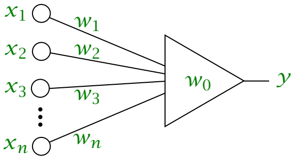
The LMS Algorithm

- The LMS Objective Function.
- Global solution.
- Pseudoinverse of a matrix.
- Optimization (learning) by gradient descent.
- LMS or Widrow-Hoff Algorithms.
- Convergence of the Batch LMS Rule.

Compiled by LaTeX at 3:50 PM on February 27, 2013.

LMS Objective Function

Again we consider the problem of programming a linear threshold function



$$y = \operatorname{sgn}(\mathbf{w}^T\mathbf{x} + w_0) = \begin{cases} 1, & \text{if } \mathbf{w}^T\mathbf{x} + w_0 > 0, \\ -1, & \text{if } \mathbf{w}^T\mathbf{x} + w_0 \leq 0. \end{cases}$$

so that it agrees with a given dichotomy of m feature vectors,

$$\mathcal{X}_m = \{(\mathbf{x}_1, \ell_1), \ldots, (\mathbf{x}_m, \ell_m)\}.$$

where $\mathbf{x}_i \in \mathbb{R}^n$ and $\ell_i \in \{-1, +1\}$, for $i = 1, 2, \ldots, m$.

LMS Objective Function (cont.)

Thus, given a dichotomy

$$\mathcal{X}_m = \{(\mathbf{x}_1, \ell_1), \ldots, (\mathbf{x}_m, \ell_m)\},$$

we seek a solution weight vector $\mathbf{w}$ and bias w_0, such that,

$$\operatorname{sgn}\left(\mathbf{w}^T \mathbf{x}_i + w_0\right) = \ell_i,$$

for $i = 1, 2, \ldots, m$.
Equivalently, using *homogeneous coordinates* (or augmented feature vectors),

$$\operatorname{sgn}\left(\widehat{\mathbf{w}}^T \widehat{\mathbf{x}}_i\right) = \ell_i,$$

for $i = 1, 2, \ldots, m$, where $\widehat{\mathbf{x}}_i = \left(1, \mathbf{x}_i^T\right)^T$.

Using *normalized coordinates*,

$$\operatorname{sgn}\left(\widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i\right) = 1, \quad \text{or,} \quad \widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i > 0,$$

for $i = 1, 2, \ldots, m$, where $\widehat{\mathbf{x}}'_i = \ell_i \widehat{\mathbf{x}}_i$.

LMS Objective Function (cont.)

Alternatively, let $\mathbf{b} \in \mathbb{R}^m$ satisfy $b_i > 0$ for $i = 1, 2, \ldots, m$. We call $\mathbf{b}$ a *margin vector*. Frequently we will assume that $b_i = 1$.

Our learning criterion is certainly satisfied if

$$\widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i = b_i$$

for $i = 1, 2, \ldots, m$. (Note that this condition is sufficient, but not necessary.) Equivalently, the above is satisfied if the expression

$$\mathcal{E}(\widehat{\mathbf{w}}) = \sum_{i=1}^{m} \left(b_i - \widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i \right)^2 .$$

equals zero. The above expression we call the *LMS objective function*, where LMS stands for *least mean square*. (N.B. we could normalize the above by dividing the right side by m.)

LMS Objective Function: Remarks

Converting the original problem of classification (satisfying a system of inequalities) into one of optimization is somewhat *ad hoc*. And there is no guarantee that we can find a $\widehat{\mathbf{w}}^\star \in \mathbb{R}^{n+1}$ that satisfies $\mathcal{E}(\widehat{\mathbf{w}}^\star) = 0$. However, this conversion can lead to practical compromises if the original inequalities possess inconsistencies.

Also, there is no unique objective function. The LMS expression,

$$\mathcal{E}(\widehat{\mathbf{w}}) = \sum_{i=1}^{m} \left(b_i - \widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i\right)^2.$$

can be replaced by numerous candidates, e.g.

$$\sum_{i=1}^{m} \left|b_i - \widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i\right|, \quad \sum_{i=1}^{m} \left(1 - \operatorname{sgn}\left(\widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i\right)\right), \quad \sum_{i=1}^{m} \left(1 - \operatorname{sgn}\left(\widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i\right)\right)^2, \quad \text{etc.}$$

However, minimizing $\mathcal{E}(\widehat{\mathbf{w}}^\star)$ is generally an easy task.

