

CLNI '93 Contribution

"Learning an Unknown Signalling Alphabet" ¹⁾by Edward C. Posner ^{2), 3)} and Eugene R. Rodemich ³⁾*ABSTRACT* (108 words)

We estimate the time it takes to learn an unknown signalling alphabet as a function of alphabet size and other parameters, a scaling law. The model assumes the signals are orthogonal vectors in a Euclidean space, with strong Gaussian noise added. The incremental learning procedure adjusts the current estimate of the alphabet by moving the signal closest to the received signal even closer to it, and the others further away. We show convergence, although in sometimes a very long time. Simulations show that the theory tracks reality. We also compare incremental with batch learning, or full maximum-likelihood over all the mass of received data. We find conditions where incremental learning is almost as good as batch, and conditions where it isn't.

¹⁾ Work Supported by Pacific Bell and NASA

^{2), 3)} Department of Electrical Engineering, Caltech

³⁾ Caltech Jet Propulsion Laboratory

LEARNING AN UNKNOWN SIGNALLING ALPHABET¹⁾
 BY EDWARD C. POSNER²⁾³⁾ AND EUGENE R. RODEMICH³⁾

SUMMARY

This paper derives a rigorous scaling law for incremental and batch learning in the following situation. There is an unknown set of n orthogonal or biorthogonal signals in n real dimensions. Presentations are made of randomly selected signals from the set with gaussian noise added. The first goal is to estimate the signal set incrementally. After each presentation, the current best estimate of the signal set is updated by moving the winning signal, the one closest to the received signal plus noise, closer to the received signal plus noise. The other $n-1$ signals are moved away. Estimates are derived for how many presentations are needed to get within a given mean-square angular error from the true signal set with high probability, and simulations are conducted to verify validity in practical situations. The incremental approach is compared with batch learning, i.e., full maximum-likelihood on the entire set of presentations, which delays making estimates until all the data (presentations) are in. The result is that batch is not much better than incremental learning in the biorthogonal case, but in the orthogonal case, batch can be truly better. Thus, incremental learning is sometimes bad.

The question first considered relates to what others sometimes call unsupervised competitive learning but is called here *incremental learning*. We have (at most) n orthonormal signals in n Euclidean dimensions. Here n is assumed known, or at least an upper bound for the number of dimensions is known. We don't know, though, what the signal set is, and maybe not even the number of signals. Random unit vectors would be approximately orthonormal, a not much different problem.

We receive random signals with zero-mean additive gaussian noise of unknown noise

¹⁾Research Supported in part by NASA through the Caltech Jet Propulsion Laboratory and in part by Pacific Dell through the Caltech Electrical Engineering Department

²⁾Department of Electrical Engineering, California Institute of Technology

³⁾Jet Propulsion Laboratory, California Institute of Technology

power (orthogonal case), and also with a random sign flip (biorthogonal case). Biorthogonal signal sets (orthogonal signals and their negatives) are more efficient of channel bandwidth in gaussian noise, so there is ample reason for considering this case. As will be seen, though, one reason for considering biorthogonality is to contrast it with the orthogonal case. We try to learn what the signalling set is on the basis of experience, that is, given random independent presentations of the signals. The results of this paper on the number of samples needed are concerned with the case of strong noise. We start with a random or any orthonormal set of unit vectors as our initial estimate. After receiving the first signal plus noise, we compute the distance to each member of the initial basis set, and find the closest. The new basis set, as we said, is obtained from the old by moving its closest member toward the received signal plus noise, and all the others away.

This modification process is much as in stochastic approximation or feature maps. There are similarities, too, with work on unsupervised learning in control theory. The estimates here are the first analytic estimates of learning time, however, in situations that we treat. Adaptive Resonance Theory or ART and non-parametric regression also touch on what we are doing.

The process of adjusting the orthonormal basis is repeated after each new received signal plus noise. The estimate of the signal set is updated each time in a potentially unending process of repeated presentations. The number of signals need not be exactly known; a method is derived for estimating the number. That is why we only need an upper bound on the number of dimensions.

If a constant update rule is used, as it is in this paper, we will see that eventually a steady state is reached with a finite mean square error between each member of the estimated orthogonal signalling set (which will not be exactly orthogonal) and the closest member of the true orthogonal set. The number of samples needed to get acceptably close (depending on a small real parameter ϵ) is called the *learning time*.

The term "unsupervised" means that we are not told what the actual signal presented was. "Competitive" means that the different signals compete with each other to be chosen if a given signal vector is a member of the estimated signalling set at a given step, then

nothing close to it can also be a member. We move closer to the assumed presented signal, and the other vectors are moved further away. A “scaling law” tells how large a sample is needed to learn, i.e., get close to the true Set, as the problem size increases. This is sometimes called the “curse of dimensionality” in statistics, for the sample size here can become large indeed, and, in batch learning, everything interacts with everything else.

