On the Normalization of the Stepsize in Nonsymmetric Stochastic Gradient Algorithms

Dirk T.M. Slock
Eurecom Institute
2229 route des Crêtes, Sophia Antipolis
F-06560 Valbonne, FRANCE

Abstract

The normalization of the stepsize in the Least Mean Square (LMS) algorithm allows for an easy control of the range of stable operation for the normalized stepsize in the normalized LMS (NLMS) algorithm, and also for an easy determination of the stepsize for maximum convergence speed. In this paper, we consider stochastic gradient algorithms in which the gradient vector differs from the data vector. For such "nonsymmetric" stochastic gradient algorithms, we propose a generalized stepsize normalization. We shall consider in detail the following three applications: the stochastic Newton scheme, the sign-data LMS algorithm, and a certain instrumental variable method recently proposed to speed up the convergence of the LMS algorithm.

1 Motivation

In order to maximize the convergence speed of a stochastic gradient algorithm, a large stepsize is needed. Also, when one wants to address the issue of the maximum stepsize at which the algorithm converges, one needs a theory that is valid beyond an infinitesimally small stepsize range. At this time, almost no exact theory for the evolution (learning curve) of the excess mean squared error (EMSE) exists. Exact results for the EMSE in the LMS algorithm actually only exists for the asymptotic case of small stepsize μ [1]. However, the outlook is more optimistic when we focus on the stepsize issue, at least for the LMS algorithm. The LMS algorithm is described by the following equations

$$\begin{aligned} \epsilon_k^p &= d_k - W_{k-1}^H X_k \\ W_k &= W_{k-1} + X_k \mu_k \epsilon_k^{pH} \end{aligned} \tag{1}$$

where the superscript H denotes Hermitian (complex conjugate) transpose, $X_k = \left[x_k^H \;\; x_{k-1}^H \cdots x_{k-N+1}^H\right]^H$ is the regression vector, x_k and d_k are the input and desired-response signals respectively and can be complex matrix valued in general (which should explain the special way in which the update equation for W_k above is written), W_k contains the set of N filter coefficients which are of dimensions commensurate with d_k and x_k, and μ_k is the stepsize which in general can be a complex matrix of appropriate dimensions. Below, we shall write the algorithm equations such that the matrix nature of the various quantities is respected, but we shall often restrict the analysis to the case where x_k is a real scalar for simplicity. Assume at first that the minimum MSE (MMSE) is zero or in other words, the desired-response $d_k = W^{oH} X_k$ is the output of an optimal FIR filter W^o. Then we get, with $\widetilde{W}_k = W^o - W_k$ the error in the estimated filter coefficients,

$$\begin{aligned} \epsilon_k^p &= \widetilde{W}_{k-1}^H X_k \\ \widetilde{W}_k &= \Phi_k \widetilde{W}_{k-1} \,, \; \Phi_k = \left[I - X_k \mu_k X_k^H\right] \end{aligned} \tag{2}$$

The eigenvalue distribution of the matrix Φ_k is

$$\begin{cases} 1 & , \text{ multiplicity } N-1 \\ \lambda_1 = 1 - \mu_k \|X_k\|^2 & , \text{ multiplicity } 1 \end{cases} \tag{3}$$

At the time instant k, the error in the filter estimate gets reduced only in the direction of X_k, by a factor $|\lambda_1|$. In order to maximize convergence speed, a constant stepsize $\mu_k \equiv \mu$ has to be chosen fairly large. However, especially for signals with a medium or high kurtosis, this will imply that $|\lambda_1| > 1$ at many time instants. So, although the algorithm will be converging on the average, certain updates at isolated points in time will actually represent diverging steps. This means an inefficient use of data. To maximize convergence speed, we clearly have to choose $\mu_k = \left(X_k^H X_k\right)^{-1} \Rightarrow \lambda_1 = 0 \,, \; \Phi_k = I - P_{X_k} = P_{X_k}^{\perp}$ where P_{X_k} is the projection matrix onto the subspace spanned by X_k and $P_{X_k}^{\perp}$ is the projection ma-

