5. Modellillesztés (folyt.)

Út az adaptív eljárásokhoz: (85) és (88) alapján: $W^* = R^{-1}P$, $\nabla(n) = 2(RW(n) - P)$. Ez utóbbi mindkét oldalát megszorozva az $\frac{1}{2}R^{-1}$ mátrixszal: $\boxed{W^* = W(n) - \frac{1}{2}R^{-1}\nabla(n)}$. (92)

Feltételezve, hogy nincs tökéletes ismeretünk az *R* mátrixról, és ebből adódóan a gradiensről, (92) átírható egy iteratív formára: $W(n+1) = W(n) - \frac{1}{2}\hat{R}^{-1}\hat{\nabla}(n)$, illetve a $0 < \mu < 1$ „bátorsági" tényező bevezetésével, visszaírva a „tökéletes" *R* mátrixot és gradienst

$$\boxed{W(n+1) = W(n) - \mu R^{-1}\nabla(n)}\ . \qquad (93)$$

Megjegyzések:

1. Ha pontosan ismerjük az *R* mátrixot és gradienst, akkor $\mu = \frac{1}{2}$ egylépéses konvergenciát biztosít tetszőleges W(*n*) kezdőpontból.
2. Mivel $\nabla(n) = 2R[W(n) - W^*]$, ezért ezt a (93) összefüggésbe behelyettesítve, és az egyenlet mindkét oldalából levonva W^* értékét:
   $W(n+1) - W^* = (1-2\mu)(W(n) - W^*) = V(n+1) = (1-2\mu)^{n+1}V(0)$, vagyis a kezdeti hiba exponenciális jelleggel csökken, ha $\mu \neq \frac{1}{2}$. Ha $0 < \mu < 0.5$, akkor monoton csökkenő hibával, ellenkező esetben pedig monoton csökkenő amplitúdójú, de lengő jellegű hibával közelítjük meg.
3. A modell-illesztés gradiens módszereit a szerint különböztetjük meg, hogy a (93) szerinti összefüggés alkalmazásához milyen előzetes ismeretek állnak rendelkezésünkre.

Az adaptív lineáris kombinátor működését leíró egyenletek, amennyiben az R és a P mátrixok ismertek:

$$\boxed{W(n+1) = W(n) - \mu R^{-1}\nabla(n)}, \text{ ill. } \boxed{V(n+1) = (1-2\mu)V(n)}\ . \qquad (94)$$

Megjegyzések:

1. A továbbiakban sorra kerülő vizsgálatok azt tárják fel, hogy milyen lehetőségeink vannak akkor, ha az R és P mátrixokra vonatkozó előzetes (a priori) ismereteink részlegesek, esetleg teljes mértékben hiányoznak, legfeljebb a folyamatban lévő mérésekre alapozhatók. Ez a gondolat végig jelen van a továbbiakban, a megértéshez fontos, hogy ezt ne hagyjuk figyelmen kívül.
2. Figyeljük meg, hogy az *R* mátrix „globális" információt hordoz a hibafelületről, a $\nabla(n)$ gradiens pedig az adott $W(n)$ paraméterérték esetén a hibafelület „lokális" jellemzése. Ezen lokális ismeret alapján „ereszkedünk" a hibafelületen az ún. gradiens eljárások alkalmazása esetén annak érdekében, hogy minél közelebb kerüljünk az optimumot (legkisebb négyzetes hibát) eredményező paraméter beállításhoz.

Az R mátrix vizsgálata: A hibafelület az R mátrixtól függ. Elöljáróban azt mutatjuk be, hogy milyen feltételek esetén lehetséges az optimum-keresést úgy megvalósítani, ahogyan egy impedancia-mérő híd esetében is szeretnénk: egyenként változtatjuk a változtatható paramétereket, mégpedig úgy, hogy mindig megkeressük a lokális minimumot, és eközben a hiba egyetlen lépés során sem nő. Ehhez egy olyan koordinátarendszerben történő keresés tartozik, amelynek tengelyei a paraboloid formájú hibafelület főtengelyeinek irányába mutatnak. Ezt a koordinátarendszert az *R* mátrix sajátvektorai jelölik ki.

