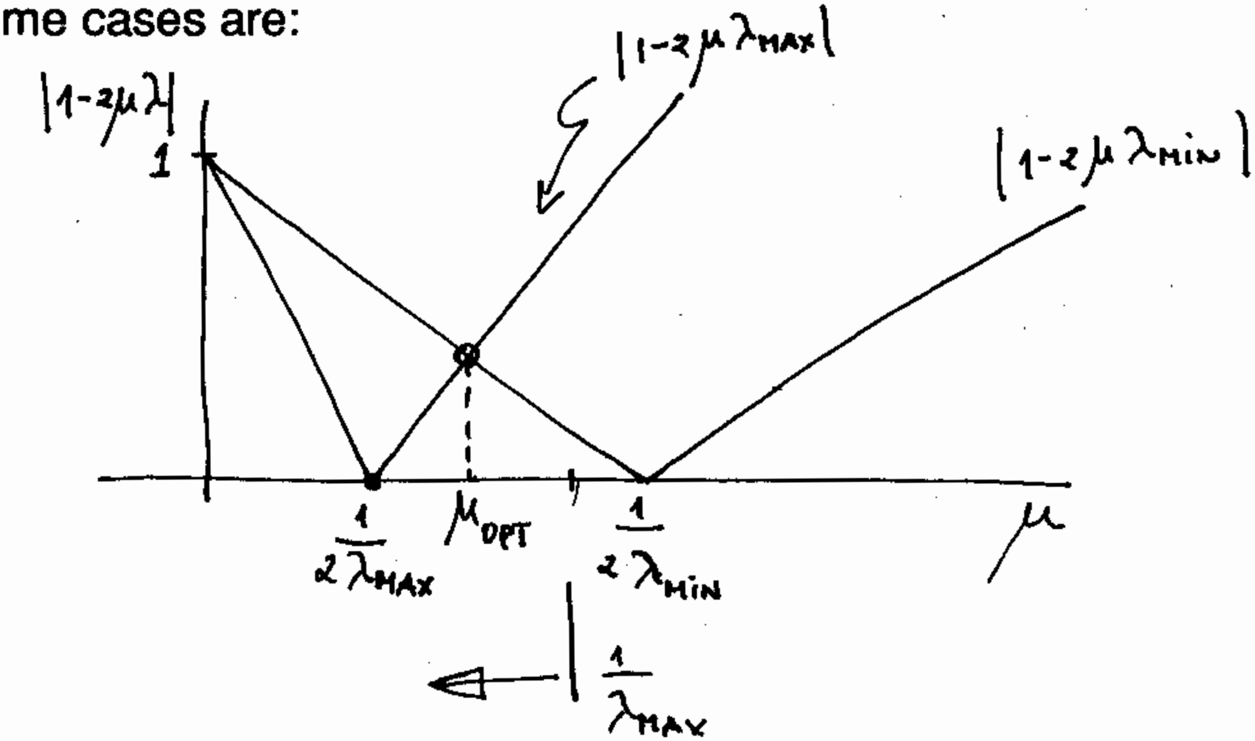
Let us order the eigenvalues

$$\lambda_{min}, \ldots \quad \lambda_{MAX}$$

The term that controls how large μ can get is λ_{max}.

$$0 < \mu < \frac{1}{\lambda_{MAX}}$$

But now if we want to find the μ that provides fastest convergence we should look at the smallest eigenvalue. The 2 extreme cases are:



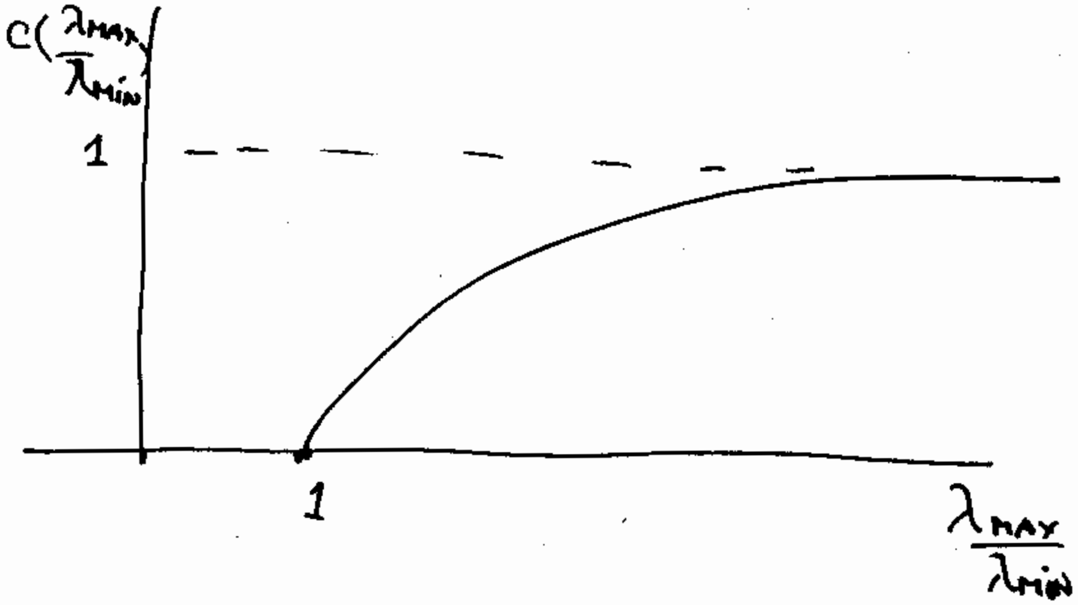
So the value of μ that produces the fastest convergence is

