

---

# Rigorous Moment-Based Automatic Modulation Classification

---

Darek T. Kawamoto

Robert W. McGwier

Hume Center, Virginia Tech, Arlington, VA 22203, USA

DAREK@VT.EDU

RWMCGW1@VT.EDU

## Abstract

In this paper we develop the connection between the high-order moments, orthogonal polynomials, and probability densities representing signal constellations with AWGN in order to improve moment-based Automatic Modulation Classification (AMC). The result is that an approximate weighted  $L^2$  distance between probability densities can be computed using a Euclidean distance on vectors consisting of series expansion coefficients. This analysis justifies the use of high-order moments in AMC. A discriminative Deep Neural Network (DNN) is trained to perform AMC, resulting in near-maximum likelihood performance at marginal SNR.

## 1. Introduction

Automatic Modulation Classification (AMC) is a critical component of spectrum awareness in cognitive radios. In order to diagnose a transmitted signal's protocol, symbol rate detection combined with AMC can be used to narrow the search down to several candidate protocols.

Moment-based classifiers, and the related cumulant classifier (Reichert, 1992; Swami & Sadler, 2000; Spooner et al., 2000; Spooner, 2001; Dobre et al., 2004; Aslam et al., 2010; Su, 2013), use the Higher Order Statistics (HOS) of the input signal to perform feature-based AMC. Historically, cumulant classifiers have been favored because of their robustness to Gaussian noise (removing the need for an SNR parameter in the classifier formulation).

The classification decisions in (Reichert, 1992; Swami & Sadler, 2000) are performed using a hierarchical decision tree, while (Spooner et al., 2000; Spooner, 2001; Dobre et al., 2004) use a weighting on the cumulants and make assumptions in order to build a Euclidean metric around the weighted cumulant vectors. The former use a series of thresholded decisions (based on the statistical cumulants

of the modulations), while the latter are able to use metric space-based methods such as k-Nearest Neighbors.

### 1.1. Preliminary Observations

The biggest difficulty in moment-based classifiers is that the input signal's empirical moments may greatly deviate from its statistical moments, especially for short-length inputs (e.g., 100 to 250 symbols). While some of this deviation is caused by AWGN, a significant amount is caused by a mismatch between the prior distribution of the symbols (used to calculate the statistical moments) and the proportions of the actually transmitted symbols (used to calculate the empirical moments). This difficulty mirrors that of density estimation of multi-modal probability distributions.

This suggests that instead of comparing empirical moments to statistical moments, we should compare the inputs with a large number of simulated or sampled training data. The challenge is to derive a reasonable metric suitable for this modulation classification task.

### 1.2. Novel Contribution

The main contribution of this work is application of orthogonal polynomials (Szego, 1939; Dunkl & Xu, 2014) to moment-based AMC. The probability densities corresponding to modulation constellations under AWGN (with noise power  $N_0 < 2$ ) can be decomposed into a series expansion via Gram-Charlier Approximation (GCA) (Kendall & Stuart, 1969) – the series' coefficients are related to the probability distribution's moments through the Hermite polynomials. Furthermore, these probability densities are contained in a weighted Hilbert space,  $L^2(\mathbb{C}, w(z) = \pi e^{z\bar{z}})$ , so that their weighted  $L^2$  distance can be computed using Euclidean distance on the series expansion coefficients. While the application of GCA to AMC is novel, the topics of GCA and Hermite polynomials are well-studied in statistics, finance, and physics.

To the authors' knowledge, this is the first work to directly link the moment-based classifier to a distribution-based classifier which compares the input data to reference modulation constellations represented by probability densities (Wang & Wang, 2010; Wang & Chan, 2012; Zhu et al., 2013).

Because an underlying assumption of this work is that the moment-estimate variances are driven by the finite-data constraint, this work makes no assumptions about the moment-estimator error distributions. Instead, to mitigate this “uneven prior” problem, we create a large set of training data and train a Deep Neural Network (DNN) to perform classification without a specific moment-estimator error model. The DNN can be thought of as dividing the input space into non-linear decision regions, and also returning approximate probabilities of the input data belonging to each modulation class  $S_m \in \mathcal{S}$ . While this DNN would not be unlike any other DNN operating on moment vectors, this work rigorously justifies the use of moments for classification through the relation to the  $L^2$  distance described above.

From the machine learning perspective, this work can be viewed as the design and justification of expert features for use in a DNN classifier. This work is complementary to other efforts (O’Shea et al., 2016) to perform modulation recognition and signal processing using automatic feature learning through deep Convolutional Neural Networks (CNN). One particular difference to highlight is that the AMC proposed in this paper disregards all notions of time-ordering in the input signal data, while that sense of time is preserved and exploited in a CNN. These authors expect that CNNs are necessary for neural-network-based demodulation and follow-on processing, which this work does not address.