$$\boxed{\varepsilon(n)=\varepsilon_{\min}+(W(n)-W^*)^T R(W(n)-W^*)=\varepsilon_{\min}+V^T(n)RV(n)} \qquad (95)$$

Fontos szerepet játszik tehát az R sajátérték/sajátvektor rendszere. Példaként a (90) szerinti esetben vizsgálódva:

$R=\begin{bmatrix} 0.5 & 0.5\cos\frac{2\pi}{N} \\ 0.5\cos\frac{2\pi}{N} & 0.5 \end{bmatrix}$. A $\det[\lambda I-R]=0$ egyenlet gyökei adják a sajátértékeket:

$$(\lambda-0.5)^2-0.25\cos^2\frac{2\pi}{N}=\lambda^2-\lambda+0.25\sin^2\frac{2\pi}{N}=0 \qquad (96)$$

A két gyök:

$$\boxed{\lambda_0=0.5+0.5\cos\frac{2\pi}{N}}, \qquad \text{ill.} \qquad \boxed{\lambda_1=0.5-0.5\cos\frac{2\pi}{N}} \qquad (97)$$

A sajátvektorok az $RQ_0=\lambda_0 Q_0$, $RQ_1=\lambda_1 Q_1$ egyenletekből származtathatók.

$$\begin{bmatrix} 0.5 & 0.5\cos\frac{2\pi}{N} \\ 0.5\cos\frac{2\pi}{N} & 0.5 \end{bmatrix}\begin{bmatrix} q_{00} \\ q_{01} \end{bmatrix}=(0.5+\cos\frac{2\pi}{N})\begin{bmatrix} q_{00} \\ q_{01} \end{bmatrix} \quad \Rightarrow \quad \boxed{q_{00}=q_{01}} \qquad (98)$$

$$\begin{bmatrix} 0.5 & 0.5\cos\frac{2\pi}{N} \\ 0.5\cos\frac{2\pi}{N} & 0.5 \end{bmatrix}\begin{bmatrix} q_{10} \\ q_{11} \end{bmatrix}=(0.5-\cos\frac{2\pi}{N})\begin{bmatrix} q_{10} \\ q_{11} \end{bmatrix} \quad \Rightarrow \quad \boxed{q_{10}=-q_{11}} \qquad (99)$$

A sajátvektorokat egységre normálva: $Q_0=\begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$, $Q_1=\begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$, lásd 25. ábra. (100)

A példa szerinti sajátvektorok tehát egymásra merőleges, a koordináta rendszer tengelyeivel 45 fokos szöget bezáró vektorok. Ezek adják meg azokat az ereszkedési irányokat, amelyek mentén történő mozgás egyetlen paraméter változtatásával lehetséges.

Általában $\det(R-\lambda I)=0 \quad \Rightarrow \quad \lambda_0,\lambda_1,...,\lambda_{N-1}$. $(R-\lambda_n I)Q_n=0$, $n=0,1,...,N-1$. A sajátvektorokat mátrixba rendezve:

$$R\left[\underbrace{Q_0 \quad Q_1 \quad \cdots \quad Q_{N-1}}_{Q}\right]=\left[\underbrace{Q_0 \quad Q_1 \quad \cdots \quad Q_{N-1}}_{Q}\right]\underbrace{diag\langle\lambda_0 \quad \lambda_1 \quad \cdots \quad \lambda_{N-1}\rangle}_{\Lambda} \qquad (101)$$

$RQ=Q\Lambda$, ill.

$$\boxed{R=Q\Lambda Q^{-1}}, \qquad (102)$$

ami *R* ún. normál formája. Mivel az *R* definíció szerint szimmetrikus mátrix, ezért $R=R^T$. Fontos tulajdonság, hogy ilyenkor a sajátvektorok ortogonálisak: $Q_i^T Q_j=0$, ha $\forall i\neq j$, egyébként $Q_i^T Q_i=c_i \quad \forall i$. Ha $Q_i^T Q_i=1 \quad \forall i$-re, akkor a sajátvektorok ortonormáltak, és $Q^T Q=I$, azaz $Q^{-1}=Q^T$. Az ortogonalitás bizonyítása: A definíció alapján $Q_i^T R^T=\lambda_i Q_i^T$, ill. $RQ_j=\lambda_j Q_j$. Az első egyenlet mindkét oldalát jobbról szorozva Q_j-vel, a második egyenlet

mindkét oldalát balról szorozva Q_i^T -vel: $Q_i^T R^T Q_j = \lambda_i Q_i^T Q_j$, ill. $Q_i^T R Q_j = \lambda_j Q_i^T Q_j$. Mivel $R = R^T$, ezért az egyenletek baloldala egyenlő, ezáltal $\lambda_i Q_i^T Q_j = \lambda_j Q_i^T Q_j$. Mivel $\lambda_i \neq \lambda_j$, ezért az egyenlőség csak akkor állhat fenn, ha $Q_i^T Q_j = 0$.