$$\mu_{OPT} = \frac{1}{\lambda_{MAX} + \lambda_{MIN}}$$

For this choice, both the slowest and fastest modes will converge at the same speed. So if the only consideration is speed of convergence, this is the value.

$$1 - 2\mu_{OPT}\lambda_{MIN} = \frac{\frac{\lambda_{MAX}}{\lambda_{MIN}} - 1}{\frac{\lambda_{MAX}}{\lambda_{MIN}} + 1} = R(\ ).$$

The ratio $\frac{\lambda_{MAX}}{\lambda_{MIN}}$ is of fundamental importance. It is called the EIGENVALUE SPREAD, and it varies from 1 to infinite.



The larger the eigenvalue spread, the slower the convergence

(in matrix theory this is called the condition number of the matrix).

Let us understand this result intuitively.

For the steepest descent

$$\xi = \xi_{MIN} + \underline{V}'^{T} \underline{\Lambda} \, V' \Rightarrow$$

$$\xi - \xi_{MIN} = V'^{T} \underline{\Lambda} \, V' = \sum_{i=0}^{L} \lambda_i \underline{v_i}'^{2}$$

i.e., the MSE is the sum of the product of the eigenvalues multiplied by the lenght of the v_i' components.

In the direction of the eigenvector v_i, we find the component of MSE just by multiplying λ_i by $v^2_{i.}$

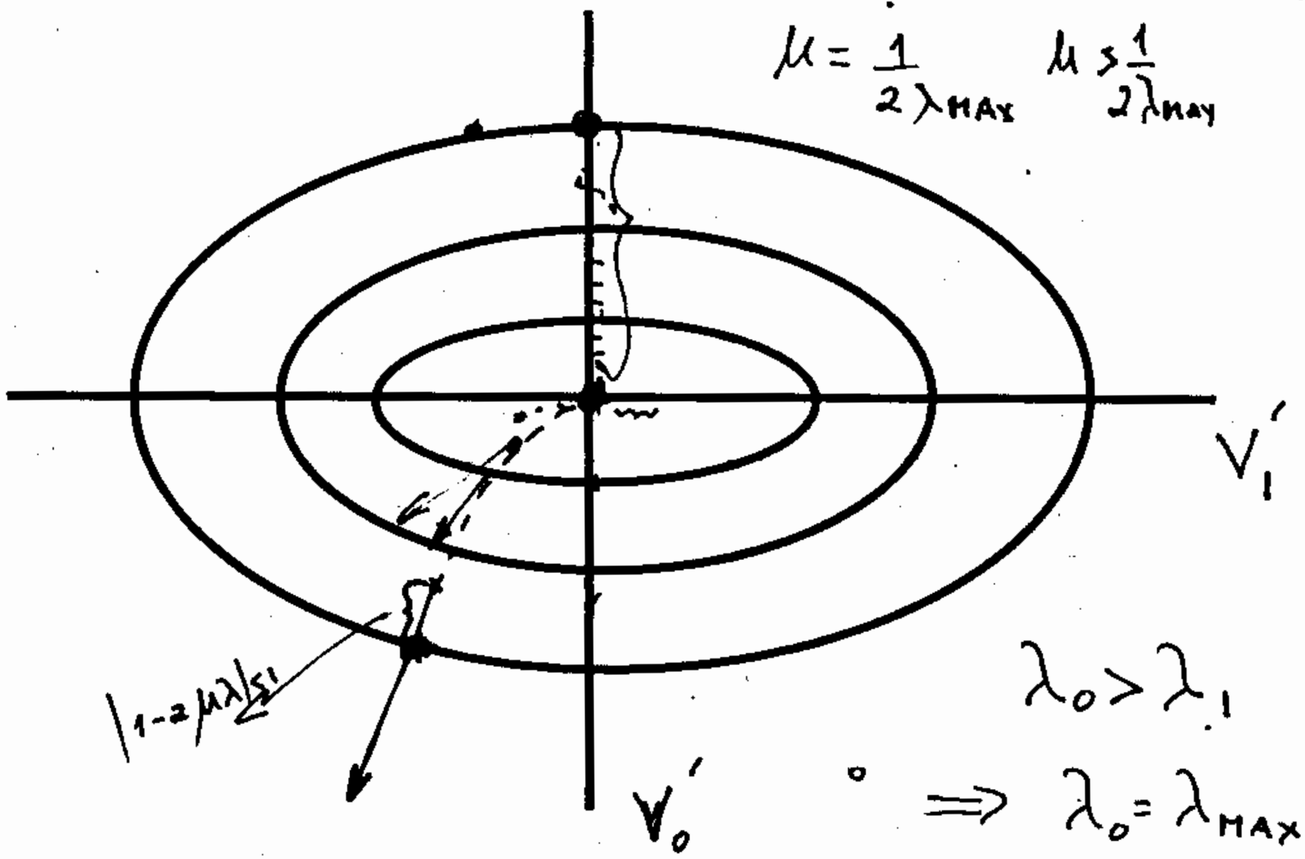
Therefore, the MSE increases most rapidly in the direction of the eigenvector corresponding to $\lambda_{max.}$ Moreover the largest m is determined by the largest eigenvalue.