### 1.3. Signal Model, Definitions, Assumptions

Throughout this paper, we will limit the modulations to linear ones (i.e., ASK, PSK, and QAM), and work with the probability densities of the noisy, post-receiver symbols. Given an SNR  $\frac{E_s}{N_0}$ , signal constellation  $S_m \in \mathcal{S}$ , and complex noise  $\omega = \omega_I + j\omega_Q$  with  $\omega_I, \omega_Q \sim \mathcal{N}(0, \frac{N_0}{2})$ , each received symbol  $Z \in \mathbb{C}$  has a probability density given by the complex Gaussian mixture,

$$f_Z(z|S_m, N_0) = \sum_{s_k \in S_m} \frac{1}{\pi N_0 |S_m|} e^{-\frac{|z-s_k|^2}{N_0}}, \quad (1)$$

where complex baseband symbols  $s_k$  iterate over the constellation points of  $S_m$ .

The basic approach is to estimate an empirical density  $\hat{f}_Z(z)$  from the moments of  $L$  received symbols, and then to find the constellation whose probability density most closely matches the estimated density (with respect to  $L^2(\mathbb{C}, w(z) = \pi e^{z\bar{z}})$  distance), i.e.,

$$\arg \min_{S_m} \|\hat{f}_Z(z) - f_Z(z|S_m, N_0)\|_2. \quad (2)$$

We generate training data by sampling from  $f_Z(z|S_m, N_0)$  and estimating the densities  $\hat{f}_Z^{(i)}(z|S_m, N_0)$  from these

samples. While a nearest-neighbor classifier could then compute

$$\arg \min_{S_m} \|\hat{f}_Z(z) - \hat{f}_Z^{(i)}(z|S_m, N_0)\|_2, \quad (3)$$

we choose to use a discriminative DNN in order to decrease computational cost and output soft-decisions.

### 1.4. Outline

The rest of this paper is structured as follows: Section 2 outlines the procedure for estimating the probability density of the input I/Q data using its input moments (GCA), and describe the metric on the corresponding Hilbert space. Section 2.5 details the method of classification using a DNN. Experiments and results are detailed in Section 3, with conclusions in Section 4. Additionally, the proof of a claim made in Section 2.3 is presented in Section 5.

## 2. Analysis – The Moment Problem

The general task of mapping a sequence of moments to its measure  $F$  is called the moment problem. On  $\mathbb{R}$ , a potential solution to the problem is Gram-Charlier Approximation (GCA). GCA approximates a probability density  $f(x)$  by computing a Radon-Nikodym derivative ( $\frac{dF}{d\Phi}$ ) from Gaussian measure to the desired measure; letting  $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ ,

$$f_X(x) = \frac{f_X(x)}{\phi(x)} \phi(x). \quad (4)$$

### 2.1. Gram-Charlier Approximation on $\mathbb{R}$

GCA can be formulated using the cumulants of a distribution (Kendall & Stuart, 1969) or from its moments (Sauer & Heydt, 1979), and can also be done in the complex domain.

In GCA, the Radon-Nikodym derivative  $\frac{f_X(x)}{\phi(x)}$  is a series expansion with respect to the Hermite polynomials  $H_n(x)$ ,

$$\frac{f_X(x)}{\phi(x)} \approx \sum_{n=0}^{\infty} \frac{E[H_n(x)]}{\sqrt{n!}} \frac{H_n(x)}{\sqrt{n!}}, \quad (5)$$

where

$$H_n(x) = (-1)^n e^{\frac{x^2}{2}} \frac{d^n}{dx^n} e^{-\frac{x^2}{2}}, \quad (6)$$

and where the  $\sqrt{n!}$  terms are normalizing factors. The formula for GCA in  $\mathbb{R}$  is then

$$f_X(x) \approx \sum_{n=0}^{\infty} \frac{E[H_n(x)] H_n(x)}{n!} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \quad (7)$$

The approximation is equal under certain convergence conditions (Section 2.3).

