Learning with Few Features and Samples

João Pinheiro, Y.Z. Janice Chen, Paulo Azevedo, Daniel Menasché, and Don Towsley, *Life Fellow, IEEE*

***Abstract*—Motivated by sensors gathering data for classification purposes, with stringent costs associated to the collection of samples and addition of features, we consider the problem of determining the optimal number of features to be transmitted through the network. Under a bivariate model, we express how the classifier expected error probability depends on the sample size and on one of the two features discriminatory power. Then, we use the characterized error probability to determine a threshold on the sample size such that if the sample size is larger than the threshold the presence of an additional feature is in fact beneficial.**

***Index Terms*—Statistical learning, samples, features, SVM.**

I. Introduction

SENSING the environment is a major challenge and a fundamental cornerstone within autonomous systems, vehicular networks, weather forecasting, military networks and many other modern initiatives. After sensors are deployed, they degrade over time. Some sensors will have their batteries depleted and others will fail. Each sensor collects features from the environment, such as temperature, light and pressure. Failures, in turn, will cause a decrease in the number of collected features, where these features may represent different modalities at the same geographical position, or have the same modality but be collected from different geographic locations.

Features and samples have costs associated with them. Collecting additional features may require the procurement of new sensors, and retrieving more samples per time unit may come at the cost of increased bandwidth and a faster depletion of limited batteries. Given such costs, which preclude the collection of a large number of features or the transmission of data at high sampling rates, we pose the following questions:

- given a small sample size, when are fewer features preferred for training a classifier, i.e., when is less better?
- how to compensate for the lack of features using samples?

Motivated by the above questions, in this letter we start our investigation focusing on the special case of a binary classifier.

Related work. Training classifiers with few samples has recently received attention from the remote sensing community [1], [2] among others [3]–[5]. Consider a system that implements a change point detection algorithm to track abrupt events in the environment. Once a change point occurs, the training of a classifier is then restricted to small sample sizes and possibly limited number of features [6]. Although there has been significant effort to characterize how samples and features impact classification accuracy [3], [6], their relationship still poses open challenges, e.g., regarding the minimum number of samples required to achieve a target error probability as a function of the number of available features.

Methodology. To illustrate the relationship between sample size, number of features, and asymptotic error probability of the trained classifier, we consider a simple workload with two features. The conditional distribution of each sample given its class is assumed to be bivariate Gaussian (e.g., as in [7]), with a correlation coefficient ρ; one of the features has variance one and the other one has variance δ^2, noting that for $\rho = 0$ the *discriminatory power* of the second feature decreases as δ grows. Although this workload is admittedly simple, this simple model already serves our purposes, namely, to show that *(a)* depending on the discriminatory power of the second feature it may be worth ignoring it, and *(b)* additional samples can compensate for the lack of a feature.

Contributions. Among our contributions, we express error probability as a function of sample size, number of features, and δ, initially assuming $\rho = 0$. Our analysis reveals the existence of a sample size threshold $n^\star$ such that one should train a classifier using a single feature when the sample size is smaller than $n^\star$, and both features otherwise. The first regime, with very small sample sizes, corresponds to overfitting, wherein fewer features provide greater generalization power and smaller classification error [8], [9]. As sample size grows, the system switches to the second regime, wherein having additional features is helpful, and there is a trade-off between samples and features, i.e., samples can compensate for features. When the number of samples is further increased, however, both samples and features are required to achieve the corresponding accuracy levels. Then, we discuss how $\rho \neq 0$ impacts our results.[1]

II. Problem Formulation

The classification process involves two stages: training the classifier, with labeled samples, and then inferring the classes of unlabeled samples. Our two main attributes of interest are dataset cardinality, denoted by n, which corresponds to the number of samples in the training set, and dataset dimensionality, denoted by d, which corresponds to the number of features in each sample. One of our goals is to assess how n and d simultaneously affect the classifier performance, as measured by its expected error probability.

