Excursions of Adaptive Algorithms via the Poisson Clumping Heuristic

William A. Sethares, *Member, IEEE*, and James A. Bucklew, *Member, IEEE*

Abstract—**This paper details the application of the Poisson clumping heuristic (PCH) to the least mean square adaptive algorithm and its signed variants. Under certain conditions on the input and disturbance statistics, the parameter estimate errors form a Markov process. The PCH asserts that large excursions of the parameter estimates occur in clumps, and that these clumps are distributed in a Poisson manner with parameter λ_b. Expressions are derived for each of the four algorithms in the scalar case, which allows calculation of λ_b in a relatively straightforward manner. These values are compared to simulations of the algorithms. Given that the results are asymptotic in b, the close agreement between simulated and theoretical values is striking, even for very modest b. The four algorithms are then compared in terms of λ_b. Some observations are made regarding the relative performance of the four variants, and no single LMS variant always outperforms the others. Suggestions are made as to how this technique might be applied in the vector case, and a crucial "monotonicity property" is verified.**

I. Introduction

In certain adaptive filtering tasks, the numerical complexity of algorithms such as recursive maximum likelihood (RML) or normalized least mean square (NLMS) is too high for a given throughput. This encourages the use of the least mean square (LMS) algorithm [1] or the simpler "signed" variants in which the n multiplications per iteration of LMS are replaced by n compare operations [2]. What is sacrificed (or gained) in terms of performance for this simplification?

This paper addresses the issue of performance of LMS and its three signed variants in terms of the Poisson clumping heuristic (PCH) [3], which gives a measure of the probability of large parameter excursions. Given a stationary input, the parameter estimate errors are an asymptotically stationary random process. Exact expressions for these asymptotic distributions are forbidding, but the PCH provides a way of gathering information about the steady state behavior in a relatively straightforward way. For a given input and disturbance distribution, the behavior of the four algorithm forms can be compared.

Large parameter errors occur due to large inputs or disturbances from the "tail" of the distributions, or due to a long string of (unlikely but inevitable) malicious events. In either case, these events are rare, and may be investigated for certain models using large deviations theory [4], [22]. The large deviation methodology can give more information (e.g., when the algorithms are convergent), though we argue that the present method appears more tractable for the study of steady state behavior. In [5], the probability density function of the LMS adaptive weights is derived under the assumption that the input process is zero mean independent identically distributed (i.i.d.) Gaussian. The present technique makes no assumptions about the input process other than that it is i.i.d. and its distribution function is known. The characteristic function of (complex) scalar LMS weights is investigated in [6], again under the Gaussian assumption, and the adaptive weights are shown to act in an essentially Gaussian fashion.

Rather than investigating this stationary distribution itself, we suppose that a measure of performance for an adaptive algorithm is stated in terms of a desired bound on the maximum parameter error b. This is apparent when one is more interested in a catastrophic error than the long term average behavior. For instance, systems which involve a feedback of the error signal back into the input of the adaptive element (e.g., [19], [20], [22]) may become unstable if the parameter error becomes too large. Mean time to instability can be predicted via the likelihood of large parameter excursions while a mean square error criterion ignores these destabilizing events since they are of low probability.

The structure of the algorithm (when driven by i.i.d. inputs) implies that the parameter estimate errors form a Markov process. This suggests the use of the PCH as a tool for understanding the behavior of the algorithm, in terms of the possibility of attaining the error b. The PCH asserts that "hits" of the error b, for b large, will tend to occur in "clumps," and that these clumps will be distributed in a Poisson manner. For certain input processes, the Poisson parameter λ_b can be calculated analytically as a function of the stepsize μ and the error b. For other inputs, numerical techniques are used to derive expressions for the hitting rate, given particular distributions on the input and disturbance, and assuming that the two processes are independent. This allows a comparison of the performance of the various algorithms, where performance is gauged by the likelihood of such large parameter excursions. Interestingly, no one of the algorithms "outperforms" the others in all cases. We have chosen to con-

Manuscript received March 20, 1990; revised February 26, 1991.

The authors are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706.

IEEE Log Number 9107650.

centrate primarily on the scalar case, since our purpose is to compare the four variants of LMS and to demonstrate the usefulness of the PCH as a tool for measuring the performance of adaptive algorithms.

Section II recalls four popular adaptive algorithms and then provides a brief introduction to the PCH by formally defining a mosaic process and the related notions of clumps and hits. Central to any application of the PCH is that the process must behave in an essentially monotonic fashion for large excursion levels. We argue that if the input process to the algorithm fulfills a persistence of excitation condition then the desired monotonicity holds.

Section III develops detailed expressions for the hitting rate λ_b of the error bound b for each of the four algorithms. This involves a formula for the stationary distributions of the parameter estimates in terms of the input distribution and the distribution of the disturbance, as well as the actual calculation of λ_b. The section concludes with an examination of the possibility of extending the results to the vector case. We derive a result for inputs and disturbances possessing spherical symmetry in their (multidimensional) probability distributions.

Section IV derives working expressions for the calculation of λ_b where the disturbance is an i.i.d. collection of Bernoulli random variables. This is our test case, for which we compare all four algorithms when excited by three different input distributions (Gaussian, double sided Rayleigh, and Laplacian). Two sets of experiments are reported. The first set verifies that the calculations of λ_b agree with straightforward simulations of the algorithms. The second set compares λ_b for the four algorithms. No one of the algorithms outperforms the others in all cases. The last section presents conclusions and areas for further investigation.

II. Problem Setup

A. Discussion of the Algorithms

The LMS algorithm is well known as a recursive gradient algorithm that tends to adjust a set of filter coefficients so as to minimize the mean squared error between an input process $\{x_k\}$ and a given signal. If w_k represents the error between the current set of estimates and the optimal weights, then the adaptive update on w_k is given by

$$w_{k+1} = w_k - \mu x_k(x_k w_k + d_k) \tag{1}$$

where μ is a step size coefficient typically chosen to be much smaller than the expected value of x_k^2, and d_k represents the essentially unpredictable (or disturbance) portion of the given signal.

