6.4 Convergence and Excess Mean Squared Error

- The LMS algorithm, unlike ~~steepest~~ descent, has noisy inputs, so trajectory is a random walk with drift.

Because of shallow gradient near optimum, the gradient noise causes $\underline{w}(n)$ to meander. The resulting misadjustment causes excess mse

$$J(n) = J_{min} + \underbrace{\underline{\epsilon}^{\dagger}(n) R \underline{\epsilon}(n)}_{J_{ex} = E[\underline{\epsilon}^{\dagger} R \underline{\epsilon}]}$$

- Haykin 9.4 provides a thorough analysis of convergence behaviour. We will take a shortcut to excess mse in class.
  - J G Proakis and J H Miller, "An Adaptive Receiver for Digital Signaling Through Channels with Intersymbol Interference" IEEE Trans Inf Th, vol IT-15, no 4, pp 484-497, July 1969.
    (adapted from Widrow's work in 1966)

- In order to obtain $J_{ex} = E[\underline{\epsilon}^+ R \underline{\epsilon}]$, we need the statistics of $\underline{\epsilon}$; i.e. its covariance matrix $E[\underline{\epsilon}\,\underline{\epsilon}^+]$. We obtain this from the update equation, a ΔE with noisy input.

Recall

$$\underline{w}(n+1) = \underline{w}(n) + \mu \underbrace{e^*(n)\underline{u}(n)}_{\text{represent as mean value plus zero mean noise}}$$

$$= (I - \mu R)\underline{w}(n) + \mu \underline{p} + \mu \underline{\gamma}(n) \qquad \underline{p} - R\underline{w}(n) + \gamma(n)$$

Shift to misadjustment vector $\underline{\epsilon}(n) = \underline{w}(n) - \underline{w}_o$

$$\underline{\epsilon}(n+1) = (I - \mu R)\underline{\epsilon}(n) + \mu \underline{\gamma}(n)$$

This is progress, but now we need the statistics of the noise $\underline{\gamma}(n)$; i.e., $E[\gamma(n)\underline{\gamma}^+(n-k)]$

- To obtain this noise covariance matrix, we'll make approximations valid at or in the neighbourhood of the convergence point
  - first, $e^*(n)\underline{u}(n) = \underbrace{(p - R\underline{w}(n))}_{=0 \text{ at convergence point}} + \underline{\gamma}(n)$

$$\approx \underline{\gamma}(n)$$

so we evaluate the covariances in $E[\underline{\gamma}(n)\underline{\gamma}^+(n)]$ at the convergence point and assume that they don't change much in that region

$$E[\underline{\gamma}(n)\underline{\gamma}^+(n)] = E[|e(n)|^2 \underline{u}(n)\underline{u}^+(n)]$$

  - next, assume $e(n)$ is independent of $\underline{u}(n)$ and $\underline{u}(n)\underline{u}^+(n)$; true if Gaussian by principle of orthogonality; approx true if not Gaussian, since at least uncorrelated; and $e(n)$ is approx Gaussian anyway by central limit theorem: $e(n) = d(n) - \underline{w}^+\underline{u}$

Using this independence, we obtain

$$E[\underline{\gamma}(n)\underline{\gamma}^t(n)] \approx E[|e(n)|^2]\, E[\underline{u}(n)\underline{u}(n)^t]$$

$$= J_{min} R$$

↑ near convergence point

- and finally, assume that $\underline{\gamma}(n)$ is a white sequence,

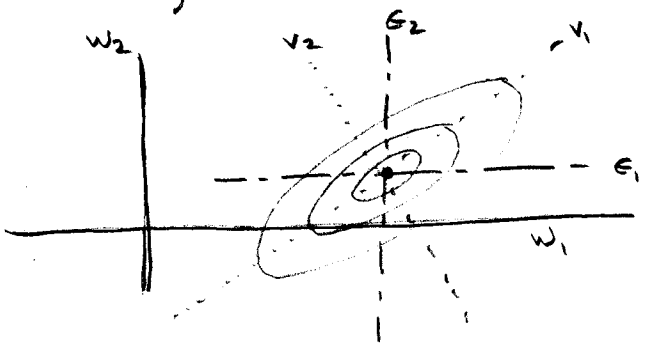
  because: - optimized estimation filters $\underline{w}$ exploit temporal correlations
  - if there were significant correlation between $\underline{\gamma}(n)$ and $\underline{\gamma}(n-k)$, the filter would adapt to use it, thereby removing it from $\underline{\gamma}(n)$