Background. We consider a binary classification problem, where each sample is drawn from one of two classes, denoted by labels -1 and +1. The set of samples comprises a dataset $\mathcal{D}$ with n ordered pairs $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, for $i = 1, \ldots, n$. The dataset $\mathcal{D}$ is used to train

João Ismael Pinheiro, Paulo Renato Azevedo, and Daniel Sadoc Menasché are with the Department of Computer Science, Federal University of Rio de Janeiro (UFRJ), RJ, Brazil e-mail: jismael@im.ufrj.br and sadoc@dcc.ufrj.br.

Y.Z. Janice Chen and Don Towsley are with the College of Information and Computer Sciences, UMass, Amherst, US e-mail: towsley@cs.umass.edu

Manuscript received December 7, 2021.

[1] A replication package in Python, together with additional details, are available at https://github.com/paulorenatoaz/slacgs.

a classifier that minimizes the empirical error probability on $\mathcal{D}$. Letting X and Y denote the random variables corresponding to the feature vector and target class, the joint probability distribution for the pair (X, Y) is assumed to be unknown.

A classifier is a function $h : \boldsymbol{x} \to h(\boldsymbol{x})$ assigning a label $h(\boldsymbol{x}) \in \{-1, +1\}$ to any feature vector $\boldsymbol{x} \in \mathbb{R}^d$. The error probability (or loss) of classifier h is denoted by $L(h)$, and is given by $L(h) = P(Y \neq h(X))$. A dictionary H (also known as hypotheses set) is a set of classifiers (e.g., linear classifiers). The learning problem consists of choosing a classifier from H to be the "solution" for our problem. It is well known that the Bayes classifier, which classifies each input according to $\text{argmax}_y P(Y = y|X = x)$, is globally optimal [10], [11]; the Bayes error is the minimal error probability. We denote by $h^{(B)}$ and $L(h^{(B)})$ the Bayes classifier and its error probability.

Bivariate Gaussian. To assess the generalization power of the considered classifiers, we sample $\mathcal{D}$ from a known probabilistic model, e.g., as in [12], [13]. In particular, we consider a 2-dimensional classification problem, wherein samples from each class are drawn from a bivariate Gaussian distribution. We assume equal priors for the two classes, i.e., $P(Y = +1) = P(Y = -1) = 0.5$. The conditional distribution of feature vector $X = (X_1, X_2)$ given Y is

$$(X_1, X_2) \sim \begin{cases} \mathcal{N}((+1, +1), \Sigma), & \text{if } Y = +1, \\ \mathcal{N}((-1, -1), \Sigma), & \text{if } Y = -1, \end{cases} \tag{1}$$

where $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes a bivariate Gaussian distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^2$ and covariance matrix Σ,

$$\Sigma = \begin{pmatrix} 1 & \rho\delta \\ \rho\delta & \delta^2 \end{pmatrix} \tag{2}$$

with $\delta \geq 1$ and $|\rho| \leq 1$. Note that δ is the conditional standard deviation of feature X_2, given Y. If $\delta > 1$ feature X_2 has a smaller discriminating power than feature X_1: the larger the value of δ, the less informative is feature X_2.

Note that we consider balanced datasets, with the same number of samples in each class.

SVM classifier. Our dictionary H is formed by all linear classifiers, i.e., straight lines. A naive approach for learning consists in searching for the line that minimizes the empirical error rate, e.g., through grid search. However, the problem may admit multiple solutions, needing a principled strategy to break ties [14]. Therefore, we consider linear support vector machines (SVMs), which have a number of desired properties, including (i) a principled strategy to determine the best discriminator, by maximizing the distance, known as the *margin*, between the separating line and the two classes of points to be separated [14]; (ii) polynomial computational cost [15], [16] and (iii) convergence to an optimal solution, under mild conditions [17]–[19]. We have also experimented with linear discriminant analysis (LDA), obtaining similar results, and opted to report SVM results, since the principles of SVM require less restrictive assumptions.