Several variants of the LMS are in common use. If it is desired to minimize the mean absolute error (rather than mean square error), then the algorithm becomes

$$w_{k+1} = w_k - \mu x_k \operatorname{sgn}(w_k x_k + d_k) \tag{2}$$

where sgn $(\cdot)$ represents the signum function. This was first proposed in [7], and has been analyzed more recently in [8], [9]. One justification for the use of this signed error form is that it is numerically simpler to implement than LMS, and claims have been made that it tends to reject disturbances better than LMS. Our results indicate that for certain distributions on the input and disturbance sequences these claims are justified, while for others they are not.

An alternative which preserves the numerical simplicity of (2) is to take the signum function of the input data stream. This is known as the signed regressor (or clipped) LMS algorithm, and was proposed in [10]. This algorithm has been analyzed more recently in [11]–[13]:

$$w_{k+1} = w_k - \mu \operatorname{sgn}(x_k)(w_k x_k + d_k). \tag{3}$$

Finally, the sign-sign version of LMS applies the signum function to both the input data and the error signal:

$$w_{k+1} = w_k - \mu \operatorname{sgn}(x_k) \operatorname{sgn}(w_k x_k + d_k). \tag{4}$$

First proposed in [14], this version has found extensive use in adaptive equalization and in adaptive pulse code modulation [15].

B. Introduction to the Poisson Clumping Heuristic

The Poisson clumping heuristic (PCH) is applicable to a large class of stationary probability models which have a certain monotonicity property for large excursions. The probability that such a process will achieve a large value is asymptotically small, and is distributed in a Poisson fashion with parameter λ_b where b is the large value of interest. Moreover, these hits of the bound b tend to occur in clumps rather than in isolation. To formalize these notions, we first define a *mosaic process* which is a formalization of the idea of "throwing sets down at random." Let $\mathfrak{B}$ be a collection of sets. Think of each $B \in \mathfrak{B}$ as being a "small" set located near the origin. Let C be a probability distribution over $\mathfrak{B}$. Given a set B and a $t \in \mathfrak{R}$, define the translation of the set B as $t + B = \{t + x: x \in B\}$. Mark out points $\{y_i\}$ on the real axis according to the events of a Poisson point process of rate λ. Define the mosaic process

$$S = \bigcup_i (y_i + B_i)$$

where the $B_i \in \mathfrak{B}$ are chosen via the distribution C. We call the y_i, centers and the $y_i + B_i$, clumps. The clump endpoints of $y_i + B_i$ are $y_i + \max\{x: x \in B_i\}$ and $y_i + \min\{x: x \in B_i\}$. S is the union of i.i.d. shaped clumps with Poisson random centers.

The probability that a stationary stochastic process (in our case, the adaptive algorithm) makes a large excursion, can be related to the centers and clumps of the mosaic process. Consider some stationary real valued random process $\{w_k\}$. Suppose we are interested in the distribution of

$$M_N \equiv \max_{k \in \{0,1,\cdots,N\}} w_k$$

for large N. Defining the random set $S_b \equiv \{k: w_k \geq b\}$, gives the fundamental relation

$$P(M_N < b) = P(S_b \cap \{0, 1, \cdots, N\} \text{ is empty}).$$

For b large, S_b is a sparse stationary random set. It is stationary and random since $\{w_k\}$ is assumed to be a stationary random process, and it is sparse in the sense that if b is large, there are very few indices contained in S_b. The PCH assumes that S_b is a mosaic process with some clump rate λ_b and some clump distribution C_b.

The clump rate λ_b is derived as follows:

$$\begin{aligned}\lambda_b &= \lim_{\delta \to 0^+} \frac{P(\text{some clump endpoint lies in } (0, \delta))}{\delta} \\ &= \lim_{\delta \to 0^+} \frac{P(\text{the clump containing 0 ends in } (0, \delta),\ 0 \in S_b)}{\delta} \\ &= P(0 \in S_b) \lim_{\delta \to 0^+} \frac{P(\text{the clump containing 0 ends in } (0, \delta) | 0 \in S_b)}{\delta}.\end{aligned}$$

In discrete time, replace "δ" with "1" to obtain

$$\begin{aligned}\lambda_b &= P(0 \in S_b) P(\text{clump ends at } 0 | 0 \in S_b) \\ &= P(w_0 \geq b) P(\text{clump ends at } 0 | 0 \in S_b) \\ &= P(w_0 \geq b) P(w_1 \leq b, w_2 \leq b, \cdots, \\ &\qquad w_n \leq b | w_0 \geq b) \qquad (5)\end{aligned}$$

for n "large" compared to a clump diameter. From the fact that the clumps are arriving as a Poisson process

$$\begin{aligned}&P(\max_{k \in \{0, 1, \cdots, N\}} w_k \leq b) \\ &\quad = P(M_N < b) \\ &\quad = P(S_b \cap \{0, 1, \cdots, N\} \text{ empty}) \\ &\quad \approx P(\text{no centers have hit in interval } [0, N]) \\ &\quad = \exp(-N\lambda_b). \qquad (6)\end{aligned}$$

Note also that the mean time to an escape over the level b of $\{w_k\}$ is the interarrival time between clumps and it is given by $1/\lambda_b$. Both of these observations are of interest in trying to characterize the large excursion behavior of adaptive algorithms. Refer to Section IV-B and IV-C for a more concrete presentation.

C. The Monotonicity Property

Our application of the PCH requires that the parameter estimate errors decrease in a monotonic fashion for large excursion levels. While this is not always true for the four algorithms above, it is the generic behavior. We now give a formal definition of the required property:

Monotonicity Property:

$$\lim_{b \to \infty} \frac{P(w_1 \leq b, w_2 \leq b, \cdots, w_n \leq b | w_0 \geq b)}{P(w_1 \leq b | w_0 \geq b)} = 1$$

$$\forall\ n = 1, 2, \cdots.$$

Think of it this way. Large excursions are the result of large input values or large disturbance values. These tend to be fairly rare events, and the possibility of multiple rare events occurring near each other in time is a "rare squared" event. Ignoring such extra-rate events justifies the approximation of $P(w_1 \leq b, w_2 \leq b, \cdots, w_n \leq b | w_0 \geq b)$ by $P(w_{k+1} \leq b | w_k \geq b)$. Moreover, as the bound increases, the approximation gets better. The simplest way to guarantee this rarity of successive large events on which the monotonicity property rests is to assume that the inputs are i.i.d.

