Naïve Bayes Data Complexity and Characterization of Optima of the Unsupervised Expected Likelihood

by

Ali Wytsma

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2017

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The naïve Bayes model is a simple model that has been used for many decades, often as a baseline, for both supervised and unsupervised learning. With a latent class variable it is one of the simplest latent variable models, and is often used for clustering. The estimation of its parameters by maximum likelihood (e.g. using gradient ascent, expectation maximization) is subject to local optima since the objective is non-concave. However, the conditions under which global optimality can be guaranteed are currently unknown. I provide a first characterization of the optima of the naïve Bayes model. For problems with up to three features, I describe comprehensive conditions that ensure global optimality. For more than three features, I show that all stationary points exhibit marginal distributions with respect to the features that match those of the training data. In a second line of work, I consider the naïve Bayes model with an observed class variable, which is often used for classification. Well known results provide some upper bounds on order of the sample complexity for agnostic PAC learning, however exact bounds are unknown. These bounds would show exactly how much data is needed for model training using a particular algorithm. I detail the framework for determining an exact tight bound on sample complexity, and prove some of the sub-theorems that this framework rests on. I also provide some insight into the nature of the distributions that are hardest to model within specified accuracy parameters.

Acknowledgements

I would like express my sincerest thanks to Professor Poupart. His guidance and support are very deeply appreciated, and I am very thankful for his remarkable enthusiasm. I would like to thank George Trimponias, with whom I collaborated significantly over the past two years. I am deeply thankful for his creativity, passion and optimism.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The naïve Bayes model is one of the most basic models in machine learning. It was introduced over half a century ago, and is widely used as a baseline model in various settings. The naïve Bayes model has many advantages, such as its simplicity and its scalability, which is due to its number of parameters being linear in the number of features.

The naïve Bayes model is effective for both supervised and unsupervised learning. When the class variable is observed, it can be used for classification. It can be used in a variety of settings, such as text classification [16], sentiment classification [30] and spam filtering [34].

When the class variable is unobserved and its parameters are estimated in an unsupervised fashion, it can be used for clustering. For instance, it is often used as a baseline for clustering data with discrete features such as gene expression [35], behavioural data [12] and text documents [22]. It is a special case of the latent Dirichlet allocation model [8] where all the words in a document are forced to be generated by the same latent topic. It can also be viewed as a discrete version of the popular Gaussian mixture model where the Gaussian components are replaced by discrete distributions.

Despite the naïve Bayes being a well established and commonly used model, some of its fundamental properties are not well understood. In this thesis, I will describe my research into the characterization of the optima of the likelihood of the naïve Bayes model in the unsupervised setting. I will then describe my research related to the data complexity of the naïve Bayes model in the supervised setting.

A common approach for unsupervised training of a naïve Bayes model for clustering consists of maximizing the likelihood of the data. Since the optimization objective is non-concave, popular algorithms such as gradient ascent and expectation maximization [14]

may get stuck in local optima. Despite the naïve Bayes model being one of the oldest and most basic models in machine learning, there is no characterization of the optima of the likelihood objective. This is a major gap in the theory of the naïve Bayes model that translates into some uncertainty about the reliability of the clusterings found in practice. With the democratization of machine learning, there is a need for software libraries that produce reliable clusterings. However, at the moment, practitioners cannot trust that a clustering produced by a naïve Bayes model is as good as possible (in terms of data likelihood).

I provide a first analysis of the stationary points of the unsupervised naïve Bayes likelihood. For up to three binary features, I show that global optimality is attained unless the optimum satisfies special degenerate properties. To support this, I provide general conditions that ensure global optimality. In all cases, including problems with more than 3 features, I show that all stationary points possess marginal distributions of the features that match those of the empirical data. This is a nice property that suggests that even when a local optimum is found, it is still a reasonable solution.

In practice these simple conditions allow a user to check if there is a chance that the local optimum that they have encountered may not be globally optimal. The user can verify if the stationary point matches the conditions placed on the parameters at spurious local optima, and if that is not the case, then the point must be globally optimal.

In the case of supervised learning with the naïve Bayes model, it is well known that the log likelihood is concave. Hence, there are no sets of parameters that could be locally optimal but not globally optimal. In practice, however, the parameters are chosen based on a set of training data. Depending on how well that sample represents the underlying distribution, the parameters that best describe the sample could lead to a different classifier than the parameters that best describe the underlying distribution. The portion of the classification error due to this effect is called the estimation error. If we are unlucky, and the observations in the sample happen to be very unusual for the underlying distribution, then the classifier we select may not generalize well beyond our sample. So, even though there is only a single parameterization that is optimal, it is optimal for the finite training sample, and not necessarily for the distribution.

In general, as our sample gets larger, by the central limit theorem the frequency of any event in this sample will tend towards its true probability in the underlying distribution. Hence, as our sample becomes larger, the probability of choosing an unrepresentative sample will become smaller. For practitioners, it would be useful to know if there is a certain sample size that will guarantee that with high probability the classifier selected will have performance close to that of the optimal classifier. This way, they could control their

sample size to make sure that the probability of encountering errors beyond a specified threshold is sufficiently small. Or, if the amount of training data available is limited, it would allow them to assess the probability that they are making errors of a certain magnitude.

I present a strategy for determining the exact minimum number of training observations needed for agnostic probably approximately correct learning using the naïve Bayes classifier and maximum likelihood. Next, I complete the first steps for implementing this strategy, in the restricted case of the single feature naïve Bayes classifier. This leads us to some interesting insights into how the estimation error changes as the underlying distribution changes. It also shows us the regions where the distributions with the greatest probability of breaching our error threshold exist.

This technique, once completed in full, would allow a user to find out exactly how many training observations they must use to probabilistically reach their desired accuracy. In practice, our results show that for certain problems and accuracy levels this could be 93% lower than the generally known bound. This means that we could greatly reduce the burden of storage and computational costs, while only sacrificing accuracy up to a specified amount. Alternatively, if the amount of data is fixed, this technique allows the user to assess with what probability they are guaranteed to reach a specified accuracy target.

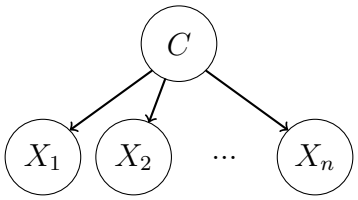
This thesis is structured as follows. Chapter 2 provides the reader with the necessary background knowledge, and discusses related work. Chapter 3 illustrates my work describing the stationary points of the log likelihood of the naïve Bayes model in the unsupervised setting. Chapter 4 details my work towards determining the minimum number of observations required for agnostic probably approximately correct classification using maximum likelihood and the naïve Bayes model. Finally, Chapter 5 concludes this thesis and provides some interesting directions for future work.

Chapter 2

Background and Related Work

The naïve Bayes model is a very simple type of Bayesian network that has been in use for over half a century [27]. As mentioned in Chapter 1, it is used in a variety of contexts, such as text classification, spam filtering, or clustering gene expressions, often as a baseline. It uses a single class variable, which I will denote C, and multiple features that are conditionally independent given the class, which I will denote $X_1, X_2, ..., X_n$. We will assume that the number of features, n, is finite.

Figure 2.1: The naïve Bayes model



The naïve Bayes model represents distributions which factorize as follows:

$$P(C = c, X_1 = x_1, ..., X_n = x_n) = P(C = c) \prod_{j=1}^{n} P(X_j = x_j | C = c)$$

We assume that we are attempting to model a distribution $\mathcal{D}$. $P_{\mathcal{D}}(event)$ will denote the probability of a certain event within the distribution. To construct the model, we will be using a training sample $\mathcal{S}$, which will consist of m i.i.d. observations from $\mathcal{D}$. The empirical distribution, or sample distribution, $P_{\mathcal{S}}(event)$ will describe the frequency of the specified event within the sample. For example, $P_{\mathcal{S}}(X = x) = \frac{\#\text{ of obs s.t. } X=x}{\#\text{ of obs}}$.

I will also use the term marginal distribution, which refers to the distribution of a subset of the variables, regardless of the values of the other variables. For example, if the system has two binary features, then $P_{\mathcal{S}}(X_1 = 1) = P_{\mathcal{S}}(X_1 = 1, X_2 = 0) + P_{\mathcal{S}}(X_1 = 1, X_2 = 1)$.

One of the most common techniques for training a naïve Bayes model on a sample $\mathcal{S}$ is finding the parameters that maximize the likelihood of $\mathcal{S}$. This is equivalent to finding the parameters that maximize the log likelihood of $\mathcal{S}$. Let the model parameters be denoted Ω, and the probability of a certain event as calculated using the model parameters be denoted $P_\Omega(event)$. In the supervised scenario, when the training data is labeled, the log likelihood of $\mathcal{S}$ is:

$$\begin{aligned}\mathcal{L}(\mathcal{S};\Omega) &= \sum_{(c,\mathbf{x})\in\mathcal{S}} log P_\Omega(C = c, \mathbf{X} = \mathbf{x}) \\ &= \sum_{(c,\mathbf{x})\in\mathcal{S}} log P_\Omega(C = c) \prod_{j=1}^{n} P_\Omega(X_j = x_j | C = c)\end{aligned}$$

In the unsupervised scenario, when the training data does not include the class variable C, which we assume to have domain $\mathcal{C}$, the log likelihood of $\mathcal{S}$ is:

$$\begin{aligned}\mathcal{L}(\mathcal{S};\Omega) &= \sum_{\mathbf{x}\in\mathcal{S}} log P_\Omega(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}\in\mathcal{S}} log \sum_{c\in\mathcal{C}} P_\Omega(C = c, \mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}\in\mathcal{S}} log \sum_{c\in\mathcal{C}} P_\Omega(C = c) \prod_{j=1}^{n} P_\Omega(X_j = x_j | C = c)\end{aligned}$$

So, to train our naïve Bayes model using maximum likelihood, we must optimize the likelihood over our parameter space. This will yield the parameters that have the highest likelihood of having generated the dataset. We can then classify new, unlabeled observations $\mathbf{x}$ using the rule:

$$y = \underset{c\in\mathcal{C}}{\operatorname{argmax}}\, P_\Omega(C = c | X_1 = x_1, ..., X_n = x_n) = \underset{c\in\mathcal{C}}{\operatorname{argmax}}\, P_\Omega(C = c) \prod_{j=1}^{n} P_\Omega(X_j = x_j | C = c)$$

Two popular techniques for optimizing the likelihood are gradient ascent [11] and expectation maximization [13]. To perform gradient ascent, we first initialize the model parameters to some point. The gradient of the objective function (in this case, the log likelihood) is then computed, and we change the parameters by moving them a small step in the direction of the gradient, such that the objective is improved. We continue this process until we reach a point that is locally optimal within the parameter space.

Expectation maximization is another optimization technique that is often used when the model objective depends on unobserved features. To perform expectation maximization, we begin by initializing the model parameters to some point. We then repeat the following two steps until convergence to a point that is locally optimal within the parameter space:

- Perform an expectation step, in which we calculate a function of the parameters that represents the expected value of the log likelihood based on the distribution of the latent features given the observed features and the current parameter settings.
- Perform a maximization step, in which we determine the model parameters that would maximize the expected log likelihood found in the previous step. We then set the parameters to be the parameters found in this step.

Gradient ascent and expectation maximization are two very popular and commonly used optimization techniques. However, gradient ascent and expectation maximization are only guaranteed to converge to a local maximum. Hence, if the objective is concave, then we know that any point that these methods converge to must be a global maximum. However, if the function is not quasi-concave, then these methods may converge to a spurious local maximum, which may have likelihood that is much lower than that at the global optimum. Furthermore, without any understanding of the shape of the objective function, and the nature of its local optima, we cannot determine, for any stationary point that is reached, whether or not it might be a spurious local maximum.

In the case of supervised learning using the naïve Bayes model, it is a simple exercise to show that the log likelihood is concave, so there are no risks of spurious maxima. In the case of unsupervised naïve Bayes learning, however, the likelihood is not concave, and hence it is valuable to understand the nature of any stationary points. Such a characterization would allow us to know the conditions under which a stationary point might not be globally optimal and how far from optimal it may be.

The problem of characterizing the stationary points of the likelihood has recently attracted considerable attention in the machine learning community. While I am not aware of such work in the case of unsupervised naïve Bayes, various characterizations exist for

other latent variable models such as mixtures of Gaussians and matrix completion. For instance, [37] provides a global analysis of Expectation Maximization for mixtures of two Gaussians. Furthermore, [23] shows that arbitrarily bad local optima exist in the likelihood of mixtures of at least three Gaussians and [1] shows that for certain data samples the number of local optima can be unbounded. In contrast, for the problem of completing a positive semidefinite matrix based on incomplete measurements, it was recently shown that no spurious local optima exist (i.e., all local optima are global optima) despite the non-convex nature of the objective [17, 7].

In a different line of work, researchers considered alternatives to maximum likelihood to obtain provable guarantees about the estimation of latent variable models. For instance, with the method of moments, it is possible to reliably estimate the underlying parameters of mixtures of Gaussians [6, 28, 21], latent Dirichlet allocation [2] and other latent variable models [3] with sufficient data and suitable minor conditions. However, techniques based on maximum likelihood tend to be more data efficient and therefore often remain the preferred choice of practitioners.

Even in cases, such as supervised naïve Bayes, where it is known that the log likelihood is concave, there is still a question of how much data is needed to ensure that our model will probabilistically achieve our desired accuracy levels. It is important that we have enough data in our training sample so that it can replicate the underlying distribution sufficiently well. However, there is also a trade off in terms of computational and storage complexity. In practice, we generally need to accept a certain probability of having some error in our learning tasks. Hence it is important to consider how much error is acceptable for the user, and how much of our resources we are willing to spend to mitigate that error.

To discuss whether or not a classifier is accurate enough, we will first define agnostic PAC learnability. A hypothesis class $\mathcal{H}$ is agnostic PAC learnable, if, $\forall \epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$, there exists an integer m and a learning algorithm such that, if our dataset contains at least m i.i.d. observations from the distribution $\mathcal{D}$, then with probability at least $1-\delta$, the algorithm will choose a classifier $h \in \mathcal{H}$ such that:

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

where $L_{\mathcal{D}}(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$.

Intuitively, this means that as long as our sample is large enough, there is a sufficiently high probability that the algorithm will choose a classifier that performs to within at least ϵ of the optimal classifier in the hypothesis class over the distribution.

We must also consider several other definitions:

Shattering: A hypothesis class $\mathcal{H}$ with domain $\mathcal{X}$ shatters a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$.

VC Dimension: The VC-dimension of a hypothesis class $\mathcal{H}$ with domain $\mathcal{X}$ is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.

There are several well known theorems that provide bounds on the sample complexity for agnostic PAC learning. One of the most well known is the Fundamental Theorem of Statistical Learning [9], which states:

Let $\mathcal{H}$ be a hypothesis class from a domain to $\{0, 1\}$ with VC dimension $d < \infty$. Then, using the 0-1 loss function, there exists a constant c such that $\mathcal{H}$ is agnostic PAC learnable to within ϵ with certainty $1 - \delta$ with sample complexity $m \leq c\frac{d+log(1/\delta)}{\epsilon^2}$.

In the case of the naïve Bayes classifier with n features, we are mapping from the domain $\{0, 1\}^n$ to $\{0, 1\}$. Therefore, any set that is shattered by this hypothesis space contains no more than 2^n elements, and so the VC dimension is finite and is at most 2^n. Hence, the Fundamental Theorem of Statistical Learning does apply in the case of the naïve Bayes classifier.

To say that a certain hypothesis class is agnostic PAC learnable means that there exists a learning algorithm such that with enough data we can ensure, with confidence $\geq 1 - \delta$, that our loss will be within ϵ of that of the optimal hypothesis. However, though we know that such an algorithm exists, we may not know whether it holds for our particular algorithm. We don't even know whether the algorithm that does satisfy agnostic PAC learnability is computationally tractable.

So, we know that there is a bound on the sample complexity for some learning algorithm, and we have an idea of how it will change as the accuracy and confidence thresholds ϵ and δ change. However, without knowing which algorithm to use, this result cannot be put into practice.

Another important consideration for those who wish to use this Theorem in practice is that it does not tell us exactly how many observations we need to use to reach our accuracy goals. This result gives us the order of the sample complexity, but unless we find exact values for the constant (for example, by examining the proof of the theorem) and find and incorporate any lower order terms that may have been dropped, we can't find an exact number of observations needed. This theorem is more useful in terms of giving a broad idea of the magnitude of data needed, for example saying that if we want to halve our accuracy threshold ϵ, then we would expect to need four times as much data to achieve the same confidence.

Furthermore, the bound identified in this Theorem is not tight. Depending on the hypothesis class and algorithm being used, it is possible that there is a tighter bound.

We will also define empirical risk minimization (ERM):

Empirical Risk Minimization: ERM is the learning algorithm that chooses the hypothesis that minimizes the loss function over the training sample.

Another well known theorem provides an empirical bound for finite hypothesis classes [9]:

If $\mathcal{H}$ is a finite hypothesis class from a domain to $\{0,1\}$, then using the ERM algorithm the class is agnostic PAC learnable with sample complexity $m \leq \lceil \frac{2log(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$.

