Bursting in the LMS Algorithm

Markus Rupp

***Abstract*—The least mean square (LMS) algorithm is known to converge in the mean and in the mean square. However, during short time periods, the error sequence can blow up and cause severe disturbances, especially for non-Gaussian processes. This contribution discusses potential short time unstable behavior of the LMS algorithm for spherically invariant random processes (SIRP) like Gaussian, Laplacian, and K_0. The result of this investigation is that the probability for bursting decreases with the step size. However, since a smaller step size also causes a slower convergence rate, one has to choose a tradeoff between convergence speed and the frequence of bursting.**

I. Introduction

For the past few decades, the convergence of the least mean square (LMS) algorithm has been analyzed, either in the mean or in the mean square, depending on the statistics of the input sequence. Since these analyses are based on ensemble averaging, it is still possible that the error sequence for a particular realization bursts up during a limited time period. In speech applications, like adaptive hybrids, hands-free telephone sets, or noise control, these bursts severely disturb a conversation. As we shall see, the probability of the occurrence of the bursting phenomenon depends strongly on the statistics of the input sequence and the chosen step size. For unbounded processes, like Gaussian, Laplacian, or K_0, the probability of failure will be nonzero, even if the step size is very small. We shall see that for step sizes that correspond to the fastest convergence speeds, bursting can happen relatively often, and thus, smaller step sizes and, therefore, lower convergence speed may be necessary.[1]

The update equation for the LMS algorithm is given by

$$\hat{\mathbf{w}}(k+1) = \hat{\mathbf{w}}(k) + \mu \mathbf{u}(k)[d(k) - \mathbf{u}^T(k)\hat{\mathbf{w}}(k)]$$

where $d(k) = \mathbf{u}^T(k)\mathbf{w} + r(k)$ is the observed plant output corrupted by additive noise $r(k)$, and $\mathbf{u}(k)$ is a vector with M samples of the input sequence. If the algorithm is written in state-space form using the weight-error vector $\epsilon(k)$ (which is the difference between the estimated taps $\hat{\mathbf{w}}(k)$ and the optimal solution $\mathbf{w}$), the update equations become

$$\epsilon(k+1) = [\mathbf{I} - \mu \mathbf{u}(k)\mathbf{u}^T(k)]\epsilon(k) - \mu \mathbf{u}(k) r(k) \qquad (1)$$

where $\mathbf{I}$ is an M-dimensional identity matrix. Since we are not interested in the effect of the noise $r(k)$, we consider only the homogeneous part of the equation and do not take the noise into account. Thus, the algorithm's contraction mapping property with respect to the weight-error vector can be applied. Deterministic descriptions including the effect of the noise can be found in [1]. The

Manuscript received February 15, 1994; revised April 5, 1995. This work was supported by a scholarship from DAAD (German Academic Exchange Service) as well as the scientifical division of NATO (Wissenschaftsausschuß der NATO). The associate editor coordinating the review of this paper and approving it for publication was Dr. Fuyun Ling.

The author is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA.

IEEE Log Number 9413867.

[1] In practice, all input sequences are bounded because of the A/D converters, and using a very small step size would be appropriate to assure convergence. A small step size, however, will not be appropriate for fast convergence, and the improving effect of the adaptive filter will be compromised. To simplify matters, we shall completely neglect the bounding effect of the converters.

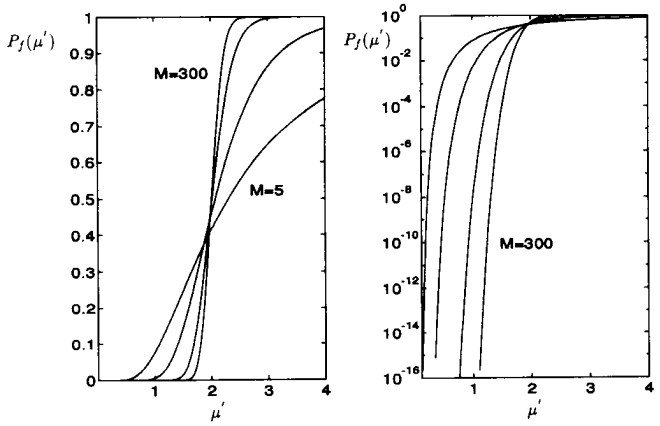

Fig. 1. Failure probability P_f as a function of μ' for a Gaussian process with filter orders $M = 5, 20, 100, 300$.