Minimizing the LMS Objective Function

Inspection suggests that the LMS Objective Function

$$\mathcal{E}\left(\widehat{\mathbf{w}}\right) = \sum_{i=1}^{m} \left(b_i - \widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i\right)^2$$

describes a parabolic function. It may have a unique global minimum, or an infinite number of global minima which occupy a connected linear set. (The latter can occur if $m < n + 1$.) Letting,

$$X = \begin{pmatrix} \widehat{\mathbf{x}}'^T_1 \\ \widehat{\mathbf{x}}'^T_2 \\ \vdots \\ \widehat{\mathbf{x}}'^T_m \end{pmatrix} = \begin{pmatrix} \ell_1 & \hat{x}'_{1,1} & \hat{x}'_{1,2} & \cdots & \hat{x}'_{1,n} \\ \ell_2 & \hat{x}'_{2,1} & \hat{x}'_{2,2} & \cdots & \hat{x}'_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_m & \hat{x}'_{m,1} & \hat{x}'_{m,2} & \cdots & \hat{x}'_{m,n} \end{pmatrix} \in \mathbb{R}^{m\times(n+1)},$$

then,

$$\mathcal{E}\left(\widehat{\mathbf{w}}\right) = \sum_{i=1}^{m} \left(b_i - \widehat{\mathbf{x}}'^T_i \widehat{\mathbf{w}}\right)^2 = \left\| \begin{pmatrix} b_1 - \widehat{\mathbf{x}}'^T_1 \widehat{\mathbf{w}} \\ b_2 - \widehat{\mathbf{x}}'^T_2 \widehat{\mathbf{w}} \\ \vdots \\ b_m - \widehat{\mathbf{x}}'^T_m \widehat{\mathbf{w}} \end{pmatrix} \right\|^2 = \left\| \mathbf{b} - \begin{pmatrix} \widehat{\mathbf{x}}'^T_1 \\ \widehat{\mathbf{x}}'^T_2 \\ \vdots \\ \widehat{\mathbf{x}}'^T_m \end{pmatrix} \widehat{\mathbf{w}} \right\|^2 = \|\mathbf{b} - X\widehat{\mathbf{w}}|^2.$$

Minimizing the LMS Objective Function (cont.)

$$
\begin{aligned}
\mathcal{E}(\widehat{\mathbf{w}}) &= \|\mathbf{b} - X\widehat{\mathbf{w}}\|^2 \\
&= (\mathbf{b} - X\widehat{\mathbf{w}})^T(\mathbf{b} - X\widehat{\mathbf{w}}) \\
&= \left(\mathbf{b}^T - \widehat{\mathbf{w}}^T X^T\right)(\mathbf{b} - X\widehat{\mathbf{w}}) \\
&= \widehat{\mathbf{w}}^T X^T X \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^T X^T \mathbf{b} - \mathbf{b}^T X \widehat{\mathbf{w}} + \mathbf{b}^T \mathbf{b} \\
&= \widehat{\mathbf{w}}^T X^T X \widehat{\mathbf{w}} - 2\mathbf{b}^T X \widehat{\mathbf{w}} + \|\mathbf{b}\|^2
\end{aligned}
$$

As an aside, note that

$$
X^T X = (\widehat{\mathbf{x}}'_1, \ldots, \widehat{\mathbf{x}}'_m) \begin{pmatrix} \widehat{\mathbf{x}}_1'^T \\ \vdots \\ \widehat{\mathbf{x}}_m'^T \end{pmatrix} = \sum_{i=1}^{m} \widehat{\mathbf{x}}'_i \widehat{\mathbf{x}}_i'^T \in \mathbb{R}^{(n+1)\times(n+1)},
$$

$$
\mathbf{b}^T X = (b_1, \ldots, b_m) \begin{pmatrix} \widehat{\mathbf{x}}_1'^T \\ \vdots \\ \widehat{\mathbf{x}}_m'^T \end{pmatrix} = \sum_{i=1}^{m} b_i \widehat{\mathbf{x}}_i'^T \in \mathbb{R}^{n+1}.
$$

Minimizing the LMS Objective Function (cont.)