1058-6393/92 $03.00 © 1992 IEEE

trix onto its orthogonal complement. This normalization ensures that the filter estimate error component in the direction of X_k gets nulled out, which is the best we can do. More generally, in order to have a well controlled convergence behavior, we should choose $\mu_k = \overline{\mu}/\|X_k\|^2$. This leads to $\lambda_1 = 1-\overline{\mu}$, which is stable for $\overline{\mu} \in (0,2)$. This normalization allows for a very much predictible convergence behavior, with λ_1 and the stable range for $\overline{\mu}$ being independent of the input signal statistics. The thus normalized LMS (NLMS) algorithm has been introduced in [2] and is also known as the *projection algorithm* [3].

1.1 Example: Sign-Data LMS

In this paper, we consider more general stochastic gradient algorithms in which the gradient vector differs from the data vector. A typical example is the *sign-data LMS* algorithm, which can be described by the following equations:

$$\begin{aligned} \epsilon_k^p &= d_k - W_{k-1}^H X_k \\ W_k &= W_{k-1} + \operatorname{sign}(X_k)\,\mu_k \epsilon_k^{pH} \end{aligned} \tag{4}$$

For noise-free system identification, we get

$$\begin{aligned} \epsilon_k^p &= \widetilde{W}_{k-1}^H X_k \\ \widetilde{W}_k &= \Phi_k \widetilde{W}_{k-1}\,,\ \Phi_k = \left[I - \operatorname{sign}(X_k)\,\mu_k X_k^H\right] \end{aligned} \tag{5}$$

The eigenvalue λ_1 of Φ_k is now:

$$\lambda_1 = 1 - \mu_k X_k^H \operatorname{sign}(X_k) = 1 - \mu_k \sum_{i=0}^{N-1} |x_k| \quad . \tag{6}$$

For maximum convergence speed, we put again $\lambda_1 = 0$ which leads to $\mu_k = \left(X_k^H \operatorname{sign}(X_k)\right)^{-1}$. So the normalized stepsize becomes $\mu_k = \overline{\mu}\left(X_k^H \operatorname{sign}(X_k)\right)^{-1}$ which leads to $|\lambda_1| = |1-\overline{\mu}| < 1$ for $\overline{\mu} \in (0,2)$, independently of the signal statistics. Nagumo and Noda [2] proposed this normalization and found:

- $\overline{\mu} = 1$ does not necessarily lead to fastest convergence
- the range of stable operation for $\overline{\mu}$ can be smaller than $(0,2)$ or even $(0,1)$

Hence, $|\lambda_1|$ is not the proper indicator for convergence.

2 General Stochastic Gradient Algorithm

In this paper, we shall consider more general stochastic gradient algorithms of the form

$$\begin{aligned} \epsilon_k^p &= d_k - W_{k-1}^H X_k \\ W_k &= W_{k-1} + Y_k \mu_k \epsilon_k^{pH} \end{aligned} \tag{7}$$

where X_k is the data vector and Y_k is (part of) the gradient vector. The filter estimate W_k converges to a filter W satisfying $\mathrm{E}\left\{\left(d_k - W^H X_k\right) Y_k\right\} = 0$ or hence $W = R_{YX}^{-1} R_{Yd}$ (where $R_{ab} = \mathrm{E}\, ab^H$). For noise-free system identification, we get

$$\begin{aligned} \epsilon_k^p &= \widetilde{W}_{k-1}^H X_k \\ \widetilde{W}_k &= \Phi_k \widetilde{W}_{k-1}\,,\ \Phi_k = \left[I - Y_k \mu_k X_k^H\right] \end{aligned} \tag{8}$$