squared L_2 norm of the weight-error vector $\epsilon(k)$, which is obtained from the homogeneous part of (1), is given by

$$\|\epsilon(k+1)\|_2^2 = \|\epsilon(k)\|_2^2 \big(1 + \kappa_{\epsilon u}(k)\mu \|\mathbf{u}(k)\|_2^2 [\mu \|\mathbf{u}(k)\|_2^2 - 2]\big).$$

The term $\kappa_{\epsilon u}(k) = \frac{[\epsilon^T(k)\mathbf{u}(k)]^2}{\|\epsilon(k)\|_2^2 \|\mathbf{u}(k)\|_2^2}$ lies between zero and one. Thus, for $0 < \mu\|\mathbf{u}(k)\|_2^2 < 2$, the update equation is a contraction mapping of the weight-error vector. Consequently, the algorithm fails at time k if $\mu\|\mathbf{u}(k)\|_2^2 > 2$, as long as $\kappa_{\epsilon u}(k) \neq 0$. Thus, the algorithm becomes unstable at time instant k if $\|\mathbf{u}(k)\|_2^2 > \frac{2}{\mu}$.

Depending on the filter structure, the occurrence of this instability can affect the output of the filter differently. In a linear combiner, where at every time instant independent vectors $\mathbf{u}(k)$ are applied, it is just a short burst, and it is very likely that this does not repeat at future time instants. In a transversal filter, however, the norm $\|\mathbf{u}(k)\|_2^2$ may not change very quickly, especially for long filters. If one sample $u(k)$ is very high, the norm may exceed the bound of 2μ for about M steps, and the bursting becomes rather perceptible. A common stability limit for the mean square approach is $\mu < 2/(\sigma_u^2 M)$ (see [2]). In order to obtain more convenient expressions, a normalized step size $\mu' = \mu M \sigma_u^2$ will be used throughout the correspondence. Since the algorithm is not necessarily contracting for $\|\mathbf{u}(k)\|_2^2 > \frac{2M\sigma_u^2}{\mu'}$, the probability P_f of algorithm failure at time instant k can be defined as a function of the normalized step size μ'

$$P_f(\mu') = \Pr\left(\|\mathbf{u}(k)\|_2^2 \geq \frac{2M\sigma_u^2}{\mu'}\right). \qquad (2)$$

Since the squared L_2 norm is never negative, a very rough approximation can be given using the Tchebycheff inequality (see [3]): $P_f(\mu') \leq \frac{\mu'}{2}$.

Since $P_f \leq 1$ even for $\mu' \geq 2$, the above inequality only provides a bound that is good for small step sizes. If, however, exact equalities are desired, the statistics of the input process have to be known. Since Gaussian statistics are very frequently used in analyzing the LMS algorithm, P_f is first discussed for these statistics in the next section. In the third section, this approach will be extended to the larger class of spherically invariant random processes (SIRP) since they closely resemble speech sequences. Examples for the special SIRP's with K_0 and Laplace density will be presented.

