An Impossibility Result for High Dimensional Supervised Learning

M. H. Rohban, P. Ishwar, B. Orten†, W. C. Karl, and V. Saligrama
ECE Department Boston University, †Turn, Inc. CA USA
Email: {mhrohban,pi,srv,wckarl}@bu.edu, burkay.orten@turn.com

***Abstract*—We study high-dimensional asymptotic performance limits of binary supervised classification problems where the class conditional densities are Gaussian with unknown means and covariances and the number of signal dimensions scales faster than the number of labeled training samples. We show that the Bayes error, namely the minimum attainable error probability with complete distributional knowledge and equally likely classes, can be arbitrarily close to zero and yet the limiting minimax error probability of every supervised learning algorithm is no better than a random coin toss. In contrast to related studies where the classification difficulty (Bayes error) is made to vanish, we hold it constant when taking high-dimensional limits. In contrast to VC-dimension based minimax lower bounds that consider the worst case error probability over *all* distributions that have a fixed Bayes error, our worst case is over the family of Gaussian distributions with constant Bayes error. We also show that a nontrivial asymptotic minimax error probability can only be attained for parametric subsets of zero measure (in a suitable measure space). These results expose the fundamental importance of prior knowledge and suggest that unless we impose strong structural constraints, such as sparsity, on the parametric space, supervised learning may be ineffective in high dimensional small sample settings.**

I. Introduction

In a number of applications ranging from medical imaging to economics, one encounters inference problems that suffer from the "curse of dimensionality", namely the situation where the observed signals are high-dimensional and we lack sufficient labeled training samples from which to accurately learn models and make reliable decisions. This may be true even when the underlying decision problem becomes "easy" with perfect knowledge of models or latent variables. For example, in detecting the presence or absence of stroke, high-dimensional tomographic X-ray projections are measured, though a stroke may affect only tissue properties in a *localized* spatial region and may be easily detectable if one knew where to look. Labeled training samples are typically limited in this context due to the high cost of engaging domain experts. Research in recent years has therefore focused on leveraging prior knowledge in the form of sparsity or other latent low-dimensional structure to improve decision making.

Our aim in this work is to expose certain fundamental limitations of supervised learning in high dimensional small sample settings and highlight the fundamental necessity of strong structural constraints (prior knowledge), such as sparsity, for attaining nontrivial asymptotic error rates. Towards this end we study the high-dimensional asymptotic performance limit of binary supervised classification where the class conditional densities are Gaussian with unknown means and covariances. If d and n respectively denote the number of signal dimensions and the number of labeled training samples, our focus is on the "big d small n" asymptotic regime where $n/d \to 0$ as $d \to \infty$. In previous related work, either $n/d \to c > 0$ as $d \to \infty$ or the classification difficulty (Bayes error) converges to zero as $d \to \infty$ [1], [2], or the focus was on special families of learning rules such as Naive-Bayes, banded covariance structure [3], and plug-in rules [4], [5]. In contrast, we hold the Bayes error constant when taking high-dimensional limits.

We establish two key results in this paper: 1) An impossibility result: When the number of signal dimensions scales faster than the number of labeled samples at constant classification difficulty, the asymptotic minimax classification error probability of *any supervised classification algorithm* cannot converge to anything less than half. 2) Necessity of "structure" in parameter set: Nontrivial asymptotic minimax error probability is attainable only for parametric subsets of zero Haar measure.

II. Related results from VC Theory

There are several well known probabilistic lower bounds for the error probability in the *distribution-free* setting of learning [6]. These bounds are typically based on the VC dimension V of a set of classifiers containing the optimal Bayes rule. For the case when the Bayes error P_e^* is zero, it is known that if $n < (V-1)/(32\epsilon)$, then for any supervised classifier g_n and all $\epsilon \leq 1/8$ and $\delta \leq 0.01$, $\sup_{p_{X,Y}:P_e^*=0} \Pr(L_{g_n} \geq \epsilon) \geq \delta$ [7], where $p_{X,Y}$ is the joint distribution of data points and their labels, $L_{g_n} = \Pr(g_n(X) \neq Y | X_1, Y_1, \ldots, X_n, Y_n)$, and $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is the training set. This bound implies that if n is small compared to V, then there exist distributions for which, with probability of at least 0.01, the conditional Bayes error of every classifier is larger than $1/8$. For the case when $P_e^* = c \neq 0$, it is known that if $n < V/(320\epsilon)$ then for all supervised learning rules g_n, and all $\epsilon, \delta \in (0, 1/64)$, $\sup_{p_{X,Y}:P_e^*=c} \Pr(L_{g_n} - c \geq \epsilon) \geq \delta$. For linear classifiers in $\mathbb{R}^d$, $V = (d+1)$ and the mentioned bounds prove the impossibility of learnability in the *distribution-free* high dimensional setting. However, these bounds are known to be pessimistic as they are proved by constructing pathological adversarial distributions $p_{X|Y}$ whose support is concentrated on V points in order to make the error of any learning rule in the hypothesis space larger than ϵ with probability of at least δ. It has been suggested that these bounds do not hold for some practical choices of $p_{X,Y}$ [8].