To be more concrete, rewrite the LMS update equation (1) as

$$w_{k+1} = (1 - \mu x_k^2) w_k - \mu x_k d_k.$$

Given that the algorithm is operating in its normal region (with w_k small), w_{k+1} can attain a large value in one of two ways. Either the input x_k can be very large (causing $(1 - \mu x_k^2)$ to be an expansion rather than a contraction), or the disturbance d_k can be large. Both occurrences are rare. On the other hand, when w_k is large, only another rare event can cause it to increase in magnitude, since $(1 - \mu x_k^2)$ is strongly contractive for normal sized x_k. Similar arguments apply to each of the variants of LMS. The exception to this rule is when the input fails to be persistently exciting [16] for the particular algorithm. This is explored in [2], where excitation conditions are derived for three of the four algorithms above.

III. Development

A. Scalar System Case

This section develops general expressions for the calculation of the Poisson parameter λ_b for each of the algorithms of interest. Under the assumption that the inputs and disturbances are independent identically distributed sequences, the coefficient error processes are Markovian.

The PCH requires that we obtain the following probability:

$$\begin{aligned}p &\equiv P(\text{clump ends at time } 0 | 0 \text{ is in clump}) \\ &= P(w_1 \leq b, w_2 \leq b, \cdots, w_n \leq b | w_0 \geq b) \qquad (7)\end{aligned}$$

where n is small compared to the average interclump distance.

All of the stochastic models we consider approximate the "monotonicity property" for small step sizes, that is, the magnitude of the parameter error tends to decrease monotonically once it has reached a large level. Hence,

we make the approximation

$$p \approx P(w_1 \le b | w_0 \ge b). \tag{8}$$

Let $F_a(\cdot)(f_a(\cdot))$ denote the distribution (density, if it exists) function of the random variable a. We let $F_w(\cdot)(f_w(\cdot))$ denote the stationary distribution (density, if it exists) of the coefficient error Markov process. Similarly define $F_{w_1|w_0}(y|z) \equiv P(w_1 \le y | w_0 = z)$, the conditional probability distribution of w_1 given that $w_0 = z$. The associated conditional density (if it exists) is denoted $f_{w_1|w_0}(y|z)$. Assume that the processes have been running for a long time and that all marginal distributions have settled into the stationary distributions. Then

$$\begin{aligned} p &\approx \int_b^\infty \frac{P(w_1 \le b | w_0 = z)\, dF_{w_0}(z)}{[1 - F_{w_0}(b)]} \\ &= \int_b^\infty \frac{P(w_1 \le b | w_0 = z)\, dF_w(z)}{[1 - F_w(b)]} \\ &= \int_b^\infty \frac{F_{w_1|w_0}(b|z)\, dF_w(z)}{[1 - F_w(b)]}. \end{aligned} \tag{9}$$

The clump rate λ_b is given by

$$\begin{aligned} \lambda_b &= P(\text{clump ends at time } 0 | 0 \text{ is in clump}) \\ &\quad \cdot P(0 \text{ is in clump}) \\ &= p \cdot P(w_0 \ge b) \\ &= p \cdot (1 - F_w(b)) \\ &\approx \int_b^\infty F_{w_1|w_0}(b|z)\, dF_w(z). \end{aligned} \tag{10}$$

Hence, in order to find expressions for the clump rate λ_b, (10) requires expressions for the transition structure $F_{w_1|w_0}(\cdot|\cdot)$ and for the stationary distribution $F_w(\cdot)$. These two requirements are interrelated. To find the stationary distribution, consider

$$F_{w_{k+1}}(y) = \int F_{w_1|w_0}(y|z)\, dF_{w_k}(z). \tag{11}$$

Taking the limit as k approaches ∞ gives

$$F_w(y) = \int F_{w_1|w_0}(y|z)\, dF_w(z). \tag{12}$$

If densities exist, then

$$f_w(y) = \int f_{w_1|w_0}(y|z) f_w(z)\, dz. \tag{13}$$

The conditional distribution corresponding to the coefficient error models (1) are given by

$$F^{\text{lms}}_{w_1|w_0}(y|z) = P(z - \mu x_0(zx_0 + d_0) \le y) \tag{14}$$

$$F^{\text{se}}_{w_1|w_0}(y|z) = P(z - \mu x_0 \operatorname{sgn}(zx_0 + d_0) \le y) \tag{15}$$

$$F^{\text{sr}}_{w_1|w_0}(y|z) = P(z - \mu \operatorname{sgn}(x_0)(zx_0 + d_0) \le y) \tag{16}$$

$$F^{\text{ss}}_{w_1|w_0}(y|z) = P(z - \mu \operatorname{sgn}(x_0) \operatorname{sgn}(zx_0 + d_0) \le y) \tag{17}$$

where the $\{x_k\}$ are the i.i.d. input sequence and the $\{d_k\}$ are the i.i.d. disturbance sequence ($\{x_k\}$ and $\{d_k\}$ are independent of each other). The superscripts lms, se, sr, and ss indicate which algorithm the expression applies to, and the notation $F_{w_1|w_0}(\cdot|\cdot)$ without any further superscripts, means that the expression is valid for all four algorithms. For simplicity, suppose that the distribution function $F_{x_0}(\cdot)$ is absolutely continuous (does not have point masses) and hence has an associated density. Note also that $F_{w_1|w_0}(\cdot|\cdot)$ depends entirely on the distributions of x_0 and d_0.

We see from expressions (10), (12), and (13) that $F_{w_1|w_0}(\cdot|\cdot)$ is the crucial quantity to compute for each of the models. Expressions (12) and (13) imply an iterative method to find the stationary distributions (or densities) of the coefficient error random processes and (10) gives the explicit expression for the clump rate.