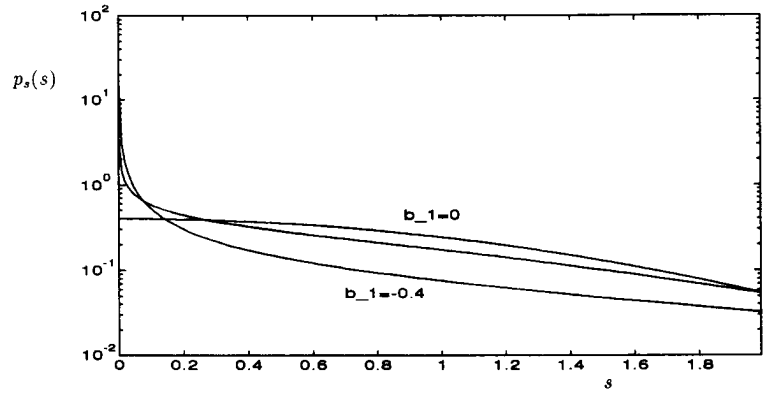

Fig. 2. Pdf of G_{01}^{10} for $b_1 = 0, -0.2$, and -0.4.

II. Gaussian Input Sequence

If it is assumed that the input sequence is an uncorrelated Gaussian random process with unit variance ($\sigma_u^2 = 1$), the statistics of M random variables are circularly symmetrical, and their joint density depends only on the distance from the origin r, i.e., the probability P_f from (2) can be written in terms of a squared radius $s = r^2$ (see [3, pp. 133])

$$P_f(\mu') = \Pr\left(s > \frac{2M}{\mu'}\right).$$

The radial density of an M-dimensional Gaussian process is given by

$$p_s(s) = \frac{s^{\frac{M}{2}-1}}{2^{\frac{M}{2}}\Gamma(\frac{M}{2})} e^{-s/2} U(s)$$

where $U(s)$ is the unit step function. Thus, the desired probability can be calculated

$$\begin{aligned} P_f(\mu') &= \frac{1}{2^{\frac{M}{2}}\Gamma(\frac{M}{2})} \int_{\frac{2M}{\mu'}}^{\infty} s^{\frac{M}{2}-1} e^{-s/2}\, ds \\ &= \frac{\Gamma\left(\frac{M}{2}, \frac{M}{\mu'}\right)}{\Gamma(\frac{M}{2})}. \end{aligned} \tag{3}$$

The function $\Gamma(a, x)$ is known as the incomplete gamma function (see [4, 8.350]) and can be computed using Matlab. Fig. 1 depicts the failure probability for various choices of the filter order M as a function of the step size μ'. It can be seen that for $\mu' \leq 1$, P_f is very small, whereas it increases drastically for $\mu' > 2$. The semilog plot on the right side emphasizes the behavior for small μ'. This can be calculated explicitly. For $\frac{M}{2}$ being an integer value, the incomplete normalized gamma function can be given as a series (see [4, 8.352.2])

$$\frac{\Gamma\left(\frac{M}{2}, \frac{M}{\mu'}\right)}{\Gamma\left(\frac{M}{2}\right)} = \sum_{k=0}^{\frac{M}{2}-1} \frac{\left(\frac{M}{\mu'}\right)^k}{k!} e^{-\frac{M}{\mu'}}. \tag{4}$$

As long as $\frac{M}{\mu'} \gg \frac{M}{2}$, the series can be approximated by its largest component

$$\frac{\Gamma\left(\frac{M}{2}, \frac{M}{\mu'}\right)}{\Gamma\left(\frac{M}{2}\right)} \approx \frac{\left(\frac{M}{\mu'}\right)^{\frac{M}{2}-1}}{\left(\frac{M}{2}-1\right)!} e^{-\frac{M}{\mu'}}.$$

Thus, for small step sizes, the probability of failure is given explicitly. If the filter order M is fixed, then the failure probability behaves essentially like $e^{-\frac{M}{\mu'}}$ or in the semilog plot like $-\frac{M}{\mu'}$. For large step sizes $\mu' > M$, the behavior can also easily be deduced from (4)

$$P_f \approx e^{-\frac{M}{\mu'}}.$$

Although this is not of practical importance, it gives an indication of the behavior of $P_f(\mu')$ for large μ'.