Note that though the parameter space for the naïve Bayes model is continuous, there are regions of the parameter space that lead to identical classifier outcomes over the domain. In fact, in our case there are only a finite number of different classifiers, since the classifiers are functions that map from the finite feature domain, $\{0,1\}^n$, to the finite class domain, $\{0,1\}$. Hence, the hypothesis class is finite.

This theorem is more useful for our goal than the previous, since it provides an exact value that can be calculated. It can give an upper bound after which we can be sufficiently certain that our accuracy threshold is achieved through ERM. It is not, however, a tight bound. And in the case of naïve Bayes (and many other models), the fact that $|\mathcal{H}|$ grows exponentially with n, and that ϵ is generally quite small, means that this upper bound will usually be quite large. However, by considering and exploiting the specific form of the model and the classifiers created, I hope to find the exact number of observations that is required, rather than an upper bound. Exploiting the structure of the hypothesis class could lead to a precise sample complexity that is much lower than the upper bound found in the general case of a finite hypothesis class.

There has been significant research for many years into sample complexity and PAC learnability of differerent hypothesis classes. However, many of these results have been related to the order of the sample complexity, both for the upper bound [25, 26] and the lower bound [15]. A further direction of study has been to see whether the order of sample complexity bounds can be tightened for specific models. For example, [4] tightens the lower bounds for several classes such as linear halfspaces, [5] improves the upper bounds for neural nets, and [32] shows a tight bound for sub-Gaussian distributions. However, these works are all focused on the order of the sample complexity, so while they give a good idea of how the data needs to grow in order to probabilistically achieve good accuracy, they cannot be used to give exact values of how many training observations are needed.

There has also been some work in bounding the sample complexity using quantities other than the ones normally used in learning theory, such as VC dimension and accuracy and confidence parameters. For example, [20] provides bounds on the sample complexity of Bayesian learning using metrics derived from information theory. [24] describes extending sample complexity results to reinforcement learning, where additional factors such as sampling models come into play.

So, most work in the field so far has been related to determining and tightening the bounds on the order of the sample complexity in various settings. I have not found any other work where the exact number of observations needed is calculated. Knowing the exact sample size would be incredibly valuable in practice. For simplicity, I begin with the simplest setting of one of the simplest models, the single feature naïve Bayes classifier, but I design my approach in such a way that it can hopefully be extended to include more features or to other models.

Chapter 3

Characterizing the Optima of the Likelihood for Unsupervised Naïve Bayes

3.1 Problem Description

In this section I focus on the problem of unsupervised classification using the naïve Bayes model. As discussed in Chapter 1, this is something that is often done in a variety of contexts.

A common approach for unsupervised training of a naïve Bayes model for clustering consists of maximizing the likelihood of the data. However, since the optimization objective is non-concave, popular algorithms such as gradient ascent and expectation maximization [14] may get stuck in local optima. Despite the naïve Bayes model being one of the oldest and most basic models in machine learning, there is no characterization of the optima of the likelihood objective. This is a major gap in the theory of the naïve Bayes model that translates into some uncertainty about the reliability of the clusterings found in practice.

I provide a first analysis of the stationary points of the unsupervised naïve Bayes likelihood. For up to three binary features, I show that global optimality is attained unless the point satisfies special degenerate properties. To support this, I provide general conditions that ensure global optimality. In all cases, including problems with more than 3 features, I show that all stationary points possess marginal distributions of the features that match those of the empirical data. This is a nice property that suggests that even when a local optimum is found, it is still a reasonable solution.

This characterization is useful in practice to verify the possibility of having achieved a local optimum after choosing parameters using a gradient based technique. By simply comparing the point reached to the forms of the spurious local optima as specified in the upcoming theorems, we can assess whether the point is globally optimal or not.

In Section 3.2, I introduce the notation and define the model and objective function. In Section 3.3 I discuss my contributions and their significance. In Sections 3.4-3.7 I prove the theorems stated in Section 3.3.

3.2 Definitions and Notation

The naïve Bayes model is a very simple type of Bayesian network with one class variable and multiple features that are conditionally independent given the class. While the naïve Bayes model is typically used for classification in supervised learning, it can also be used for clustering by unsupervised learning where each class corresponds to a different cluster.

For a class variable C with domain $\mathcal{C} = \{1, 2, ..., |\mathcal{C}|\}$ and features $\mathbf{X} = (X_1, X_2, ..., X_n)$, the naïve Bayes model represents a joint distribution that factorizes as follows:

$$P_\Omega(C = c, X_1 = x_1, ..., X_n = x_n) = P_\Omega(C = c) \prod_{j=1}^{n} P_\Omega(X_j = x_j | C = c)$$

where x_j is the observed value of feature X_j, and $P_\Omega(Event)$ is the likelihood of $Event$ based on the model parameters Ω. The classifier assigns class y to an observation $\mathbf{x}$ based on the rule:

$$y = \operatorname*{argmax}_{c \in \mathcal{C}} P_\Omega(C = c | X_1 = x_1, ..., X_n = x_n) = \operatorname*{argmax}_{c \in \mathcal{C}} P_\Omega(C = c) \prod_{j=1}^{n} P_\Omega(X_j = x_j | C = c)$$

Without loss of generality, we assume binary features[1]. Let us define some important symbols and functions.

Let $\Omega = (\theta, \Phi)$ denote the parameters of the naïve Bayes model, where θ_c is the probability of class c (i.e., $\theta_c = P_\Omega(C = c)$) and $\phi_{c,j}$ is the probability that feature j is 0 given class c (i.e., $\phi_{c,j} = P_\Omega(X_j = 0 | C = c)$). Note that $\theta_{|\mathcal{C}|} = 1 - \theta_1 - ... - \theta_{|\mathcal{C}|-1}$. In this work,

[1]Categorical features can always be converted to binary features. For instance, the conversion of categorical features to binary features is common practice in industrial recommender systems.

we consider the problem of learning the parameters of the naïve Bayes model by maximum log likelihood. When the class is unobserved, the log likelihood of a sample $\mathcal{S}$ is:

$$\begin{aligned}\mathcal{L}(\mathcal{S};\Omega) &= \sum_{\mathbf{x}\in\mathcal{S}} log P_\Omega(\mathbf{X}=\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{S}} log \sum_{c\in C} P_\Omega(C=c,\mathbf{X}=\mathbf{x}) \\ &= \sum_{\mathbf{x}\in\mathcal{S}} log \sum_{c\in C} P_\Omega(C{=}c) \prod_{j=1}^{n} P_\Omega(X_j{=}x_j|C{=}c) = \sum_{\mathbf{x}\in\mathcal{S}} log \sum_{c\in C} \theta_c \prod_{j=1}^{n} \phi_{c,j}^{1-x_j}(1{-}\phi_{c,j})^{x_j}\end{aligned}$$

Note that the above objective is not concave in Ω due to the sum over the hidden classes. For example, consider the case of a single binary feature and a single binary class. If the empirical probabilities are $P_\mathcal{S}(X=0)=\frac{5}{16}$ and $P_\mathcal{S}(X=1)=\frac{11}{16}$, then the points $(\theta,\phi_{0,1},\phi_{1,1})=(\frac{1}{4},\frac{1}{2},\frac{1}{4})$ and $(\theta,\phi_{0,1},\phi_{1,1})=(\frac{3}{4},\frac{1}{4},\frac{1}{2})$ both generate distributions that exactly match the sample empirical distribution, and have log likelihood -0.2697. If we take the point $(\theta,\phi_{0,1},\phi_{1,1})=(\frac{1}{2},\frac{3}{8},\frac{3}{8})$ that is midway between them, however, it generates a distribution with log likelihood -0.2734. This simple construction shows that the log likelihood may be non-concave.

As a result, algorithms such as gradient ascent and expectation maximization could be subject to local optima. While EM optimizes a lower bound of the log likelihood at each step, it converges to a stationary point (i.e., point at which the derivative is zero) of the log likelihood [36]. Hence, my analysis of the stationary points of the log likelihood will help practitioners to understand when the parameters found by gradient ascent and EM are local and global optima.

I will also use $P_\mathcal{S}(Event)$ to denote the probability of $Event$ based on its frequency within the sample $\mathcal{S}$. For example, $P_\mathcal{S}(\mathbf{X}=\mathbf{x})=\frac{\#\text{ of observations with }\mathbf{X}=\mathbf{x}}{\#\text{ of observations in }\mathcal{S}}$. Define, $\lambda_\mathbf{x}=\frac{P_\mathcal{S}(\mathbf{X}=\mathbf{x})}{P_\Omega(\mathbf{X}=\mathbf{x})}$, the ratio between the probability of $(\mathbf{X}=\mathbf{x})$ based on the sample and the parameters.

It will also be more convenient to work with the average log likelihood $A\mathcal{L}(\mathcal{S};\Omega)$, which is equivalent to the log likelihood objective up to a constant scaling factor:

$$\begin{aligned}A\mathcal{L}(\mathcal{S};\Omega) = \frac{1}{\text{size of }\mathcal{S}}\mathcal{L}(\mathcal{S};\Omega) &= \frac{1}{\text{size of }\mathcal{S}}\sum_{\mathbf{x}\in\mathcal{S}} log P_\Omega(\mathbf{X}=\mathbf{x}) \\ &= \sum_{\mathbf{x}\in\{0,1\}^n} P_\mathcal{S}(\mathbf{X}=\mathbf{x}) log P_\Omega(\mathbf{X}=\mathbf{x})\end{aligned} \tag{3.1}$$

Hence, to perform clustering by maximum likelihood with the naïve Bayes model, we must find the parameters Ω that maximize the objective (3.1).

3.3 Contributions

This work is focused on the case of unsupervised classification using the naïve Bayes model. As discussed in Section 1, this is something that is often done in a variety of contexts for clustering.

We suppose that the observed data shows the values for all of the features, which are assumed to be binary, but does not show which class each observation belongs to. We then want to find the naïve Bayes parameters that will maximize the log likelihood of the observations as defined in Equation (3.1). As discussed in the previous section, this optimization problem is non-concave due to the sum over the classes within the log portion of the equation, however I show that the stationary points still satisfy some nice properties.

For this work, I have restricted the scope to include only points that are stationary w.r.t. each of the parameters. It is important to note that in practice, this optimization problem is constrained since each parameter must be between 0 and 1. In optimization, a point that is on the boundary of the feasible region may have a non-zero derivative in the direction normal to that boundary. Hence, this work shows how any local optima must either be on the boundary of the parameter space (i.e. at least one of the parameters is 0 or 1), or they must belong to the forms specified. An analysis of the form of local optima on the boundary of the parameter space would require use of a more general optimality condition, such as KKT conditions, and would be a promising direction for future research.

I will use the language "interior of the parameter space" to represent the portion of the parameter space where no parameter is 0 or 1.

My first contribution (formalized in Theorem 1 in Section 3.4) reveals an interesting property of the log likelihood of the unsupervised naïve Bayes model with any number of binary features. At any point that is stationary w.r.t. all of the parameters (i.e. any local optimum in the interior of the parameter space), the marginal distribution of any feature must match its empirical distribution. Hence, if Ω is a stationary point of the log likelihood for a sample $\mathcal{S}$, then $P_\Omega(X_j = 0) = P_\mathcal{S}(X_j = 0)\ \forall j$. Thus, even when the naïve Bayes estimate does not match the distribution seen in the sample, it still satisfies the nice property of having the correct distribution for any individual feature.

My second contribution (formalized in Theorem 2 in Section 3.5), relates to the log likelihood of the unsupervised naïve Bayes model with a single binary feature. I show that any point that is stationary w.r.t. all of the parameters (i.e. any local optimum in the interior of the parameter space) is globally optimal, and in fact that it exactly replicates the empirical distribution. Hence, if Ω is a stationary point of the log likelihood for a sample $\mathcal{S}$, then $P_\Omega(\mathbf{X} = \mathbf{x}) = P_\mathcal{S}(\mathbf{X} = \mathbf{x})\forall \mathbf{x}$.

My third contribution (formalized in Theorem 3 in Section 3.6) consists of a description of the stationary points of the log likelihood of the unsupervised naïve Bayes model with two binary features. It shows that any point Ω that is stationary w.r.t. all of the parameters (i.e. any local optimum in the interior of the parameter space) is either globally optimal and matches the distribution of the sample $\mathcal{S}$, or has a specific form. Furthermore, I show that if the features are independent in the empirical distribution, then all stationary points in the interior of the parameter space are globally optimal. The possible stationary points are summarized in Table 3.1.

Table 3.1: Stationary points of the log likelihood of the unsupervised naïve Bayes model with two features in the interior of the parameter space

Stationary points	**Optimality**
$\phi_{c,j} = P_{\mathcal{S}}(X_j = 0)$ $\forall j, \forall c \in \mathcal{C}$ s.t. $\theta_c > 0$	Globally optimal if features are independent in empirical distribution
All others	Always globally optimal, and $P_{\Omega}(\mathbf{x}) = P_{\mathcal{S}}(\mathbf{x})$ $\forall \mathbf{x}$

My fourth contribution (formalized in Theorem 4 in Section 3.7) consists of a description of the stationary points of the log likelihood for the 3 binary feature case of the naïve Bayes model. If a point is stationary w.r.t. all of the parameters (i.e. if it is a local optimum in the interior of the parameter space), then either it is globally optimal and it matches the distribution of the sample $\mathcal{S}$, or it takes one of 3 forms. Furthermore, I show that all stationary points in the interior of the parameter space are globally optimal if the features are independent in the empirical distribution. The stationary points are summarized in Table 3.2.

For up to three features, I show that if the features of the empirical distribution are independent, then any stationary point of the likelihood in the interior of the parameter space will be a global optimum, and no spurious local optima are possible for classification. So, if the features are independent, then algorithms such as gradient ascent and expectation maximization can be used without any risk of converging to suboptimal solutions in the interior of the parameter space.

The three feature case (formalized in Theorem 4) is particularly interesting because for a system with n features, the derivatives are formulated as a system of polynomials of order $n - 1$. So in the 1 and 2 feature cases, we have conditions for stationarity that can be solved linearly. In the 3 feature case, however, I show that we can create a decoupled system of equations for each class, and that by doing novel combinations of these equations,

Table 3.2: Stationary points of the log likelihood of the unsupervised naïve Bayes model with three features in the interior of the parameter space

Stationary Points	**Optimality**
$\phi_{c,j} = P_{\mathcal{S}}(X_j = 0)\ \forall c \in \mathcal{C}$ s.t. $\theta_c > 0$ $\forall j$ except at most one	Globally optimal if features are independent in empirical distribution
$\exists j^*$ s.t. $\phi_{c,j} = P_{\mathcal{S}}(X_j = 0 \mid X_{j^*} = 0)$ $\forall j \neq j^*, \forall c \in \mathcal{C}$ s.t. $\theta_c \phi_{c,j^*} > 0$	Globally optimal if features are independent in empirical distribution
$\exists j^*$ s.t. $\phi_{c,j^*} = P_{\mathcal{S}}(X_{j^*} = 0)$ $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0$ and $P_\Omega(\bigwedge_{j \neq j^*} X_j = x_j) = P_{\mathcal{S}}(\bigwedge_{j \neq j^*} X_j = x_j)$	Globally optimal if features are independent in empirical distribution
All others	Always globally optimal, and $P_\Omega(\mathbf{x}) = P_{\mathcal{S}}(\mathbf{x})\ \forall \mathbf{x}$

we can still generate linear systems. This observation, along with some empirical testing, encourages us to believe that these results could be extended to an arbitrary number of features.

My results show that even though spurious local optima can occur when the features are not independent, they must take specific forms. This provides a simple test to confirm whether the stationary point is a local optimum: we just need to check whether it has one of the specific forms described in the above tables. If not, then we can be certain that the stationary point is a global optimum.

For example, in the two feature case, if we arrive at a stationary point Ω, then we only need to check for each feature whether $\phi_{c,j}$ is the same for each class c with $\theta_c > 0$, or if some of the parameters are 0 or 1. In this case, we may be at a local optimum. Otherwise, we can guarantee that we have attained a global optimum. Similar conditions can be drawn for the 3 feature case. Hence, if our stationary point is not globally optimal, we will be able to detect this by comparing it to the forms described in the tables above.