B. Classical LMS

A more detailed expression for the conditional distribution $F^{\text{lms}}_{w_1|w_0}(y|z)$ is from (14):

$$\begin{aligned} F^{\text{lms}}_{w_1|w_0}(y|z) &= \int P(z - \mu x^2 z - \mu x d_0 \le y)\, dF_{x_0}(x) \\ &= \int_0^\infty \left(1 - F_{d_0}\left(\frac{z - y - \mu x^2 z}{\mu x}\right)\right) dF_{x_0}(x) \\ &\quad + \int_{-\infty}^0 F_{d_0}\left(\frac{z - y - \mu x^2 z}{\mu x}\right) dF_{x_0}(x). \end{aligned} \tag{18}$$

C. Signed Error LMS

Let $1_A(x)$ denote the indicator function for the set A, i.e.,

$$1_A(x) = \begin{cases} 1 & \text{for } x \in A \\ 0 & \text{for } x \notin A. \end{cases} \tag{19}$$

For a real nonrandom parameter d:

$$\begin{aligned} &P(z - \mu x_0 \operatorname{sgn}(zx_0 + d) \le y) \\ &\quad = P\left(x_0 \operatorname{sgn}(zx_0 + d) \ge \frac{z - y}{\mu}\right) \\ &\quad = 1_{[0,\infty)}(z)\left[P\left(x_0 \ge \frac{z - y}{\mu}, x_0 \ge -\frac{d}{z}\right)\right. \\ &\qquad \left. + P\left(-x_0 \ge \frac{z - y}{\mu}, x_0 \le -\frac{d}{z}\right)\right] \\ &\qquad + 1_{(-\infty,0)}(z)\left[P\left(x_0 \ge \frac{z - y}{\mu}, x_0 \le -\frac{d}{z}\right)\right. \\ &\qquad \left. + P\left(-x_0 \ge \frac{z - y}{\mu}, x_0 \ge -\frac{d}{z}\right)\right]. \end{aligned} \tag{20}$$

Hence,

$$F^{se}_{w_1|w_0}(y|z) = P(z - \mu x_0 \operatorname{sgn}(zx_0 + d_0) \le y)$$
$$= \int P(z - \mu x_0 \operatorname{sgn}(zx_0 + d) \le y)\, dF_{d_0}(d) \tag{21}$$

where we may now substitute (20) into (21) to obtain a final expression.

Unlike standard LMS, the signed error algorithm cannot converge to zero even when there is no disturbance. Thus there always exists a nontrivial stationary distribution. In this case, a closed form analytic solution can be found for at least one example.

No Disturbance Case: Suppose $P(d_k = 0) = 1$ for all k. Then (20) and (21) imply

$$F^{se}_{w_1|w_0}(y|z) = 1_{[0,\infty)} P\left(|x_0| \ge \frac{z - y}{\mu}\right) + 1_{(-\infty,0)} P\left(|x_0| \le \frac{y - z}{\mu}\right). \tag{22}$$

Hence, if densities exist

$$f^{se}_{w_1|w_0}(y|z) = 1_{[0,\infty)} f_{|x_0|}\left(\frac{z - y}{\mu}\right)\frac{1}{\mu} + 1_{(-\infty,0)} f_{|x_0|}\left(\frac{y - z}{\mu}\right)\frac{1}{\mu} \tag{23}$$

which when substituted into (13) gives

$$f_w(y) = \int_0^\infty f_{|x_0|}\left(\frac{z - y}{\mu}\right)\frac{1}{\mu} f_w(z)\, dz + \int_{-\infty}^0 f_{|x_0|}\left(\frac{y - z}{\mu}\right)\frac{1}{\mu} f_w(z)\, dz. \tag{24}$$

Using (10) and (22), the expression for the clump rate is

$$\lambda_b = \int_b^\infty P\left(|x_0| \ge \frac{z - b}{\mu}\right) dF_w(z). \tag{25}$$

Example: Suppose that $|x_0|$ has the probability density $\exp(-x)$ for $x \ge 0$. It is then easy to check that a solution of (24) is obtained if $f_w(y) = (1/2\mu)\exp(-|y|/\mu)$. Substituting into (25) shows that $\lambda_b = 1/4 \exp(-b/\mu)$.

D. Signed Regressor LMS

In order to get more detailed expressions, consider the special case where $\{x_k\}$ and $\{d_k\}$ are symmetric sequences of random variables. If x_1 has a symmetric distribution about zero, then $|x_1|$ and $\operatorname{sgn}(x_1)$ are independent random variables. Note also that $d_1 \operatorname{sgn}(x_1)$ has the same distribution as d_1. Hence, from (16),

$$F^{sr}_{w_1|w_0}(y|z) = P(z - \mu z|x_0| + \mu d_0 \operatorname{sgn}(x_0) \le y)$$
$$= \int P(z - \mu z|x_0| + \mu d \le y)\, dF_{d_0}(d). \tag{26}$$

It is easy to check that

$$P(z(1 - \mu|x_0|) \le y - \mu d) = 1_{[0,\infty)}(z)\left(1 - F_{|x_0|}\left(\frac{1}{\mu}\left(1 - \frac{y - \mu d}{z}\right)\right)\right) + 1_{(-\infty,0)} F_{|x_0|}\left(\frac{1}{\mu}\left(1 - \frac{y - \mu d}{z}\right)\right). \tag{27}$$

Substituting this last expression into (26) and taking a derivative with respect to the y variable yields

$$f^{sr}_{w_1|w_0}(y|z) = \int f_{|x_0|}\left(\frac{1}{\mu}\left(1 - \frac{y - \mu d}{z}\right)\right)\frac{1}{|z|\mu}\, dF_{d_0}(d). \tag{28}$$

E. Sign-Sign LMS

The conditional distribution $F^{ss}_{w_1|w_0}(y|z)$ for the sign-sign LMS is from (17)

$$\begin{aligned} F^{ss}_{w_1|w_0}(y|z) &= P\left(\operatorname{sgn}(x_0)\operatorname{sgn}(zx_0 + d_0) \ge \frac{z - y}{\mu}\right) \\ &= \int_0^\infty P\left(\operatorname{sgn}(zx + d_0) \ge \frac{z - y}{\mu}\right) dF_{x_0}(x) \\ &\quad + \int_{-\infty}^0 P\left(\operatorname{sgn}(zx + d_0) \le \frac{y - z}{\mu}\right) \cdot dF_{x_0}(x) \\ &= \int_0^\infty 1_{(-1,1]}\left(\frac{z - y}{\mu}\right) P(\operatorname{sgn}(zx + d_0) \ge 0) \\ &\quad \cdot dF_{x_0}(x) + 1_{(-\infty,-1)}\left(\frac{z - y}{\mu}\right) \\ &\quad + \int_{-\infty}^0 1_{(-1,1)}\left(\frac{y - z}{\mu}\right) \\ &\quad \cdot P(\operatorname{sgn}(zx + d_0) < 0)\, dF_{x_0}(x). \end{aligned} \tag{29}$$

