

Discrimination on the Grassmann Manifold: Fundamental Limits of Subspace Classifiers

Matthew Nokleby*, Miguel Rodrigues[†], and Robert Calderbank*

*Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA

[†]Department of Electronic and Electrical Engineering, University College London, UK

Email: {matthew.nokleby, robert.calderbank}@duke.edu, m.rodrigues@ucl.ac.uk

Abstract—Repurposing tools and intuitions from Shannon theory, we derive fundamental limits on the reliable classification of high-dimensional signals from low-dimensional features. We focus on the classification of linear and affine subspaces and suppose the features to be noisy linear projections. Leveraging a syntactic equivalence of discrimination between subspaces and communications over vector wireless channels, we derive asymptotic bounds on classifier performance. First, we define the *classification capacity*, which characterizes necessary and sufficient relationships between the signal dimension, the number of features, and the number of classes to be discriminated, as all three quantities approach infinity. Second, we define the *diversity-discrimination tradeoff*, which characterizes relationships between the number of classes and the misclassification probability as the signal-to-noise ratio approaches infinity. We derive inner and outer bounds on these measures, revealing precise relationships between signal dimension and classifier performance.

I. INTRODUCTION

The classification of high-dimensional signals arises in practical applications, from face and digit recognition to tumor classification [1]–[3]. In order to reduce the computational burden of classification, it is common to construct first a low-dimensional representation of the signal to be classified, a process often called *feature extraction*. Standard techniques include linear discriminant analysis (LDA) and principal component analysis (PCA) [4], as well as their myriad variations. In principle, feature extraction degrades classifier performance, and fewer features extracted leads to more classification errors.

In this paper, we present a precise, information-theoretic treatment of the relationship between the number of features extracted and classifier error performance. We consider N -dimensional signals, drawn from k -dimensional linear and affine subspaces, which are classified from M linear features corrupted by Gaussian noise. To characterize classifier performance, we define two performance measures inspired by Shannon theory:

- The *classification capacity*, which characterizes the number of unique classes that can be discerned, as a function of the number of linear features taken, in the limit of high signal dimension. Just as the usual Shannon capacity captures the phase transition of the error probability as a function of the information rate as the code length goes

to infinity, the classification capacity captures the phase transition of the misclassification probability in terms of the (logarithm of the) number of classes, normalized by the number of measurements as the signal dimension goes to infinity.

- The *diversity-discrimination tradeoff* (DDT), which characterizes the relationship between the number of classes and the misclassification probability in the limit of high SNR. Just as the diversity-multiplexing tradeoff from vector wireless channels [5] defines the number of codewords and the error probability in terms of a region of achievable exponent pairs in the SNR, the DDT defines a region of achievable exponent pairs for the number of classes and the misclassification probability.

Much of our analysis owes to a syntactic equivalence between the classification of high-dimensional signals and the communication of data over vector wireless channels. In particular, as we identified in an earlier work [6], the classification of subspaces is nearly equivalent to communication over non-coherent vector channels. The capacity, as well as capacity-achieving techniques, are known [7]–[9], and we leverage this knowledge in bounding the classification capacity and DDT. Similarly, the classification of affine spaces is related to communication over coherent vector channels, and we develop novel adaptations of existing techniques [5], [10] to derive our results.

In Section II we specify the signal model, introduce the classification problem, and define the classification capacity and the diversity-discrimination tradeoff. In Section III we discuss the classification of linear subspaces, presenting bounds on the classification capacity and DDT. In Section IV we discuss the classification of *affine* subspaces, again presenting capacity and DDT bounds. In Section V we examine classifier performance numerically, focusing on regimes in which the capacity and DDT bounds are not tight. Section VI concludes the paper.

Due to space limitations, we only sketch the proofs. A full version of this work can be found at [11].

II. PRELIMINARIES

A. System Model

We suppose the classifier operates on noisy, linear projections of the signal of interest. These projections, denoted by

The work of Matthew Nokleby and Robert Calderbank was supported in part by Air Force Office of Scientific Research grant FA 9550-13-01-0076 under the Complex Networks Program. This work also was supported in part by the Royal Society International Exchanges Scheme IE120996.