Linear SVM relies on a single parameter, the misclassification cost, appearing as a regularization term in the Lagrange formulation, and denoted by γ [20]. Throughout this work, we let $\gamma = 1$, its default value; we experimented with other values, with results remaining roughly the same.

TABLE I
SVM ERROR PROBABILITIES (MULTIPLIED BY 100) FOR $\rho = 0$

		$E_1(n)$,	$E_2(n;\delta)$, 2 features					
n	1 feature	δ=1	δ=2	δ=3	δ=4	δ=5	δ=6	δ=7
2	20.50	14.99	23.29	26.72	28.20	28.93	29.36	29.63
4	18.83	13.91	21.73	24.47	25.62	26.20	26.52	26.73
8	17.63	11.98	18.95	21.02	21.82	22.21	22.42	22.54
16	16.87	10.24	16.28	17.92	18.55	18.86	19.01	19.11
32	16.39	9.18	14.74	16.22	16.78	17.05	17.20	17.29
64	16.15	8.57	13.96	15.39	15.94	16.20	16.34	16.43
129	16.01	8.24	13.57	14.99	15.53	15.79	15.93	16.02
256	15.94	8.06	13.38	14.79	15.33	15.59	15.73	15.82
512	15.90	7.96	13.28	14.69	15.23	15.49	15.63	15.72
1024	15.88	7.91	13.23	14.64	15.18	15.44	15.58	15.67
$n^\star$			12	28	46	70	98	130
Linear regression of error probability: $E(n) = \alpha + \beta/n$.								
α	0.159	0.080	0.132	0.146	0.151	0.154	0.155	0.156
β	0.142	0.333	0.468	0.522	0.542	0.552	0.556	0.558
$n^\star$			12	30	52	78	106	136

III. Analytical Model

Next, we consider the Bayes error corresponding to an infinite sample size ($n = \infty$). We begin by considering the case where only feature x_1 is available. We have a 1-dimensional classification problem, wherein we search for the best cut point c. The error probability, under the setting of Section II, is $L_1(c) = (1 + \Phi(c - 1) - \Phi(c + 1))/2$ where $\Phi(\cdot)$ is the CDF of a Gaussian with mean zero and variance one. The optimal separator is $\tilde{c} = 0$, and $L_1(0) = L_1(\tilde{h}_1) = 1 - \Phi(1) = 0.1586553$.

Next, we consider the case where both features are available. Given a classifier characterized by the line $x_2 = a + bx_1$, its theoretical error rate under the setting of Section II is

$$L_2(a, b) = \tag{3}$$
$$= \frac{1}{2}\left(1 - \Phi\left(\frac{|a + b - 1|}{\sqrt{S}}\right)\right) + \frac{1}{2}\left(1 - \Phi\left(\frac{|a - b + 1|}{\sqrt{S}}\right)\right),$$

where $S = (\rho\delta - b)^2 + \delta^2(1 - \rho^2)$. It can be verified that the optimal solution is given by a straight line through the origin, $\tilde{a} = 0$, with a slope of $\tilde{b} = \delta(\delta - \rho)/(\delta\rho - 1)$. As the classifier $h^{(B)}$ belongs to the dictionary formed by straight lines in $\mathbb{R}^2$,

$$L(h^{(B)}) = L_2\left(0, \frac{\delta(\delta - \rho)}{\delta\rho - 1}\right) = 1 - \Phi\left(\frac{1}{\delta}\sqrt{\frac{\delta^2 - 2\rho\delta + 1}{1 - \rho^2}}\right). \tag{4}$$

Interestingly, taking the partial derivative of the above expression with respect to ρ and setting it to zero we conclude that the upper bound of $L(h^{(B)})$ is achieved at $\rho = 1/\delta$,

$$0 \leq L(h^{(B)}; \rho) \leq L(h^{(B)}; 1/\delta) = 1 - \Phi(1). \tag{5}$$

The lower bound on the error is achieved for $\rho = 1$ and $\rho = -1$, corresponding to scenarios where X_2 is a deterministic function of X_1. As ρ increases from -1 to $1/\delta$, the error increases; as ρ is further increased, the error decreases to zero for large enough n (see also Section IV-D).