Note that

$$P(\operatorname{sgn}(zx + d_0) > 0) = P((zx + d_0) > 0) = P(d_0 > -zx) = 1 - F_{d_0}(-zx)$$

and similarly

$$P(\operatorname{sgn}(zx + d_0) < 0) = F_{d_0}(-zx).$$

Hence we obtain the expression

$$F^{ss}_{w_1|w_0}(y|z) = 1_{(-\infty,-1)}\left(\frac{z - y}{\mu}\right) + 1_{[-1,1]}\left(\frac{z - y}{\mu}\right) \cdot \left[\int_0^\infty (1 - F_{d_0}(-zx))\, dF_{x_0}(x) + \int_{-\infty}^0 F_{d_0}(-zx)\, dF_{x_0}(x)\right]. \tag{30}$$

F. Vector System Case

In this subsection, we indicate the possibility of extending these techniques beyond the scalar case by working out an example involving LMS for certain inputs and disturbances. The main drawback is that we assume the inputs are i.i.d. vectors. This assumption is not strictly correct in the vector case although it is a common simplifying assumption, see [17].

Consider the expression for the sum of squares coefficient error of the LMS algorithm:

$$\begin{aligned} w_{k+1}^T w_{k+1} &= w_k^T w_k - 2\mu x_k^T w_k w_k^T x_k \\ &\quad + \mu^2 x_k^T x_k (x_k^T w_k + d_k)^2 \\ &= w_k^T w_k + (\mu^2 x_k^T x_k - 2\mu)(x_k^T w_k + d_k)^2. \end{aligned} \tag{31}$$

We need to compute

$$\begin{aligned} &P(w_1^T w_1 \le y | w_0^T w_0 = z) \\ &\quad = \int P(w_1^T w_1 \le y | w_0^T w_0 = z,\ x_0^T x_0 = x)\, dF_{x_0^T x_0}(x). \end{aligned} \tag{32}$$

Suppose that x_0 has a spherically symmetric distribution in n-dimensional space. Then $x_0^T w_0 = \|x_0\| \cdot \|w_0\| \cos(\theta)$ where θ is the angle between the two vectors. Since x_0 and w_0 are independent and the angle of x_0 is uniformly distributed with respect to any fixed angle, then it is true that θ is also uniformly distributed on an interval of length 2π, say $[0, 2\pi]$. Then (32) becomes

$$\begin{aligned} &\int P(z + (\mu^2 x - 2\mu)(\sqrt{xz}\cos(\theta) + d_0)^2 \le y)\, dF_{x_0^T x_0}(x) \\ &\quad = \int_{2/\mu}^{\infty} P\left((\sqrt{xz}\cos(\theta) + d_0)^2 \le \frac{y-z}{\mu^2 x - 2\mu}\right) \\ &\qquad \cdot dF_{x_0^T x_0}(x) \\ &\qquad + \int_{-\infty}^{2/\mu} P\left((\sqrt{xz}\cos(\theta) + d_0)^2 \ge \frac{y-z}{\mu^2 x - 2\mu}\right) \\ &\qquad \cdot dF_{x_0^T x_0}(x) \\ &\quad = \int_{2/\mu}^{\infty} \int \left[F_{\cos(\theta)}\left(\sqrt{\frac{y-z}{xz(\mu^2 x - 2\mu)}} - \frac{d}{\sqrt{xz}}\right)\right. \\ &\qquad \left. - F_{\cos(\theta)}\left(-\sqrt{\frac{y-z}{xz(\mu^2 x - 2\mu)}} - \frac{d}{\sqrt{xz}}\right)\right] \\ &\qquad \cdot dF_{d_0}(d)\, dF_{x_0^T x_0}(x) + \int_{-\infty}^{2/\mu} \int \\ &\qquad \cdot \left[1 - \left[F_{\cos(\theta)}\left(\sqrt{\frac{y-z}{xz(\mu^2 x - 2\mu)}} - \frac{d}{\sqrt{xz}}\right)\right.\right. \\ &\qquad \left.\left. - F_{\cos(\theta)}\left(-\sqrt{\frac{y-z}{xz(\mu^2 x - 2\mu)}} - \frac{d}{\sqrt{xz}}\right)\right]\right] \\ &\qquad \cdot dF_{d_0}(d)\, dF_{x_0^T x_0}(x). \end{aligned} \tag{33}$$

Note that the distribution function of $\cos(\theta)$ can be written as

$$F_{\cos(\theta)}(a) = \begin{cases} 1 & a > 1 \\ \frac{1}{2} + \frac{1}{\pi}\arcsin(a) & -1 < a < 1 \\ 0 & a < -1 \end{cases} \tag{34}$$

for real arguments a, and should be taken to be zero for imaginary ones.

IV. Results

This section conducts a series of specific computational examples to verify that the PCH is applicable to the various adaptive schemes and to compare the clump rate λ_b for the different algorithms.

A. Example Setup

Suppose that the disturbance sequence is an i.i.d. collection of Bernoulli random variables:

$$F_{d_0}(x) = \begin{cases} 1 & \text{if } d_0 > \epsilon \\ \frac{1}{2} & \text{if } -\epsilon < d_0 \le \epsilon \\ 0 & \text{if } d_0 \le -\epsilon \end{cases} \tag{35}$$

where ϵ is a positive real number.