3.4 Estimated Marginals Match Empirical Marginals

Theorem 1. *Suppose we estimate a distribution with binary features and an unobserved class using the naïve Bayes formulation. If Ω is a stationary point of the likelihood (Eq. 3.1) in the interior of the parameter space for a sample $\mathcal{S}$, then the marginal probabilities for*

each feature as estimated using Ω match those of the sample:

$$P_\Omega(X_j = x_j) = P_\mathcal{S}(X_j = x_j) \;\; \forall j, \forall \mathbf{x} \in \{0,1\}^n$$

Proof. Note that we can write:

$$\begin{aligned} A\mathcal{L}(\mathcal{S}, \Omega) &= \sum_{\mathbf{x}\in\{0,1\}^n} P_\mathcal{S}(\mathbf{X} = \mathbf{x}) log P_\Omega(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}\in\{0,1\}^n} P_\mathcal{S}(\mathbf{X} = \mathbf{x}) log \sum_{c\in\mathcal{C}} \theta_c \prod_{j\in\{1,\dots,n\}} \phi_{c,j} \\ &= \sum_{\mathbf{x}\in\{0,1\}^n} P_\mathcal{S}(\mathbf{X} = \mathbf{x}) \\ &\quad \cdot log(\sum_{c\in|\mathcal{C}|-1} \theta_c \prod_{j\in\{1,\dots,n\}} \phi_{c,j} + (1 - \theta_1 - \dots - \theta_{|\mathcal{C}|-1}) \prod_{j\in\{1,\dots,n\}} \phi_{|\mathcal{C}|,j}) \end{aligned}$$

If Ω is a stationary point in the interior of the parameter space, then by setting the derivatives of (3.1) to 0, we see that it must satisfy the following equations $\forall j$ and $\forall c \in \mathcal{C}$:

$$0 = \frac{\delta A\mathcal{L}(\mathcal{S}, \Omega)}{\delta \theta_c} = \sum_{\mathbf{x}\in\{0,1\}^n} \lambda_\mathbf{x} [P_\Omega(\mathbf{x}|c) - P_\Omega(\mathbf{x} \mid C = |\mathcal{C}|)] \tag{3.2}$$

$$0 = \frac{\delta A\mathcal{L}(\mathcal{S}, \Omega)}{\delta \phi_{c,j}} = \theta_c \sum_{\mathbf{x}\in\{0,1\}^n} \lambda_\mathbf{x} \frac{(-1)^{x_j} P_\Omega(\mathbf{x}|c)}{\phi_{c,j}^{1-x_j}(1 - \phi_{c,j})^{x_j}} \tag{3.3}$$

If we rearrange Equation (3.2), we get that:

$$\sum_{\mathbf{x}\in\{0,1\}^n} \lambda_\mathbf{x} P_\Omega(\mathbf{x}|\tilde{c}) = \sum_{\mathbf{x}\in\{0,1\}^n} \lambda_\mathbf{x} P_\Omega(\mathbf{x}|\hat{c}) \;\; \forall \tilde{c}, \hat{c} \in \mathcal{C}$$

Combining this with the fact that $\sum_{c \in \mathcal{C}} \theta_c = 1$ reveals:

$$\begin{aligned}
\sum_{\mathbf{x} \in \{0,1\}^n} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|\tilde{c}) &= \sum_{c \in \mathcal{C}} \theta_c \sum_{\mathbf{x} \in \{0,1\}^n} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|\tilde{c}) \\
&= \sum_{c \in \mathcal{C}} \theta_c \sum_{\mathbf{x} \in \{0,1\}^n} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c) \\
&= \sum_{\mathbf{x} \in \{0,1\}^n} \lambda_{\mathbf{x}} \sum_{c \in \mathcal{C}} \theta_c P_\Omega(\mathbf{x}|c) \\
&= \sum_{\mathbf{x} \in \{0,1\}^n} P_{\mathcal{S}}(\mathbf{x}) \\
&= 1
\end{aligned} \tag{3.4}$$

If we rearrange Equation (3.3), we find that, $\forall c$ s.t. $\theta_c > 0$:

$$\begin{aligned}
\sum_{\substack{\mathbf{x} \in \{0,1\}^n \\ x_j = 1}} \lambda_{\mathbf{x}} \frac{P_\Omega(\mathbf{x}|c)}{(1 - \phi_{c,j})} &= \sum_{\substack{\mathbf{x} \in \{0,1\}^n \\ x_j = 0}} \lambda_{\mathbf{x}} \frac{P_\Omega(\mathbf{x}|c)}{\phi_{c,j}} \\
\sum_{\substack{\mathbf{x} \in \{0,1\}^n \\ x_j = 1}} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c) &= \frac{(1 - \phi_{c,j})}{\phi_{c,j}} \sum_{\substack{\mathbf{x} \in \{0,1\}^n \\ x_j = 0}} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c)
\end{aligned} \tag{3.5}$$

By combining Equations (3.4) and (3.5), we see that if $\theta_c > 0$, then for any feature j:

$$\begin{aligned}
1 &= \sum_{\mathbf{x} \in \{0,1\}^n} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c) \\
&= \sum_{\substack{\mathbf{x} \in \{0,1\}^n \\ x_j = 0}} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c) + \sum_{\substack{\mathbf{x} \in \{0,1\}^n \\ x_j = 1}} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c) \\
&= (1 + \frac{(1 - \phi_{c,j})}{\phi_{c,j}}) \sum_{\substack{\mathbf{x} \in \{0,1\}^n \\ x_j = 0}} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c) \\
\Rightarrow \phi_{c,j} &= \sum_{\substack{\mathbf{x} \in \{0,1\}^n \\ x_j = 0}} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c)
\end{aligned} \tag{3.6}$$

Finally, note that the naïve Bayes estimate of the marginal probability $P_\Omega(X_j = 0)$ is of

the form:

$$\begin{aligned}\sum_{c\in\mathcal{C}}\theta_c\phi_{c,j} &= \sum_{c\in\mathcal{C}}\theta_c \sum_{\substack{\mathbf{x}\in\{0,1\}^n\\ x_j=0}} \lambda_{\mathbf{x}} P_\Omega(\mathbf{x}|c)\\ &= \sum_{\substack{\mathbf{x}\in\{0,1\}^n\\ x_j=0}} \lambda_{\mathbf{x}} \sum_{c\in\mathcal{C}} \theta_c P_\Omega(\mathbf{x}|c)\\ &= \sum_{\substack{\mathbf{x}\in\{0,1\}^n\\ x_j=0}} P_{\mathcal{S}}(\mathbf{x})\\ &= P_{\mathcal{S}}(X_j = 0)\end{aligned}$$

Hence, at any stationary point, the marginal probability estimated using the naïve Bayes parameters matches the marginal probability from the sample. □

3.5 Optima in the One Feature Case

Theorem 2. *Suppose we estimate a distribution with a single binary feature and an unobserved class using the naïve Bayes formulation. Then every stationary point Ω of the likelihood (Eq. 3.1) in the interior of the parameter space for a sample $\mathcal{S}$ is globally optimal and exactly matches the empirical distribution:*

$$\sum_{\mathbf{x}\in\{0,1\}^n} P_{\mathcal{S}}(\mathbf{x}) log P_\Omega(\mathbf{x}) \geq \sum_{\mathbf{x}\in\{0,1\}^n} P_{\mathcal{S}}(\mathbf{x}) log P_{\tilde{\Omega}}(\mathbf{x}) \;\; \forall \tilde{\Omega} \in [0,1]^{2|\mathcal{C}|-1}$$

and

$$P_\Omega(\mathbf{x}) = P_{\mathcal{S}}(\mathbf{x}) \;\; \forall \mathbf{x}$$

Proof. If Ω is a stationary point in the interior of the parameter space for a single feature,

then choose a class $\hat{c}$ s.t. $\theta_{\hat{c}} > 0$ and Equation (3.3) simplifies to:

$$\begin{aligned}
0 &= \lambda_0 - \lambda_1 \\
\frac{P_{\mathcal{S}}(X_1 = 0)}{\sum_{c \in \mathcal{C}} \theta_c \phi_{c,1}} &= \frac{P_{\mathcal{S}}(X_1 = 1)}{\sum_{c \in \mathcal{C}} \theta_c (1 - \phi_{c,1})} \\
\frac{P_{\mathcal{S}}(X_1 = 0)}{\sum_{c \in \mathcal{C}} \theta_c \phi_{c,1}} &= \frac{1 - P_{\mathcal{S}}(X_1 = 0)}{\sum_{c \in \mathcal{C}} \theta_c (1 - \phi_{c,1})} \\
P_{\mathcal{S}}(X_1 = 0) \sum_{c \in \mathcal{C}} \theta_c (1 - \phi_{c,1}) &= (1 - P_{\mathcal{S}}(X_1 = 0)) \sum_{c \in \mathcal{C}} \theta_c \phi_{c,1} \\
P_{\mathcal{S}}(X_1 = 0) &= \sum_{c \in \mathcal{C}} \theta_c \phi_{c,1}
\end{aligned}$$

Hence, if there is a single feature, then the distribution defined by any stationary point matches the empirical distribution.

Furthermore, as shown in [18], the maximum value that the log likelihood can take is if the modeled distribution exactly matches the empirical distribution. Since we know that any stationary point matches the empirical distribution, it must also be globally optimal. □

3.6 Optima in the Two Feature Case

Theorem 3. *Suppose we estimate a distribution with two binary features and an unobserved class using the naïve Bayes formulation. Then the possible stationary points Ω of the likelihood (Eq. 3.1) that are in the interior of the parameter space for a sample $\mathcal{S}$ are described in the table below:*

Stationary Points	**Optimality**
$\phi_{c,j} = P_{\mathcal{S}}(X_j = 0)$ $\forall j, \forall c \in \mathcal{C}$ *s.t.* $\theta_c > 0$	*Globally optimal if features are independent in empirical distribution, spurious maximum otherwise*
All others	*Always globally optimal, and* $P_\Omega(\mathbf{x}) = P_{\mathcal{S}}(\mathbf{x})\ \forall \mathbf{x}$

Furthermore, if the features are independent then every stationary point in the interior of the parameter space will be globally optimal and exactly match the empirical distribution.

Proof. In the two feature case, Equation (3.3) simplifies to the following two equations

which must hold $\forall c$ s.t. $\theta_c > 0$:

$$0 = (\lambda_{00} - \lambda_{01} - \lambda_{10} + \lambda_{11})\phi_{c,2} + (\lambda_{01} - \lambda_{11}) \tag{3.7}$$
$$0 = (\lambda_{00} - \lambda_{01} - \lambda_{10} + \lambda_{11})\phi_{c,1} + (\lambda_{10} - \lambda_{11}) \tag{3.8}$$

Hence, if we take $\tilde{c}, \hat{c} \in \mathcal{C}$ s.t. $\theta_{\tilde{c}}, \theta_{\hat{c}} > 0$, then $\phi_{\tilde{c},2}$ and $\phi_{\hat{c},2}$ must satisfy the same linear equation, (3.7). Note that though λ_{**} depends on the θ and ϕ variables, it is not specific to a particular class (the denominator, $P_\Omega(\mathbf{X} = **)$, sums over all of the classes). So, the λ coefficients in those equations are the same for every c s.t $\theta_c > 0$. So either $\phi_{\tilde{c},2}$ and $\phi_{\hat{c},2}$ are equal, or the coefficients in Equation (3.7) are 0.

If the coefficients in (3.7) are 0, then this means that $\lambda_{01} = \lambda_{11}$ and $\lambda_{00} = \lambda_{10}$. However, this would also mean that $\lambda_{10} = \lambda_{11}$ by Equation (3.8).

Hence, if the coefficients in (3.7) are 0, then we have that $\lambda_{00} = \lambda_{01} = \lambda_{10} = \lambda_{11}$. Since the numerator of each of these expressions is a probability, they add up to 1. The same can be said for the denominators. Hence, we must have that $\lambda_{\mathbf{x}} = 1 \; \forall \mathbf{x}$. This implies that $P_{\mathcal{S}}(\mathbf{x}) = P_\Omega(\mathbf{x}) \forall \mathbf{x}$, and so the probabilities generated by the Bayesian approximation exactly match those in the empirical distribution.

If the coefficients are not 0, then we must have $\phi_{\tilde{c},1} = \phi_{\hat{c},1}$ and $\phi_{\tilde{c},2} = \phi_{\hat{c},2}$ $\forall \tilde{c}, \hat{c} \in \mathcal{C}$ s.t. $\theta_{\tilde{c}}, \theta_{\hat{c}} > 0$.

Then, using Theorem 1 we see that $\forall c$ s.t. $\theta_c > 0$, $\phi_{c,j} = P_{\mathcal{S}}(X_j = 0)$. Hence, the probabilities estimated using the naïve Bayes parameters will be the product of each of the true marginal probabilities. For example, if $\theta_1 > 0$:

$$\begin{aligned} P_\Omega(X_1 = 0, X_2 = 0) &= \sum_{c \in \mathcal{C}} \theta_c \phi_{c,1} \phi_{c,2} \\ &= \phi_{1,1}\phi_{1,2} \\ &= P_{\mathcal{S}}(X_1 = 0) P_{\mathcal{S}}(X_2 = 0) \end{aligned}$$

In this case we find that if the features are independent in the empirical distribution, then the modeled joint probabilities will match the empirical joint probabilities. Hence, if the features are independent in the empirical distribution, then any stationary point of the likelihood will be globally optimal. If the features have dependencies, we find that the only stationary points that might not be globally optimal are those where $\phi_{c,j} = P_{\mathcal{S}}(X_j = 0)$ $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0, \forall j \in \{1, 2\}$, which intuitively means that the features are not informative of the class.

Furthermore, note that for any underlying distribution, by using two classes and setting $\theta_1 = P_{\mathcal{S}}(X_1 = 0)$, $\phi_{1,1} = 1$, $\phi_{1,2} = P_{\mathcal{S}}(X_2 = 0 | X_1 = 0)$, $\phi_{2,1} = 0$, $\phi_{2,2} = P_{\mathcal{S}}(X_2 = 0 | X_1 = 1)$,

then the modeled distribution will exactly match the empirical distribution. Hence, there is always at least one point where the modeled distribution exactly matches the empirical distribution. If, at any stationary point, the distribution does not match the empirical distribution, then the likelihood will be lower than that at the point that is constructed to match, by [18], and so this must be a spurious maximum. Hence if the features are not independent, then stationary points of the form $\phi_{c,1} = P_{\mathcal{S}}(X_1 = 0)$ and $\phi_{c,2} = P_{\mathcal{S}}(X_2 = 0)$ $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0$ will be spurious local maxima. □

3.7 Optima in the Three Feature Case

Theorem 4. *Suppose we estimate a distribution with three binary features and an unobserved binary class using the naïve Bayes formulation. Then the possible stationary points Ω of the likelihood (3.1) in the interior of the parameter space for a sample $\mathcal{S}$ are described in the table below:*

Stationary Points	**Optimality**
$\phi_{c,j} = P_{\mathcal{S}}(X_j = 0)\ \forall c \in \mathcal{C}$ *s.t.* $\theta_c > 0$ $\forall j$ *except at most one*	*Globally optimal if features are independent in empirical distribution*
$\exists j^*$ *s.t.* $\phi_{c,j} = P_{\mathcal{S}}(X_j = 0 \mid X_{j^*} = 0)$ $\forall j \neq j^*, \forall c \in \mathcal{C}$ *s.t.* $\theta_c \phi_{c,j^*} > 0$	*Globally optimal if features are independent in empirical distribution*
$\exists j^*$ *s.t.* $\phi_{c,j^*} = P_{\mathcal{S}}(X_{j^*} = 0)$ $\forall c \in \mathcal{C}$ *s.t.* $\theta_c > 0$ *and* $P_\Omega(\bigwedge_{j \neq j^*} X_j = x_j) = P_{\mathcal{S}}(\bigwedge_{j \neq j^*} X_j = x_j)$	*Globally optimal if features are independent in empirical distribution*
All others	*Always globally optimal, and* $P_\Omega(\mathbf{x}) = P_{\mathcal{S}}(\mathbf{x})\ \forall \mathbf{x}$

Furthermore, if the features are independent then every stationary point in the interior of the parameter space will be globally optimal and exactly match the empirical distribution.

In the three feature case, the system of equations that a stationary point in the interior of the parameter space must satisfy, generated by derivatives as expressed in Equation (3.6) have the form:

$$1 = (\lambda_{000} - \lambda_{001} - \lambda_{010} + \lambda_{011})\phi_{c,2}\phi_{c,3} + (\lambda_{001} - \lambda_{011})\phi_{c,2} + (\lambda_{010} - \lambda_{011})\phi_{c,3} + \lambda_{011} \quad (3.9)$$

$$1 = (\lambda_{100} - \lambda_{101} - \lambda_{110} + \lambda_{111})\phi_{c,2}\phi_{c,3} + (\lambda_{101} - \lambda_{111})\phi_{c,2} + (\lambda_{110} - \lambda_{111})\phi_{c,3} + \lambda_{111} \quad (3.10)$$

$$1 = (\lambda_{000} - \lambda_{001} - \lambda_{100} + \lambda_{101})\phi_{c,1}\phi_{c,3} + (\lambda_{001} - \lambda_{101})\phi_{c,1} + (\lambda_{100} - \lambda_{101})\phi_{c,3} + \lambda_{101} \quad (3.11)$$

$$1 = (\lambda_{010} - \lambda_{011} - \lambda_{110} + \lambda_{111})\phi_{c,1}\phi_{c,3} + (\lambda_{011} - \lambda_{111})\phi_{c,1} + (\lambda_{110} - \lambda_{111})\phi_{c,3} + \lambda_{111} \quad (3.12)$$

$$1 = (\lambda_{000} - \lambda_{010} - \lambda_{100} + \lambda_{110})\phi_{c,1}\phi_{c,2} + (\lambda_{010} - \lambda_{110})\phi_{c,1} + (\lambda_{100} - \lambda_{110})\phi_{c,2} + \lambda_{110} \quad (3.13)$$

$$1 = (\lambda_{001} - \lambda_{011} - \lambda_{101} + \lambda_{111})\phi_{c,1}\phi_{c,2} + (\lambda_{011} - \lambda_{111})\phi_{c,1} + (\lambda_{101} - \lambda_{111})\phi_{c,2} + \lambda_{111} \quad (3.14)$$

The complete proof of this theorem will follow after I first prove three lemmas. The idea behind the proof is that even though the derivatives w.r.t. $\phi_{c,j}$ will no longer be linear in $\phi_{c,j}$, under certain conditions we can do variable elimination to cancel out the higher order terms, and generate a system that is linear. Then, for a class $\hat{c}$, we will have a system with 3 equations that are linear in the three variables $\phi_{\hat{c},j}$, and under certain conditions we can uniquely solve for them.