$\mathbf{y} \in \mathbb{R}^M$, are related to the signal $\mathbf{x} \in \mathbb{R}^N$ as follows:

$$\mathbf{y} = \Phi \mathbf{x} + \mathbf{z}, \quad (1)$$

where $\Phi \in \mathbb{R}^{M \times N}$ is a random measurement matrix describing the linear features, and $\mathbf{z} \in \mathbb{R}^M$ is white Gaussian noise having mean zero and covariance $\sigma^2 \mathbf{I}_{M \times M}$ for some $\sigma^2 > 0$. We suppose $M \leq N$ throughout.

Define $\text{SNR} = 1/\sigma^2$. To ensure that the SNR remains meaningful, both the signal of interest and the measurement matrix must obey energy constraints:

$$\mathbb{E}[\|\Phi\|^2] \leq 1, \mathbb{E}[\|\mathbf{x}\|^2] \leq M, \quad (2)$$

where $\|\cdot\|$ denotes the Euclidean norm and represents the induced operator norm when applied to a matrix. We further constrain the distribution of Φ to be invariant to rotations; that is $p(\Phi) = p(\Phi \mathbf{V})$ for all unitary $\mathbf{V} \in \mathbb{R}^{N \times N}$.

The classification task is to identify, from the feature vector $\mathbf{y}$, the probability distribution from which $\mathbf{x}$ arose. Formally, we suppose that $\mathbf{x} \sim q$, where the distribution q is an element drawn uniformly from the *class alphabet* $\mathcal{P}$. The class alphabet contains the distributions among which $\mathbf{x}$ is to be discriminated. We suppose the classifier knows $\mathcal{P}$ perfectly. Upon obtaining $\mathbf{y}$, the classifier generates an estimate $\hat{q} \in \mathcal{P}$ of the underlying distribution. The average misclassification probability is

$$P_e = \frac{1}{L} \sum_{q \in \mathcal{P}} \Pr(\hat{q} \neq q | \mathbf{x} \sim q), \quad (3)$$

where $L = |\mathcal{P}|$. Because Φ is random, P_e is computed according to the distribution $p(\Phi)$.

The classification capacity and DDT characterize the asymptotic error performance maximized over a particular family of classification problems. To describe these families, we define the *constraint sets* $\mathcal{Q}$, which specify the permissible distributions in a classification task. Thus $\mathcal{P} \subset \mathcal{Q}$. In this paper we consider two families of classification problems, each defined by a different constraint set.

In Section III, we study classifier performance over low-dimensional linear subspaces. We describe the subspaces via zero-mean Gaussian distributions having the form

$$q = \mathcal{N}(\mathbf{0}, \mathbf{U}\mathbf{U}^T), \quad (4)$$

where $\mathbf{U} \in \mathbb{R}^{N \times k}$ is a matrix whose column space is the k -dimensional subspace associated with the class. Imposing the energy constraint on these distributions, and ensuring that the subspaces have k non-trivial dimensions, the associated constraint set is

$$\mathcal{Q}_{\text{linear}} = \{q : E_q[\|\mathbf{x}\|^2] \leq M, q = \mathcal{N}(\mathbf{0}, \mathbf{U}\mathbf{U}^T), \mathbf{U} \in \mathbb{R}^{N \times k}, \alpha \leq \text{eig}(\mathbf{U}^T \mathbf{U}) \leq M/k\}, \quad (5)$$

for fixed $0 < \alpha \leq M/k$, where $\text{eig}(\cdot)$ returns the vector of eigenvalues, and where the final inequality holds element-wise.

As mentioned in the introduction, the analysis for linear subspaces has antecedents in capacity results for non-coherent vector channels. To see this, we rewrite the measurement

model in (1) when the classes conform to (5), which yields

$$\mathbf{y} = \Phi \mathbf{U} \mathbf{h} + \mathbf{z}, \quad (6)$$

where $\mathbf{h} \in \mathbb{R}^k$ is i.i.d. Gaussian with mean zero and unit variance. We can think of $\mathbf{h}$ as a Gaussian process that “drives” the subspace associated with each class. Taking the transpose of (6) yields

$$\mathbf{y}^T = \mathbf{h}^T \mathbf{U}^T \Phi^T + \mathbf{z}^T.$$

Comparing with the system models of [7], [8], one sees that, except for the presence of Φ and the real-valued channel model, the preceding equation is identical to a non-coherent channel having k transmit antennas, a single receive antenna, and a coherence time of M . The driving force $\mathbf{h}$ takes the place of the unknown channel matrix, and $\mathbf{U}^T \Phi^T$ takes the place of the codeword. In Section III we exploit this equivalence to prove bounds on the classification capacity and the DDT.