Since the stability limit is stated as $\mu' = 2$. P_f for this step size is of special interest and needs further calculation. As can be observed from Fig. 1, the value for $\mu' = 2$ tends to 0.5 as M increases. We prove in the Appendix that indeed P_f tends to 0.5 for $\mu' = 2$ as M goes to infinity; in other words, the probability for short-time instability at this limit equals at most 0.5 even for large filter orders. The value for maximal convergence speed $\mu' = 1$ has a very low probability of failure, and this probability vanishes as M tends to infinity. However, if statistics other than Gaussian are applied, the situation can become much worse, and the occurrence of bursting becomes more likely.

III. SIRP Input Sequence

Spherically invariant processes are defined as being those for which all joint density functions are only dependent on the radius and not on any angles (see also [5] and [6]). A suitable description of these pdf's can be given in terms of Meijer's G-functions [4]. They are defined by a Mellin-Barnes integral

$$\begin{aligned} &G_{pq}^{mn}\left(z\middle|\begin{matrix}a_p\\ b_q\end{matrix}\right) \\ &= \frac{1}{2\pi j}\int_C z^s \frac{\prod_{l=1}^{m}\Gamma(b_l - s)\prod_{l=1}^{n}\Gamma(1 - a_l + s)}{\prod_{l=m+1}^{q}\Gamma(1 - b_l + s)\prod_{l=n+1}^{p}\Gamma(a_l - s)}\, ds \end{aligned} \tag{5}$$

where the parameters are divided into two subsets a_p and b_q. In the case where only simple poles appear, the integral can be expressed as a sum of hypergeometric functions and can be computed numerically. Applications of G-functions to LMS, NLMS, and delayed update LMS (DLMS) algorithms can be found in [7] and [8]. We consider only two types of G-functions: G_{01}^{10} and G_{02}^{20} with one and two parameters, respectively. Both are known very well and can be expressed explicitly as

$$\begin{aligned} G_{01}^{10}(x|_{b_1}) &= x^{b_1} e^{-x} \\ G_{02}^{20}(x|_{b_1\, b_2}) &= 2x^{\frac{b_1+b_2}{2}} K_{b_1-b_2}\left(2\sqrt{x}\right) \end{aligned}$$

where $K_n(x)$ is the modified Bessel or McDonald function of the second kind for order n (see [4, 8.432]). Since G-functions have the advantage that almost every linear operation is simply a change in the parameter sets, this family of functions is very suitable for calculating the failure rates. As shown in [5] and [6], pdf's of bandlimited speech signals can be described by chosing the two parameters b_1 and b_2. A suitable form for pdf's is

$$p_x(x) = A G_{pq}^{mn}\left(\lambda x^2\middle|\begin{matrix}a_p\\ b_q\end{matrix}\right)$$

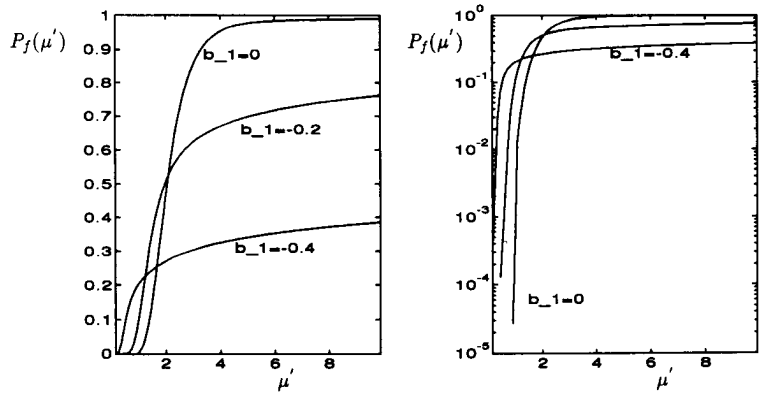

Fig. 3. Failure probability P_f as a function of μ' for G_{01}^{10}densities with $b_1 = 0, -0.2, -0.4$, and filter order $M = 20$.