As alluded to earlier, the Hermite polynomials form an orthogonal basis in the weighted Hilbert space  $L^2(\mathbb{R}, e^{-\frac{x^2}{2}})$ . Consequently,

$$\int_{\mathbb{R}} H_m(x)H_n(x) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \begin{cases} 0, & m \neq n, \\ n!, & m = n. \end{cases} \quad (8)$$

## 2.2. Gram-Charlier Approximation on $\mathbb{C}$

GCA can be extended into  $\mathbb{C}$  as well. Letting  $z := x + jy$  and the overline denoting complex conjugation ( $\bar{z} := z - jy$ ), we series expand with respect to complex Hermite polynomials related to the complex Gaussian measure,

$$H_{p,q}(z) := (-1)^{p+q} e^{z\bar{z}} \frac{\partial^{p+q}}{\partial_z^p \partial_{\bar{z}}^q} e^{-z\bar{z}}, \quad (9)$$

$$\phi(z) = \frac{1}{\pi} e^{-z\bar{z}} \quad (10)$$

$$f_Z(z) = \frac{f_Z(z)}{\phi(z)} \phi(z), \text{ with} \quad (11)$$

$$\frac{f_Z(z)}{\phi(z)} \approx \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \frac{E[\overline{H_{p,q}(z)}]}{\sqrt{p!q!}} \frac{H_{p,q}(z)}{\sqrt{p!q!}}, \quad (12)$$

$$f_Z(z) \approx \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \frac{E[\overline{H_{p,q}(z)}]}{\sqrt{p!q!}} \frac{H_{p,q}(z)}{\sqrt{p!q!}} \frac{1}{\pi} e^{-z\bar{z}}. \quad (13)$$

In the sequel, we will deal exclusively with probability densities in  $\mathbb{C}$ , and the complex Hermite polynomials  $H_{p,q}(z)$ . The complex Hermite polynomials obey simple recurrence relations (Dunkl & Xu, 2014),

$$H_{p+1,q}(z) = zH_{p,q}(z) - \frac{\partial H_{p,q}(z)}{\partial_{\bar{z}}}, \text{ and} \quad (14)$$

$$H_{p,q+1}(z) = \bar{z}H_{p,q}(z) - \frac{\partial H_{p,q}(z)}{\partial_z}, \quad (15)$$

and are orthogonal,

$$\int_{\mathbb{C}} H_{p,q}(z)H_{r,s}(z) \frac{1}{\pi} e^{-z\bar{z}} dz = \begin{cases} 0, & p \neq r \text{ or } q \neq s, \\ p!q!, & p = r \text{ and } q = s. \end{cases} \quad (16)$$

The first few complex Hermite polynomials are

$$\begin{aligned} H_{0,0}(z) &= 1, & H_{1,0}(z) &= z, \\ H_{1,1}(z) &= |z|^2 - 1, & H_{2,0}(z) &= z^2, \\ H_{2,1}(z) &= z^2\bar{z} - 2z, & H_{3,0}(z) &= z^3, \\ H_{2,2}(z) &= |z|^4 - 4|z|^2 + 2, & H_{4,0}(z) &= z^4. \end{aligned}$$

Noting that the series expansion depends only on coefficients of the orthonormal basis functions  $\frac{H_{p,q}(z)}{\sqrt{p!q!}}$ , we will operate with vectors of series expansion coefficients

$$h_{p,q}(f_Z) := \frac{E[\overline{H_{p,q}(z)}]}{\sqrt{p!q!}}. \quad (17)$$

For example,

$$h_{2,1}(f_Z) = \frac{E[\overline{H_{2,1}(z)}]}{\sqrt{2}} = \frac{E[\bar{z}^2 z - 2\bar{z}]}{\sqrt{2}} = \frac{E[\bar{z}^2 z] - 2E[\bar{z}]}{\sqrt{2}}. \quad (18)$$

These coefficients are easy to compute from the input data's empirical moments.

Rewriting (13), we have

$$f_Z(z) \approx \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} h_{p,q}(f_Z) \frac{H_{p,q}(z)}{\sqrt{p!q!}} \frac{1}{\pi} e^{-z\bar{z}}. \quad (19)$$

We will see that the Euclidean distance on these coefficients equals the weighted  $L^2$  distance between two approximate probability densities due to the Hilbert space these densities belong to.

## 2.3. GCA Convergence

In general, GCA may not converge nicely (or at all!). This issue comes from the Radon-Nikodym derivative and the span of the weighted Hermite polynomial basis functions; the Hilbert space associated with GCA on  $\mathbb{C}$  contains all functions which are square-integrable with respect to a weight function  $w(z) = \pi e^{z\bar{z}}$ , i.e.,  $L^2(\mathbb{C}, w(z) = \pi e^{z\bar{z}})$  (Itô, 1952; Dunkl & Xu, 2014).

It can be proven (Section 5) that Gaussian mixtures of (1) with noise power  $N_0 < 2$  are contained in this Hilbert space. Despite this restriction, this limit does not inherently prohibit classification attempts at very low SNRs. Since integrability is determined by tail behavior controlled by the noise variance, simply rescaling the input signal by  $\frac{1}{\sqrt{G}}$  results in a new tail which decays faster by a factor of  $e^{-G|z|^2}$ . The real difficulty lies in properly estimating the moments, and from the decreased distance between densities associated with each modulation.