There also exist impossibility results for the *fixed distribution* setting [9] where a fixed and known distribution p_X is assumed and classifiers are assumed to belong to a

specific set $\mathcal{C}$. For $\Pr(L_{g_n} \leq \epsilon) \geq 1-\delta$ it is *necessary* that $n \geq \log((1-\delta)n_c)$ where n_c is the ϵ covering number of $\mathcal{C}$ with respect to p_X [9].

The learning scenario discussed in this paper differs from both these settings. We consider $p_{X|Y,\theta}$ to be a *Gaussian distribution with an unknown set of parameters* θ. Hence, unlike the second setting, p_X is not a fixed distribution, but belongs to a *family* of Gaussian distributions parameterized by different choices of θ. However, this family is much more restricted than the set of *all* distributions, which is assumed in the distribution-free setting. In addition, we establish a stronger notion of impossibility that corresponds to taking $\epsilon = \frac{1}{2}$ and $\delta = 1$ asymptotically, i.e., the worst-case error probability within the family is not less than half asymptotically. It should also be noted that our notion of impossibility is not available in the distribution-free setting, and additional effort is required to establish it in the fixed-distribution setting.

III. Problem Formulation

Binary classification with Gaussian class conditional densities: Let $X \in \mathbb{R}^d$ denote the observed signal (data), $Y \in \{-1,+1\}$ the latent binary class label with $p_Y(+1) = p_Y(-1) = 1/2$, and $p_{X|Y,\theta}(x|y,\theta) = \mathcal{N}(\mu_y, \Sigma)(x)$, where $\mu_{+1}, \mu_{-1} \in \mathbb{R}^d$, $\mu_{+1} \neq \mu_{-1}$ are mean vectors for classes $+1$ and -1 respectively, Σ is a covariance matrix common to both classes, and $\theta := (\mu_{+1}, \mu_{-1}, \Sigma)$ denotes the tuple of parameters. Thus $(X,Y)|\theta \sim p_{X|Y,\theta}(x|y,\theta)p_Y(y)$. The minimum probability of error classification rule, henceforth referred to as the *optimum* rule, is the maximum aposteriori probability (MAP) rule and is given by

$$\begin{aligned} \widehat{y}^*(x) &= \operatorname*{arg\,max}_{y\in\{-1,+1\}} p_{Y|X,\theta}(y|x,\theta) \\ &= \operatorname{sign}\left(\Delta^T \Sigma^+ (x-\mu)\right) \end{aligned} \tag{1}$$

where $\mu := \frac{1}{2}(\mu_{+1}+\mu_{-1})$, $\Delta := \mu_{+1}-\mu_{-1}$, and Σ^+ is the pseudoinverse of Σ. The error probability of the optimum rule (Bayes error) is given by

$$P_e^* = Q\left(\frac{1}{2}\left\|\left(\Sigma^+\right)^{\frac{1}{2}}(\mu_{+1}-\mu_{-1})\right\|\right) =: Q(\alpha/2) \tag{2}$$

where $Q(t) := \int_t^\infty \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}t^2\}dt$, $\alpha = \|(\Sigma^+)^{\frac{1}{2}}(\mu_{+1}-\mu_{-1})\|$, and the minimum attainable error probability with complete distributional knowledge and equally likely classes, i.e., the "classification difficulty", is given by $Q(\alpha/2)$. We will assume that $\alpha > 0$ so that the Bayes error is nontrivial, i.e, strictly less than $1/2$.

Supervised classification rules: Let $\mathcal{T}_n := \{(X_i, Y_i), i = 1,\ldots,n\}$ be a set of n labeled training samples that are independent and identically distributed (iid) according to $p_{X|Y,\theta}\cdot p_Y$ where θ belongs to a *known* set of feasible parameters Θ. Let X_0 be a test sample (independent of $\mathcal{T}_n$) whose *true* class label Y_0 is unobservable and needs to be estimated. A supervised classification rule is a measurable mapping

$$\widehat{y}_{n,d} : \mathbb{R}^d \times \mathcal{T}_n \to \{-1,+1\}$$

from the test data space to the set of class labels constructed using an algorithm that has access to the set of training samples $\mathcal{T}_n$ and knowledge of the form of $p_{X,Y|\theta}(x,y|\theta)$ and Θ but no direct knowledge of θ itself.