Megjegyzések:

1. Mivel $V^T RV$ pozitív definit, ezért a sajátértékek nem negatívak.
2. Az *R* korrelációs mátrix sajátvektorai a hibafelület főtengelyeit jelölik ki.

$$\varepsilon(n) = \varepsilon_{\min} + (W(n) - W^*)^T R(W(n) - W^*) = \varepsilon_{\min} + V^T(n) R V(n) = \varepsilon_{\min} + \underbrace{V^T(n) Q}_{V^T(n)} \Lambda \underbrace{Q^T V(n)}_{V^T(n)} =$$

$$= \varepsilon_{\min} + \left[Q^T V(n)\right]^T \Lambda \left[Q^T V(n)\right] = V'^T(n) \Lambda V'(n). \qquad (103)$$

$$\text{Ezzel } \nabla(n) = 2\Lambda V'(n) = 2\begin{bmatrix} \lambda_0 v_0' & \lambda_1 v_1' & \cdots & \lambda_{N-1} v_{N-1}' \end{bmatrix}^T. \qquad (104)$$

A viszonyokat a 26. ábra illusztrálja: a sajátvektorok mátrixával transzformált paraméter-hiba vektorok koordinátái a paraboloid főtengelyei mentén értelmezhetők. Az optimalizálás egyváltozós optimalizálások sorozataként is végrehajtható. Ezt mutatja be a következő példa, amelyben az optimum megközelítését a gradiens mentén történő ereszkedéssel oldjuk meg:

Példa:

Egyváltozós eset: $w(n+1) = w(n) + \mu(-\nabla(n))$, $\nabla(n) = 2\lambda(w(n) - w^*)$, amivel $w(n+1) - w^* = (1 - 2\mu\lambda)(w(n) - w^*)$, ill. $V(n+1) = rV(n) = r^{n+1} V(0)$. Ahhoz, hogy az eljárás konvergáljon $|r| = |1 - 2\mu\lambda| < 1$ szükséges. Ebből:

$$\boxed{0 < \mu < \frac{1}{\lambda}} \qquad (105)$$

Ha $0 < \mu < \frac{1}{2\lambda}$, akkor túlcsillapított, ha $\mu = \frac{1}{2\lambda}$, akkor kritikusan csillapított, ha $\frac{1}{2\lambda} < \mu < \frac{1}{\lambda}$, akkor alulcsillapított az iterációs eljárás.

Megjegyzés: Vegyük észre, hogy az egyváltozós esetben $R = \lambda$, tehát a (94) összefüggés első eleme $W(n+1) = W(n) - 2\mu(W(n) - W^*)$ alakú, amelynek mindkét oldalából levonva W^*-ot a (94) összefüggés második elemét kapjuk.

Többváltozós eset: $V'(n+1) = (I - 2\mu\Lambda)^{n+1} V'(n)$. Ahhoz, hogy az eljárás konvergáljon

$$\boxed{0 < \mu < \frac{1}{\lambda_{\max}}} \qquad (106)$$

szükséges. Vegyük észre, hogy ilyenkor *N* egyváltozós esettel van dolgunk, ahol a legmeredekebb ereszkedés azon tengely mentén történik, amelyikhez a legnagyobb sajátérték tartozik. Ha a sajátértékek nem ismertek, akkor a $\lambda_{\max} < \sum_{i=1}^{N-1} \lambda_i = tr[\Lambda] = tr[R]$ alapján

$$\boxed{0 < \mu < \frac{1}{tr[R]}} \qquad (107)$$

választással élhetünk.

