Abstract

Yes or no is perhaps the most common answer we provide each day. Indeed, binary answers to well structured questions are the building blocks of our knowledge. I started my research career in drafting such answers in various circumstances within the domains of statistics and machine learning. From statisticians and computer scientists’ point of view, classification is a well defined field. But more broadly, discretization is a powerful convention to help us understand the real-world social, economic and scientific situations. Also, the clean and tractable finite sample results from classification literature motivates me to investigate the explicit interplay among parameters in other fields. In this essay, I include my selected works regarding binary status in high dimensional statistics, statistical learning theory and social networks. In the first chapter, I introduce Regularized Optimal Affine Discriminant (ROAD), a high dimensional classification method explicitly using covariance information. In the second chapter, novel performance bounds of oracle type for asymmetric errors under the Neyman-Pearson context are derived. In the third chapter, I study the problem of information aggregation in social networks, where the focus is to determine aggregate learning status in any finite population network.
Acknowledgements

Firstly, I offer my sincerest gratitude towards my advisors Professor Jianqing Fan and Professor Phillippe Rigollet, who jointly advised me over the past four years. They have been very patient with my slow start and tolerant during my naive adventures; on one hand, they guided my research and helped me plan every important step, and, on the other hand, they always encouraged independence and creativity. I am also deeply thankful to my undergraduate advisor, Professor Don Fraser, who introduced me to the field of Statistics. Without the guidance and support from my advisors, my academic career would be drastically different and I would not be embarking on my current career path. Their knowledge and personality have deeply influenced my traits as a young researcher. In addition, I am honored to have Professor Erhan Cinlar and Professor Sebastien Bubeck on my dissertation committee; their comments provided critical insight improving my work at Princeton.

There are a few people in my family to whom I would like express my deepest appreciation. My late grandparents took care of me when I was young. I learned many from my parents, who are among the finest human beings on the planet, about how to be a good MAN before anything else. And my wife, Luo Xing, makes much sacrifice over the years. Furthermore, my extended family, including my uncles, aunts and cousins, all stand behind me in my pursuit of an academic career.

My coauthors (besides advisers) also deserve a thank you note. Professor Yang Feng at Columbia University taught me much data analysis, among other things. He also serves as a reader of my Ph.D. dissertation. Lei Qi, who is in finance industry currently, was a very considerate coauthor and coworker. Ming Yang, who is to become Professor of Finance at Duke University, is my most critical and sharp peer. Yao Zeng, Harvard Ph.D. student in Economics, motivated me in the least favorable situation. I enjoyed all these exciting collaborations.
Professor Jianqing Fan’s statistics lab provided a good environment for statistics research, even though Princeton University does not have a department dedicated to Statistics. I feel very proud and fortunate to be part of this lab.

Friendships at Princeton are memorable...
To my family.
# Contents

Abstract ............................................................... iii  
Acknowledgements ................................................... iv  

1 Regularized Optimal Affine Discriminant (ROAD) ................. 1  
  1.1 Introduction ...................................................... 1  
  1.2 Naive Bayes and Fisher Discriminant .......................... 5  
  1.3 Regularized Optimal Affine Discriminant ....................... 8  
      1.3.1 ROAD .................................................... 8  
      1.3.2 Variants of ROAD ...................................... 9  
  1.4 Constrained Coordinate Descent ............................... 11  
  1.5 Asymptotic Property ............................................. 14  
      1.5.1 Risk Approximation .................................... 14  
      1.5.2 Screening-based ROAD (S-ROAD) ....................... 16  
      1.5.3 Continuous Piecewise Linear Solution Path .......... 18  
  1.6 Numerical Investigation ........................................ 19  
      1.6.1 Equal Correlation Setting, Sparse Fixed Signal .......... 20  
      1.6.2 The Effect of $\gamma$ .................................. 23  
      1.6.3 Block Diagonal Correlation Setting, Sparse Fixed Signal 23  
      1.6.4 Block-Diagonal Negative Correlation Setting, Sparse Fixed Signal 25  
      1.6.5 Random Correlation Setting, Double Exponential Signal .... 26  
      1.6.6 Real Data .............................................. 26  
  1.7 Discussion ....................................................... 27
1.7.1 Extension to Multi-Class ........................................... 29
1.8 Proofs ................................................................. 30
  1.8.1 Proof of Theorem 1 ............................................. 30
  1.8.2 Proof of Theorem 2 ............................................. 33
  1.8.3 Proof of Theorem 5 ............................................. 34

2 Neyman-Pearson Classification .................................................. 37
  2.1 Introduction ................................................................ 37
  2.2 Binary classification .................................................. 38
    2.2.1 Classification risk and classifiers ............................. 38
    2.2.2 The Neyman-Pearson paradigm ................................. 41
  2.3 Neyman-Pearson classification via convex optimization .......... 44
    2.3.1 Conservativeness on type I error ............................. 44
    2.3.2 Convexified NP classifier ...................................... 46
  2.4 Performance Bounds .................................................... 47
    2.4.1 Control of type I error .......................................... 48
    2.4.2 Simultaneous control of the two errors ...................... 48
    2.4.3 A Different Sampling Scheme ................................ 50
    2.4.4 Price to pay for being conservative ......................... 52
  2.5 Chance constrained optimization .................................... 55
  2.6 Appendix ............................................................... 60
    2.6.1 Proof of Theorem 2.4.1 ....................................... 60
    2.6.2 Proof of Proposition 2 .......................................... 63
    2.6.3 Proof of Theorem 2.4.2 ....................................... 64
    2.6.4 Proof of Corollary 2.4.1 ...................................... 66
    2.6.5 Proof of Theorem 2.4.3 ....................................... 67
    2.6.6 Proof of Proposition 3 .......................................... 68
    2.6.7 Technical lemmas on Binomial distributions ................ 70
3 Learning in Social Networks

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>75</td>
</tr>
<tr>
<td>3.2 Model</td>
<td>82</td>
</tr>
<tr>
<td>3.3 Finite Population Learning</td>
<td>89</td>
</tr>
<tr>
<td>3.4 Perfect Learning and the Rates</td>
<td>94</td>
</tr>
<tr>
<td>3.4.1 Perfect Learning</td>
<td>95</td>
</tr>
<tr>
<td>3.4.2 Learning Rates</td>
<td>100</td>
</tr>
<tr>
<td>3.5 Learning under Multiple Equilibria</td>
<td>104</td>
</tr>
<tr>
<td>3.6 Remarks and Further Research</td>
<td>105</td>
</tr>
<tr>
<td>3.7 Appendix</td>
<td>107</td>
</tr>
</tbody>
</table>

Bibliography 119
Chapter 1

Regularized Optimal Affine Discriminant (ROAD)

1.1 Introduction

Technological innovations have had deep impact on society and on various areas of scientific research. High-throughput data from microarray and proteomics technologies are frequently used in many contemporary statistical studies. In the case of microarray data, the dimensionality is frequently in thousands or beyond, while the sample size is typically in the order of tens. The large-p-small-n scenario poses challenges for the classification problems. We refer to [34] for an overview of statistical challenges associated with high dimensionality.

When the feature space dimension $p$ is very high compared to the sample size $n$, the Fisher discriminant rule performs poorly due to diverging spectra as demonstrated by [12]. These authors showed that the independence rule in which the covariance structure is ignored performs better than the naive Fisher rule (NFR) in the high dimensional setting. [31] demonstrated further that even for the independence rules, a procedure using all the features can be as poor as random guessing due to noise accumulation in estimating population centroids in high-dimensional feature space. As a result, [31] proposed the Features Annealed Independence Rule (FAIR) that selects a subset of important features.
for classification. [29] reported that for microarray data, ignoring correlations between genes leads to better classification results. [74] proposed the Nearest Shrunken Centroid (NSC) which likewise employs the working independence structure. Similar problems are also studied in the machine learning community such as [27] and [54].

In microarray studies, correlation among different genes is an essential characteristic of the data and usually not negligible. Other examples include proteomics, and metabolomics data where correlation among biomarkers is commonplace. More details can be found in [5]. Intuitively, the independence assumption among genes leads to loss of critical information and hence is suboptimal. We believe that in many cases, the crucial point is not whether to consider correlations, but how we can incorporate the covariance structure into the analysis with a bullet proof vest against diverging spectra and significant noise accumulation effect.

The setup of the objective classification problem is now introduced. We assume in the following that the variability of data under consideration can be described reasonably well by the means and variances. To be more precise, suppose that random variables representing two classes \( C_1 \) and \( C_2 \) follow \( p \)-variate normal distributions: \( X|Y=1 \sim \mathcal{N}_p(\mu_1, \Sigma) \) and \( X|Y=2 \sim \mathcal{N}_p(\mu_2, \Sigma) \) respectively. Moreover, assume \( P(Y=1) = 1/2 \). This Gaussian discriminant analysis setup is known for its good performance despite its rigid model structure. For any linear discriminant rule

\[
\delta_w(X) = \mathbb{I}\{w^T(X - \mu_a) > 0\},
\]

where \( \mu_a = (\mu_2 + \mu_1)/2 \), and \( \mathbb{I} \) denotes the indicator function with value 1 corresponds to assigning \( X \) to class \( C_2 \) and 0 class \( C_1 \), the misclassification rate of the (pseudo) classifier \( \delta_w \) is

\[
W(\delta_w) = \frac{1}{2}P_2(\delta_w(X) = 0) + \frac{1}{2}P_1(\delta_w(X) = 1) = 1 - \Phi(w^T\mu_d/(w^T\Sigma w)^{1/2}),
\]

where \( \mu_d = (\mu_2 - \mu_1)/2 \), and \( P_i \) is the conditional distribution of \( X \) given its class label \( i \). We will focus on such linear classifier \( \delta_w(\cdot) \), and the mission is to find a good data
projection direction \( \mathbf{w} \). Note that the Fisher discriminant

\[
\delta_F(\mathbf{X}) = \mathbb{I}\{(\Sigma^{-1} \mu_d)^T (\mathbf{X} - \mu_a) > 0\}
\]

is the Bayes rule. There are two fundamental difficulties in applying the Fisher discriminant whose missclassification rate is

\[
1 - \Phi \left( (\mu_d^T \Sigma^{-1} \mu_d)^{1/2} \right).
\]

The first difficulty arises from the noise accumulation effect in estimating the population centroids \([31]\) when \( p \) is large. The second challenge is more severe: estimating the inverse of covariance matrix \( \Sigma \) when \( p > n \) \([12]\). As a result, much previous researches focus on the independence rules, which act as if \( \Sigma \) is diagonal. However, correlation matters!

To illustrate this point, consider a case when \( p = 2 \). These two features can be selected from the original thousands of features, and we can estimate the correlation between two variables with reasonable accuracy. Let

\[
\Sigma = \begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix},
\]

where \( \rho \in [0, 1) \) and \( \mu_d = (\mu_1, \mu_2)^T \). Without loss of generality, assume \( |\mu_1| \geq |\mu_2| > 0 \). The misclassification rate of Fisher discriminant depends on

\[
\Delta_p(\rho) = \mu_d^T \Sigma^{-1} \mu_d = \frac{1}{1 - \rho^2}(\mu_1^2 + \mu_2^2 - 2\rho \mu_1 \mu_2).
\]

Note that

\[
\Delta_p'(\rho) > 0 \Leftrightarrow \mu_1 \mu_2 \rho^2 - (\mu_1^2 + \mu_2^2) \rho + \mu_1 \mu_2 < 0.
\]

Therefore, when \( \mu_1 \mu_2 < 0 \), \( \Delta_p'(\rho) > 0 \) for all \( \rho \in [0, 1) \). On the other hand, when \( \mu_1 \mu_2 > 0 \), \( \Delta_p(\rho) \) decreases on \( \rho \in (0, \frac{\mu_2}{\mu_1}) \), and increases on \( (\frac{\mu_2}{\mu_1}, 1) \). Notice that when \( \rho \to 1 \), \( \Delta_p \to \infty \) regardless of signs for \( \mu_1 \mu_2 \), which in turn leads to vanishing classification error. On the other hand, if we use independence rule (also called naive Bayes rule), the optimal
misclassification rate

\[ 1 - \Phi \left( \frac{\|\mu_d\|_2^2}{(\mu_d^T \Sigma \mu_d)^{1/2}} \right) \]  

(1.6)

depends on \( \Gamma(\rho) = \|\mu_d\|_2^4 / \mu_d^T \Sigma \mu_d \), which is monotonically decreasing for \( \rho \in [0, 1) \), with the limit \( (\mu_1^2 + \mu_2^2)^2 / (\mu_1 + \mu_2)^4 \) that is smaller than unity when \( \mu_1 \) and \( \mu_2 \) have the same sign. Hence, the optimal classification error using the independence rule actually increases as correlation among features increases.

The above simple example shows that by incorporating correlation information, the gain in terms of classification error can be substantial. Elaboration on this point in more realistic scenarios is provided in Section 2. Now it seems wise to use at least a part of covariance structure to improve the performance of a classifier. So there is a need to estimate the covariance matrix \( \Sigma \). Without structural assumptions on \( \Sigma \), the pooled sample covariance \( \hat{\Sigma} \) is one natural estimate. But for \( p > n \), it is not considered as a good estimate of \( \Sigma \) in general. We are lucky here because our mission is not constructing a good estimate of the covariance matrix, but finding a good direction \( w \) that leads to a good classifier. To mimic the optimal data projection direction \( \Sigma^{-1}\mu_d \), we do not adopt a direct plug-in approach, simply because it is unlikely that a product is a good estimate when at least one of its components is not. Instead, we find the data projection direction \( w \) by directly minimizing the classification error subject to a capacity constraint on \( w \).

From a broad spectrum of simulated and real data analysis, we are convinced that this approach leads to a robust and efficient sparse linear classifier.

Admittedly, our work is far from the first to use covariance for classification; support vector machines \([77]\), for example, implicitly utilize covariance between covariates. Another notable work is “shrunken centroids regularized discriminant analysis” (SCRDA) \([44]\), which calls for a version of regularized sample covariance matrix \( \hat{\Sigma}_{reg} \), and soft-thresholds on \( \hat{\Sigma}_{reg}^{-1} \hat{x}_i \). \([70]\) consider a sparse linear discriminant analysis, assuming the sparsity on both the covariance matrix and the mean difference vector so that they can be regularized. They show that such a regularized estimator is asymptotically optimal under some conditions. However, to the best of our knowledge, this work is the first
to select features by directly optimizing the misclassification rates, to explicitly use un-
regularized sample covariance information, and to establish the oracle inequality and risk
approximation theory.

There is a huge literature on high dimensional classification. Examples include prin-
cipal component analysis in [7] and [83], partial least squares in [60], [48] and [15], and
sliced inverse regression in [55] and [6].

The rest of our paper is organized as follows. Section 1.2 provides some insights on
the performances of naive Bayes, Fisher discriminant and restricted Fisher discriminants.
In Section 1.3, we propose the Regularized Optimal Affine Discriminant (ROAD) and
variants of ROAD. An efficient algorithm Constrained Coordinate Descent (CCD) is con-
structed in Section 1.4. Main risk approximation results and continuous piecewise linear
property of the solution path are established in Section 1.5. We conduct simulation and
empirical studies in Section 1.6. A discussion is given in Section 1.7, and all proofs are
relegated to the appendix.

1.2 Naive Bayes and Fisher Discriminant

To compare the naive Bayes and Fisher discriminant at the population level, we assume
without loss of generality that variables have been marginally standardized so that $\Sigma$ is a
correlation matrix. Recall that the naive Bayes discriminant has error rate (1.6) and the
Fisher discriminant has error rate (1.4). Let $\Gamma_p = \|\mu_d\|^2/\mu_d^T \Sigma \mu_d$ and $\Delta_p = \mu_d^T \Sigma^{-1} \mu_d$.
Denote by $\{\lambda_i\}^p_{i=1}$ the eigenvalues and $\{\xi_i\}^p_{i=1}$ eigenvectors of the matrix $\Sigma$. Decompose

$$\mu_d = a_1 \xi_1 + \cdots + a_p \xi_p, \quad (1.7)$$

where $\{a_i\}^p_{i=1}$ are the coefficients of $\mu_d$ in this new orthonormal basis $\{\xi_i\}^p_{i=1}$. Using the
decomposition (1.7), we have

$$\Delta_p = \sum_{j=1}^p a_j^2 / \lambda_j, \quad \Gamma_p = \left( \sum_{j=1}^p a_j^2 \right)^2 / \sum_{j=1}^p \lambda_j a_j^2. \quad (1.8)$$
The relative efficiency of Fisher discriminant over naive Bayes is characterized by \( \Delta_p/\Gamma_p \).

By the Cauchy-Schwartz inequality,

\[
\Delta_p/\Gamma_p \geq 1.
\]

The naive Bayes method performs as well as the Fisher discriminant only when \( \mu_d \) is an eigenvector of \( \Sigma \).

In general, \( \Delta_p/\Gamma_p \) can be much larger than unity. Since \( \Sigma \) is the correlation matrix, \( \sum_{j=1}^{p} \lambda_j = \text{tr}(\Sigma) = p \). If \( \mu_d \) is equally loaded on \( \xi_j \), then the ratio

\[
\Delta_p/\Gamma_p = p^{-2} \sum_{j=1}^{p} \lambda_j \sum_{j=1}^{p} \lambda_j^{-1} = p^{-1} \sum_{j=1}^{p} \lambda_j^{-1}.
\]

(1.9)

More generally, if \( \{a_j\}_{j=1}^{p} \) are realizations from a distribution with the second moment \( \sigma^2 \), then by the law of large numbers,

\[
\sum_{j=1}^{p} a_j^2 \lambda_j^{-1} \approx \sigma^2 \sum_{j=1}^{p} 1/\lambda_j, \quad p^{-1} \sum_{j=1}^{p} a_j^2 \approx \sigma^2, \quad \sum_{j=1}^{p} \lambda_j a_j^2 \approx \sigma^2 \sum_{j=1}^{p} \lambda_j.
\]

Hence, (1.9) holds approximately in this case. In other words, the right hand side of (1.9) is approximately the relative efficiency of the Fisher discriminant over the naive Bayes. Now suppose further that half of the eigenvalues of \( \Sigma \) are \( c \) and the other half are \( 2-c \). Then, the right hand side of (1.9) is \( (c^{-1} + (2-c)^{-1})/2 \). For example when the condition number is 10, this ratio is about 3. A high ratio translates into a large difference in error rates: \( 1 - \Phi(\Gamma_p^{1/2}) \) for independence rule is much larger than \( 1 - \Phi(3\Gamma_p^{1/2}) \) for Fisher discriminant. For example, when \( \Gamma_p^{1/2} = 0.5 \), we have 30.9% and 6.7% error rates respectively for the naive Bayes and Fisher discriminant.

To put the above arguments under a visual inspection, consider a case in which \( p = 1000, \mu_d = (\mu_s^T, 0^T)^T \) with \( \mu_s = (1, 1, 1, 1, 2, 2, 2, 2, 2)^T/10 \) and \( \Sigma \) equals the equi-correlation matrix with pairwise correlation \( \rho \). The vector \( \mu_d \) simulates the case in which 10 genes out of 1000 express mean differences. Figure 1.1 depicts the theoretical error rates of the Fisher discriminant and the naive Bayes rule as functions of \( \rho \).
Figure 1.1: Misclassification rates of Fisher discriminant, naive Bayes and restricted Fisher rules (10 and 20 features, respectively) against $\rho$.

It is not surprising that the Fisher discriminant rule performs significantly better than the naive Bayes as $\rho$ deviates away from zero. The error rate of the naive Bayes actually increases with $\rho$, whereas the error rate of the Fisher discriminant tends to zero as $\rho$ approaches 1. This phenomenon is the same as what was shown analytically through the toy example in Section 1. To mimic Fisher discriminant by a plug-in estimator, we need to estimate $\Sigma^{-1}\mu_d$ with reasonable accuracy. This mission is difficult if not impossible. On the other hand, imitating a weaker oracle is more manageable. For example, when the samples are of reasonable size, we can select the 10 variables with differences in means by applying a two-sample $t$-test. Restricting to the best linear classifiers based on these $s = 10$ variables, we have the optimal error rate

$$1 - \Phi((\mu^T_s \Sigma^{-1}_s \mu_s)^{1/2}),$$

where the classification rule is $\delta_{w^R}$ and $w^R = ((\Sigma^{-1}_s \mu_s)^T, 0^T)^T$. The performance of this oracle classifier is depicted by the sub-Fisher (10 features) in Figure 1.1. It performs much better than the naive Bayes method. One can also employ the naive Bayes rule to the restricted feature space, but this method has exactly the same performance as the naive Bayes method in the whole space. Thus, the restricted Fisher discriminant outperforms
both the naive Bayes method with restricted features and the naive Bayes method using all features.

Mimicking the performance of the restricted Fisher discriminant is feasible. Instead of estimating a $1000 \times 1000$ covariance matrix, we only need to gauge a $10 \times 10$ submatrix. However, this restricted Fisher rule is not powerful enough, as shown in Figure 1. We can improve its performance by including 10 most correlated variables to each of those selected features to further account for the correlation effect, giving rise to a 20-dimensional feature space. Since the variables are equally correlated in this example, we are free to choose any 10 variables among the other 990. The performance of such an enlarged restricted Fisher discriminant is represented by sub-Fisher (20 features) in Figure 1. It performs closely to the Fisher discriminant which uses the whole feature space, and it is feasible to implement with finite samples.

1.3 Regularized Optimal Affine Discriminant

The misclassification rate of Fisher discriminant is $1 - \Phi(\Delta_p^{1/2})$, where $\Delta_p = \mu_d^T \Sigma^{-1} \mu_d$. However, for high dimensional data, it is impossible to achieve such a performance empirically. Among other reasons, the estimated covariance matrix $\hat{\Sigma}$ is ill-conditioned or not invertible. One solution is to focus only on the $s(<< p)$ most important features for classification. Ideally, the best $s$ features should be the ones with the largest $\Delta_s$ among all $\binom{p}{s}$ possibilities, where $\Delta_s$ is the counterpart of $\Delta_p$ when only $s$ variables are considered. Naive search for the best subset of size $s$ is NP-hard. Thus, we develop a regularized method to circumvent these two problems.

1.3.1 ROAD

Recall that by (1.2), minimizing the classification error $W(\delta_w)$ is the same as maximizing $w^T \mu_d / (w^T \Sigma w)^{1/2}$, which is equivalent to minimizing $w^T \Sigma w$ subject to $w^T \mu_d = 1$. We would like to add a penalty function for capacity control. There are many ways to do regularization; for the literature on penalized methods, refer to LASSO [75], SCAD [33],
Elastic net [82], MCP [79] and related methods [81, 84]. As our primary interest is classification error (the risk of the procedure), an $L_1$ constraint $\|w\|_1 \leq c$ is added for regularization, so the problem can be recast as

$$w_c = \arg\min_{\|w\|_1 \leq c, w^T\mu_d = 1} w^T\Sigma w.$$  \hspace{1cm} (1.10)

We name the classifier $\delta_{w_c}(\cdot)$ the Regularized Optimal Affine Discriminant (ROAD). The existence of a feasible solution in (1.10) dictates

$$c \geq 1/\max_{1 \leq i \leq p} |\mu_{d,i}|.$$ \hspace{1cm} (1.11)

When $c$ is small, we obtain a sparse solution and achieve feature selection using covariance information. When $c \geq \|\Sigma^{-1}\mu_d\|_1/\mu_d^T\Sigma^{-1}\mu_d$, the $L_1$ constraint is no longer binding and $\delta_{w_c}$ reduces to the Fisher discriminant, which can be denoted by $\delta_{w_c} (\approx \delta_F)$. Therefore we have provided a family of linear discriminants, indexed by $c$, using only one feature to all features. In some applications such as portfolio selection, the choice of $c$ reflects the investor’s tolerance upper bound on gross exposure. In other applications, when the user does not have a such a preference, the choice of $c$ can be data-driven. To accommodate both application scenarios, we propose a coordinate descent algorithm (Section 1.4) to implement our ROAD proposal.

### 1.3.2 Variants of ROAD

At the sample level, NSC [74] and FAIR [31] both use shrunken versions of standardized mean difference to find the $s$ features. In the same spirit, we consider the following Diagonal Regularized Optimal Affine Discriminant (D-ROAD) $\delta_{w_c}$, where

$$w_c = \arg\min_{\|w\|_1 \leq c, w^T\mu_d = 1} w^T\text{diag}(\Sigma)w.$$ \hspace{1cm} (1.12)
The D-ROAD will be compared with NSC [74] and FAIR [31] in the simulation studies, and all these independence based rules will be compared with ROAD and its two variants defined below.

A screening-based variant (to be proposed) of ROAD aims at mimicking the performance of sub-Fisher (10 features) in Figure 1. A fast way to select features is the independence screening, which uses the marginal information such as the two-sample $t$-test. We can also enlarge the selected feature subspace by incorporating the features which are most correlated to what have been chosen. This additional variant of ROAD tracks the performance of sub-Fisher (20 features) in Figure 1. We will refer to the two variants of ROAD as S-ROAD1 and S-ROAD2. More description of these procedures, along with their theoretical properties and numerical investigations, will be detailed in Sections 5 and 6.

A hint of the rationale behind including correlated features that do not show a difference in means between the two classes, is revealed through the two-feature example in the introduction. Suppose $\mu_2 = 0$. Then, by (1.5), the power of the discriminant using two features is $1 - \Phi(\Delta_2^{1/2})$ where $\Delta_2 = \mu_1^2/(1 - \rho^2)$, whereas with the first feature alone the misclassification rate is $1 - \Phi(\Delta_1^{1/2})$ where $\Delta_1 = \mu_1^2$. Therefore when the correlation $|\rho|$ is large, using two correlated features is far more powerful than employing only one feature, even though the second feature has no marginal discrimination power. More intuition is granted by this observation: at the population level, the best $s$ features are not necessarily those with largest standardized mean differences. In other words, with the two class Gaussian model in mind, when $\Sigma$ is the correlation matrix, the most powerful $s$ features for classification are not necessarily the coordinates of $\mu_d$ with largest absolute values. This is illustrated by the next stylized example.
Let \( X|Y = 0 \sim \mathcal{N}(\mu_1, \Sigma) \) and \( X|Y = 1 \sim \mathcal{N}(\mu_2, \Sigma) \), where \( \mu_1 = (0, 0, 0)^T \), \( \mu_2 = (4, 0.5, 1)^T \), and
\[
\Sigma = \begin{pmatrix}
1 & -0.25 & 0 \\
-0.25 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}.
\]
Suppose the objective is to choose 2 out of 3 variables for classification. If we rank features by marginal information, for example by the absolute value of standardized mean differences, then we would choose the 1st and 3rd features. On the other hand, denote \( \mu_{d,ij} \) the mean difference vector for features \( i \) and \( j \), \( \Sigma_{ij} \) the covariance matrix of features \( i \) and \( j \), then the classification power using features \( i \) and \( j \) depends on \( \Gamma_{ij} = \mu_{d,ij}^{T} \Sigma_{ij}^{-1} \mu_{d,ij} \).
Simple calculation leads to
\[
\Gamma_{12} = 18.4 > 17 = \Gamma_{13}.
\]
Hence the most powerful two features for classification are not the 1st and 3rd.

### 1.4 Constrained Coordinate Descent

With a Lagrangian argument, we reformulate problem (1.10) as
\[
\w_{\lambda} = \arg\min_{w^T \mu_i = 1} \frac{1}{2} w^T \Sigma w + \lambda \|w\|_1.	ag{1.13}
\]
In this section, we propose a Constrained Coordinate Descent (CCD) algorithm that is tailored for solving our minimization problem with linear constraints. Optimization (1.13) is a constrained quadratic programming problem and can be solved by existing softwares such as MOSEK. Although these softwares are well regarded in practice, they are slow for our application. The structure of (1.13) could be exploited in order to obtain a more efficient algorithm. In line with the LARS algorithm, we will exploit the fact that the solution path has a piecewise-linear property.

In the compressed sensing literature, it is common to replace an affine constraint by a quadratic penalty. We borrow this idea and consider the following approximation to
\[ \tilde{w}_{\lambda, \gamma} = \arg\min_{w} \frac{1}{2} w^T \Sigma w + \lambda \|w\|_1 + \frac{1}{2} \gamma (w^T \mu_d - 1)^2. \]  

(1.14)

In practice, we replace \( \Sigma \) by the pooled sample covariance \( \hat{\Sigma} \) and \( \mu \) by the sample mean difference vector \( \hat{\mu}_d \). By Theorem 6.7 in [65], we have

\[ \tilde{w}_{\lambda, \gamma} \to \bar{w}_\lambda \text{ when } \gamma \to \infty. \]

Note that we do not have to enforce the affine constraint strictly, because it only serves to normalize our problem. In the optimization problem (1.14), when \( \lambda = 0 \), the solution \( \tilde{w}_{0, \gamma} \) is always in the direction of \( \Sigma^{-1} \mu_d \), the Fisher discriminant, regardless of the value of \( \gamma \). In addition, this observation is confirmed in the data analysis (Section 6.2) by the insensitivity of choice for \( \gamma \). Therefore we hold \( \gamma \) as a constant in practice.

We solve (1.14) by coordinate descent. Non-gradient algorithms seem to be less popular for convex optimization. For instance, the popular textbook *Convex Optimization* by [16] does not even have a section on these methods. Coordinate descent method is an algorithm, in which the \( p \) search directions are just unit vectors \( e_1, \ldots, e_p \), where \( e_i \) denotes the \( i \)th element in the standard basis of \( \mathbb{R}^p \). These unit vectors are used as search directions in each search cycle until some convergence criterion is met.

What makes the coordinate descent algorithm particularly attractive for (1.14) is that there is an explicit formula for each coordinate update. For a given \( \gamma \), fix \( \tau \) and \( K \), then do the optimization on a grid (of log-scale) of \( \lambda \) values: \( \tau \lambda_{\max} = \lambda_K < \lambda_{K-1} < \cdots < \lambda_1 = \lambda_{\max} \). The \( \lambda_{\max} \) is the minimum \( \lambda \) value such that no variables enter the model; this is analogous to the minimum requirement on \( c \) in (1.11). In our implementation, we take \( \tau = 0.001 \) and \( K = 100 \). The problem is solved backwards from \( \lambda_{\max} \). When \( \lambda = \lambda_{i+1} \), we use the solution from \( \lambda = \lambda_i \) as the initial value. This kind of “warm start” is very effective in improving computational efficiency.

Consider a coordinate descent step to solve (1.14). Without loss of generality, suppose that \( \tilde{w}_j \) for all \( j \geq 2 \) are given, and we need to optimize with respect to \( w_1 \). The objective

\[ \tilde{w}_{\lambda, \gamma} \to \bar{w}_\lambda \text{ when } \gamma \to \infty. \]
function now becomes
\[ g(w_1) = \frac{1}{2} \left( w_1^T \bar{w}_2 \right) \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} w_1 \\ \bar{w}_2 \end{pmatrix} + \lambda |w_1| + \lambda |\bar{w}_2|_1 + \frac{1}{2} \gamma ( w^T \mu_d - 1)^2. \]

When \( w_1 \neq 0 \), we have
\[
g'(w_1) = \Sigma_{11} w_1 + \Sigma_{12} \bar{w}_2 + \lambda \text{sign}(w_1) + \gamma ( w^T \mu_d - 1) \mu_{d1}
\]
\[
= (\Sigma_{11} + \gamma \mu_{d1}^2) w_1 + (\Sigma_{12} + \gamma \mu_{d1} \mu_{d2}^T) \bar{w}_2 + \lambda \text{sign}(w_1) - \gamma \mu_{d1}.
\]

By simple calculation [28], the coordinate-wise update has the form
\[
\tilde{w}_1 = S \left( \frac{\gamma \mu_{d1} - (\Sigma_{12} + \gamma \mu_{d1} \mu_{d2}^T) \bar{w}_2}{\Sigma_{11} + \gamma \mu_{d1}^2} \right),
\]
where \( S(z, \lambda) = \text{sign}(z)(|z| - \lambda)^+ \) is the soft-thresholding operator.

Now, we consider the convergence property of the coordinate descent algorithm. Here, although the objective function is not strictly convex, it is strictly convex in each of the coordinates.

To show \( g(w_1) \) is strictly convex in \( w_1 \), we decompose it as follows:
\[
g(w_1) = g_1(w_1) + g_2(w_1),
\]
where \( g_2(w_1) = \lambda |w_1| \) and
\[
g_1(w_1) = \frac{1}{2} \left( w_1^T \bar{w}_2 \right) \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} w_1 \\ \bar{w}_2 \end{pmatrix} + \lambda |\bar{w}_2|_1 + \frac{1}{2} \gamma ( w^T \mu_d - 1)^2.
\]

Note that \( g_1(w_1) \) is a quadratic function of \( w_1 \) and \( g_1''(w_1) = \Sigma_{11} + \gamma \mu_{d1}^2 > 0 \) for all \( w_1 \in \mathbb{R} \). Therefore, the function \( g_1(\cdot) \) is strictly convex on \( \mathbb{R} \). Also, it is clear that \( g_2 \) is convex on \( \mathbb{R} \). Therefore \( g = g_1 + g_2 \) is a strictly convex function on \( \mathbb{R} \).

Combining the coordinate-wise strict convexity with the fact that the non-differentiable part of the objective function is separable, Theorem 5.1 of [76] guarantees
that coordinate descent algorithms converge to coordinate-wise minima. Moreover, since all directional derivatives exist, every coordinate-wise minimum is also a local minimum. A similar study on the convergence of the coordinate descent algorithm can be found in [18].