Okay, so how do we minimize

$$\mathcal{E}(\widehat{\mathbf{w}}) = \widehat{\mathbf{w}}^T X^T X \widehat{\mathbf{w}} - 2\mathbf{b}^T X \widehat{\mathbf{w}} + \|\mathbf{b}\|^2?$$

Using calculus (e.g., Math 121), we can compute the gradient of $\mathcal{E}(\widehat{\mathbf{w}})$, and algebraically determine a value of $\widehat{\mathbf{w}}^\star$ which makes each component vanish. That is, solve

$$\nabla\mathcal{E}(\widehat{\mathbf{w}}) = \begin{pmatrix} \frac{\partial \mathcal{E}}{\partial \widehat{w}_0} \\ \frac{\partial \mathcal{E}}{\partial \widehat{w}_1} \\ \vdots \\ \frac{\partial \mathcal{E}}{\partial \widehat{w}_n} \end{pmatrix} = 0.$$

It is straightforward to show that

$$\nabla\mathcal{E}(\widehat{\mathbf{w}}) = 2X^T X\widehat{\mathbf{w}} - 2X^T\mathbf{b}.$$

Minimizing the LMS Objective Function (cont.)

Thus,

$$\nabla \mathcal{E}(\widehat{\mathbf{w}}) = 2X^T X \widehat{\mathbf{w}} - 2X^T \mathbf{b} = 0$$

if

$$\widehat{\mathbf{w}}^\star = (X^T X)^{-1} X^T \mathbf{b} = X^\dagger \mathbf{b},$$

where the matrix,

$$X^\dagger \stackrel{\text{def}}{=} (X^T X)^{-1} X^T \in \mathbb{R}^{(n+1)\times m}$$

is called the *pseudoinverse* of X. If $X^T X$ is singular, one defines

$$X^\dagger \stackrel{\text{def}}{=} \lim_{\epsilon \to 0} (X^T X + \epsilon I)^{-1} X^T.$$

Observe that if $X^T X$ is nonsingular,

$$X^\dagger X = (X^T X)^{-1} X^T X = I.$$

Example

The following example appears in R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Second Edition, Wiley, NY, 2001, p. 241.
Given the dichotomy,

$$\mathcal{X}_4 = \left\{ \left((1,2)^T, 1\right), \left((2,0)^T, 1\right), \left((3,1)^T, -1\right), \left((2,3)^T, -1\right) \right\}$$

we obtain,

$$X = \begin{pmatrix} \widehat{\mathbf{x}}_1'^T \\ \widehat{\mathbf{x}}_2'^T \\ \widehat{\mathbf{x}}_3'^T \\ \widehat{\mathbf{x}}_4'^T \end{pmatrix} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \\ -1 & -3 & -1 \\ -1 & -2 & -3 \end{pmatrix}.$$

Whence,

$$X^T X = \begin{pmatrix} 4 & 8 & 6 \\ 8 & 18 & 11 \\ 6 & 11 & 14 \end{pmatrix}, \quad \text{and,} \quad X^\dagger = (X^T X)^{-1} X^T = \begin{pmatrix} \frac{5}{4} & \frac{13}{12} & \frac{3}{4} & \frac{7}{12} \\ -\frac{1}{2} & -\frac{1}{6} & -\frac{1}{2} & -\frac{1}{6} \\ 0 & -\frac{1}{3} & 0 & -\frac{1}{3} \end{pmatrix}.$$

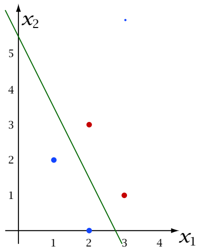
Example (cont.)

Letting, $\mathbf{b} = (1, 1, 1, 1)^T$, then

$$\widehat{\mathbf{w}} = X^{\dagger}\mathbf{b} = \begin{pmatrix} \frac{5}{4} & \frac{13}{12} & \frac{3}{4} & \frac{7}{12} \\ -\frac{1}{2} & -\frac{1}{6} & -\frac{1}{2} & -\frac{1}{6} \\ 0 & -\frac{1}{3} & 0 & -\frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{11}{3} \\ -\frac{4}{3} \\ -\frac{2}{3} \end{pmatrix}.$$

Whence,

$$w_0 = \frac{11}{3} \quad \text{and} \quad \mathbf{w} = \left(-\frac{4}{3}, -\frac{2}{3}\right)^T$$



Method of Steepest Descent

An alternative approach is the *method of steepest descent*.