In Section IV, we study classifier performance over low-dimensional *affine* subspaces, or linear subspaces that are translated by a non-zero point in $\mathbb{R}^N$. We again describe the subspaces via Gaussian distributions, but in this case the distributions have nonzero means corresponding to the translation of the linear subspace:

$$q = \mathcal{N}(\mu, \mathbf{U}\mathbf{U}^T). \quad (7)$$

As before $\mathbf{U} \in \mathbb{R}^{N \times k}$ defines a k -dimensional subspace, and $\mu \in \mathbb{R}^N$ specifies its translation. The associated constraint set is

$$\mathcal{Q}_{\text{affine}} = \{q : E_q[\|\mathbf{x}\|^2] \leq M, q = \mathcal{N}(\mu, \mathbf{U}\mathbf{U}^T), \mu \in \mathbb{R}^N, \mathbf{U} \in \mathbb{R}^{N \times k}, \alpha \leq \text{eig}(\mathbf{U}^T \mathbf{U}) \leq M/k\}, \quad (8)$$

where again $0 < \alpha \leq M/k$.

B. Classification Capacity

By analogy with the sequence of codebooks defined under Shannon theory, we characterize the classification capacity in terms of a sequence of class alphabets. Fix a sequence of distribution constraint sets $\{\mathcal{Q}^{(M)}\}$ indexed by the number of measurements M . Let $\{\mathcal{P}^{(M)}\}$ denote a sequence of class alphabets, where each alphabet $\mathcal{P}^{(M)} \subset \mathcal{Q}^{(M)}$ contains $L^{(M)}$ distributions (or classes) with support on $N^{(M)}$ -dimensional space. Each class corresponds to a $k^{(M)}$ -dimensional linear or affine subspace. We keep α fixed.

The classification capacity characterizes classifier performance as M , $N^{(M)}$, $k^{(M)}$ and $L^{(M)}$ approach infinity. We suppose the preceding quantities scale as follows. First,

$$\lim_{M \rightarrow \infty} \frac{N^{(M)}}{M} = \nu, \quad \lim_{M \rightarrow \infty} \frac{k^{(M)}}{M} = \kappa, \quad (9)$$

meaning the ambient signal dimension and subspace dimension scale linearly in M . Also,

$$\lim_{M \rightarrow \infty} \frac{\log(L^{(M)})}{M} = \rho, \quad (10)$$

which implies that the number of classes grows exponentially in M . By analogy with communications theory, the quantity ρ can be interpreted as the “rate” of the sequence of class

alphabets, or the average number of bits discerned by the classifier per measurement. Finally, let $P_e^{(M)}$ denote the average misclassification probability associated with the class alphabet $\mathcal{P}^{(M)}$.

Definition 1: Fix the sequence of constraint sets $\{\mathcal{Q}^{(M)}\}$. We say that the rate ρ is *achievable* if there exists a sequence of alphabets $\{\mathcal{P}^{(M)}\}$ such that $\mathcal{P}^{(M)} \subset \mathcal{Q}^{(M)}$, $\rho^{(M)} = \frac{\log(L^{(M)})}{M} \rightarrow \rho$, and $P_e^{(M)} \rightarrow 0$ as $M \rightarrow \infty$.

Definition 2: The *classification capacity*, denoted by C , is the supremum over achievable ρ .

We can bound the classification capacity via the mutual information between the classes and the feature vector.

Lemma 1: Fix the sequence of distribution constraint sets $\{\mathcal{Q}^{(M)}\}$. When the following limit exists, the classification capacity satisfies

$$C \leq \lim_{M \rightarrow \infty} \sup_{\substack{p(\Phi), p(q) \\ E[\|\Phi\|^2] \leq 1, q \in \mathcal{Q}^{(M)}}} \frac{I(q; \mathbf{y})}{M}, \quad (11)$$

where $I(q; \mathbf{y})$ is the mutual information between the classes and the feature vector, and where we treat the classes as a random variable distributed according to $p(q)$.