In each coordinate update, the computational complexity is $O(p)$. A complete cycle through all $p$ variables costs $O(p^2)$ operations. From our experience, CCD converges quickly after a few cycles if “warm start” is used for the initial solution. Let $C$ denote the average number of cycles until convergence for each $\lambda$. Then our algorithm CCD enjoys computational complexity $O(Ckp^2)$. The D-ROAD can be similarly implemented by replacing the covariance matrix with its diagonal.

### 1.5 Asymptotic Property

#### 1.5.1 Risk Approximation

Let $\hat{w}_c$ be a sample version of $w_c$ in (1.10),

$$
\hat{w}_c \in \arg\min_{\|w\|_1 \leq c, w^T \mu_d = 1} w^T \tilde{\Sigma} w. \tag{1.15}
$$

The fact that $\tilde{\Sigma}$ is only positive semi-definite leads to potential non-uniqueness of $\hat{w}_c$.

Now, we have three different classifiers: $\delta_{w_{\infty}} = \mathbb{I}\{w_{\infty}^T(X - \mu_a) > 0\}$, $\delta_{w_c} = \mathbb{I}\{w_c^T(X - \mu_a) > 0\}$ and $\hat{\delta}_{w_c} = \mathbb{I}\{\hat{w}_c^T(X - \hat{\mu}_a) > 0\}$. The first two are oracle classifiers, requiring the knowledge of unknown parameters $\mu_1$, $\mu_2$ and $\Sigma$, while the third one is the feasible classifier, ROAD, based on the sample. Their classification errors are given by (1.2). Explicitly, the error rates are respectively $W(\delta_{w_{\infty}})$ [see (1.4)], $W(\delta_{w_c})$, and $W(\hat{\delta}_{w_c})$. By (1.2), an obvious estimator of the misclassification rate of $\hat{\delta}_{w_c}$ is

$$
W_n(\hat{\delta}_{w_c}) = 1 - \Phi \left( \frac{\hat{w}_c^T \mu_d}{(\hat{w}_c^T \tilde{\Sigma} \hat{w}_c)^{1/2}} \right). \tag{1.16}
$$

Two questions arise naturally:

(i) how close is $W(\hat{\delta}_{w_c})$, the misclassification error of $\hat{\delta}_{w_c}$, to that of its oracle $W(\delta_{w_c})$?
(ii) does \( W_n(\hat{\delta}_{\omega_c}) \) estimate \( W(\hat{\delta}_{\omega_c}) \) well?

Theorem 1 addresses these two questions. We introduce an intermediate optimization problem for convenience:

\[
\mathbf{w}_c^{(1)} = \underset{\| \mathbf{w} \| \leq c, \mathbf{w}^T \hat{\mu}_d = 1}{\text{argmin}} \mathbf{w}^T \Sigma \mathbf{w}.
\]

**Theorem 1.** Let \( s_c = \| \mathbf{w}_c \|_0 \), \( s_c^{(1)} = \| \mathbf{w}_c^{(1)} \|_0 \), and \( \hat{s}_c = \| \hat{\mathbf{w}}_c \|_0 \). Assume that \( \lambda_{\text{min}}(\Sigma) \geq \sigma_0^2 > 0 \), \( \| \Sigma - \bar{\Sigma} \|_\infty = O_p(a_n) \) and \( \| \hat{\mu}_d - \mu_d \|_\infty = O_p(a_n) \) for a given sequence \( a_n \rightarrow 0 \).

Then, we have

\[
W(\hat{\delta}_{\omega_c}) - W(\delta_{\omega_c}) = O_p(d_n),
\]

and

\[
W_n(\hat{\delta}_{\omega_c}) - W(\hat{\delta}_{\omega_c}) = O_p(b_n),
\]

where \( b_n = (c^2 \lor s_c \lor s_c^{(1)}) a_n \) and \( d_n = b_n \lor (\hat{s}_c a_n) \).

**Remark 1.** In Theorem 1, \( \| \cdot \|_\infty \) is the element wise super-norm. When \( \hat{\Sigma} \) is the sample covariance, under some mild moment conditions, we have \( \| \hat{\Sigma} - \Sigma \|_\infty = O_p(\sqrt{(\log p)/n}) \); hence we can take \( a_n = \sqrt{(\log p)/n} \). The first result in Theorem 1 shows the difference between the misclassification rate of \( \hat{\delta}_{\omega_c} \) and its oracle version \( \delta_{\omega_c} \); the second result says about the error in estimating the true misclassification rate of ROAD.

**Remark 2.** In view of (1.2), one intends to choose a \( \mathbf{w} \) that makes \( \mathbf{w}^T \Sigma \mathbf{w} \) small and \( \mathbf{w}^T \mu_d \) large. A compromise of these dual objectives leads to a utility function

\[
U(\mathbf{w}) = -\mathbf{w}^T \Sigma \mathbf{w} + \xi \mathbf{\mu}_d^T \mathbf{w},
\]

as a proxy of the objective function (1.2) for a fixed \( \xi \). For any \( \xi > 0 \), the optimal choice \( \mathbf{w}^* \in \text{argmin} \; U(\mathbf{w}) \) leads to the Fisher discriminant rule. Consider also the regularized versions

\[
\mathbf{w}_c^* = \underset{\| \mathbf{w} \| \leq c}{\text{argmin}} U(\mathbf{w}), \quad \text{and} \quad \hat{\mathbf{w}}_c^* = \underset{\| \mathbf{w} \| \leq c}{\text{argmin}} \hat{U}(\mathbf{w}),
\]

where \( \hat{U}(\mathbf{w}) \) is the utility function with \( \Sigma \) and \( \mu_d \) estimated by \( \hat{\Sigma} \) and \( \hat{\mu}_d \). Then, it is easy to see the following utility approximation: for any \( \| \mathbf{w} \| \leq c \)

\[
|U(\mathbf{w}) - \hat{U}(\mathbf{w})| \leq \| \hat{\Sigma} - \Sigma \|_\infty c^2 + \xi \| \hat{\mu}_d - \mu_d \|_\infty
\]

and

\[
|U(\hat{\mathbf{w}}_c^*) - U(\mathbf{w}_c^*)| \leq 2\left( \| \hat{\Sigma} - \Sigma \|_\infty c^2 + \xi \| \hat{\mu}_d - \mu_d \|_\infty \right).
\]
Remark 3. The most prominent technical challenge of our original problem (1.10) is due to different dualities of penalization problems. For the population version (1.10), it can be reduced, by the Lagrange multiplier method, to the utility $U(w)$ optimization problem in Remark 2 with a given $\xi > 0$, while for the sample version (1.15), it can be reduced to the utility $\hat{U}(w)$ optimization problem with a different $\hat{\xi}$. Therefore, the problem is not the same as the utility optimization problem in Remark 2: $\hat{\xi}$ is hard to bound. In fact, it is much harder and yields more complicated results.

We now show how different the data projection direction in the regularized oracle can be from that in the Fisher discriminant. To gain better insight, we reformulate the $L_1$ constraint problem as the following penalized version:

$$w^\lambda = \arg\min \limits_{w, \mu_d^T w = 1} w^T \Sigma w + \lambda \|w\|_1.$$  \hspace{1cm} (1.17)

The following characterizes its convergence to the Fisher discriminant weight $w_\infty$ as $\lambda \to 0$.

Theorem 2. Let $s$ be the size of the set $\{k : (\Sigma^{-1} \mu_d)_k \neq 0\}$. Then, we have

$$\|w^\lambda - w_\infty\|_2 \leq \frac{\lambda \sqrt{s}}{\lambda_{\min}(\Sigma)},$$

where $w_\infty = \Sigma^{-1} \mu_d / (\mu_d^T \Sigma^{-1} \mu_d)$ is the normalized Fisher discriminant, optimizing (1.17) with $\lambda = 0$.

1.5.2 Screening-based ROAD (S-ROAD)

Following the idea of Sure Independence Screening in [32], we pre-screen all the features before hitting the ROAD. The advantage of this two-step procedure is that we have a control on the total number of features used in the final classification rule. A popular method for independent feature selection is the two-sample $t$-test [31, 74], which is a specific case of marginal screening in [32]. The sure screening property of such a method was demonstrated in [31], which selects consistently the features with different means in the same settings as ours.
Once the features are selected, we can hit the ROAD, producing the vanilla Screening-based Regularized Optimal Affine Discriminant (S-ROAD1):

1. Employ a screening method to get \( k \) features.

2. Apply ROAD to the \( k \) selected features.

In the first step, we use the \( t \)-statistics as the screening criteria and determine a data-driven threshold. This idea is motivated by a FDR criterion for choosing marginal screening threshold in [80]. A random permutation \( \pi \) of \( \{1, \cdots, n\} \) is used to decouple \( X_i \) and \( Y_i \) so that the resulting data \((X_{\pi(i)}, Y_i)\) follow a null model, by which we mean that features have no prediction power for the class label. More specifically, the screening step is carried out as follows:

(i) Calculate the \( t \)-statistic \( t_j \) for each feature \( j \), where \( j = 1, \cdots, p \).

(ii) For the permuted data pairs \((X_{\pi(i)}, Y_i)\), recalculate the \( t \)-statistic \( t_j^* \), for \( j = 1, \cdots, p \).

(Intuitively, if \( j \) is the index of an important feature, \(|t_j|\) should be larger than most of \(|t_j^*|\), because the random permutation is meant to eliminate the prediction power of features.)

(iii) For \( q \in [0, 1] \), let \( \omega(q) \) be the \( q^{th} \) quantile of \(|t_j^*|, j = 1, 2, \cdots, p\). Then, the selected set \( \mathcal{A} \) is defined as

\[
\mathcal{A} = \{ j | |t_j| \geq \omega(q) \}.
\]

The choice of threshold is made to retain the features whose \( t \)-statistics are significant in the two sample \( t \)-test. Alternatively, if the user knows his \( k \), (due to budget constraints, etc.), then he can just rank \(|t_j|\)'s and choose the threshold accordingly.

The S-ROAD1 tracks the performance of oracle procedures like sub-Fisher (10 features) in Figure 1. The feature space gotten by step (1) can be expanded by including those features which are most correlated with what have already been selected. This additional variant, S-ROAD2, aims at achieving the performance of sub-Fisher (20 features) type of procedure in Figure 1.
To elaborate on the theoretical properties of S-ROAD1, assume with no loss of generality that the first $k$ variables are selected in the screening step. Denote by $\Sigma_k$ the upper left $k \times k$ block of $\Sigma$ and $\mu_k$ the first $k$ coordinates of $\mu_d$. Let

$$\beta_c = \arg\min_{\|\beta\|_1 \leq c, \beta^T \mu_k = 1} \beta^T \Sigma_k \beta.$$  

The quantities $\hat{\beta}_c$ and $\beta_c^{(1)}$ are defined similarly to $\hat{w}_c$ and $w_c^{(1)}$ (defined right before Theorem 1). Then denote by $y_c = (\beta_c^T, 0^T)^T$, $\hat{y}_c = (\hat{\beta}_c^T, 0^T)^T$ and $y_c^{(1)} = (w_c^{(1)}, 0^T)^T$.

The next two theorems can be verified along lines similar to Theorems 1 and 2. Hence, the proofs are omitted.

**Theorem 3.** If $\|\Sigma_k - \Sigma_k\|_\infty = O_p(\sqrt{\log k}/n)$, $\|\mu_k - \mu_k\|_\infty = O_p(\sqrt{\log k}/n)$, and $\lambda_{\min}(\Sigma_k) \geq \delta_0 > 0$, then we have

$$W(\hat{\delta}_{y_c}) - W(\delta_{y_c}) = O_p(e_n),$$

and

$$W_n(\hat{\delta}_{y_c}) - W(\delta_{y_c}) = O_p(e_n),$$

where $e_n = (c^2 \lor k)\sqrt{\log k/n}$.

This result is cleaner than Theorem 1, as the rate does not involve $s_c$ and $\hat{s}_c$: they are simply replaced by the upper bound $k$. Accurate bounds for $s_c$ and $\hat{s}_c$ are of interest for future exploration, but they are beyond the scope of this paper.

**Theorem 4.** Let $y_k^\lambda = \arg\min_{y: \mu_k^T y = 1, y \in M_k} R(y) + \lambda \|y\|_1$ where $M_k$ is the subspace in $R^p$ with the last $p - k$ components being zero, and $y^0 = ((\Sigma_k^{-1} \mu_k)^T/(\mu_k^T \Sigma_k^{-1} \mu_k), 0^T)^T$. Then we have

$$\|y_k^\lambda - y^0\|_2 \leq \frac{\lambda \sqrt{k}}{\lambda_{\min}(\Sigma_k)}.$$  

### 1.5.3 Continuous Piecewise Linear Solution Path

We use the word “linear” when referring to “affine”, in line with the status quo in the statistical community. Continuous piecewise linear paths are of much interest to statisticians, as the property reduces the computational complexity of solutions and justifies the
linear interpolations of solutions at discrete points. Previous well known investigations include [30] and [64]. Our setup differs from others mainly in that in addition to a complexity penalty, there is also an affine constraint. Our proof calls in point set topology, and is purely geometrical, in a spirit very different from the existing ones. In particular, we stress that the continuity property is intuitively correct, but it is far from a trivial consequence of the assumptions. The authors also believe that the claim holds true even if the $p-1$ dimensional affine subspace constraint is replaced by more generic ones, though the technicality of the proof must be more involved.

**Theorem 5.** Let $\mu_d \in \mathbb{R}^p$ be a constant, and $\Sigma$ be a positive definite matrix of dimension $p \times p$. Let

$$w_c = \arg\min_{||w||_1 \leq c, w^T \mu_d = 1} w^T \Sigma w,$$

then $w_c$ is a continuous piecewise linear function in $c$.

**Proposition 1.** $W(\delta_{w_c})$ is a Lipschitz function in $c$.

**Proof.** Recall that

$$W(\delta_{w_c}) = 1 - \Phi\left(1/(R(w_c))^{1/2}\right).$$

By Theorem 5 and the fact that composition of Lipschitz functions is again Lipschitz, the conclusion holds.

### 1.6 Numerical Investigation

In this section, several simulation and real data studies are conducted. We compare ROAD and its variants S-ROAD1 (Screening-based ROAD version 1), S-ROAD2 (Screening-based ROAD version 2) and D-ROAD (Diagonal ROAD) with NSC (Nearest Shrunken Centroid), SCRDA (Shrunken Centroids Regularized Discriminant Analysis), FAIR (Feature Annealed Independence Rule), NB (Naive Bayes), NFR (Naive Fisher Rule, which uses the generalized inverse of the sample covariance matrix), as well as the Oracle.

In all simulation studies, the number of variables is $p = 1000$, and the sample size of the training and testing data is $n = 300$ for each class. Each simulation is repeated 100 times to test the stability of the method. Without loss of generality, the mean vector of
the first class $\mu_1$ is set to be 0. We use five-fold cross-validation to choose the penalty parameter $\lambda$.

### 1.6.1 Equal Correlation Setting, Sparse Fixed Signal

In this subsection, we consider the setting where $\Sigma_{i,i} = 1$ for all $i = 1, \cdots, p$ and $\Sigma_{i,j} = \rho$ for all $i, j = 1, \cdots, p$ and $i \neq j$, and take $\mu_2$ to be a sparse vector: $\mu_2 = (1^T_{10}, 0^T_{990})^T$, where $1_d$ is a length $d$ vector with all entries 1, $0_d$ is a length $d$ vector with all entries 0, where the sparsity size is $s_0 = 10$. Also, we fix $\gamma = 10$ in (1.14) for this simulation. Sensitivity of the performance due to the choice of $\gamma$ will be investigated in the next subsection.

The solution paths for ROAD and D-ROAD of one realization are rendered in Figure 1.2. It is clear from the figure that, as the penalty parameter decreases (index increases), both ROAD and D-ROAD use more features. Also, the cutoff point for D-ROAD, where the number of features starts to increase dramatically, tends to come later than that for ROAD.

The simulation results for the pairwise correlations ranging from 0 to 0.9 are shown in Tables 1.1 and 1.2. We would like to mention that the results for NFR (Naive Fisher
Table 1.1: Equal correlation setting, fixed signal: Median of the percentage for testing classification error and standard deviations (in parentheses). Signal all equal to 1. $s_0 = 10$.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>ROAD</th>
<th>S-ROAD1</th>
<th>S-ROAD2</th>
<th>D-ROAD</th>
<th>SCRDA</th>
<th>NSC</th>
<th>FAIR</th>
<th>NB</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6.0(1.2)</td>
<td>6.0(1.1)</td>
<td>6.0(1.2)</td>
<td>5.7(1.1)</td>
<td>5.9(1.0)</td>
<td>5.7(1.0)</td>
<td>11.2(1.4)</td>
<td>5.5(1.1)</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>6.3(2.5)</td>
<td>12.2(5.0)</td>
<td>8.8(2.4)</td>
<td>11.6(5.1)</td>
<td>10.3(1.4)</td>
<td>11.1(3.0)</td>
<td>12.4(1.4)</td>
<td>26.8(10.1)</td>
<td>5.0(0.9)</td>
</tr>
<tr>
<td>0.2</td>
<td>5.3(1.0)</td>
<td>16.0(6.3)</td>
<td>8.7(2.5)</td>
<td>16.1(7.5)</td>
<td>8.5(1.2)</td>
<td>14.5(4.3)</td>
<td>17.3(1.7)</td>
<td>34.8(11.6)</td>
<td>4.0(0.8)</td>
</tr>
<tr>
<td>0.3</td>
<td>4.2(0.9)</td>
<td>19.1(7.9)</td>
<td>7.8(2.6)</td>
<td>19.1(9.4)</td>
<td>6.6(1.1)</td>
<td>17.1(5.5)</td>
<td>20.8(1.7)</td>
<td>39.3(12.3)</td>
<td>3.2(0.7)</td>
</tr>
<tr>
<td>0.4</td>
<td>3.2(0.8)</td>
<td>22.8(9.4)</td>
<td>6.5(2.6)</td>
<td>22.2(9.9)</td>
<td>4.8(1.0)</td>
<td>20.5(6.1)</td>
<td>23.2(1.8)</td>
<td>41.6(11.3)</td>
<td>2.0(0.6)</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0(0.6)</td>
<td>25.8(11.0)</td>
<td>4.8(1.4)</td>
<td>25.2(10.2)</td>
<td>2.9(0.7)</td>
<td>23.2(6.0)</td>
<td>25.3(1.6)</td>
<td>43.5(11.1)</td>
<td>1.3(0.5)</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0(0.4)</td>
<td>18.3(12.4)</td>
<td>3.3(1.3)</td>
<td>28.1(10.3)</td>
<td>1.5(0.5)</td>
<td>25.8(5.7)</td>
<td>26.8(1.8)</td>
<td>44.4(12.1)</td>
<td>0.7(0.3)</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3(0.2)</td>
<td>15.5(13.6)</td>
<td>1.7(1.0)</td>
<td>29.1(10.1)</td>
<td>0.5(0.3)</td>
<td>27.0(8.2)</td>
<td>28.2(2.0)</td>
<td>45.2(12.3)</td>
<td>0.2(0.2)</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0(0.1)</td>
<td>5.0(14.0)</td>
<td>0.3(0.4)</td>
<td>29.5(9.9)</td>
<td>0.0(0.1)</td>
<td>28.3(8.7)</td>
<td>29.2(2.0)</td>
<td>46.2(10.3)</td>
<td>0.0(0.1)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0(0.0)</td>
<td>0.6(14.8)</td>
<td>0.0(0.1)</td>
<td>30.3(7.6)</td>
<td>0.0(0.2)</td>
<td>29.9(8.0)</td>
<td>30.2(1.9)</td>
<td>46.8(8.8)</td>
<td>0.0(0.0)</td>
</tr>
</tbody>
</table>

Rule) are not included in these (and the subsequent) tables because the test classification error is always around 50%, i.e., it is about the same as random guess. Also in the tables are the screening-based versions of the ROAD. S-ROAD1 refers to the vanilla version where we first apply the two-sample $t$-test to select any features with the corresponding $t$-test statistic with absolute value larger than the maximum absolute $t$-test statistic value calculated on the permuted data. S-ROAD2 does the same except for each variable in S-ROAD1’s pre-screened set, it adds an additional variable which is most correlated with that variable. Figure 1.3, a graphical summary of Table 1, presents the median test errors for different methods. We can see from Table 1.1 and Figure 1.3 that the oracle classification error decreases as $\rho$ increases. This phenomenon is due to a similar reason to the two-dimensional showcase in the introduction. When $\rho$ goes to 1, all the variables contribute in the same way to boost the classification power. ROAD performs reasonably close to the Oracle, while working independence based method such as D-ROAD, NSC, FAIR and NB fail when $\rho$ is large. The huge discrepancy shows the advantage of employing the correlation structure. Since SCRDA also employ the correlation structure, it does not fail when $\rho$ is large. However, ROAD still outperforms SCRDA in all the correlation settings. S-ROAD1 and S-ROAD2 both have misclassification rates similar to that of ROAD. It is worth to emphasize that the merits of the screening based ROADs mainly lie in the computation cost, which is reduced significantly by the pre-screening step.
Table 1.2: Equal correlation setting, fixed signal: Median of number of nonzero coefficients and standard deviations (in parentheses). Signal all equal to 1. $s_0 = 10$. 

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>ROAD</th>
<th>S-ROAD1</th>
<th>S-ROAD2</th>
<th>D-ROAD</th>
<th>SCRDA</th>
<th>NSC</th>
<th>FAIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16.00(24.16)</td>
<td>10.00(1.31)</td>
<td>17.00(4.31)</td>
<td>29.50(58.54)</td>
<td>10.00(13.25)</td>
<td>10.00(44.86)</td>
<td>11.00(1.62)</td>
</tr>
<tr>
<td>0.1</td>
<td>117.50(30.50)</td>
<td>11.00(3.32)</td>
<td>21.00(4.15)</td>
<td>14.00(122.02)</td>
<td>1000.00(345.48)</td>
<td>35.50(117.32)</td>
<td>10.00(0.27)</td>
</tr>
<tr>
<td>0.2</td>
<td>130.50(33.33)</td>
<td>11.00(6.99)</td>
<td>22.00(8.98)</td>
<td>15.50(111.42)</td>
<td>1000.00(0.00)</td>
<td>95.00(120.17)</td>
<td>10.00(0.69)</td>
</tr>
<tr>
<td>0.3</td>
<td>136.50(36.16)</td>
<td>11.00(11.56)</td>
<td>22.00(10.38)</td>
<td>17.50(106.16)</td>
<td>1000.00(0.00)</td>
<td>103.50(117.52)</td>
<td>9.00(1.19)</td>
</tr>
<tr>
<td>0.4</td>
<td>135.00(34.43)</td>
<td>10.00(14.21)</td>
<td>22.00(17.07)</td>
<td>10.00(98.10)</td>
<td>1000.00(0.00)</td>
<td>70.00(131.65)</td>
<td>8.00(1.33)</td>
</tr>
<tr>
<td>0.5</td>
<td>138.50(38.17)</td>
<td>9.00(21.71)</td>
<td>22.00(21.56)</td>
<td>10.00(105.33)</td>
<td>1000.00(0.00)</td>
<td>65.00(137.97)</td>
<td>7.00(1.30)</td>
</tr>
<tr>
<td>0.6</td>
<td>148.00(49.74)</td>
<td>10.50(27.92)</td>
<td>22.00(31.88)</td>
<td>10.00(110.23)</td>
<td>1000.00(0.00)</td>
<td>38.00(141.91)</td>
<td>6.00(1.30)</td>
</tr>
<tr>
<td>0.7</td>
<td>170.50(52.29)</td>
<td>11.00(37.37)</td>
<td>22.00(41.76)</td>
<td>1.00(118.43)</td>
<td>1000.00(0.00)</td>
<td>27.50(140.10)</td>
<td>5.00(1.20)</td>
</tr>
<tr>
<td>0.8</td>
<td>203.00(27.72)</td>
<td>12.00(50.36)</td>
<td>24.00(59.23)</td>
<td>1.00(143.83)</td>
<td>1000.00(10.92)</td>
<td>15.00(157.98)</td>
<td>5.00(1.29)</td>
</tr>
<tr>
<td>0.9</td>
<td>151.50(8.02)</td>
<td>14.00(55.32)</td>
<td>28.00(50.45)</td>
<td>1.00(153.27)</td>
<td>1000.00(56.30)</td>
<td>14.00(225.38)</td>
<td>3.00(1.08)</td>
</tr>
</tbody>
</table>

The ROAD is a very robust estimator. It performs well even when all the variables are independent, in which case there could be a lot of noise for fitting the covariance matrix. Table 1.1 indicates that ROAD has almost the same performance as D-ROAD, NSC and FAIR under the independence assumption, i.e. $\rho = 0$. As $\rho$ increases, the edge of ROAD becomes more substantial. In general, the ROAD is recommended on the grounds that even with pairwise correlation of about 0.1 (which is quite common in microarray data as well as financial data), the gain is substantial.

Another interesting observation is that the D-ROAD performs similarly to NSC and FAIR in terms of classification error. An intuitive explanation is that they are all “sparse” independence rules. NSC uses soft-thresholding on the standardized sample mean difference, and its equivalent LASSO derivation can be found in [78]. FAIR selects features with large marginal $t$-statistics in absolute values, while D-ROAD is another L1 penalized independence rule, whose implementation is different from NSC.

Table 2 summarizes the number of features selected by different classifiers. Note that ROAD mimics Fisher discriminant coordinate $\Sigma^{-1}\mu_d$, which has $p = 1000$ nonzero entries under our simulated model. Therefore, the large number of features selected by ROAD is not out of expectation.
1.6.2 The Effect of $\gamma$

Under the settings of the previous subsection, we look into the variation of the ROAD performance as $\gamma$ changes. In Table 1.3, the number of active variables varies; however, the median classification error remains about the same for a broad range of $\gamma$ values. The reason is that the cross validation step chooses the “best” $\lambda$ according to a specific $\gamma$. Therefore, the final performance remains almost unchanged. Since our primary concern is the classification error, we fix $\gamma = 10$ for simplicity in the subsequent simulations and in the real data analysis.

1.6.3 Block Diagonal Correlation Setting, Sparse Fixed Signal

In this subsection, we follow the same setup as in Section 1.6.1 except that the covariance matrix $\Sigma$ is taken to be block diagonal. The first block is a $20 \times 20$ equi-correlated matrix and the second block is a $(p - 20) \times (p - 20)$ equi-correlated matrix, both with pairwise
Table 1.3: Equal correlation setting; signals all equal to 1; \( s_0 = 10 \). Results for different \( \gamma \).

<table>
<thead>
<tr>
<th>Median classification error (in percentage)</th>
<th>( \rho = 0 )</th>
<th>( \rho = 0.5 )</th>
<th>( \rho = 0.9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROAD(_{\gamma=0.01})</td>
<td>5.8(1.2)</td>
<td>2.7(0.6)</td>
<td>0.2(0.2)</td>
</tr>
<tr>
<td>ROAD(_{\gamma=0.1})</td>
<td>6.0(1.2)</td>
<td>2.0(0.6)</td>
<td>0.2(0.1)</td>
</tr>
<tr>
<td>ROAD(_{\gamma=1})</td>
<td>6.0(1.3)</td>
<td>2.0(0.6)</td>
<td>0.0(0.1)</td>
</tr>
<tr>
<td>ROAD(_{\gamma=10})</td>
<td>6.0(1.2)</td>
<td>2.0(0.6)</td>
<td>0.0(0.0)</td>
</tr>
<tr>
<td>ROAD(_{\gamma=100})</td>
<td>6.2(1.2)</td>
<td>2.3(0.6)</td>
<td>0.0(0.1)</td>
</tr>
</tbody>
</table>

Table 1.4: Block diagonal correlation setting, sparse fixed signal: Median of the percentage for testing classification error and standard deviations (in parentheses). Signal all equal to 1. \( s_0 = 10 \).

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>ROAD</th>
<th>S-ROAD1</th>
<th>S-ROAD2</th>
<th>D-ROAD</th>
<th>SCRDA</th>
<th>NSC</th>
<th>FAIR</th>
<th>NB</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6.0(1.2)</td>
<td>6.0(1.2)</td>
<td>6.0(1.2)</td>
<td>5.7(1.1)</td>
<td>6.0(0.1)</td>
<td>5.5(0.3)</td>
<td>5.7(1.0)</td>
<td>11.2(1.4)</td>
<td>5.5(1.1)</td>
</tr>
<tr>
<td>0.1</td>
<td>10.8(3.6)</td>
<td>13.0(4.8)</td>
<td>10.3(3.0)</td>
<td>12.8(4.4)</td>
<td>13.0(0.3)</td>
<td>12.5(0.8)</td>
<td>12.7(1.5)</td>
<td>25.7(7.6)</td>
<td>8.8(1.2)</td>
</tr>
<tr>
<td>0.2</td>
<td>10.7(4.1)</td>
<td>18.0(5.7)</td>
<td>9.7(3.6)</td>
<td>17.7(5.9)</td>
<td>14.2(1.1)</td>
<td>17.2(0.4)</td>
<td>17.7(1.6)</td>
<td>34.4(7.9)</td>
<td>8.8(1.2)</td>
</tr>
<tr>
<td>0.3</td>
<td>9.5(3.8)</td>
<td>23.2(5.5)</td>
<td>8.8(4.0)</td>
<td>23.2(5.6)</td>
<td>12.7(0.9)</td>
<td>20.0(0.8)</td>
<td>20.4(1.6)</td>
<td>38.3(7.5)</td>
<td>7.7(1.0)</td>
</tr>
<tr>
<td>0.4</td>
<td>8.0(3.3)</td>
<td>29.7(4.2)</td>
<td>7.5(4.2)</td>
<td>29.3(4.1)</td>
<td>11.0(1.2)</td>
<td>23.8(1.3)</td>
<td>23.2(1.8)</td>
<td>41.0(6.9)</td>
<td>6.6(1.1)</td>
</tr>
<tr>
<td>0.5</td>
<td>6.2(2.6)</td>
<td>30.1(3.9)</td>
<td>5.7(0.9)</td>
<td>30.0(3.1)</td>
<td>8.7(0.4)</td>
<td>26.2(1.7)</td>
<td>25.1(1.7)</td>
<td>42.2(6.6)</td>
<td>5.0(1.0)</td>
</tr>
<tr>
<td>0.6</td>
<td>4.2(0.9)</td>
<td>30.3(4.2)</td>
<td>4.0(0.8)</td>
<td>30.3(2.2)</td>
<td>6.4(0.1)</td>
<td>26.5(1.2)</td>
<td>26.8(1.8)</td>
<td>43.6(7.0)</td>
<td>3.5(0.7)</td>
</tr>
<tr>
<td>0.7</td>
<td>2.3(0.7)</td>
<td>30.0(6.4)</td>
<td>2.2(0.7)</td>
<td>30.6(2.1)</td>
<td>2.5(0.7)</td>
<td>28.1(3.2)</td>
<td>28.2(2.0)</td>
<td>44.2(6.5)</td>
<td>1.8(0.6)</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8(0.4)</td>
<td>29.8(9.8)</td>
<td>0.7(0.4)</td>
<td>30.6(2.1)</td>
<td>0.6(0.4)</td>
<td>29.2(1.6)</td>
<td>29.2(2.0)</td>
<td>44.8(5.7)</td>
<td>0.7(0.3)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.0(0.1)</td>
<td>29.8(12.8)</td>
<td>0.0(0.1)</td>
<td>30.6(1.9)</td>
<td>0.2(0.2)</td>
<td>29.2(1.2)</td>
<td>30.2(1.9)</td>
<td>45.2(4.9)</td>
<td>0.0(0.1)</td>
</tr>
</tbody>
</table>

correlation \( \rho \). In other words, \( \Sigma_{i,i} = 1 \) for all \( i = 1, \ldots, p \), \( \Sigma_{i,j} = \rho \) for all \( i, j = 1, \ldots, 20 \) and \( i \neq j \), \( \Sigma_{i,j} = \rho \) for all \( i, j = 21, \ldots, p \) and \( i \neq j \), and the rest elements are zeros. As before, we examine the performances of various estimators when \( \rho \) varies. The percentage for testing error and the number of selected features in the estimators are shown in Tables 1.4 and 1.5, respectively.

In this block-diagonal setting, we have observed similar results to those in Section 1.6.1: ROAD and S-ROAD2 perform significantly better than the other methods. One interesting phenomenon is that S-ROAD1 does not perform well when \( \rho \) is large. The reason is that the current true model has 20 important features, and by looking only at marginal contribution, S-ROAD1 misses some important variables, as shown in Table 1.4.
1.6.4 Block-Diagonal Negative Correlation Setting, Sparse Fixed Signal

In this subsection, we again follow a similar setup as in Section 1.6.1. Here, the covariance matrix $\Sigma$ is taken to be block diagonal with each block size equals to 10. Each block is an equi-correlated matrix with pairwise correlation $\rho = -0.1$. In other words, $\Sigma = \text{diag}(\Sigma_0, \cdots, \Sigma_0)$, where $\Sigma_0$ is a $10 \times 10$ equi-correlated matrix with correlation $-0.1$. Here, $\mu_2 = 0.5 \times (1_5^T, 0_5^T, 1_5^T, 0_{985}^T)^T$ and the sparsity size is $s_0 = 10$. As before, we examine the performances of various estimators when $\rho$ varies. The percentage for testing error and the number of selected features in the estimators are shown in Table 1.6.
Table 1.7: Random correlation setting, double exponential signal: Median error (in percentage) and number of nonzero coefficients with standard deviations in parentheses.