Megjegyzés: Ha ismernénk Λ-t, azaz lenne „globális” információnk a hibafelületről akkor nyilvánvalóan a skalár μ helyett $\mu = \frac{1}{2}\Lambda^{-1}$ mátrixot alkalmaznánk, hiszen ezzel egylépéses

konvergenciát tudnánk biztosítani. Ehelyett a „lokális” információra, a gradiensre alapozva, annak irányában igyekszünk a hibafelület minimumát elérni.

Iteratív modellillesztési módszerek: Az alábbiakban néhány klasszikus szélsőérték keresési eljárást foglalunk össze, amelyeket négyzetes kritériumok, és paramétereiben lineáris modellek esetén négyzetes hibafelületek esetén előszeretettel alkalmazunk. Ezek tekinthető tanuló eljárásoknak is, mert minden lépésben „informálódnak” az aktuális viszonyokról, esetünkben a hibafelület gradienséről, és annak függvényében lépnek tovább. Természetesen alkalmazhatunk másfajta módszereket is, ahol például a $W(n)$ értékeket véletlen módon vagy más stratégiával választjuk ki, és ezt követően vizsgáljuk a hibát. Ha a korábbinál kisebb hibát kapunk, akkor a kiválasztott érték lesz az új javaslat, ellenkező esetben elvetjük azt (Monte-Carlo módszerek, genetikus algoritmusok). Az ilyen módszerek azonban inkább akkor merülnek fel, ha (1) nem négyzetes hibakritériumot használunk, ill. ha (2) a modellünk paramétereiben nem lineáris. Ezekben a helyzetekben ugyanis a hibafelület nem paraboloid, lokális minimumai lehetnek, amelyek esetén a „lokális” információra építő gradiens eljárások könnyen leállhatnak a lokális minimumok valamelyikében.

Iteratív modellillesztés Newton módszerrel:
Erre a Wiener-Hopf egyenletből kiindulva jutunk, a korábbiakban megismertük, itt csak a felsorolás teljessége érdekében szerepel. Feltételezzük, hogy ismerjük az R és a P mátrixot. Ebből adódóan a módszer inkább csak elvi jelentőségű, mert a gyakorlatban nem elvárható előzetes ismereteket tételez fel. Mégis ki kell emelni, mert irányt mutat a közelítő eljárások megtervezéséhez. Rendre két összefüggést adunk meg. Az első a paraméter vektort adja meg következő iterációs lépésben, míg a második a paraméter-hiba alakulását a kiindulási paraméter-hibából.

$$W(n+1) = W(n) - \mu R^{-1}\nabla(n), \tag{108}$$

$$V(n+1) = (1-2\mu)^{n+1}V(0). \tag{109}$$

Jól látható, hogy $\mu = 0.5$ esetén egylépéses a konvergencia.

Iteratív modellillesztés a legmeredekebb lejtő módszerével:
Ez már egy praktikus módszer, amelyik nem feltételezi az R és a P mátrixok ismeretét, de azt igen, hogy a gradienst „lokális” információk alapján meg tudjuk határozni:

$$\nabla(n) = \frac{\partial\varepsilon(n)}{\partial W(n)} \approx \frac{\Delta\varepsilon(n)}{\Delta W(n)} = \hat{\nabla}(n) \tag{110}$$

Ez praktikusan azt igényli, hogy az n-edik iterációs lépéshez elvégzünk egy olyan mérési sorozatot, hogy $W(n)$ kis megváltozása különböző bemenőjel-értékek mellett $\varepsilon(n)$ mekkora megváltozását eredményezi, majd ezeket a megváltozásokat átlagoljuk (amivel közelítjük a várható-érték képzést), és ezzel $\nabla(n)$ egy (reményeink szerint igen jó) becslését kapjuk.

$$W(n+1) = W(n) - \mu\nabla(n) \tag{111}$$

$$V'(n+1) = (I-2\mu\Lambda)^{n+1}V'(0) \tag{112}$$

Megjegyzések:
1. A gradiens mentén történő ereszkedés eredményét a főtengely irányú koordináta-rendszereben „látványosabban” tudjuk érzékeltetni.
2. A gradiens mentén történő ereszkedés eredménye természetesen nem függ attól, hogy milyen vonatkoztatási (koordináta) rendszert alkalmazunk.