Proof sketch: The lemma is an immediate consequence of Fano's inequality. ■

C. Diversity-Discrimination Tradeoff

The *diversity-multiplexing tradeoff* (DMT) was introduced in the context of wireless communications to characterize the high-SNR performance of fading coherent MIMO channels. The spatial flexibility afforded by multiple antennas can increase simultaneously the information rate and decrease the probability of error, and [5] characterizes the tradeoff between these two desiderata. We define a similar characterization in the context of classification, capturing the relationship between the increase of discernible classes and the decay of misclassification probability as the SNR approaches infinity.

As before, we define a sequence of classification alphabets, but here the signal dimensions remain fixed while the SNR goes to infinity. Fix a distribution constraint set $\mathcal{Q}$. Let $\{\mathcal{P}^{(\text{SNR})}\}$ be a sequence of class alphabets indexed by the SNR, where each alphabet $\mathcal{P}^{(\text{SNR})} \subset \mathcal{Q}$, and let $P_e^{(\text{SNR})}$ be the average misclassification probability associated with the class alphabet $\mathcal{P}^{(\text{SNR})}$. The dimensions N , M , and k remain fixed, but the number of classes $L^{(\text{SNR})}$ can vary with the SNR. We define the discrimination gain, the diversity gain, and finally the diversity-discrimination tradeoff as follows.

Definition 3: Fix the distribution constraint set $\mathcal{Q}$. A sequence of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$, where $\mathcal{P}^{(\text{SNR})} \subset \mathcal{Q}$, is said to have discrimination gain r and diversity gain d if

$$\lim_{\text{SNR} \rightarrow \infty} \frac{\log(L^{(\text{SNR})})}{\frac{1}{2} \log(\text{SNR})} \geq r \quad (12)$$

and

$$\lim_{\text{SNR} \rightarrow \infty} -\frac{\log(P_e^{(\text{SNR})})}{\frac{1}{2} \log(\text{SNR})} \geq d. \quad (13)$$

We say the sequence of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$ has diversity-discrimination function $d(r)$.

In other words, the sequence of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$ has approximately $\text{SNR}^{r/2}$ classes and a misclassification probability of approximately $\text{SNR}^{-d/2}$ at high SNR.

Definition 4: The *diversity-discrimination tradeoff* is defined as the supremum over the DDT functions of all permissible sequences of class alphabets $\{\mathcal{P}^{(\text{SNR})}\}$, where $\mathcal{P}^{(\text{SNR})} \subset \mathcal{Q}$, or

$$d^*(r) = \sup_{\{\mathcal{P}^{(\text{SNR})}\}, \mathcal{P}^{(\text{SNR})} \subset \mathcal{Q}} d(r). \quad (14)$$

III. LINEAR SUBSPACES

Here we characterize classifier performance over low-dimensional subspaces represented by zero-mean Gaussian distributions having low-rank covariance matrices. We prove upper and lower bounds on the classification capacity and the DDT, showing that the relationship between the number of classes to discriminate and the misclassification error depends in precise ways on the subspace dimension, the number of linear features acquired, and the SNR.

Theorem 1: Let $\{\mathcal{Q}_{\text{subspace}}^{(M)}\}$ be a sequence of distribution constraint sets, where $\mathcal{Q}_{\text{subspace}}^{(M)}$ is defined in (5), with $k(M)/M$ and $N(M)/M$ approaching $\kappa \leq 1$ and ν as shown in (9) and with fixed $0 < \alpha \leq 1/\kappa$. Then, the classification capacity is

$$\begin{aligned} \frac{\min\{\kappa, 1 - \kappa\}}{2} \log(\text{SNR}) + O(1) &\leq C \\ &\leq \frac{1 - \kappa}{2} \log(\text{SNR}) + O(1). \end{aligned} \quad (15)$$

Proof sketch: To upper bound the capacity, we bound the mutual information $I(q; \mathbf{y})$ and invoke Lemma 1. Exploiting the fact that Φ is invariant to rotations, we bound the mutual information in terms of the log-determinant of Wishart matrices, from which we show that the mutual information obeys

$$I(q; \mathbf{y}) \leq \frac{M - k}{2} \log(\text{SNR}) + O(M).$$

Invoking Lemma 1 establishes the outer bound.