<table>
<thead>
<tr>
<th>Error</th>
<th>ROAD</th>
<th>S-ROAD1</th>
<th>S-ROAD2</th>
<th>D-ROAD</th>
<th>SCRDA</th>
<th>NSC</th>
<th>FAIR</th>
<th>NB</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>error</td>
<td>2.0(0.6)</td>
<td>11.0(5.2)</td>
<td>5.8(3.9)</td>
<td>17.0(2.2)</td>
<td>5.2(1.1)</td>
<td>16.2(1.3)</td>
<td>17.0(1.6)</td>
<td>46.2(2.4)</td>
<td>1.3(0.5)</td>
</tr>
<tr>
<td>nonzero</td>
<td>83.00(39.54)</td>
<td>4.00(8.13)</td>
<td>9.00(10.69)</td>
<td>1.00(3.89)</td>
<td>1000.00(0.00)</td>
<td>4.00(0.58)</td>
<td>1.00(0.17)</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

1.6.5 Random Correlation Setting, Double Exponential Signal

To evaluate the stability of the ROAD, we take a random matrix $\Sigma$ as the correlation structure, and use a signal $\mu$ whose nonzero entries come from a double exponential distribution. A random covariance matrix $\Sigma$ is generated as follows:

(i) For a given integer $m$ (here we take $m = 10$), generate a $p \times m$ matrix $\Omega$ where $\Omega_{i,j} \sim \text{Unif}(-1, 1)$. Then the matrix $\Omega\Omega^T$ is positive semi-definite.

(ii) Denote $c_{\Omega} = \min_i (\Omega\Omega^T)_{ii}$. Let $\Xi = \Omega\Omega^T + c_{\Omega}I$, where $I$ is the identity matrix. It is clear that $\Xi$ is positive definite.

(iii) Normalize the matrix $\Xi$ to get $\Sigma$ whose diagonal elements are unity.

For the signal, we take $\mu$ to be a sparse vector with sparsity size $s = 10$, and the nonzero elements are generated from the double exponential distribution with density function

$$f(x) = \exp(-2|x|).$$

Table 1.7 summaries the results. It shows that even under random correlation setting and random signals, our procedure ROAD still outperforms other competing classification rules such as SCRDA, NSC and FAIR in terms of the classification error.

1.6.6 Real Data

Though the ROAD seems to perform best in a broad spectrum of idealized experiments, it has to be tested against reality. We now evaluate the performance of our newly proposed estimator on three popular gene expression data sets: “Leukemia” [41], “Lung Cancer” [42], and “Neuroblastoma data set” [61]. The first two data sets come with predetermined,
separate training and test sets of data vectors. The Leukemia data set contains \( p = 7,129 \) genes for \( n_1 = 27 \) acute lymphoblastic leukemia (ALL) and \( n_2 = 11 \) acute myeloid leukemia (AML) vectors in the training set. The test set includes 20 ALL and 14 AML vectors. The Lung Cancer data set contains \( p = 12,533 \) genes for \( n_1 = 16 \) adenocarcinoma (ADCA) and \( n_2 = 16 \) mesothelioma training vectors, along with 134 ADCA and 15 mesothelioma test vectors. The Neuroblastoma data set, obtained via the MicroArray Quality Control phase-II (MAQC-II) project, consists of gene expression profiles for \( p = 10,707 \) genes from 251 patients of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004. We analyzed the gene expression data with the 3-year event-free survival (3-year EFS), which indicates whether a patient survived 3 years after the diagnosis of neuroblastoma. There are 239 subjects with the 3-year EFS information available (49 positives and 190 negatives). We randomly select 83 subjects (19 positives and 64 negatives, which are about one third of the total subjects) as the training set and the rest as the test set. The readers can find more details about the data sets in the original papers.

Following [29] and [31], we standardized each sample to zero mean and unit variance. The classification results for ROAD, S-ROAD1, S-ROAD2, SCRDA, FAIR, NSC and NB are shown in Tables 1.8, 1.9 and 1.10. For the leukemia and lung cancer data, ROAD performs the best in terms of classification error. For the neuroblastoma data, NB performs best, however, it makes use of all 10,707 genes, which is not very desirable. In contrast, ROAD has a competitive performance in terms of classification error and it only selects 33 genes. Although SCRDA has a close performance, the number of selected variables varies a lot for the three data set (264, 2410, 1). Overall, ROAD is a robust classification tool for high-dimensional data.

### 1.7 Discussion

With a simple two-class gaussian model, we explored the bright side of using correlation structure for high dimensional classification. Targeting directly on the classification er-
Table 1.8: Classification error and number of selected genes by various methods of leukemia data. Training and testing samples are of sizes 38 and 34, respectively.

<table>
<thead>
<tr>
<th></th>
<th>ROAD</th>
<th>S-ROAD1</th>
<th>S-ROAD2</th>
<th>SCRDA</th>
<th>FAIR</th>
<th>NSC</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Error</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Testing Error</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>No. of selected genes</td>
<td>40</td>
<td>49</td>
<td>66</td>
<td>264</td>
<td>11</td>
<td>24</td>
<td>7129</td>
</tr>
</tbody>
</table>

Table 1.9: Classification error and number of selected genes by various methods of lung cancer data. Training and testing samples are of sizes 32 and 149, respectively.

<table>
<thead>
<tr>
<th></th>
<th>ROAD</th>
<th>S-ROAD1</th>
<th>S-ROAD2</th>
<th>SCRDA</th>
<th>FAIR</th>
<th>NSC</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Error</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Testing Error</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>No. of selected genes</td>
<td>52</td>
<td>56</td>
<td>54</td>
<td>2410</td>
<td>31</td>
<td>38</td>
<td>12533</td>
</tr>
</tbody>
</table>

ror, ROAD employs un-regularized pooled sample covariance matrix and sample mean difference vector without suffering from curse of dimensionality and noise accumulation.

The sparsity of chosen features is evident in simulations and real data analysis; however, we have not discovered intuitively good conditions on $\Sigma$ and $\mu_d$, such that a certain desirable sparsity pattern of $\hat{w}_c$ follows. We resolve a part of the problem by introducing screening-based variants of ROAD, but the precise control of the sparsity size is worth for further investigation. Furthermore, we can explore the conditions for the model selection consistency.


In this paper, we have restricted ourselves to the linear rules. They can be easily extended to nonlinear discriminants via transformations such as low order polynomials or spline basis functions. One may also use the popular “kernel tricks” in the machine learning community. See, for example, [47] for more details. After the features are

Table 1.10: Classification error and number of selected genes by various methods of neuroblastoma data. Training and testing samples are of sizes 83 and 163, respectively.

<table>
<thead>
<tr>
<th></th>
<th>ROAD</th>
<th>S-ROAD1</th>
<th>S-ROAD2</th>
<th>SCRDA</th>
<th>FAIR</th>
<th>NSC</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Error</td>
<td>3</td>
<td>22</td>
<td>14</td>
<td>16</td>
<td>15</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Testing Error</td>
<td>33</td>
<td>47</td>
<td>37</td>
<td>37</td>
<td>44</td>
<td>35</td>
<td>32</td>
</tr>
<tr>
<td>No. of selected genes</td>
<td>33</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>18</td>
<td>41</td>
<td>10707</td>
</tr>
</tbody>
</table>
transformed, we can hit the ROAD. One essential technical challenge of the current paper is rooted in a stochastic linear constraint. The precise role of this constraint has not been completely pinned down. In the following, a preliminary proposal is provided for extending ROAD to multi-class settings.

1.7.1 Extension to Multi-Class

In this section, we outline an extension of ROAD to multi-class classification problems. Suppose that there are $K$ classes, and for $j = 1, \cdots, K$, the $j$th class has mean $\mu_j$ and common covariance $\Sigma$. Denote the overall mean of features by $\mu_a = K^{-1} \sum_{j=1}^{K} \mu_j$.

Fisher’s reduced rank approach to multi-class classification is a minimum distance classifier in some lower dimensional projection space. The first step is to find $s \leq K - 1$ discriminant coordinates $(w_1^*, \cdots, w_s^*)$ that separate the population centroids $\{\mu_j\}_{j=1}^{K}$ the most in the projected space $S = \text{span}\{w_1^*, \cdots, w_s^*\}$. Then the population centroids $\mu_j$’s and new observation $X$ are both projected onto $S$. The observation $X$ will be assigned to the class whose projected centroid is closest to the projection of $X$ onto $S$.

Note that it is usually not necessary to compute all $K - 1$ discriminant coordinates whose span is that of all $K$ population centroids; the process can stop as long as the projected population centroids are well spread out in $S$.

We adopt the above procedure for multi-class classification. However, the large-$p$-small-$n$ scenario demands regularization in selecting discriminant coordinates. Indeed, in the Fisher’s proposal the first discriminant coordinate $w_1^*$ is the solution of

$$\max_w \frac{w^T B w}{w^T \Sigma w},$$

where $B = \Psi^T \Psi$, and the $j$th column of $\Psi^T$ is $(\mu_j - \mu_a)$. Note that a multiple of $B$ is the between-class variance matrix. The second discriminant coordinate $w_2^*$ is the maximizer of $w^T B w / (w^T \Sigma w)$ with constraint $w_1^T \Sigma w = 0$, and the subsequent discriminant coordinates are determined analogously.
Since solving (1.18) is the same as looking for the eigenvector of $\Sigma^{-1/2}B\Sigma^{-1/2}$ corresponding to the largest eigenvalue, diverging spectrum and noise accumulation have to be considered when we work on the sample. To address these issues, we regularize $w$ as in the binary case,

$$\min_{\|w\|_1 \leq c, w^TBw=1} w^T\Sigma w,$$  \hspace{1cm} (1.19)

whose solution is the first regularized discriminant coordinate $\bar{w}_1^*$. Here, equation (1.19) is related to the null space method in [51]. The second regularized discriminant coordinate is obtained by solving (1.19) with additional constraint $\bar{w}_1^*^T\Sigma w = 0$. Other regularized discriminant coordinates can be found similarly. With these $s \leq K-1$ regularized discriminant coordinates, the classifier is now based on the minimum distance to the projected centroids in the $s$-dimensional space spanned by $\{\bar{w}_j^*\}_{j=1}^s$.

The implementation and theoretical properties for multi-class ROAD are interesting topics for future research.

1.8 Proofs

1.8.1 Proof of Theorem 1

We now show first part of the theorem. Let $f_0(w) = w^T\mu_d/(w^T\Sigma w)^{1/2}$, $f_1(w) = w^T\hat{\mu}_d/(w^T\Sigma w)^{1/2}$, and $f_2(w) = w^T\hat{\mu}_d/(w^T\hat{\Sigma} w)^{1/2}$. Then, it follows easily that

$$|f_0(w_c) - f_2(\hat{w}_c)| \leq \Lambda_1 + \Lambda_2,$$

where $\Lambda_1 = |f_0(w_c) - f_1(w_c^{(1)})|$ and $\Lambda_2 = |f_1(w_c^{(1)}) - f_2(\hat{w}_c)|$. We now bound both terms separately in the following two steps.
Step 1 (bound $\Lambda_1$): For any $\mathbf{w}$, we have

$$|f_0(\mathbf{w}) - f_1(\mathbf{w})| \leq \left| \frac{\mathbf{w}^T \mu_d}{(\mathbf{w}^T \Sigma \mathbf{w})^{1/2}} - \frac{\mathbf{w}^T \hat{\mu}_d}{(\mathbf{w}^T \Sigma \mathbf{w})^{1/2}} \right| \leq \frac{\|\mathbf{w}\|_1 \|\mu_d - \hat{\mu}_d\|_\infty}{\|\mathbf{w}\|_2 \lambda_{\min}^{1/2}(\Sigma)} \leq \sqrt{\|\mathbf{w}\|_0 \|\mu_d - \hat{\mu}_d\|_\infty} \sigma_0 = \sqrt{\|\mathbf{w}\|_0 O_p(a_n)}. \quad (1.20)$$

Since $\mathbf{w}_c^{(1)}$ maximizes $f_1(\cdot)$, it follows that

$$f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c^{(1)}) = f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c) + [f_1(\mathbf{w}_c) - f_1(\mathbf{w}_c^{(1)})] \leq f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c), \quad (1.21)$$

and similarly noticing $\mathbf{w}_c$ maximizing $f_0(\cdot)$, we have

$$f_1(\mathbf{w}_c^{(1)}) - f_0(\mathbf{w}_c) = f_1(\mathbf{w}_c^{(1)}) - f_0(\mathbf{w}_c^{(1)}) + [f_0(\mathbf{w}_c^{(1)}) - f_0(\mathbf{w}_c)] \leq f_1(\mathbf{w}_c^{(1)}) - f_0(\mathbf{w}_c^{(1)}). \quad (1.22)$$

Combining the results of (1.21) and (1.22) and using (1.20), we conclude that

$$\Lambda_1 = |f_0(\mathbf{w}_c) - f_1(\mathbf{w}_c^{(1)})| = O_p \left( (s_c \lor s_c^{(1)}) a_n \right).$$

By the Lipschitz property of $\Phi$,

$$|\Phi(f_1(\mathbf{w}_c^{(1)})) - \Phi(f_0(\mathbf{w}_c))| = O_p \left( (s_c \lor s_c^{(1)}) a_n \right).$$

Step 2 (bound $\Lambda_2$): Note that $\mathbf{w}_c^{(1)}$ and $\hat{\mathbf{w}}_c$ both are in the set \{ $\mathbf{w} : \mathbf{w}^T \mu_d = 1, \|\mathbf{w}\|_1 \leq 1$ \}. Therefore, by definition of minimizers, we have

$$\mathbf{w}_c^{(1)^T} \Sigma \mathbf{w}_c^{(1)} - \hat{\mathbf{w}}_c^T \Sigma \hat{\mathbf{w}}_c \leq 0, \text{ and } \hat{\mathbf{w}}_c^T \Sigma \hat{\mathbf{w}}_c - \mathbf{w}_c^{(1)^T} \Sigma \mathbf{w}_c^{(1)} \leq 0.$$
Consequently,

\[
\begin{align*}
\mathbf{w}_c^{(1)T} \Sigma \mathbf{w}_c^{(1)} - \hat{\mathbf{w}}_c^T \hat{\Sigma} \hat{\mathbf{w}}_c &= \left[ \mathbf{w}_c^{(1)T} \Sigma \mathbf{w}_c^{(1)} - \hat{\mathbf{w}}_c^T \hat{\Sigma} \hat{\mathbf{w}}_c \right] + \hat{\mathbf{w}}_c^T \hat{\Sigma} \hat{\mathbf{w}}_c - \hat{\mathbf{w}}_c^T \hat{\Sigma} \hat{\mathbf{w}}_c \\
&\leq \hat{\mathbf{w}}_c^T (\Sigma - \hat{\Sigma}) \hat{\mathbf{w}}_c \\
&\leq \| \Sigma - \hat{\Sigma} \|_\infty \| \hat{\mathbf{w}}_c \|_1^2 \\
&\leq c^2 \| \Sigma - \hat{\Sigma} \|_\infty \\
&= O_p(a_n c^2).
\end{align*}
\] (1.23)

By the same argument, we also have

\[
\begin{align*}
\hat{\mathbf{w}}_c^T \hat{\Sigma} \hat{\mathbf{w}}_c - \mathbf{w}_c^{(1)T} \Sigma \mathbf{w}_c^{(1)} &= \left[ \hat{\mathbf{w}}_c^T \hat{\Sigma} \hat{\mathbf{w}}_c - \mathbf{w}_c^{(1)T} \hat{\Sigma} \mathbf{w}_c^{(1)} \right] + \mathbf{w}_c^{(1)T} \hat{\Sigma} \mathbf{w}_c^{(1)} - \mathbf{w}_c^{(1)T} \Sigma \mathbf{w}_c^{(1)} \\
&\leq \mathbf{w}_c^{(1)T} (\hat{\Sigma} - \Sigma) \mathbf{w}_c^{(1)} \\
&\leq c^2 \| \Sigma - \hat{\Sigma} \|_\infty \\
&= O_p(a_n c^2).
\end{align*}
\] (1.24)

Combination of (1.23) and (1.24) leads to

\[
| \hat{\mathbf{w}}_c^T \hat{\Sigma} \hat{\mathbf{w}}_c - \mathbf{w}_c^{(1)T} \Sigma \mathbf{w}_c^{(1)} | = O_p(a_n c^2).
\]

Let \( g(x) = \Phi(x^{-1/2}) \). The function \( g \) is Lipschitz on \((0, \infty)\), as \( g'(x) \) is bounded on \((0, \infty)\). Hence, \( | \Phi(f_2(\hat{\mathbf{w}}_c)) - \Phi(f_0(\mathbf{w}_c^{(1)})) | = O_p(a_n c^2) \). Thus,

\[
| W_n(\hat{\delta}_{\mathbf{w}_c}, \theta) - W(\delta_{\mathbf{w}_c}, \theta) | \leq | \Phi(f_2(\hat{\mathbf{w}}_c)) - \Phi(f_0(\mathbf{w}_c^{(1)})) | + | \Phi(f_1(\hat{\mathbf{w}}_c^{(1)})) - \Phi(f_0(\mathbf{w}_c)) | \\
= O_p \left( (s_c \lor s_c^{(1)}) a_n \right) + O_p(a_n c^2) \\
= O_p(b_n).
\]

We now prove the second result of the Theorem. Since \( | \hat{\mathbf{w}}_c^T \Sigma \hat{\mathbf{w}}_c - \hat{\mathbf{w}}_c^T \hat{\Sigma} \hat{\mathbf{w}}_c | = O_p(a_n c^2) \), we have

\[
| \Phi(f_1(\hat{\mathbf{w}}_c)) - \Phi(f_2(\hat{\mathbf{w}}_c)) | = O_p(a_n c^2). \] (1.25)
By (1.20), (1.25), and the first part of the Theorem, we have

\[
|W(\hat{\delta}_{w_c}, \theta) - W(\delta_{w_c}, \theta)| \\
= |\Phi(f_0(\hat{w}_c)) - \Phi(f_0(w_c))| \\
\leq |\Phi(f_0(\hat{w}_c)) - \Phi(f_1(\hat{w}_c))| + |\Phi(f_1(\hat{w}_c)) - \Phi(f_2(\hat{w}_c))| + |\Phi(f_2(\hat{w}_c)) - \Phi(f_0(w_c))| \\
= O_p(\hat{s}_c a_n) + O_p(a_n c^2) + O_p(b_n) \\
= O_p(d_n).
\]

This completes the proof of Theorem.

1.8.2 Proof of Theorem 2

Let \( w^\lambda = w_\infty + \gamma^\lambda \). Then, from the definition of \( w^\lambda \), we have

\[
\gamma^\lambda = \arg\min_{\mu_d^T w_\infty + \mu_d^T \gamma = 1} R(w_\infty + \gamma) + \lambda \|w_\infty + \gamma\|_1 \\
= \arg\min_{\mu_d^T \gamma = 0} f(\gamma),
\]

where \( f(\gamma) = R(\gamma) + \lambda \sum_{k \in K} |\gamma_k| + \lambda \sum_{k \in K} (|w_k^\infty + \gamma_k| - |w_k^\infty|) \). In the last statement, we used the fact that

\[
w_\infty^T \Sigma \gamma = \mu_d^T \gamma / (\mu_d^T \Sigma^{-1} \mu_d) = 0.
\]

We write \( \gamma \) for \( \gamma^\lambda \) for short in what follows.

By (1.26), we have \( f(\gamma) \leq f(0) = 0 \). This implies that

\[
R(\gamma) \leq \lambda \sum_{k \in K} (|w_k^\infty| - |w_k^\infty + \gamma_k|) \leq \lambda \sum_{k \in K} |\gamma_k| \leq \lambda \sqrt{s} \|\gamma\|_2.
\]

On the other hand, \( R(\gamma) \geq \lambda_{\min}(\Sigma) \|\gamma\|_2^2 \). Bringing the upper and lower bound of \( R(\gamma) \) together, we conclude that

\[
\|\gamma\|_2 \leq \frac{\lambda \sqrt{s}}{\lambda_{\min}(\Sigma)}.
\]

The proof is now complete.
1.8.3 Proof of Theorem 5

By the positive definiteness of $\Sigma$, $\Sigma^{-1}$ and $\Sigma^{-\frac{1}{2}}$ are also positive definite. Let $v = \Sigma^{1/2}w$, then the transformation $v \mapsto w$ is linear. Define

$$v_c = \arg\min_{\|\Sigma^{-1/2}v\|_1 \leq c} v^T v,$$

where $\mu_d = \Sigma^{-1/2} \mu_d$. It is enough to show that $v_c$ is piecewise linear in $c$.

Let $\Omega_c = \{v : \|\Sigma^{-1/2}v\|_1 \leq c\}$ and $S = \{v : v^T \mu_d = 1\}$. When $c$ is small, the solution set is $\emptyset$; when $c$ is large, the constraint $\Omega_c$ is inactive. Denote by “a” the smallest “c” such that $\Omega_c \cap S \neq \emptyset$, and by “b” the smallest such that $v_c$ are the same for all $c \geq b$. Hence we are interested in $c \in [a, b]$, when changes in $c$ actually affects the solution.

Let $P$ be the projection of the origin $O$ onto the hyperplane $S$ in the $p$ dimensional space. Let

$$F_c = \{S^0_{1,c}, \ldots, S^0_{j_0,c}; S^1_{1,c}, \ldots, S^1_{j_1,c}; \ldots; S^{p-1}_{1,c}, \ldots, S^{p-1}_{j_{p-1},c}\},$$

where $S^i_{j,c}$ denotes an $i$-dimensional face of $\Omega_c$, i.e., $S^0_{j,c}$ represents a vertex, $S^1_{j,c}$ an edge, and $S^{p-1}_{j,c}$ a facet. It is clear that $F_c$ is a finite set.

Define a mapping $\varphi : [a, b] \to \mathbb{Z} \times \mathbb{Z}$, where $\varphi(c) = (i, j)$ such that i) $v_c \in S^i_{j,c}$ and ii) $i$ is minimal. By definition, this mapping is single valued.

For any $c_0 \in (a, b]$, denote $D_{c_0} = \{(i, j)|\forall \epsilon > 0, \exists c \in (c_0 - \epsilon, c_0) \text{ s.t. } \varphi(c) = (i, j)\}$. The set $D_{c_0}$ is non-empty because the collection $\{(i, j) \in \mathbb{Z} \times \mathbb{Z}|S^i_{j,c} \in F_c\}$ is finite. Then the theorem follows from compactness of $[a, b]$ and Lemma 2, Remark 4 and Lemma 3.

**Lemma 1.** $\forall c_0 \in (a, b]$, $\exists \epsilon > 0$ such that $\forall (i, j) \in D_{c_0}$ and $\forall c \in (c_0 - \epsilon, c_0)$, $P^i_{j,c} \in S^i_{j,c} \cap S$, where $P^i_{j,c}$ is the projection of $P$ onto $S \cap \tilde{S}^i_{j,c}$, and $\tilde{S}^i_{j,c}$ denotes the $i$-dimensional affine space in which $S^i_{j,c}$ embeds, and $\tilde{S}^i_{j,c}$ is the interior of $S^i_{j,c}$, where the topology is the natural subspace topology restricted to $S^i_{j,c}$.

**Proof.** Fix $c_0 \in (a, b]$. For any $(i, j) \in D_{c_0}$ and $\epsilon > 0$, by the definition of $D_{c_0}$, there exists $c' \in [c_0 - \epsilon, c_0)$ such that $\varphi(c') = (i, j)$. The minimality of $i$ in the definition for $\varphi$ implies that $v_{c'} = P^i_{j,c'} \in S^i_{j,c'}$, which in the interior of $S^i_{j,c'}$. Therefore, $P^i_{j,c'} \in S^i_{j,c'} \cap S$. By arbitrariness of $\epsilon$, $\exists (c_n) \not\to c_0$ such that $P^i_{j,c_n} \in S^i_{j,c_n} \cap S$ for all $n$. 

34
It can also be shown that \( \{ c | P^i_{j,c} \in S^\iota_j \cap S \} \) is connected: let \( P^i_{j,c_1} \in S^\iota_j \cap S \), \( P^i_{j,c_2} \in S^\iota_j \cap S \), \( c_1 < c_2 \). For any \( c_3 \in (c_1, c_2) \), \( P^i_{j,c_3} \) is on the line segment with endpoints \( P^i_{j,c_1} \) and \( P^i_{j,c_2} \) because \( S^\iota_j \) are parallel affine subspace in \( \mathbb{R}^p \). Let \( S^\iota_j,cone := \bigcup_{c \geq 0} S^\iota_j \). Since \( S^\iota_j \) is a cone, \( c \) is not correct, then pick any \( (i,j) \) such that for any \( \epsilon \) take \( \{ c \} \subseteq S^\iota_j \cap S \). Hence, \( \exists \epsilon > 0 \) such that for all \( c \in [c_0 - \epsilon, c_0) \), \( P^i_{j,c} \in S^\iota_j \). Take \( \epsilon = \min_{(i,j) \in D_{c_0}} \epsilon_{ij} \), the claim follows.

\[ \square \]

**Lemma 2.** \( \forall c_0 \in (a, b], D_{c_0} \) is a singleton, and \( \exists \epsilon' > 0 \) such that \( v_c \) is linear in \( c \) on \( (c_0 - \epsilon', c_0) \).

**Proof.** Fix \( c_0 \in (a, b] \). We claim that for some \( (i, j) \in D_{c_0} \), there exists positive \( \epsilon' (\leq \epsilon \) that validates Lemma 1) such that for any \( c \in (c_0 - \epsilon', c_0) \), \( v_c = P^i_{j,c} \). Assume that the claim is not correct, then pick any \( (i, j) \in D_{c_0} \), there exists a sequence \( \{ c_k \} (c_k \neq c_{k'} \text{ if } k \neq k') \) converging to \( c_0 \) from the left s.t. \( v_{c_k} \neq P^i_{j,c_k} \). Without loss of generality, take \( \{ c_k \} \subset (c_0 - \epsilon, c_0) \). Lemma 1 implies that \( P^i_{j,c_k} \in S^\iota_j \cap S \). If \( v_{c_k} \in S^\iota_j \), we would have \( v_{c_k} = P^i_{j,c_k} \). Hence \( v_{c_k} \notin S^\iota_j \). By finiteness of the index pairs in \( F_c \), there exists \( (i', j') \neq (i, j) \) such that \( \varphi(c) = (i', j') \) for \( c \in \{ c_k \} \), where \( \{ c_k \} \) is some subsequence of \( \{ c_k \} \). This implies \( (i', j') \in D_{c_0} \), which together with Lemma 1 implies \( v_c = P^i_{j,c} \) for \( c \in \{ c_k \} \). Therefore

\[
\| P^i_{j',c} - P \|_2 < \| P^i_{j,c} - P \|_2
\]

for \( c \in \{ c_k \} \).

On the other hand, because \( (i, j) \in D_{c_0} \), there exist infinitely many \( c' \in (c_0 - \epsilon, c_0) \) such that \( \| P^i_{j',c'} - P \|_2 \geq \| P^i_{j,c} - P \|_2 \). Therefore,

\[
g(c) = \| P - P^i_{j,c} \|^2 - \| P - P^i_{j',c} \|^2
\]

changes signs infinitely many times on \( (c_0 - \epsilon, c_0) \). This leads to a contradiction because \( P^i_{j,c} \) and \( P^i_{j',c} \) are both linear functions of \( c \). Hence, the conclusion holds.

To show that \( D_{c_0} \) is a singleton, suppose it has two distinct elements \( (i, j) \) and \( (i', j') \). We have shown that \( v_c = P^i_{j,c} \) and \( v_c = P^i_{j',c} \) for all \( c \) in a left neighborhood of \( c_0 \) (not including \( c_0 \)). Also we have \( P^i_{j,c} \in S^\iota_j \) and \( P^i_{j',c} \in S^\iota_j \) by Lemma 1. This can be true only when \( S^\iota_j \subset S^\iota_{j',c} \) (or vice versa), but then \( i < i' \), contradicting with minimality in definition of \( D_{c_0} \).

\[ \square \]

**Remark 4.** Similarly, \( \forall c_0 \in [a, b], \exists \epsilon' > 0 \) such that \( v_c \) is linear in \( c \) on \( (c_0, c_0 + \epsilon') \).
Lemma 3. \( v_c \) is a continuous function of \( c \) on \([a, b] \).

Proof. The continuity follows from two parts i) and ii).

i) \( \forall c_0 \in [a, b), \exists \epsilon > 0 \) such that \( v_c \) is continuous on \([c_0, c_0 + \epsilon)\). Indeed, let

\[
h(c) = \min_{\|\Sigma^{-1/2} v\|_1 \leq c, \nu^T \mu_d = 1} v^T v.
\]

We know that the mapping \( c \mapsto v_c(= P^i_{j,c}) \) is linear and hence continuous on \((c_0, c_0 + \epsilon)\) for some small \( \epsilon > 0 \). It only remains to show that the mapping is right continuous at \( c_0 \). Notice here \( h(c) = \|P^i_{j,c}\|^2 \) for \( c \in (c_0, c_0 + \epsilon) \). Let \( L = \lim_{c \downarrow c_0} P^i_{j,c} \). It is clear that \( L \in S^i_{j,c_0} \). Because \( L \in \Omega_{c_0} \cap S \), \( h(c) \leq \|L\|^2 \). This inequality has to take the equal sign because \( h(\cdot) \) is monotone decreasing, and \( h(c) = \|P^i_{j,c}\|^2 \rightarrow \|L\|^2 \) as \( c \) approaches \( c_0 \) from the right. Because \( v_{c_0} \) is unique, \( v_{c_0} = L = \lim_{c \downarrow c_0} P^i_{j,c} = \lim_{c \downarrow c_0} v_c \).

ii) \( \forall c_0 \in (a, b], \exists \epsilon > 0 \) such that \( v_c \) is continuous on \((c_0 - \epsilon, c_0]\). Again, it remains to show that there is no jump at \( c_0 \). Let \((i_{c_0}, j_{c_0}) = \varphi(c_0)\). Clearly \( P^i_{j_{c_0},c_0} \in S^i_{j_{c_0},c_0} \). Introduce a notion of parallelism of affine subspaces in \( \mathbb{R}^p \). We denote \( S^i_{j,c_0,c} \parallel S \), if only by translation, \( S^i_{j,c_0,c} \) becomes a subset of \( S \) (or vice versa in other situations); use the notation \( \sim S^i_{j,c_0,c} \parallel S \) otherwise.

If \( S^i_{j,c_0,c} \parallel S \), for \( c \) in some left neighborhood of \( c_0 \), \( P^i_{j,c_0} \) exists and \( P^i_{j,c_0} \in S^i_{j,c_0,c} \). Note \( P^i_{j,c_0} \in \Omega_{c} \cap S \), and \( \|P^i_{j,c_0}\|^2 \rightarrow \|P^i_{j,c_0}\|^2 \) as \( c \) approaches \( c_0 \) from the left. Since \( h(\cdot) \) is monotone decreasing, obviously \( h(c) \rightarrow \|P^i_{j,c_0}\|^2 = h(c_0) \). This shows the left continuity of \( h \) at \( c_0 \). Suppose \( D_{c_0} = \{ (i, j) \} \), then we know on a left neighborhood of \( c_0 \) (not including \( c_0 \)), \( v_c = P^i_{j,c} \). Let \( E = \lim_{c \uparrow c_0} P^i_{j,c} \), then \( E \in \Omega_{c_0} \cap S \). Note that \( \|P^i_{j,c_0}\|^2 \geq \|P^i_{j,c}\|^2 \) for all \( c \) in \( c_0 \)'s left neighborhood, so we have \( \|P^i_{j,c_0}\|^2 \geq \|E\|^2 \). On the other hand, \( \|P^i_{j,c_0}\|^2 \leq \|E\|^2 \) by the definition of \( P^i_{j,c_0} \). Also, consider the uniqueness of distance minimizing point in \( \Omega_{c_0} \cap S \) to origin \( O \), \( E = P^i_{j,c_0} \), and hence \( v_c \) has left continuity at \( c_0 \).

If \( S^i_{j,c_0} \parallel S \), \( \exists Q \in \Omega_{c_0 - \epsilon/2} \cap S \) such that \( Q \neq P^i_{j,c_0} \). When \( c \) goes from \( c_0 - \epsilon/2 \) to \( c_0 \), there exists a point \( Q_c \in \Omega_c \cap S \) moving on the line segment from \( Q \) to \( P^i_{j,c_0} \). Therefore, \( h(\cdot) \) is left continuous at \( c_0 \). Replace \( P^i_{j,c_0} \) by \( Q_c \) in the previous paragraph, the left continuity of \( v_c \) at \( c_0 \) follows from the same argument.

\[\square\]
Chapter 2

Neyman-Pearson Classification

2.1 Introduction

The Neyman-Pearson (NP) paradigm in statistical learning extends the objective of classical binary classification in that, while the latter focuses on minimizing classification error that is a weighted sum of type I and type II errors, the former minimizes type II error with an upper bound $\alpha$ on type I error. With slight abuse of language, in verbal discussion we do not distinguish type I/II error from probability of type I/II error.

For learning with the NP paradigm, it is essential to avoid one kind of error at the expense of the other. As an illustration, consider the following problem in medical diagnosis: failing to detect a malignant tumor has far more severe consequences than flagging a benign tumor. So it makes sense to put priority on controlling the false negative rate. Other scenarios include spam filtering, machine monitoring, target recognition, etc.