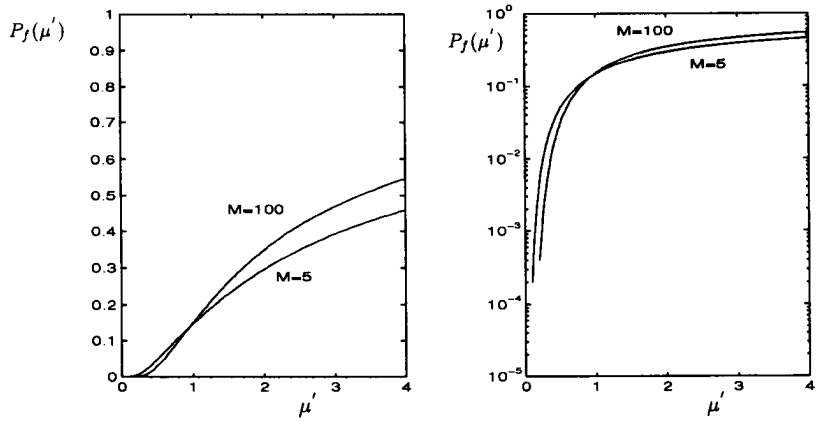

Fig. 4. Failure probability P_f as a function of μ' for a Laplace process with filter orders $M = 5, 100$.

where A and λ are normalization factors in order for the above to be a valid pdf with $\sigma_x = 1$ and can be calculated from the parameters a_p, b_q as

$$\lambda = (-1)^{n+m-q} \frac{\prod_{l=1}^{q} \left(b_l + \frac{1}{2}\right)}{\prod_{l=1}^{p} \left(a_l + \frac{1}{2}\right)},$$

$$A = \lambda^{\frac{1}{2}} \frac{\prod_{l=m+1}^{q} \Gamma\left(\frac{1}{2} - b_l\right) \prod_{l=n+1}^{p} \Gamma\left(\frac{1}{2} + a_l\right)}{\prod_{l=1}^{m} \Gamma\left(\frac{1}{2} + b_l\right) \prod_{l=1}^{n} \Gamma\left(\frac{1}{2} - a_l\right)}.$$

The radial density function $p_{r_M}(r)$ describes the density of the L_2-norm of the $(1 \times M)$ vector $\mathbf{u}(k)$ and can be given explicitly as

$$p_{r_M}(r) = 2A \frac{\sqrt{\pi}}{\Gamma\left(\frac{M}{2}\right)} G_{p+1\,q+1}^{m+1\,n} \left(\lambda r^2 \middle| \begin{matrix} a_p, 0 \\ \frac{M-1}{2}, b_q \end{matrix} \right).$$

Hence, the probability of failure can be written as

$$\begin{aligned} P_f(\mu') &= 1 - \int_0^{\sqrt{\frac{2M}{\mu'}}} p_{r_M}(r)\, dr \\ &= 1 - A \frac{\sqrt{\pi}}{\Gamma\left(\frac{M}{2}\right)} \int_0^{\frac{2M}{\mu'}} s^{-\frac{1}{2}} G_{p+1\,q+1}^{m+1\,n} \left(\lambda s \middle| \begin{matrix} a_p, 0 \\ \frac{M-1}{2}, b_q \end{matrix} \right) ds \\ &= 1 - A \frac{\sqrt{\pi\lambda}}{\Gamma\left(\frac{M}{2}\right)} \int_0^{\frac{2M}{\mu'}} G_{p+1\,q+1}^{m+1\,n} \left(\lambda s \middle| \begin{matrix} a_p - \frac{1}{2}, -\frac{1}{2} \\ \frac{M}{2} - 1, b_q - \frac{1}{2} \end{matrix} \right) ds \\ &= 1 - A \frac{\sqrt{\pi\lambda}}{\Gamma\left(\frac{M}{2}\right)} \frac{2M}{\mu'} G_{p+2\,q+2}^{m+1\,n+1} \left(\lambda \frac{2M}{\mu'} \middle| \begin{matrix} 0, a_p - \frac{1}{2}, -\frac{1}{2} \\ \frac{M}{2} - 1, b_q - \frac{1}{2}, -1 \end{matrix} \right). \end{aligned} \tag{6}$$