To lower bound the capacity, we construct a random ensemble of classes in which each $\mathbf{U}_l \in \mathbb{R}^{N \times k}$ has elements drawn i.i.d. from the standard Gaussian distribution. We invoke the Bhattacharya bound to estimate the pairwise misclassification probability between Gaussian classes. When $0 \leq \kappa \leq 1/2$, this error is approximately $2^{-\frac{\kappa M}{2} \log(\text{SNR})}$; when $1/2 \leq \kappa < 1$, this error is approximately $2^{-\frac{(1-\kappa)M}{2} \log(\text{SNR})}$. Taking the union bound over approximately $2^{\rho M}$ classes establishes the lower bound and completes the proof. ■

Theorem 2: Let $\mathcal{Q}_{\text{subspace}}$ be a constraint set defined as in (5), with $0 < \alpha \leq 1/\kappa$. Then, the DDT is upper bounded by

$$d(r) \leq k \left[1 - \frac{r}{M - k} \right]^+. \quad (16)$$

The DDT is bounded below by

$$d(r) \geq [\min\{k - r, M - k - r\}]^+. \quad (17)$$

Proof sketch: To prove the upper bound, we adapt the outage-based argument of [9]; when $\|\mathbf{h}\|^2$ is small, the mutual information is low, and by Fano's inequality the probability

of error remains nonzero as $\text{SNR} \rightarrow \infty$. To prove the inner bound, we again employ the Gaussian i.i.d. ensemble from the proof of Theorem 1; direct analysis of the misclassification probability yields the result. ■

The upshot of Theorems 1 and 2 is that if the number of classes scales faster than $\text{SNR}^{\frac{M-k}{2}}$ —whether in the sense of increasing signal dimension or SNR—the error probability remains bounded away from zero. For $k \geq M/2$, there exists an ensemble of classes such that, when the number of classes scales slower than $\text{SNR}^{\frac{M-k}{2}}$, the error approaches zero. Furthermore, in Section V we show empirically that, even when $k < M/2$, the misclassification probability decays in the SNR if the number of classes scales slower than $\text{SNR}^{\frac{M-k}{2}}$. We thus conjecture that the outer bound on the classification capacity is tight.

IV. AFFINE SUBSPACES

Next we characterize classifier performance over low-rank affine subspaces represented by Gaussian distributions having low-rank covariances but nonzero mean. We again prove upper and lower bounds on the classification capacity and DDT. In this case we can derive nearly-matching upper and lower bounds on the classification capacity. However, it is difficult to define a useful “outage” event as we did in the case of linear subspaces, so we derive only a trivial outer bound on the DDT.

Theorem 3: Let $\{\mathcal{Q}_{\text{affine}}^{(M)}\}$ be a sequence of constraint sets, with $\mathcal{Q}_{\text{affine}}^{(M)}$ as defined in (8), with $k(M)/M$ and $N(M)/M$ approaching $0 \leq \kappa \leq \nu$ and ν as shown in (9), and with fixed $0 < \alpha < 1/\kappa$. Then, the classification capacity satisfies

$$C = \frac{1-\kappa}{2} \log(\text{SNR}) + O(1). \quad (18)$$

Proof sketch: To prove the upper bound, we observe that the entropy of a Gaussian distribution is independent of the mean. Therefore, the calculation of the mutual information from the proof of Theorem 1 applies directly to the classification of affine subspaces.

To prove the lower bound, we consider a random ensemble of classes. Each class shares a common covariance matrix of rank k , whereas each class has a unique mean drawn i.i.d. from a Gaussian distribution. The received signal is then Gaussian-distributed points corrupted by noise and occluded by k -dimensional self-noise. Again invoking the Bhattacharya bound, we find that the pairwise misclassification probability is approximately $2^{-\frac{(1-\kappa)M}{2} \log(\text{SNR})}$. Taking the union bound over $2^{\rho M}$ classes shows that the error probability goes to zero when $\rho < \frac{1-\kappa}{2} \log(\text{SNR})$, as was claimed. ■

A. Diversity-Discrimination Tradeoff

Theorem 4: Let $\mathcal{Q}_{\text{affine}}$ be a constraint set defined as in (8), with $0 < \alpha \leq 1/\kappa$. Then, the DDT is upper bounded by

$$d(r) \leq \begin{cases} \infty, & r \leq M - k \\ 0 & r \geq M - k \end{cases}, \quad (19)$$

and it is lower bounded by

$$d(r) \geq [M - k - r]^+. \quad (20)$$

Proof: If the discrimination gain is higher than $M - k$, then the mutual information is too small relative to the number of classes, and Fano’s inequality guarantees errors with nonzero probability. This proves the outer bound.