We again consider the issue of finding the stepsize sequence that will lead to fastest convergence, and the related issue of stepsize normalization such that we can specify bounds on the normalized stepsize that will guarantee algorithm convergence. For reasons similar to those that led to the normalized LMS algorithm, we shall limit the scope for possible stepsize normalizations:

- normalization based on convergence dynamics considerations (and not measurement noise amplification, double-talk detection etc.). The normalization should furthermore only depend on the instantaneous dynamics, hence on Φ_k
- the normalized stepsize should be independent of the "level" of X_k and Y_k, and hence should only depend on the angle θ_k between X_k and Y_k

So we shall restrict the stepsize sequence to be of the form

$$\mu_k = \overline{\mu}_k f(X_k, Y_k) = \overline{\mu}\frac{g(\cos\theta_k)}{\|X_k\|\,\|Y_k\|} \tag{9}$$

where $\cos\theta_k = \overline{Y}_k^H \overline{X}_k$ and $\overline{X}_k = X_k \left(X_k^H X_k\right)^{-H/2}$, $\overline{Y}_k = Y_k \left(Y_k^H Y_k\right)^{-H/2}$ are normalized vectors. The state transition matrix becomes

$$\Phi_k = I - \overline{\mu}\,\overline{Y}_k\, g(\cos\theta_k)\,\overline{X}_k^H \quad . \tag{10}$$

In the *instrumental variable* method [4], the gradient vector Y_k has the same structure as the data vector X_k, except that it is filled up with the instrumental variables y_k: $Y_k = \left[y_k^H\ y_{k-1}^H \cdots y_{k-N+1}^H\right]^H$. Under certain conditions, we can apply a Law of Large Numbers to arrive at

$$\begin{aligned} \cos\theta_k &= \overline{X}_k^H \overline{Y}_k = \rho_{xy} + \mathcal{O}\left(\tfrac{1}{\sqrt{N}}\right) \\ \rho_{xy} &= R_{xx}^{-1/2} R_{xy} R_{yy}^{-H/2} \end{aligned} \tag{11}$$

where $\mathcal{O}(c)$ is a zero mean random variable with standard deviation $O(c)$. So for high filter orders, the angle between the vectors X_k and Y_k gets highly concentrated around a value determined by the crosscorrelation between the two signals.

2.1 Projection Algorithm

With the restrictions on the normalization put forward in equation (9), the normalization design issue becomes an issue of determining the function $g(\cos\theta_k) = ?$. For the convergence of the mean of $\widetilde{W}_k$, we need to consider the eigenvalues of Φ_k. This matrix has one nontrivial eigenvalue $\lambda_1 = 1 - \overline{\mu} g(\cos\theta_k)\cos\theta_k$ while the other eigenvalues equal 1. Considering λ_1, the proper stepsize normalization would be

$$g(\cos\theta_k) = 1/\cos\theta_k \;\Rightarrow\; \lambda_1 = 1 - \overline{\mu} \;. \tag{12}$$

This leads to an *oblique projection*

$$\Phi_k = I - \overline{\mu}\, P_{Y_k;X_k},\; P_{Y_k;X_k} = Y_k \left(X_k^H Y_k\right)^{-1} X_k^H \tag{13}$$

where $P_{Y_k;X_k}$ is the projection onto Y_k along the orthogonal complement of X_k. While the *a priori* error ϵ_k^p is computed using the previous filter estimate W_{k-1}, the *a posteriori* error ϵ_k is computed using the updated filter estimate W_k:

$$\epsilon_k = d_k - W_k^H X_k = \lambda_1 \epsilon_k^p \tag{14}$$

The update of the filter estimate results in a reduction of the error signal by a factor λ_1. However, we know from the sign-data LMS example that $|\lambda_1|$ is not the proper indicator for convergence.