Lemma 1 will show an important property that will be useful in the subsequent lemmas. Lemma 2 will show that if the derivatives are such that we are unable to create a linear system, then if the features are independent every stationary point of the likelihood in the interior of the parameter space will be globally optimal. Lemma 3 will show that if we are able to create a linear system, then if the features are independent every stationary point of the likelihood in the interior of the parameter space is globally optimal.

Lemma 1. *If a point Ω satisfies $\phi_{\hat{c},j} = \phi_{\tilde{c},j}\ \forall \hat{c}, \tilde{c} \in \mathcal{C}$ s.t. $\theta_{\hat{c}}, \theta_{\tilde{c}} > 0$ for every feature j except at most one, then the point will be globally optimal if the features are independent.*

Proof. Start by noting that for a feature j, if $\phi_{\tilde{c},j} = \phi_{\hat{c},j}\ \forall \tilde{c}, \hat{c} \in \mathcal{C}$ s.t. $\theta_{\tilde{c}}, \theta\hat{c} > 0$, then by Theorem 1 $\phi_{\tilde{c},j} = P_{\mathcal{S}}(X_j = 0)$. Without loss of generality, assume this property is satisfied at a point Ω for every feature except perhaps feature 1. Then, $\forall \mathbf{x} \in \{0,1\}^n$:

$$\begin{aligned}
P_\Omega(\mathbf{x}) =& \sum_{c \in \mathcal{C}} \theta_c P_\Phi(\mathbf{x}|c) \\
=& \sum_{\substack{c \in \mathcal{C} \\ \theta_c > 0}} \theta_c \phi_{c,1}^{1-x_1}(1-\phi_{c,1})^{x_1} \phi_{c,2}^{1-x_2}(1-\phi_{c,2})^{x_2} \cdot \ldots \cdot \phi_{c,n}^{1-x_n}(1-\phi_{c,n})^{x_n} \\
=& P_{\mathcal{S}}(X_2 = x_2) \cdot \ldots \cdot P_{\mathcal{S}}(X_n = x_n) \sum_{\substack{c \in \mathcal{C} \\ \theta_c > 0}} \theta_c \phi_{c,1}^{1-x_1}(1-\phi_{c,1})^{x_1} \\
=& P_{\mathcal{S}}(X_2 = x_2) \cdot \ldots \cdot P_{\mathcal{S}}(X_n = x_n) P_{\mathcal{S}}(X_1 = x_1)
\end{aligned}$$

Hence, if the features are independent then $P_\Omega(\mathbf{x}) = P_{\mathcal{S}}(\mathbf{x})\ \forall \mathbf{x} \in \{0,1\}^n$. □

Lemma 2. *Suppose that at least 1 of the 3 pairs of Equations (3.9)(3.10), (3.11)(3.12), and (3.13)(3.14) cannot be combined to create an equation linear in $\phi_{c,j} \forall j$ and the features of the empirical distribution are independent. Then the distribution defined by any solution to this system of equations matches the empirical distribution and therefore is globally optimal.*

Proof. Based on Equation (3.6), as shown during the proof of Theorem 1, the following equations must hold at any stationary point in the interior of the parameter space $\forall c \in \mathcal{C}$

s.t. $\theta_c > 0$, $\forall j$:

$$1 = \sum_{\substack{\mathbf{x}\in\{0,1\}^n \\ x_j=0}} \frac{\lambda_{\mathbf{x}} P_\Phi(\mathbf{x}|c)}{\phi_{c,j}} = \sum_{\substack{\mathbf{x}\in\{0,1\}^n \\ x_j=1}} \frac{\lambda_{\mathbf{x}} P_\Phi(\mathbf{x}|c)}{(1-\phi_{c,j})}$$

In the 3 feature case, this yields the following system of equations, $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0$:

$$\begin{aligned}
1 =& \lambda_{000}\phi_{c,2}\phi_{c,3} + \lambda_{001}\phi_{c,2}(1-\phi_{c,3}) + \lambda_{010}(1-\phi_{c,2})\phi_{c,3} + \lambda_{011}(1-\phi_{c,2})(1-\phi_{c,3}) \\
1 =& \lambda_{100}\phi_{c,2}\phi_{c,3} + \lambda_{101}\phi_{c,2}(1-\phi_{c,3}) + \lambda_{110}(1-\phi_{c,2})\phi_{c,3} + \lambda_{111}(1-\phi_{c,2})(1-\phi_{c,3}) \\
1 =& \lambda_{000}\phi_{c,1}\phi_{c,3} + \lambda_{001}\phi_{c,1}(1-\phi_{c,3}) + \lambda_{100}(1-\phi_{c,1})\phi_{c,3} + \lambda_{101}(1-\phi_{c,1})(1-\phi_{c,3}) \\
1 =& \lambda_{010}\phi_{c,1}\phi_{c,3} + \lambda_{011}\phi_{c,1}(1-\phi_{c,3}) + \lambda_{110}(1-\phi_{c,1})\phi_{c,3} + \lambda_{111}(1-\phi_{c,1})(1-\phi_{c,3}) \\
1 =& \lambda_{000}\phi_{c,1}\phi_{c,2} + \lambda_{010}\phi_{c,1}(1-\phi_{c,2}) + \lambda_{100}(1-\phi_{c,1})\phi_{c,2} + \lambda_{110}(1-\phi_{c,1})(1-\phi_{c,2}) \\
1 =& \lambda_{001}\phi_{c,1}\phi_{c,2} + \lambda_{011}\phi_{c,1}(1-\phi_{c,2}) + \lambda_{101}(1-\phi_{c,1})\phi_{c,2} + \lambda_{111}(1-\phi_{c,1})(1-\phi_{c,2})
\end{aligned}$$

Note that though these equations must all hold at any stationary point in the interior of the parameter space, there is some redundancy amongst them, for example the final one can be deduced from the prior equations.

If we rearrange the equations, we get a system with the following form. Unlike the two feature case, however, those equations are not linear in $\phi_{c,j}$.

$$1 = (\lambda_{000} - \lambda_{001} - \lambda_{010} + \lambda_{011})\phi_{c,2}\phi_{c,3} + (\lambda_{001} - \lambda_{011})\phi_{c,2} + (\lambda_{010} - \lambda_{011})\phi_{c,3} + \lambda_{011} \quad (3.9)$$

$$1 = (\lambda_{100} - \lambda_{101} - \lambda_{110} + \lambda_{111})\phi_{c,2}\phi_{c,3} + (\lambda_{101} - \lambda_{111})\phi_{c,2} + (\lambda_{110} - \lambda_{111})\phi_{c,3} + \lambda_{111} \quad (3.10)$$

$$1 = (\lambda_{000} - \lambda_{001} - \lambda_{100} + \lambda_{101})\phi_{c,1}\phi_{c,3} + (\lambda_{001} - \lambda_{101})\phi_{c,1} + (\lambda_{100} - \lambda_{101})\phi_{c,3} + \lambda_{101} \quad (3.11)$$

$$1 = (\lambda_{010} - \lambda_{011} - \lambda_{110} + \lambda_{111})\phi_{c,1}\phi_{c,3} + (\lambda_{011} - \lambda_{111})\phi_{c,1} + (\lambda_{110} - \lambda_{111})\phi_{c,3} + \lambda_{111} \quad (3.12)$$

$$1 = (\lambda_{000} - \lambda_{010} - \lambda_{100} + \lambda_{110})\phi_{c,1}\phi_{c,2} + (\lambda_{010} - \lambda_{110})\phi_{c,1} + (\lambda_{100} - \lambda_{110})\phi_{c,2} + \lambda_{110} \quad (3.13)$$

$$1 = (\lambda_{001} - \lambda_{011} - \lambda_{101} + \lambda_{111})\phi_{c,1}\phi_{c,2} + (\lambda_{011} - \lambda_{111})\phi_{c,1} + (\lambda_{101} - \lambda_{111})\phi_{c,2} + \lambda_{111} \quad (3.14)$$

Now, note that if we separate these equations into groups of two by common non-linear term, then under certain conditions we can combine the two equations to create an equation that is linear in two parameters. We would like to use this to create a system of 3 linear equations with 3 variables.

For example, consider combining Equations (3.9) and (3.10) (though we could have chosen any set of two corresponding equations). Then under the following circumstances we may not be able to combine them to yield a linear equation:
1. All of the coefficients for one of the equations are 0.

2. The coefficients of the first equation can all be multiplied by the same non-zero constant to yield the coefficients of the second equation.

Start by considering case 1: suppose that all of the coefficients in (3.10) are 0. This is equivalent to saying that $1 = \lambda_{111} = \lambda_{101} = \lambda_{110} = \lambda_{100}$. Then Equations (3.11)-(3.14) become:

$$\begin{aligned}
0 =&(\lambda_{000} - \lambda_{001})\phi_{c,1}\phi_{c,3} + (\lambda_{001} - 1)\phi_{c,1} \\
0 =&(\lambda_{010} - \lambda_{011})\phi_{c,1}\phi_{c,3} + (\lambda_{011} - 1)\phi_{c,1} \\
0 =&(\lambda_{000} - \lambda_{010})\phi_{c,1}\phi_{c,2} + (\lambda_{010} - 1)\phi_{c,1} \\
0 =&(\lambda_{001} - \lambda_{011})\phi_{c,1}\phi_{c,2} + (\lambda_{011} - 1)\phi_{c,1}
\end{aligned}$$

So, if $\phi_{c,1} \neq 0$, then we have two linear equations each in $\phi_{c,2}$ and $\phi_{c,3}$. Note that the only way that both of the equations for one of the variables can be redundant is if $1 = \lambda_{011} = \lambda_{001} = \lambda_{010} = \lambda_{000}$, and in this case the stationary point will be globally optimal. Otherwise, due to linearity, if a solution does exist then it will be unique for $\phi_{c,2}$ and $\phi_{c,3}$, $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0$. Let's denote these unique solutions as $\bar{\phi_2}$ and $\bar{\phi_3}$.

So, $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0$, we must have that either $\phi_{c,1} = 0$ or $\phi_{c,2} = \bar{\phi_2}, \phi_{c,3} = \bar{\phi_3}$.

Then $P_\Omega(\mathbf{X} = 000) = \sum_{c \in \mathcal{C}} \theta_c \phi_{c,1}\phi_{c,2}\phi_{c,3} = \bar{\phi_2}\bar{\phi_3}P_\mathcal{S}(X_1 = 0)$ and similarly $P_\Omega(\mathbf{X} = 001) = \bar{\phi_2}(1 - \bar{\phi_3})P_\mathcal{S}(X_1 = 0)$, $P_\Omega(\mathbf{X} = 011) = (1 - \bar{\phi_2})\bar{\phi_3}P_\mathcal{S}(X_1 = 0)$ and $P_\Omega(\mathbf{X} = 011) = (1 - \bar{\phi_2})(1 - \bar{\phi_3})P_\mathcal{S}(X_1 = 0)$.

Now, by solving one of the linear equations in $\phi_{c,2}$ with non-zero coefficients, we find that:

$$\begin{aligned}
\bar{\phi_2}(\lambda_{000} - \lambda_{010}) &= 1 - \lambda_{010} \\
\bar{\phi_2}\left(\frac{P_\mathcal{S}(\mathbf{X} = 000)}{\bar{\phi_2}\bar{\phi_3}P_\mathcal{S}(X_1 = 0)} - \frac{P_\mathcal{S}(\mathbf{X} = 010)}{(1 - \bar{\phi_2})\bar{\phi_3}P_\mathcal{S}(X_1 = 0)}\right) &= 1 - \frac{P_\mathcal{S}(\mathbf{X} = 010)}{(1 - \bar{\phi_2})\bar{\phi_3}P_\mathcal{S}(X_1 = 0)} \\
\bar{\phi_3} &= \frac{P_\mathcal{S}(X_1 = 0, X_3 = 0)}{P_\mathcal{S}(X_1 = 0)}
\end{aligned}$$

Doing the same for one of the linear equations in $\phi_{c,3}$, we find that $\bar{\phi_2} = \frac{P_\mathcal{S}(X_1=0,X_2=0)}{P_\mathcal{S}(X_1=0)}$.

So, if all of the coefficients for the second equation are 0, then at any stationary point in the interior of the parameter space $P_\Omega(\mathbf{X} = 0ab) = \frac{P_\mathcal{S}(X_1=0,X_2=a)P_\mathcal{S}(X_1=0,X_3=b)}{P_\mathcal{S}(X_1=0)}$.

Even if all of the coefficients for one of the equations are 0, we can still deduce some information about the stationary points, and furthermore we can say that if the features are

independent that any stationary point of the log likelihood in the interior of the parameter space will be globally optimal.

Now, consider the second case where we cannot create a linear system of equations: when one equation is a non-zero multiple of its related equation.

Taking Equations (3.9) and (3.10) again as examples, this is equivalent to saying that $\exists \alpha \neq 0$ s.t. $\alpha(\lambda_{011} - 1) = \lambda_{111} - 1, \alpha(\lambda_{010} - 1) = \lambda_{110} - 1, \alpha(\lambda_{001} - 1) = \lambda_{101} - 1$ and $\alpha(\lambda_{000} - 1) = \lambda_{100} - 1$.

Substituting these into Equations (3.11)-(3.14) yields:

$$\begin{aligned}
0 =&(1-\alpha)(\lambda_{000} - \lambda_{001})\phi_{c,1}\phi_{c,3} + (1-\alpha)(\lambda_{001} - 1)\phi_{c,1} + \alpha(\lambda_{000} - \lambda_{001})\phi_{c,3} + \alpha(\lambda_{001} - 1) \\
0 =&(1-\alpha)(\lambda_{010} - \lambda_{011})\phi_{c,1}\phi_{c,3} + (1-\alpha)(\lambda_{011} - 1)\phi_{c,1} + \alpha(\lambda_{010} - \lambda_{011})\phi_{c,3} + \alpha(\lambda_{011} - 1) \\
0 =&(1-\alpha)(\lambda_{000} - \lambda_{010})\phi_{c,1}\phi_{c,2} + (1-\alpha)(\lambda_{010} - 1)\phi_{c,1} + \alpha(\lambda_{000} - \lambda_{010})\phi_{c,2} + \alpha(\lambda_{010} - 1) \\
0 =&(1-\alpha)(\lambda_{001} - \lambda_{011})\phi_{c,1}\phi_{c,2} + (1-\alpha)(\lambda_{011} - 1)\phi_{c,1} + \alpha(\lambda_{001} - \lambda_{011})\phi_{c,2} + \alpha(\lambda_{011} - 1)
\end{aligned}$$

Combining the first two equations and the last two equations yields:

$$\begin{aligned}
0 =&[(\lambda_{000} - \lambda_{001} - \lambda_{010} + \lambda_{011})\phi_{c,2} + \lambda_{010} - \lambda_{011}][(1-\alpha)\phi_{c,1} + \alpha] \\
0 =&[(\lambda_{000} - \lambda_{001} - \lambda_{010} + \lambda_{011})\phi_{c,3} + \lambda_{001} - \lambda_{011}][(1-\alpha)\phi_{c,1} + \alpha]
\end{aligned}$$

Suppose that $(\alpha - 1)\phi_{c,1} = \alpha \; \forall c \in \mathcal{C}$ s.t. $\theta_c > 0$. Then, by Theorem 1, $\phi_{c,1} = P_{\mathcal{S}}(X_1 = 0) \; \forall c \in \mathcal{C}$ s.t. $\theta_c > 0$.

We know that $\alpha(\lambda_{000} - 1) = \lambda_{100} - 1$. But then using $(\alpha - 1)P_{\mathcal{S}}(X_1 = 0) = \alpha$, this can be rewritten as:

$$\begin{aligned}
1 &= P_{\mathcal{S}}(X_1 = 0)\lambda_{000} + P_{\mathcal{S}}(X_1 = 1)\lambda_{100} \\
&= \frac{P_{\mathcal{S}}(X_1 = 0)P_{\mathcal{S}}(\mathbf{X} = 000)}{\sum_{c \in \mathcal{C}} \theta_c \phi_{c,1}\phi_{c,2}\phi_{c,3}} + \frac{P_{\mathcal{S}}(X_1 = 1)P_{\mathcal{S}}(\mathbf{X} = 100)}{\sum_{c \in \mathcal{C}} \theta_c (1 - \phi_{c,1})\phi_{c,2}\phi_{c,3}} \\
&= \frac{P_{\mathcal{S}}(X_2 = 0, X_3 = 0)}{\sum_{c \in \mathcal{C}} \theta_c \phi_{c,2}\phi_{c,3}}
\end{aligned}$$

Then, at any stationary point of this nature, $P_{\Omega}(\mathbf{X} = abc) = P_{\mathcal{S}}(X_1 = a)P_{\mathcal{S}}(X_2 = b, X_3 = c)$. If the features are independent, then any stationary point of this nature will be globally optimal.

Next, suppose that $\exists \hat{c} \in \mathcal{C}$ s.t. $\theta_{\hat{c}} > 0$ and $(\alpha - 1)\phi_{\hat{c},1} \neq \alpha$. Without loss of generality, assume $\hat{c} = 1$. Then, $\phi_{1,2}$ and $\phi_{1,3}$ are uniquely determined.