We begin by representing Taylor's theorem for functions of more than one variable: let $\mathbf{x} \in \mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}$, so

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n) \in \mathbb{R}.$$

Now let $\delta\mathbf{x} \in \mathbb{R}^n$, and consider

$$f(\mathbf{x} + \delta\mathbf{x}) = f(x_1 + \delta x_1, \ldots, x_n + \delta x_n).$$

Define $F : \mathbb{R} \to \mathbb{R}$, such that,

$$F(s) = f(\mathbf{x} + s\,\delta\mathbf{x}).$$

Thus,

$$F(0) = f(\mathbf{x}), \quad \text{and,} \quad F(1) = f(\mathbf{x} + \delta\mathbf{x}).$$

Method of Steepest Descent (cont.)

Taylor's theorem for a single variable (Math 21/22),

$$F(s) = F(0) + \frac{1}{1!}F'(0)s + \frac{1}{2!}F''(0)s^2 + \frac{1}{3!}F'''(0)s^3 + \cdots \,.$$

Our plan is to set $s = 1$ and replace $F(1)$ by $f(\mathbf{x} + \delta\mathbf{x})$, $F(0)$ by $f(\mathbf{x})$, etc.
To evaluate $F'(0)$ we will invoke the multivariate chain rule, e.g.,

$$\frac{d}{ds}\, f\big(u(s), v(s)\big) = \frac{\partial f}{\partial u}(u, v)\, u'(s) + \frac{\partial f}{\partial v}(u, v)\, v'(s).$$

Thus,

$$\begin{aligned}
F'(s) = \frac{dF}{ds}(s) &= \frac{df}{ds}(x_1 + s\,\delta x_1, \ldots, x_n + s\,\delta x_n) \\
&= \frac{\partial f}{\partial x_1}(\mathbf{x} + s\,\delta\mathbf{x})\frac{d}{ds}(x_1 + s\,\delta x_1) + \cdots + \frac{\partial f}{\partial x_n}(\mathbf{x} + s\,\delta\mathbf{x})\frac{d}{ds}(x_n + s\,\delta x_n) \\
&= \frac{\partial f}{\partial x_1}(\mathbf{x} + s\,\delta\mathbf{x}) \cdot \delta x_1 + \cdots + \frac{\partial f}{\partial x_n}(\mathbf{x} + s\,\delta\mathbf{x}) \cdot \delta x_n.
\end{aligned}$$

Method of Steepest Descent (cont.)

Thus,

$$F'(0) = \frac{\partial f}{\partial x_1}(\mathbf{x}) \cdot \delta x_1 + \cdots + \frac{\partial f}{\partial x_n}(\mathbf{x}) \cdot \delta x_n = \nabla f(\mathbf{x})^T \delta \mathbf{x}.$$

Method of Steepest Descent (cont.)

Thus, it is possible to show

$$\begin{aligned} f(\mathbf{x} + \delta\mathbf{x}) &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T \delta\mathbf{x} + \mathcal{O}\left(\|\delta\mathbf{x}\|^2\right) \\ &= f(\mathbf{x}) + \|\nabla f(\mathbf{x})\| \|\delta\mathbf{x}\| \cos\theta + \mathcal{O}\left(\|\delta\mathbf{x}\|^2\right), \end{aligned}$$

where θ defines the angle between $\nabla f(\mathbf{x})$ and $\delta\mathbf{x}$. If $\|\delta\mathbf{x}\| \ll 1$, then

$$\delta f = f(\mathbf{x} + \delta\mathbf{x}) - f(\mathbf{x}) \approx \|\nabla f(\mathbf{x})\| \|\delta\mathbf{x}\| \cos\theta.$$

Thus, the greatest *reduction* δf occurs if $\cos\theta = -1$, that is if $\delta\mathbf{x} = -\eta\nabla f$, where $\eta > 0$. We thus seek a local minimum of the LMS objective function by taking a sequence of steps

$$\widehat{\mathbf{w}}(t+1) = \widehat{\mathbf{w}}(t) - \eta\nabla\mathcal{E}\left(\widehat{\mathbf{w}}(t)\right).$$