2.2 Convergence of the Learning Curve

However, we are not just interested in the convergence of the mean, but especially in the convergence of the learning curve, and hence in the convergence of the second-order moments of $\widetilde{W}_k$. To illustrate how this comes about, consider the system identification setup: the desired response $d_k = W^{oH} X_k + n_k$ is the sum of the output of an optimal filter W^o plus some independent zero-mean i.i.d. measurement noise with variance $\xi^o = \mathrm{E}\, n_k^2$. So the filter estimation error system for algorithm (7) is

$$\begin{aligned} \epsilon_k^p &= \widetilde{W}_{k-1}^H X_k + n_k \\ \widetilde{W}_k &= \Phi_k \widetilde{W}_{k-1} - Y_k\, \mu_k\, n_k^H \;. \end{aligned} \tag{15}$$

With $COV_k = \mathrm{E}\left(\widetilde{W}_k \widetilde{W}_k^H\right)$ and introducing the independence assumption (treating X_k and $\widetilde{W}_{k-1}$ as independent), the learning curve becomes

$$\begin{aligned} \xi_k &= \mathrm{E}\,(\epsilon_k^p)^2 = \mathrm{trace}\,(R_{XX}\, COV_{k-1}) + \xi^o \\ COV_k &= \mathrm{E}\,\left(\Phi_k\, COV_{k-1}\, \Phi_k^H\right) + \xi^o\, \mathrm{E}\,\left(\mu_k^2\, Y_k Y_k^H\right) \end{aligned} \tag{16}$$

So the dynamics of the learning curve are determined by the second-order statistics of Φ_k. However, for the purpose of normalization, we should omit the averaging operation in $\mathrm{E}\,\left(\Phi_k\, COV_{k-1}\, \Phi_k^H\right)$. So the instantaneous dynamics for second-order statistics governed by $\Phi_k \Phi_k^H$. The eigenvalues of $\Phi_k \Phi_k^H$ equal the (singular values of $\Phi_k)^2$. The singular value distribution of Φ_k is :

$$\begin{cases} 1 & \text{multiplicity } N-2 \\ \sigma_1 > 1 > \sigma_2 & \text{when } Y_k \neq \alpha_k X_k \end{cases} \tag{17}$$

When Y_k is proportional to X_k, then $\sigma_1 = |\lambda_1|$, $\sigma_2 = 1$, but this is basically the case of the LMS algorithm which we shall not consider further. For any vector V, $\|\Phi_k V\| \leq \sigma_1 \|V\|$, so a worst case analysis of the convergence may want to concentrate on σ_1. However, σ_1 is not the proper convergence indicator either: $\sigma_1 > 1$ whenever Y_k and X_k are not proportional, but convergence occurs nevertheless.

3 The Proper Normalization

3.1 A First Attempt

So we have for the two non-trivial singular values of Φ_k

- $\sigma_1 > 1$: in one direction, the error $\widetilde{W}_{k-1}$ gets amplified,
- $\sigma_2 < 1$: in another direction, $\widetilde{W}_{k-1}$ gets reduced.

Any particular direction in $\Re^N$ gets sometimes amplified, sometimes reduced, so we have to consider an averaged action. Proposed convergence measure, representative of the averaged convergence action:

$$\frac{1}{N}\,\mathrm{trace}\,\left(\Phi_k \Phi_k^H\right) = \frac{2}{N}\frac{\sigma_1^2 + \sigma_2^2}{2} + \frac{N-2}{N}. \tag{18}$$

In order to find the normalization that leads to the maximum convergence speed, consider minimizing the convergence measure (18) with $\overline{\mu} = 1$ (recall: $\Phi_k = I - \overline{Y}_k\, g(\cos\theta_k) \overline{X}_k^H$):

$$\min_g \frac{\sigma_1^2 + \sigma_2^2}{2} \;\Rightarrow\; \begin{cases} g\,(\cos\theta_k) = \cos\theta_k \;, \\ \Phi_k \;=\; I - P_{Y_k}\, P_{X_k} \;. \end{cases} \tag{19}$$