We used typical properties of the G-functions as described in [5] and [6] to derive the last two equations. For the two specific G-functions under consideration, we finally get

$$P_f(\mu') = 1 - A \frac{\sqrt{\pi\lambda}}{\Gamma\left(\frac{M}{2}\right)} \frac{2M}{\mu'} G_{2\,3}^{2\,1} \left(\frac{2M\lambda}{\mu'} \middle| \begin{matrix} 0, -\frac{1}{2} \\ \frac{M}{2} - 1, b_1 - \frac{1}{2}, -1 \end{matrix} \right)$$

for the G_{01}^{10} function and for the G_{02}^{20}

$$P_f(\mu') = 1 - A \frac{\sqrt{\pi\lambda}}{\Gamma\left(\frac{M}{2}\right)} \frac{2M}{\mu'} G_{2\,4}^{3\,1} \left(\frac{2M\lambda}{\mu'} \middle| \begin{matrix} 0, -\frac{1}{2} \\ \frac{M}{2} - 1, b_1 - \frac{1}{2}, b_2 - \frac{1}{2}, -1 \end{matrix} \right).$$

Both expressions look very similar; however, depending on the parameters b_1 and b_2, their behavior can be quite different.

Next, we shall present some quantitative examples. Note that the G_{01}^{10} function is completely parametrized by only one variable b_1. A spherically invariant process can be obtained only for $-0.5 < b_1 \leq 0$ (see [5]). Fig. 2 depicts the pdf's for three different values of b_1. For $b_1 = 0$, the G-function coincides with the Gaussian density, whereas for $b_1 < 0$, the density functions become singular at zero. This range of b_1 is of particular interest since measured speech pdf's show a clear singularity at zero as well. Fig. 3 depicts the failure probability versus the step size μ' for several choices of b_1. As the figure clearly demonstrates, the more negative the parameter is, the higher the failure probability for reasonable step sizes μ'.

Fig. 4 presents the failure behavior for a Laplace density. This pdf can be described by a G_{02}^{20} function with parameters $b_1 = 0$, and $b_2 = 0.5$. Unlike the K_0 density (which is obtained with $b_1 = b_2 = 0$), the pdf does not become singular at zero. A more extreme behavior can thus be expected from the K_0 density. In both cases, the failure rates do not vary much with the filter order. The step sizes for fastest convergence have been calculated in [7].

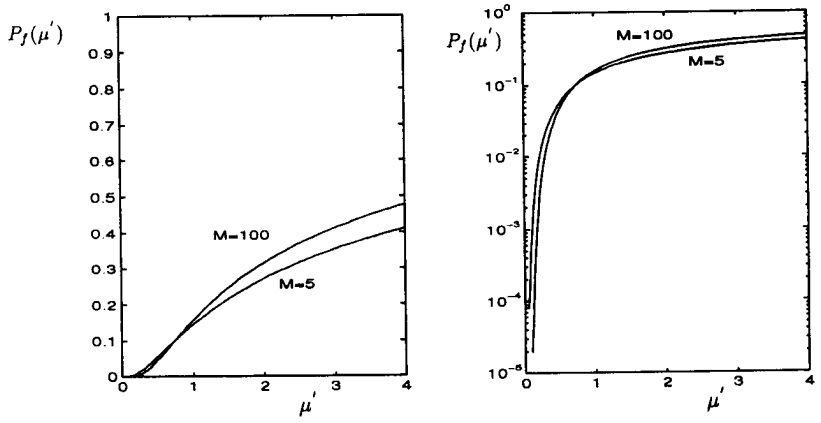

Fig. 5. Failure probability P_f as a function of μ' for a K_0 process with filter orders $M = 5$, 100.