Constant difficulty parameter sets: We are interested in taking constant-difficulty high-dimensional limits. To this end we define $\Theta_0(\alpha)$ to be the set of all θ values for which the Bayes error is equal to $Q(\alpha/2)$ (see (2)):

$$\Theta_0(\alpha) := \left\{(\mu_{+1},\mu_{-1},\Sigma) : \left\|\left(\Sigma^+\right)^{\frac{1}{2}}(\mu_{+1}-\mu_{-1})\right\| = \alpha\right\}.$$

A canonical subset of Θ_0 of particular interest is one in which the covariance is spherical and the means are constrained to be on opposites ends of a d-dimensional sphere:

$$\Theta_{\text{Sphere}}(\alpha) := \left\{\left(\mathbf{h}, -\mathbf{h}, \beta^2\mathbf{I}\right) : \|\mathbf{h}\| = 1, \beta = 2/\alpha\right\}.$$

Clearly, $\Theta_{\text{Sphere}}(\alpha) \subseteq \Theta_0(\alpha)$. This special parametric set corresponds to the scenario in which X can be represented as $X = Y\mathbf{h} + Z$ where Z is white Gaussian noise that is independent of Y.

Error probabilities: Let

$$P_{e|\theta}(\widehat{y}_{n,d}) := \Pr\left(\widehat{y}_{n,d}(X_0) \neq Y_0\right)$$

denote the expected error probability of the classifier $\widehat{y}_{n,d}$ averaged across training samples $\mathcal{T}_n$ for some parameter θ and let

$$P_e(n,d,\Theta,\widehat{y}_{n,d}) := \sup_{\theta\in\Theta} P_{e|\theta}(\widehat{y}_{n,d})$$

be the maximum expected error probability of the classifier $\widehat{y}_{n,d}$ over the parameter set Θ which depends on the number of labeled training samples n and the number of signal dimensions d.

Goal: We aim to gain an understanding of how $P_e(n,d,\Theta,\widehat{y}_{n,d})$ behaves in the constant-difficulty high dimensional setting where $d,n \to \infty$, $n/d, n/\text{rank}(\Sigma^+) \to 0$, and $\Theta \subseteq \Theta_0(\alpha)$ is a sequence of constant-difficulty parameter sets.

IV. Main Results

Theorem 1. *For any sequence of classifiers $\widehat{y}_{n,d}$, we have*

$$\liminf_{(d,n/d)\to(\infty,0)} P_e(n,d,\Theta_{\textit{Sphere}},\widehat{y}_{n,d}) \geq \frac{1}{2}$$

Corollary 1. *For any sequence of parameter sets Θ with $\Theta_{\textit{Sphere}} \subseteq \Theta$, and any sequence of classifiers $\widehat{y}_{n,d}$, we have*

$$\liminf_{(d,n/d)\to(\infty,0)} P_e(n,d,\Theta,\widehat{y}_{n,d}) \geq \frac{1}{2}$$

Corollary 2. *Let $\Theta_{\textit{subset}} := \{(\mathbf{h},-\mathbf{h},\beta^2\mathbf{I}) \in \Theta_{\textit{Sphere}}, \mathbf{h} \in \mathcal{H} \subseteq \mathcal{S}^{d-1}\}$ where $\mathcal{S}^{d-1}$ is the unit $(d-1)$-sphere in $\mathbb{R}^d$. Let $\text{vol}(\mathcal{H}) \triangleq \Pr_{H\sim U(\mathcal{S}^{d-1})}(H \in \mathcal{H})$, where $U(\mathcal{S}^{d-1})$ denotes the uniform distribution over $\mathcal{S}^{d-1}$. If for a sequence of classifiers $\widehat{y}_{n,d}$,*

$$\limsup_{(d,n/d)\to(\infty,0)} P_e(n,d,\Theta_{\textit{Sphere}},\widehat{y}_{n,d}) = \frac{1}{2}$$

and

$$\lim_{d\to\infty} \text{vol}(\mathcal{H}) > 0$$

then

$$\limsup_{(d,n/d)\to(\infty,0)} P_e(n,d,\Theta_{subset},\widehat{y}_{n,d}) \geq \frac{1}{2}.$$

The proofs of the theorem and the two corollaries are presented in Section VI.