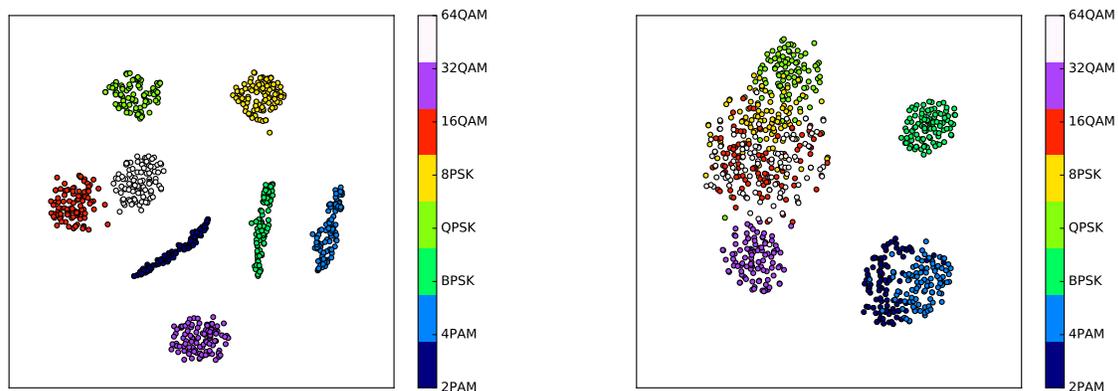
## 2.4. Distance Calculations in $L^2(\mathbb{C}, w(z))$

The distance between two functions  $f_1, f_2 \in L^2(\mathbb{C}, w(z))$  is the weighted  $L^2$  norm of the difference of the two functions:

$$d(f_1, f_2) = \sqrt{\int_{\mathbb{C}} |f_1(z) - f_2(z)|^2 \pi e^{z\bar{z}} dz}. \quad (20)$$

Substituting (19) into (20), and making use of the orthogonality of the Hermite polynomials (with respect to Gaussian measure), it can be proven that

$$d(f_1, f_2) = \sqrt{\sum_{p=0}^{\infty} \sum_{q=0}^{\infty} |h_{p,q}(f_1) - h_{p,q}(f_2)|^2}. \quad (21)$$


 (a) High SNR, lots of data (20 dB  $\frac{E_s}{N_0}$ ,  $L = 2048$  symbols).

 (b) Low SNR, less data (3 dB  $\frac{E_s}{N_0}$ ,  $L = 128$  symbols).

Figure 1. t-SNE visualization of the GCA coefficient space.

Table 1. Cluster variance of modulation classes in GCA coefficient space

L \ Class	2ASK	4ASK	BPSK	QPSK	8PSK	16QAM	32QAM	64QAM
32	56.8359	59.7063	43.4253	43.332	48.4357	52.0161	67.6009	52.5992
64	30.8399	33.614	19.1733	25.0875	24.1494	27.0839	34.7037	28.3482
128	14.9264	15.0719	11.4357	11.1725	11.6224	12.8272	18.7215	13.6556
256	6.96395	7.38893	5.5833	6.15215	6.00063	6.7035	9.08378	7.34727
512	3.50094	3.74949	2.61287	2.86368	3.09839	3.40288	4.69743	3.58065
1024	1.74854	1.89098	1.43434	1.54149	1.49448	1.80567	2.20783	1.83182
2048	0.82347	0.907244	0.631166	0.766772	0.747173	0.848353	1.10696	0.883359
4096	0.494611	0.499778	0.348313	0.377318	0.367536	0.437135	0.560456	0.47068
8192	0.237483	0.243161	0.167552	0.194374	0.187227	0.234803	0.306519	0.221274
16384	0.10472	0.117317	0.078921	0.0898927	0.0912061	0.111279	0.152445	0.113516
32768	0.0574201	0.0604335	0.0416373	0.0482941	0.046277	0.0519672	0.0711747	0.0560899
65536	0.0284362	0.0297853	0.0203381	0.0224413	0.0241345	0.0263447	0.0366153	0.0278138
131072	0.0135647	0.0152332	0.0102348	0.0117713	0.0126825	0.0137095	0.018084	0.0141807

Since estimating an infinite number of moments is impractical, we truncate the series coefficients to obtain a lower bound on the distance (by Bessel's inequality):

$$d(f_1, f_2) \geq \sqrt{\sum_{p,q} |h_{p,q}(f_1) - h_{p,q}(f_2)|^2}. \quad (22)$$

This Euclidean metric on the (sub-)set of GCA coefficients enables use of metric space-based classifiers (e.g., k-NN, SVM, etc.).