In this above analysis, we considered the case $n = \infty$. The opposite extreme, $n = 2$, with a dataset comprised of two samples, one in each class, is also amenable to analytical treatment. In this case, the expected error for the one-feature case is computed through an integral, $\int_{c=-\infty}^{+\infty} L_1(c) e^{-c^2} \pi^{-1/2} dc \approx$

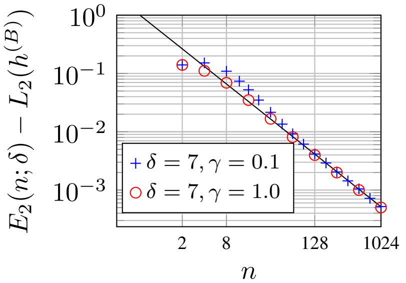

Fig. 1. Error probability decays as $1/n$. Figure shows SVM error probability minus asymptotic loss of Bayes classifier (dots, from Table I minus eq. (4)).

0.2070336 (to be compared against 0.205 shown in the top left of Table I). For the two-feature case, the error is expressed as a more complex double integral involving the perpendicular bisector of the points. The results suggest that when the sample size is small (e.g., two observations) and the variability is high ($\delta \geq 2$), it may be advantageous to use fewer features, as shown in Table I and further detailed next.

IV. Analyzing the Role of Samples and Features

A. *When is it better to use fewer features?*

When the sample size is small, using fewer features simplifies training and avoids overfitting. To illustrate this point, we compare the two scenarios, namely with both features and with a single feature available, indicating how error probability decays as a function of the sample size, for $\rho = 0$ (see Table I).

More on methodology. Each loss in Table I is obtained from 32,000 datasets sampled from the same distribution. Confidence interval lengths of the error probability estimates were negligible, on the order of 10^{-5}. Recall that each dataset $\mathcal{D}$ contains $n/2$ samples from each class, and that for each of those datasets SVM produced a corresponding separating line (in the case of two features) or cut point (in the case of a single feature). Using the results reported in Sec. III the corresponding error probabilities were computed. Finally, each set of m error probabilities is averaged to produce the numbers in the table.

Fewer features can be better. Let $E_d(n;\delta)$ be the expected error probability when SVM is applied to a dataset with n samples and d features. Table I shows that $E_1(2) < E_2(2;\delta)$ and, for $\delta \geq 2$, there is a point $n^\star(\delta)$ at which $E_1(n^\star(\delta)) = E_2(n^\star(\delta);\delta)$. Table I indicates that $n^\star(\delta)$ is an increasing function of δ, i.e., the less informative the second feature is, the larger the sample size is needed to justify the use of such an additional feature. Indeed, when n is small, there is overfitting, i.e., the total number of parameters to be estimated along the classification process is close to the total number of available samples, leaving us with fewer degrees of freedom. In face of overfitting, whenever the less discriminating feature is discarded the cut point generalizes better than the line.

Regularization. The classical variance-bias tradeoff [21]–[23] and the more recent "double descent" literature [12], [13], [24]–[26] suggest that as the number of parameters in the model increase the error probability peaks around the point where model capacity roughly equals the number of samples. The isotropic case $\rho = 0$, $\delta = 1$ is discussed in [24]. In Table I we consider the non-isotropic case, and show that reducing the number of features has the effect of regularization, naturally leading to better generalization capability, which is critical for small sample sizes as overfitting must be avoided.