To prove the inner bound, we analyze the same random ensemble of classes used in Theorem 3. Observing that the pairwise error probability is approximately $2^{-\frac{(1-\kappa)M}{2} \log(\text{SNR})} = \text{SNR}^{-\frac{M-k}{2}}$, we take the union bound over $\text{SNR}^{\frac{r}{2}}$ classes, which yields the result. ■

As before, the upshot is that the number of classes cannot scale any faster than $\text{SNR}^{\frac{M-k}{2}}$ without incurring nonzero misclassification error. Furthermore, provided the number of classes scales slower than $\text{SNR}^{\frac{M-k}{2}}$, the error decays to zero in both the SNR and in M . The capacity-achieving ensemble has a single covariance common to the classes, while the discrimination power resides in the classes’ separate means. This ensemble is similar to the assumptions behind LDA.

Theorem 4 leaves open the possibility that a better ensemble of classes might achieve better decay of misclassification probability in the SNR. In Section V we evaluate this question empirically. We choose means drawn from a lattice, rather than a random codebook, in order to increase the distance between means. However, we find that this ensemble of classes does not achieve error performance any better than the lower bound of Theorem 4. We therefore believe the lower bound is tight.

V. NUMERICAL RESULTS

A. Linear Subspaces

Here we examine the empirical performance of classifiers over linear subspaces. Our particular aim is to see whether the ensemble of i.i.d. Gaussian classes, described in the proof of Theorem 1 achieves near-capacity performance for $\kappa < \frac{1}{2}$. Due to computational limitations, it is difficult to examine empirical performance as the signal dimension grows large, so instead of testing the classification capacity directly, we examine the DDT performance. Noting that the maximum discrimination gain corresponds to the prelog of the classification capacity, we conjecture that when an ensemble has nonzero diversity gain at r , it achieves classification rates $\rho \approx \frac{r}{2} \log_2(\text{SNR})$.

We choose subspaces described by matrices $\mathbf{U}_l \in \mathbb{R}^{N \times k}$ having i.i.d. standard Gaussian entries, and we choose $\Phi \in \mathbb{R}^{M \times N}$ to be drawn at random from the appropriate Stiefel manifold. Finally, we use the maximum-likelihood classifier on the linear features, which chooses the class that minimizes the Mahalanobis distance to the feature vector. First, we select $N = 5$, $M = 3$, $k = 1$, and discrimination gains $r \in \{0, 0.75, 1.5, 1.8\}$. In Figure 1 we plot the misclassification probability as a function of the SNR, averaged over 10^5 realizations. We also plot the high-SNR error slopes predicted by the outer bound of Theorem 2. We observe decaying misclassification probability for all values of r ; furthermore, we observe rates of decay consistent with the outer bound of the DDT. We conjecture that, in the regime $\kappa < \frac{1}{2}$, the outer bounds of both the classification capacity and the DDT are tight. If so, the classification capacity outer bound is tight in all regimes.

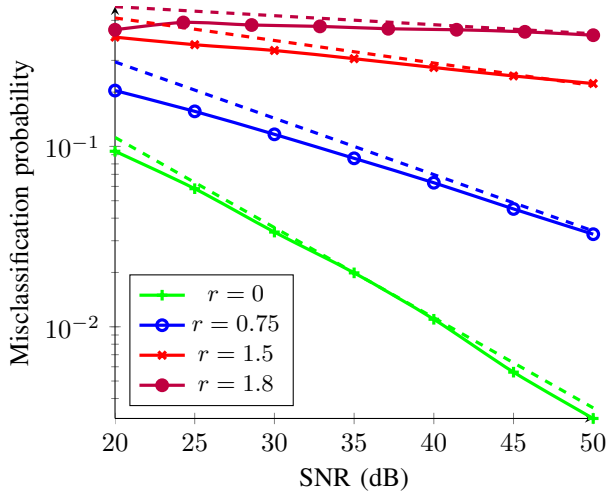


Fig. 1. Misclassification probability vs. SNR for linear subspaces. Dashed lines indicate high-SNR slopes predicted by the DDT upper bound from Theorem 2.