Training an LTU using Steepest Descent

We now return to our original problem. Given a dichotomy

$$\mathcal{X}_m = \{(\mathbf{x}_1, \ell_1), \ldots, (\mathbf{x}_m, \ell_m)\}$$

of m feature vectors $\mathbf{x}_i \in \mathbb{R}^n$ with $\ell_i \in \{-1, 1\}$ for $i = 1, \ldots, m$, we construct the set of normalized, augmented feature vectors

$$\widehat{\mathcal{X}}'_m = \{(\ell_i, \ell_i x_{i,1}, \ldots, \ell_i x_{i,n})^T \in \mathbb{R}^{n+1} | i = 1, \ldots, m\}.$$

Given a margin vector, $\mathbf{b} \in \mathbb{R}^m$, with $b_i > 0$ for $i = 1, \ldots, m$, we construct the LMS objective function,

$$\mathcal{E}(\widehat{\mathbf{w}}) = \frac{1}{2} \sum_{i=1}^{m} (\widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i - b_i)^2 = \frac{1}{2} \|X\widehat{\mathbf{w}} - \mathbf{b}\|^2,$$

(the factor of $\frac{1}{2}$ is added with some foresight), and then evaluate its gradient

$$\nabla\mathcal{E}(\widehat{\mathbf{w}}) = \sum_{i=1}^{m} (\widehat{\mathbf{w}}^T \widehat{\mathbf{x}}'_i - b_i)\widehat{\mathbf{x}}'_i = X^T(X\widehat{\mathbf{w}} - \mathbf{b}).$$

Training an LTU using Steepest Descent (cont.)

Substitution into the steepest descent update rule,

$$\widehat{\mathbf{w}}(t+1) = \widehat{\mathbf{w}}(t) - \eta \nabla \mathcal{E}\left(\widehat{\mathbf{w}}(t)\right),$$

yields the *batch LMS update rule*,

$$\widehat{\mathbf{w}}(t+1) = \widehat{\mathbf{w}}(t) + \eta \sum_{i=1}^{m} \left(b_i - \widehat{\mathbf{w}}(t)^T \widehat{\mathbf{x}}'_i\right) \widehat{\mathbf{x}}'_i.$$

Alternatively, one can abstract the *sequential LMS*, or *Widrow-Hoff rule*, from the above:

$$\widehat{\mathbf{w}}(t+1) = \widehat{\mathbf{w}}(t) + \eta \left(b - \widehat{\mathbf{w}}(t)^T \widehat{\mathbf{x}}'(t)\right) \widehat{\mathbf{x}}'(t).$$

where $\widehat{\mathbf{x}}'(t) \in \widehat{\mathcal{X}}'_m$ is the element of the dichotomy that is presented to the LTU at epoch t. (Here, we assume that b is fixed; otherwise, replace it by $b(t)$.)

Sequential LMS Rule

The sequential LMS rule

$$\widehat{\mathbf{w}}(t+1) = \widehat{\mathbf{w}}(t) + \eta\left[b - \widehat{\mathbf{w}}(t)^T \widehat{\mathbf{x}}'(t)\right]\widehat{\mathbf{x}}'(t),$$

resembles the sequential perceptron rule,

$$\widehat{\mathbf{w}}(t+1) = \widehat{\mathbf{w}}(t) + \frac{\eta}{2}\left[1 - \operatorname{sgn}\left(\widehat{\mathbf{w}}(t)^T \widehat{\mathbf{x}}'(t)\right)\right]\widehat{\mathbf{x}}'(t).$$

Sequential rules are well suited to *real-time implementations*, as only the current values for the weights, i.e. the configuration of the LTU itself, need to be stored. They also work with dichotomies of infinite sets.

Convergence of the batch LMS rule

Recall, the LMS objective function in the form

$$\mathcal{E}(\widehat{\mathbf{w}}) = \frac{1}{2}\left\| X\widehat{\mathbf{w}} - \mathbf{b} \right\|^2,$$

has as its gradient,

$$\nabla\mathcal{E}(\widehat{\mathbf{w}}) = X^T X\widehat{\mathbf{w}} - X^T\mathbf{b} = X^T(X\widehat{\mathbf{w}} - \mathbf{b}).$$

Substitution into the rule of steepest descent,

$$\widehat{\mathbf{w}}(t+1) = \widehat{\mathbf{w}}(t) - \eta\,\nabla\mathcal{E}\left(\widehat{\mathbf{w}}(t)\right),$$

yields,

$$\widehat{\mathbf{w}}(t+1) = \widehat{\mathbf{w}}(t) - \eta\, X^T\left(X\widehat{\mathbf{w}}(t) - \mathbf{b}\right)$$

Convergence of the batch LMS rule (cont.)