In the learning context, as true errors are inaccessible, we cannot enforce almost surely the desired upper bound for type I error. The best we can hope is that a data dependent classifier has type I error bounded with high probability. Ideally, a good classification rule $\hat{f}$ in NP context should satisfy two properties. The first is that type I error of the classifier $\hat{f}$ is bounded from above by a pre-specified level with pre-specified high probability; the second is that $\hat{f}$ has good performance bounds for excess type II error. As will be illustrated, it is unlikely that these two goals can be fulfilled simultaneously. Following the original spirit of NP paradigm, we put priority on type I error and insist on the pre-
specified upper bound $\alpha$. Our proposed learning procedure meets the conservative attitude on type I error, and has good performance bound measured by the excess $\varphi$-type II error. We also discuss the general consequence of being conservative in NP learning.

The paper is organized as follows. In Section 2.2, the classical setup for binary classification is reviewed and the main notation is introduced. A parallel between binary classification and statistical hypothesis testing is drawn in Section 2.3 with emphasis on the NP paradigm in both frameworks. The main propositions and theorems are stated in Section 2.4 while proofs and technical results are relegated to the Appendix. Finally, Section 2.5 illustrates an application of our results to chance constrained optimization.

In the rest of the paper, we denote by $x_j$ the $j$-th coordinate of a vector $x \in \mathbb{R}^d$.

## 2.2 Binary classification

### 2.2.1 Classification risk and classifiers

Let $(X,Y)$ be a random couple where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of covariates and $Y \in \{-1, 1\}$ is a label that indicates to which class $X$ belongs. A classifier $h$ is a mapping $h : \mathcal{X} \to [-1, 1]$ whose sign returns the predicted class given $X$. An error occurs when $-h(X)Y \geq 0$ and it is therefore natural to define the classification loss by $1_{I(-h(X)Y \geq 0)}$, where $1_{I(\cdot)}$ denotes the indicator function.

The expectation of the classification loss with respect to the joint distribution of $(X,Y)$ is called (classification) risk and is defined by

$$R(h) = \mathbb{P}(-h(X)Y \geq 0).$$

Clearly the indicator function is not convex, and for computational convenience, a common practice is to replace it by a convex surrogate [see, e.g. 10, and references therein].
To this end, we rewrite the risk function as

\[ R(h) = \mathbb{E}[\varphi(-h(X)Y)], \quad (2.2.1) \]

where \( \varphi(z) = \mathbb{1}(z \geq 0) \). Convex relaxation can be achieved by simply replacing the indicator function by a convex surrogate.

**Definition 2.2.1.** A function \( \varphi : [-1, 1] \to \mathbb{R}^+ \) is called a convex surrogate if it is non-decreasing, continuous and convex and if \( \varphi(0) = 1 \).

Commonly used examples of convex surrogates are the hinge loss \( \varphi(x) = (1 + x)_+ \), the logit loss \( \varphi(x) = \log_2(1 + e^x) \) and the exponential loss \( \varphi(x) = e^x \).

For a given choice of \( \varphi \), define the \( \varphi \)-risk

\[ R_\varphi(h) = \mathbb{E}[\varphi(-Yh(X))] . \]

Hereafter, we assume that \( \varphi \) is fixed and refer to \( R_\varphi \) as the risk when there is no confusion. In our subsequent analysis, this convex relaxation will also be the ground to analyze a stochastic convex optimization problem subject to stochastic constraints. A general treatment of such problems can be found in Section 2.5.

Because of overfitting, it is unreasonable to look for mappings minimizing empirical risk over all classifiers. Indeed, one could have a small empirical risk but a large true risk. Hence, we resort to regularization. There are in general two ways to proceed. The first is to restrict the candidate classifiers to a specific class \( \mathcal{H} \), and the second is to change the objective function by, for example, adding a penalty term. The two approaches can be combined, and sometimes they are obviously equivalent.

In this paper, we pursue the first idea by defining the class of candidate classifiers as follows. Let \( h_1, \ldots, h_M, M \geq 2 \) be a given collection of classifiers. In our setup, we allow
$M$ to be large. In particular, our results remain asymptotically meaningful as long as $M = o(e^n)$. Such classifiers are usually called base classifiers and can be constructed in a very naive manner. Typical examples include decision stumps or small trees. While the $h_j$’s may have no satisfactory classifying power individually, for over two decades, boosting type of algorithms have successfully exploited the idea that a suitable weighted majority vote among these classifiers may result in low classification risk [66]. Consequently, we restrict our search for classifiers to the set of functions consisting of convex combinations of the $h_j$’s:

$$\mathcal{H}^{\text{conv}} = \{h_\lambda = \sum_{j=1}^{M} \lambda_j h_j, \lambda \in \Lambda\},$$

where $\Lambda$ denotes the flat simplex of $\mathbb{R}^M$ and is defined by $\Lambda = \{\lambda \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^{M} \lambda_j = 1\}$. In effect, classification rules given by the sign of $h \in \mathcal{H}^{\text{conv}}$ are exactly the set of rules produced by the weighted majority votes among the base classifiers $h_1, \ldots, h_M$.

By restricting our search to classifiers in $\mathcal{H}^{\text{conv}}$, the best attainable $\varphi$-risk is called oracle risk and is abusively denoted by $R_\varphi(\mathcal{H}^{\text{conv}})$. As a result, we have $R_\varphi(h) \geq R_\varphi(\mathcal{H}^{\text{conv}})$ for any $h \in \mathcal{H}^{\text{conv}}$ and a natural measure of performance for a classifier $h \in \mathcal{H}^{\text{conv}}$ is given by its excess risk defined by $R_\varphi(h) - R_\varphi(\mathcal{H}^{\text{conv}})$.

The excess risk of a data driven classifier $h_n$ is a random quantity and we are interested in bounding it with high probability. Formally, the statistical goal of binary classification is to construct a classifier $h_n$ such that the oracle inequality

$$R_\varphi(h_n) \leq R_\varphi(h_{\mathcal{H}^{\text{conv}}}) + \Delta_n(\mathcal{H}^{\text{conv}}, \delta)$$

holds with probability $1 - \delta$, where $\Delta_n(\cdot, \cdot)$ should be as small as possible.
In the scope of this paper, we focus on candidate classifiers in the class $\mathcal{H}^{\text{conv}}$. Some of the following results such as Theorem 2.4.1 can be extended to more general classes of classifiers with known complexity such as classes with bounded VC-dimension, as for example in [20]. However, our main argument for bounding $\varphi$-type II error (defined in next subsection) relies on Proposition 2 which, in turn, depends heavily on the convexity of the problem, and it is not clear how it can be extended to more general classes of classifiers.

2.2.2 The Neyman-Pearson paradigm

We make the convention that when $h(X) \geq 0$ the predicted class is $+1$, and $-1$ otherwise. Under this convention, the risk function in classical binary classification can be expressed as a convex combination of type I error $R^-(h) = \mathbb{P}(-Yh(X) \geq 0 | Y = -1)$ and type II error $R^+(h) = \mathbb{P}(-Yh(X) > 0 | Y = 1)$:

$$R(h) = \mathbb{P}(Y = -1)R^-(h) + \mathbb{P}(Y = 1)R^+(h).$$

While the goal of classical binary classification is $\min_{h \in \mathcal{H}} R(h)$, where $\mathcal{H}$ is the set of candidate classifiers, the NP classification targets on

$$\min_{\substack{h \in \mathcal{H} \\colon R^-(h) \leq \alpha}} R^+(h).$$

More generally, we can define the $\varphi$-type I and $\varphi$-type II errors respectively by

$$R^-_\varphi(h) = \mathbb{E}[\varphi(-Yh(X)) | Y = -1] \quad \text{and} \quad R^+_\varphi(h) = \mathbb{E}[\varphi(-Yh(X)) | Y = 1].$$

Our main theorems concern about $R^-_\varphi(\cdot)$ and $R^+_\varphi(\cdot)$, but we will come back and address how convexification and conservativeness affect $R^-(\cdot)$ and $R^+(\cdot)$. 

Following the NP paradigm, for a given class $\mathcal{H}$ of classifiers, we seek to solve the constrained minimization problem:

$$\min_{h \in \mathcal{H}} \min_{R_C(h) \leq \alpha} R_C^+(h),$$

where $\alpha \in (0, 1)$, the significance level, is a constant specified by the user.

NP classification is closely related to the NP approach to statistical hypothesis testing. We now recall a few key concepts about the latter. Many classical works have addressed the theory of statistical hypothesis testing, in particular [53] provides a thorough treatment of the subject.

Statistical hypothesis testing bears strong resemblance with binary classification if we assume the following model. Let $P^-$ and $P^+$ be two probability distributions on $\mathcal{X} \subset \mathbb{R}^d$. Let $p \in (0, 1)$ and assume that $Y$ is a random variable defined by

$$Y = \begin{cases} 
1 \text{ with probability } p, \\
-1 \text{ with probability } 1 - p.
\end{cases}$$

Assume further that the conditional distribution of $X$ given $Y$ is given by $P_Y$. Given such a model, the goal of statistical hypothesis testing is to determine whether $X$ was generated from $P^-$ or $P^+$. To that end, we construct a test $\phi : \mathcal{X} \to [0, 1]$ and the conclusion of the test based on $\phi$ is that $X$ is generated from $P^+$ with probability $\phi(X)$ and from $P^-$ with probability $1 - \phi(X)$. Note that randomness here comes from an exogenous randomization process such as flipping a biased coin. Two kinds of errors arise: type I error occurs when rejecting $P^-$ when it is true, and type II error occurs when accepting $P^-$ when it is false. The Neyman-Pearson paradigm in hypothesis testing amounts to choosing $\phi$ that solves
the following constrained optimization problem

\[
\begin{align*}
\text{maximize} \quad & \mathbb{E}[\phi(X)|Y = 1], \\
\text{subject to} \quad & \mathbb{E}[\phi(X)|Y = -1] \leq \alpha,
\end{align*}
\]

where \( \alpha \in (0, 1) \) is the significance level of the test. In other words, we specify a significance level \( \alpha \) on type I error, and minimize type II error. We call a solution to this problem a most powerful test of level \( \alpha \). The Neyman-Pearson Lemma gives mild sufficient conditions for the existence of such a test.

**Theorem 2.2.1** (Neyman-Pearson Lemma). Let \( P^- \) and \( P^+ \) be probability distributions possessing densities \( p^- \) and \( p^+ \) respectively with respect to some measure \( \mu \). Let \( f_k(x) = \mathbb{1}(L(x) \geq k) \), where the likelihood ratio \( L(x) = p^+(x)/p^-(x) \) and \( k \) is such that \( P^-(L(X) > k) \leq \alpha \) and \( P^-(L(X) \geq k) \geq \alpha \). Then,

- \( f_k \) is a level \( \alpha = \mathbb{E}[\varphi_k(X)|Y = -1] \) most powerful test.

- For a given level \( \alpha \), the most powerful test of level \( \alpha \) is defined by

\[
\phi(X) = \begin{cases} 
1 & \text{if} \quad L(X) > k \\
0 & \text{if} \quad L(X) < k \\
\frac{\alpha - P^-(L(X) > k)}{P^-(L(X) = k)} & \text{if} \quad L(X) = k.
\end{cases}
\]

Notice that in the learning framework, \( \phi \) cannot be computed since it requires the knowledge of the likelihood ratio and of the distributions \( P^- \) and \( P^+ \). Therefore, it remains merely a theoretical proposition. Nevertheless, the result motivates the NP paradigm pursued here.
2.3 Neyman-Pearson classification via convex optimization

Recall that in NP classification with a loss function $\varphi$, the goal is to solve the problem (2.2.3). This cannot be done directly as conditional distributions $P^{-}$ and $P^{+}$, and hence $R_{\varphi}^{-}$ and $R_{\varphi}^{+}$, are unknown. In statistical applications, information about these distributions is available through two i.i.d. samples $X_{1}^{-}, \ldots, X_{n^{-}}^{-}$, $n^{-} \geq 1$ and $X_{1}^{+}, \ldots, X_{n^{+}}^{+}$, $n^{+} \geq 1$, where $X_{i}^{-} \sim P^{-}, i = 1, \ldots, n^{-}$ and $X_{i}^{+} \sim P^{+}, i = 1, \ldots, n^{+}$. We do not assume that the two samples $(X_{1}^{-}, \ldots, X_{n^{-}}^{-})$ and $(X_{1}^{+}, \ldots, X_{n^{+}}^{+})$ are mutually independent. Presently the sample sizes $n^{-}$ and $n^{+}$ are assumed to be deterministic and will appear in the subsequent finite sample bounds. A different sampling scheme, where these quantities are random, is investigated in subsection 2.4.3.

2.3.1 Conservativeness on type I error

While the binary classification problem has been extensively studied, theoretical proposition on how to implement the NP paradigm remains scarce. To the best of our knowledge, [20] initiated the theoretical treatment of the NP classification paradigm and an early empirical study can be found in [22]. The framework of [20] is the following. Fix a constant $\varepsilon_{0} > 0$ and let $\mathcal{H}$ be a given set of classifiers with finite VC dimension. They study a procedure that consists of solving the following relaxed empirical optimization problem

$$\min_{h \in \mathcal{H}} \quad \hat{R}^{+}(h), \quad \hat{R}^{-}(h) \leq \alpha + \varepsilon_{0}/2$$

(2.3.1)

where

$$\hat{R}^{-}(h) = \frac{1}{n^{-}} \sum_{i=1}^{n^{-}} \mathbb{I}(h(X_{i}^{-}) \geq 0), \quad \text{and} \quad \hat{R}^{+}(h) = \frac{1}{n^{+}} \sum_{i=1}^{n^{+}} \mathbb{I}(h(X_{i}^{+}) \leq 0)$$

44
denote the empirical type I and empirical type II errors respectively. Let \( \hat{h} \) be a solution to (2.3.1). Denote by \( h^* \) a solution to the original Neyman-Pearson optimization problem:

\[
h^* \in \arg\min_{\substack{\hat{h} \in \mathcal{H} \\ R^-(\hat{h}) \leq \alpha}} R^+(\hat{h}),
\]

The main result of \cite{20} states that, simultaneously with high probability, the type II error \( R^+(\hat{h}) \) is bounded from above by \( R^+(h^*) + \varepsilon_1 \), for some \( \varepsilon_1 > 0 \) and the type I error of \( \hat{h} \) is bounded from above by \( \alpha + \varepsilon_0 \). In a later paper, \cite{21} considers problem (2.3.1) for a data-dependent family of classifiers \( \mathcal{H} \), and bound estimation errors accordingly. Several results for traditional statistical learning such as PAC bounds or oracle inequalities have been studied in \cite{67} and \cite{68} in the same framework as the one laid down by \cite{20}. A noteworthy departure from this setup is \cite{69} where sensible performance measures for NP classification that go beyond analyzing separately two kinds of errors are introduced. Furthermore, \cite{?} develops a general solution to semi-supervised novelty detection by reducing it to NP classification. Recently, \cite{46} transposed several results of \cite{20} and \cite{68} to NP classification with convex loss.

The present work departs from previous literature in our treatment of type I error. In fact, the classifiers in all the papers mentioned above take a compromise on the pre-determined upper bound on type I error, i.e., they ensure that \( \mathbb{P}(R^-(\hat{h}) > \alpha + \varepsilon_0) \) is small, for some \( \varepsilon_0 > 0 \). However, it is our primary interest to make sure that \( R^-(\hat{h}) \leq \alpha \) with high probability, following the original principle of the Neyman-Pearson paradigm that type I error should be controlled by a pre-specified level \( \alpha \). As we follow an empirical risk minimization procedure, to control \( \mathbb{P}(R^-(\hat{h}) > \alpha) \), it is necessary to have \( \hat{h} \) be a solution to some program with a strengthened constraint on empirical type I error. If our concern is only on type I error, we can just do so. However, we also want to evaluate the excess
type II error. Our conservative attitude on type I error faces new technical challenges which we summarize here. In the approach of [20] and of [68], the relaxed constraint on the type I error is constructed such that the constraint $\hat{R}^- (h) \leq \alpha + \varepsilon_0 / 2$ on type I error in (2.3.1) is satisfied by $h^*$ (defined in (2.3.2)) with high probability, and that this classifier accommodates the excess type II error well. As a result, the control of type II error mainly follows as a standard exercise to control suprema of empirical processes. This is not the case here; we have to develop methods to control the optimum value of an optimization problem under a stochastic constraint. Such methods have consequences not only in NP classification but also on chance constraint programming as explained in Section 2.5.

2.3.2 Convexified NP classifier

Concerned about computational feasibility, our proposed classifier is the solution to a convex program, which is an empirical form NP classification problem (2.2.3) where the distribution of the observations is unknown. In view of the arguments presented in the previous subsection, we cannot simply replace the unknown risk functions by their empirical counterparts. The treatment of the convex constraint should be done carefully and we proceed as follows.

For any classifier $h$ and a given convex surrogate $\varphi$, define $\hat{R}^-_\varphi$ and $\hat{R}^+_{\varphi}$ to be the empirical counterparts of $R^-_\varphi$ and $R^+_{\varphi}$ respectively by

$$\hat{R}^-_\varphi (h) = \frac{1}{n^-} \sum_{i=1}^{n^-} \varphi (h(X^-_i)),$$

and

$$\hat{R}^+_{\varphi} (h) = \frac{1}{n^+} \sum_{i=1}^{n^+} \varphi (-h(X^+_i)).$$

Moreover, for any $a > 0$, let $\mathcal{H}^{\varphi,a} = \{ h \in \mathcal{H}^{\text{conv}} : R^-_{\varphi} (h) \leq a \}$ be the set of classifiers in $\mathcal{H}^{\text{conv}}$ whose convexified type I errors are bounded from above by $a$, and let $\mathcal{H}^{\varphi,a}_{n^-} = \{ h \in$
\( \mathcal{H}^{\text{conv}} : \hat{R}_\varphi(h) \leq a \) be the set of classifiers in \( \mathcal{H}^{\text{conv}} \) whose empirical convexified type I errors are bounded by \( a \). To make our analysis meaningful, we assume that \( \mathcal{H}^{\varphi, \alpha} \neq \emptyset \).

We are now in a position to construct a classifier in \( \mathcal{H}^{\text{conv}} \) according to the Neyman-Pearson paradigm. For any \( \tau > 0 \) such that \( \tau \leq \alpha \sqrt{n} \), define the convexified NP classifier \( \tilde{h}^\tau \) as any classifier that solves the following optimization problem

\[
\min_{h \in \mathcal{H}^{\text{conv}}} \begin{cases} 
\hat{R}_\varphi(h) \leq \alpha - \tau \sqrt{n}, \\
\hat{R}_\varphi(h) \leq a - \tau / \sqrt{n}.
\end{cases}
\]  

(2.3.3)

Note that this problem consists of minimizing a convex function subject to a convex constraint and can therefore be solved by standard algorithms such as [see, e.g., 17, and references therein].

In the next section, we present a series of results on type I and type II errors of classifiers that are more general than \( \tilde{h}^\tau \).

### 2.4 Performance Bounds

In this section, we will first evaluate our proposed classifier \( \tilde{h}^\tau \) against \( \varphi \) I/II errors. These benchmarks are necessary because \( \tilde{h}^\tau \) is constructed based on them. Moreover, in view of the decision theory framework, such errors are just expected loss with a general loss function \( \varphi \), which are interesting to investigate. As the true type I and type II errors are usually the main concern in statistical learning, we will also address the effect of convexification in terms of the excess type II error. Interestingly, given that we want to be conservative on type I error, neither working on \( \varphi \) errors nor working on true errors leads to a most desirable type II error. The price to pay for being conservative will be characterized explicitly.
2.4.1 Control of type I error

First, we identify classifiers $h$ such that $R_\varphi(h) \leq \alpha$ with high probability. This is done by enforcing its empirical counterpart $\hat{R}_\varphi(h)$ be bounded from above by the quantity

$$
\alpha_\kappa = \alpha - \kappa / \sqrt{n},
$$

for a proper choice of positive constant $\kappa$.

**Theorem 2.4.1.** Fix constants $\delta, \alpha \in (0, 1), L > 0$ and let $\varphi : [-1, 1] \to \mathbb{R}^+$ be a given $L$-Lipschitz convex surrogate. Define

$$
\kappa = 4\sqrt{2L} \sqrt{\log \left( \frac{2M}{\delta} \right)}.
$$

Then for any (random) classifier $h \in \mathcal{H}^{\text{conv}}$ that satisfies $\hat{R}_\varphi(h) \leq \alpha_\kappa$, we have

$$
R^-(h) \leq R_\varphi^-(h) \leq \alpha.
$$

with probability at least $1 - \delta$. Equivalently

$$
\mathbb{P} \left[ \mathcal{H}_{n-}^{\varphi, \alpha_\kappa} \subset \mathcal{H}_{n-}^{\varphi, \alpha} \right] \geq 1 - \delta. \quad (2.4.1)
$$

2.4.2 Simultaneous control of the two errors

Theorem 2.4.1 guarantees that any classifier that satisfies the strengthened constraint on the empirical $\varphi$-type I error will have $\varphi$-type I error and true type I error bounded from above by $\alpha$. We now check that the constraint is not too strong so that the $\varphi$-type II error is overly deteriorated. Indeed, an extremely small $\alpha_\kappa$ would certainly ensure a good control of type I error but would deteriorate significantly the best achievable $\varphi$-type II error. Below, we show not only that this is not the case for our approach but also that the convexified NP classifier $\tilde{h}_\tau$ defined in subsection 2.3.2 with $\tau = \alpha_\kappa$ suffers only a small
degradation of its $\varphi$-type II error compared to the best achievable. Analogous to classical binary classification, a desirable result is that with high probability,

$$R^+_{\varphi}(\tilde{h}^{\kappa}) - \min_{h \in H_{\varphi,\alpha}^{-\nu}} R^+_{\varphi}(h) \leq \tilde{\Delta}_n(\mathcal{F}),$$

(2.4.2)

where $\tilde{\Delta}_n(\mathcal{F})$ goes to 0 as $n = n^- + n^+ \to \infty$.

The following proposition is pivotal to our argument.

**Proposition 2.** Fix constant $\alpha \in (0,1)$ and let $\varphi : [-1,1] \to \mathbb{R}^+$ be a given continuous convex surrogate. Assume further that there exists $\nu_0 > 0$ such that the set of classifiers $H_{\varphi,\nu_0}$ is nonempty. Then, for any $\nu \in (0,\nu_0)$,

$$\min_{h \in H_{\varphi,\alpha-\nu}} R^+_{\varphi}(h) - \min_{h \in H_{\varphi,\alpha}} R^+_{\varphi}(h) \leq \varphi(1) \frac{\nu}{\nu_0 - \nu}.$$

This proposition ensures that if the convex surrogate $\varphi$ is continuous, strengthening the constraint on type I error ($\varphi$-type I error) does not increase too much the best achievable $\varphi$-type II error. We should mention that the proof does not use the Lipschitz property of $\varphi$, but only that it is uniformly bounded by $\varphi(1)$ on $[-1,1]$. This proposition has direct consequences on chance constrained programming as discussed in Section 2.5.

The next theorem shows that the NP classifier $\tilde{h}^{\kappa}$ defined in subsection 2.3.2 is a good candidate to perform classification with the Neyman-Pearson paradigm. It relies on the following assumption which is necessary to verify the condition of Proposition 2.

**Assumption 1.** There exists a positive constant $\varepsilon < 1$ such that the set of classifiers $H_{\varphi,\varepsilon^\alpha}$ is nonempty.

Note that this assumption can be tested using (2.4.1) for large enough $n^-$. Indeed, it follows from this inequality that with probability $1 - \delta$,

$$H_{n^-}^{\varphi,\varepsilon^\alpha-\kappa/\sqrt{n^-}} \subset H_{n^-}^{\varphi,\varepsilon^\alpha-\kappa/\sqrt{n^-}+\kappa/\sqrt{n^-}} = H_{n^-}^{\varphi,\varepsilon^\alpha}.$$
Thus, it is sufficient to check if \( \mathcal{H}_{n^{-}}^{\varphi,\varepsilon,\alpha} - \kappa/\sqrt{n} \) is nonempty for some \( \varepsilon > 0 \). Before stating our main theorem, we need the following definition. Under Assumption 1, let \( \bar{\varepsilon} \) denote the smallest \( \varepsilon \) such that \( \mathcal{H}_{n^{-}}^{\varphi,\varepsilon,\alpha} \neq \emptyset \) and let \( n_0 \) be the smallest integer such that

\[
n_0 \geq \left( \frac{4\kappa}{(1 - \bar{\varepsilon})\alpha} \right)^2.
\]

(2.4.3)

**Theorem 2.4.2.** Let \( \varphi, \kappa, \delta \) and \( \alpha \) be the same as in Theorem 2.4.1, and \( \tilde{h}^{\kappa} \) denote any solution to (2.3.3). Moreover, let Assumption 1 hold and assume that \( n^{-} \geq n_0 \) where \( n_0 \) is defined in (2.4.3). Then, the following hold with probability \( 1 - 2\delta \),

\[
R^{-}(\tilde{h}^{\kappa}) \leq R_{\varphi}^{-}(\tilde{h}^{\kappa}) \leq \alpha
\]

(2.4.4)

and

\[
R_{\varphi}^{+}(\tilde{h}^{\kappa}) - \min_{h \in \mathcal{H}_{n^{-}}^{\varphi,\alpha}} R_{\varphi}^{+}(h) \leq \frac{4\varphi(1)\kappa}{(1 - \bar{\varepsilon})\alpha \sqrt{n^{-}}} + \frac{2\kappa}{\sqrt{n^{+}}}.
\]

(2.4.5)

In particular, there exists a constant \( C > 0 \) depending on \( \alpha, \varphi(1) \) and \( \bar{\varepsilon} \), such that (2.4.5) yields

\[
R_{\varphi}^{+}(\tilde{h}^{\kappa}) - \min_{h \in \mathcal{H}_{n^{-}}^{\varphi,\alpha}} R_{\varphi}^{+}(h) \leq C \left( \sqrt{\frac{\log(2M/\delta)}{n^{-}}} + \sqrt{\frac{\log(2M/\delta)}{n^{+}}} \right).
\]

Note here that Theorem 4.2 is not exactly of the type (2.4.2). The right hand side of (2.4.5) goes to zero if both \( n^{-} \) and \( n^{+} \) go to infinity. Inequality (2.4.5) conveys a message that accuracy of the estimate depends on information from both classes of labeled data. This concern motivates us to consider a different sampling scheme.

### 2.4.3 A Different Sampling Scheme

In this subsection (only), we consider a model for observations that is more standard in statistical learning theory [see, e.g., 14, 26].

Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be \( n \) independent copies of the random couple \((X, Y) \in \mathcal{X} \times \{-1, 1\}\). Denote by \( P_X \) the marginal distribution of \( X \) and by \( \eta(x) = E[Y|X = x] \)
the regression function of $Y$ onto $X$. Denote by $p$ the probability of positive label and observe that

$$p = \Pr[Y = 1] = \mathbb{E}(\Pr[Y = 1|X]) = \frac{1 + \mathbb{E}[\eta(X)]}{2}.$$ 

In what follows, we assume that $P_X(\eta(X) = -1) \lor P_X(\eta(X) = 1) < 1$ so that $p \in (0, 1)$.

Let $N^- = \text{card}\{Y_i : Y_i = -1\}$ be the random number of instances labeled $-1$ and $N^+ = n - N^- = \text{card}\{Y_i : Y_i = 1\}$. In this setup, the NP classifier is defined as in subsection 2.3.2 where $n^-$ and $n^+$ are replaced by $N^-$ and $N^+$ respectively. To distinguish this classifier from $\tilde{h}^\tau$ previously defined, we denote the NP classifier obtained with this sampling scheme by $\tilde{h}^\tau_n$.

Let the event $\mathcal{F}$ be defined by

$$\mathcal{F} = \{R^-_\varphi(\tilde{h}^\kappa_n) \leq \alpha\} \cap \\{R^+_\varphi(\tilde{h}^\kappa_n) - \min_{h \in H_n, \alpha} R^+_{\varphi}(h) \leq \frac{4\varphi(1)\kappa}{(1 - \bar{\varepsilon})\alpha \sqrt{N^-}} + \frac{2\kappa}{\sqrt{N^+}}\}.$$ 

Denote $\mathcal{B}_{n^-} = \{Y_1 = \cdots = Y_{n^-} = -1, Y_{n^-+1} = \cdots = Y_n = 1\}$. Although the event $\mathcal{B}_{n^-}$ is different from the event $\{N^- = n^-\}$, symmetry leads to the following key observation:

$$\Pr(\mathcal{F}|N^- = n^-) = \Pr(\mathcal{F}|\mathcal{B}_{n^-}).$$

Therefore, under the conditions of Theorem 2.4.2, we find that for $n^- \geq n_0$ the event $\mathcal{F}$ satisfies

$$\Pr(\mathcal{F}|N^- = n^-) \geq 1 - 2\delta.$$ (2.4.6)

We obtain the following corollary of Theorem 2.4.2.

**Corollary 2.4.1.** Let $\varphi$, $\kappa$, $\delta$, and $\alpha$ be the same as in Theorem 2.4.1, and $\tilde{h}^\kappa_n$ be the NP classifier obtained with the current sampling scheme. Then under Assumption 1, if $n > 2n_0/(1 - p)$, where $n_0$ is defined in (2.4.3), we have with probability $(1 - 2\delta)(1 - e^{-n(1 - p)^2})$,

$$R^-_\varphi(\tilde{h}^\kappa_n) \leq R^-_\varphi(\tilde{h}^\kappa_n) \leq \alpha.$$ (2.4.7)
and
\[ R^+_\varphi(\tilde{h}_n^\kappa) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) \leq \frac{4\varphi(1)\kappa}{(1 - \bar{\varepsilon})\alpha\sqrt{N}} + \frac{2\kappa}{\sqrt{N^+}}. \] (2.4.8)

Moreover, with probability \(1 - 2\delta - e^{-n(1-p)^2\alpha^2} - e^{-np^2\alpha^2}\), we have simultaneously (2.4.7) and
\[ R^+_\varphi(\tilde{h}_n^\kappa) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) \leq \frac{4\sqrt{2}\varphi(1)\kappa}{(1 - \bar{\varepsilon})\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\kappa}{\sqrt{np^2}}. \] (2.4.9)

2.4.4 Price to pay for being conservative

We have shown that the computational feasible classifier \(\tilde{h}^\kappa\) satisfies oracle inequalities which take the optimal \(\varphi\)-type II errors as the benchmark. In this subsection, the excess type II error will be measured, and we will characterize the price to pay by being conservative on type I error.

Much like its counterparts in classical binary classification, the next strikingly simple relation addresses the consequence of convexification in the NP paradigm.

**Theorem 2.4.3.** Let \(\tilde{h}\) be any classifier, then
\[ R^+(\tilde{h}) - \min_{R^-(h) \leq \alpha} R^+(h) \leq R^\varphi_+(\tilde{h}) - \inf_{R^-(h) \leq \alpha} R^\varphi_+(h). \]

This theorem applies to any classifier; in particular, it holds for our proposed \(\tilde{h}^\kappa\). As the proof of Theorem 2.4.3 indicates, \(\min_{R^-(h) \leq \alpha} R^+(h) = \inf_{R^-(h) \leq \alpha} R^\varphi_+(h)\). So the bound in the theorem can be very tight, depending on the nature of \(\tilde{h}\).

Now relax the range of base classifiers \(h_1, \ldots, h_M\) to be \([-B, B]\). Denote by \(\mathcal{H}^{\varphi,\alpha}_B\) the set of convex combinations of the base classifiers that have \(\varphi\)-type I error bounded from above by \(\alpha\).

Therefore, we have the following observation:
\[ R^+(\tilde{h}^\kappa) - \min_{R^-(h) \leq \alpha} R^+(h) \leq T_1 + T_2 + T_3. \]
where

\[ T_1 = R^+_{\varphi}(\tilde{h}^\kappa) - \min_{h \in \mathcal{H}^\phi,\alpha} R^+_{\varphi}(h), \]

\[ T_2 = \min_{h \in \mathcal{H}^\phi,\alpha} R^+_{\varphi}(h) - \inf_{R^-(h) \leq \alpha, -B \leq h \leq B} R^+_{\varphi}(h), \]

\[ T_3 = \inf_{R^-(h) \leq \alpha, -B \leq h \leq B} R^+_{\varphi}(h) - \inf_{R^-(h) \leq \alpha} R^+_{\varphi}(h). \]

With the new set of base classifiers taking ranges in \([-B, B]\), Theorem 2.4.2 holds if we replace \(\kappa\) by \(\kappa_B = 4\sqrt{2L_B B \log(2M/\delta)}\), where \(L_B\) is the Lipschitz constant of \(\varphi\) on \([-B, B]\). Therefore, the convergence rate of \(T_1\) is explicitly controlled. We can see that with a fixed sample size, choosing a set of base classifiers with smaller range will result in a tighter bound for the excess \(\varphi\)-type II error. However, if one concerns more about the true type II error, choosing a smaller \(B\) should not be a better option, because only signs matters for true type I and II errors. This intuition is reflected in the term \(T_3\). When \(B\) increases, \(T_3\) decreases. More specifically, it can be shown that

\[ T_3 = \left( P^+(\mathcal{X}^+)\varphi(-B) + P^+(\mathcal{X}^-)\varphi(0) \right) - P^+(\mathcal{X}^-) = P^+(\mathcal{X}^+)\varphi(-B), \]

where \(\mathcal{X}^+ \subset \mathcal{X}\) is the part of feature space mapped to label +1 by the optimal NP classifier that solves \(\min_{R^-(h) \leq \alpha} R^+(h)\), and \(\mathcal{X}^-\) is the part that mapped to label \(-1\); this is what NP Lemma says when there is no need for randomization. Therefore, \(T_3\) diminishes towards 0 as \(B\) increases, and the trade-off between \(T_1\) and \(T_3\) is very clear. When \(\varphi(x) = (1 + x)_+\) is the hinge loss, the best trade-off occurs at \(B \in (0, 1)\). When \(B(\geq 1)\) goes to infinity, \(T_3 = 0\) stays the same while the upper bound of \(T_1\) blows up.