V. Discussion

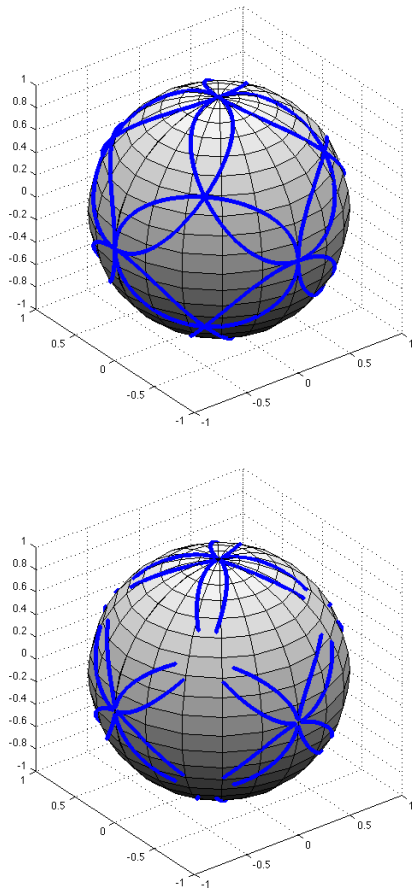

Fig. 1. Exponential sparsity class $\mathcal{H}_{exp}$ (solid curves, top figure) and polynomial sparsity class $\mathcal{H}_{poly}$ (solid curves, bottom figure) for $d = 3$.

Theorem 1 informs us that even in the "easier" scenario where the covariance is spherical and perfectly known, the worst case classification performance of *any* sequence of supervised classifiers is asymptotically no better than the classifier which flips an unbiased coin to make decisions. It is interesting to note that this conclusion holds for any arbitrarily small positive Bays error $Q(\alpha/2)$.

An immediate corollary of this theorem is the impossibility of attaining less than half asymptotic error probability for larger parametric sets, specifically parameter sets that contain Θ_{Sphere} such as Θ_0. These parameter sets contain more unknowns (degrees of freedom), e.g., the covariance, but have the same classification difficulty $Q(\alpha/2)$.

These results are consistent with previous results concerning the asymptotic behavior of the so-called plug-in family of classifiers where Maximum Likelihood (ML) estimates of parameters are plugged into the MAP rule given in (1)):

- Plug-in classification for Θ_0: In [5] it was shown that for plug-in classifiers that use ML estimates of the means and covariance, the performance is asymptotically no better than half.

- Sensing-aware classification: Here it is assumed that data is generated according to $X = \mathbf{h}W + Z$, where $\mathbf{h}$ is the "sensing subspace", W is a class-dependent scalar latent variable and Z is white Gaussian noise independent of the class labels and latent variables. Assuming Gaussian class conditional densities for W, the class conditional density of X would be Gaussian with covariance $\gamma^2\mathbf{h}\mathbf{h}^\top + \beta^2\mathbf{I}$. Also the mean of X would be in the same direction $\mathbf{h}$ for the two classes. In [5] it was shown that the asymptotic probability of error of the ML projection classifier which is based on projecting data along the ML estimate of $\mathbf{h}$ is also no better than half. This can be seen as a special case of the Corollary 1, by considering $\Theta_{\text{Sensing Aware}} \supseteq \Theta_{\text{Sphere}}$ where

$$\Theta_{\text{Sensing Aware}}(\alpha) := \big\{ \big(m_1\mathbf{h}, m_2\mathbf{h}, \gamma^2\mathbf{h}\mathbf{h}^\top + \beta^2\mathbf{I}\big) : \|\mathbf{h}\| = 1, \gamma \geq 0, \beta > 0, |m_1 - m_2| = \alpha\sqrt{\gamma^2+\beta^2}\big\}.$$

Another important consequence of the Theorem 1 is Corollary 2. This result can be interpreted as implying that nontrivial asymptotic minimax error probability is attainable only for parametric subsets of zero Haar measure. This highlights the *necessity* of some weak form of sparsity in the set of feasible $\mathbf{h}$ values in order to attain non-trivial asymptotic error probability. This is consistent with previous results that have shown that the supervised classification error can in fact converge to the Bayes error when $\mathbf{h}$ belongs to specific sparsity classes [4], [5].

Specifically, in [5] it was shown that if the magnitudes of the components of $\mathbf{h}$ decay exponentially (or even polynomially) when reordered according to decreasing values, then $\mathbf{h}$ can be estimated consistently (in the mean square sense) by soft-thresholding the ML estimate of $\mathbf{h}$ even in the case that d grows sub-exponentially in terms of $n/\log(n)$. Then it can be shown that a classifier based on projecting data onto the estimation of $\mathbf{h}$ attains the Bayes error asymptotically.

To represent these two sparsity classes, $\mathcal{H}_{exp}$ and $\mathcal{H}_{poly}$ were defined in [5] as bellow:

$$\mathcal{H}_{exp} = \big\{\mathbf{h} : \big|h_{(k)}\big| = M_1(d)\alpha^k, 0 < \alpha < 1\big\}$$
$$\mathcal{H}_{poly} = \big\{\mathbf{h} : \big|h_{(k)}\big| = M_2(d)k^{-\beta}, \beta > 0.5\big\}$$

where $(h_{(1)}, \ldots, h_{(d)})$ are the components of $\mathbf{h}$ in decreasing order of magnitude. Since there is only one degree of freedom in each set, the Haar measure of these two sets vanish when $d \geq 3$. Figure 1 illustrates these two sets (dark solid curves) on the unit sphere in 3-dimensional space. These sets satisfy the necessary conditions suggested by the Corollary 2 and in fact achieve the Bayes error asymptotically which is better than half. To summarize, it is essential to have strong structural assumptions on the feasible set of parameters in order to obtain non-trivial asymptotic classification performance in high-dimensional small sample settings.