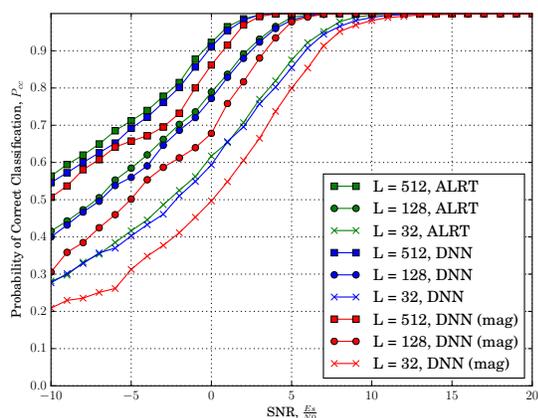
## 2.5. Automatic Modulation Classification

The developments in this section suggest that received symbols can be analyzed in terms of its moments to discriminate between various modulation constellations. We model the probability density function of received sym-

bols as a convolution of the modulation constellation with AWGN, as in (1). We treat the received symbols as unordered data, compute the GCA series coefficients  $h_{p,q}$ , and compare them to entries in a reference library.

We populate the reference library with the GCA series coefficients of many generated example signals from each modulation (with AWGN). Rather than implement a k-Nearest Neighbors classifier, whose complexity grows with the size of the reference library, we choose to use the library to train a discriminative Deep Neural Network (DNN), which implements a non-linear classifier.

As a normalizing step, we divide the data by its standard deviation in order to ensure GCA convergence of linear modulation constellations under AWGN (see Section 5). After this step, we calculate the empirical moments and



(a) Performance vs Length and SNR

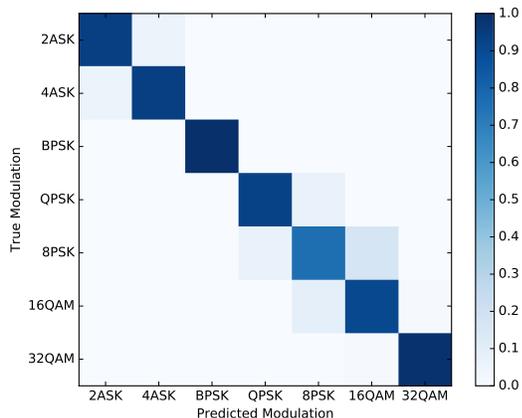

 (b) Confusion Matrix, Coherent DNN with  $L = 128$ ,  $\frac{E_s}{N_0} = 3\text{dB}$ 

Figure 2. 7-Class AMC Performance

GCA series coefficients  $h_{p,q}$  as defined in (17). The real and imaginary components of the coefficients are split to form real-valued inputs to a DNN classifier. In a variant of this classifier, the magnitudes of the complex coefficients are used instead (to make the classifier invariant to phase-offsets).

Figure 1 shows t-SNE visualizations (Maaten & Hinton, 2008) of the GCA coefficient space for 8 different modulation types. At high SNR (20 dB  $\frac{E_s}{N_0}$ ) and with many received symbols (2048) to estimate the constellations' probability density function, the classes are well-separated even in the low-dimension manifold plotted in Figure 1(a). On the other hand, at lower SNR (3 dB  $\frac{E_s}{N_0}$ ) and with only a few received symbols (128), the two-dimensional visualization fails to keep each class separated.

While t-SNE can be used as a rough indicator of class separability, it does not adequately show the scale of the clusters of space around each modulation. Table 1 shows the cluster variance of each modulation class as a function of  $L$ , the number of received symbols, at 3 dB  $\frac{E_s}{N_0}$ . This table shows that cluster variance roughly decreases by a factor of 2 for every factor of 2 increase in the length of received data used in estimating the moments. This is consistent with the moment-estimator variances decreasing as the amount of data collected increases.

### 3. Experimental Results

#### 3.1. Experiment Design

The DNN architecture consists of 4 layers of ReLU6 activations (of widths 200, 100, 100, and 100, respectively) and one final soft-max layer, trained with backpropagation

against the cross-entropy cost function in the TensorFlow framework (Abadi et al., 2015). The DNN input consist of the coefficients  $h_{p,q}$ , and the output layer width is the number modulation classes to decide from. Because we are unable to leverage the SNR-invariant nature of cumulants, we create a new DNN for each SNR / input length pair.

Each DNN is parameterized by  $L$ , the number of received symbols, and SNR  $\frac{E_s}{N_0}$ . The amount training data  $N$  generated for each modulation type is set such that  $NL = 1280000$ ; the training data is used in several thousand iterations of supervised training.