By substituting $0 = (\lambda_{000} - \lambda_{001} - \lambda_{010} + \lambda_{011})\phi_{1,2} + \lambda_{010} - \lambda_{011}$ into Equation (3.9), we see that $(\lambda_{001} - \lambda_{011})\phi_{1,2} = 1 - \lambda_{011}$, and substituting this back shows $(\lambda_{000} - \lambda_{010})\phi_{1,2} = 1 - \lambda_{010}$. This shows that $\phi_{1,2}(\lambda_{001} - 1) = (\phi_{1,2} - 1)(\lambda_{011} - 1)$ and $\phi_{1,2}(\lambda_{000} - 1) = (\phi_{1,2} - 1)(\lambda_{010} - 1)$.

Similarly for $\phi_{1,3}$, we see that $\phi_{1,3}(\lambda_{010} - 1) = (\phi_{1,3} - 1)(\lambda_{011} - 1)$ and $\phi_{1,3}(\lambda_{000} - 1) = (\phi_{1,3} - 1)(\lambda_{001} - 1)$.

This shows that (3.9) is a multiple of (3.10), (3.11) is a multiple of (3.12) and (3.13) is a multiple of (3.14).

Note that if $\phi_{1,2} = 0$, then we have that $1 = \lambda_{010} = \lambda_{011} = \lambda_{110} = \lambda_{111}$. This, however, would mean that Equation (3.12) would have all 0 coefficients, which is a case that we have already dealt with. A similar case arises if $\phi_{1,2} = 1$ or if $\phi_{1,3} \in \{0, 1\}$. So, assume $0 < \phi_{1,2}, \phi_{1,3} < 1$.

Define $\beta = \frac{-\phi_{1,2}}{1-\phi_{1,2}}$ and $\gamma = \frac{-\phi_{1,3}}{1-\phi_{1,3}}$.

Next, note that $\lambda_{\mathbf{x}}$ can be defined in terms of $\lambda_{000}, \alpha, \beta$ and γ $\forall \mathbf{x}$. We can in fact substitute these values into Equation (3.9), which must hold $\forall c$ s.t. $\theta_c > 0$. This becomes:

$$\begin{aligned}
0 =& (1-\beta)(1-\gamma)(\lambda_{000} - 1)\phi_{c,2}\phi_{c,3} + \gamma(1-\beta)(\lambda_{000} - 1)\phi_{c,2} + \beta(1-\gamma)(\lambda_{000} - 1)\phi_{c,3} + \beta\gamma \\
=& (\lambda_{000} - 1)[\phi_{c,2}\phi_{c,3} + \frac{\gamma}{1-\gamma}\phi_{c,2} + \frac{\beta}{1-\beta}\phi_{c,3} + \frac{\beta}{1-\beta}\frac{\gamma}{1-\gamma}] \\
=& (\lambda_{000} - 1)(\phi_{c,2} + \frac{\beta}{1-\beta})(\phi_{c,3} + \frac{\gamma}{1-\gamma})
\end{aligned}$$

Doing the same substitution for Equations (3.11) and (3.13) yields:

$$\begin{aligned}
0 =& (\lambda_{000} - 1)(\phi_{c,2} + \frac{\beta}{1-\beta})(\phi_{c,3} + \frac{\gamma}{1-\gamma}) \\
0 =& (\lambda_{000} - 1)(\phi_{c,1} + \frac{\alpha}{1-\alpha})(\phi_{c,3} + \frac{\gamma}{1-\gamma}) \\
0 =& (\lambda_{000} - 1)(\phi_{c,1} + \frac{\alpha}{1-\alpha})(\phi_{c,2} + \frac{\beta}{1-\beta})
\end{aligned}$$

If $\lambda_{000} = 1$, then we can easily verify that $\lambda_{\mathbf{x}} = 1$ $\forall \mathbf{x}$, and therefore this stationary point is globally optimal.

Note that for class 1, we know that $\phi_{1,2} = \frac{-\beta}{1-\beta}$ and $\phi_{1,3} = \frac{-\gamma}{1-\gamma}$.

For class 2, if $\phi_{2,2} = \phi_{1,2}$ and $\phi_{2,3} = \phi_{1,3}$, then by Lemma 1, if the features are independent the point is globally optimal. If this is not the case, then we must have $\phi_{2,1} = \frac{-\alpha}{1-\alpha}$, and either $\phi_{2,2} = \phi_{1,2}$ or $\phi_{2,3} = \phi_{1,3}$.

Suppose $\phi_{2,1} = \frac{-\alpha}{1-\alpha}$, and $\phi_{2,2} = \phi_{1,2}$. Then:

$$\phi_{2,1}(\lambda_{000} - \lambda_{010}) = 1 - \lambda_{010}$$
$$P_{\mathcal{S}}(\mathbf{X} = 000) - \frac{\phi_{2,2} P_{\mathcal{S}}(\mathbf{X} = 010)}{1 - \phi_{2,2}} = \theta_1 \phi_{1,1} \phi_{1,3} + \theta_2 \phi_{2,1} \phi_{2,3} - \frac{P_{\mathcal{S}}(\mathbf{X} = 010)}{1 - \phi_{2,2}}$$
$$P_{\mathcal{S}}(X_1 = 0, X_3 = 0) = \theta_1 \phi_{1,1} \phi_{1,3} + \theta_2 \phi_{2,1} \phi_{2,3}$$

We obtain $P_\Omega(\mathbf{X} = abc) = P_{\mathcal{S}}(X_2 = b) P_{\mathcal{S}}(X_1 = a, X_3 = c)$. Similarly, had we chosen $\phi_{2,3} = \phi_{1,3}$, we would have found that $P_\Omega(\mathbf{X} = abc) = P_{\mathcal{S}}(X_3 = c) P_{\mathcal{S}}(X_1 = a, X_2 = b)$. If the features are independent, then these points match the empirical distribution.

So we have now proven that if we are unable to create a linear equation out of each of the 3 pairs of equations (3.9)(3.10), (3.11)(3.12) and (3.13)(3.14), and if the features are linearly independent, then every stationary point of the likelihood in the interior of the parameter space is globally optimal and replicates the sample distribution. □

Lemma 3. *Suppose in that we have successfully combined each pair of equations* (3.9)(3.10), (3.11)(3.12) *and* (3.13)(3.14) *to create 3 equations that are linear in* $\phi_{c,j} \forall j$ *and the features of the empirical distribution are independent. Then all solutions of this linear system match the empirical distribution and are therefore globally optimal.*

Proof. Assume we have manipulated the equations to create the linear system. We will then be able to use standard results to show when a unique solution exists. Start by noting that, for a pair of equations, say (3.9) and (3.10), if either one of them has a coefficient for the non-linear term that is non-zero, then we can multiply each equation by the appropriate amount and take the difference. For now, assume that in each pair of equations at least one of them has a non-zero non-linear coefficient.

Then, define:
$\alpha_1 = \lambda_{000} - \lambda_{001} - \lambda_{010} + \lambda_{011}$
$\beta_1 = \lambda_{100} - \lambda_{101} - \lambda_{110} + \lambda_{111}$
$\alpha_2 = \lambda_{000} - \lambda_{001} - \lambda_{100} + \lambda_{101}$
$\beta_2 = \lambda_{010} - \lambda_{011} - \lambda_{110} + \lambda_{111}$
$\alpha_3 = \lambda_{000} - \lambda_{010} - \lambda_{100} + \lambda_{110}$
$\beta_3 = \lambda_{100} - \lambda_{110} - \lambda_{110} + \lambda_{111}$

Then by doing β_1(3.9)$-\alpha_1$(3.10), β_2(3.11)$-\alpha_2$(3.12) and β_3(3.13)$-\alpha_3$(3.14), we get the

following linear system of three equations with three $\phi_{c,j}$ variables:

$$\begin{aligned}
&\beta_1(\lambda_{011}-1)-\alpha_1(\lambda_{111}-1)\\
&\qquad=[\beta_1(\lambda_{001}-\lambda_{011})-\alpha_1(\lambda_{101}-\lambda_{111})]\phi_{c,2}+[\beta_1(\lambda_{010}-\lambda_{011})-\alpha_1(\lambda_{110}-\lambda_{111})]\phi_{c,3}\\
&\beta_2(\lambda_{101}-1)-\alpha_2(\lambda_{111}-1)\\
&\qquad=[\beta_2(\lambda_{001}-\lambda_{101})-\alpha_2(\lambda_{011}-\lambda_{111})]\phi_{c,1}+[\beta_2(\lambda_{100}-\lambda_{101})-\alpha_2(\lambda_{110}-\lambda_{111})]\phi_{c,3}\\
&\beta_3(\lambda_{110}-1)-\alpha_3(\lambda_{111}-1)\\
&\qquad=[\beta_3(\lambda_{010}-\lambda_{110})-\alpha_3(\lambda_{011}-\lambda_{111})]\phi_{c,1}+[\beta_3(\lambda_{100}-\lambda_{110})-\alpha_3(\lambda_{101}-\lambda_{111})]\phi_{c,2}
\end{aligned}$$

By taking the determinant, we find that the matrix representation of this system will be full rank unless:
$[\beta_1(\lambda_{010}-\lambda_{011})-\alpha_1(\lambda_{110}-\lambda_{111})][\beta_2(\lambda_{001}-\lambda_{101})-\alpha_2(\lambda_{011}-\lambda_{111})][\beta_3(\lambda_{100}-\lambda_{110})-\alpha_3(\lambda_{101}-\lambda_{111})]$
$=-[\beta_1(\lambda_{001}-\lambda_{011})-\alpha_1(\lambda_{101}-\lambda_{111})][\beta_2(\lambda_{100}-\lambda_{101})-\alpha_2(\lambda_{110}-\lambda_{111})][\beta_3(\lambda_{010}-\lambda_{110})-\alpha_3(\lambda_{011}-\lambda_{111})]$

If we insert the values of α_k, β_k, then this is equivalent to:

$$\begin{aligned}
&[(\lambda_{100}-\lambda_{101})(\lambda_{010}-\lambda_{011})-(\lambda_{000}-\lambda_{001})(\lambda_{110}-\lambda_{111})]\\
&[(\lambda_{010}-\lambda_{110})(\lambda_{001}-\lambda_{101})-(\lambda_{000}-\lambda_{100})(\lambda_{011}-\lambda_{111})]\\
&[(\lambda_{001}-\lambda_{011})(\lambda_{100}-\lambda_{110})-(\lambda_{000}-\lambda_{010})(\lambda_{101}-\lambda_{111})]\\
&=-\\
&[(\lambda_{100}-\lambda_{101})(\lambda_{010}-\lambda_{011})-(\lambda_{000}-\lambda_{001})(\lambda_{110}-\lambda_{111})]\\
&[(\lambda_{010}-\lambda_{110})(\lambda_{001}-\lambda_{101})-(\lambda_{000}-\lambda_{100})(\lambda_{011}-\lambda_{111})]\\
&[(\lambda_{001}-\lambda_{011})(\lambda_{100}-\lambda_{110})-(\lambda_{000}-\lambda_{010})(\lambda_{101}-\lambda_{111})]
\end{aligned}$$

If this relation does not hold, then the system is full rank and has at most one solution, and so there is only a single possible value for $(\phi_{c,1}, \phi_{c,2}, \phi_{c,3})$, regardless of the class c. Hence, $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0$, $\phi_{c,j} = P_{\mathcal{S}}(F_j = 0)$, and by Lemma 1, if the features are independent then every stationary point in the interior of the parameter space is globally optimal.

If this relation does hold, then at least one of the expressions must be zero. Say, for example, $(\lambda_{010}-\lambda_{110})(\lambda_{001}-\lambda_{101}) = (\lambda_{000}-\lambda_{100})(\lambda_{011}-\lambda_{111})$. If there is a factor that is zero on both sides, say $\lambda_{010} = \lambda_{110}$ and $\lambda_{000} = \lambda_{100}$, then one equation (in this case, (3.13)) will be redundant. This is a case that we have already dealt with in Lemma 2.

Otherwise, we have $\frac{(\lambda_{010}-\lambda_{110})}{(\lambda_{000}-\lambda_{100})} = \frac{(\lambda_{011}-\lambda_{111})}{(\lambda_{001}-\lambda_{101})}$ and $\frac{(\lambda_{010}-\lambda_{110})}{(\lambda_{011}-\lambda_{111})} = \frac{(\lambda_{000}-\lambda_{100})}{(\lambda_{001}-\lambda_{101})}$. Intuitively, what these relationships mean is that the $\phi_{c,1}$ term will cancel out when we combine either

(3.11) and (3.12) or (3.13) and (3.14). Since we have already dealt with the case where one equation is a multiple of the other in Lemma 2, assume that they are not. Then combining (3.11) and (3.12) will create an equation that is linear in $\phi_{c,3}$ with a non-zero coefficient, and similarly for $\phi_{c,2}$. Hence, $\phi_{c,2}$ and $\phi_{c,3}$ are uniquely determined $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0$. By Lemma 1, if the features are linearly independent then this stationary point will match the sample distribution.

The final case that we must verify is when both of the non-linear coefficients for a set of equations are zero, so we cannot multiply across both equations. Suppose $\lambda_{000} = \lambda_{001} + \lambda_{010} - \lambda_{011}$ and $\lambda_{100} = \lambda_{101} + \lambda_{110} - \lambda_{111}$. Then Equations (3.9) and (3.10) become:

$$0 = (\lambda_{001} - \lambda_{011})\phi_{c,2} + (\lambda_{010} - \lambda_{011})\phi_{c,3} + \lambda_{011} - 1$$
$$0 = (\lambda_{101} - \lambda_{111})\phi_{c,2} + (\lambda_{110} - \lambda_{111})\phi_{c,3} + \lambda_{111} - 1$$

Since we have already dealt in Lemma 2 with the case where one equation is a multiple of the other, or where either equation is redundant we know that these equations must have a unique solution. Hence, $\phi_{c,2} = P_{\mathcal{S}}(X_2 = 0)$ and $\phi_{c,3} = P_{\mathcal{S}}(X_3 = 0)$ $\forall c \in \mathcal{C}$ s.t. $\theta_c > 0$. By Lemma 1 if the features are linearly independent then this stationary point will match the sample distribution. □

We can now complete the proof of Theorem 4.

Proof. Start by attempting to create a system of 3 equations linear in $\phi_{c,1}$, $\phi_{c,2}$ and $\phi_{c,3}$ based Equations (3.9)-(3.14), which are generated from the derivatives. If the terms $\lambda_{\mathbf{x}}$ are such that we are unable to do so and the features of the empirical distribution are independent then by Lemma 2 any stationary point of the likelihood in the interior of the parameter space will match the empirical distribution.

Otherwise, assume that we have created a linear system out of Equations (3.9)-(3.14). By Lemma 3, if the features are independent then any stationary point of the log likelihood in the interior of the parameter space will be globally optimal and will match the sample distribution.

The characterization of the stationary points in the interior of the parameter space that may not be globally optimal is created by looking at the cases where a linear system cannot be created as identified in the proof of Lemma 2, as well as the solutions to the linear system as identified in the proof of Lemma 3. □

3.8 Example of Spurious Local Optima

I will now illustrate an example. I will use the two feature case, and describe the optima in the interior of the parameter space.

Suppose that we are observing two features, $\mathbf{X} = (X_1, X_2)$, and that we have a sample $\mathcal{S}$ with 16 observations. Suppose the observations are distributed as follows:

$P_{\mathcal{S}}(\mathbf{X} = (0,0))$	$P_{\mathcal{S}}(\mathbf{X} = (0,1))$	$P_{\mathcal{S}}(\mathbf{X} = (1,0))$	$P_{\mathcal{S}}(\mathbf{X} = (1,1))$
$\frac{3}{16}$	$\frac{2}{16}$	$\frac{6}{16}$	$\frac{5}{16}$

Then, we have that $P_{\mathcal{S}}(X_1 = 0) = \frac{5}{16}$ and $P_{\mathcal{S}}(X_2 = 0) = \frac{9}{16}$.

By the result from Theorem 3, we know that any spurious local optimum in the interior of the parameter space must have the form $\phi_{c,j} = P_{\mathcal{S}}(X_j = 0)$. Indeed, if construct a parameterization where $\phi_{c,1} = \frac{5}{16}$ and $\phi_{c,2} = \frac{9}{16}$, then this point will be stationary regardless of the values of any θ parameters. Hence, any point Ω' of this form will be locally optimal. Furthermore Ω' encode the following distribution:

$P_{\Omega'}(\mathbf{X} = (0,0))$	$P_{\Omega'}(\mathbf{X} = (0,1))$	$P_{\Omega'}(\mathbf{X} = (1,0))$	$P_{\Omega'}(\mathbf{X} = (1,1))$
$\frac{45}{256}$	$\frac{35}{256}$	$\frac{99}{256}$	$\frac{77}{256}$

and Ω' will have average log likelihood -0.5668.

Now, however, consider the point $\Omega^* = (\theta, \phi_{0,1}, \phi_{0,2}, \phi_{1,1}, \phi_{1,2}) = (\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{1}{4}, \frac{1}{2})$.