VI. PROOFS

A. Proof of Theorem 1

Proof: The key proof-idea is to randomize the selection of $\theta \in \Theta_{\text{Sphere}}$. The worst error probability over $\theta \in \Theta_{\text{Sphere}}$ is not smaller than the average (expected) error probability when θ is random. For any fixed value of θ, the training and test samples are totally independent but if θ is random, they can become dependent *through* θ. Then the expected error probability (with respect to both training data and θ) of any supervised classification rule $\widehat{y}_{n,d}$ cannot be smaller than that of the θ-distribution-induced MAP rule based on $\mathcal{T}_n$. If the randomizing distribution for θ is carefully selected then the lower bound can be made to converge to $1/2$ as n and d scale. Now we work out the details of this proof-idea.

Let $\mathcal{S}^{d-1}$ denote the unit $(d-1)$-sphere. For every $\mathbf{h} \in \mathcal{S}^{d-1}$, $\|\mathbf{h}\| = 1$ and $(\mathbf{h}, -\mathbf{h}, \beta^2\mathbf{I}) \in \Theta_{\text{Sphere}}$. Let $\theta \sim (H, -H, \beta^2\mathbf{I})$ where $H \sim \text{Uniform}(\mathcal{S}^{d-1})$ which is independent of data labels Y_i. Then, for every realization $H = \mathbf{h}$, $\theta \in \Theta_{\text{Sphere}}$ and conditioned on $H = \mathbf{h}$ (equivalently conditioned on a realization of θ), the training and test samples have the joint distribution that was described earlier. By defining

$$\widehat{y}_{\text{MAP}}(X_0) \triangleq \underset{y_0 \in \{-1,+1\}}{\arg\max}\ p_{Y_0|X_0,\mathcal{T}_n}(y_0|x_0, \mathcal{T}_n, \Theta_{\text{Sphere}}) \quad (3)$$

and using the notation $P_e \triangleq P_e(n, d, \Theta_{\text{Sphere}}, \widehat{y}_{n,d})$, we have :

$$\begin{aligned} P_e &\geq \mathbb{E}_{\theta \in \Theta_{\text{Sphere}}}\left[P_{e|\theta}(\widehat{y}_{n,d})\right] \\ &= \mathbb{E}_{\theta \in \Theta_{\text{Sphere}}, \mathcal{T}_n}\left[\Pr\left(\widehat{y}_{n,d}(X_0) \neq Y_0 | \mathcal{T}_n, \theta\right)\right] \\ &\overset{(i)}{\geq} \mathbb{E}_{\theta \in \Theta_{\text{Sphere}}, \mathcal{T}_n}\left[\Pr\left(\widehat{y}_{\text{MAP}}(X_0) \neq Y_0 | \mathcal{T}_n, \theta\right)\right] \end{aligned} \quad (4)$$

where (i) holds because the MAP rule minimizes the error probability. Using the notation $\widehat{y}_{\text{MAP}} \triangleq \widehat{y}_{\text{MAP}}(X_0)$, we have:

$$\begin{aligned} \widehat{y}_{\text{MAP}} &= \underset{y_0 \in \{-1,+1\}}{\arg\max}\ p_{Y_0|X_0,\mathcal{T}_n}(y_0|x_0, \mathcal{T}_n, \Theta_{\text{Sphere}}) \\ &= \underset{y_0 \in \{-1,+1\}}{\arg\max}\ p_{X_0,Y_0,\mathcal{T}_n}(x_0, y_0, \mathcal{T}_n | \Theta_{\text{Sphere}}) \\ &= \underset{y_0 \in \{-1,+1\}}{\arg\max}\ \mathbb{E}_{\theta \in \Theta_{\text{Sphere}}}\left[p_{X_0,Y_0,\mathcal{T}_n|\theta}(x_0, y_0, \mathcal{T}_n|\theta)\right] \\ &\overset{(i)}{=} \underset{y_0 \in \{-1,+1\}}{\arg\max}\ \mathbb{E}_{\theta \in \Theta_{\text{Sphere}}}\left[\prod_{i=0}^{n} p_{X_i,Y_i|\theta}(x_i, y_i|\theta)\right] \\ &\overset{(ii)}{=} \underset{y_0 \in \{-1,+1\}}{\arg\max}\ \mathbb{E}_{\theta \in \Theta_{\text{Sphere}}}\left[\prod_{i=0}^{n} p_{X_i|Y_i,\theta}(x_i|y_i, \theta)\right] \\ &\overset{(iii)}{=} \underset{y_0 \in \{-1,+1\}}{\arg\max}\ \mathbb{E}_H\left[\prod_{i=0}^{n} \exp\left\{-\frac{1}{2\beta^2}\left\|x_i - y_i H\right\|^2\right\}\right] \\ &= \underset{y_0 \in \{-1,+1\}}{\arg\max}\ \mathbb{E}_H\Big[\exp\Big\{-\frac{1}{2\beta^2}\sum_{i=0}^{n}\|x_i\|^2 + \\ &\qquad \|y_i H\|^2 - 2y_i x_i^T H\Big\}\Big] \\ &\overset{(iv)}{=} \underset{y_0 \in \{-1,+1\}}{\arg\max}\ \mathbb{E}_H\left[\exp\left\{\frac{1}{\beta^2} H^T\left(\sum_{i=0}^{n} y_i x_i\right)\right\}\right] \\ &\overset{(v)}{=} \underset{y_0 \in \{-1,+1\}}{\arg\max}\ \frac{1}{\beta^2}\left\|\sum_{i=0}^{n} y_i x_i\right\| \\ &= \text{sign}\left(x_0^\top\left(\sum_{i=1}^{n} y_i x_i\right)\right). \end{aligned}$$