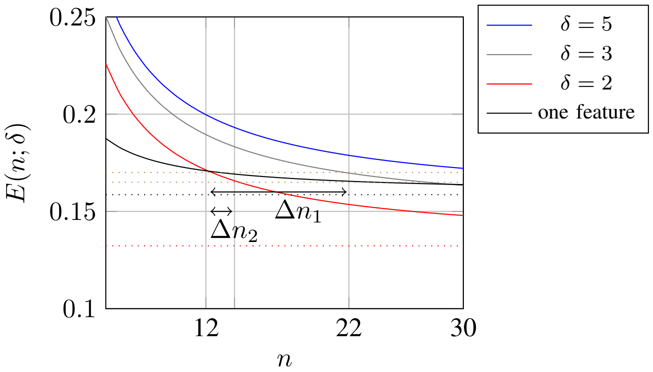

Fig. 2. For small values of n, it is beneficial to use a single feature. For intermediary values, samples can compensate for features. For large values of n, features and samples may be needed depending on the target error levels.

B. *How does error probability decay as sample size grows?*

Next, our goal is to analytically express how error probability decays as sample size grows. To that aim, we begin by considering two extremes, namely large and small sample sizes, and proceed by searching for a curve that best matches the decay of the error probability between such extremes.

Linear regression. We leverage the error probabilities observed for $8 \leq n \leq 1024$ to obtain, through least squares regression, the following approximation,

$$E_1(n) = \alpha_1 + \beta_1/n, \quad E_2(n;\delta) = \alpha_2(\delta) + \beta_2(\delta)/n. \tag{6}$$

Table I reports the values of the intercepts and slopes for the single-feature model and for various values of δ defining the two-feature model. As expected, the intercepts agree up to two decimal digits with the asymptotic values reported in the line corresponding to $n = 1,024$ in Table I. In addition, intercepts and slopes increase with δ as the problem becomes more challenging. Given these values, we readily obtain a closed-form expression to approximate the optimal threshold. It follows from (6) that $n^\star(\delta) = (\beta_2(\delta) - \beta_1)/(\alpha_1 - \alpha_2(\delta))$. Table I shows the values of $n^\star$ obtained from the above model, indicating its close agreement with SVM simulations.

Power law tail. For $\delta \geq 5$ the values of $n^\star$ are greater than 70, motivating an analysis of the rate at which error decays for large n. Figure 1 illustrates in log-log scale how the error probability decays, for $\delta = 7$ and $\gamma \in \{0.1, 1\}$. We empirically observed that the loss asymptotically decays towards Bayes error as n^{-1}. This rate is in agreement with [14], [25], [27], [28]: the loss decays as $O(1/n)$ in the "realizable case", and as $O(1/\sqrt{n})$ in the "agnostic case" [29]–[31].

C. *When can samples compensate for features?*

Figure 2, obtained using the proposed approximation, illustrates how the error probability decays as a function of the number of samples. In particular, it shows that for $\delta = 2$ it is worthwhile to leverage both features for $n \geq 12$ (in agreement with Table I). In addition, it also shows that an increase of n, by $\Delta n_2 = 2$, already causes a decrease of the error probability

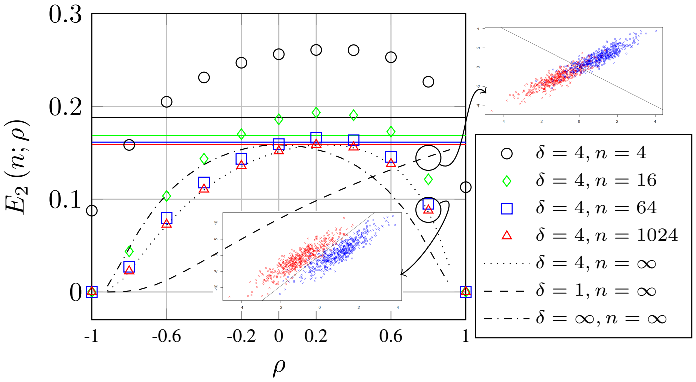

Fig. 3. Error as a function of ρ. Horizontal (resp., concave) lines correspond to one (resp., two) features. Error peaks at $\rho = 1/\delta$ (see Sec. III): for $\rho \approx 1/\delta$, the second feature is beneficial only for large n.

by 3.25%. To achieve a similar decrease with a single feature requires an increase in the number of samples by $\Delta n_1 = 10$. Note also that if one aims to decrease the error probability beyond 0.159 (horizontal dashed line in Fig. 2 corresponding to the asymptotic error probability with a single feature) one needs to use two features, eventually reaching the asymptotic error probability of 0.132 (also marked as an horizontal line).