Here is another point of view: ϵ_k^p measures the component of $\widetilde{W}_{k-1}$ along X_k. Hence after the update, that component should be reduced since we are using ϵ_k^p to update the filter estimate $\Rightarrow$

$$\min_g \|\Phi_k X_k\|^2 \;\Rightarrow\; g\,(\cos\theta_k) = \cos\theta_k \tag{20}$$

leading to the same normalization. The normalizations we proposed so far can be summarized as follows.

criterion	λ_1	$\frac{\sigma_1^2+\sigma_2^2}{2}$
$g(\cos\theta_k)$	$\frac{1}{\cos\theta_k}$	$\cos\theta_k$
μ_k $(\overline{\mu}=1)$	$(X_k^H Y_k)^{-1}$	$(Y_k^H Y_k)^{-1}(Y_k^H X_k)(X_k^H X_k)^{-1}$
λ_1	0	$\sin^2\theta_k < 1$
$\frac{\sigma_1^2+\sigma_2^2}{2}$	$\frac{1}{2\cos^2\theta_k}$	$1-\frac{1}{2}\cos^2\theta_k < 1$
Φ_k	$I - P_{Y_k;X_k}$	$I - P_{Y_k}P_{X_k}$
stepsize	audacious	cautious

Note that the stepsize that minimizes $\frac{\sigma_1^2+\sigma_2^2}{2}$, leads to $|\lambda_1| < 1$ and $\frac{\sigma_1^2+\sigma_2^2}{2} < 1$, while the stepsize that minimizes $|\lambda_1|$ leads to $\lambda_1 = 0$, but $\frac{\sigma_1^2+\sigma_2^2}{2} > 1$ when $\theta_k > \frac{\pi}{4}$. $g(\cos\theta_k) = \cos\theta_k$ leads to a cautious stepsize since the stepsize gets reduced as the gradient vector Y_k and the data vector X_k (in which direction the error $\widetilde{W}_k$ gets measured) are less aligned. $g(\cos\theta_k) = 1/\cos\theta_k$ on the other hand leads to a audacious stepsize since the stepsize increases as Y_k and X_k become less aligned.

3.2 Counterexample: Stochastic Newton

Is $\frac{\sigma_1^2+\sigma_2^2}{2}$ the right criterion for convergence (if it is, then $\frac{\sigma_1^2+\sigma_2^2}{2} = 1$ should be the boundary between convergence and divergence)? Consider the stochastic Newton algorithm:

$$\begin{aligned} \epsilon_k^p &= d_k - W_{k-1}^H X_k \\ W_k &= W_{k-1} + R_{XX}^{-1} X_k \mu_k \epsilon_k^{pH} \end{aligned} \tag{21}$$

This is a special case of the general stochastic gradient algorithm with $Y_k = R_{XX}^{-1} X_k$. For noise-free system identification, the error system becomes

$$\begin{aligned} \widetilde{W}_k &= \Phi_k \widetilde{W}_{k-1} \\ \Phi_k &= I - R_{XX}^{-1} X_k \mu_k X_k^H \end{aligned} \tag{22}$$

What is the proper stepsize normalization in this case? Consider the following parameter transformation: $U = R_{XX}^{H/2} W$, then the error system gets transformed into

$$\begin{aligned} \widetilde{U}_k &= \Phi_k \widetilde{U}_{k-1} \\ \Phi_k &= I - \mu_k Z_k Z_k^H \ , \ Z_k = R_{XX}^{-1/2} X_k \end{aligned} \tag{23}$$

which is a symmetric system! Hence, the optimal normalized stepsize is

$$\mu_k = (Z_k^H Z_k)^{-1} = (X_k^H R_{XX}^{-1} X_k)^{-1} = (X_k^H Y_k)^{-1} \tag{24}$$

resulting from minimizing $|\lambda_1|$ instead of $\frac{\sigma_1^2+\sigma_2^2}{2}$.