Note that \(\mathcal{H}^\varphi_{\mathcal{X},\alpha} \subset \{ h : R^-(h) \leq \alpha, -B \leq h \leq B \}\), so \(T_2\) reflects the price to pay for being conservative on type I error. It also reflect the bias for choosing a specific candidate pool of classifiers, i.e., convex combinations of base classifiers. As long as
the base classifiers are rich enough, the latter bias should be small. However in our belief, the price to pay for being conservative is unavoidable. Even if we do not resort to convexification, getting the best insurance on type I error still demands a high premium on type II error.

The same attitude is shared in the seminal paper [20], where it was claimed without justification that if we use $\alpha' < \alpha$ for the empirical program, “it seems unlikely that we can control the estimation error $R^+(\hat{h}) - R^+(h^*)$ in a distribution independent way”. The following proposition confirms this opinion in a certain sense.

Fix $\alpha \in (0, 1), n^- \geq 1, n^+ \geq 1$ and $\alpha' < \alpha$. Let $\hat{h}(\alpha')$ be the classifier defined as any solution of the following optimization problem:

$$\min_{\hat{h} \in \mathcal{H}} \hat{R}^+(h).$$

The following negative result holds not only for this estimator but also for the oracle $h^*(\alpha')$ defined as the solution of

$$\min_{h \in \mathcal{H}} R^+(h).$$

Note that $h^*(\alpha')$ is not a classifier but only a pseudo-classifier since it depends on the unknown distribution of the data.

**Proposition 3.** There exist base classifiers $h_1, h_2$ and a probability distribution for $(X, Y)$ for which, regardless of the sample sizes $n^-$ and $n^+$, any pseudo-classifier $h_\lambda = \lambda h_1 + (1 - \lambda)h_2, 0 \leq \lambda \leq 1$, such that $R^-(h_\lambda) < \alpha$, it holds

$$R^+(h_\lambda) - \min_{R^-(h) \leq \alpha'} R^+(h) \geq \alpha.$$

In particular, the excess type II risk of $h^*(\alpha - \varepsilon_{n^-}), \varepsilon_{n^-} > 0$ does not converge to zero as sample sizes increase even if $\varepsilon_{n^-} \to 0$. Moreover, when $\alpha \leq 1/2$ for any (pseudo-)classifier
\( h_\lambda (0 \leq \lambda \leq 1) \) such that \( R^-(h_\lambda) < \alpha \), it holds

\[
R^+(h_\lambda) - \min_{R^-(h_\lambda) \leq \alpha, \lambda \in [0,1]} R^+(h_\lambda) \geq \alpha.
\]  

(2.4.10)

with probability at least \( \alpha \land 1/4 \). In other words, if we let \( \mathcal{A} = \{h_\lambda : R^-(h_\lambda) < \alpha, \lambda \in [0,1]\} \), and \( \mathcal{B} = \{h_\lambda : R^+(h_\lambda) - \min_{R^-(h_\lambda) \leq \alpha, \lambda \in [0,1]} R^+(h_\lambda) \geq \alpha, \lambda \in [0,1]\} \), then \( \mathbb{P}(\mathcal{A} \subset \mathcal{B}) \geq \alpha \land 1/4 \). In particular, the excess type II risk of \( \hat{h}(\alpha - \varepsilon_\text{n}^-) \), \( \varepsilon_\text{n}^- > 0 \) does not converge to zero with positive probability, as sample sizes increase even if \( \varepsilon_\text{n}^- \to 0 \).

The proof of this result is postponed to the appendix. The fact that the oracle \( h^*(\alpha - \varepsilon_\text{n}^-) \) satisfies the lower bound indicates that the problem comes from using a strengthened constraint. Note that the condition \( \alpha \leq 1/2 \) is purely technical and can be removed. Nevertheless, it is always the case in practice that \( \alpha \leq 1/2 \). When the number of base classifiers is great then two, we believe that similar counterexamples can be still constructed, though the technicality will be more involved.

In view of this negative result and our previous discussion, we have to accept the price to pay for being conservative on type I error, and our classifier \( \tilde{h}^\kappa \) is no exception. As such conservativeness follows from the original spirit of the Neyman-Pearson paradigm, we need to pay whatever we have to pay. The positive sides are that our proposed procedure is computationally feasible, and it attains good rates under a different (but still meaningful) criterion.

### 2.5 Chance constrained optimization

Implementing the Neyman-Pearson paradigm for the convexified binary classification bears strong connections with chance constrained optimization. A recent account of such problems can be found in [11, Chapter 2] and we refer to this book for references and
applications. A chance constrained optimization problem is of the following form:

\[
\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{P}\{F(\lambda, \xi) \leq 0\} \geq 1 - \alpha,
\]  

(2.5.1)

where \( \xi \in \Xi \) is a random vector, \( \Lambda \subset \mathbb{R}^M \) is convex, \( \alpha \) is a small positive number and \( f \) is a deterministic real valued convex function. Problem (2.5.1) can be viewed as a relaxation of robust optimization. Indeed, for the latter, the goal is to solve the problem

\[
\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \sup_{\xi \in \Xi} F(\lambda, \xi) \leq 0,
\]  

(2.5.2)

and this essentially corresponds to (2.5.1) for the case \( \alpha = 0 \). For simplicity, we take \( F \) to be scalar valued but extensions to vector valued functions and conic orders are considered in [see, e.g., 11, Chapter 10]. Moreover, it is standard to assume that \( F(\cdot, \xi) \) is convex almost surely.

Problem (2.5.1) may not be convex because the chance constraint \( \{\lambda \in \Lambda : \mathbb{P}\{F(\lambda, \xi) \leq 0\} \geq 1 - \alpha\} \) is not convex in general and thus may not be tractable. To solve this problem, [62] and [52] have derived sufficient conditions on the distribution of \( \xi \) for the chance constraint to be convex. On the other hand, [19] initiated a different treatment of the problem where no assumption on the distribution of \( \xi \) is made, in line with the spirit of statistical learning. In that paper, they introduced the so-called scenario approach based on a sample \( \xi_1, \ldots, \xi_n \) of independent copies of \( \xi \). The scenario approach consists of solving

\[
\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad F(\lambda, \xi_i) \leq 0, \quad i = 1, \ldots, n.
\]  

(2.5.3)

[19] showed that under certain conditions, if the sample size \( n \) is bigger than some \( n(\alpha, \delta) \), then with probability \( 1 - \delta \), the optimal solution \( \hat{\lambda}^{sc} \) of (2.5.3) is feasible for (2.5.1). The
authors did not address the control of the term \( f(\hat{\lambda}^{sc}) - f^* \) where \( f^* \) denotes the optimal objective value in (2.5.1). However, in view of Proposition 3, it is very unlikely that this term can be controlled well.

In an attempt to overcome this limitation, a new *analytical* approach was introduced by [59]. It amounts to solving the following convex optimization problem

\[
\min_{\lambda \in \Lambda, t \in \mathbb{R}^s} f(\lambda) \quad \text{s.t.} \quad G(\lambda, t) \leq 0, \tag{2.5.4}
\]

in which \( t \) is some additional instrumental variable and where \( G(\cdot, t) \) is convex. The problem (2.5.4) provides a conservative convex approximation to (2.5.1), in the sense that every \( x \) feasible for (2.5.4) is also feasible for (2.5.1). [59] considered a particular class of conservative convex approximation where the key step is to replace \( \mathbb{P}\{F(\lambda, \xi) \geq 0\} \) by \( \mathbb{E}\varphi(F(\lambda, \xi)) \) in (2.5.1), where \( \varphi \) a nonnegative, nondecreasing, convex function that takes value 1 at 0. [59] discussed several choices of \( \varphi \) including hinge and exponential losses, with a focus on the latter that they name *Bernstein Approximation*.

The idea of a conservative convex approximation is also what we employ in our paper. Recall that \( P^- \) the conditional distribution of \( X \) given \( Y = -1 \). In a parallel form of (2.5.1), we cast our target problem as

\[
\min_{\lambda \in \Lambda} R^+(h_\lambda) \quad \text{s.t.} \quad P^-(h_\lambda(X) \leq 0) \geq 1 - \alpha, \tag{2.5.5}
\]

where \( \Lambda \) is the flat simplex of \( \mathbb{R}^M \).

Problem (2.5.5) differs from (2.5.1) in that \( R^+(h_\lambda) \) is not a convex function of \( \lambda \). Replacing \( R^+(h_\lambda) \) by \( R^+_\varphi(h_\lambda) \) turns (2.5.5) into a standard chance constrained optimization problem:

\[
\min_{\lambda \in \Lambda} R^+_\varphi(h_\lambda) \quad \text{s.t.} \quad P^-(h_\lambda(X) \leq 0) \geq 1 - \alpha. \tag{2.5.6}
\]
However, there are two important differences in our setting, so that we cannot use directly Scenario Approach or Bernstein Approximation or other analytical approaches to (2.5.1). First, $R^+_{\phi}(f_\lambda)$ is an unknown function of $\lambda$. Second, we assume minimum knowledge about $P^-$. On the other hand, chance constrained optimization techniques in previous literature assume knowledge about the distribution of the random vector $\xi$. For example, [59] require that the moment generating function of the random vector $\xi$ is efficiently computable to study the Bernstein Approximation.

Given a finite sample, it is not feasible to construct a strictly conservative approximation to the constraint in (2.5.6). On the other hand, it is possible to ensure that if we learned $\hat{h}$ from the sample, this constraint is satisfied with high probability $1 - \delta$, i.e., the classifier is approximately feasible for (2.5.6). In retrospect, our approach to (2.5.6) is an innovative hybrid between the analytical approach based on convex surrogates and the scenario approach.

We do have structural assumptions on the problem. Let $g_j, j \in \{1, \ldots, M\}$ be arbitrary functions that take values in $[-1, 1]$ and $F(\lambda, \xi) = \sum_{j=1}^{N} \lambda_j g_j(\xi)$. Consider a convexified version of (2.5.1):

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{E}[\varphi(F(\lambda, \xi))] \leq \alpha,$$

where $\varphi$ is a $L$-Lipschitz convex surrogate, $L > 0$. Suppose that we observe a sample $(\xi_1, \ldots, \xi_n)$ that are independent copies of $\xi$. We propose to approximately solve the above problem by

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \sum_{i=1}^{n} \varphi(F(\lambda, \xi_i)) \leq n\alpha - \kappa \sqrt{n},$$

58
for some $\kappa > 0$ to be defined. Denote by $\tilde{\lambda}$ any solution to this problem and by $f^*_\varphi$ the value of the objective at the optimum in (2.5.7). The following theorem summarizes our contribution to chance constrained optimization.

Theorem 2.5.1. Fix constants $\delta, \alpha \in (0, 1/2)$, $L > 0$ and let $\varphi : [-1, 1] \to \mathbb{R}^+$ be a given $L$-Lipschitz convex surrogate. Define

$$\kappa = 4\sqrt{2L} \sqrt{\log \left(\frac{2M}{\delta}\right)}.$$

Then, the following hold with probability at least $1 - 2\delta$

(i) $\tilde{\lambda}$ is feasible for (2.5.1).

(ii) If there exists $\varepsilon \in (0, 1)$ such that the constraint $\mathbb{E}[\varphi(F(\lambda, \xi))] \leq \varepsilon \alpha$ is feasible for some $\lambda \in \Lambda$, then for

$$n \geq \left(\frac{4\kappa}{(1 - \varepsilon)\alpha}\right)^2,$$

we have

$$f(\tilde{\lambda}) - f^*_\varphi \leq \frac{4\varphi(1)\kappa}{(1 - \varepsilon)\alpha \sqrt{n}}.$$

In particular, as $M$ and $n$ go to infinity with all other quantities kept fixed, we obtain

$$f(\tilde{\lambda}) - f^*_\varphi = O\left(\sqrt{\frac{\log M}{n}}\right).$$

The proof essentially follows that of Theorem 2.4.2 and we omit it. The limitations of Theorem 2.5.1 include rigid structural assumptions on the function $F$ and on the set $\Lambda$. While the latter can be easily relaxed using more sophisticated empirical process theory, the former is inherent to our analysis.
2.6 Appendix

2.6.1 Proof of Theorem 2.4.1

We begin with the following lemma, which is extensively used in the sequel. Its proof relies on standard arguments to bound suprema of empirical processes. Recall that \( \{h_1, \ldots, h_M\} \) is family of \( M \) classifiers such that \( h_j : \mathcal{X} \to [-1, 1] \) and that for any \( \lambda \) in the simplex \( \Lambda \subset \mathbb{R}^M \), \( h_\lambda \) denotes the convex combination defined by

\[
h_\lambda = \sum_{j=1}^{N} \lambda_j h_j.
\]

The following standard notation in empirical process theory will be used. Let \( X_1, \ldots, X_n \in \mathcal{X} \) be \( n \) i.i.d random variables with marginal distribution \( P \). Then for any measurable function \( f : \mathcal{X} \to \mathbb{R} \), we write

\[
P_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \quad \text{and} \quad P(f) = \mathbb{E}f(X) = \int f dP.
\]

Moreover, the Rademacher average of \( f \) is defined as

\[
R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i),
\]

where \( \varepsilon_1, \ldots, \varepsilon_n \) are i.i.d. Rademacher random variables such that \( P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2 \) for \( i = 1, \ldots, n \).

**Lemma 4.** Fix \( L > 0, \delta \in (0, 1) \). Let \( X_1, \ldots, X_n \) be \( n \) i.i.d random variables on \( \mathcal{X} \) with marginal distribution \( P \). Moreover, let \( \varphi : [-1,1] \to \mathbb{R} \) an \( L \)-Lipschitz function. Then, with probability at least \( 1 - \delta \), it holds

\[
\sup_{\lambda \in \Lambda} |(P_n - P)(\varphi \circ h_\lambda)| \leq \frac{4\sqrt{2}L}{\sqrt{n}} \sqrt{\log \left( \frac{2M}{\delta} \right)}.
\]

60
Proof. Define \( \bar{\varphi}(\cdot) \equiv \varphi(\cdot) - \varphi(0) \), so that \( \bar{\varphi} \) is an \( L \)-Lipschitz function that satisfies \( \bar{\varphi}(0) = 0 \). Moreover, for any \( \lambda \in \Lambda \), it holds

\[
(P_n - P)(\varphi \circ h_\lambda) = (P_n - P)(\bar{\varphi} \circ h_\lambda).
\]

Let \( \Phi : \mathbb{R} \to \mathbb{R}_+ \) be a given convex increasing function. Applying successively the symmetrization and the contraction inequalities [see, e.g., 50, Section 2], we find

\[
\mathbb{E}\Phi \left( \sup_{\lambda \in \Lambda} |(P_n - P)(\bar{\varphi} \circ h_\lambda)| \right) \leq \mathbb{E}\Phi \left( 2 \sup_{\lambda \in \Lambda} |R_n(\bar{\varphi} \circ h_\lambda)| \right) \leq \mathbb{E}\Phi \left( 4L \sup_{\lambda \in \Lambda} |R_n(h_\lambda)| \right).
\]

Observe now that \( \lambda \mapsto |R_n(h_\lambda)| \) is a convex function and Theorem 32.2 in [63] entails that

\[
\sup_{\lambda \in \Lambda} |R_n(h_\lambda)| = \max_{1 \leq j \leq M} |R_n(h_j)|.
\]

We now use a Chernoff bound to control this quantity. To that end, fix \( s, t > 0 \), and observe that

\[
\mathbb{P} \left( \sup_{\lambda \in \Lambda} |(P_n - P)(\varphi \circ h_\lambda)| > t \right) \leq \frac{1}{\Phi(st)} \mathbb{E}\Phi \left( s \sup_{\lambda \in \Lambda} |(P_n - P)(\bar{\varphi} \circ h_\lambda)| \right) \leq \frac{1}{\Phi(st)} \mathbb{E}\Phi \left( 4Ls \max_{1 \leq j \leq M} |R_n(h_j)| \right). \tag{2.6.1}
\]

Moreover, since \( \Phi \) is increasing,

\[
\mathbb{E}\Phi \left( 4Ls \max_{1 \leq j \leq M} |R_n(h_j)| \right) = \mathbb{E} \max_{1 \leq j \leq M} \Phi \left( 4Ls |R_n(h_j)| \right) \leq \sum_{j=1}^M \mathbb{E} \left[ \Phi \left( 4Ls R_n(h_j) \right) \vee \Phi \left( -4Ls R_n(h_j) \right) \right] \leq 2 \sum_{j=1}^M \mathbb{E}\Phi \left( 4Ls R_n(h_j) \right). \tag{2.6.2}
\]
Now choose $\Phi(\cdot) = \exp(\cdot)$, then
\[
\mathbb{E}\Phi(4LsR_n(h_j)) = \prod_{i=1}^{n} \mathbb{E}\cosh \left( \frac{4Ls h_j(X_i)}{n} \right) \leq \exp \left( \frac{8L^2 s^2}{n} \right),
\]
where $\cosh$ is the hyperbolic cosine function and where in the inequality, we used the fact that $|h_j(X_i)| \leq 1$ for any $i, j$ and $\cosh(x) \leq \exp(x^2/2)$. Together with (2.6.1) and (2.6.2), it yields
\[
\mathbb{P} \left( \sup_{\lambda \in \Lambda} |(P_n - P)(\varphi \circ h_\lambda)| > t \right) \leq 2M \inf_{s > 0} \exp \left( \frac{8L^2 s^2}{n} - st \right) \leq 2M \exp \left( -\frac{nt^2}{32L^2} \right).
\]
Choosing
\[
t = \frac{4\sqrt{2L}}{\sqrt{n}} \sqrt{\log \left( \frac{2M}{\delta} \right)},
\]
completes the proof of the Lemma. \hfill \Box

We now proceed to the proof of Theorem 2.4.1. Note first that from the properties of $\varphi$, $R^-(h) \leq R^-_{\varphi}(h)$. Next, we have for any data-dependent classifier $h \in \mathcal{H}_{\text{conv}}$ such that $\hat{R}^-_{\varphi}(h) \leq \alpha$:
\[
R^-(h) \leq \hat{R}^-_{\varphi}(h) + \sup_{h \in \mathcal{H}_{\text{conv}}} \left| \hat{R}^-_{\varphi}(h) - R^-(h) \right| \leq \alpha - \frac{\kappa}{\sqrt{n^-}} + \sup_{h \in \mathcal{H}_{\text{conv}}} \left| \hat{R}^-_{\varphi}(h) - R^-(h) \right|.
\]
Lemma 4 implies that, with probability $1 - \delta$
\[
\sup_{h \in \mathcal{H}_{\text{conv}}} \left| \hat{R}^-_{\varphi}(h) - R^-(h) \right| = \sup_{\lambda \in \Lambda} \left| (P^-_{n^-} - P^-)(\varphi \circ h_\lambda) \right| \leq \frac{\kappa}{\sqrt{n^-}}.
\]
The previous two displays imply that $R^-(h) \leq \alpha$ with probability $1 - \delta$, which completes the proof of Theorem 2.4.1.
2.6.2 Proof of Proposition 2

The proof of this proposition builds upon the following lemma.

Lemma 5. Let $\gamma(\alpha) = \inf_{h_\lambda \in \mathcal{H}^{\varphi, \alpha}} R_{\varphi}^+(h_\lambda)$, then $\gamma$ is a non-increasing convex function on $[0, 1]$.

**Proof.** First, it is clear that $\gamma$ is a non-increasing function of $\alpha$ because for $\alpha' > \alpha$, 
\[
\{h_\lambda \in \mathcal{H}^{\text{conv}} : R_{\varphi}^+(h_\lambda) \leq \alpha\} \subset \{h_\lambda \in \mathcal{H}^{\text{conv}} : R_{\varphi}^+(h_\lambda) \leq \alpha'\}.
\]

We now show that $\gamma$ is convex. To that end, observe first that since $\varphi$ is continuous on $[-1, 1]$, the set $\{\lambda \in \Lambda : h_\lambda \in \mathcal{H}^{\varphi, \alpha}\}$ is compact. Moreover, the function $\lambda \mapsto R_{\varphi}^+(h_\lambda)$ is convex. Therefore, there exists $\lambda^* \in \Lambda$ such that
\[
\gamma(\alpha) = \inf_{h_\lambda \in \mathcal{H}^{\varphi, \alpha}} R_{\varphi}^+(h_\lambda) = \min_{h_\lambda \in \mathcal{H}^{\varphi, \alpha}} R_{\varphi}^+(h_\lambda) = R_{\varphi}^+(h_{\lambda^*}).
\]

Now, fix $\alpha_1, \alpha_2 \in [0, 1]$. From the above considerations, there exist $\lambda_1, \lambda_2 \in \Lambda$ such that $\gamma(\alpha_1) = R_{\varphi}^+(h_{\lambda_1})$ and $\gamma(\alpha_2) = R_{\varphi}^+(h_{\lambda_2})$. For any $\theta \in (0, 1)$, define the convex combinations $\bar{\alpha}_\theta = \theta \alpha_1 + (1 - \theta) \alpha_2$ and $\bar{\lambda}_\theta = \theta \lambda_1 + (1 - \theta) \lambda_2$. Since $\lambda \mapsto R_{\varphi}^+(h_\lambda)$ is convex, it holds
\[
R_{\varphi}^-(h_{\lambda_{\theta}}) \leq \theta R_{\varphi}^-(h_{\lambda_1}) + (1 - \theta) R_{\varphi}^-(h_{\lambda_2}) \leq \theta \alpha_1 + (1 - \theta) \alpha_2 = \bar{\alpha}_\theta,
\]
so that $h_{\lambda_{\theta}} \in \mathcal{H}^{\varphi, \bar{\alpha}_\theta}$. Hence, $\gamma(\bar{\alpha}_\theta) \leq R_{\varphi}^+(h_{\lambda_{\theta}})$. Together with the convexity of $\varphi$, it yields
\[
\gamma(\theta \alpha_1 + (1 - \theta) \alpha_2) \leq R_{\varphi}^+(h_{\lambda_{\theta}}) \leq \theta R_{\varphi}^+(h_{\lambda_1}) + (1 - \theta) R_{\varphi}^+(h_{\lambda_2}) = \theta \gamma(\alpha_1) + (1 - \theta) \gamma(\alpha_2).
\]

$\square$
We now complete the proof of Proposition 2. For any \( x \in [0, 1] \), let \( \gamma(x) = \inf_{h \in H} R^x(h) \) and observe that the statement of the proposition is equivalent to

\[
\gamma(\alpha - \nu) - \gamma(\alpha) \leq \varphi(1) \frac{\nu}{\nu_0 - \nu}, \quad 0 < \nu < \nu_0.
\] (2.6.3)

Lemma 5 together with the assumption that \( H^{\varphi, \alpha - \nu_0} \neq \emptyset \) imply that \( \gamma \) is a non-increasing convex real-valued function on \( [\alpha - \nu_0, 1] \) so that

\[
\gamma(\alpha - \nu) - \gamma(\alpha) \leq \nu \sup_{g \in \partial \gamma(\alpha - \nu)} |g|,
\]

where \( \partial \gamma(\alpha - \nu) \) denotes the sub-differential of \( \gamma \) at \( \alpha - \nu \). Moreover, since \( \gamma \) is a non-increasing convex function on \( [\alpha - \nu_0, \alpha - \nu] \), it holds

\[
\gamma(\alpha - \nu_0) - \gamma(\alpha - \nu) \geq (\nu - \nu_0) \sup_{g \in \partial \gamma(\alpha - \nu)} |g|.
\]

The previous two displays yield

\[
\gamma(\alpha - \nu) - \gamma(\alpha) \leq \nu \frac{\gamma(\alpha - \nu_0) - \gamma(\alpha - \nu)}{\nu - \nu_0} \leq \nu \varphi(1) \frac{\nu_0 - \nu}{\nu - \nu_0}.
\]

2.6.3 Proof of Theorem 2.4.2

Define the events \( \mathcal{E}^- \) and \( \mathcal{E}^+ \) by

\[
\mathcal{E}^- = \bigcap_{h \in H^{\text{conv}}} \left\{ |\hat{R}_{\varphi}^- (h) - R_{\varphi}^- (h)| \leq \frac{\kappa}{\sqrt{n}} \right\},
\]

\[
\mathcal{E}^+ = \bigcap_{h \in H^{\text{conv}}} \left\{ |\hat{R}_{\varphi}^+(h) - R_{\varphi}^+(h)| \leq \frac{\kappa}{\sqrt{n}} \right\}.
\]

Lemma 4 implies

\[
\mathbb{P}(\mathcal{E}^-) \wedge \mathbb{P}(\mathcal{E}^+) \geq 1 - \delta.
\] (2.6.4)
Note first that Theorem 2.4.1 implies that (2.4.4) holds with probability $1 - \delta$. Observe now that the l.h.s of (2.4.5) can be decomposed as

$$R^\ast_\varphi(\tilde{h}^\kappa) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R^\ast_{\varphi}(h) = A_1 + A_1 + A_3,$$

where

$$A_1 = \left( R^\ast_\varphi(\tilde{h}^\kappa) - \hat{R}^\ast_\varphi(\tilde{h}^\kappa) \right) + \left( \hat{R}^\ast_\varphi(\tilde{h}^\kappa) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R^\ast_{\varphi}(h) \right)$$

$$A_2 = \min_{h \in \mathcal{H}^{\varphi, \alpha}} R^\ast_{\varphi}(h) - \min_{h \in \mathcal{H}^{\varphi, \alpha}_{2\kappa}} R^\ast_{\varphi}(h)$$

$$A_3 = \min_{h \in \mathcal{H}^{\varphi, \alpha}_{2\kappa}} R^\ast_{\varphi}(h) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R^\ast_{\varphi}(h).$$

To bound $A_1$ from above, observe that

$$A_1 \leq 2 \sup_{h \in \mathcal{H}^{\varphi, \alpha}_{\kappa}} |\hat{R}^\ast_{\varphi}(h) - R^\ast_{\varphi}(h)| \leq 2 \sup_{h \in \mathcal{H}^{\varphi, \alpha}_{\kappa}} |\hat{R}^\ast_{\varphi}(h) - R^\ast_{\varphi}(h)|.$$

Therefore, on the event $\mathcal{E}^+$ it holds

$$A_1 \leq \frac{2\kappa}{\sqrt{n^+}}.$$

We now treat $A_2$. Note that $A_2 \leq 0$ on the event $\mathcal{H}^{\varphi, \alpha_{2\kappa}} \subset \mathcal{H}^{\varphi, \alpha_{\kappa}}$. But this event contains $\mathcal{E}^-$ so that $A_2 \leq 0$ on the event $\mathcal{E}^-$. Finally, to control $A_3$, observe that under Assumption 1, Proposition 2 can be applied with $\nu = 2\kappa/\sqrt{n^-}$ and $\nu_0 = (1 - \bar{\varepsilon})\alpha$. Indeed, the assumptions of the theorem imply that $\nu \leq \nu_0/2$. It yields

$$A_3 \leq \frac{4\varphi(1)\kappa}{(1 - \bar{\varepsilon})\alpha \sqrt{n^-}}.$$
Combining the bounds on $A_1$, $A_2$ and $A_3$ obtained above, we find that (2.4.5) holds on the event $\mathcal{E}^- \cap \mathcal{E}^+$ that has probability at least $1 - 2\delta$ in view of (2.6.4).

The last statement of the theorem follows directly from the definition of $\kappa$.

2.6.4 Proof of Corollary 2.4.1

Now prove (2.4.8),

\[
\mathbb{P}(\mathcal{F}) = \sum_{n^- = 0}^{n} \mathbb{P}(\mathcal{F}|N^- = n^-) \mathbb{P}(N^- = n^-) \\
\geq \sum_{n^- = n_0}^{n} \mathbb{P}(\mathcal{F}|N^- = n^-) \mathbb{P}(N^- = n^-) \\
\geq (1 - 2\delta) \mathbb{P}(N^- \geq n_0),
\]

where in the last inequality, we used (2.4.6). Applying now Lemma 6, we obtain

\[
\mathbb{P}(N^- \geq n_0) \geq 1 - e^{-\frac{n(1-p)^2}{2}}.
\]

Therefore,

\[
\mathbb{P}(\mathcal{F}) \geq (1 - 2\delta)(1 - e^{-\frac{n(1-p)^2}{2}}),
\]

which completes the proof of (2.4.8).

The proof of (2.4.9) follows by observing that

\[
\left\{ R_{\phi}^*(\tilde{h}_n) - \min_{h \in \mathcal{H}_{\phi,n}} R_{\phi}^*(h) > \frac{4\sqrt{2}\varphi(1)\kappa}{(1 - \varepsilon)\alpha \sqrt{n(1-p)}} + \frac{2\sqrt{2}\kappa}{\sqrt{n p}} \right\} \subset \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 = (\mathcal{A}_1 \cap \mathcal{A}_2^c) \cup \mathcal{A}_2 \cup \mathcal{A}_3,
\]
where
\[ A_1 = \left\{ R^+_{\varphi}(\hat{h}_n^\kappa) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_{\varphi}(h) > \frac{4\varphi(1)\kappa}{(1 - \bar{\varepsilon})\alpha\sqrt{N}} + \frac{2\kappa}{\sqrt{N^+}} \right\} \subset \mathcal{F}^c, \]
\[ A_2 = \{ N^- < n(1 - p)/2 \}, \]
\[ A_3 = \{ N^+ < np/2 \}. \]

Since \( A_2 \subset \{ N^+ \geq n_0 \} \), we find
\[
\mathbb{P}(A_1 \cap A_2^c) \leq \sum_{n^- \geq n_0} \mathbb{P}(\mathcal{F}^c | N^- = n^-) \mathbb{P}(N^- = n^-) \leq 2\delta.
\]

Next, using Lemma 6, we get
\[
\mathbb{P}(A_2) \leq e^{-\frac{np(1-p)^2}{2}} \quad \text{and} \quad \mathbb{P}(A_3) \leq e^{-\frac{np^2}{2}}.
\]

Hence, we find
\[
\mathbb{P} \left\{ R^+_{\varphi}(\hat{h}_n^\kappa) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_{\varphi}(h) > \frac{4\sqrt{2}\varphi(1)\kappa}{(1 - \bar{\varepsilon})\alpha\sqrt{n(1 - p)}} + \frac{2\sqrt{2}\kappa}{\sqrt{np}} \right\} \leq 2\delta + e^{-\frac{np(1-p)^2}{2}} + e^{-\frac{np^2}{2}},
\]
which completes the proof of the corollary.

2.6.5 Proof of Theorem 2.4.3

First observe that for any \( \hat{h} \), \( R^+_{\varphi}(\hat{h}) \leq R^+_{\varphi}(\hat{\hat{h}}) \). Then the result follows from the claim that
\[
\min_{R^-(h) \leq \alpha} R^+_{\varphi}(h) = \inf_{R^-(h) \leq \alpha} R^+_{\varphi}(h).
\]

It is clear \( \min_{R^-(h) \leq \alpha} R^+_{\varphi}(h) \leq \inf_{R^-(h) \leq \alpha} R^+_{\varphi}(h) \), it remains to prove the other direction. By the Neyman-Pearson Lemma, We can decompose the feature space \( \mathcal{X} \) into a disjoint union of \( \mathcal{X}^+ \) and \( \mathcal{X}^- \), and the optimal (pseudo) classifier that solves \( \min_{R^-(h) \leq \alpha} R^+(h) \)
assigns label $+1$ for any $x \in \mathcal{X}^+$, and $-1$ for any $x \in \mathcal{X}^-$. Note that if any two classifiers $g_1$ and $g_2$ have the same signs, i.e., $\text{sgn}(g_1) = \text{sgn}(g_2)$, then $R^-(g_1) = R^-(g_2)$ and $R^+(g_1) = R^+(g_2)$. On the other hand, for $\varphi$-type I and II errors, values of classifiers do matter.

Let $\bar{h}_{B,\epsilon}(x) = B \cdot I(x \in \mathcal{X}^+) + (-\epsilon) \cdot I(x \in \mathcal{X}^-)$. Then clearly for any $B, \epsilon > 0$, $\bar{h}_{B,\epsilon}$ solves $\min_{R^-(h) \leq \alpha} R^+(h)$. Also, for any $B, \epsilon > 0$,

$$\inf_{R^-(h) \leq \alpha} R^+(h) \leq R^+(\bar{h}_{B,\epsilon}) = P^+(\mathcal{X}^+) \varphi(-B) + P^+(\mathcal{X}^-) \varphi(\epsilon).$$

Taking the limit, we have

$$\lim_{B \to \infty, \epsilon \to 0} R^+(\bar{h}_{B,\epsilon}) = \lim_{B \to \infty, \epsilon \to 0} P^+(\mathcal{X}^+) \varphi(-B) + P^+(\mathcal{X}^-) \varphi(\epsilon) = P^+(\mathcal{X}^-) = R^+(\bar{h}_{B,\epsilon}).$$

Therefore, $\inf_{R^-(h) \leq \alpha} R^+(h) \leq \min_{R^-(h) \leq \alpha} R^+(h)$, which completes the proof.