B. Affine Subspaces

Next we examine the empirical performance of classifiers over affine subspaces. In particular, we wish to address experimentally the gap between the inner and outer bounds on the DDT. In an effort to improve diversity performance over the random ensemble proposed in Theorem 4, we choose the means of each class from the integer lattice in $\mathbb{R}^N$, scaling them in order to meet the power constraint. This choice guarantees a minimum separation of the means deterministically, rather than the average separation provided by the i.i.d. Gaussian ensemble. As before we choose a common, arbitrarily-chosen k -dimensional subspace. Finally, we choose Φ randomly from the appropriate Stiefel manifold.

We choose $N = 4$, $M = 2$, $k = 1$, and discrimination gains $r \in \{0, 0.5, 0.9\}$. In Figure 2 we plot the misclassification probability as a function of the SNR, averaged over 10^5 realizations. The error slopes are no better than the high-SNR slopes predicted by the lower bound of Theorem 4. Thus, while it is possible that we have simply studied a suboptimal ensemble of classes, we conjecture that the inner bound on the DDT is tight.

VI. CONCLUSION

We have studied fundamental limits associated with the classification of linear and affine spaces from noisy, linear features. Our approach is inspired by a connection between the information theory of wireless channels and the information theory of this machine learning domain, tracing back to a formal equivalence between the identification of codewords over non-coherent channels and the classification of low-dimensional subspaces represented by low-rank Gaussian distributions. We have introduced fundamental performance limits reminiscent of wireless information theory: the classification capacity and the diversity-discrimination tradeoff, which govern classifier

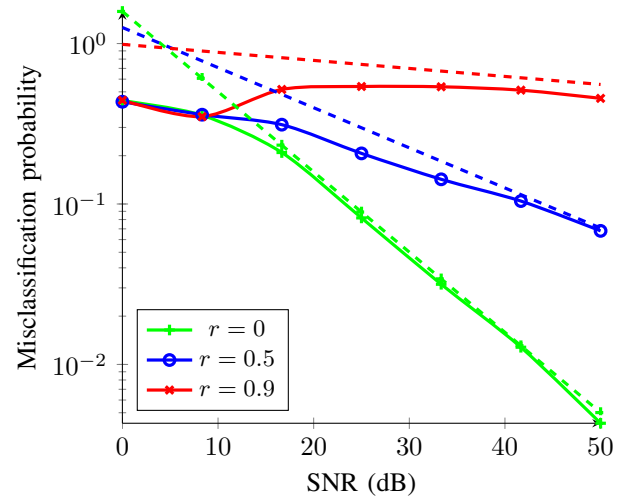


Fig. 2. Misclassification probability vs. SNR for affine subspaces. Dashed lines indicated the high-SNR slopes predicted by the DDT lower bound from Theorem 4.

performance in the asymptotic regimes of large signal dimensionality and high SNR, respectively. We proved inner and outer bounds on these quantities, and further showed that these bounds match in many regimes. Our analysis revealed that random ensembles of subspaces drawn from the Grassmann manifold are capacity and DDT optimal in these regimes. Furthermore, via numerical evaluation, we conjectured that these ensembles are optimal even in regimes in which the inner and outer bounds are not tight.

REFERENCES

- [1] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.
- [2] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994.
- [3] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, vol. 24, no. 3, pp. 227–235, 2000.
- [4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [5] L. Zheng and D. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.
- [6] M. Nokleby, M. Rodrigues, and R. Calderbank, "Information-theoretic limits on the classification of Gaussian mixtures: Classification on the Grassmann manifold," in *Proc. Information Theory Workshop (ITW)*, 2013.
- [7] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Info. Theory*, vol. 45, no. 1, pp. 139–157, 1999.
- [8] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Info. Theory*, vol. 48, no. 2, pp. 359–383, 2002.
- [9] —, "The diversity-multiplexing tradeoff for non-coherent multiple antenna channels," in *Proc. Allerton*, vol. 40, no. 2, 2002, pp. 834–843.
- [10] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.
- [11] M. Nokleby, M. Rodrigues, and R. Calderbank, "Discrimination on the Grassmann manifold: Fundamental limits of subspace classifiers," *submitted to IEEE Trans. Info. Theory*, <http://arxiv.org/abs/1404.5187>.