For convergence one needs to take into consideration the largest change in the gradient.

$$v_i'(k+1) = (1-2\mu\lambda_i)\, v_i'(k)$$

$\mu < \frac{1}{2\lambda_{MAX}}$ $\mu = \frac{1}{\lambda_{MAX} + \lambda_{MIN}}$

$\mu = \frac{1}{2\lambda_{MAX}}$ $\mu \leq \frac{1}{2\lambda_{MAX}}$



$\lambda_0 > \lambda_1$

$\Longrightarrow \lambda_0 = \lambda_{MAX}$

$\mu < \frac{1}{\lambda_{MAX}}$

Consequences:

- When eccentricity is large, the gradient direction can be quite different from the direction of the minimum.
- The length of the gradient is very different in the v_o,v_1 plane.We must constraint the mu by the largest value, so we need much more steps when we travel along the axis of the smallest eigenvalue.
- For small eigenvalue spread, the gradient points almost towards the minimum, and none of this happens.

Can also understand now μ_{opt}. Consider seed value in the minor axis of ellipse.

If μ is $< 1/2\lambda_{max}$ no overshoot will occur. When $\mu = 1/2\lambda_{max}$ the convergence is in one step. When $\mu \underset{<}{=} 1/2\lambda_{max}$ but smaller than $1/\lambda_{max}$ the algorithm overshoots the minimum.

μ_{opt} is the value that makes the error overshoot in λ_{max} in order to make it converge faster in the λ_{min} direction.

What happens if R is singular?

This means that at least one of the λ_i is zero.Therefore,

$$V_i'(k+1) = (1-2\cancel{\lambda_i\mu})V_i'(k) = V_i'(k)$$

in other words, the associated filter coefficient is unchanged. The adaptive time constant will be infinite.However, this does not mean that the filter does not produce a useful solution.

If R has a zero eigenvalue $\lambda_i = 0$ and v_i is the associated eigenvector, then

$$\bar{R}\cdot\gamma\bar{v}_i = \gamma\bar{R}\cdot\bar{v}_i \equiv 0$$

(we say that v_i belongs to the NULL space of R).

We want to find w* such that RW* = P. Now if we substitute w* by $\bar{w}^* + \gamma\bar{v}_i$

$$R(\bar{w}^* + \gamma\bar{v}_i) = \bar{R}\bar{w}^* + \cancelto{\equiv 0}{\gamma R\cdot\bar{v}_i} = P$$

This means that we can add any vector from the null space of R without disturbing the Wiener solution.

There are 2 implications:

1. W* is not unique.
2. The convergence of the null space modes are IRRELEVANT, if the goal is to obtain A solution instead of THE solution.
3. We can practically be interested in $\lambda_{min} \neq 0$.

RELATION OF EIGENVALUE SPREAD WITH INPUT SIGNAL PARAMETERS.

First note that

$$\det R = \prod_{i=0}^{L-1} \lambda_i$$

so $$L\,r(0) = L\sigma_x^2 = \sum_{i=0}^{L-1} \lambda_i$$

i.e., if the determinant is non zero, the sum of the eigenvalues is equal to L (filter order) times the signal power.Therefore,

$$\lambda_{MAX} < L\sigma_x^2$$

A tighter bound is valid for square R with entries r_{ij}.

$$\lambda_{MAX} < \max_{j} \sum_{i=0}^{L} |r_{ij}|$$

OR

$$< \max_{i} \sum_{j=0}^{L} |r_{ij}|$$

Also it can be proven that

$$\min S(e^{jw}) \leq \lambda_i \leq \max S(e^{jw})$$

where $S(e^{jw})$ is the power spectral density of the input. We know that

$$S(e^{jw}) = \sum_{i=-\infty}^{\infty} R_i e^{-jwi}$$

so in the case of very large filter lengths, λ_{max}, and λ_{min} approach the maximum and the minimum in the spectrum

$$\lambda_{MAX} \longrightarrow \max_{w} S(e^{jw})$$

$$\lambda_{MIN} \longrightarrow \min_{w} S(e^{jw})$$

So signals that have large spectral dynamic range (ratio of largest to smallest values) will have large eigenvalue spread, so they will show slow convergence.