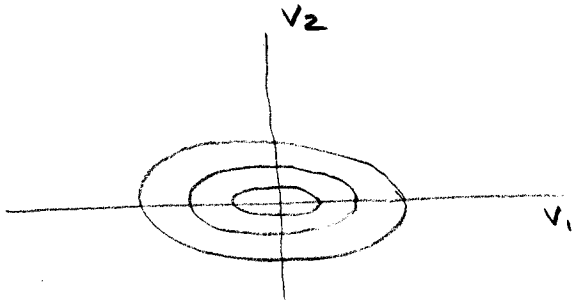
- We now have a model for the noise driving the ΔE

$$\underline{\epsilon}(n+1) = (I - \mu R)\,\underline{\epsilon}(n) + \mu\,\underline{\gamma}(n)$$

and we could obtain $R_\epsilon = E[\underline{\epsilon}\,\underline{\epsilon}^t]$ from it — but...

- it's easier and more instructive to transform to the eigenvector basis





$$\underline{\epsilon} = Q\underline{v} \qquad Q = [q_1|q_2|\cdots|q_M] \qquad \underline{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_M \end{bmatrix}$$

v_i is the coefficient of mode i

- In terms of coefficients in the eigenvector basis

$$Q\underline{v}(n+1) = (I-\mu R)Q\underline{v}(n) + \mu\underline{\gamma}(n)$$

$$\underline{v}(n+1) = (I-\mu\Lambda)\underline{v}(n) + \mu \underbrace{Q^T\underline{\gamma}(n)}_{\underline{\nu}(n)}$$

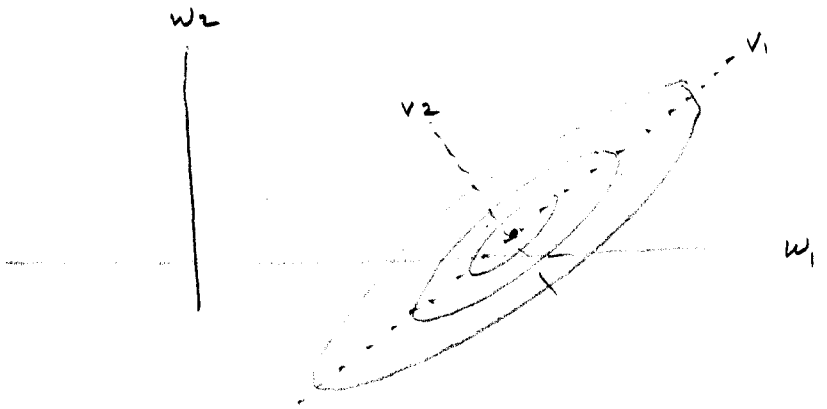
and the transformed noise covariance matrix is

$$E[\underline{\nu}\underline{\nu}^T] = Q^T E[\underline{\gamma}\underline{\gamma}^T]Q = J_{min}Q^TRQ = \underbrace{J_{min}\Lambda}_{\text{diagonal}}$$

so transformed noise components are uncorrelated, and have variances $\sigma_{\nu_i}^2 = J_{min}\lambda_i$

- Transformation has expressed misadjustment dynamics as a set of uncoupled modes:

$$\boxed{v_i(n+1) = (1-\mu\lambda_i)v_i(n) + \mu\nu_i(n)} \qquad i = 1..M$$



- Two observations:
  - $(1-\mu\lambda_i)$ is smallest for the dominant mode, the one with λ_{max}, so it decays most quickly, less opportunity for noise buildup
  - but $\sigma_{\nu_i}^2$ is proportional to λ_i, so greatest for λ_{max}, because steepest in that direction, more error variability.