Ω^* has average log likelihood -0.5674, and encodes the distribution:

$P_{\Omega^*}(\mathbf{X} = (0,0))$	$P_{\Omega^*}(\mathbf{X} = (0,1))$	$P_{\Omega^*}(\mathbf{X} = (1,0))$	$P_{\Omega^*}(\mathbf{X} = (1,1))$
$\frac{3}{16}$	$\frac{2}{16}$	$\frac{6}{16}$	$\frac{5}{16}$

Ω^* and Ω' are both stationary points in the interior of the parameter space. However, Ω^* is globally optimal and exactly matches the empirical distribution, whereas Ω' is spuriously locally optimal, and only matches the marginal empirical distribution for each individual feature.

This is an example of how spurious local optima can occur. In this case, if we were to use gradient ascent to converge to a point in the interior of the parameter space, then we would check whether the point satisfies $\phi_{c,1} = \frac{5}{16}$ and $\phi_{c,2} = \frac{9}{16}$ for each class c. If this is not the case, then we know that we have achieved a global optimum. On the other hand, if the ϕ values are as specified, then by simply checking whether the features are independent in the distribution of our sample, we can finally verify whether we are in a spurious local optimum.

3.9 Conclusion of Work Describing Unsupervised Likelihood Optima

In this Chapter I proved several interesting results related to the stationary points of the likelihood of the unsupervised naïve Bayes model, which is commonly used as a baseline for clustering.

First of all, I showed that for any number of features that at any stationary point in the interior of the parameter space, the marginal probabilities for single feature will exactly match those from the sample. I then provided the first description of the optima for cases with up to three features. I showed that global optimality is generally attained unless special conditions are met. Using these conditions, it is possible to assess, at any stationary point that might be encountered during training, whether it is possible that the algorithm might be stuck at a spurious maximum.

This is the first detailed description of the optima of the likelihood of the unsupervised naïve Bayes model, and leads to some interesting future work which will be described in the thesis conclusion in Chapter 5.

Chapter 4

Sample Complexity of the Naïve Bayes Classifier

When modeling a distribution, the choice of optimal parameters will depend on what is seen in the training sample. Hence, when building a model it is desirable to have a probabilistic guarantee about how well the sample represents the underlying distribution.

This chapter begins an investigation into the sample complexity of the naïve Bayes model. The goal of this work is to show that given confidence and accuracy thresholds δ, ϵ for agnostic PAC learning, we can create an algorithm to determine the value m that represents the minimum sample size required to achieve the specified accuracy with sufficient confidence using maximum likelihood. This will be described in more detail in Section 4.1. In fact, this result has not yet been fully proven, but I did complete some proofs of intermediate results. I also developed a conjecture, which, if proven, would determine an exact sample complexity which is much lower than the upper bounds found using the known thoery.

This result, once the proof is complete, will allow users to save on storage and computational costs. They will be able to know exactly how much data they need to probabilistically achieve their accuracy goals, and use only that amount.

4.1 Problem Description

Suppose the naïve Bayes model is being used for supervised classification. There is some set of naïve Bayes parameters that will most accurately represent the true underlying

distribution. However, during training we select the parameters that best describe the sample. Depending on the sample used, we may end up choosing parameters that lead to different classifications than the optimal classifier.

In general, as the sample size gets larger the probability of choosing a non-optimal classifier will become smaller, since by the central limit theorem the distribution of different feature and class values in the sample will converge towards their true probabilities. However, there are also costs for using more data, for example higher computation time, computation power and data storage requirements. In cases where there are lots of data available and we are striving for probabilistically near-optimal learning, it is possible that we could achieve our accuracy goals using only a portion of the data. In this case, it would be advantageous to know exactly how much data we need to guarantee that we are within a specified accuracy margin of the optimal classifier with sufficiently high probability. This will allow us to balance strong model performance with reasonable computational costs. In this chapter, I outline my progress so far on determining bounds on the amount of data needed for probabilistically near-optimal learning.

Since our learning task is agnostic, we will consider the estimation error, that is the error that is incremental to that of the optimal naïve Bayes classifier. Error will be measured by the misclassification percentage over the true distribution.

For this work, we want to know, for given confidence and accuracy thresholds $\delta, \epsilon \in (0, 1)$, can we find the minimum integer m such that if the training set contains at least m observations, then the probability of having estimation error greater than ϵ is less than δ, regardless of the true underlying distribution?

To simplify the analysis, I started by considering the case of a single binary feature and a binary class.

4.2 Definitions and Notation

Let X be the single binary feature, and C be the binary class.

I will use a representation of a naïve Bayes classifier with the lowest misclassification error over the distribution throughout this work. The parameters for this classifier will be $\theta^* = P(C = 0)$, $\phi_0^* = P(X = 0|C = 0)$ and $\phi_1^* = P(X = 0|C = 1)$.

To describe the sample, I will use m to represent the number of observations. For $c \in \{0, 1\}$, k_c will represent the number of observations of class c. Similarly for $c, x \in \{0, 1\}$, k_{cx} will represent the number of observations of class c with feature x.

For agnostic P.A.C. learning, ϵ will be the threshold that we want to bound the error by and δ will be the probability with which we allow ourselves to exceed that threshold. So, we want to find m such that, if the sample $\mathcal{S}$ has at least m i.i.d. observations (x, c) from the distribution $\mathcal{D}$, then with probability at least $1-\delta$, we select $h \in \mathcal{H}$ that satisfies:

$P_{\mathcal{D}}(h(x) \neq c) \leq \min_{h' \in \mathcal{H}} P_{\mathcal{D}}(h'(x) \neq c) + \epsilon$

4.3 Problem Formulation

Consider that for the distribution, there are parameters $(\theta^*, \phi_0^*, \phi_1^*)$ that define an optimal naïve Bayes classifier as selected using ERM and misclassification loss. If our sample generates a different set of parameters for the classifier, then we might have more misclassification error than the optimal classifier. Since the training sample is finite, any algorithm might choose such a sub-optimal parameterization.

We want to be guaranteed that with high probability, the classifier we choose will be nearly optimal, regardless of what the true underlying distribution is. Moreover, we want to find the minimum sample size m that yields this guarantee.

Formally, for user defined ϵ and δ values and where err is the estimation error, the optimization problem is:

$$min(m \in \mathbb{N} \quad s.t. \quad \min_{(\theta^*, \phi_0^*, \phi_1^*) \in [0,1]^3} P(err < \epsilon) \geq 1 - \delta) \tag{4.1}$$

Where $P(err < \epsilon)$ will be defined later in this section. Note that to use many conventional search algorithms over m, we will also need to show that $\min_{(\theta^*, \phi_0^*, \phi_1^*) \in [0,1]^3} P(err < \epsilon)$ is monotonically increasing w.r.t. m. This will be done on page 41.

To solve this problem, we first need a way to determine, if an optimal classifier for the underlying distribution was $(\theta^*, \phi_0^*, \phi_1^*)$ and we took a sample of size m, what would be the probability that we would end up with estimation error $< \epsilon$? However, when we actually build our model, the distribution is unknown, and we want our model to be sufficiently accurate regardless of what the true distribution is. So, if we can define that probability $\forall(\theta^*, \phi_0^*, \phi_1^*) \in [0, 1]^3$, then we can minimize over the cube to find which set of parameters yields the lowest probability of being within ϵ of the optimal classifier. Then we know that regardless of what the true underlying distribution is, the probability of choosing a sufficiently good model is at least as high as it is at the minimum.

Once we determine how to evaluate the probability $P(err < \epsilon)$ for a fixed $\theta^*, \phi_0^*, \phi_1^*$ and m, then the main difficulty will be figuring out how to minimize this over the space of parameters $(\theta^*, \phi_0^*, \phi_1^*) \in [0,1]^3$. If we had a simple method for finding this minimum, then we could iterate over m values (or, use a more efficient algorithm such as binary search) until we find a value m^* such that $\min_{(\theta^*,\phi_0^*,\phi_1^*)\in[0,1]^3} P(err < \epsilon) \geq 1 - \delta$ is satisfied at $m = m^*$, but not at $m = m^* - 1$.

Since we know exactly how to solve the problem once we can solve $\min_{(\theta^*,\phi_0^*,\phi_1^*)\in[0,1]^3} P(err < \epsilon)$ for a fixed m, the focus of this section will primarily be on deriving the function $P(err < \epsilon)$ for a fixed $\theta^*, \phi_0^*, \phi_1^*$ and m, and discussing its properties to give us some insights into how to minimize it.

Start by defining the misclassification error of the optimal classifier:

$$\begin{aligned} err_{opt} = &min\{P(C = 0, X = 0), P(C = 1, X = 0)\} \\ &+ min\{P(C = 0, X = 1), P(C = 1, X = 1)\} \end{aligned}$$

This is because when the classifier sees $X = 0$, it will select the most likely class. So, any observations with $X = 0$ and the less likely class, given $X = 0$, will be misclassified. This will happen with probability $min\{P(C = 0, X = 0), P(C = 1, X = 0)\}$, and similarly for $X = 1$.

Note that since $(\theta^*, \phi_0^*, \phi_1^*)$ is an optimal classifier, if $P(C = 0, X = 0) \geq P(C = 1, X = 0)$ then the classifier must have $\theta^*\phi_0^* \geq (1 - \theta^*)\phi_1^*$, and vice versa. So, $P(C = 0, X = 0) \geq P(C = 1, X = 0) \Leftrightarrow \theta^* \geq \frac{\phi_1^*}{\phi_0^* + \phi_1^*}$.

Similarly, $P(C = 0, X = 1) \geq P(C = 1, X = 1) \Leftrightarrow \theta^* \geq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*}$.

Furthermore, for this work, I will assume that $(\theta^*, \phi_0^*, \phi_1^*)$ perfectly represents our distribution, since any single feature distribution can be expressed using naïve Bayes parameters. In future, however, I would like to expand this work to cover cases with multiple features. In this case, not every distribution can necessarily be represented using naïve Bayes parameters, so a different approach to the problem formulation will need to be used.

The misclassification error of the optimal classifier is:

$$err_{opt} = \begin{cases} P(C=0,X=0)+P(C=0,X=1)\, if\ \theta^* \leq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}\ and\ \theta^* \leq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*} \\ P(C=1,X=0)+P(C=0,X=1)\, if\ \theta^* \geq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}\ and\ \theta^* \leq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*} \\ P(C=0,X=0)+P(C=1,X=1)\, if\ \theta^* \leq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}\ and\ \theta^* \geq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*} \\ P(C=1,X=0)+P(C=1,X=1)\, if\ \theta^* \geq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}\ and\ \theta^* \geq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*} \end{cases}$$

$$= \begin{cases} \theta^* & if\ \theta^* \leq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}\ and\ \theta^* \leq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*} \\ \phi_1^*+\theta^*(1-\phi_0^*-\phi_1^*) & if\ \theta^* \geq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}\ and\ \theta^* \leq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*} \\ 1-\phi_1^*+\theta^*(\phi_0^*+\phi_1^*-1) & if\ \theta^* \leq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}\ and\ \theta^* \geq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*} \\ 1-\theta^* & if\ \theta^* \geq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}\ and\ \theta^* \geq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*} \end{cases}$$

Suppose we create a naïve Bayes model based on a sample from a distribution. The sample is described by: m, the number of observations, k_c the number of observations of class c, and k_{cx} the number of observations of class c with feature x.

For a model trained using any ERM rule and the misclassification loss, it will classify new observations based on the following probabilities:

$$P_{NB}(C=0|X=0,k_{00},k_{10}) = \begin{cases} 1\, if\ k_{00} > k_{10} \\ \frac{1}{2}\, if\ k_{00} = k_{10} \\ 0\, if\ k_{00} < k_{10} \end{cases}$$

$$P_{NB}(C=0|X=1,k_{01},k_{11}) = \begin{cases} 1\, if\ k_{01} > k_{11} \\ \frac{1}{2}\, if\ k_{01} = k_{11} \\ 0\, if\ k_{01} < k_{11} \end{cases}$$

Then the classifier error depends on the relative values of k_{00}, k_{01}, k_{10} and k_{11} from the sample, as well as the underlying distribution:

$$\begin{aligned} err_{tot} =& P(X=0)[P_{NB}(C=0|X=0,k_{00},k_{10})P(C=1|X=0) \\ &+ P_{NB}(C=1|X=0,k_{00},k_{10})P(C=0|X=0)] \\ &+ P(X=1)[P_{NB}(C=0|X=1,k_{01},k_{11})P(C=1|X=1) \\ &+ P_{NB}(C=1|X=1,k_{01},k_{11})P(C=0|X=1)] \end{aligned}$$

Based on the values of k_{cx}, the total classifier error is as follows:

1. $k_{00} < k_{10}, k_{01} < k_{11} \rightarrow err_{tot} = \theta^*$
2. $k_{00} = k_{10}, k_{01} < k_{11} \rightarrow err_{tot} = \theta^* + \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
3. $k_{00} > k_{10}, k_{01} < k_{11} \rightarrow err_{tot} = \theta^* + \phi_1^* - \theta^*(\phi_0^* + \phi_1^*)$
4. $k_{00} < k_{10}, k_{01} = k_{11} \rightarrow err_{tot} = \frac{1}{2} - \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
5. $k_{00} = k_{10}, k_{01} = k_{11} \rightarrow err_{tot} = \frac{1}{2}$
6. $k_{00} > k_{10}, k_{01} = k_{11} \rightarrow err_{tot} = \frac{1}{2} + \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
7. $k_{00} < k_{10}, k_{01} > k_{11} \rightarrow err_{tot} = 1 - \theta^* - \phi_1^* + \theta^*(\phi_0^* + \phi_1^*)$
8. $k_{00} = k_{10}, k_{01} > k_{11} \rightarrow err_{tot} = 1 - \theta^* - \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
9. $k_{00} > k_{10}, k_{01} > k_{11} \rightarrow err_{tot} = 1 - \theta^*$

For the estimation error, we will need to subtract the approximation error from the total error. So we will need to consider the nine different circumstances for the sample parameters k_{cx}, as listed above, as well as the four possible cases for the optimal error, based on the underlying distribution. So, there will be four regions, each with nine subcases.

Region A: $\theta^* \leq \frac{\phi_1^*}{\phi_0^* + \phi_1^*}$ **and** $\theta^* \leq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*}$

1. $k_{00} < k_{10}, k_{01} < k_{11} \rightarrow err = 0$
2. $k_{00} = k_{10}, k_{01} < k_{11} \rightarrow err = \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
3. $k_{00} > k_{10}, k_{01} < k_{11} \rightarrow err = \phi_1^* - \theta^*(\phi_0^* + \phi_1^*)$
4. $k_{00} < k_{10}, k_{01} = k_{11} \rightarrow err = \frac{1}{2} - \theta^* - \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
5. $k_{00} = k_{10}, k_{01} = k_{11} \rightarrow err = \frac{1}{2} - \theta^*$
6. $k_{00} > k_{10}, k_{01} = k_{11} \rightarrow err = \frac{1}{2} - \theta^* + \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
7. $k_{00} < k_{10}, k_{01} > k_{11} \rightarrow err = 1 - 2\theta^* - \phi_1^* + \theta^*(\phi_0^* + \phi_1^*)$
8. $k_{00} = k_{10}, k_{01} > k_{11} \rightarrow err = 1 - 2\theta^* - \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
9. $k_{00} > k_{10}, k_{01} > k_{11} \rightarrow err = 1 - 2\theta^*$

Region B: $\theta^* \geq \frac{\phi_1^*}{\phi_0^* + \phi_1^*}$ **and** $\theta^* \leq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*}$

1. $k_{00} < k_{10}, k_{01} < k_{11} \rightarrow err = \theta^*(\phi_0^* + \phi_1^*) - \phi_1^*$
2. $k_{00} = k_{10}, k_{01} < k_{11} \rightarrow err = \frac{1}{2}(\theta^*(\phi_0^* + \phi_1^*) - \phi_1^*)$
3. $k_{00} > k_{10}, k_{01} < k_{11} \rightarrow err = 0$
4. $k_{00} < k_{10}, k_{01} = k_{11} \rightarrow err = \frac{1}{2} - \theta^* + \frac{3}{2}(\theta^*(\phi_0^* + \phi_1^*) - \phi_1^*)$
5. $k_{00} = k_{10}, k_{01} = k_{11} \rightarrow err = \frac{1}{2} - \theta^* + \theta^*(\phi_0^* + \phi_1^*) - \phi_1^*$
6. $k_{00} > k_{10}, k_{01} = k_{11} \rightarrow err = \frac{1}{2} - \theta^* + \frac{1}{2}(\theta^*(\phi_0^* + \phi_1^*) - \phi_1^*)$
7. $k_{00} < k_{10}, k_{01} > k_{11} \rightarrow err = 1 - 2\theta^* + 2(\theta^*(\phi_0^* + \phi_1^*) - \phi_1^*)$
8. $k_{00} = k_{10}, k_{01} > k_{11} \rightarrow err = 1 - 2\theta^* + \frac{3}{2}(\theta^*(\phi_0^* + \phi_1^*) - \phi_1^*)$
9. $k_{00} > k_{10}, k_{01} > k_{11} \rightarrow err = 1 - 2\theta^* + (\theta^*(\phi_0^* + \phi_1^*) - \phi_1^*)$

Region C: $\theta^* \leq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}$ **and** $\theta^* \geq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*}$