### 2.6.6 Proof of Proposition 3

Let the base classifiers be defined as

$$h_1(x) = -1 \quad \text{and} \quad h_2(x) = I(x \leq \alpha) - I(x > \alpha), \quad \forall x \in [0,1]$$

For any $\lambda \in [0,1]$, denote the convex combination of $h_1$ and $h_2$ by $h_\lambda = \lambda h_1 + (1 - \lambda) h_2$, i.e.,

$$h_\lambda(x) = (1 - 2\lambda) I(x \leq \alpha) - I(x > \alpha).$$

Suppose the conditional distributions of $X$ given $Y = 1$ or $Y = -1$, denoted respectively by $P^+$ and $P^-$, are both uniform on $[0,1]$. Recall that $R^-(h_\lambda) = P^-(h_\lambda(X) \geq 0)$ and $R^+(h_\lambda) = P^+(h_\lambda(X) < 0)$. Then, we have

$$R^-(h_\lambda) = P^-(h_\lambda(X) \geq 0) = \alpha I(\lambda \leq 1/2). \quad (2.6.5)$$
Therefore, for any $\tau \in [0, \alpha]$, we have

$$\{ \lambda \in [0, 1] : R^- (h_\lambda) \leq \tau \} = \begin{cases} [0, 1] & \text{if } \tau = \alpha, \\ (1/2, 1] & \text{if } \tau < \alpha. \end{cases}$$

Observe now that

$$R^+(h_\lambda) = P^+(h_\lambda(X) < 0) = (1 - \alpha) I(\lambda < 1/2) + I(\lambda \geq 1/2). \quad (2.6.6)$$

For any $\tau \in [0, \alpha]$, it yields

$$\inf_{\lambda \in [0, 1] : R^- (h_\lambda) \leq \tau} R^+(h_\lambda) = \begin{cases} 1 - \alpha & \text{if } \tau = \alpha, \\ 1 & \text{if } \tau < \alpha. \end{cases}$$

Consider now a classifier $\tilde{h}_\lambda$ such that $R^- (\tilde{h}_\lambda) \leq \tau$ for some $\tau < \alpha$. Then from (2.6.5), we see that must have $\lambda > 1/2$. Together with (2.6.6), this implies that $R^+(\tilde{h}_\lambda) = 1$. It yields

$$R^+(\tilde{h}_\lambda) - \min_{\lambda : R^- (h_\lambda) \leq \alpha} R^+(h_\lambda) = 1 - (1 - \alpha) = \alpha.$$

This completes the first part of the proposition. Moreover, in the same manner as (2.6.5), it can be easily proved that

$$\hat{R}^- (h_\lambda) = \frac{1}{n^+} \sum_{i=1}^{n^-} I(h_\lambda(X^-_i) \geq 0) = \alpha_{n^-} - \alpha \leq 1/2), \quad (2.6.7)$$

where

$$\alpha_{n^-} = \frac{1}{n^-} \sum_{i=1}^{n^-} I(\lambda \leq \alpha) \quad (2.6.8)$$
If a classifier $\hat{h}_\lambda$ is such that $\hat{R}^{-}(\hat{h}_\lambda) < \alpha_n$, then (2.6.7) implies that $\lambda > 1/2$. Using again (2.6.6), we find also that $R^+(\hat{h}_\lambda) = 1$. It yields

$$R^+(\hat{h}_\lambda) - \min_{\lambda : \hat{R}^{-}(\hat{h}_\lambda) \leq \alpha} R^+(h_\lambda) = 1 - (1 - \alpha) = \alpha.$$ 

It remains to show that $\hat{R}^{-}(\hat{h}_\lambda) < \alpha_n$ with positive probability for any classifier such that $\hat{R}^{-}(\hat{h}_\lambda) \leq \tau$ for some $\tau < \alpha$. Note that a sufficient condition for a classifier $\hat{h}_\lambda$ to satisfy this constraint is to have $\alpha \leq \alpha_n$. It is therefore sufficient to find a lower bound on the probability of the event $\mathcal{A} = \{\alpha_n \geq \alpha\}$. Such a lower bound is provided by Lemma 7, which guarantees that $\mathbb{P}(\mathcal{A}) \geq \alpha \land 1/4$.

2.6.7 Technical lemmas on Binomial distributions

The following lemmas are purely technical and arise from the fact that we observe binary data. They are used in two unrelated results.

**Lemma 6.** Let $N$ be a binomial random variables with parameters $n \geq 1$ and $q \in (0, 1)$. Then, for any $t > 0$ such that $t \leq nq/2$, it holds

$$\mathbb{P}(N \geq t) \geq 1 - e^{-\frac{nq^2}{2}}.$$ 

**Proof.** Note first that $n - N$ has binomial distribution with parameters $n \geq 1$ and $1 - q$. Therefore, we can write $n - N = \sum_{i=1}^{n} Z_i$ where $Z_i$ are i.i.d. Bernoulli random variables with parameter $1 - q$. Thus, using Hoeffding’s inequality, we find that for any $s \geq 0$,

$$\mathbb{P}(n - N - n(1 - q) \geq s) \leq e^{-\frac{2s^2}{n}}.$$ 

Applying the above inequality with $s = n - n(1 - q) - t \geq nq/2 \geq 0$ yields

$$\mathbb{P}(N \geq t) = \mathbb{P}(n - N - n(1 - q) \leq n - n(1 - q) - t) \geq 1 - e^{-\frac{nq^2}{4}}.$$ 

70
The next lemma provides a lower bound on the probability that a binomial distribution exceeds its expectation. Our result is uniform in the size of the binomial and it can be easily verified that it is sharp by considering sizes $n = 1$ and $n = 2$ and by looking at Figure 2.1. In particular, we do resort to Gaussian approximation which improves upon the lower bounds that can be derived from the inequalities presented in [71].

![Figure 2.1: Tail probabilities $\mathbb{P}(N \geq nq)$ where $N$ is a binomial random variable with parameters $n$ and $q$.](image-url)
Lemma 7. Let $N$ be a binomial random variable with parameters $n \geq 1$ and $0 < q \leq 1/2$. Then, it holds

$$\mathbb{P}(N \geq nq) \geq q \wedge (1/4).$$

Proof. We introduce the following local definition, which is limited to the scope of this proof. Fix $n \geq 1$ and for any $q \in (0, 1)$, let $P_q$ denote the distribution of a binomial random variable with parameters $n$ and $q$. Note first that if $n = 1$, the result is trivial since

$$P_q(N \geq q) = \mathbb{P}(Z \geq q) = \mathbb{P}(Z = 1) = q,$$

where $Z$ is a Bernoulli random variable with parameter $q$.

Assume that $n \geq 2$. Note that if $q \leq 1/n$, then $P_q(N \geq nq) \geq \mathbb{P}(Z = 1) = q$, where $Z$ is a Bernoulli random variable with parameter $q$. Moreover, for any any integer $k$ such that $k/n < q \leq (k+1)/n$, we have

$$P_q(N \geq nq) = P_q(N \geq k + 1) \geq P_{n/k}(N \geq k + 1).$$

The above inequality can be easily proved by taking the derivative over the interval $(k/n, (k+1)/n]$, of the function

$$q \mapsto \sum_{j=k+1}^{n} \binom{n}{j} q^j (1-q)^j.$$

We now show that

$$P_{n/k}(N \geq k + 1) \geq P_{n/(k+1)}(N \geq k), \quad 2 \leq k \leq n/2.$$

Let $U_1, \ldots, U_n$ be $n$ i.i.d. random variables uniformly distributed on the interval $[0, 1]$ and denote by $U_{(k)}$ the corresponding $k$th order statistic such that $U_{(1)} \leq \ldots \leq U_{(n)}$. 


Following [35, Section 7.2], it is not hard to show that

\[ P_k(n \geq k + 1) = \mathbb{P}(U_{k+1} \leq \frac{k}{n}) = n \binom{n-1}{k} \int_0^\frac{n}{k} t^k(1-t)^{n-k-1} dt , \]

and in the same manner,

\[ P_{k-1}(n \geq k) = \mathbb{P}(U_k \leq \frac{k-1}{n}) = n \binom{n-1}{k-1} \int_0^\frac{n}{k-1} t^{k-1}(1-t)^{n-k} dt . \]

Note that

\[ \binom{n-1}{k-1} = \binom{n-1}{k} \frac{k}{n-k} , \]

so that (2.6.10) follows if we prove

\[ k \int_0^\frac{k}{n} t^{k-1}(1-t)^{n-k} dt \leq (n-k) \int_0^\frac{k}{n} t^k(1-t)^{n-k-1} dt . \quad (2.6.11) \]

We can establish the following chain of equivalent inequalities.

\[ k \int_0^\frac{k}{n} t^{k-1}(1-t)^{n-k} dt \leq (n-k) \int_0^\frac{k}{n} t^k(1-t)^{n-k-1} dt \]
\[ \iff \int_0^\frac{k}{n} \frac{dt}{t} (1-t)^{n-k} dt \leq - \int_0^\frac{k}{n} t^{k-1}(1-t)^{n-k} dt + k \int_0^\frac{k}{n-1} t^{k-1}(1-t)^{n-k} dt \]
\[ \iff \int_0^\frac{k}{n} \frac{dt}{t} \left[ t^{k-1}(1-t)^{n-k} \right] dt \leq k \int_0^\frac{k}{n} t^{k-1}(1-t)^{n-k} dt \]
\[ \iff \left( \frac{k}{n} \right)^k \left( 1 - \frac{k}{n} \right)^{n-k} \leq k \int_0^\frac{k}{n-1} t^{k-1}(1-t)^{n-k} dt \]

We now study the variations of the function \( t \mapsto b(t) = t^{k-1}(1-t)^{n-k} \) on the interval \([ (k-1)/n, k/n ] \). Taking derivative, it is not hard to see that function \( b \) admits a unique local optimum, which is a maximum, at \( t_0 = \frac{k-1}{n-1} \) and that \( t_0 \in ((k-1)/n, k/n) \) because \( k \leq n \). Therefore, the function is increasing on \([ (k-1)/n, t_0 ] \) and decreasing on \([ t_0, k/n ] \).
It implies that
\[
\int_{\frac{k}{n}}^{\frac{k+1}{n}} b(t) \, dt \geq \frac{1}{n} \min \left[ b\left(\frac{k-1}{n}\right), b\left(\frac{k}{n}\right) \right].
\]

Hence, the proof of (2.6.11) follows from the following two observations:
\[
\left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} = \frac{k}{n} \left(\frac{k}{n}\right)^{k-1} \left(1 - \frac{k}{n}\right)^{n-k} = \frac{k}{n} b\left(\frac{k}{n}\right),
\]
and
\[
\left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} \leq \frac{k}{n} \left(\frac{k-1}{n}\right)^{k-1} \left(1 - \frac{k-1}{n}\right)^{n-k} = \frac{k}{n} b\left(\frac{k-1}{n}\right).
\]

While the first equality above is obvious, the second inequality can be obtained by an equivalent statement is
\[
\left(\frac{k}{n}\right)^{k-1} \left(\frac{n-k}{n}\right)^{n-k} \leq \left(\frac{k-1}{n}\right)^{k-1} \left(\frac{n-k+1}{n}\right)^{n-k}
\]
\[
\Leftrightarrow \left(\frac{k}{k-1}\right)^{k-1} \left(\frac{n-k}{n-k+1}\right)^{n-k} \leq 1
\]

Since the function \( t \mapsto \left(\frac{t+1}{t}\right)^t \) is increasing on \([0, \infty)\), and \( k \leq n-k+1 \), the result follows.

To conclude the proof of the Lemma, note that (2.6.9) and (2.6.10) imply that for any \( q > 1/n \),
\[
P_q(N \geq nq) \geq P_{\frac{1}{2}}(N \geq 2) = 1 - \left(\frac{n-1}{n}\right)^n - \left(\frac{n-1}{n}\right)^{n-1} \geq 1 - \left(\frac{1}{2}\right)^2 = 1 - \frac{1}{4} = \frac{3}{4},
\]
where, in the last inequality, we used the fact that the function
\[
t \mapsto 1 - \left(\frac{t-1}{t}\right)^t - \left(\frac{t-1}{t}\right)^{t-1}
\]
is increasing on \([1, \infty)\).
Chapter 3
Learning in Social Networks

3.1 Introduction

The effectiveness of information aggregation has been long and widely recognized as a central theme for good decision making at both individual and aggregate levels. Boosted by the Internet and online social networks, this theme is especially relevant in the modern world in communication and decision making, where people communicate with their friends in social networks before making specific decisions, through extremely efficient, open and multi-dimensional approaches. Recently, [1] provide a fascinating model to study communication in social networks and its implications for information aggregation. They define an intuitive concept of asymptotic learning, which means as the population of a network diverges, the probability that a large fraction of agents take “correct” actions converges to one or eventually exceeds a high threshold. Given agents communicate either truthfully or strategically, they establish equilibrium conditions under which asymptotic learning occurs. They also discuss the welfare implications of asymptotic learning, and investigate the impacts of specific types of cost structures and social cliques.
Motivated by the asymptotic learning concept, we ask the following questions. Can we define a good communication learning concept regarding a finite population network? If so, does such learning occur in a given finite social network? What are necessary and sufficient conditions to guarantee such learning? Can we write down clean and tractable rates at which a society achieves long run asymptotic learning? These questions are relevant and important, because it is common practice for people to assess the effectiveness of information aggregation of given communication approaches, in given organizations, regions or nations with relatively stationary population. Such assessment regarding finite population networks naturally offers a solid foundation for people to understand the quality of social learning when the society evolves. As of now, current researchers in social networks have not provided desirable answers to these questions, and it was admitted by economists that they need new inputs to address them ([43]; [4]; [49]).

Based on an information exchange game in social networks modified from [1], we propose a finite population learning concept, which captures the level of aggregation of disperse information in any given communication network. In the model, there is an underlying state. People in a social network do not know the underlying state, but they have a common prior on the distribution of the state. After receiving private signals correlated with the underlying state, they exchange information simultaneously in the network, at times specified by a homogenous Poisson process, until taking an irreversible action to exit the network. Upon each person’s exit, she makes a guess of the underlying state. People’s payoffs depend on the waiting time before guesses and the expected squared distance between their guesses and the underlying state. Finite population learning involves three parameters, \( \epsilon, \bar{\epsilon}, \) and \( \delta \) for a given social network \( G^n \) of population size \( n \); rigorously, it is called \( (\epsilon, \bar{\epsilon}, \delta) \)-learning. The parameter \( \epsilon \) is the precision under which an agent’s decision is considered “correct”, \( 1 - \bar{\epsilon} \) represents the fraction of agents in the network who
make the approximately correct decision, and $1 - \delta$ represents the probability at which such a fraction of agents make the approximately correct decision. We think of these three parameters as tolerance parameters of finite population learning. To contrast with asymptotically driven concepts, $(\epsilon, \bar{\epsilon}, \delta)$-learning is simply referred as finite population learning in verbal discussions.

We derive sufficient and necessary conditions for the occurrence of finite population learning under any given equilibrium. Intuitively, finite population learning is more likely to occur when the number of signals an agent obtains when she exits under equilibrium is larger, or the tolerances of learning are higher. Interestingly, the impact of the information precisions on finite population learning is ambiguous, which parallels the well-known Hirshleifer effect and subsequent work on the social value of information but stems from a new and different mechanism. We also provide sufficient and necessary conditions for the occurrence of finite population learning under any equilibrium, namely, without knowledge of particular equilibrium. Our conditions are easy to check, and they are helpful in the sense that they exhibit explicit interplays among parameters, such as tolerances, information precisions and information-sensitiveness of the decision problem.

A straightforward advantage of our tractable and transparent conditions for finite population learning over existing literature is that these conditions lead to meaningful comparative statics regarding the effectiveness of information aggregation in networks, as highlighted above. In these conditions, the underlying forces that shape the effectiveness of information aggregation in a given finite communication network are explicitly displayed in a single formula. Compared to the asymptotic learning results as known in previous literature, our conditions for finite population learning involves only one equilibrium outcome, which is the number of signals an agent obtains when she exits under equilibrium. More importantly, different from our finite population learning concept in
which the total amount of information is fixed, the existing asymptotic learning literature employs an implicit assumption that the total amount of information grows linearly with the population size. Hence, the learning status with respect to a sequence of networks with growing population reflects not only the effectiveness of information aggregation of certain network structures, but also an increased endowment of total information. Our finite population learning concept overcomes this defect and disentangles the effectiveness of information aggregation from the growth of information endowment.

The finite population learning concept enables us to investigate the rate at which a sequence of communication networks \( \{G^n\}_{n=1}^{\infty} \), which is referred as a society, reaches perfect learning. Perfect learning occurs if all communication networks in a society achieve finite population learning under vanishing tolerances as population grows. We say \( \delta \)-perfect learning occurs along society \( \{G^n\}_{n=1}^{\infty} \) if on the one hand \((\epsilon, \bar{\epsilon}, \delta_n)\)-learning occurs for each network \( G^n \) in the society, and on the other hand \( \delta_n \) goes to zero as \( n \) goes to infinity. The learning rate is characterized by the sequence \( \{\delta_n\}_{n=1}^{\infty} \). Faster learning rate means the sequence vanishes faster. Intuitively, faster learning rate implies perfect learning is reached at a higher quality. It is helpful to distinguish our learning rate concept from the speed of convergence to a pre-defined consensus in existing social learning literature, which mainly concerns about the time towards a consensus in a circumstance where people make repeated decisions and learn from others’ previous decisions to help make their own future decisions. In such a context, the observable sequence of aggregate decisions naturally reveals the dynamics of information aggregation along the time dimension. In our story of direct communication, however, although people communicate with each other repeatedly, they only make a single decision, and different people may go through varying communication rounds before their decisions. This makes the time dynamics of information aggregation largely unobservable, and calls for alternative dimensions to look
into the information dynamics. As will be analyzed in depth, our concept of learning rate gears towards the nature of direct communication, and helps investigate the speed at which the long run distribution of social learning is reached in different societies with different patterns of population growth.

We have derived conditions for societies to reach $\delta$-perfect learning at a certain desired rate $\{\delta_n\}_{n=1}^{\infty}$. Given a particular sequence of networks and their associated equilibria, we define an *equilibrium informed agent* as one who obtains an unbounded number of signals as the population goes to infinity. The $\delta$-perfect learning occurs if almost all agents in the society are equilibrium informed. Alternatively, without involving any equilibrium, we define a *socially informed agent* as one who has an unbounded number of neighbors in a finite distance as the population goes to infinity. The $\delta$-perfect learning occurs if almost all agents are socially informed. We also explicitly explore the achievable fastest learning rate for perfect learning in a given society. Under some circumstances, achievable learning rate could be in the order of $\frac{1}{\sqrt{n}} \exp(-n)$. This implies that a society with growing population might achieve a desirable level of finite population learning very quickly. Our new learning results lead to sharper social and economic implications and offer new insights on communication the associated information exchange in social networks.

**Relation to Literature.** Our work lies in the category of Bayesian social learning in social networks, in which decision makers in a social network update their information according to the Bayes’ rule. General Bayesian social learning is divided into two subcategories, namely Bayesian observational learning and Bayesian communication learning. In Bayesian observational learning, agents observe past actions of their neighbors. From these observed actions, agents update their beliefs and make inferences. Herd behavior is a very typical consequence of observational learning. In literature, [8], [13] and [72] are early
attempts to model herd effects through Bayesian observational learning. \cite{9} and Smith and \cite{73} relax the assumption of full observation network topology and study Bayesian observational learning with sampling of past actions. Recently, \cite{3} and \cite{58} investigate how detailed network structures could add new interesting insights.

Our work belongs to Bayesian communication learning, which means that agents cannot directly observe actions of others but can communicate with each other before making a decision. Consequently, agents update their beliefs and make inferences based on the information given by others. New interesting considerations arise in Bayesian communication learning, for example, agents may not want to truthfully reveal their information to others through communication. \cite{23} pioneers the research in strategic communication, and \cite{1} is an interesting piece that look into how communication learning shapes information aggregation in social networks. Other research like \cite{36} and \cite{45} also look into strategic communication in social networks, but their focus is not on information aggregation.

Also in literature, there is a branch that applies various non-Bayesian updating methods to investigate information aggregation and social learning. \cite{24} develops a tractable non-Bayesian learning model which are frequently employed in research of social networks today. Essentially, the De-Groot model is pertaining to observational learning, in which agents make today’s decisions by taking the average of neighbors’ beliefs revealed in their decisions yesterday. \cite{25} and \cite{38}, \cite{40}, \cite{39}, \cite{37} apply the De-Groot model to financial networks and general social networks, respectively. By a field experiment, \cite{57} compare a non-Baysian model of communication with a model in which agents communicate their signals and update information based on Bayes’ rule. Their evidence are generally in favor of the Bayesian communication learning approach. We do not discuss non-Bayesian updating in our paper.
Our paper is most closely related to [1]. Compared to their work, we employ a simplified framework for network communication and exploit more undeveloped mechanisms. In particular, we mainly focus on the effect of social learning and information aggregation in finite population communication networks that allows for clear comparative statics with respect to learning, and discuss the rates of learning as the population increases. As of now, researchers have not provided desirable results in finite population communication network as well as results regarding learning rates. To the best of our knowledge our work is the first attempt to address these questions with clear answers.

Our work is also related to [40], [39] and [37], in particular regarding the investigation of learning rate. [40], [39] and [37] employ the De-Groot model to analyze the impacts of homophily in social networks, which refers to the tendency of agents to associate relatively more with those who are similar to them, on the learning rate in the context of observational learning. Our results of learning rate are different from theirs in two aspects. First, our focus is on Bayesian communication learning rather than non-Bayesian observational learning. Second, as discussed before, our concept of learning rate is based on perfect learning as the population in networks diverges, rather than the time towards a consensus in their model. An appealing feature of [40], [39] and [37] is that their results of learning rate is based on certain statistics of networks rather than the full network structures, which could lead to potentially more empirical traction.

The rest of the paper is organized as follows. Section 2 introduces the information exchange game and characterizes its equilibrium. Our new finite population learning concept is proposed in Section 3. Section 4 discusses the dynamics of learning and addresses learning rates explicitly. Section 5 briefly introduces a conservative learning status in mul-
multiple equilibria. In the final section, we discuss possible directions for further research. All proofs are attached in the appendix.

3.2 Model

In this section, we present our model of information exchange in social networks, which is closely related to [1] but has different focus. We make the following assumptions to simplify the analysis and focus on finite population learning, which will be rigorously defined in the next section. First, we assume mandatory communication, which means that no agent holds her information to herself. When communication times arrive, an agent has to send her information set to all of her direct neighbors. Second, we assume truthful communication, which means whenever an agent sends information, she has to send unmanipulated information, whether it is her own private information or obtained information originated from other agents. Literally, these two assumptions are equivalent to assumption of truthful communication in [1], but as our focus in this paper is not to uncover similarity and differences between truthful and strategic communication, we do not plan to relax this assumption and make the model more general. We will first analyze communication and information aggregation in one given finite population network, and then consider the limit and the path as the population grows to infinity. In this course, we assume that existing links are kept when a network grows.

Before we formally introduce the game, we would like to illustrate how information flows with an example. For simplicity, suppose there are four agents in a network. At time $t = 0$, each agent $i$ receives some private signal $s_i$. So the total information endowment in the system is $\{s_1, s_2, s_3, s_4\}$. Communication occurs at $t = 1, 2$. Due to the structure of the graph in our example, there is no need to consider beyond the second communication round.
We will study two cases, and focus on agent 1’s information set. In the first case, no agent exits after time $t = 0$. So the information flow is as follows:

After the first round of communication, i.e., $t = 1$, agent 1 has signals $\{s_1, s_2, s_3\}$. Also note that at this time agent 3 has $\{s_3, s_4\}$. At $t = 2$, agent 3 sends the newly grabbed signal $s_4$ to agent 1. So agent 1’s information set enriches to $\{s_1, s_2, s_3, s_4\}$.

In the second case, suppose agent 3 exits after time $t = 0$, then she is still obliged to send her own signal to neighbors, but she does not have any incentive to receive others’ signal. Therefore, the information follow is as follows.
Note that as agent 3 does not receive signal from agent 4, she does not have any new information to send to agent 1 at the second communication round. Therefore, agent 1’s information set is still \( \{s_1, s_2, s_3\} \). By contrasting the two cases in this toy example, we see that agents’ decision will affect the information flow in the network.

Now we formally introduce the information exchange game. Suppose we are interested in a social network with agents \( N^n = \{1, 2, ..., n\} \). To model communication in the network, we organize these agents in a directed graph \( G^n = (N^n, \mathcal{E}^n) \), in which each node \( i \in N^n \) represents an agent. We allow directed graphs to have multi-arcs, so that two agents can communicate to each other. An ordered pair \((j, i) \in \mathcal{E}^n\) means agent \( j \) can send information to agent \( i \) directly. The goal of every agent is to estimate \( \theta \in \mathbb{R} \), which represents an underlying state of the world. Agents’ pre-knowledge of \( \theta \) is captured by a common prior \( \theta \sim N(0, 1/\rho) \). At time \( t = 0 \), agent \( i \) receives her private signal \( s_i = \theta + z_i \). All \( z_i \sim N(0, 1/\bar{\rho}) \) are independent and they are also independent of \( \theta \). The distributions of \( z_i \)'s are common knowledge. Our results are not affected if the means of \( \theta \) and \( z_i \) are changed to non-zero values.

In this network, agents exchange their information as follows. Suppose agents live in a world with continuous time \( t \in [0, \infty) \). Waiting induces a common exponential
discount with rate \( r > 0 \). Instead of communicating at fixed times, all agents communicate simultaneously at some points in time that follow a homogeneous Poisson process with rate \( \lambda > 0 \), which is independent of \( \theta \) and \( z_i \). This Poisson clock is also common knowledge. After communication, agents update beliefs according to the Bayes' rule. For example, the posterior distribution of \( \theta \) on \( k \) distinct signals is Gaussian with precision \( \rho + k\bar{\rho} \). So more private information, i.e., a higher \( k \), will increase the precision and lead to a better guess. Hence, there is a natural trade-off between getting more information and acting earlier, which makes an optimal stopping problem for each agent \( i \). We call the incentive to get more information \emph{information effect}, and the incentive to act earlier \emph{discount effect}. In this course, at any given time \( t \), each agent \( i \) either makes a guess \( x_i \) of the fundamental \( \theta \), or “wait” for more information. Just as illustrated in the four agents’ example, we assume that after agents make guesses and exit, they do not receive new information, but they continue to transmit information that they have already obtained when new rounds of communication take place.

We introduce a few more notations to facilitate the discussion. \( B^n_{i,s} \) denotes the \( s \)-step neighborhood of agent \( i \), which is the set of agents who can send their private signals to agent \( i \) in at most \( s \) rounds of communication, provided that no agent exits. Let \( I^n_{i,t} \) denote the information set of agent \( i \) at time \( t \). We next specify the payoff structure and the optimization problem faced by agents. Suppose agent \( i \) takes action \( x_i \) at time \( t \) when the realization of the underlying state is \( \theta \), then her instantaneous payoff of taking an action \( x_i \) is

\[
u^n_i(x_i) = \psi - (x_i - \theta)^2,
\]

where \( \psi \) is a real-valued constant that captures the information sensitiveness of the decision problem, which we will elaborate later. It is worth noting that in [1], the constant
\(\psi\) in agents’ instantaneous payoff function is implicitly assumed to satisfy a constraint \(\psi(\rho + \bar{\rho}) > 1\). We relax this assumption and allow \(\psi\) to take any real value.

At time \(t\) with information set \(I_{i,t}^n\), agent \(i\)’s optimal expected instantaneous payoff of taking an action is

\[
U_{i,t}^n(I_{i,t}^n) = \max_{x_i} \mathbb{E}(u_i^n(x_i)|I_{i,t}^n).
\]

It is easy to see that agent \(i\)’s optimal guess is \(x_{i,t}^{n,*} = \mathbb{E}[\theta|I_{i,t}^n]\) if she decides to act at time \(t\). Thanks to the normality assumption of the fundamental and signals, the optimal expected instantaneous payoff of agent \(i\) taking an action after observing \(k\) distinct signals can be calculated explicitly:

\[
\mathbb{E}[\psi - (x_{i,t}^{n,*} - \theta)^2|I_{i,t}^n] = \psi - \frac{1}{\rho + \bar{\rho}k}.
\]

At any time \(t\) with information set \(I_{i,t}^n\), before trying to make a best guess and exit, agent \(i\) has to make a decision about whether to exit. Note that agents benefit from waiting only if they get more signals. Due to discount in time, each agent should make a guess and exit precisely when communications take place. Moreover, each agent can only get finite number of signals even if they decide to wait forever, because there are \(n\) number of signals in total in the network. Therefore, we actually just need to consider strategy profiles in which every agent exits at a finite communication round.

Denote by \(l^n = (l_1^n, \ldots, l_n^n)\), where each \(l_i^n\) is agent \(i\)’s communication round before exit. Let \(\tau_i^n\) be the physical time until \(l_i^n\) rounds of communication. Agent \(i\)’s optimal stopping criterion is

\[
U_i^n(l_i^n, l_{-i}^n) = \mathbb{E} \left\{ e^{-r\tau_i^n} \max_{x_i} \mathbb{E}[\psi - (x_i - \theta)^2|I_i^n(l^n)] \right\},
\]
where \( I^n_i(l^n) \) is agent \( i \)'s information set upon exit if every agent chooses exit rounds according to \( l^n \). Given \( G^n \) and \( l^n \), it can be shown that

\[
U^n_i(l^n_i, l^n_{-i}) = \left( \frac{\lambda}{\lambda + r} \right)^{l^n_i} \left( \psi - \frac{1}{\rho + \bar{\rho}k^n_{i,l^n}} \right),
\]

where \( k^n_{i,l^n} \) is the number of signals agent \( i \) get upon exit if every agent acts according to \( l^n \).

The above analysis allows us to consider a much simpler game than what is studied in [1] with no sacrifice in economic and social intuitions. With this reduction, we can just consider the following strategy game.

**Definition 1.** Given \( G^n \), the game \( \Gamma_{info}(G^n) \) is a triple \( \{N^n, L^n, U^n\} \), in which

(a) \( N^n \) is the set of agents, i.e., \( N^n = \{1, 2, \ldots, n\} \);

(b) \( L^n \) is the collection of agents’ strategy spaces. For any agent \( i \in N^n \), her strategy space \( L^n_i \in L^n \) is a finite set

\[
L^n_i = \{0, 1, 2, \ldots, (L^n_i)_{max}\},
\]

where

\[
(L^n_i)_{max} = \max_{j \in G^n} dist^n(j,i);
\]

(c) \( U^n_i \in U^n \) is the payoff function for agent \( i \):

\[
U^n_i(l^n_i, l^n_{-i}) = \left( \frac{\lambda}{\lambda + r} \right)^{l^n_i} \left( \psi - \frac{1}{\rho + \bar{\rho}k^n_{i,l^n}} \right).
\]

Nash Equilibria (NE) of \( \Gamma(G^n) \): \( \Sigma^{n,*} = \{\sigma^{n,*}\} \)

Now we provide an example of the network games. On the 4 agent graph displayed previously, suppose \( \lambda = r, \psi = 1/2 \) and \( \rho = \bar{\rho} = 1/2 \). The decision problem for agents 2 and 4 are degenerate. They should exit right away because they won’t get any new
signals due to graph structure. The payoff matrix for agent 1 (row) and 3 (column) is as follows:

<table>
<thead>
<tr>
<th></th>
<th>0 Step</th>
<th>1 Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Step</td>
<td>0,0</td>
<td>0,1/12</td>
</tr>
<tr>
<td>1 Step</td>
<td>1/8,0</td>
<td>1/8,1/12</td>
</tr>
<tr>
<td>2 Step</td>
<td>1/16,0</td>
<td>3/40,1/12</td>
</tr>
</tbody>
</table>

Therefore, there is one NE of the game. Namely, agents 2 and 4 exit immediately after they receive their private signals, while agent 1 exits after the second communication round and agent 3 exits after the first communication round.

Our model of information exchange in social networks differs from [1] in various aspects. First, we assume mandatory and truthful communication throughout the analysis, which simplifies the framework and allows us to focus on finite population learning. We leave strategic information exchange to future works. Second, we work with a much simpler game without sacrifice in social and economic intuitions. Third, we relax the restrictions on the parameter $\psi$ in the payoff function. This is important because it captures the information sensitiveness of the decision problem. Interestingly, due to the nature of the information exchange game, the information sensitiveness of the decision problem is not monotone in $\psi$. When $\psi$ takes negative or very small positive value, any agent $i$ would like to wait forever regardless of the information she could potentially obtain from the communication network, in which case the decision problem is information irrelevant. When $\psi$ is positive and not too small, information is relevant. Specifically, when $\psi$ is moderate, information effect dominates, and thus the decision problem is more information sensitive; while when $\psi$ is large, the discount effect dominates, and thus the decision problem is less information sensitive. We therefore hold that discussion of information aggregation in networks should be broken down to information sensitive or less informa-
tion sensitive cases. In the analysis below, we will see that information sensitiveness of the decision problem plays a role in shaping the information exchange process and the level of information aggregation.

### 3.3 Finite Population Learning

In this section, we measure the level of information aggregation in any given communication network. Related recent research on learning in social networks focuses on asymptotic learning, which means that as the fraction of agents taking the correct action converging to one (or a small left neighborhood of one) as the population of the social network grows large ([1]; [3]). However, as discussed in [4], people are also very interested in the information dynamics away from long run limit. For example, does a well-defined learning status occur in a given social network of a certain size? What are appropriate conditions for such learning? We look into these questions in this section and provide corresponding results.

In pursuing our goal, we introduce a new concept of learning in social networks. We will argue that this is an appropriate concept to capture the information dynamics away from long run limit.