Cost to transfer observations. Fig. 2 shows that the curves corresponding to error probabilities with a single feature and two features cross at a point whose abscissa, $n^\star$, grows with δ. In this paper, we assume that the cost of an observation does not depend on the number of features. Alternatively, observations containing two features may correspond to an increased cost, e.g., measured in bits to be transmitted through the network [32]. In that case, the unit of the x axis of Fig. 2 can be replaced by a cost, e.g., in bits, which impacts the curve corresponding to two features by stretching it horizontally. This, in turn, would cause $n^\star$ to move right, motivating the use of a single feature in a broader set of scenarios.

D. *Beyond independence: how ρ impacts $n^\star$?*

Next, we consider correlated features. When the covariance matrix of features given classes is the same for the two classes (see (2)) our numerical experiments indicate that as ρ increases from -1 to +1 the threshold $n^\star$ first increases and then decreases. This is in agreement with the concave behavior of $L(h^{(B)})$ characterized in Section III, and depicted in Fig. 3 for $\delta = 4$. In Fig. 3 the horizontal (resp., concave) lines correspond to the error obtained using a single feature (resp., two features). For $\rho = 1$ and $\rho = -1$ having the second feature is always helpful, i.e., $n^\star = 0$. As $\rho \to 1/\delta$, the second feature is beneficial for $n \geq 1024$.

Intuition. When $\rho = 1/\delta$ the best separator of the two classes corresponds to a vertical line, $X_1 = 0$, i.e., feature X_2 is useless. Indeed, recall that in general the best separator is $X_2 = \tilde{b}X_1$. As ρ approaches $1/\delta$ from below (resp., above), $\tilde{b}$ tends to $-\infty$ (resp., $+\infty$) and the contribution of X_2 to the best separator decreases, causing a corresponding increase in the classification error. When $\rho = \pm 1$, in contrast, the best slope equals $\tilde{b} = \pm\delta$, i.e., X_2 contributes to the best separator given by $X_1 = \pm X_2/\delta$ (see Section III).

Figure 3 also shows that when $\rho \neq 0$ an increase in δ may be beneficial. Indeed, for $\delta_1 < \delta_2$ and $\rho_0 = (\delta_1 + \delta_2)/(2\delta_1\delta_2)$ we have that $E_2(\infty; \delta_1, \rho) \lesseqgtr E_2(\infty; \delta_2, \rho)$ if $\rho \lesseqgtr \rho_0$. As illustrated in the scatter plots in Figure 3, if $\rho > 0$ an increase in the dispersion of the points may simplify their separation.

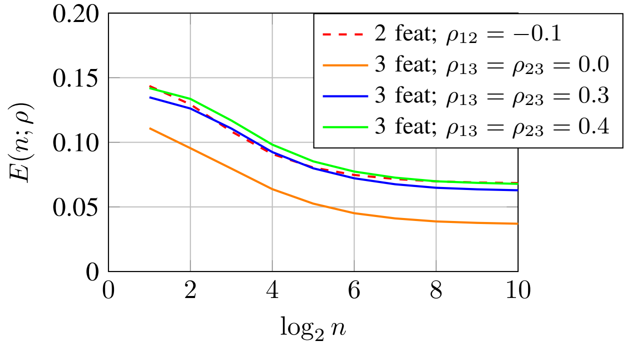

Fig. 4. Error as a function of n, varying $\rho_{13} = \rho_{23}$.

Different covariance matrices. When the covariance matrices are different, in contrast, the classification of samples may be exclusively based on covariances, rendering the two features necessary. This is the case, for instance, if features are functionally dependent on each other, e.g., $X_1 = -X_2$ and $X_1 = X_2$ for classes 1 and 2, respectively. In that case, $n^\star = 0$, samples cannot compensate for features and the classes are not linearly separable.