1. $k_{00} < k_{10}, k_{01} < k_{11} \rightarrow err = 2\theta^* - 1 + \phi_1^* - \theta^*(\phi_0^* + \phi_1^*)$
2. $k_{00} = k_{10}, k_{01} < k_{11} \rightarrow err = 2\theta^* - 1 + \frac{3}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
3. $k_{00} > k_{10}, k_{01} < k_{11} \rightarrow err = 2\theta^* - 1 + 2(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
4. $k_{00} < k_{10}, k_{01} = k_{11} \rightarrow err = \theta^* - \frac{1}{2} + \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
5. $k_{00} = k_{10}, k_{01} = k_{11} \rightarrow err = \theta^* - \frac{1}{2} + \phi_1^* - \theta^*(\phi_0^* + \phi_1^*)$
6. $k_{00} > k_{10}, k_{01} = k_{11} \rightarrow err = \theta^* - \frac{1}{2} + \frac{3}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
7. $k_{00} < k_{10}, k_{01} > k_{11} \rightarrow err = 0$
8. $k_{00} = k_{10}, k_{01} > k_{11} \rightarrow err = \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
9. $k_{00} > k_{10}, k_{01} > k_{11} \rightarrow err = \phi_1^* - \theta^*(\phi_0^* + \phi_1^*)$

Region D: $\theta^* \geq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}$ **and** $\theta^* \geq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*}$

1. $k_{00} < k_{10}, k_{01} < k_{11} \rightarrow err = 2\theta^* - 1$
2. $k_{00} = k_{10}, k_{01} < k_{11} \rightarrow err = 2\theta^* - 1 + \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
3. $k_{00} > k_{10}, k_{01} < k_{11} \rightarrow err = 2\theta^* - 1 + \phi_1^* - \theta^*(\phi_0^* + \phi_1^*)$
4. $k_{00} < k_{10}, k_{01} = k_{11} \rightarrow err = \theta^* - \frac{1}{2} - \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
5. $k_{00} = k_{10}, k_{01} = k_{11} \rightarrow err = \theta^* - \frac{1}{2}$
6. $k_{00} > k_{10}, k_{01} = k_{11} \rightarrow err = \theta^* - \frac{1}{2} + \frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
7. $k_{00} < k_{10}, k_{01} > k_{11} \rightarrow err = -(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
8. $k_{00} = k_{10}, k_{01} > k_{11} \rightarrow err = -\frac{1}{2}(\phi_1^* - \theta^*(\phi_0^* + \phi_1^*))$
9. $k_{00} > k_{10}, k_{01} > k_{11} \rightarrow err = 0$

So now, if we have a specific distribution parameterized by $(\theta^*, \phi_0^*, \phi_1^*)$, a sample of size m and a bound ϵ on the estimation error, then first we can use the parameters to figure out which of the four regions we are operating in (based on $\theta^* \geq / \leq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}$ and $\theta^* \geq / \leq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*}$). Next, we look through the nine subcases to identify which ones satisfy $err \leq \epsilon$. Then, to find the probability that $err \leq \epsilon$, we must simply determine the probability of selecting a sample such that the k_{cx} values satisfy one of the relationships defined between the k_{cx} identified in the previous step. These probabilities can be expressed as products of binomials.

As an example, consider the arbitrary case where $\epsilon = 0.05$ and $(\theta^*, \phi_0^*, \phi_1^*) = (0.1, 0.2, 0.7)$. Then this point belongs to Region A, since $\theta^* \leq \frac{\phi_1^*}{\phi_0^*+\phi_1^*}$ and $\theta^* \leq \frac{1-\phi_1^*}{2-\phi_0^*-\phi_1^*}$. Furthermore, by putting the values of $(\theta^*, \phi_0^*, \phi_1^*)$ into the error expressions for the nine subcases, we see

that the only one that satisfies $err < \epsilon$ is the first one, where $k_{00} < k_{10} \wedge k_{01} < k_{11}$. Hence, $P(err < \epsilon) = P(k_{00} < k_{10} \wedge k_{01} < k_{11})$. This can be expressed as:

$$\begin{aligned}
&P(k_{00} < k_{10} \wedge k_{01} < k_{11}) \\
&\quad = \sum_{k_0=0}^{m} \sum_{k_{00}=0}^{k_0} \sum_{k_{10}=0}^{m-k_0} 1_{\{k_{00}<k_{10}\wedge k_{01}<k_{11}\}} \cdot \binom{m}{k_0} \theta^{*k_0} (1-\theta^*)^{m-k_0} \\
&\quad\quad \cdot \binom{k_0}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \binom{m-k_0}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}} \\
&\quad = \sum_{k_0=0}^{\lfloor \frac{m-2}{2} \rfloor} \sum_{k_{00}=0}^{\lfloor \frac{k_0-1}{2} \rfloor} \sum_{k_{10}=0}^{\lfloor \frac{m-k_0-1}{2} \rfloor} \binom{m}{k_0} \theta^{*k_0} (1-\theta^*)^{m-k_0} \\
&\quad\quad \cdot \binom{k_0}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \binom{m-k_0}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}}
\end{aligned}$$

In general, at any particular point in the parameter space, the probability of choosing a sample of size m that yields a model with low error is:

$$\begin{aligned}
P(err < \epsilon) = &\sum_{k_0=0}^{m} \sum_{k_{00}=0}^{k_0} \sum_{k_{10}=0}^{m-k_0} 1_{\{err<\epsilon | k_{00},k_{01},k_{10},k_{11},\theta^*,\phi_0^*,\phi_1^*\}} \cdot \binom{m}{k_0} \binom{k_0}{k_{00}} \binom{m-k_0}{k_{10}} \\
&\cdot \theta^{*k_0} (1-\theta^*)^{m-k_0} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}}
\end{aligned} \tag{4.2}$$

Where $1_{\{err<\epsilon | k_{00},k_{01},k_{10},k_{11},\theta^*,\phi_0^*,\phi_1^*\}}$ is an indicator function that returns 1 whenever k_{00}, k_{01}, k_{10}, k_{11} satisfy a relationship that yields $err < \epsilon$ for the given $(\theta^*, \phi_0^*, \phi_1^*)$.

Hence, for any given $\epsilon, (\theta^*, \phi_0^*, \phi_1^*)$ and m, we can find the probability that a sample of size m will have estimation error $< \epsilon$ if the underlying distribution is defined by $(\theta^*, \phi_0^*, \phi_1^*)$. If we minimize over the cube of all possible values of $(\theta^*, \phi_0^*, \phi_1^*)$, then we can say that no matter what the true underlying distribution is, the probability of having having $err < \epsilon$ for a sample of size m is at least as great as the value at the minimum.

Note that as we vary $(\theta^*, \phi_0^*, \phi_1^*)$, and move throughout the cube, the binomial terms above will change continuously. However, when we hit certain surfaces, then this might change the number of subcases which satisfy $err < \epsilon$, and so the values of $k_{00}, k_{01}, k_{10}, k_{11}$ that we consider acceptable in the indicator function will change as well. When this happens, then there will suddenly be additional or fewer terms that are included in the

summation. As such, the function (4.2) is continuous over the cube except for certain surfaces of discontinuity which correspond to $err = \epsilon$ for any of the subcases.

Furthermore, note that for any point $(\theta^*, \phi_0^*, \phi_1^*)$ that falls within Region A, there is a symmetric point $(1 - \theta^*, \phi_1^*, \phi_0^*)$ that will fall within Region D, and will have the same probability of $err < \epsilon$. Similarly for a point $(\theta^*, \phi_0^*, \phi_1^*)$ that falls within Region B, there is a symmetric point $(1 - \theta^*, \phi_1^*, \phi_0^*)$ that will fall within Region C whose distribution leads to an equal probability of choosing a good sample. So, to find the minimum value over the cube, we can simply find the minimum value over Regions A and B.

Once we know the minimum probability of choosing a good classifier for a fixed m, we will want to find the minimum m such that this probability is sufficiently high. To show that this is a feasible approach to take, we must also show that this probability is monotonically increasing with m.

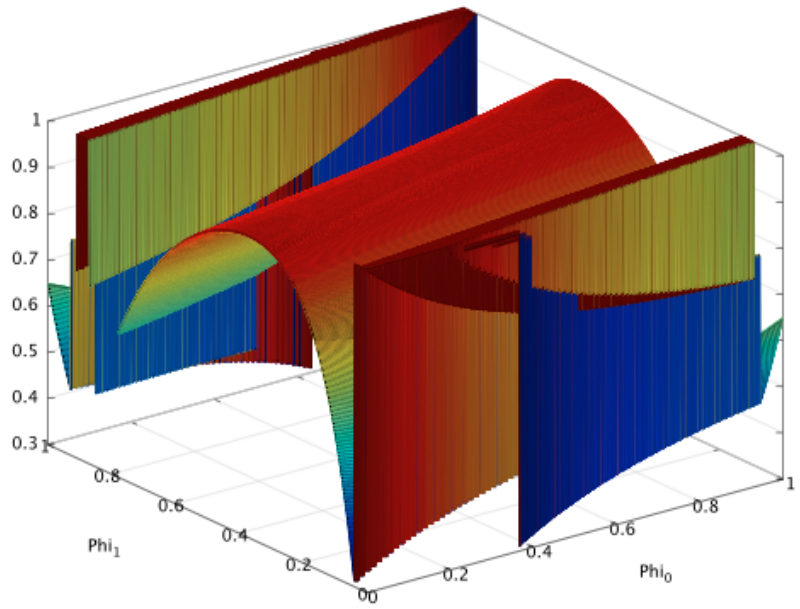
This follows directly from the Central Limit Theorem [19]. This theorem states that the distribution of the mean of sequences of m i.i.d. random variables each with mean μ and finite variance σ^2 converges in probability to a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Hence, in our samples, as m gets higher, then the frequency of any particular feature value will approach a normal distribution about its expected value based on the true distribution, with lower and lower variance. Since the variance lowers as m becomes higher, the probability of having the frequency of a certain feature value less than a fixed distance, such as ϵ, away from its expected value will be higher. $P(err < \epsilon)$ is constructed by having products of the probabilities of having sample frequencies within a fixed distance from their means. Hence, as m increases, so will $P(err < \epsilon)$.

Figure 4.1 provides a visualization of the surface $P(err < \epsilon)$ for fixed $\theta^* = 0.10$ and $m = 10$ and for a given value $\epsilon = 0.17$ as we vary ϕ_0^* and ϕ_1^*. This helps illustrate the symmetry in the surface, and also the way that it has a finite number of lines of discontinuity, with continuous surfaces between them.

The colours are more red where $P(err < \epsilon)$ is high (i.e. we have high probability of selecting a good classifier, and more blue where $P(err < \epsilon)$ is low (i.e. we have low probability of selecting a good classifier).

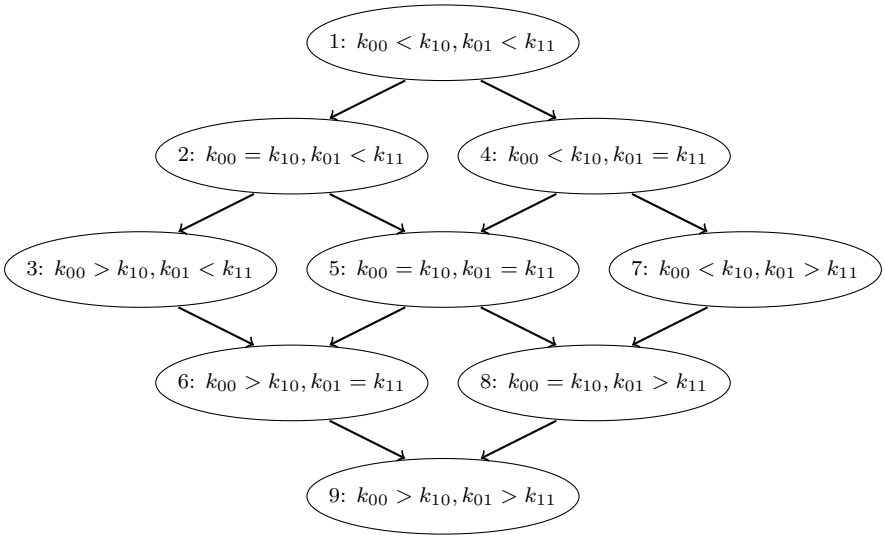
Figure 4.1: Surface illustrating $P(err < \epsilon)$ for $m = 10$, $\theta^* = 0.10$ and $\epsilon = 0.17$.



Note that, based on ϵ, within any of the four regions, we will always be summing over the conditions for the one or more of the 9 settings of k_{00}, k_{01}, k_{10} and k_{11} that yield error $< \epsilon$. However, we can also show that there is a hierarchy amongst the 9 subcases, so if certain subcases satisfy $err < \epsilon$, then there are others that are necessarily smaller, and they must also be included. For example, in Region A, subcase 1 has $err = 0$, so it will always be included. Furthermore, in Region A the error for subcase 2 is half that of subcase 3, so if subcase 3 satisfies $err < \epsilon$, so will subcase 2.

Fully, for Region A, the relationships are as follows in Figure 4.2, where each node indicates the number of a subcase, and an arrow from subcase x to subcase y indicates that the error for subcase x is less than or equal to that for subcase y.

Figure 4.2: Hierarchy of subcase errors in Region A, where a child node has error at least as great as its parents.



Hence, we can see that for any node that has $err < \epsilon$, all of its ancestor nodes must also have $err < \epsilon$. So, depending on the values of $(\theta^*, \phi_0^*, \phi_1^*)$ and ϵ, two example configurations that could lead to $err < \epsilon$ are highlighted below in Figures 4.3 and 4.4.

Figure 4.3: Example of subcases with $err < \epsilon$ in Region A highlighted

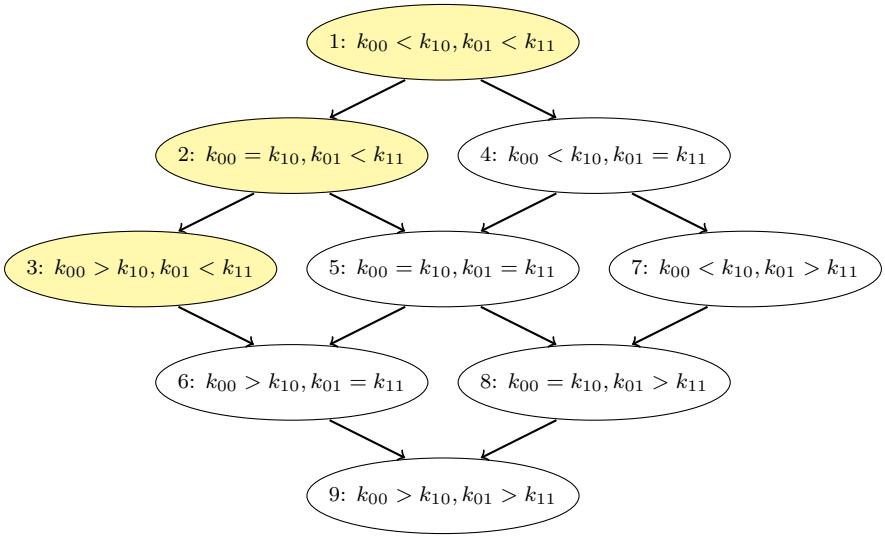
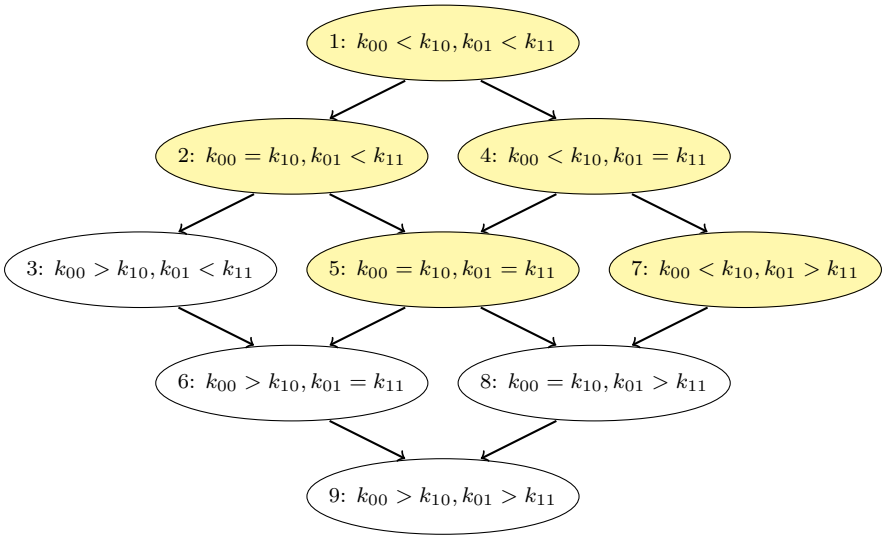


Figure 4.4: Example of subcases with $err < \epsilon$ in Region A highlighted

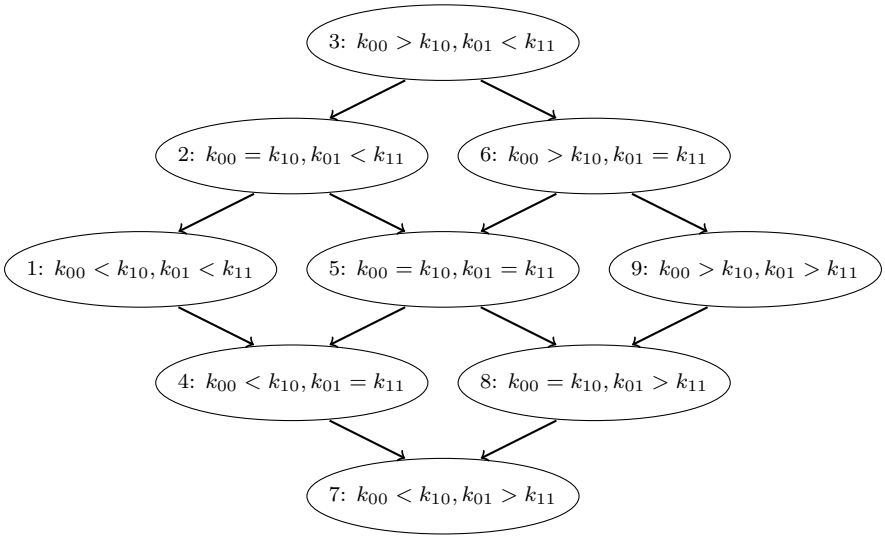


At some places, the conditions can be described using one or two relations that must

both be satisfied simultaneously. For example, in the case where ϵ is so small that $err > \epsilon$ in every subcase except subcase 1, where $err = 0$, then the condition that must be satisfied is $(k_{00} < k_{10} \wedge k_{01} < k_{11})$. In the case described in Figure 4.3, the condition would be $(k_{01} < k_{11})$. These cases where no OR conditions are needed are called "consolidated conditions". If we need to use OR conditions in our logical statement, then these are "unconsolidated conditions". An example of unconsolidated conditions is in Figure 4.4, where the condition is $(k_{00} < k_{10} \vee (k_{00} = k_{10} \wedge k_{01} \leq k_{11}))$.