In the above derivation, (i) is because the training and test samples are conditionally independent given θ, (ii) is because $p_{Y_i}(+1) = p_{Y_i}(-1) = 0.5$ for all i, (iii) is because $\theta = (H, -H, \beta^2\mathbf{I}) = (\mu_{+1}, \mu_{-1}, \Sigma)$, (iv) is because $\|H\| = 1$ and $y_i = \pm 1$ for all i, and (v) is due to the following result.

Lemma 1. *Let H be uniformly distributed on $\mathcal{S}^{d-1}$ and $f(x) := \mathbb{E}_H[\exp\{H^\top x\}]$. Then $f(x)$ is a radial function that is convex and nondecreasing in $\|x\|$.*

Proof: Since H is uniformly distributed on $\mathcal{S}^{d-1}$, $f(x)$ is a radial function. Consider $x = tu$ where $t \in \mathbb{R}$ and $\|u\| = 1$. If $g(t) := f(tu)$ then $g(0) = 1$ and g is symmetric since the distribution of H is spherically symmetric. We also have $g'(t) = \mathbb{E}_H[(H^\top u)\exp\{tH^\top u\}]$ so that $g'(0) = 0$, again because the distribution of H is spherically symmetric. Finally, $g''(t) = \mathbb{E}_H[(H^\top u)^2 \exp\{tH^\top u\}] \geq 0$ which shows that $g(t)$ is convex for $t \geq 0$. Since $g(t)$ is convex for $t \geq 0$ and $g'(0) = 0$, it is nondecreasing for $t \geq 0$. ∎

Continuing the proof, we have :

$$\begin{aligned} P_e &\overset{(i)}{\geq} \mathbb{E}_{\theta, \mathcal{T}_n}\left[P\left(\widehat{y}_{\text{MAP}}(X_0, \mathcal{T}_n) \neq Y_0 | \mathcal{T}_n, \theta\right)\right] \\ &= \mathbb{E}_{H, \mathcal{T}_n}\left[\Pr\left(\text{sign}\left(X_0^\top \sum_{i=1}^{n} Y_i X_i\right) \neq Y_0 \middle| \mathcal{T}_n, H\right)\right] \\ &= \mathbb{E}_{H, \mathcal{T}_n}\left[Q\left(\frac{-H^\top\left(\sum_{i=1}^{n} Y_i X_i\right)}{\beta\left\|\left(\sum_{i=1}^{n} Y_i X_i\right)\right\|}\right)\right] \\ &= \mathbb{E}_{H, \mathcal{T}_n}\left[Q\left(\frac{-(1 + H^\top V)}{\beta\sqrt{1 + 2H^\top V + \|V\|^2}}\right)\right] \end{aligned} \quad (5)$$

where $V \sim \mathcal{N}(\mathbf{0}, \frac{\beta^2}{n}\mathbf{I}_d)$ and is independent of H. This follows as we can write X_i as $X_i = Y_i H + Z_i$, where Z_i are white Gaussian noise with variance β^2, which are independent of H and labels Y_i. Hence $\frac{1}{n}\sum_{i=1}^{n} X_i Y_i = H + \frac{1}{n}\sum_{i=1}^{n} Y_i Z_i$. Finally, by taking $V = \frac{1}{n}\sum_{i=1}^{n} Y_i Z_i$, it follows that $V \sim \mathcal{N}(\mathbf{0}, \frac{\beta^2}{n}\mathbf{I}_d)$. Note that (i) is proved in equation (4).