In addition to regularization by dropout, we introduce Gaussian noise to the DNN inputs. For the experiments performed in this work, we add noise with  $\sigma = 0.01$  to the first 500 iterations of training. More analysis of the bias-variance tradeoff is necessary to improve training methods at low SNR where the signal space may not be adequately sampled in training.

For each experiment, the DNN classification results are compared to an Average Likelihood Ratio Test (ALRT) classifier (Sills, 1999), as well as the magnitude-only DNN.

#### 3.2. AMC Experiments

Several experiments were designed to explore the effectiveness of the moment-based DNN classifier. 8th degree series expansion coefficients ( $h_{p,q}, p + q \leq 8$ ) were calculated from the empirical moments, and simulations are run for a various number of received symbols  $L$ . In all of the experiments, all of the QAMs are rectangular.

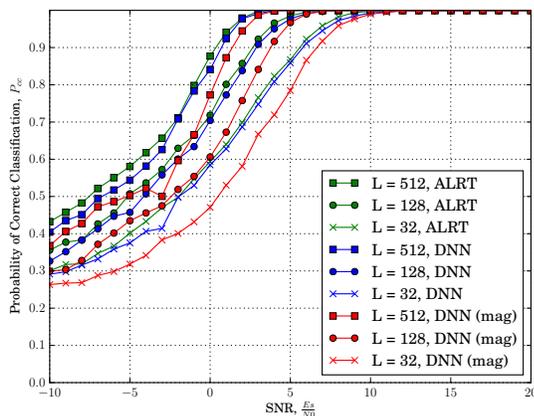


Figure 3. 4-Class AMC Performance

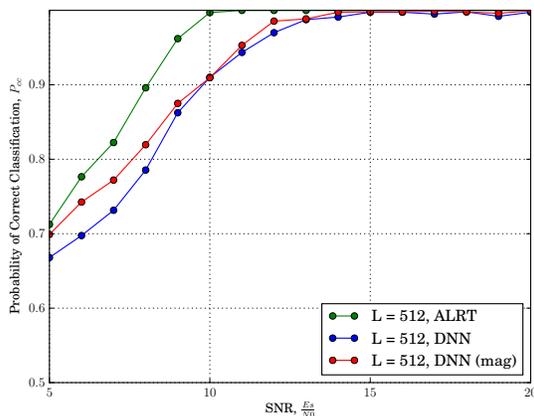


Figure 4. 16QAM vs. 64QAM performance

### 3.2.1. 7-CLASS AMC PROBLEM

In the 7-class AMC experiment, DNNs were trained for discrimination between 7 modulations: 2ASK, 4ASK, BPSK, QSPK, 8PSK, 16QAM, and 32QAM. The overall classification performance is plotted in Figure 2(a), along with a confusion matrix for the experiment with  $L = 128$  and  $\frac{E_s}{N_0} = 3\text{dB}$  in Figure 2(b).

For this experiment, the coherent moment-based DNN classifier performance closely tracks that of ALRT, trailing by only a few percent. Additionally, the magnitude-only DNN trails in performance by several dB. The confusion matrix shows that the most confusing modulation to the classifier is 8PSK, its noisy constellation having similar features to both QPSK and 16QAM. This shows that while the probability of correct classification,  $P_{CC}$ , is a good indication of classifier performance, performance for each modulation class depends on its complexity and similarity to others.

### 3.2.2. 4-CLASS AMC PROBLEM

In the 4-class AMC experiment, the modulations to be discriminated were BPSK, QSPK, 8PSK, and 16QAM. The overall classification performance  $P_{cc}$  is plotted in Figure 3. While the number of classes has decreased, the overall performance  $P_{CC}$  has also decreased because of the increased relative confusion between the modulations of interest; refer to the confusion matrix in Figure 2(b). The 4-class problem represents a more realistic modulation recognition problem, as the efficiencies and benefits of PSK and QAM often outweigh that of ASK.

The moment-based DNN AMC outperforms the cumulant classifiers in (Swami & Sadler, 2000). This is because of

the increased number of moments used in the classifier (leading to better probability density approximation), as well as the non-linear decision-making power of the DNN (as opposed to the static thresholds). This performance comes at the cost of increased computational complexity.

### 3.2.3. 16QAM vs. 64QAM AND OVERFITTING ISSUES

8-class experiments including 64QAM were performed but were not much more interesting than the 7-class experiments, except for high degrees of confusion between 16QAM and 64QAM at low amounts of data.

DNNs were trained and for discrimination between only 16QAM and 64QAM; performance is plotted in Figure 4. Even with 512 received symbols, the classifiers only perform at about 90% at 10 dB SNR. One potential explanation for the difficulty is that the probability density corresponding to the 64QAM constellation requires more data to properly approximate; this notion is supported by the t-SNE plots in Figure 1 and the cluster variances in Table 1.