**Definition 2.** Given a social network $G^n$, the information exchange game $Γ_{info}(G^n)$ and an equilibrium profile $σ^{n,*}$, for a triple $(ε, 1 − \bar{ε}, δ)$, we say $G^n$ achieves $(ε, 1 − \bar{ε}, δ)$-learning under $σ^{n,*}$ if

$$P_{σ^{n,∗}} \left( \frac{1}{n} \sum_{i=1}^{n} (1 - M_i^{n,ε}) \geq \bar{ε} \right) \leq δ,$$

where $M_i^{n,ε} = 1(|x_i - θ| \leq ε)$, $x_i$ is agent $i$’s action upon exit, and $P_{σ^{n,∗}}$ denotes the conditional probability given $σ^{n,*}$.

In this definition, the parameter $ε$ tunes what is the approximately correct decision for individual agents, $1 − \bar{ε}$ controls the fraction of agents who make the approximately
correct decision, and $1 - \delta$ represents the probability at which such a fraction of agents make the approximately correct decision. Different from [1], we highlight the difference between $\varepsilon$ and $\bar{\varepsilon}$, because these two parameters capture different tolerances. Concretely, $\varepsilon$ is at the individual level while $\bar{\varepsilon}$ is at the aggregate level. These two tolerances are not necessarily equal, and we will highlight the importance of distinguishing them below.

An important difference between our definition of $(\varepsilon, \bar{\varepsilon}, \delta)$-learning and the $(\varepsilon, \delta)$-asymptotic learning in [1] is that, while their learning concept is defined on a society $\{G^n\}_{n=1}^{\infty}$ that is an infinite sequence of communication networks, ours is defined on one social network $G^n$ with finite population $n$. This single network based definition allows us to study whether learning occurs in a given social network, besides exploring the long run limit behavior along the society. To capture this essence, we call the new learning concept finite population learning in verbal discussions.

A natural question to ask is whether such finite population learning occurs in a given communication network. If so, under what conditions? The following proposition provides a necessary condition and a sufficient condition for $(\varepsilon, \bar{\varepsilon}, \delta)$-learning in a given social network under any equilibrium profile. When there is no confusion, we refer to the information exchange game $\Gamma_{info}(G^n)$ simply as $G^n$. Denote by $erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$ the error function of the normal distribution.

**Proposition 4.** For a given social network $G^n$ under any equilibrium $\sigma^*(=\sigma^{n,*})$,

(a) $(\varepsilon, \bar{\varepsilon}, \delta)$-learning does not occur if

$$
\frac{1}{n} \sum_{i=1}^{n} erf \left( \sqrt{\frac{\rho + \overline{\rho} k_{i}^{n,\sigma^*}}{2}} \right) < (1 - \bar{\varepsilon})(1 - \delta).
$$

(3.3.1)
(b) \((\varepsilon, \bar{\varepsilon}, \delta)\)-learning occurs if

\[
\frac{1}{n} \sum_{i=1}^{n} \text{erf} \left( \varepsilon \sqrt{\frac{\rho + \hat{\rho} k_{i_n,\sigma}^*}{2}} \right) \geq 1 - \bar{\varepsilon} \delta. \tag{3.3.2}
\]

This proposition provides clear conditions for the occurrence of finite population learning. Our conditions are more operative and transparent than their asymptotic counterparts in [1]. Recall that in the Proposition 1 of [1], two equilibrium-specific variables, the \(k\)-radius set \(V_{k,\sigma}^n\) and an extra variable \(k\), are needed to characterize the conditions, while in our conditions we characterize the finite population learning with only one equilibrium outcome \(k_{i_n,\sigma}^*\), which is the number of signals obtained by agent \(i\) when she takes an action under the equilibrium profile \(\sigma^{n,*}\). To check conditions in [1], one need to first find a \(k\) that satisfies the condition for the error function, and then check if the set \(V_{k,\sigma}^n\) satisfies the limit condition under equilibrium. In this process, the final determination of \(k\) and its relation to the equilibrium are unclear. In contrast, our conditions only require one equilibrium outcome \(k_{i_n,\sigma}^*\), and the set \(\{k_{i_n,\sigma}^*\}_{i=1}^{n}\) is directly induced by an equilibrium \(\sigma^{n,*}\) in a communication network \(G^n\). Hence, our conditions not only allow us to investigate the effect of learning in a given communication network with finite population, but also offer a more interpretable link between the communication equilibrium and its corresponding learning result.

Our necessary and sufficient conditions also allow us to untangle the interplay among parameters involved in learning on a given communication network. For example, we are able to answer the following question. Given the tolerances \(\varepsilon, \bar{\varepsilon}, \delta\) and the information precisions \(\rho\) and \(\hat{\rho}\), how does the change of \(k_{i_n,\sigma}^*\) affect the occurrence of finite population learning in a given social network \(G^n\)? When \(k_{i_n,\sigma}^*\)'s are sufficiently small to validate condition (3.3.1), finite population learning does not occur. Similarly, when some of
\(k^{n,\sigma^\ast}_i\)'s are sufficiently large so that the condition \((3.3.2)\) is satisfied, finite population learning occurs. In particular, Similar marginal interpretations also apply to parameters \(\varepsilon, \bar{\varepsilon}, \delta, \rho\) and \(\bar{\rho}\). Generally, finite population learning in a given social network \(G^n\) is more likely to occur when the equilibrium induces larger numbers of signals obtained by agents, for a fixed set of tolerances and information precisions; it is also more likely to occur when the tolerances and the information precisions are larger. As interplays among the parameters \(\varepsilon, \bar{\varepsilon}, \delta, \rho, \bar{\rho}\) and \(k^{n,\sigma^\ast}_i\) are clear through \((3.3.1)\) and \((3.3.2)\), the two conditions provide various comparative statics that help us better understand learning in different social circumstances. Importantly, the total amount of information is fixed in any finite population network, which ensures that these comparative statics indeed disentangle the effectiveness of information aggregation from the endowment of information, so that the net effect of information aggregation is transparent.

It is interesting to note that \((1 - \bar{\varepsilon})(1 - \delta) < 1 - \bar{\varepsilon}\delta\) for any \(0 < \bar{\varepsilon}, \delta < 1\). This gap indicates that failure of condition \((3.3.1)\) does not necessarily lead to condition \((3.3.2)\), and vice versa. Therefore, there exist circumstances under which the occurrence of finite population learning is undetermined in view of conditions \((3.3.1)\) and \((3.3.2)\). Two perspectives help understand this gap. First, we use Markov’s inequality to get tractable forms of the necessary and the sufficient conditions. Sharper inequalities may lead to weaker conditions and thus probably fill part of the gap, but they are likely to make these conditions intractable and less transparent. Secondly and more importantly, as we discussed above, our finite population learning conditions involve equilibrium outcome \(k^{n,\sigma^\ast}_i\)'s in a clean and simple formula. The cost for enjoying this clarity is that we did not fully utilize the set of equilibrium outcomes \(\{k^{n,\sigma^\ast}_i\}_{i=1}^n\).

Also, a beauty of symmetry arises in our necessary and sufficient conditions for finite population learning. The parameters \(\bar{\varepsilon}\) and \(\delta\) are completely interchangeable in these con-
ditions, which was not expected at the first place as they captures tolerances in different categories. On the other hand, in our two conditions, parameter $\varepsilon$ stands in a position that is unchangeable with $\overline{\varepsilon}$ and $\delta$, which hints that $\varepsilon$ and $\overline{\varepsilon}$ plays different roles in finite population learning.

Conditions (3.3.1) and (3.3.2) have powerful implications. The next corollary establishes a necessary condition and a sufficient condition for finite population learning in any given communication network and under any equilibrium. The proof is straightforward so that we omit it, but the result is non-trivial.

**Corollary 1.** For any social network $G^n$ and any equilibrium $\sigma^{n,*}$,

(a) $(\varepsilon, \overline{\varepsilon}, \delta)$-learning does not occur if

$$erf \left( \varepsilon \sqrt{\frac{\rho + \overline{\rho} n}{2}} \right) < (1 - \overline{\varepsilon})(1 - \delta). \quad (3.3.3)$$

(b) $(\varepsilon, \overline{\varepsilon}, \delta)$-learning occurs if

$$erf \left( \varepsilon \sqrt{\frac{\rho + \overline{\rho} n}{2}} \right) \geq 1 - \overline{\varepsilon} \delta. \quad (3.3.4)$$

For any social network $G^n$ and equilibrium $\sigma^{n,*}$, each agent can get at most $n$ signals before she takes an action, in other words, $k_i^{n,\sigma^*} \leq n$ for every agent $i$. Hence, the left hand side of condition (3.3.3) is an upper bound of the left hand side of condition (3.3.1). This implies that when the condition (3.3.3) is satisfied, the condition (3.3.1) is also satisfied for any $G^n$ and $\sigma^{n,*}$, and thus $(\varepsilon, \overline{\varepsilon}, \delta)$-learning does not occur whatever $G^n$ and $\sigma^{n,*}$ are. Similarly, because every agent at least has her own private signal, $k_i^{n,\sigma^*} \geq 1$ for all $i$, the left hand side of (3.3.4) is a lower bound for the left hand side of (3.3.2). Hence, when the condition (3.3.4) is satisfied, the condition (3.3.2) must be satisfied too for any $G^n$ and $\sigma^{n,*}$, and thus finite population learning occurs whatever $G^n$ and $\sigma^{n,*}$ are.
Corollary 1 is interesting because it tells that under some circumstances, we can determine the occurrence of finite population learning without knowing either the structure of the social network or the equilibrium. Intuitively, if any one parameter of the tolerances, information precisions or population size is too low, such that the condition (3.3.3) is satisfied, we may conclude that finite population learning does not occur no matter how effective the communication network is organized. Conversely, if any one of the tolerances or information precisions is sufficiently large such that condition (3.3.4) holds, we know that finite population learning surely occurs even if all agents are isolated. In this sense, Corollary 1 extends the perspective in the Proposition 2 in [1] that we can determine learning without knowledge of particular equilibrium, to finite population learning.

3.4 Perfect Learning and the Rates

Based on the analysis of finite population learning, we consider the effect of information aggregation and learning as population in communication networks grows. Our approach to address the limiting behavior of learning is different from asymptotic learning in existing literature ([1]; [3]). In particular, we highlight finite population learning as the foundation of asymptotic learning. Consequently, we are able to check learning status all along the path to the limit, and the probabilistic tolerance parameters naturally induce learning rates. This concept of learning rate is different from what is employed in [40], [39], [37] that focuses on the time dimension rather than population. As suggested by [4], even if asymptotic learning occurs in the long run, people are also interested in investigating the rates at which the long run distribution is reached in different societies with different patterns of population growth. Our results on learning rates are of the same spirit.
3.4.1 Perfect Learning

Similar to asymptotic learning in [1], we consider a sequence of communication networks \( \{G^n\}_{n=1}^{\infty} \), which is called a society. However, different from asymptotic learning in existing literature, we are interested in whether social networks in a society achieve finite population learning under diminishing tolerances as population grows. Recall that we have three tolerance parameters \( \varepsilon, \bar{\varepsilon}, \delta \) in the \( (\varepsilon, \bar{\varepsilon}, \delta) \)-learning definition. To inquire the limiting behavior in a society, we can focus on one parameter at a time, keeping the other two fixed. In this section, we allow \( \delta \) depend on the population size \( n \). Asymptotic results concerning \( \varepsilon \) and \( \bar{\varepsilon} \) are interesting for further investigation. The following definition introduces \( \delta \)-perfect learning on a given society \( \{G^n\}_{n=1}^{\infty} \).

**Definition 3.** We say \( \delta \)-perfect learning occurs in society \( \{G^n\}_{n=1}^{\infty} \) under equilibria \( \{\sigma^{n,*}\}_{n=1}^{\infty} \) if there exists a vanishing positive sequence \( \{\delta_n\}_{n=1}^{\infty} \) such that \( (\varepsilon, \bar{\varepsilon}, \delta_n) \)-learning occurs in \( G^n \) under its associated \( \sigma^{n,*} \) for all \( n \).

Compared to the perfect asymptotic learning concept in [1], our definition of perfect learning is both stronger and more general for the following reasons. First, we require the networks in the society to achieve learning not only in the limit but also all along the path towards the limit. Second, by focusing on different parameters \( \varepsilon, \bar{\varepsilon} \) and \( \delta \), we could address three different kinds of asymptotic learning. As discussed in previous section, these three parameters exhibit different impacts on finite population learning, so that they can play different roles in perfect learning. But for the rest of this section, we will focus on \( \delta \)-perfect learning. Third and most importantly, this definition allows us to investigate learning rates, which is the focus of the next subsection.

The analysis of perfect learning relies heavily on the properties of \( (\varepsilon, \bar{\varepsilon}, \delta) \)-learning uncovered in the previous section. Previously, we fixed \( \delta \) and derived sufficient conditions for \( (\varepsilon, \bar{\varepsilon}, \delta) \)-learning in a single network. Now we let \( \delta \) vary, and would like to find conditions.
under which there exits a vanishing positive sequence \( \{\delta_n\}_{n=1}^{\infty} \), such that \((\varepsilon, \bar{\varepsilon}, \delta_n)\)-learning occurs in \( G^n \) under its associated equilibrium \( \sigma^{n,*} \) for all \( n \).

In the following, we will derive two sufficient conditions for \( \delta \)-perfect learning. The first condition, stated as Proposition 5, relies on the equilibrium outcome \( k_i^{n,\sigma^*} \). The second condition, stated as Proposition 6, relies only on the formation of the society.

To deliver the first sufficient condition, we first define an \textit{equilibrium informed agent} in a society, based on the central quantity \( k_i^{n,\sigma^*} \).

**Definition 4 (Equilibrium Informed Agent).** For agent \( i \) in a given society \( \{G^n\}_{n=1}^{\infty} \), she is equilibrium informed with respect to \( \{G^n\}_{n=1}^{\infty} \) under equilibria \( \{\sigma^{n,*}\}_{n=1}^{\infty} \) if

\[
\lim_{n \to \infty} k_i^{n,\sigma^*} = \infty.
\]

Moreover, we denote by \( EI^{n,*} \) the set of equilibrium informed agents in the network \( G^n \) under equilibrium \( \sigma^{n,*} \).

The quantity \( k_i^{n,\sigma^*} \) for an equilibrium informed agent is not bounded from above along the society. In economic terms, an agent has equilibrium informed status means that she enjoys increasing information advantage as population grows.

The next proposition offers a sufficient condition for \( \delta \)-perfect learning. In a similar spirit, we have a more general sufficient condition, Lemma 9, in the Appendix. We omit the proof of Proposition 5 as it can be viewed as a corollary to Lemma 9.

**Proposition 5.** The \( \delta \)-perfect learning occurs in a society \( \{G^n\}_{n=1}^{\infty} \) under equilibria \( \{\sigma^{n,*}\}_{n=1}^{\infty} \) if

\[
\lim_{n \to \infty} \frac{1}{n} |EI^{n,*}| = 1.
\]

Proposition 5 states that perfect learning occurs when almost all agents are equilibrium informed. Intuitively, this indicates that almost all agents enjoy the information advantage as the population grows, and that the agents who cannot enjoy such benefit are negligible
in the society. This is consistent with the idea of social learning that successful learning allows individuals have sufficient information to make a good decision, and that such individuals represent an overwhelming proportion of the society.

Proposition 5 provides a transparent sufficient condition for perfect learning. We can do so because our perfect learning concept is powered by finite population learning, a sufficient condition of which only involves one set of equilibrium variables: \( k^{n,\sigma^*}_i \).

Next we consider the second sufficient condition that relies only on formation of the society. To streamline the presentation in the main texts, we assume that each agent enjoys a positive payoff even if she exits at the beginning. We will hold this assumption for the rest of this section. In the appendix, we relax this assumption and discuss all possible cases.

**Assumption 1.** \((\rho + \bar{\rho})\psi > 1\).

Before looking into the next sufficient condition for perfect learning, we first point out an important observation which states that, although the number of signals an agent gets in equilibrium may diverge to infinity with growth of the communication network, the equilibrium communication steps will not follow such a growth trend.

**Lemma 8.** Under Assumption 1, for any agent \( i \), the communication rounds she optimally experiences before taking an action in any social network \( G^n \) along a society \( \{G^n\}_{n=1}^\infty \) is bounded from above by a constant independent of \( n \). Mathematically,

\[
s^{n,\sigma^*}_i \leq s^n_i < \ln \left[ 1 - \frac{1}{(\rho + \bar{\rho})\psi} \right] / \ln \left( \frac{\lambda}{\lambda + r} \right),
\]

in which \( s^n_i \) stands for the optimal communication rounds for agent \( i \) given that other agents never exit.

A more general version (without Assumption 1) of Lemma 8 with its associated proof is included in the Appendix as Lemma 10. A key idea behind Lemma 8 is that even
expecting infinite number of signals does not justify the discount of waiting for the next communication step, after agent \( i \) gets sufficiently large number of signals within some finite communication steps. From condition (3.4.1), we see that the upper bound is exclusively determined by parameters of the information exchange game. Specifically, the upper bound is increasing in \( r \) while decreasing in \( \rho \), \( \bar{\rho} \), \( \psi \) and \( \lambda \).

Lemma 8 plays an important role in shaping our next sufficient condition that bypasses equilibrium and directly links perfect learning to network formations. Recall Proposition 5 which states that almost all agents’ \( k_i^{n,\sigma^*} \)’s go to infinity along the society is sufficient for perfect learning. On the other hand, from Lemma 8 we know that no agents have optimal communication steps \( s_i^{n,\sigma^*} \) going to infinity along the society. By combining the two observations, loosely speaking, we know that it should be the case that almost all agents get infinite number of signals within finite communication steps. In other words, if an agent obtains more signals along the society, it should not be through waiting for increasingly more communication steps. This consideration leads to our following definition of a socially informed agent.

**Definition 5 (Socially Informed Agent).** For each agent \( i \) in a given society \( \{G^n\}_{n=1}^\infty \), let \( S_i = \min\{s_0 \in \mathbb{N} : \lim_{n \to \infty} |B^n_{i,s_0}| = \infty\} \), where \( B^n_{i,s} \) is the set of agent \( i \)'s \( s \) degree neighbors in \( G^n \). Agent \( i \) is socially informed with respect to \( \{G^n\}_{n=1}^\infty \) if \( S_i \) is finite, and if there exists \( N \in \mathbb{N} \) such that for all social networks \( G^n \in \{G^n\}_{n=1}^\infty \) with \( n \geq N \), we have

\[
\psi - \frac{1}{\rho + \bar{\rho}|B^n_{i,S_i}|} > 0 \tag{3.4.2}
\]

and

\[
\left( \frac{\lambda}{\lambda + r} \right)^{S_i} \left( \psi - \frac{1}{\rho + \bar{\rho}|B^n_{i,S_i}|} \right) > \left( \frac{\lambda}{\lambda + r} \right)^s \left( \psi - \frac{1}{\rho + \bar{\rho}|B^n_{i,s}|} \right) \quad \text{for all } s < S_i, s \in \mathbb{N} \cup \{0\} \tag{3.4.3}
\]

Moreover, we denote by \( SI^n \) the set of socially informed agents in the network \( G^n \).
In Definition 5, condition (3.4.2) is automatically satisfied in view of Assumption 1. Intuitively, a socially informed agent can be reached by an infinite number of neighbors after some finite communication steps, which is captured as $S_i$. Furthermore, condition (3.4.3) ensures that this agent strictly prefers to wait at least until the arrival of such communication step $S_i$, given other agents never exit. Therefore with compatible incentive, agent $i$ is guaranteed to obtain an infinite number of signals from finite communication steps, if other agents never exit. Note also that the definition of a socially informed agent does not require knowledge of any specific equilibrium.

With the help of socially informed agents, we bypass equilibrium and state the following sufficient condition for perfect learning.

**Proposition 6.** The $\delta$-perfect learning occurs in a society $\{G^n\}_{n=1}^{\infty}$ under any equilibria $\{\sigma^n, *\}_{n=1}^{\infty}$ if

$$\lim_{n \to \infty} \frac{1}{n} |S\Gamma^n| = 1.$$

Proposition 6 is interesting because we can determine the occurrence of perfect learning through knowledge on the formation of society alone. Given our tractable conditions for socially informed agents, it is easy to check whether a given society sufficiently supports perfect learning under any equilibrium. Given the difficulty of explicitly solving for equilibria of the information exchange game, Proposition 6 is of more value. Although Proposition 2 of [1] has similar advantage as our Proposition 6, our sufficient condition based on socially informed agent is more transparent and powerful. This is largely because we employ a more general concept of perfect learning, and we highlight the observation documented in Lemma 8 that no agent would like to wait too long in case she eventually exits. Proposition 6 has more power in the sense that it can be easily interpreted to accommodate the criteria for perfect asymptotic learning based on information maven and social connector ([1]).
3.4.2 Learning Rates

In this subsection we define the learning rate for $\delta$-perfect learning. It is natural to expect similar concepts of learning rates for $\varepsilon$-perfect learning and $\bar{\varepsilon}$-perfect learning.

**Definition 6.** Suppose $\delta$-perfect learning occurs in $\{G^n\}^\infty_{n=1}$ under equilibria $\{\sigma^{n,*}\}^\infty_{n=1}$, then we call the corresponding sequence of tolerances $\{\delta_n\}^\infty_{n=1}$ the learning rate.

Intuitively, a desirable concept of learning rates must capture the quality of perfect learning. More concretely, faster learning rate should imply higher quality of perfect learning. Therefore, the sequence of tolerances themselves has natural advantages over other candidates. Recall that in finite population learning, the three tolerance parameters are natural measures of learning quality. With $\varepsilon$ and $\bar{\varepsilon}$ fixed, $\delta$ should be considered as the unique measure of learning quality, and smaller $\delta$ indicates better quality of finite population learning. Regarding perfect learning, learning rate as a measure of learning quality should be defined globally along the path to the limit, so any point measure is undesirable. In short, we have argued that a faster convergence rate of the vanishing sequence $\{\delta_n\}^\infty_{n=1}$ indicates higher quality of perfect learning.

It is worth highlighting the difference between our learning rate concept and the speed of convergence to a pre-defined consensus mainly employed in observational learning problems, which concerns about the time towards a consensus in a circumstance where people make repeated decisions and learn from others’ previous decisions to help make their own future decisions ([40], [39], [37]). In observational learning problems with repeated decisions, the observable sequence of aggregate decisions naturally reveals the time dynamics of information aggregation. However, in our story of direct communication, the dynamics of information aggregation along the time dimension is largely unobservable, which calls for alternative dimensions to look into the information dynamics. Thanks to our established results of finite population learning, the status of social learning through
communication can be determined in any finite population network, which in turn offers a natural standpoint to look into the dynamics of information aggregation and the evolution of social learning along the population dimension. The feature also distinguishes our work from existing literature on learning rate of similar spirit. For example, [2] attempt to define and investigate an asymptotic learning based rate in an observational learning context similar to [3]. Although their concept also captures a convergence rate of probabilities, it does not characterize the learning status in every social network along the society.

In what follows we try to explore the achievable fastest learning rate. With finite population learning as the foundation for perfect learning, we are able to discuss, along the path of perfect learning, achievable smallest $\delta_n$ for every communication network $G^n$ to reach finite population learning. This consideration also helps us find the size of population needed to achieve a given tolerance.

Next, we argue that it is not trivial to construct concretely the smallest sequence $\{\delta_n\}_{n=1}^\infty$ for perfect learning, while keeping other parameters fixed. Recall the sufficient condition part of Proposition 4, which implies $\delta$-perfect learning occurs with rates $\{\delta_n\}_{n=1}^\infty$ if

$$
\frac{1}{n} \sum_{i=1}^{n} \text{erf} \left( \varepsilon \sqrt{\rho + \hat{k}_{n,\sigma^*}^2} \right) \geq 1 - \delta_n \bar{\varepsilon} 
$$

holds for every communication network $G^n$ in the society $\{G^n\}_{n=1}^\infty$. Theoretically, if we can directly solve the inequalities with respect to $\delta_n$, the achievable fastest learning rate $\{\delta_n\}_{n=1}^\infty$ is constructed. However, two technical problems prevent us from directly doing so. First, we cannot read off a transparent rate out of the error function. Second, without specific knowledge of network formations, the relation between $k_i^{n,\sigma^*}$ and $n$ is hard to be pinned down generally. Hence, in the following we first discuss learning rates on
some examples, and then propose an approach to obtain a conservative estimate of the achievable fastest learning rate.

**Example 1 (Isolated Agents).** When all agents are isolated from each other in a communication network $G^n$, we have $k_i^{n,\sigma^*} = 1$ for every agent $i$.

In Example 1, the negative condition (3.3.1) is reduced to

$$erf\left(\varepsilon\sqrt{\frac{\rho + \bar{\rho}}{2}}\right) < (1 - \bar{\varepsilon})(1 - \delta_n).$$

If $erf\left(\varepsilon\sqrt{\frac{\rho + \bar{\rho}}{2}}\right) < (1 - \bar{\varepsilon})$, it is clear that for any vanishing sequence $\{\delta_n\}_{n=1}^{\infty}$, the above inequality must be satisfied for large $n$. This tells us that in fairly general circumstances, purely isolated society cannot achieve $\delta$-perfect learning. In this case, learning rate is irrelevant.

**Example 2 (Complete Graph).** When the communication network $G^n$ is a complete graph and the benefit of getting $n - 1$ new signals justifies the discount of one communication step, we have $k_i^{n,\sigma^*} = n$ for every agent $i$.

In Example 2, we have

$$erf\left(\varepsilon\sqrt{\frac{\rho + \bar{\rho}n}{2}}\right) \geq 1 - \delta_n\bar{\varepsilon},$$

as a sufficient condition for $\delta$-perfect learning, which translates to

$$\delta_n \geq \frac{1}{\bar{\varepsilon}} \left[1 - erf\left(\varepsilon\sqrt{\frac{\rho + \bar{\rho}n}{2}}\right)\right]. \tag{3.4.5}$$

The sequence of the right hand sides of inequality (3.4.5) can serve as the learning rate.

At the cost of getting a conservative estimate, we approximate the error function in order to get more transparent learning rate. Note that the error function $erf$ has an
approximation:

\[
1 - erf(x) < \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2}.
\]

Therefore a sufficient condition for \(\delta\)-perfect learning would be

\[
\delta_n \geq \frac{1}{\sqrt{\pi}} \frac{1}{\bar{\varepsilon}} \varepsilon \sqrt{\rho + \bar{\rho} n} \exp \left[-\frac{\varepsilon^2(\rho + \bar{\rho} n)}{4}\right].
\]

Keep other parameters fixed, and focus on the relations between population size \(n\) and \(\delta_n\). We see that \(\delta_n\) could decrease in the order of \(\frac{1}{\sqrt{n}} \exp(-n)\). This says that when population grows, the probability that at least \(\bar{\varepsilon}\) fraction of people make the wrong decision decreases very quickly to zero. In fact, it decreases faster than any polynomial rate. This is not totally surprising as \(k_{n,\sigma^*} = n\) for all agents.

Following the idea of error function approximations, we go beyond Example 2 to consider a more general case in which the \(k_{n,\sigma^*}^{i}\)'s of all agents in communication network \(G^n\) with equilibrium \(\sigma^{n,\ast}\) is lower bounded by a deterministic function \(f(n)\). That is, suppose there exists a deterministic function \(f\) such that \(k_{n,\sigma^*}^{i} \geq f(n)\) for every agent \(i\) in every communication network \(G^n\) of a society. Through the same approximation argument as above, we get a sufficient condition for \(\delta\)-perfect learning

\[
\delta_n \geq \frac{1}{\sqrt{\pi}} \frac{1}{\bar{\varepsilon}} \varepsilon \sqrt{\rho + \bar{\rho} f(n)} \exp \left[-\frac{\varepsilon^2(\rho + \bar{\rho} f(n))}{4}\right]. \tag{3.4.6}
\]

If \(f(n)\) converges to infinity as \(n\) goes to infinity, the right hand side of inequality (3.4.6) converges to 0. Keeping other parameters fixed, this implies \(\delta_n\) could decrease in the order of \(\frac{1}{\sqrt{f(n)}} \exp(-f(n))\). Formally, we summarize these discussions with the next proposition.

**Proposition 7.** Suppose there exists a deterministic function \(f(n)\) such that \(\lim_{n \to \infty} f(n) = \infty\), and \(k_{n,\sigma^*}^{i} \geq f(n)\) for any agent \(i\) in communication network \(G^n\) with associated com-
communication equilibrium \( \sigma^{n,*} \) along the society \( \{G^n\}_{n=1}^\infty \), then \( \delta \)-perfect learning could occur with learning rate \( \{\delta_n\}_{n=1}^\infty \), where each \( \delta_n \) is in the order of \( \frac{1}{\sqrt{f(n)}} \exp(-f(n)) \).

Finally, we offer an example to illustrate the potential power of Proposition 7.

**Example 3.** Suppose \( f(n) = C \cdot n \) where \( 0 < C < 1 \), then \( \delta \)-perfect learning could occur with learning rate \( \{\delta_n\}_{n=1}^\infty \), where each \( \delta_n \) is in the order of \( \frac{1}{\sqrt{n}} \exp(-n) \).

An interpretation of this example is that, even if agents in a communication network are sparse in the sense that each of them can only get a small proportion of information in the entire population, perfect learning can still be reached at a fast rate. This example can represent a scenario in which communication networks in a society consist of dispersed social groups while agents within these social groups are very closely connected. This may lead to interesting results pertaining to social cliques or homophily as discussed in [1] and [40], [39], [37].

### 3.5 Learning under Multiple Equilibria

As the information exchange game exhibits strategic complementarity, it is expected that multiple equilibria might emerge under some circumstances. An interesting perspective in investigating finite population learning is to measure the effect of learning against multiple equilibria. As opposed to Corollary 1 and Proposition 6 that offer conditions for learning under a given equilibrium, in this subsection learning status is evaluated against all equilibria. We provide the following generalized (conservative) version of finite population learning to accommodate multiple equilibria without equilibrium selection.

**Definition 7.** Denote by \( \Sigma^{n,*} = \{\sigma^{n,*}\} \) the set of equilibria of \( \Gamma_{info}(G^n) \), \( (\varepsilon, \bar{\varepsilon}, \delta) \)-learning occurs if

\[
\sup_{\sigma^{n,*} \in \Sigma^{n,*}} P_{\sigma^{n,*}} \left( \frac{1}{n} \sum_{i=1}^{n} (1 - M_i^{n,\varepsilon}) \geq \bar{\varepsilon} \right) \leq \delta.
\]
This definition offers a conservative standard to evaluate finite population learning in the sense that the least favorable equilibrium determines the learning status. When $\Sigma_{n,*}$ is a singleton, the above definition reduces to Definition 2. The proof of Proposition 4 can be recycled to get the next corollary.

**Corollary 2.** Given an information exchange game $\Gamma_{info}(G^n)$,

(a) $(\varepsilon, \bar{\varepsilon}, \delta)$-learning does not occur if

$$
\min_{\sigma^n, \bar{\sigma}^n \in \Sigma_{n,*}} \frac{1}{n} \sum_{i=1}^{n} \text{erf} \left( \frac{\varepsilon}{\sqrt{\rho + \bar{\rho} k_i^{n, \sigma^n}}} \right) < (1 - \bar{\varepsilon})(1 - \delta). 
$$

(b) $(\varepsilon, \bar{\varepsilon}, \delta)$-learning occurs if

$$
\min_{\sigma^n, \bar{\sigma}^n \in \Sigma_{n,*}} \frac{1}{n} \sum_{i=1}^{n} \text{erf} \left( \frac{\varepsilon}{\sqrt{\rho + \bar{\rho} k_i^{n, \sigma^n}}} \right) \geq 1 - \bar{\varepsilon}\delta.
$$

### 3.6 Remarks and Further Research

We have proposed a finite population learning concept to capture the level of information aggregation in any given communication network. In our framework, one equilibrium outcome, i.e., the number of signals obtained by an agent when she makes a decision, plays a key role. This equilibrium outcome is computable ([56]), which also allows us to numerically check the learning status. Different from existing literature that mainly address the learning behavior at the limit, this new concept helps reveal explicit interplays among time discount, frequency of communication, information precision and information sensitivity of the decision problem in a finite communication network. As the total amount of information fixed in a given finite network, our approach enables meaningful comparative statics regarding the effectiveness of information aggregation in networks. We also provide conditions for learning under a particular equilibrium, under any equilibrium, and
under all equilibria, respectively. Thanks to its tractability and transparency, the finite population learning concept offers a solid foundation to investigate long run dynamics of learning behavior and the associated learning rates as population diverges.

However, our analysis is also subject to certain limitations, which suggest directions for future research. In our model, complete information on the structure of a given communication network is required in determining the number of signals obtained by an agent, and in checking its corresponding learning status. In some circumstances, researchers do not want to assume such specific information, rather they want to get some understanding of the learning status regarding a large class of networks. This goal calls for some new criteria that can determine the learning status for given classes of networks with given finite population; ideally, these criteria should be formulated in terms of some functions of summary statistics of these networks. Relaxing the knowledge on specific network structure may also lead to more general results about the learning rates. However, this task is challenging within the current finite population learning framework. Specifically, our established conditions for finite population learning require all agents' exact numbers of signals upon their exits. Only knowing some commonly used summary statistics of the associated graphs can hardly help offer good estimates of these numbers of signals, mostly because these numbers of signals are also affected by other parameters not directly related to the network structure, such as the information precisions and the tolerances of learning. Therefore, the learning status of a certain class of communication networks is largely undetermined if we just consider properties of the graphs. To address this issue, we would like to have novel statistical properties of communication networks that are more friendly to analysis of communication learning, and more effective conditions for finite population learning are also needed. [40], [39], [37]
are promising attempts towards this direction, but their results as of now are still limited to the context of non-Bayesian observational learning.