Multi-dimensional extension. Next, we consider the three dimensional case, i.e., with three features. We illustrate an example where behavior is similar to the two dimensional case, and another where we identify more complex behavior.

First, we consider $\sigma_1 = \sigma_2 = 1$, $\sigma_3 = 2$, and $\rho_{12} = 0$ while $\rho_{13} = \rho_{23}$. As the correlation between the third feature and others increases, it is beneficial to first sample two features, three features, and then two features again. This is similar in spirit to the behavior presented in Figure 3, although more difficult to visualize and interpret. We observed this pattern across a wide variety of parameters.

Second, we consider the case where $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\rho_{12} = -0.1$, and $\rho_{13} = \rho_{23} = 0.3$. In this scenario, determining whether the additional feature X_3 is useful cannot be resolved by a simple threshold strategy. Indeed, the error curves using 2 and 3 features, as functions of n, cross at two points: $n \approx 7$ and $n \approx 25$. When $\rho_{13} = \rho_{23} = 0.4$, a similar pattern emerges, with the error curves crossing once for $1 < n < 2$ and again at $n \approx 2^8$. These multiple crossings stand in stark contrast to the two-dimensional case, where at most one crossing was observed. A deeper analysis of this behavior is left for future work.

V. Discussion and Conclusion

Aiming at classification tasks, the decay of the error probability as a function of the sample size is at the core of statistical learning theory (SLT) [32]–[40]. In the realm of SLT, some of the questions addressed in this letter have been considered in more complex settings [8], [9], precluding a simple analysis *on the joint effect of variability and dependence.* Those works are complementary to ours. Our work is part of a research agenda on distributed learning under constraints. In this work we provide a principled understanding of the role of samples and features for classification. In [32] we considered estimation tasks, leaving additional tasks as subject for future work.

References

[1] S. Zhang, X. Kang, P. Duan, B. Sun, and S. Li, “Polygon structure-guided hyperspectral image classification with single sample for strong geometric characteristics scenes,” *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[2] F. He, K. Lv, J. Yang, and X. Huang, “One-Shot Distributed Algorithm for PCA With RBF Kernels,” *IEEE Signal Processing Letters*, vol. 28, pp. 1465–1469, 2021.

[3] B. Hanczar and E. R. Dougherty, “The reliability of estimated confidence intervals for classification error rates when only a single sample is available,” *Pattern Recognition*, vol. 46, no. 3, pp. 1067–1077, 2013.

[4] E. G. Miller, N. E. Matsakis, and P. A. Viola, “Learning from one example through shared densities on transforms,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, vol. 1. IEEE, 2000, pp. 464–471.

[5] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, “Generalizing from a few examples: A survey on few-shot learning,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[6] H. Noh, R. Rajagopal, and A. Kiremidjian, “Sequential structural damage diagnosis algorithm using a change point detection method,” *Journal of Sound and Vibration*, vol. 332, no. 24, pp. 6419–6433, 2013.

[7] P. Willett, P. F. Swaszek, and R. S. Blum, “The good, bad and ugly: distributed detection of a known signal in dependent gaussian noise,” *IEEE Transactions on signal processing*, vol. 48, no. 12, pp. 3266–3279, 2000.

[8] G. Hughes, “On the mean accuracy of statistical pattern recognizers,” *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[9] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, “Optimal number of features as a function of sample size for various classification rules,” *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005.

[10] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.

[11] A. Faragó and G. Lugosi, “Strong universal consistency of neural network classifiers,” *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1146–1151, 1993.

[12] P. Nakkiran, P. Venkat, S. M. Kakade, and T. Ma, “Optimal regularization can mitigate double descent,” in *International Conference on Learning Representations*, 2020.

[13] S. d’Ascoli, L. Sagun, and G. Biroli, “Triple descent and the two kinds of overfitting: where & why do they appear?” in *NeurIPS*, 2020.