One additional concern in the 16QAM vs. 64QAM experiment is that the magnitude-only DNN often outperforms the coherent DNN. The reason for this is overfitting in the training process caused by the modulation sub-space not being sampled densely enough. At low amounts of data ( $L = 128, 512$ ), the cluster variances are very large and the classes overlap in the GCA coefficient space. During the training phase, the training accuracy often exceeded the test accuracy by more than several percent. This suggests that performance may be increased by better regularization of the network, by adding and using a holdout training set for cross-validation, or perhaps by significantly increasing the amount of training data.

## 4. Conclusions

In this paper we have introduced the techniques of orthogonal polynomials and Gram-Charlier Approximation to the Automatic Modulation Classification problem, and have shown how moment-based classifiers can nearly match maximum-likelihood techniques at marginal SNR.

This work represents an improvement and exposition on previous moment-based classifiers, and links them to distribution-based classifiers. As the amount of data used in classification increases to infinity, the moment-estimator variance approaches 0 so that performance will very closely approach that of maximum-likelihood methods (depending on how long the GCA series is carried out) even at finite SNR. This rigor comes at the cost of computational complexity.

One way to decrease the computational complexity involved is to compute a smaller subset of the moments involved. It is possible to avoid computation of specific cross-moments,  $E[z^p \bar{z}^q]$ , by subtracting the appropriate Hermite polynomials  $H_{p,q}(z)$  from all of the bases of the series expansion, and re-normalizing. An analysis of the space to determine the most influential cross-moments needs to be conducted in order to make this simplification.

Alternatively, the Hermite polynomials can be used to inform an orthogonalization of a very small set of cross-moments (as in previous moment-based literature), enabling metric-space based algorithms to be employed in a more informed manner.

Finally, these authors fully expect that these techniques can be applied, with slight modification and an appropriate decrease in performance, directly to pre-receiver symbols; more substantial modification will be necessary to cope with non-AWGN channels and noise models.

## 5. Proof of GCA Convergence for Gaussian Mixtures

Here, we will prove that the Gaussian mixture of (1) resides in the Hilbert space  $L^2(\mathbb{C}, \pi e^{z\bar{z}})$ . Expanding the square of the mixture, we have

$$\begin{aligned} & \int_{\mathbb{C}} |f_Z(z|S_m)|^2 \pi e^{z\bar{z}} dz \\ &= \int_{\mathbb{C}} \left| \sum_{s_k \in S_m} \frac{1}{\pi N_0 |S_m|} e^{-\frac{|z-s_k|^2}{N_0}} \right|^2 \pi e^{z\bar{z}} dz \\ &= \int_{\mathbb{C}} \sum_{s_j, s_k \in S_m} \frac{1}{\pi^2 N_0^2 |S_m|^2} e^{-\frac{|z-s_j|^2 - |z-s_k|^2}{N_0}} \pi e^{z\bar{z}} dz \\ &\propto \sum_{s_j, s_k \in S_m} \int_{\mathbb{C}} e^{-\frac{|z-s_j|^2 - |z-s_k|^2}{N_0}} e^{z\bar{z}} dz. \end{aligned}$$

It is sufficient to prove that  $\int_{\mathbb{C}} e^{-\frac{|z-s_j|^2 - |z-s_k|^2}{N_0}} e^{z\bar{z}} dz < \infty$  for all  $s_j, s_k \in S_m$ . Expanding the numerator of the exponent,

$$\begin{aligned} & \int_{\mathbb{C}} e^{-\frac{|z-s_j|^2 - |z-s_k|^2}{N_0}} e^{z\bar{z}} dz \\ &= \int_{\mathbb{C}} e^{-\frac{(2|z|^2 + |s_j|^2 + |s_k|^2 - 2\Re[\bar{s}_j z + \bar{s}_k z])}{N_0}} e^{z\bar{z}} dz \\ &= \int_{\mathbb{C}} e^{-\frac{(2|z|^2 + |s_j|^2 + |s_k|^2 - 2\Re[\bar{s}_j z + \bar{s}_k z]) + N_0 |z|^2}{N_0}} dz \end{aligned}$$

And this integral converges for all  $N_0 < 2$ . In an intuitive sense, this means the GCA can represent a skinnier Gaussian  $f_Z(z)$  (or mixture thereof) from a wider one  $\phi(z)$ . On the other hand, if the density to estimate is too wide the series will fail to converge; the polynomials cannot approximately “upweight” the tails of the base Gaussian  $\phi(z)$  by an exponential amount across the entire support  $\mathbb{C}$ .