Another line of generalization is to make our model more flexible and realistic. For example, the current setting assumes that agents have private signals with the same precision, which amounts to assuming that the total amount of information grows linearly with the population size when we consider the perfect learning. Hence, the improved learning status along a society reflects not only the level of information aggregation, but also an increased endowment of total information. Therefore, the net effect of information aggregation is not totally transparent. It might be interesting to allow the total information to increase in a nonlinear (e.g., log rate) fashion with the population size, and allow non-uniform distribution of signal precisions among agents. Also, even when we focus on certain classes of networks without specifying complete network structure, it is still assumed that any agent in the communication network knows the complete network structure. This assumption can be relaxed by limiting agents’ knowledge on the network to a certain local neighborhood, and infer other parts of the network according to her local knowledge. Keeping the Bayesian learning paradigm, other potential generalizations of our current work include considering the implications of correlated private information among agents, and heterogeneous characteristics of agents that are widely addressed in non-Bayesian learning frameworks.

3.7 Appendix

This appendix provides all proofs and generalized results of corresponding parts in the main text.

Proof of Proposition 4. On the one hand, to prevent \((\varepsilon, \bar{\varepsilon}, \delta)\)-learning, it is enough to show that a lower bound of \(P_{\sigma^{n,*}}\left(\frac{1}{n} \sum_{i=1}^{n} (1 - M^{n,\varepsilon}_i) \geq \bar{\varepsilon}\right)\) is greater than \(\delta\). It follows
from Markov inequality,

\[
\mathbb{P}_{\sigma^{n,*}} \left( \frac{1}{n} \sum_{i=1}^{n} M_{i}^{n,\varepsilon} > 1 - \bar{\varepsilon} \right) \leq \frac{\mathbb{E}_{\sigma^{n,*}} M_{i}^{n,\varepsilon}}{n(1 - \bar{\varepsilon})} = \frac{n - \sum_{i=1}^{n} erf \left( \frac{\sqrt{\frac{n}{\rho_{k}^{n,\sigma^{*}}}}}{} \right)}{n(1 - \bar{\varepsilon})}.
\]

This implies that

\[
\mathbb{P}_{\sigma^{n,*}} \left( \frac{1}{n} \sum_{i=1}^{n} (1 - M_{i}^{n,\varepsilon}) \geq \bar{\varepsilon} \right) = \mathbb{P}_{\sigma^{n,*}} \left( \frac{1}{n} \sum_{i=1}^{n} M_{i}^{n,\varepsilon} \leq 1 - \bar{\varepsilon} \right) > 1 - \frac{n - \sum_{i=1}^{n} erf \left( \frac{\sqrt{\frac{n}{\rho_{k}^{n,\sigma^{*}}}}}{} \right)}{n(1 - \bar{\varepsilon})}.
\]

Therefore, it is enough to take

\[
1 - \frac{n - \sum_{i=1}^{n} erf \left( \frac{\sqrt{\frac{n}{\rho_{k}^{n,\sigma^{*}}}}}{} \right)}{n(1 - \bar{\varepsilon})} > \delta,
\]

which concludes that the condition (3.3.1) is a necessary condition for \((\varepsilon, \bar{\varepsilon}, \delta)\)-learning. On the other hand, similarly, to ensure \((\varepsilon, \bar{\varepsilon}, \delta)\)-learning we have

\[
\mathbb{P}_{\sigma^{n,*}} \left( \frac{1}{n} \sum_{i=1}^{n} (1 - M_{i}^{n,\varepsilon}) \geq \bar{\varepsilon} \right) \leq \frac{\mathbb{E}_{\sigma^{n,*}} (\sum_{i=1}^{n} (1 - M_{i}^{n,\varepsilon}))}{n \bar{\varepsilon}} = \frac{n - \sum_{i=1}^{n} erf \left( \frac{\sqrt{\frac{n}{\rho_{k}^{n,\sigma^{*}}}}}{} \right)}{n \bar{\varepsilon}}.
\]

To demand the right hand side of the above inequality smaller than or equal to \(\delta\), it is sufficient and necessary to assume the condition (3.3.2) in the proposition. This concludes the part that the condition (3.3.2) is a sufficient condition for \((\varepsilon, \bar{\varepsilon}, \delta)\)-learning. \(\square\)

The following provides a more general sufficient condition for \(\delta\)-perfect learning, given the knowledge of equilibrium. Let \(f_1 \geq f_2 \geq \ldots \geq f_J\), where each \(f_j(n)\) is a monotone increasing function (not necessarily strictly increasing) on \(n\). Let \(\{b_{i}^{j} : j = 1, \ldots, J\}\) be such that

\[
\frac{\{|i : k_{i}^{n,\sigma^{*}} \geq f_1(n)\}}{n} \geq b_{i}^{1},
\]
\[
\left| \left\{ i : f_1(n) > k_i^{n, \sigma^*} \geq f_2(n) \right\} \right| \geq b_n^2,
\]
and up until
\[
\left| \left\{ i : f_{J-1}(n) > k_i^{n, \sigma^*} \geq f_J(n) \right\} \right| \geq b_n^J.
\]
Clearly, \( b_n^1, \ldots, b_n^J \in (0, 1) \) and \( 0 \leq b_n^1 + \ldots + b_n^J \leq 1 \). The rest agents \( i \)'s are such that \( k_i^{n, \sigma^*} \geq 1 \) but their \( k_i^{n, \sigma^*} \) doesn’t satisfy any of the inequalities above. Their fraction is at most \( 1 - (b_n^1 + \ldots + b_n^J) \).

**Lemma 9.** \( \delta \)-perfect learning occurs if

(a) \( \lim_{n \to \infty} \sum_{j=1}^J b_n^j = 1 \),

(b) for each \( j \in \{1, \ldots, J\} \), \( \lim_{n \to \infty} b_n^j \left( 1 - \text{erf} \left( \frac{\sqrt{\rho + \bar{\rho} f_j(n)}}{2} \right) \right) = 0 \).

**Proof of Lemma 9.** Recall that a sufficient condition for \((\epsilon, \bar{\epsilon}, \delta_n)^\)-learning is

\[
\frac{1}{n} \sum_{i=1}^n \text{erf} \left( \epsilon \sqrt{\frac{\rho + \bar{\rho} k_i^{n, \sigma^*}}{2}} \right) \geq 1 - \delta_n \epsilon.
\]

Then by the definition of \( b_n^j \) and \( f_j \), it is enough to have

\[
\sum_{j=1}^J \text{erf} \left( \epsilon \sqrt{\frac{\rho + \bar{\rho} f_j(n)}}{2} \right) \cdot b_n^j + \left( 1 - \sum_{j=1}^J b_n^j \right) \text{erf} \left( \epsilon \sqrt{\frac{\rho + \bar{\rho}}{2}} \right) \geq 1 - \delta_n \epsilon.
\]

This translates to

\[
\delta_n \geq \frac{1}{\epsilon} \left[ \sum_{j=1}^J b_n^j \left( 1 - \text{erf} \left( \epsilon \sqrt{\frac{\rho + \bar{\rho} f_j(n)}}{2} \right) \right) + \left( 1 - \sum_{j=1}^J b_n^j \right) \left( 1 - \text{erf} \left( \epsilon \sqrt{\frac{\rho + \bar{\rho}}{2}} \right) \right) \right] .
\]

To ensure \( \lim_{n \to \infty} \delta_n = 0 \), it is necessary to have \( \lim_{n \to \infty} \sum_{j=1}^J b_n^j \). Moreover, for each \( j \in \{1, \ldots, J\} \),

\[
\lim_{n \to \infty} b_n^j \left( 1 - \text{erf} \left( \epsilon \sqrt{\frac{\rho + \bar{\rho} f_j(n)}}{2} \right) \right) = 0.
\]

A remark is that if \( f_j \) does not increase strictly for \( n \geq N^* \), \( b_n^i \) needs to decrease to 0. \( \Box \)
Regarding learning rate, we briefly mention that allowing more than one tolerances among \( \varepsilon, \bar{\varepsilon}, \delta \) to vary with population size \( n \) also leads to interesting results. In particular, from the proof of Lemma 9, a sufficient condition for \((\varepsilon, \bar{\varepsilon}, \delta_n)\)-learning is

\[
\delta_n \bar{\varepsilon}_n \geq \sum_{j=1}^{J} b_n^j \left( 1 - erf \left( \varepsilon \sqrt{\frac{\rho + \bar{\rho} f_j(n)}{2}} \right) \right) + \left( 1 - \sum_{j=1}^{J} b_n^j \right) \left( 1 - erf \left( \varepsilon \sqrt{\frac{\rho + \bar{\rho}}{2}} \right) \right).
\]

In this condition, the role of \( \delta_n \) and that of \( \bar{\varepsilon}_n \) are completely interchangeable, which implies that we can trade in equal numeric values the probabilistic confidence with the fraction of agents who make the wrong decisions.

The following provides a generalized version of Lemma 8 when Assumption 1 is relaxed, and its corresponding proof.

**Lemma 10 (Generalized Lemma 8).** For any agent \( i \), either the communication step she optimally experiences before taking an action in any social network \( G^n \) along a society \( \{G^n\}_{n=1}^{\infty} \) is bounded from above by a constant independent of \( n \), or she waits forever and never exits. Specifically, in any \( G^n \):

(a) If \((\rho + \bar{\rho})\psi > 1\), then for any agent \( i \)

\[
s_i^{n,\sigma^*} \leq s_i^n < \ln \left[ 1 - \frac{1}{(\rho + \bar{\rho})\psi} \right] / \ln \left( \frac{\lambda}{\lambda + r} \right),
\]

in which \( s_i^n \) stands for the optimal communication rounds for agent \( i \) given that other agents never exit.

(b) If \((\rho + \bar{\rho})\psi \leq 0\), then for any agent \( i \)

\[
s_i^{n,\sigma^*} = s_i^n = \infty.
\]

(c) If \( 0 < (\rho + \bar{\rho})\psi \leq 1 \), then there are two subcases.

(c.1) For agent \( i \) with

\[
\lim_{n \to \infty} |B_i^n| < \frac{1 - \rho \psi}{\rho \psi},
\]

110
we have

$$s_i^{n,\sigma^*} = s_i^n = \infty.$$  

(c.2) For agent $i$ with

$$\lim_{n \to \infty} |B_n^i| \geq \frac{1 - \rho \psi}{\bar{\rho} \psi},$$

we have either

$$s_i^{n,\sigma^*} \leq s_i^n \leq s_i^{\{G_n\}_{n=1}^\infty},$$

or

$$s_i^{n,\sigma^*} = \infty, s_i^n \leq s_i^{\{G_n\}_{n=1}^\infty},$$

or

$$s_i^{n,\sigma^*} = s_i^n = \infty,$$

where $s_i^{\{G_n\}_{n=1}^\infty}$ is a constant that depends on the society and agent $i$’s position in the society, and it does not change with $n$.

**Proof of Lemma 10.** We proceed case by case.

**Case (A).** $(\rho + \bar{\rho})\psi > 1$.

In this case, agent $i$ enjoys a positive payoff $\psi - \frac{1}{\rho + \bar{\rho}}$ if she exists at $t = 0$ and does not communicate with anyone else. Note that her expected payoff at $t = 0$ by taking $s_i^n$ communication steps is strictly upper bounded by $\left(\frac{\lambda}{\lambda + r}\right)^{s_i^n} \psi$. Therefore, it is suboptimal for her to choose a $s_i^n$ such that

$$\left(\frac{\lambda}{\lambda + r}\right)^{s_i^n} \psi \leq \psi - \frac{1}{\rho + \bar{\rho}},$$

which implies

$$s_i^n < \ln \left[1 - \frac{1}{(\rho + \bar{\rho})\psi}\right] / \ln \left(\frac{\lambda}{\lambda + r}\right)$$

is necessary for agent $i$’s optimality. $s_i^{n,\sigma^*} \leq s_i^n$ comes from the fact that other agents do not necessarily wait forever in an equilibrium, so that it may be optimal for agent $i$ to exit earlier too.

**Case (B).** $(\rho + \bar{\rho})\psi \leq 0$. 

111
Now agent $i$ always gets a negative payoff whenever she exists. Moreover, if she tries to take an action at any time $t$, it is always preferable for her to wait for another communication step and take an action, in order to discount the negative payoff. In this sense, she optimally chooses to wait forever and never exits, both in equilibrium and given that other agents never exit. In other words, $s^n_{i,\sigma^*} = s^n_i = \infty$.

**Case (c.1).** $0 < (\rho + \bar{\rho})\psi \leq 1$ and $\lim_{n \to \infty} |B^n_i| < \frac{1-\rho\psi}{\rho\psi}$.

Again, agent $i$ always gets a negative payoff whenever she exists. Hence, this corresponds to case (b) and $s^n_{i,\sigma^*} = s^n_i = \infty$.

**Case (c.2).** $0 < (\rho + \bar{\rho})\psi \leq 1$ and $\lim_{n \to \infty} |B^n_i| \geq \frac{1-\rho\psi}{\rho\psi}$.

On the one hand, for any $G^n$ with $|B^n_i| < \frac{1-\rho\psi}{\rho\psi}$, agent $i$ always gets a negative payoff whenever she exists. Hence, this again corresponds to case (b) and $s^n_{i,\sigma^*} = s^n_i = \infty$.

On the other hand, for any $G^n$ with $|B^n_i| \geq \frac{1-\rho\psi}{\rho\psi}$, we consider the communication step $(S^n_{i})_{\text{max}}$ in which agent $i$ obtains signals from all her neighbors $B^n_i$ for the first time, provided others never exit. In other words, $(S^n_{i})_{\text{max}}$ is the distance between agent $i$ and her farthest neighbor in $G^n$. As agent $i$ must enjoy a nonnegative payoff if she exactly exits after the communication step $(S^n_{i})_{\text{max}}$ in this case, so waiting longer from this point should only make her worse off because of the discount. Hence

$$s^n_i \leq (S^n_{i})_{\text{max}} \quad (3.7.3)$$

is necessary for agent $i$’s optimality. Note that, $(S^n_{i})_{\text{max}}$ is non-decreasing in $n$ for any given $i$, and $|B^n_{i,s}|$ is strictly monotone increasing in $s$ when $s \leq (S^n_{i})_{\text{max}}$ for any given $n$ and $G^n$.

Also, for the given communication network $G^n$, there exists at least one communication step $s$ such that after that agent $i$ gets positive payoff, given that other agents never exit. Hence, it is suboptimal for her to wait longer if

$$\left(\frac{\lambda}{\lambda + r}\right)\psi \leq \psi - \frac{1}{\rho + \bar{\rho}|B^n_{i,s}|},$$

which implies

$$|B^n_{i,s}| < \frac{\lambda + r - \rho r\psi}{\bar{\rho} r\psi} \quad (3.7.4)$$

112
is necessary for agent $i$’s optimality (optimal for no longer waiting).

Now we consider two sub-cases. First, we discuss the case when $\lim_{n \to \infty} |B^n_i| < \infty$. In this case, we must have $\lim_{n \to \infty} (S^n_i)_{\text{max}} < \infty$, because $(S^n_i)_{\text{max}} \leq |B^n_i|$ for any $i$ and $n$. Recall that $(S^n_i)_{\text{max}}$ is non-decreasing in $n$ for any $i$, we further have $(S^n_i)_{\text{max}} \leq \lim_{n \to \infty} (S^n_i)_{\text{max}}$. Hence, by (3.7.3) we obtain

$$s^n_i \leq \lim_{n \to \infty} (S^n_i)_{\text{max}} < \infty$$  \hspace{1cm} (3.7.5)

for all $G^n$ satisfying $|B^n_i| \geq \frac{1-\rho \psi}{\rho \psi}$ and $\lim_{n \to \infty} |B^n_i| < \infty$. We denote $\lim_{n \to \infty} |B^n_i|$ as $s^n_{1i}(G^n)_{n=1}$, which is a constant that depends on the society and agent $i$’s position in the society and does not change with respect to $n$.

Second, we discuss the case when $\lim_{n \to \infty} |B^n_i| = \infty$. Now there should be either $\lim_{n \to \infty} (S^n_i)_{\text{max}} < \infty$ or $\lim_{n \to \infty} (S^n_i)_{\text{max}} = \infty$. If it is the former case, it corresponds to the case when $\lim_{n \to \infty} |B^n_i| < \infty$ and we have $s^n_i \leq \lim_{n \to \infty} (S^n_i)_{\text{max}} = s^n_{1i}(G^n)_{n=1}$ for all $G^n$. If it is the latter case, as $(S^n_i)_{\text{max}}$ is non-decreasing in $n$ for any given $i$ and $|B^n_{i,s}|$ is strictly monotone increasing in $s$ when $s \leq (S^n_i)_{\text{max}}$ for any given $n$ and $G^n$, there exists a largest $G^N$ with its associated $(S^N_i)_{\text{max}}$ that satisfies condition (3.7.4). Hence, by (3.7.3) and (3.7.4) we obtain

$$s^n_i \leq (S^N_i)_{\text{max}}$$  \hspace{1cm} (3.7.6)

for all $G^n$ satisfying $|B^n_i| \geq \frac{1-\rho \psi}{\rho \psi}$, $\lim_{n \to \infty} |B^n_i| = \infty$ and $\lim_{n \to \infty} (S^n_i)_{\text{max}} = \infty$. We denote $(S^N_i)_{\text{max}}$ as $s^n_{2i}(G^n)_{n=1}$, which is again a constant that depends on the society and agent $i$’s position in the society and does not change with respect to $n$.

To sum up, we denote $s_i^{(G^n)_{n=1}} = \max\{s^n_{1i}(G^n)_{n=1}, s^n_{2i}(G^n)_{n=1}\}$ and it finally reaches $s^n_i \leq s_i^{(G^n)_{n=1}}$ for agent $i$ in such $G^n$ with $|B^n_i| \geq \frac{1-\rho \psi}{\rho \psi}$, where $s_i^{(G^n)_{n=1}}$ is a constant that depends on the society and agent $i$’s position in the society and does not change with respect to $n$.

In terms of $s^n_{i,\sigma^*}$ in this case, in an equilibrium, agent $i$ gets weakly fewer signals than that she can get when other agents never exit, so that there can be either the case in which she gets positive payoff and takes an action weakly earlier, namely, $s^n_{i,\sigma^*} \leq S^n_i$, or the case in which she cannot get enough signals to ensure a positive payoff so that she optimally wait forever, namely, $s^n_{i,\sigma^*} = \infty$. This concludes the proof. \hspace{1cm} \square
Proof of Proposition 6. By Lemma 9, it suffices to show that the number of signals obtained under any equilibrium by any socially informed agent $i$ goes to infinity, as $n$ goes to infinity along the society. In other words, we want to show that $\lim_{n \to \infty} k^{n,\sigma^*}_i = \infty$ under any equilibrium $\sigma$ for any socially informed agent $i$. In the following we consider a fixed socially informed agent $i$. Recall that in Definition 5, the communication step $S_i$ is defined as the smallest positive integer satisfying $\lim_{n \to \infty} |B^n_{i,S_i}| = \infty$.

We first discuss the case when $S_i = 1$. Since any agent still communicates with others even after she exits, $B^n_{i,1} = B^n_{i,1}$ under any equilibrium $\sigma$ for any $n$. As agent $i$ is socially informed, we have for sufficiently large $n$

$$\psi - \frac{1}{\rho + \bar{\rho}|B^n_{i,1}|} > 0 \text{ under any } \sigma^{n,*}$$

and

$$\left(\frac{\lambda}{\lambda + r}\right) \left(\psi - \frac{1}{\rho + \bar{\rho}|B^n_{i,1}|}\right) > \psi - \frac{1}{\rho + \bar{\rho}} \text{ under any } \sigma^{n,*}.$$

The above display implies $k_i^{n,\sigma^*} \geq |B^n_{i,1}|$ under any $\sigma^{n,*}$ for sufficiently large $n$. As a consequence, $\lim_{n \to \infty} k_i^{n,\sigma^*} \geq \lim_{n \to \infty} |B^n_{i,1}| = \lim_{n \to \infty} |B^n_{1,1}| = \infty$ under any $\sigma^{n,*}$.

The following discussion is on the cases when $S_i \geq 2$. We proceed through three steps.

Step 1. We claim when $S_i \geq 2$, for sufficiently large network $G^n$, there exists at least one path $\{j_{S_i-1}, j_{S_i-2}, \ldots, j_1, i\}$ from $j_{S_i-1}$ to $i$ such that

$$\lim_{n \to \infty} |B^n_{j_{S_i-1},s}| = \infty \text{ for all } s \in \{1, \ldots, S_i - 1\}. \quad (3.7.7)$$

Now we construct the path $\{j_{S_i-1}, j_{S_i-2}, \ldots, j_1, i\}$ that satisfies condition (3.7.7). Because $S_i$ is the smallest integer $j$ such that $\lim_{n \to \infty} |B^n_{j,1}| = \infty$, $B^n_{i,S_i-1} \setminus B^n_{i,S_i-2}$, the set of points that of distance $S_i - 1$ to $i$, must be finite in the limit, i.e., $\lim_{n \to \infty} |B^n_{i,S_i-1} \setminus B^n_{i,S_i-2}| < \infty$. Therefore, there is at least one agent $j$ of distance $S_i - 1$ to $i$, such that $\lim_{n \to \infty} |B^n_{j,1}| = \infty$. We denote one of such agents $j$ as $j_{S_i-1}$. If $S_i = 2$, the desired path has been constructed. When $S_i \geq 3$, choose any path $\{j_{S_i-1}, j_{S_i-2}, \ldots, j_1, i\}$ from the chosen $j_{S_i-1}$ to $i$. Clearly, $j_{S_i-s} \in B^n_{i,S_i-s}$. Moreover, condition (3.7.7) is satisfied in view of $\lim_{n \to \infty} |B^n_{j_{S_i-1},1}| = \infty$. 

114
We next argue that when \( S_i \geq 2 \), the agent \( j_{S_i-1} \) on the path \( \{j_{S_i-1}, j_{S_i-2}, ..., j_1, i\} \) will not exit before she experiences \( s \) communication steps under any equilibrium \( \sigma^{n,*} \) provided that \( n \) is sufficiently large. It is worth noting that agent \( j_{S_i-1} \) does not necessarily get a positive payoff when she experiences \( s \) communication steps in equilibrium.

We will see this by induction from \( j_{S_i-1} \) to \( j_1 \) sequentially. We first show that agent \( j_{S_i-1} \) will not exit before she experiences her first communication step in any equilibrium \( \sigma^{n,*} \) provided that \( n \) is sufficiently large. It requires that there exists \( N \) such that for all social networks \( G^n \in \{G^n\}_{n=1}^{\infty} \) and its associated equilibrium \( \sigma_{n,*} \) with \( n \geq N \),

\[
\left( \frac{\lambda}{\lambda + r} \right) \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{j_{S_i-1},1}^{n,\sigma^{n,*}}|} \right) > \psi - \frac{1}{\rho + \bar{\rho}} \quad \text{under any } \sigma^{n,*}. \quad (3.7.8)
\]

To validate condition (3.7.8), recall condition (3.4.3) from Definition 5 for \( s = S_i - 1 \), which states that there exists \( N \) such that for all social networks \( G^n \in \{G^n\}_{n=1}^{\infty} \) with \( n \geq N \) it holds

\[
\left( \frac{\lambda}{\lambda + r} \right) \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{i,S_i-1}^{n}|} \right) > \psi - \frac{1}{\rho + \bar{\rho}|B_{i,S_i-1}^{n}|}. \quad (3.7.9)
\]

By the definition of \( S_i \), the construction of \( j_{S_i-1} \) and the fact that \( B_{j_{S_i-1},1}^{n,\sigma^{n,*}} = B_{j_{S_i-1},1}^{n} \) under any equilibrium \( \sigma^{n,*} \) with any \( n \), we know that \( \lim_{n \to \infty} |B_{j_{S_i-1},1}^{n,\sigma^{n,*}}| = \lim_{n \to \infty} |B_{j_{S_i-1},1}^{n}| = \infty \) under any \( \sigma^{n,*} \) and \( \lim_{n \to \infty} |B_{i,S_i-1}^{n}| < \infty \). Also we have \( |B_{i,S_i-1}^{n}| \geq 1 \). Note that the right hand side of condition (3.7.9) is greater than or equal to the right hand side of condition (3.7.8), we obtain easily that (3.7.8) holds for sufficiently large \( n \). Hence we get that agent \( j_{S_i-1} \) will not exit before she experiences her first communication step under any \( \sigma^{n,*} \) provided that \( n \) is sufficiently large.

We then show that agent \( j_{S_i-2} \) (for \( S_i \geq 3 \)) will not exit before she experiences her second communication step under any equilibrium for sufficiently large \( n \). It requires that there exists \( N \) such that for all social networks \( G^n \in \{G^n\}_{n=1}^{\infty} \) and its associated
equilibrium $\sigma^{n,*}$ with $n \geq N$,
\[
\left( \frac{\lambda}{\lambda + r} \right)^2 \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{j_i,s_i-2}|} \right) > \psi - \frac{1}{\rho + \bar{\rho}} \tag{3.7.10}
\]
and
\[
\left( \frac{\lambda}{\lambda + r} \right)^2 \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{i,S_i-2}|} \right) > \left( \frac{\lambda}{\lambda + r} \right) \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{i,S_i-1}|} \right). \tag{3.7.11}
\]

To validate (3.7.10) and (3.7.11), we used again the condition (3.4.3) from Definition 5 for $s = S_i - 2$ and $s = S_i - 1$, which state that there exists $N$ such that for all social networks $G^n \in \{G^n\}_{n=1}^{\infty}$ with $n \geq N$ we have
\[
\left( \frac{\lambda}{\lambda + r} \right)^2 \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{j_i,s_i-2}|} \right) > \psi - \frac{1}{\rho + \bar{\rho}|B_{i,S_i-2}|} \tag{3.7.12}
\]
and
\[
\left( \frac{\lambda}{\lambda + r} \right)^2 \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{i,S_i-2}|} \right) > \left( \frac{\lambda}{\lambda + r} \right) \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{i,S_i-1}|} \right). \tag{3.7.13}
\]

Similarly, by the definition of $S_i$ and the construction of $j_{S_i-1}$ and $j_{S_i-2}$, we know that $\lim_{n \to \infty} |B_{j_{S_i-2}}| = \lim_{n \to \infty} |B_{i,S_i}| = \infty$, $\lim_{n \to \infty} |B_{i,S_i-1}| < \infty$, $\lim_{n \to \infty} |B_{i,S_i-2}| < \infty$, and $\lim_{n \to \infty} |B_{j_{S_i-2},1}| < \infty$ under any equilibrium $\sigma$. Also we have $|B_{i,S_i-2}| \geq 1$ and $B_{j_{S_i-2},1} \subseteq B_{j_{S_i-2},1} \subseteq B_{i,S_i-1}$ (and thus $|B_{i,S_i-1}| \geq |B_{j_{S_i-2},1}| \geq |B_{j_{S_i-2},1}|$) for any $n$ under any equilibrium $\sigma^{n,*}$. Note that the right hand side of condition (3.7.12) is greater than or equal to the right hand side of condition (3.7.10), and the right hand side of condition (3.7.13) is greater than or equal to the right hand side of condition (3.7.11). Then it can be verified that the following two conditions hold for sufficiently large $n$, the right hand sides of which are the same as those in conditions (3.7.10) and (3.7.11):
\[
\left( \frac{\lambda}{\lambda + r} \right)^2 \left( \psi - \frac{1}{\rho + \bar{\rho}|B_{j_{S_i-2},2}|} \right) > \psi - \frac{1}{\rho + \bar{\rho}} \tag{3.7.14}
\]
and
\[
\left( \frac{\lambda}{\lambda + r} \right)^2 \left( \psi - \frac{1}{\rho + \tilde{\rho}|B_{j_{S_{i-2},1}}^n|} \right) > \left( \frac{\lambda}{\lambda + r} \right) \left( \psi - \frac{1}{\rho + \tilde{\rho}|B_{j_{S_{i-2},1}}^{n,\sigma^*}|} \right).
\] (3.7.15)

Furthermore, recall that we have already shown that agent \( j_{S_{i-1}} \) will not exit before she experiences her first communication step under any equilibrium \( \sigma^{n,*} \) provided that \( n \) is sufficiently large, which implies that \( B_{j_{S_{i-1},1}}^{n,\sigma^*} \subseteq B_{j_{S_{i-2},2}}^n \) under any \( \sigma \) for sufficiently large \( n \), and thus \( \lim_{n \to \infty} |B_{j_{S_{i-1},1}}^{n,\sigma^*}| \geq \lim_{n \to \infty} |B_{j_{S_{i-1},1}}^n| = \infty \) under any equilibrium \( \sigma^{n,*} \). Also we know that \( \lim_{n \to \infty} |B_{j_{S_{i-1},2}}^n| < \infty \). Together with conditions (3.7.14) and (3.7.15), these facts validate conditions (3.7.10) and (3.7.11). Hence we get that agent \( j_{S_{i-2}} \) will not exit before she experiences her second communication step in any \( \sigma^{n,*} \) provided that \( n \) is sufficiently large.

The arguments above for \( j_{S_{i-2}} \) can be extended successively to \( j_1 \). Hence, under any equilibrium \( \sigma^{n,*} \), no \( j_{S_{i-s}} \) in the established path \( \{ j_{S_{i-1}}, j_{S_{i-2}}, ..., j_1, i \} \) will exit before she experiences \( s \) communication steps under any equilibrium \( \sigma^{n,*} \) provided that \( n \) is sufficiently large. A byproduct is that \( \lim_{n \to \infty} |B_{j_{S_{i-s},s}}^{n,\sigma^*}| = \infty \) under any \( \sigma^{n,*} \), for \( s \in \{1, 2, ..., S_{i-1}\} \).

**Step 3.** Finally, we argue that the socially informed agent \( i \) will not exit before she experiences \( S_i \) communication steps under any equilibrium \( \sigma^{n,*} \) when \( n \) is sufficiently large. It requires that there exists \( N \in \mathbb{N} \) such that for all social networks \( G^n \in \{G^n\}_{n=1}^\infty \) with \( n \geq N \), we have
\[
\psi - \frac{1}{\rho + \tilde{\rho}|B_{i,S_i}^n|} > 0 \text{ under any } \sigma^{n,*}
\] (3.7.16)
and
\[
\left( \frac{\lambda}{\lambda + r} \right)^{S_i} \left( \psi - \frac{1}{\rho + \tilde{\rho}|B_{i,S_i}^n|} \right) > \left( \frac{\lambda}{\lambda + r} \right)^s \left( \psi - \frac{1}{\rho + \tilde{\rho}|B_{i,s}^{n,\sigma^*}|} \right) \text{ under any } \sigma^{n,*}
\] (3.7.17)
for all \( s < S_i, s \in \mathbb{N} \cup \{0\} \).

Recall that we have already shown that agent \( j_{S_{i-s}} \) in the constructed path will not exit before she experiences \( S_i - s \) communication steps for \( s \in \{1, 2, ..., S_{i-1}\} \), under any equilibrium \( \sigma^{n,*} \) provided that \( n \) is sufficiently large, which implies that \( B_{j_{S_{i-1},1}}^{n,\sigma^*} \subseteq B_{j_{S_{i-2},2}}^n \subseteq ... \).
... \subseteq B_{j_i, S_i - 1}^{n, \sigma^*} \subseteq B_{i, S_i}^{n, \sigma^*} under any \sigma for sufficiently large $n$, and thus $\lim_{n \to \infty} \left| B_{i, S_i}^{n, \sigma^*} \right| \geq \lim_{n \to \infty} \left| B_{j_i, S_i - 1}^{n, \sigma^*} \right| \geq \cdots \geq \lim_{n \to \infty} \left| B_{j_{S_i - 2}, 2}^{n, \sigma^*} \right| \geq \lim_{n \to \infty} \left| B_{j_{S_i - 1}, 1}^{n, \sigma^*} \right| = \infty$ under any $\sigma$. Also, we have $B_{i, s}^{n, \sigma^*} \subseteq B_{i, s}^n$ and thus $\left| B_{i, s}^{n, \sigma^*} \right| \leq \left| B_{i, s}^n \right|$, under any $\sigma$ for $s \in \{1, 2, \ldots, S_i - 1\}$, which implies the right hand sides of condition (3.4.3) are greater than or equal to the right hand sides of condition (3.7.17), for $s \in \{1, 2, \ldots, S_i - 1\}$. Moreover, we know that $\lim_{n \to \infty} \left| B_{i, s}^{n, \sigma^*} \right| \leq \lim_{n \to \infty} \left| B_{i, s}^n \right| < \infty$ for $s \in \{1, 2, \ldots, S_i - 1\}$ by the definition of $S_i$. Together with conditions (3.4.2) and (3.4.3) in Definition 5, these facts validate conditions (3.7.16) and (3.7.17). Hence we get that the socially informed agent $i$ will not exit before she experiences $S_i$ communication steps and she can enjoy a positive payoff when she experiences $S_i$ communication steps, under any $\sigma^{n, \ast}$ provided that $n$ is sufficiently large. This further implies $k_{i, S_i}^{n, \sigma^*} \geq \left| B_{i, S_i}^{n, \sigma^*} \right|$ under any $\sigma$ with sufficiently large $n$, which finally leads to $\lim_{n \to \infty} \left| k_{i, S_i}^{n, \sigma^*} \right| \geq \lim_{n \to \infty} \left| B_{i, S_i}^{n, \sigma^*} \right| = \lim_{n \to \infty} \left| B_{i, S_i}^n \right| = \infty$ under any $\sigma^*$ when $S_i \geq 2$. This concludes the proof. \qed


122


[80] D. S. Zhao and Y. Li. Principled sure independence screening for cox models with ultra-high-dimensional covariates. 2010. manuscript.