Lemma 2. *$W = \frac{1 + H^\top V}{\sqrt{1 + 2H^\top V + \|V\|^2}} \xrightarrow{p} 0$, as $d \to \infty$.*

Proof: Since H and V are independent, $\mathbb{E}(H^\top V) = \mathbb{E}(H^\top)\,\mathbb{E}(V) = 0$ and

$$\begin{aligned} \text{Var}(H^\top V) &= \mathbb{E}(H^\top V V^\top H) = \mathbb{E}_H\,\mathbb{E}_{V|H}(H^\top V V^\top H) \\ &= \mathbb{E}_H\left(\frac{\beta^2}{n} H^\top H\right) = \frac{\beta^2}{n} \longrightarrow 0 \end{aligned}$$

as $n, d \to \infty$. As a result $1 + H^\top V \xrightarrow{p} 1$. Next, we will show that $\text{Var}(1 + 2H^\top V + \|V\|^2) = \mathcal{O}\left(\frac{d}{n^2}\right)$. First, observe that $\mathbb{E}(1 + 2H^\top V + \|V\|^2) = 1 + \beta^2\frac{d}{n}$. Thus

$$\begin{aligned} \text{Var}(1 + 2H^\top V + \|V\|^2) = &\mathbb{E}\left((1 + 2H^\top V + \|V\|^2)^2\right) \\ &- \left(1 + \beta^2\frac{d}{n}\right)^2 \end{aligned}$$

Furthermore, we have

$$\mathbb{E}\left((1+2H^\top V+\|V\|^2)^2\right) = \underbrace{4\,\mathbb{E}(H^\top V V^\top H)}_{4\beta^2/n} + \mathbb{E}(\|V\|^4) + \underbrace{4\,\mathbb{E}(H^\top V)}_{0} + \underbrace{2\,\mathbb{E}(\|V\|^2)}_{2\beta^2\frac{d}{n}} + \underbrace{4\,\mathbb{E}(H^\top V\|V\|^2)}_{0} + 1$$

It remains to calculate $\mathbb{E}(\|V\|^4)$. Note that $\mathbb{E}(\|V\|^4) = \mathrm{Var}(V^\top V) + \beta^4\frac{d^2}{n^2}$. But we know that for a Gaussian random variable $\epsilon \sim \mathcal{N}(\mu,\Sigma)$, and an arbitrary matrix Λ, we have $\mathrm{Var}(\epsilon^\top\Lambda\epsilon) = 2\,\mathrm{tr}(\Lambda\Sigma\Lambda\Sigma) + 4\mu^\top\Lambda\Sigma\Lambda\mu$. Therefore, $\mathbb{E}(\|V\|^4) = 2\beta^4\frac{d}{n^2} + \beta^4\frac{d^2}{n^2}$. Finally, we have

$$\mathrm{Var}(1+2H^\top V+\|V\|^2) = 4\frac{\beta^2}{n} + 2\beta^4\frac{d}{n^2} = \mathcal{O}\left(\frac{d}{n^2}\right)$$

As a result, $\frac{n}{d}(1+2H^\top V+\|V\|^2) \xrightarrow{p} \beta^2$, because it converges in L^2. We have

$$W = \sqrt{\frac{n}{d}}\frac{1+H^\top V}{\sqrt{\frac{n}{d}\left(1+2H^\top V+\|V\|^2\right)}}$$

Note that the numerator and denominator go to 1 and β in probability, respectively. Therefore, using Slutsky's Theorem, the whole fraction goes to $1/\beta$ in probability. But $\sqrt{\frac{n}{d}}$ goes to zero, therefore, $W \xrightarrow{p} 0$. ∎

Using Slutsky's theorem, $Q\left(-W/\beta\right) \xrightarrow{p} \frac{1}{2}$. Since $0 \le Q(.) \le 1$, using Dominated Convergence Theorem, $\lim\limits_{(d,n/d)\to(\infty,0)} \mathbb{E}\left[Q\left(-W/\beta^2\right)\right] = \frac{1}{2}$. Taking limit inferior of both sides of (5), we finally conclude that for any $\widehat{y}_{n,d}$

$$\liminf_{(d,n/d)\to(\infty,0)} P_e(n,d,\Theta_{\text{Sphere}},\widehat{y}_{n,d}) \ge \frac{1}{2}$$

∎

B. Proof of Corollary 1

For any classifier $\widehat{y}_{n,d}$, we have :

$$P_e(n,d,\Theta,\widehat{y}_{n,d}) = \sup_{\theta\in\Theta} P_{e|\theta}(\widehat{y}_{n,d}) \ge \sup_{\theta\in\Theta_{\text{Sphere}}} P_{e|\theta}(\widehat{y}_{n,d}) = P_e(n,d,\Theta_{\text{Sphere}},\widehat{y}_{n,d})$$

because $\Theta_{\text{Sphere}} \subseteq \Theta$. By taking limit inferior of two sides, the Corollary is proved.