It is easier to work with the transition densities instead of distributions for finding the stationary densities. For LMS, expression (18) may be differentiated with respect to the y variable to give

$$\begin{aligned} f_{w_1|w_0}^{\text{lms}}(y|z) &= \frac{1_{(0,\infty)}(\epsilon^2\mu^2 - 4\mu z(y-z))}{\sqrt{\epsilon^2\mu^2 - 4\mu z(y-z)}} \\ &\quad \cdot \left[f_{x_0}\left(\frac{-\epsilon\mu + \sqrt{\epsilon^2\mu^2 - 4\mu z(y-z)}}{2\mu z}\right)\right. \\ &\quad + f_{x_0}\left(\frac{-\epsilon\mu - \sqrt{\epsilon^2\mu^2 - 4\mu z(y-z)}}{2\mu z}\right) \\ &\quad + f_{x_0}\left(\frac{\epsilon\mu + \sqrt{\epsilon^2\mu^2 - 4\mu z(y-z)}}{2\mu z}\right) \\ &\quad \left. + f_{x_0}\left(\frac{\epsilon\mu - \sqrt{\epsilon^2\mu^2 - 4\mu z(y-z)}}{2\mu z}\right)\right]. \end{aligned} \tag{36}$$

The transition density for signed error LMS is

$$\begin{aligned} f_{w_1|w_0}^{\text{se}}(y|z) &= 1_{(0,\infty)}(z)\left[\frac{1}{\mu}\left(f_{x_0}\left(\frac{z-y}{\mu}\right)\right.\right. \\ &\quad \left.\left. + f_{x_0}\left(-\frac{z-y}{\mu}\right)\right) 1_{(\epsilon/z,\infty)}\left(\frac{z-y}{\mu}\right)\right. \\ &\quad \left. + \frac{1}{2\mu}\left(f_{x_0}\left(\frac{z-y}{\mu}\right) + f_{x_0}\left(-\frac{z-y}{\mu}\right)\right)\right. \\ &\quad \left. \cdot 1_{(-\epsilon/z,\epsilon/z)}\left(\frac{z-y}{\mu}\right)\right] \\ &\quad + 1_{(-\infty,0)}(z)\left[\frac{1}{\mu}\left(f_{x_0}\left(\frac{z-y}{\mu}\right)\right.\right. \end{aligned}$$

$$+ f_{x_0}\left(-\frac{z-y}{\mu}\right)\Big) 1_{(-\infty,-|\epsilon/z|)}\left(\frac{z-y}{\mu}\right) + \frac{1}{2\mu}\left(f_{x_0}\left(\frac{z-y}{\mu}\right) + f_{x_0}\left(-\frac{z-y}{\mu}\right)\right) \cdot 1_{(-|\epsilon/z|,|\epsilon/z|)}\left(\frac{z-y}{\mu}\right)\Big]. \quad (37)$$

For the signed regressor LMS algorithm with symmetric input density we have

$$f^{\mathrm{sr}}_{w_1|w_0}(y|z) = \frac{1}{\mu|z|}\left(f_{|x_0|}\left(\frac{1}{\mu}\left(1 - \frac{y-\mu\epsilon}{z}\right)\right) + f_{|x_0|}\left(\frac{1}{\mu}\left(1 - \frac{y+\mu\epsilon}{z}\right)\right)\right). \quad (38)$$

Finally the expression for the sign-sign version of LMS is

$$f^{\mathrm{ss}}_{w_1|w_0}(y|z) = 1_{(0,\infty)}(z)\Big[\delta(y-(z+\mu))\frac{1}{2}\Big[F_{x_0}\left(\frac{\epsilon}{z}\right) - F_{x_0}\left(-\frac{\epsilon}{z}\right)\Big] + \delta(y-(z-\mu)) \cdot \Big[1 - \frac{1}{2}\Big[F_{x_0}\left(\frac{\epsilon}{z}\right) - F_{x_0}\left(-\frac{\epsilon}{z}\right)\Big]\Big]\Big] + 1_{(-\infty,0)}(z)\Big[\delta(y-(z+\mu)) \cdot \Big[1 - \frac{1}{2}\Big[F_{x_0}\left(\frac{-\epsilon}{z}\right) - F_{x_0}\left(\frac{\epsilon}{z}\right)\Big]\Big] + \delta(y-(z-\mu)) \cdot \frac{1}{2}\Big[F_{x_0}\left(-\frac{\epsilon}{z}\right) - F_{x_0}\left(\frac{\epsilon}{z}\right)\Big]\Big]. \quad (39)$$

We note that the sign-sign version is different in a substantive way from the other three cases. The most obvious difference is that the values of the error always remain on a lattice, e.g., if $w_0 = 0$ then $w_k \in \mu\mathbb{Z}\ \forall\ k$, where $\mathbb{Z}$ denotes the integer lattice. The Markov chain describing the error propagation no longer has a stationary distribution since the chain is periodic. This is because at even time instants $w_{2k} \in 2\mu\mathbb{Z}\ \forall\ k$ and at the odd time instants $w_{2k+1} \in 2\mu\mathbb{Z} + \mu\ \forall\ k$. Hence the chain alternates between odd and even multiples of μ. To handle this situation, let $[b]$ denote the largest element of the lattice $\mu\mathbb{Z}$ not larger than b. The "monotonicity property" holds as before, so the relevant quantity to compute is

$$\begin{aligned} P(w_1 < b|w_0 \geq b) &= P(w_1 < b,\ w_0 = [b+\mu]\,|w_0 \geq b) \\ &= \frac{P(w_1 = [b],\ w_0 = [b+\mu],\ w_0 \geq b)}{P(w_0 \geq b)} \\ &= \frac{P(w_1 = [b]\,|\,w_0 = [b+\mu])P(w_0 = [b+\mu])}{P(w_0 \geq b)}. \end{aligned} \quad (40)$$

This implies that

$$\begin{aligned} \lambda_b &= P(w_1 < b|w_0 \geq b)P(w_0 \geq b) \\ &= P(w_1 = [b]|w_0 = [b+\mu])P(w_0 = [b+\mu]) \\ &= \left[1 - \frac{1}{2}P\left(-\frac{\epsilon}{[b+\mu]} \leq x_0 \leq \frac{\epsilon}{[b+\mu]}\right)\right] \cdot P(w_0 = [b+\mu]). \end{aligned} \quad (41)$$

Although there is no stationary distribution, there is one for the even time instants and one for the odd time instants. The average of these two gives a stationary distribution in the "average sense," and the calculation of $P(w_0 = [b+\mu])$, and λ_b is then straightforward.

B. *Verification of Model and Discussion*

The PCH is a heuristic, it is not a theorem. The assertion that the calculated λ_b values give information about the behavior of the algorithms must be verified experimentally. Towards this end, we calculated λ_b for particular input/disturbance distributions, and then compared these values with direct simulations of the algorithms operated with the same input/disturbance distributions. Overall, the comparisons are striking, far exceeding our expectations. Since the PCH is an asymptotic result (in b), we expect to see a close match for large b. As the evidence shows, however, the calculated and simulated value of λ_b agree remarkably well at the very modest levels of only two times the algorithm step size. Also presented is one case where the λ_b's fail to agree until 6–7 times the step size. This relatively poor performance is attributed to a failure of the monotonicity assumption at the small levels.