- Which effect will win?

- Now we can obtain the coefficient variance in each mode from the transformed AE:

$$v_i(n+1) = (1-\mu\lambda_i)\, v_i(n) + \mu z_i(n)$$

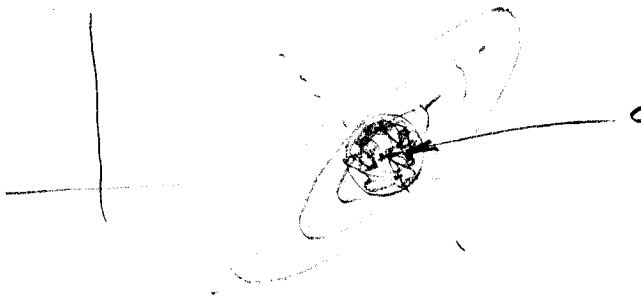
$$|v_i(n+1)|^2 = (1-\mu\lambda_i)^2 |v_i(n)| + \mu^2 |z_i(n)|^2 + 2(1-\mu\lambda_i)\, \mathrm{Re}\left[v_i(n) z_i^*(n)\right]$$

$$\sigma_{v_i}^2 = (1-\mu\lambda_i)^2 \sigma_{v_i}^2 + \mu^2 \sigma_{z_i}^2$$

$$\sigma_{v_i}^2 = \frac{\mu^2 \sigma_{z_i}^2}{1-(1-\mu\lambda_i)^2} = \frac{\mu^2 J_{min}\lambda_i}{1-(1-\mu\lambda_i)^2}$$

$$\approx \frac{\mu J_{min}}{2} \quad \text{for } \mu\lambda_i \ll 1$$

Amazing — same for all modes!



Therefore

$$E[\underline{v}\,\underline{v}^+] = \frac{\mu J_{min}}{2} I$$

and the quantity we were after all along...

$$E[\underline{\epsilon}\,\underline{\epsilon}^+] = Q\, E[\underline{v}\,\underline{v}^+] Q^+ = \frac{\mu J_{min}}{2} I$$

- Finally, we obtain the excess mse due to tap wander. We could do it like this:

$$J_{ex} = E[\underline{\epsilon}^+ R \underline{\epsilon}] = tr[R\, E[\underline{\epsilon}\underline{\epsilon}^+]]$$

$$= \mu \frac{J_{min}}{2} tr[R] = \mu \frac{J_{min}}{2} \sum_{i=1}^{M} \lambda_i = \frac{1}{2} \mu J_{min} M r_u(0)$$

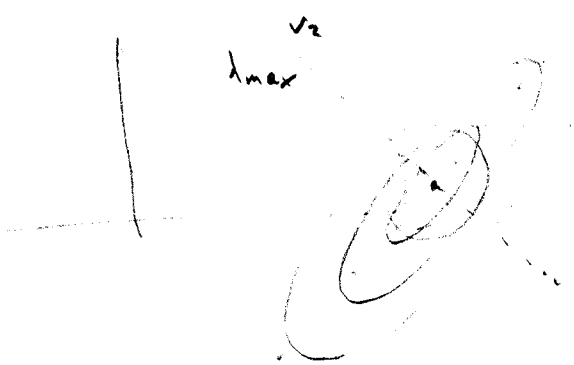
but it's more instructive to continue in the eigenvector basis

$$J_{ex} = E[\underline{\epsilon}^+ R \underline{\epsilon}] = E[\underline{v}^+ Q^+ R Q \underline{v}] = E[\underline{v}^+ \Lambda \underline{v}]$$

$$= \sum_{i=1}^{M} \lambda_i \sigma_{v_i}^2 \quad \text{so modes are weighted by eigenvals}$$

$$= \mu \frac{J_{min}}{2} \sum_{i=1}^{M} \lambda_i$$

The λ_{max} mode contributes most to J_{ex}. Why?