## References

- Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, Corrado, Greg S., Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Goodfellow, Ian, Harp, Andrew, Irving, Geoffrey, Isard, Michael, Jia, Yangqing, Jozefowicz, Rafal, Kaiser, Lukasz, Kudlur, Manjunath, Levenberg, Josh, Mané, Dan, Monga, Rajat, Moore, Sherry, Murray, Derek, Olah, Chris, Schuster, Mike, Shlens, Jonathon, Steiner, Benoit, Sutskever, Ilya, Talwar, Kunal, Tucker, Paul, Vanhoucke, Vincent, Vasudevan, Vijay, Viégas, Fernanda, Vinyals, Oriol, Warden, Pete, Wattenberg, Martin, Wicke, Martin, Yu, Yuan, and Zheng, Xiaoqiang. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <http://tensorflow.org/>. Software available from tensorflow.org.
- Aslam, M. W., Zhu, Z., and Nandi, A. K. Automatic digital modulation classification using genetic programming with k-nearest neighbor. In *MILITARY COMMUNICATIONS CONFERENCE, 2010 - MILCOM 2010*, pp. 1731–1736, Oct 2010. doi: 10.1109/MILCOM.2010.5680232.
- Dobre, O. A., Bar-Ness, Y., and Su, Wei. Robust QAM modulation classification algorithm using cyclic cumulants. In *Wireless Communications and Networking Conference, 2004. WCNC. 2004 IEEE*, volume 2, pp. 745–748 Vol.2, March 2004. doi: 10.1109/WCNC.2004.1311279.
- Dunkl, Charles F and Xu, Yuan. *Orthogonal polynomials of several variables*. Number 155. Cambridge University Press, 2014.

- Itô, Kiyosi. Complex multiple wiener integral. In *Japanese journal of mathematics: transactions and abstracts*, volume 22, pp. 63–86. The Mathematical Society of Japan, 1952.
- Kendall, M.G. and Stuart, A. *The Advanced Theory of Statistics: Distribution theory*. The Advanced Theory of Statistics. Hafner Publishing Company, 1969. ISBN 9780852641415.
- Maaten, Laurens van der and Hinton, Geoffrey. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.
- O’Shea, Timothy J, Corgan, Johnathan, and Clancy, T. Charles. Convolutional radio modulation recognition networks. *arXiv preprint arXiv:1602.04105*, 2016.
- Reichert, J. Automatic classification of communication signals using higher order statistics. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 5, pp. 221–224 vol.5, Mar 1992. doi: 10.1109/ICASSP.1992.226530.
- Sauer, P. and Heydt, G. A convenient multivariate Gram-Charlier type a series. *IEEE Transactions on Communications*, 27(1):247–248, Jan 1979. ISSN 0090-6778. doi: 10.1109/TCOM.1979.1094247.
- Sills, J. A. Maximum-likelihood modulation classification for PSK/QAM. In *Military Communications Conference Proceedings, 1999. MILCOM 1999. IEEE*, volume 1, pp. 217–220 vol.1, 1999. doi: 10.1109/MILCOM.1999.822675.
- Spooner, C. M. On the utility of sixth-order cyclic cumulants for RF signal classification. In *Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on*, volume 1, pp. 890–897 vol.1, Nov 2001. doi: 10.1109/ACSSC.2001.987051.
- Spooner, C. M., Brown, W. A., and Yeung, G. K. Automatic radio-frequency environment analysis. In *Signals, Systems and Computers, 2000. Conference Record of the Thirty-Fourth Asilomar Conference on*, volume 2, pp. 1181–1186 vol.2, Oct 2000. doi: 10.1109/ACSSC.2000.910700.
- Su, W. Feature space analysis of modulation classification using very high-order statistics. *IEEE Communications Letters*, 17(9):1688–1691, September 2013. ISSN 1089-7798. doi: 10.1109/LCOMM.2013.080613.130070.
- Swami, A. and Sadler, B.M. Hierarchical digital modulation classification using cumulants. *Communications, IEEE Transactions on*, 48(3):416–429, Mar 2000. ISSN 0090-6778. doi: 10.1109/26.837045.
- Szego, Gabor. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.
- Wang, F. and Chan, C. Variational-distance-based modulation classifier. In *2012 IEEE International Conference on Communications (ICC)*, pp. 5635–5639, June 2012. doi: 10.1109/ICC.2012.6364879.
- Wang, F. and Wang, X. Fast and robust modulation classification via Kolmogorov-Smirnov test. *IEEE Transactions on Communications*, 58(8):2324–2332, August 2010. ISSN 0090-6778. doi: 10.1109/TCOMM.2010.08.090481.
- Zhu, Z., Nandi, A. K., and Aslam, M. W. Robustness enhancement of distribution based binary discriminative features for modulation classification. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Sept 2013. doi: 10.1109/MLSP.2013.6661989.