For all experiments, the adaptive gain was set at $\mu = 0.2$. The integration step size used in the calculation of the stationary densities and the calculation of λ_b was 0.01. In the simulations, hits and clumps were counted over a period of 1 million iterations, allowing a direct comparison between simulated and calculated values of λ_b. The disturbance distribution was an i.i.d. Bernoulli random variable, and the input distribution was chosen from among Gaussian, Laplace, double-sided Rayleigh, and uniform.

In Fig. 1, curves (*a*), (*b*), and (*c*) show three typical comparisons of the simulated versus computed λ_b. As expected, the curves approach each other for larger values of b, which correspond to large parameter excursions. Somewhat unexpected is the close correlation at small values—the simulations match the theoretical values to within a few percent even for bounds b that are only twice the algorithm stepsize. We observed this close correlation in numerous experiments (not just those reported), and this gives us confidence to use the calculated values of λ_b as a measure of the algorithm's performance.

Indeed, the calculated λ_b give information that is difficult to gather via direct simulation. To attain a certain degree of confidence in a particular set of simulation data,

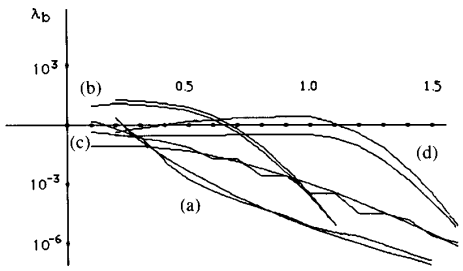


Fig. 1. Model verification; λ_b calculated and simulated for (a) LMS with Gaussian inputs, (b) SE with Gaussian inputs, (c) SS with Gaussian inputs, (d) (failure of) SE with uniform inputs.

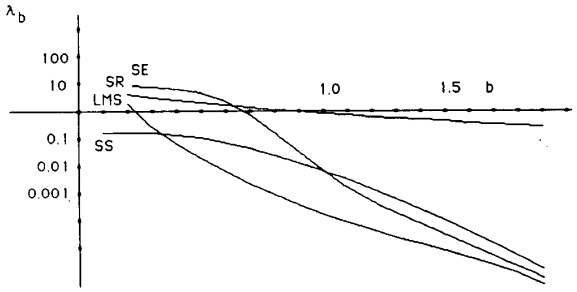


Fig. 2. Gaussian inputs.

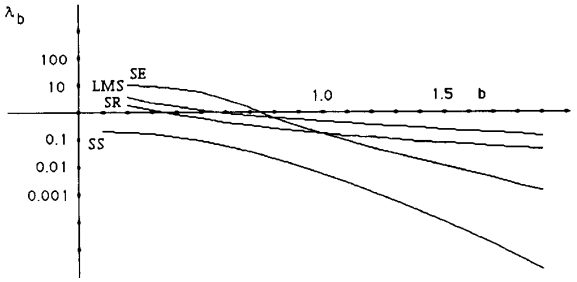


Fig. 3. Laplace inputs.

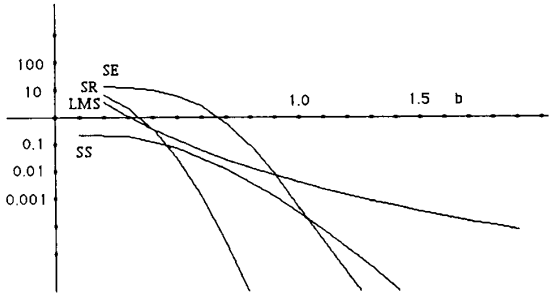


Fig. 4. Rayleigh inputs.

many hits of a bound b are required. As the bound grows larger, these hits become rarer, and the required number of simulation timesteps increases dramatically. The calculation of λ_b, on the other hand, is insensitive to the particular bound chosen. Thus, λ_b grows more accurate precisely as simulation begins to fail. Thus a real strength of the proposed method is that asymptotic data is easily gathered.

Recall that λ_b represents the probability that a particular bound b will be attained. The inverse, $1/\lambda_b$ represents the average time until this bound is attained. Consider an application (such as adaptive control [18], the adaptive hybrid problem in telephony [19], active noise cancellation [20], or in the IIR whitener/predictor [21]), in which the identification portion of the system is embedded inside a feedback loop. In each of these cases, the stability of the entire system requires that parameter error estimates within the adaptive portion remain moderate, since such a system can be destabilized by large excursions of the parameter estimates. On the other hand, if properties of the inputs and disturbances are known or can be measured, then the PCH could be used to estimate the mean time until failure of the adaptive system.

In Fig. 1, the pair of curves labeled (d) shows that the simulated curve may differ significantly from the calculated λ_b. The source of this failure is that the monotonicity property on which the PCH is based does not hold for small b. Consider (2) (the SE algorithm) with Bernoulli disturbances $d_k = \pm(1/2)$ and uniform $[-\frac{1}{2}, \frac{1}{2}]$ inputs. For small parameter estimate errors w_k, with $w_k < 1$, sgn $(w_k x_k + d_k)$ is equally likely to be $+1$ as -1, irrespective of the sign of w_k, implying that the w_k process is not monotonic in this region. Consequently, one cannot expect the PCH to return accurate values of λ_b for b smaller than (at least) $1 + \mu$. This is, indeed, when we observe the curves begin to coalesce. Situation (d) provides a useful warning that estimates based on the PCH must be examined carefully to ascertain if the desired monotonicity property is fulfilled in the region for which λ_b is calculated.

C. A Comparison of the Four Algorithms

Given the striking agreement between the calculated and simulated values of λ_b, we approach the comparison of the algorithms with some measure of confidence. Each of the four algorithms was excited by inputs from each of three distributions: Gaussian (mean = 0, variance = 1), double sided Rayleigh (with parameter $\alpha = 1$), and Laplacian (with $\alpha = 1$). This was enough to demonstrate conclusively that no single algorithm was always better (had smaller λ_b) than any other algorithm.