They are $\mu' \approx 1/2$ and $1/3$ for the Laplacian and the K_0 density, respectively. For these values, the failure rates are approximately 2 and 5%, respectively.

However, even for small step sizes, the probability for bursting has increased drastically for all three SIRP sequences discussed. Thus, when using speech signals, the designer of an adaptive filter has to choose a tradeoff between convergence speed and bursting rate. The given figures show how various input statistics alter the bursting rates. If compared with the Gaussian input, the figures enable us to find step sizes for a given statistic so that the same failure rates are obtained. If bursting is to be prevented completely, a time-varying step size $\mu(k) = \alpha/\|\mathbf{u}(k)\|_2^2$ is preferred. Because of its contraction mapping property for $0 < \alpha < 2$, the algorithm can then guarantee convergence at every time step k for every particular realization, and thus, a burst-free behavior results.

APPENDIX

We wish to prove that

$$\lim_{M\to\infty} \frac{\Gamma\left(\frac{M}{2}, \frac{M}{2}\right)}{\Gamma\left(\frac{M}{2}\right)} = \frac{1}{2}. \tag{7}$$

The filter order M is assumed to be an even number. Thus, substituting $\mu' = 2$ in (4) yields

$$\frac{\Gamma\left(\frac{M}{2}, \frac{M}{2}\right)}{\Gamma\left(\frac{M}{2}\right)} = \sum_{k=0}^{\frac{M}{2}-1} \frac{\left(\frac{M}{2}\right)^k}{k!} e^{-\frac{M}{2}}. \tag{8}$$

Recall that Bernoulli trials can be written as Poisson trials as the number of trials n tends to infinity, the probability p of a single event goes to zero, but the product np goes to a fixed number a (Poisson theorem [3])

$$\binom{n}{k} p^k (1-p)^{n-k} \underset{n\to\infty}{\longrightarrow} \frac{a^k}{k!} e^{-a} \tag{9}$$

and for $np \gg 1, n \gg 1$ [3]

$$\sum_{k=0}^{k_2} \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{1}{2} + \operatorname{erf}\left(\frac{k_2 - np}{\sqrt{np(1-p)}}\right).$$

Now, substituting $np = \frac{M}{2}$ and $k_2 = \frac{M}{2} - 1$ in the above, we get

$$\lim_{M\to\infty} \frac{\Gamma\left(\frac{M}{2}, \frac{M}{2}\right)}{\Gamma\left(\frac{M}{2}\right)} = \lim_{M\to\infty} \sum_{k=0}^{\frac{M}{2}-1} \frac{\left(\frac{M}{2}\right)^k}{k!} e^{-\frac{M}{2}}$$

$$= \lim_{M\to\infty} \frac{1}{2} + \operatorname{erf}\left(\frac{-1}{\sqrt{\frac{M}{2}}}\right) = \frac{1}{2}.$$

REFERENCES

[1] A. H. Sayed and M. Rupp, "On the robustness, convergence, and minimax performance of instantaneous-gradient adaptive filters," in *Proc. Asilomar Conf.*, Oct. 1994.

[2] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[3] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw Hill, 1987.

[4] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*,. New York: Academic, 1980.

[5] H. Brehm, "Description of spherically invariant random processes by means of G-functions," *Springer Lecture Notes in Mathematics*, vol. 969, pp. 39–73, 1982.

[6] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Processing*, vol. 12, no. 2, pp. 119–141, 1987.

[7] M. Rupp, "The behavior of LMS and NLMS algorithms in the presence of spherically invariant processes," *IEEE Trans. Signal Processing*, vol. 41, no. 3, pp. 1149–1160, Mar. 1993.

[8] M. Rupp and R. Frenzel, "The behavior of LMS and NLMS algorithms with delayed coefficient update in the presence of spherically invariant processes," *IEEE Trans. Signal Processing*, vol. 42, no. 3, pp. 668–672, Mar. 1994.