3.3 Second Attempt

With an arbitrary parameter transformation $U = A\,W$, the state transition matrix becomes $\Phi_k = I - A\,Y_k\,\mu_k\,X_k^H\,A^{-1}$. The resulting nontrivial eigenvalue $\lambda_1 = 1 - \mu_k X_k^H Y_k$ remains unchanged! However, such a parameter transformation alters $\frac{\sigma_1^2+\sigma_2^2}{2}$. For a final convergence measure, we should avoid pessimism. So if there exists a parameter transformation for which $\frac{\sigma_1^2+\sigma_2^2}{2} < 1$, then we should not conclude divergence. So consider as final criterion for the optimal stepsize:

$$\min_{\mu_k}\left\{\min_{A}\frac{\sigma_1^2+\sigma_2^2}{2}\right\} \tag{25}$$

With $B = A^H A$, the optimization problem (25) leads to

$$\begin{aligned} & \min_{B=B^H>0}(X_k^H B^{-1} X_k)(Y_k^H B\,Y_k) \\ \approx\ & \min_{B=B^H>0}(\mathrm{E}\,X_k^H B^{-1} X_k)(\mathrm{E}\,Y_k^H B\,Y_k) \\ =\ & \min_{B=B^H>0}(\mathrm{trace}\ B^{-1} R_{XX})(\mathrm{trace}\ B\,R_{YY}) \end{aligned} \tag{26}$$

where the approximation holds for large filter orders (using the law of large numbers we have invoked before). Carrying out the optimization leads to

$$\begin{aligned} & B^*\,R_{YY}\,B^* = R_{XX} \\ & B^* = R_{XX}^{1/2}\left(R_{XX}^{1/2} R_{YY} R_{XX}^{1/2}\right)^{-1/2} R_{XX}^{1/2} \\ & \mu_k^* = (Y_k^H B^* Y_k)^{-1}(Y_k^H X_k)(X_k^H B^{-*} X_k)^{-1} \end{aligned} \tag{27}$$

where all matrix square-roots in the explicit solution for B^* should be taken to be symmetric. If we introduce a normalized stepsize $\mu_k = \overline{\mu}\mu_k^*$ based on the optimal stepsize μ_k^* derived above, then obviously $\overline{\mu} = 1$ is optimal, while $\overline{\mu} \in (0,2)$ is the stable range (meaning that $\min_A \frac{\sigma_1^2+\sigma_2^2}{2} < 1$ in that range). We also have for large filter orders that

$$\begin{aligned} & (Y_k^H B^* Y_k)^{-1}(Y_k^H X_k)(X_k^H B^{-*} X_k)^{-1} \\ & \geq (Y_k^H Y_k)^{-1}(Y_k^H X_k)(X_k^H X_k)^{-1} \end{aligned} \tag{28}$$

so that $\mu_k(B = I) \leq \mu_k(B = B^*)$: stability does not get compromized by replacing B^* by the simpler identity matrix, leading to the normalized stepsize we found in our first attempt.

4 Examples

4.1 Example 1: Stochastic Newton

With $Y_k = R^{-1}X_k$ for in fact any $R = R^H > 0$, we have $R_{YY} = R^{-1}R_{XX}R^{-1}$ and hence $B^* = R$. Thus

$$\mu_k^* = \left(Y_k^H B^* Y_k\right)^{-1}\left(Y_k^H X_k\right)\left(X_k^H B^{-*} X_k\right)^{-1} = \left(X_k^H Y_k\right)^{-1} \tag{29}$$

which is the projection algorithm solution we have found before (see (24)).

4.2 Example 2: Sign-Data LMS

In this case $Y_k = \text{sign}(X_k)$. For Gaussian x_k: $R_{YY} = \frac{2}{\pi}\arcsin\left(\frac{1}{R_{XX}(1,1)}R_{XX}\right)$. Simulations on Gaussian AR processes with the optimal B^* appear to confirm that $\overline{\mu} \in (0,2)$ is the stepsize interval for convergence, and $\overline{\mu} = 1$ is optimal.