C. Proof of Corollary 2

Let $\mathrm{vol}(\mathcal{H}) \triangleq \Pr_{H\sim U(\mathcal{S}^{d-1})}(H\in\mathcal{H})$, where $U(\mathcal{S}^{d-1})$ denotes the uniform distribution over the unit $(d-1)$-sphere in d dimensional space. Suppose that $\limsup\limits_{(d,n/d)\to(\infty,0)} P_e(n,d,\Theta_{\text{subset}},\widehat{y}_{n,d}) < \frac{1}{2}$. Hence $\limsup\limits_{(d,n/d)\to(\infty,0)} \mathbb{E}_{H\sim U(\mathcal{H})}(P_{e|\theta}(\widehat{y}_{n,d})) < \frac{1}{2}$. For the specified sequence of classifiers $\widehat{y}_{n,d}$, which satisfies the conditions of the corollary, we have

$$\begin{aligned}
&\limsup_{(d,n/d)\to(\infty,0)} \mathbb{E}_{H\sim U(\mathcal{S}^{d-1})}(P_{e|\theta}(\widehat{y}_{n,d})) \\
&= \limsup_{(d,n/d)\to(\infty,0)} \mathrm{vol}(\mathcal{H})\,\mathbb{E}_{H\sim U(\mathcal{H})}(P_{e|\theta}(\widehat{y}_{n,d})) \\
&\quad + \mathrm{vol}(\bar{\mathcal{H}})\,\mathbb{E}_{H\sim U(\bar{\mathcal{H}})}(P_{e|\theta}(\widehat{y}_{n,d})) \\
&\le \limsup_{(d,n/d)\to(\infty,0)} \mathrm{vol}(\mathcal{H}) \limsup_{(d,n/d)\to(\infty,0)} \mathbb{E}_{H\sim U(\mathcal{H})}(P_{e|\theta}(\widehat{y}_{n,d})) \\
&\quad + \limsup_{(d,n/d)\to(\infty,0)} \mathrm{vol}(\bar{\mathcal{H}}) \limsup_{(d,n/d)\to(\infty,0)} \mathbb{E}_{H\sim U(\bar{\mathcal{H}})}(P_{e|\theta}(\widehat{y}_{n,d})) \\
&< \frac{1}{2}\left(\lim_{(d,n/d)\to(\infty,0)} \mathrm{vol}(\mathcal{H}) + \lim_{(d,n/d)\to(\infty,0)} \mathrm{vol}(\bar{\mathcal{H}})\right) \\
&< \frac{1}{2}
\end{aligned}$$

which is a contradiction.

VII. Concluding Remarks

That prior knowledge such as sparsity improves inference in high dimensional small sample settings is folklore. The results presented here show that in fact such knowledge is absolutely indispensable in that otherwise the asymptotic performance degenerates to a random coin toss. The results presented here focused on supervised binary classification with Gaussian class conditional densities and equally likely classes. One could expect similar conclusions to hold in more complex inference problems. However the proof techniques used in this work may not generalize to more complex and non-Gaussian settings.

Acknowledgment

This article is based upon work supported by the U.S. AFOSR and U.S. NSF under award numbers #FA9550-10-1-0458 (subaward #A1795) and #1218992 respectively.

References

[1] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *Annals of Statistics*, vol. 32, no. 3, pp. 962–994, 2004.

[2] A. Singh, R. D. Nowak, and A. R. Calderbank, "Detecting weak but hierarchically-structured patterns in networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 749–756.

[3] P. J. Bickel and E. Levina, "Some theory for fishers linear discriminant function, naive bayes, and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.

[4] J. Shao, Y. Wang, X. Deng, and S. Wang, "Sparse linear discriminant analysis by thresholding for high dimensional data," *Annals of Statistics*, vol. 39, no. 2, pp. 1241–1265, 2012.

[5] B. Orten, P. Ishwar, W. C. Karl, and V. Saligrama, "Sensing structure in learning-based binary classification of high-dimensional data: Opportunities and perils," in *Annual Allerton Conference on Communication, Control and Computing*, 2011.

[6] L. Devroye and G. Lugosi, "Lower bounds in pattern recognition and learning," *Pattern Recognition*, vol. 28, no. 7, pp. 1011–1018, 1995.

[7] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant, "A general lower bound on the number of examples needed for learning," *Journal of Information and Computation*, vol. 82, no. 3, pp. 247–261, 1989.

[8] E. Oblow, "Implementing valiants learnability theory using random sets," *Machine Learning*, vol. 8, no. 1, pp. 45–74, 1992.

[9] G. M. Benedek and A. Itai, "Learnability with respect to fixed distributions," *Journal of Theoretical Computer Science*, vol. 86, no. 2, pp. 377–389, 1991.