Because the circular blur climbs higher up the walls in the λ_{max} direction

In any case,

$$\boxed{J_{ex} = \frac{1}{2} \mu J_{min} M r_u(0)}$$

A surprisingly simple result! Next, we'll extract some meaning from it.

- Effect of step size parameter
  - μ too small? slow convergence $n_c \approx \frac{1}{\mu \lambda_{min}}$
  - μ too big? $J_{ex} = \frac{\mu}{2} J_{min} M r_u(0)$ too big
- An easy guideline: allow 10% excess mse

$$J_{ex} = 0.1 J_{min}$$

$$\mu = \frac{0.2}{M r_u(0)}$$ easy to set, since we can measure $r_u(0)$

It doesn't sacrifice too much speed:

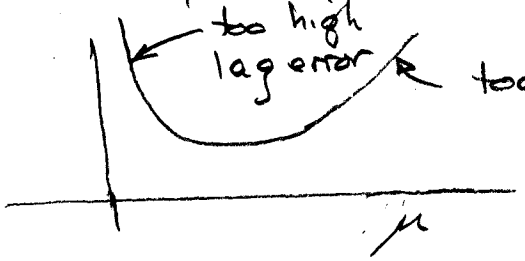
$$n_c \approx \frac{1}{\mu \lambda_{min}} = \frac{M r_u(0)}{0.2 \lambda_{min}} = \frac{\sum_i \lambda_i}{0.2 \lambda_{min}}$$

so $$\frac{\lambda_{max}}{0.2 \lambda_{min}} \leq n_c \leq \frac{M \lambda_{max}}{0.2 \lambda_{min}}$$

It still depends on the eigenvalue ratio.

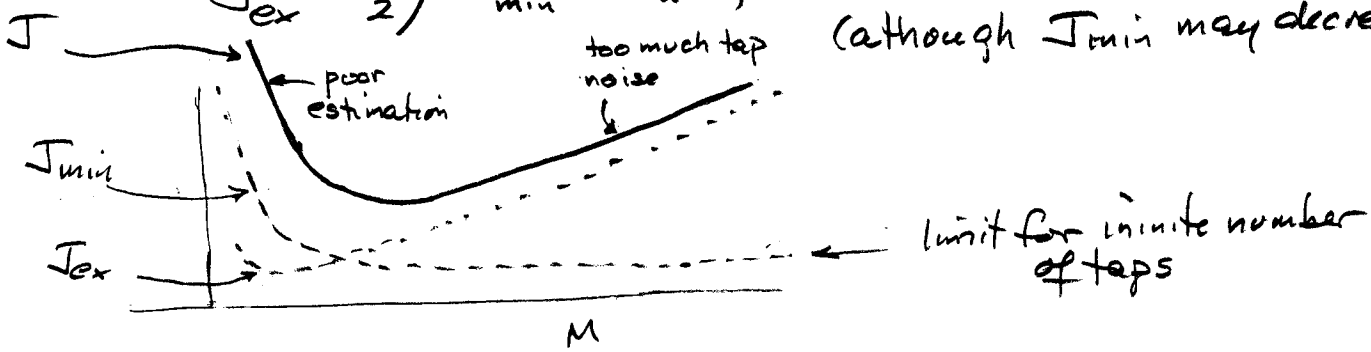
- But we don't have to keep μ constant, which leads to another strategy:
  gearshifting – start with large μ for initial convergence
  – then reduce it for small J_{ex}

- Tracking – if statistics drift, too small μ may cause filter to lag too much



- Effect of filter length
  - Normally, we expect to improve performance if we increase filter length M — with diminishing returns, of course
  - In adaptation, it's different:

$$J_{ex} = \frac{1}{2}\mu J_{min} M r_u(0)$$ increases with M! (although J_{min} may decrease)



  - The growth with M is caused by the fact that every tap has same jitter variance $\mu \frac{J_{min}}{2}$ (same as $\sigma_{v_i}^2$, circular) so additional taps add more misadjustment error

Read Haykin 9.6, 9.7

Browse Haykin 9.11, especially (9.121), (9.122), (9.142) (9.144)