[14] G. Blanchard, O. Bousquet, and P. Massart, “Statistical performance of support vector machines,” *The Annals of Statistics*, vol. 36, no. 2, pp. 489–531, 2008.

[15] L. Bottou and C.-J. Lin, “Support vector machine solvers,” *Large scale kernel machines*, vol. 3, no. 1, pp. 301–320, 2007.

[16] N. List and H. U. Simon, “SVM-optimization and steepest-descent line search,” in *Proceedings of the 22nd Annual Conference on Computational Learning Theory*. Citeseer, 2009.

[17] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, “A dual coordinate descent method for large-scale linear SVM,” in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 408–415.

[18] T. Glasmachers, “Universal consistency of multi-class support vector classification,” *Advances in Neural Information Processing Systems*, vol. 23, pp. 739–747, 2010.

[19] R. Vert, J.-P. Vert, and B. Schölkopf, “Consistency and convergence rates of one-class SVMs and related algorithms.” *Journal of Machine Learning Research*, vol. 7, no. 5, 2006.

[20] “R documentation,” 2021, https://www.rdocumentation.org/packages/e1071.

[21] R. M. Hogarth and N. Karelaia, “Ignoring information in binary choice with continuous variables: When is less “more”?” *Journal of Mathematical Psychology*, vol. 49, no. 2, pp. 115–124, 2005.

[22] D. G. Goldstein and G. Gigerenzer, “Fast and frugal forecasting,” *International journal of forecasting*, vol. 25, no. 4, pp. 760–772, 2009.

[23] P. Domingos, “A unified bias-variance decomposition,” in *Proceedings of 17th International Conference on Machine Learning*, 2000, pp. 231–238.

[24] Z. Deng, A. Kammoun, and C. Thrampoulidis, “A model of double descent for high-dimensional binary linear classification,” *Information and Inference: A Journal of the IMA*, 2021.

[25] M. Loog, T. Viering, and A. Mey, “Minimizers of the empirical risk and risk monotonicity,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 7478–7487, 2019.

[26] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[27] C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker, “Learning curves: Asymptotic values and rate of convergence,” in *Advances in Neural Information Processing Systems*, 1994, pp. 327–334.

[28] H. S. Seung, H. Sompolinsky, and N. Tishby, “Statistical mechanics of learning from examples,” *Physical review A*, vol. 45, no. 8, p. 6056, 1992.

[29] N. Srebro and S. Ben-David, “Learning bounds for support vector machines with learned kernels,” in *International Conference on Computational Learning Theory*. Springer, 2006, pp. 169–183.

[30] R. F. de Mello, C. Manapragada, and A. Bifet, “Measuring the shattering coefficient of decision tree models,” *Expert Systems with Applications*, vol. 137, pp. 443–452, 2019.

[31] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[32] Y.-Z. J. Chen, D. S. Menasché, and D. Towsley, “To collaborate or not in distributed statistical estimation with resource constraints?” in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2021, pp. 1–6.

[33] C. Giraud, “Introduction to high-dimensional statistics,” *Monographs on Statistics and Applied Probability*, vol. 139, p. 139, 2015.

[34] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data*. AMLBook New York, NY, USA:, 2012, vol. 4.

[35] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[36] R. F. de Mello, “On the shattering coefficient of supervised learning algorithms,” *arXiv preprint arXiv:1911.05461*, 2019.

[37] R. F. Mello and M. A. Ponti, *Machine learning: a practical approach on the statistical learning theory*. Springer, 2018.

[38] T. Cover and J. Thomas, *Elements of information theory*. John Wiley & Sons, 1999.

[39] J. Gareth, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: with applications in R*. Spinger, 2013.

[40] V. Berisha, A. Wisler, A. O. Hero, and A. Spanias, “Empirically estimable classification bounds based on a nonparametric divergence measure,” *IEEE Transactions on Signal Processing*, vol. 64, no. 3, pp. 580–591, 2015.