We may note that in this case of the sign-data LMS algorithm, the matrix R_{YY} is quite "similar" to R_{XX}. On the other hand, the use of the optimal B^* leads to an algorithm of which the computational complexity is no longer $O(N)$ (apart from the problem that quite some knowledge about the input process is required). Hence one may wonder how much we lose by replacing the optimal B^* by I, which leads to a computationally simple algorithm that does not require any knowledge about the input signal statistics. When $Y_k = \left[y_k^H \cdots y_{k-N+1}^H\right]^H$, we have the following general result

$$\begin{aligned} h &= \frac{\mu_k(B=B^*)}{\mu_k(B=I)} \\ &\approx \frac{(\text{trace } R_{XX})(\text{trace } R_{YY})}{(\text{trace } B^{-*}R_{XX})(\text{trace } B^* R_{YY})} \\ &\to \frac{1}{\cos^2\phi} \end{aligned} \tag{30}$$

where the limit holds for large filter orders and ϕ is the angle between the square-roots of the power spectral densities of the processes x_k and y_k. For AR(1) processes with pole a and filter order $N = 10$, we have calculated :

$$\max_{a\in(-1,1)} h = 1.06 \tag{31}$$

which is rather close to 1. It appears that for the case of the sign-data LMS algorithm, using $B = I$ leads to only little suboptimality.

4.3 Example 3: Prewhitened IV LMS

In this case, $Y_k = \left[y_k^H \cdots y_{k-N+1}^H\right]^H$ where $y_k = \text{innov}\{x_k\}$ is the whitened version of the input signal. This instrumental variable (IV) method was proposed in [5] as a simple faster converging alternative to the conventional LMS algorithm. Of course, the innovations of the input signal have to be generated by some second (in general low order) adaptive filter. We shall assume for simplicity that this whitening filter works perfectly. We get $R_{YY} = I$ (we can assume normalization of the innovations w.l.o.g.). Hence

$$B^* = R_{XX}^{1/2} . \tag{32}$$

Limited simulation experience with the optimal B^* appears again to confirm that $\overline{\mu} \in (0,2)$ is the stepsize interval for convergence, and $\overline{\mu} = 1$ is optimal. Since the use of the optimal B^* leads again to a complicated algorithm, requiring quite some a priori knowledge about the input signal, it is of interest to investigate how much we lose by using the suboptimal $B = I$. We find the following upper bound

$$h = \frac{N\sum_{i=1}^{N}\lambda_i}{\left(\sum_{i=1}^{N}\sqrt{\lambda_i}\right)^2} \le \frac{\sigma_x^2}{\sigma_y^2} \tag{33}$$

where the λ_i are the eigenvalues of R_{XX}. However, this upper bound turns out to be quite loose in general. In fact, it can be shown that

$$h = O(\frac{\sigma_x}{\sigma_y}) \tag{34}$$

whenever the indicated ratio does not get too large. For example: AR(1) with $N = 10, a = \pm 0.9$: $h = 2.15$, $\frac{\sigma_x}{\sigma_y} = 2.29$

References

[1] V. Solo. "The Error Variance of LMS with Time-Varying Weights". *IEEE Trans. Signal Processing*, 40:803–813, April 1992.

[2] J.I. Nagumo and A. Noda. " A Learning Method for System Identification". *IEEE Trans. Autom. Cont.*, AC-12:282–287, June 1967.

[3] G.C. Goodwin and K.S. Sin. *Adaptive Filtering Prediction and Control.* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[4] T. Söderström and P.G. Stoica. *Instrumental Variable Methods for System Identification.* Springer-Verlag, Berlin, 1983.

[5] M. Mboup, M. Bonnet, and N. Bershad. "Coupled adaptive Prediction and System Identification: A Statistical Model and Transient Analysis". In *Proc. IEEE Int. Conf. ASSP*, San Fransisco, CA, March 1992.