The algorithm is said to *converge to a fixed point* $\widehat{\mathbf{w}}^\star$, if for every finite initial value $\left\|\widehat{\mathbf{w}}(0)\right\| < \infty$,

$$\lim_{t\to\infty} \widehat{\mathbf{w}}(t) = \widehat{\mathbf{w}}^\star.$$

The fixed points $\widehat{\mathbf{w}}^\star$ satisfy $\nabla\mathcal{E}(\widehat{\mathbf{w}}^\star) = 0$, whence

$$X^T X \widehat{\mathbf{w}}^\star = X^T \mathbf{b}.$$

The update rule becomes,

$$\begin{aligned} \widehat{\mathbf{w}}(t+1) &= \widehat{\mathbf{w}}(t) - \eta\, X^T\big(X\widehat{\mathbf{w}}(t) - \mathbf{b}\big) \\ &= \widehat{\mathbf{w}}(t) - \eta\, X^T X\big(\widehat{\mathbf{w}}(t) - \widehat{\mathbf{w}}^\star\big) \end{aligned}$$

Let $\delta\widehat{\mathbf{w}}(t) \stackrel{\text{def}}{=} \widehat{\mathbf{w}}(t) - \widehat{\mathbf{w}}^\star$. Then,

$$\delta\widehat{\mathbf{w}}(t+1) = \delta\widehat{\mathbf{w}}(t) - \eta\, X^T X \delta\widehat{\mathbf{w}}(t) = \big(I - \eta\, X^T X\big)\delta\widehat{\mathbf{w}}(t).$$

Convergence of the batch LMS rule (cont.)

Convergence $\widehat{\mathbf{w}}(t) \to \widehat{\mathbf{w}}^\star$ occurs if $\delta\widehat{\mathbf{w}}(t) = \widehat{\mathbf{w}}(t) - \widehat{\mathbf{w}}^\star \to 0$. Thus we require that $\left\|\delta\widehat{\mathbf{w}}(t+1)\right\| < \left\|\delta\widehat{\mathbf{w}}(t)\right\|$. Inspecting the update rule,

$$\delta\widehat{\mathbf{w}}(t+1) = \left(I - \eta\, X^T X\right)\delta\widehat{\mathbf{w}}(t),$$

this reduces to the condition that all the eigenvalues of

$$I - \eta\, X^T X$$

have magnitudes less than 1.

We will now evaluate the eigenvalues of the above matrix.

Convergence of the batch LMS rule (cont.)

Let $S \in \mathbb{R}^{(n+1)\times(n+1)}$ denote the *similarity transform* that reduces the symmetric matrix $X^T X$ to a diagonal matrix $\Lambda \in \mathbb{R}^{(n+1)\times(n+1)}$. Thus, $S^T S = S S^T = I$, and

$$S\, X^T X\, S^T = \Lambda = \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_n),$$

The numbers, $\lambda_0, \lambda_1, \ldots, \lambda_n$, represent the eigenvalues of $X^T X$. Note that $0 \leq \lambda_i$ for $i = 0, 1, \ldots, n$. Thus,

$$\begin{aligned} S\,\delta\widehat{\mathbf{w}}(t+1) &= S(I - \eta X^T X) S^T S\,\delta\widehat{\mathbf{w}}(t), \\ &= (I - \eta\Lambda) S\,\delta\widehat{\mathbf{w}}(t). \end{aligned}$$

Thus convergence occurs if

$$\|S\,\delta\widehat{\mathbf{w}}(t+1)\| < \|S\,\delta\widehat{\mathbf{w}}(t)\|,$$

which occurs if the eigenvalues of $I - \eta\Lambda$ all have magnitudes less than one.

Convergence of the batch LMS rule (cont.)

The eigenvalues of $I - \eta\Lambda$ equal $1 - \eta\lambda_i$, for $i = 0, 1, \ldots, n$. (These are in fact the eigenvalues of $I - \eta X^T X$.) Thus, we require that

$$-1 < 1 - \eta\lambda_i < 1, \quad \text{or} \quad 0 < \eta\lambda_i < 2,$$

for all i. Let $\lambda_{\max} = \max_{0 \le i \le n} \lambda_i$ denote the largest eigenvalue of $X^T X$, then convergence requires that

$$0 < \eta < \frac{2}{\lambda_{\max}} \quad .$$