Consider Figs. 2–4, which display calculated λ_b values for the four algorithms. While LMS was the clear winner for Gaussian inputs (as one might expect), it is the clear loser for large excursions of the parameter estimate errors when the input comes from a Laplacian distribution. Similarly, the SR algorithm had the smallest λ_b when excited by Rayleigh inputs, but the largest for Gaussian inputs. While the SE algorithm has uniformly the largest λ_b for small b over all three cases, for large b it tends towards zero faster than LMS and SR when the inputs were chosen from the Laplacian distribution.

The SS algorithm has uniformly the smallest λ_b over the input distributions considered for small b. For large b, the SS algorithm also performs remarkably well, though it is only rarely "the best." If this is truly the generic behavior of this algorithm (and not a fiction of just the particular input/disturbance distributions we examined), then this may account for the algorithms popularity when used in a nonstationary environment. Speaking loosely, one might think of a nonstationary process as one which moves from one "stationary" distribution to another. An algorithm that was able to perform well over large classes of stationary distributions would thus be highly desirable.

V. Conclusions

The Poisson clumping heuristic allows a comparison of the LMS adaptive algorithm and its signed variants in terms of the likelihood that the parameter estimate errors in these processes will achieve large bounds. Such hits of large values are rare events which are hard to simulate using straightforward techniques. The PCH gives a way of calculating these rare events that is essentially independent of the bound.

This is useful in two ways. First, large parameter excursions imply that the output of the adaptive filter will be far from its desired value, and so provides a measure of performance for the various algorithms. Second, in adaptive systems which involve a feedback of the estimated output back into the input of the estimator, there is a possibility that large parameter errors will destabilize the system. If a bound on this stability region is known, and if the input/disturbance statistics are known or can be measured, then the PCH can be used to determine the mean time until failure of the adaptive system.

Although the PCH is a heuristic (and not a theorem), there are substantial theoretical underpinnings for its use in the adaptive setting. For example, Markov chains with a countable state space and certain strong recurrence conditions can be shown (via regeneration theory arguments) to exhibit exponential waiting times between visits to small probability sets [3]. Our chains have continuous state spaces (except for SS), but they exhibit a strong recurrence property (e.g., the "monotonicity property"). Moreover, a fine quantization of the state space (in order to make it countable) is unlikely to introduce quantitatively different behavior from the unquantized versions. It would be of great theoretical interest, however, to verify these statements rigorously.

References

[1] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, Jr., "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, no. 8, pp. 1151-1162, Aug. 1976.

[2] W. A. Sethares and C. R. Johnson, Jr., "A comparison of two quantized state adaptive algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, Jan. 1989.

[3] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*. New York: Springer, 1989.

[4] T. Brennan, "Asymptotics of divergent LMS," presented at the 1990 Int. Symp. Inform. Theory, San Diego, CA, Jan. 1990.

[5] N. J. Bershad and L. Z. Qu, "On the probability density function of the complex scalar LMS adaptive weights," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, Jan. 1989.

[6] N. J. Bershad and L. Z. Qu, "On the joint characteristic function of the LMS adaptive filter weights," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, Dec. 1984.

[7] M. A. Aizerman, E. M. Braverman, and R. I. Rozonoer, "The method of potential functions for restoring the characteristic of a function from randomly observed points," *Automat. Remote Contr.*, vol. 25, no. 12, Dec. 1964.

[8] N. A. Verhoeckx and T. A. C. M. Claasen, "Some considerations on the design of adaptive filters with the sign algorithm," *IEEE Trans. Commun.*, vol. 32, no. 3, Mar. 1984.

[9] A. Gersho, "Adaptive filtering with binary reinforcement," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 2, pp. 191-198, Mar. 1984.

[10] J. L. Moschner, "Adaptive equalization via fast quantized state methods," Tech. Dep. 6796-1, Inform. Syst. Lab., Stanford University, 1970.

[11] D. L. Duttweiler, "Adaptive filter performance with nonlinearities," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, Aug. 1982.

[12] N. J. Bershad, "On the optimum data nonlinearity in LMS adaptation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 1, Jan. 1986.

[13] W. A. Sethares, I. M. Y. Mareels, B. D. O. Anderson, and C. R. Johnson, Jr., "Excitation conditions for sign-regressor LMS," *IEEE Trans. Circuits Syst.*, June 1988.

[14] R. W. Lucky, "Techniques for adaptive equalization of digital communication systems," *Bell Syst. Tech. J.*, vol. 45, Feb. 1966.

[15] N. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[16] R. R. Bitmead, "Persistence of excitation and the convergence of adaptive schemes," *IEEE Trans. Inform. Theory*, vol. 30, Mar. 1984.

[17] J. E. Mazo, "On the independence theory of equalizer convergence," *Bell Syst. Tech. J.*, vol. 58, no. 5, May-June 1979.

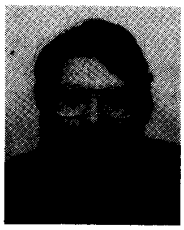
[18] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[19] W. A. Sethares, C. R. Johnson, Jr., and C. Rohrs, "Bursting in adaptive hybrids," *IEEE Trans. Commun.*, Aug. 1989.

[20] L. J. Eriksson, M. C. Allie, and R. A. Greiner, "The selection and application of an IIR adaptive filter for use in active sound attenuation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 4, Apr. 1987.

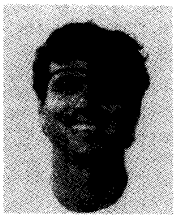
[21] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[22] R. R. Bitmead and P. E. Caines, "Escape time formulation of robust stochastic adaptive control," in *Proc. 27th Conf. Decision Contr.* (Austin, TX), Dec. 1988.



William A. Sethares (S'84-M'86) received the B.A. degree in mathematics from Brandeis University, Waltham, MA, and the M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY.

He has worked at the Raytheon Company as a Systems Engineer and is currently on the faculty of the Department of Electrical and Computer Engineering, University of Wisconsin-Madison. His research interests include adaptive systems in signal processing, communications, and control.



James A. Bucklew (S'75-M'79) received the Ph.D. degree from Purdue University, West Lafayette, IN, in 1979.

He is currently a Professor in the Department of Electrical and Computer Engineering and the Department of Mathematics at the University of Wisconsin-Madison. His research interests are in the applications of probability to signal processing and communications problems.

Dr. Bucklew received the Presidential Young Investigator Award in 1984 and is currently the Associate Editor-at-Large for the IEEE Transactions on Information Theory.