Iteratív modellillesztés a pillanatnyi deriváltra alapozva (az ún. LMS módszer):
(LMS: Least-Mean-Square). A hiba pillanatértékéből indulunk ki:

$\varepsilon(n)=[y(n)-X^T(n)W(n)]^T[y(n)-X^T(n)W(n)]=e^T(n)e(n)$. Ennek deriválásával becsüljük a gradienst:

$$\hat{\nabla}(n)=\frac{\partial\varepsilon(n)}{\partial W(n)}=-2X(n)y(n)+2X(n)X^T(n)W(n)=-2X(n)e(n) \tag{113}$$

Amivel

$$\boxed{W(n+1)=W(n)+2\mu X(n)e(n)}. \tag{114}$$

Ez egy nagyon széles körben használt összefüggés, különösen nagyobb méretű paraméter-vektorok esetén. A μ bátorsági tényező azonban nagy körültekintéssel, és tipikusan kis értékre választandó, hiszen a (113) szerinti gradiens igencsak közelítő: az aktuális $y(n), X(n)$ függvénye, miközben a tényleges gradiens (113) várható értéke. A kis bátorsági tényezővel együtt jár a sok („apró") iterációs lépés, ami lehetőséget ad sok $\{y(n), X(n)\}$ érték „megismerésére", és ezzel az elmaradt várható érték képzés „kiváltására".

Megjegyzések:

1. A neurális hálózatok térhódításának kezdetén az LMS eljárást nagyon széles körben használták a méretes adaptív lineáris kombinátorokat használó hálózatok tanítására.
2. Általános tapasztalat, hogy ha eléggé kis μ értékkel dolgozunk, akkor elég jól megközelíthetjük az optimális paraméter-vektort, nagyobb μ esetén a megmaradó paraméter-hiba nagyobb lesz. Ennek az az oka, hogy ilyenkor a paraboloid legalsó pontja környezetében ide-oda ugrálunk a pillanatnyi derivált szerint, és a nem eléggé kis μ miatt képtelenek vagyunk még lejjebb ereszkedni. Mindenképpen célszerű tehát a minimum környezetében a μ érték további csökkentése.

A paraméter-hiba kifejezését a (114) összefüggésből úgy származtatjuk, hogy mindkét oldalából levonjuk W^*-ot, ill. $y(n)=X^T(n)W^*$ feltételezéssel/közelítéssel élünk. Ez utóbbival azt feltételezzük, hogy a modellillesztés „tökéletesen sikerült.

$$W(n+1)-W^*=W(n)-W^*+2\mu X(n)[X^T(n)W^*-X^T(n)W(n)]=$$
$$=[I-2\mu X(n)X^T(n)][W(n)-W^*]$$

amiből:

$$\boxed{V(n+1)=[\prod_{i=0}^{n}(I-2\mu X(i)X^T(i))]V(0)} \tag{115}$$

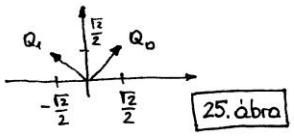
A (115) összefüggés arra mutat rá, hogy a paraméter-hiba csökkenéséhez hogyan járul hozzá a μ bátorsági tényező és az $X(n)$ ún. regressziós vektor. Nyilvánvalóan a mátrix szorzatnak kontraktívnak, azaz a paraméter-hiba vektor hosszát csökkentő hatásúnak kell lennie. Célszerű, ha ez a hatás minden lépésben érvényesül.

Iteratív modell-illesztés α -LMS módszerrel:

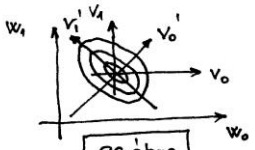
A (114) összefüggésben célszerű lehet az $X(n)$ regressziós vektor normálása, hiszen e nélkül a paraméter vektor korrekciója nagymértékben függ a „jelszinttől". (114), ill. (115) megfelelője:

$$\boxed{W(n+1)=W(n)+\frac{\alpha}{X^T(n)X(n)}X(n)e(n)} \tag{116}$$

$$\boxed{V(n+1)=[\prod_{i=0}^{n}(I-\frac{\alpha}{X^T(i)X(i)}X(i)X^T(i))]V(0)} \tag{117}$$



25. ábra



26. ábra