A possible interpretation of the biorthogonal case could arise from Code Division Multiple Access (CDMA) as could be used for advanced digital cellular radio telephone. In this, a number n of users are each given one waveform to signal with by binary antipodal signalling. The n waveforms are orthogonal to each other, to avoid mutual interference. The waveforms are constructed from n dimensions corresponding to n “chips,” which can be thought of as short intervals where the waveform can take one of two values, say a phase of $+90^\circ$ or -90° . However, the dimensions can more generally correspond to samples at the Nyquist rate for the baseband channel. The learning then refers to someone who does not know the orthogonal signalling set trying to figure it out to get free service by choosing a waveform as far from others in the set as possible. Another interpretation of incremental learning could be in language acquisition where the signal set corresponds to the phonemes of a language. Batch learning, which we also consider, could perhaps be considered like some birdsong learning in some species.

The biorthogonal signal set which we consider corresponds to using the assigned waveform or its negative to transmit random binary data. The fact that we are presented with only one waveform at a time would correspond to light traffic, or to the learning receiver being closer to one transmitter than to the others. The fact that the noise is strong in the total received signal is quite reasonable, and is the case we consider in this paper. For if the signalling set is known, the noise power added to the matched filter output of the detector can be small for low error probability, and yet the noise vector added to the signal vector can have large mean square, which is n times the variance of the matched filter output. Moreover, if there is error-correcting coding (11) top of the CDMA, even the noise on the matched filter output need not be small: it can be moderate, say a signal-to-noise ratio of **1(0 dB)** or less. This explains why some of our results are derived for the case where the noise vector mean square is comparable to or greater than the number of dimensions.

It is a reasonable operating region for CDMA.

We develop a formal model for the problem, or actually for a slightly more general one. The probability distributions of the received signal-plus-noise are evaluated. The effect of the received signal on the current estimated orthonormal set is determined, and this gets converted into an expected change, by evaluating certain high-dimensional integrals. Some products of expectations needed to be evaluated. We introduce a matrix M whose eigenvalues help determine the rate of convergence of learning. The special case of equal signal probabilities is worked in detail. It is particularized to the case of strong noise, for which the convergence time is evaluated.

The variance of the error angles is found and an example worked out. A typical result is that the incremental learning time T_1 and angular variance $v^2 (< \frac{1}{n})$ are related in strong noise by

$$T_1 \approx \frac{1}{v^2} \log\left(\frac{1}{nv^2}\right) \cdot \frac{9}{4} \frac{\sigma^8}{(\log n)^3}.$$

The effect of error angle on the CDMA model is estimated, and the problem of not being close to the correct signal set at the start considered. The more general case of unequal probability of presentation is also briefly considered. We then allow a smaller number of orthonormal signals to be present (and their number learned) than the dimensionality of the space. This may be a more realistic assumption than assuming the number of signals known. The case of 110 signals, i.e., detecting whether signals are being used at all or whether one is receiving pure noise, is also covered.

The above is the basic main theme, but we also consider some side themes of interest. In the description just presented, the signalling set used was biorthogonal, not orthogonal. A given orthonormal basis vector or its negative was chosen with equal probability for each presentation. Thus, the orthonormal set is not strictly speaking unique. It is only defined up to a sign flip on each basis vector. We make the modifications necessary to do the strict orthogonal case. The two cases are compared; there is found to be not much saving in incremental learning in knowing that a particular sign of the unknown signal vector will always be chosen.

We then tackle an even harder problem, learning for the orthogonal case using full maximum-likelihood batch instead of incremental learning. Here “batch” is identified with estimating the signal set after all presentations are made. Both the orthogonal and the biorthogonal maximum-likelihood cases are done, and the four results compared: orthogonal and biorthogonal, incremental and batch. The maximum-likelihood estimate in the orthogonal case has a learning time proportional (for large noise σ^2) to σ^6 rather than to σ^8 , as it is for the incremental orthogonal case. This suggests that the incremental algorithm does not fully capture the constraint of exact orthogonality. In the biorthogonal case, though, the maximum-likelihood estimate has the same σ^8 behavior as for the incremental algorithm, but the coefficient of σ^8 in the batch case is smaller by the factor

$$(4(\log 7)^{3/772}) \times (1/\log(1/nv^2)),$$

where v^2 is the small desired variance after learning.

We finally develop in our most recent work more accurate estimates valid for ImseII-able sample sizes. These refined estimates are compared with simulations and the results shown to agree well with our theory.