Similarly, the subcases of Region B follow the hierarchy in Figure 4.5:

Figure 4.5: Hierarchy of subcase errors in Region B, where a child node has error at least as great as its parents.



The subcase hierarchy for Regions C and D are not shown, since, as previous demonstrated, these Regions are symmetric to Regions A and B.

Based on the derivations from this section, I have formulated the problem of finding the minimum amount of data that guarantees agnostic P.A.C. learning as an optimization problem, as stated in Equation (4.1). I detailed how the most challenging part of this optimization task will be the subtask of minimizing (4.2). I have begun to describe this function, and showed how it has surfaces of discontinuity over the cube of the parameter space, but is otherwise continuous.

4.4 Finding Distributions Unlikely to Yield Good Classifiers

In the previous section, I formulated the problem by describing how the most challenging part of this problem is finding the distribution(s) that minimize the probability of choosing a sample that leads to an adequately accurate model. I showed that this will involve minimizing over the unit cube (i.e. the parameter space), and that the objective function (4.2) has surfaces of discontinuity throughout it but is otherwise continuous.

Before proceeding, I would like to define a few concepts that will be prevalent for the rest of the section:

Log-concave function: A function $f : \mathbb{R}^n \to \mathbb{R}_+$ is log-concave if its domain is a convex set and, $\forall a, b \in domain(f)$, $\forall \alpha \in (0,1)$: $f(\alpha a + (1-\alpha)b) \geq f(a)^\alpha f(b)^{1-\alpha}$

Quasi-concave function: A function $f : \mathcal{W} \to \mathbb{R}$ defined on a convex subset $\mathcal{W}$ of a real vector space is quasi-concave if, $\forall a, b \in \mathcal{W}$, $\forall \alpha \in (0,1)$: $f(\alpha a + (1-\alpha)b) \geq min\{f(a), f(b)\}$.

A well known property of log-concave functions is that they are also quasi-concave [10], and a well known property of quasi-concave function is that their minima must be achieved at the boundary of their domain [10].

My goal when I began working on the proof was to complete the following steps:

- Show that in the continuous areas, the function (4.2) is monotonic or log-concave w.r.t. one of its variables
- Show that the function (4.2) is log-concave w.r.t. one of its variables over the continuous areas of the surfaces of discontinuity
- Show that the function (4.2) is log-concave w.r.t. one of its variables over the continuous areas of the faces of the cube
- Identify where, at the intersections of the surfaces of discontinuities and faces, a minimum could occur

The first step would show that in any continuous region of the cube, any minimum must be achieved at its borders (i.e. the surfaces of discontinuity or the cube borders). This is because if, in a certain area, the function is monotonic w.r.t. one of its variables, then by changing that variable and holding the others constant we can move in such a way that the objective function will only get smaller or remain the same. Hence, we would

know that the minimum must be achieved at the border of this continuous area, which is either the cube's border or one of the surfaces of discontinuity. The same can be said if the function is log-concave w.r.t. one of its variables.

We were able to prove this in the consolidated cases (i.e. cases when the conditions on the sample parameters can be expressed without OR statements). The extension to non-consolidated cases is still to follow. A deeper description will follow later in this Section, and the proof will follow in the next Section.

Next, in the second step, I wanted to show that the function is log-concave w.r.t. one of its variables over the continuous areas of the surfaces of discontinuity. If I am able to accomplish this, then we will know that the functions minimum over these surfaces of discontinuity must occur at their borders, i.e. at the borders of the cube or where a surface of discontinuity insects another. However, I have not, as of yet, completed this proof.

In the third step I hope to show, again using log-concavity, that the minimum of the function over any face of the cube must be achieved at it's boundaries or where it meets a surface of discontinuity. Hence, we would know that the minimum must be achieved at the intersection of surfaces of discontinuity or faces of the cube.

Finally, I would optimize over each of these possible intersections to obtain a list of points that could potentially yield the minimum.

Then once I have a finite list of points that are candidate minima for the objective function over the cube, we can simply iterate through them, calculating the probability defined by the objective function (4.2) at each one, to find the true minimum.

For the rest of the discussion, we will focus on step 1, which I was able to prove in the consolidated case. Though I have not yet completed the later steps, this first step is still interesting since it allows us to get a fuller understanding of where within the parameter cube we might encounter the distributions that have the lowest probability of yielding samples that will lead to a sufficiently accurate classifier.

Before introducing the results, I will formally define several terms:

Consolidated area: The consolidated area of the function $P(err < \epsilon)$, as defined in Equation (4.2), is the area where θ^*, ϕ_0^* and ϕ_1^* are such that the conditions on k_{00}, k_{01}, k_{10} and k_{11} can be expressed without the use of use of logical OR conditions. These conditions are determined by looking at the set of conditions on page 38 that satisfy $err < \epsilon$.

Continuous area: The continuous area of the function $P(err < \epsilon)$, as defined in Equation (4.2), is the area where the function is continuous w.r.t. θ^*, ϕ_0^* and ϕ_1^*. This occurs everywhere except where θ^*, ϕ_0^* and ϕ_1^* are such that one of the error expressions for the corresponding regions as defined on page 38 is exactly equal to ϵ.

The results are as follows:

Theorem 5. *In the consolidated continuous area of Region A, the objective function* (4.2) *is monotonically decreasing w.r.t.* θ^**. That is to say, the probability of choosing a sample that yields a classifier that is accurate within* ϵ *of optimal is monotonically decreasing w.r.t.* θ^**. Under the same conditions in Region D,* (4.2) *is monotonic increasing w.r.t.* θ^**.*

Theorem 6. *In the consolidated continuous area of Region B, the objective function* (4.2) *is monotonically increasing w.r.t.* ϕ_0^* *and monotonically decreasing w.r.t.* ϕ_1^**. That is to say, the probability of choosing a sample that yields a classifier that is accurate within* ϵ *of optimal is monotonically increasing w.r.t.* ϕ_0^* *and monotonically decreasing w.r.t.* ϕ_1^**. Under the same conditions in Region C,* (4.2) *is monotonically decreasing w.r.t.* ϕ_0^* *and monotonically increasing w.r.t.* ϕ_1^**.*

Theorem 7. *In the consolidated continuous area of Regions B and C, the objective function* (4.2) *is log concave w.r.t.* θ^**. That is to say, the probability of choosing a sample that yields a classifier that is accurate within* ϵ *of optimal is log-concave w.r.t.* θ^**.*

The proofs for Theorems 5 and 6 were primarily developed by my collaborator George Trimponias of Huawei. Since they are not my original work I will omit them from the body of this thesis. I will nonetheless include them for the reader's reference in Appendix A. However, I will still discuss their impact here, since this discussion was done collaboratively and supports the understanding of the structure of the objective function. The proof of Theorem 7 will be included in Section 4.5.

Based on Theorem 5, we know that whenever we are within the consolidated continuous area of Region A or Region D, the function will be monotonic w.r.t. θ^*. If we start at any point inside this area, then by keeping ϕ_0^* and ϕ_1^* constant and varying θ^*, we will be able to move to the edge of the continuous area (we must eventually hit either a surface of discontinuity or the border of the cube), while either maintaining or lowering the objective function (4.2). Hence, within the consolidated areas of Regions A and D, the minimum value of $P(err < \epsilon)$ must be achieved either on the borders of the cube or at one of the surfaces of discontinuity.

Theorems 6 and 7 describe the characteristics of the consolidated continuous area of Regions B and C. In Theorem 6, I show that within this area the objective function (4.2) is monotonic w.r.t. both ϕ_0^* and ϕ_1^*. Similarly to the results for Theorem 5, we know that by keeping θ^* and one of the ϕ^* variables constant, and then varying the other, we will again move to the edge of the continuous area, while either maintaining or lowering the objective function (4.2). So again, within the consolidated areas of region B and C, we

know that the minimum value $P(err < \epsilon)$ must be achieved on either the borders of the cube or at one of the surfaces of discontinuity.

Theorem 7 shows that within the consolidated continuous area of Regions B and C, that (4.2) is log-concave w.r.t. θ^*. As previously mentioned, it is well known that functions that are log-concave with a closed domain achieve their minimum at the border of their domain [10]. Hence, at any point in the consolidated continuous area of Regions B and C, by keeping ϕ_0^* and ϕ_1^* constant then the minimum along that line must be achieved by either setting θ^* to it's minimum or maximum value within that area. Hence, the minimum of the objective is again achieved on either the borders of the cube or at one of the surfaces of discontinuity.

The combination of Theorem 5 with either Theorem 6 or Theorem 7 is sufficient to complete step 1 from the proof strategy within the consolidated continuous areas. Though the proof for Theorem 6 is much shorter and perhaps simpler than the proof for Theorem 7, I pursued the proof for Theorem 7 nonetheless because it allows us to more fully characterize how the probability of choosing a sample that yields a bad classifier changes throughout the cube. It also allows us to completely understand how that probability changes as θ^* changes over the consolidated continuous regions of the cube. Furthermore, I believe that the proof of Theorem 7 will help us in later, as of yet incomplete, advances of this work. Hence, the proof of Theorem 7 is important for understanding and future work, and is presented in the following Section.

4.5 Proof of Theorem 7

This section will detail the proof of Theorem 7. As mentioned previously, the proofs for Theorems 5 and 6 were done by my collaborator George and are included in Appendix A.

First, however, I will give a brief overview of the idea behind the proof. Theorem 7 states that the function (4.2) is log concave with respect to θ^* in certain circumstances. Based on a result from [29] , for a mixture of consecutive binomial terms, if the coefficients form a log concave sequence, then the mixture is log concave with respect to its parameter.

In this case, the function is of the form:

$$\sum_{k_0=0}^{m-1} \binom{m}{k_0} \theta^{*k_0} (1-\theta^*)^{m-k_0} \sum_{k_{00}=max(1,2k_0-m+1)}^{k_0} \sum_{k_{10}=0}^{min(k_{00}-1,m-2k_0+k_{00}-1)} \binom{k_0}{k_{00}} \binom{m-k_0}{k_{10}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}}$$

This is simply a mixture of binomial terms $\binom{m}{k_0}\theta^{*k_0}(1-\theta^*)^{m-k_0}$ with coefficients:

$$c_{k_0} = \sum_{k_{00}=max(1,2k_0-m+1)}^{k_0} \sum_{k_{10}=0}^{min(k_{00}-1,m-2k_0+k_{00}-1)} \binom{k_0}{k_{00}}\binom{m-k_0}{k_{10}} \cdot \phi_0^{*k_{00}}(1-\phi_0^*)^{k_0-k_{00}}\phi_1^{*k_{10}}(1-\phi_1^*)^{m-k_0-k_{10}}$$

Hence, to show that the entire function is log concave with respect to θ^*, it is sufficient to show that the series of coefficients c_{k_0} is log concave. That is, it is sufficient to show that $c_{k_0}^2 \geq c_{k_0-1} \cdot c_{k_0+1}\ \forall k_0$.

In the formal proof I will show that when we multiply two expressions of the form c_{k_0} together, we get a sum of the form shown in Lemma 4. In these expressions, τ is the sum of the k_{00} values from both of the c_{k_0} terms, and τ_1 is the k_{00} value attributed to the first c_{k_0} term. Similarly, λ is the sum of the k_{10} values from both of the c_{k_0} terms, and λ_1 is the k_{10} value attributed to the first c_{k_0} term.

Lemma 4 shows a combinatorial inequality that is sufficient to show that the series c_{k_0} is log concave. The full derivation of this inequality follows in the proof of Theorem 7. The proof of Lemma 4 follows at the end of this section.

Lemma 4. *Suppose $m \in \mathbb{N}$. Then, $\forall k_0, \tau, \lambda$ s.t. $1 \leq k_0 \leq \frac{m-2}{2}$, $2 \leq \tau \leq 2k_0$, $0 \leq \lambda \leq \tau-2$:*

$$\sum_{\tau_1=max(1,\tau-k_0)}^{min(\tau-1,k_0)} \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1} \sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)} \binom{m-k_0}{\lambda_1}\binom{m-k_0}{\lambda-\lambda_1} \tag{4.3}$$
$$\geq \sum_{\tau_1=max(1,\tau-k_0-1)}^{min(k_0-1,\tau-1)} \binom{k_0-1}{\tau_1}\binom{k_0+1}{\tau-\tau_1} \sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)} \binom{m-k_0+1}{\lambda_1}\binom{m-k_0-1}{\lambda-\lambda_1}$$

Theorem 7. *In the consolidated continuous area of Regions B and C, the objective function (4.2) is log concave w.r.t. θ^*. That is to say, the probability of choosing a sample that yields a classifier that is accurate within ϵ of optimal is log-concave w.r.t. θ^*.*

Proof. Start by noting that if we prove this result in Region B, then by symmetry the result will also hold in Region C. In the consolidated area of region B, the objective function will have the form:

$$P(err < \epsilon) = \sum_{k_0=0}^{m} \sum_{k_{00}=0}^{k_0} \sum_{k_{10}=0}^{m-k_0} 1_{\{k_{00} \dagger k_{10} \wedge k_{01} \ddagger k_{11}\}} \cdot \binom{m}{k_0} \binom{k_0}{k_{00}} \binom{m-k_0}{k_{10}}$$
$$\cdot \theta^{*k_0} (1-\theta^*)^{m-k_0} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}}$$

Where † means that, if there is a relational requirement between k_{00} and k_{10}, it will be either $k_{00} > k_{10}$ or $k_{00} \geq k_{10}$, and ‡ indicates that any relational requirement between k_{01} and k_{11} will be either $k_{01} < k_{11}$ or $k_{01} \leq k_{11}$.

Depending on the accuracy threshold ϵ, and the values of the parameters $(\theta^*, \phi_0^*, \phi_1^*)$, different sample characteristics will yield a model that is sufficiently accurate. Because we are in a consolidated area of Region B, we know that the condition on the sample parameters can be expressed using † and ‡.

For this proof I will explicitly consider the case where the condition is of the form $\{k_{00} > k_{10} \wedge k_{01} < k_{11}\}$, however the exact same proof method can be used in any of the consolidated cases.

In this case, the objective function will have the form:

$$P(err < \epsilon) = \sum_{k_0=0}^{m} \sum_{k_{00}=0}^{k_0} \sum_{k_{10}=0}^{m-k_0} 1_{\{k_{00} > k_{10} \wedge k_{01} < k_{11}\}} \cdot \binom{m}{k_0} \binom{k_0}{k_{00}} \binom{m-k_0}{k_{10}}$$
$$\cdot \theta^{*k_0} (1-\theta^*)^{m-k_0} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}}$$

Note that the condition $k_{11} > k_{01}$ implies $k_{11} \geq 1$, which means that we must have $k_1 \geq 1$ and $k_0 < m$.

Now, note that the second condition can be rewritten:

$$\begin{aligned} k_{01} < k_{11} &\Leftrightarrow k_0 - k_{00} < m - k_0 - k_{10} \\ &\Leftrightarrow k_{10} < m - 2k_0 + k_{00} \end{aligned}$$

Furthermore, since $k_{10} \geq 0$, we must have that $m - 2k_0 + k_{00} > 0$, and hence $k_{00} > 2k_0 - m$.

Replacing this into the objective function yields:

$$
\begin{aligned}
P(err < \epsilon) = &\sum_{k_0=0}^{m} \sum_{k_{00}=0}^{k_0} \sum_{k_{10}=0}^{m-k_0} 1_{\{k_{10}<k_{00} \wedge k_{10}<m-2k_0+k_{00}\}} \cdot \binom{m}{k_0}\binom{k_0}{k_{00}}\binom{m-k_0}{k_{10}} \\
&\cdot \theta^{*k_0}(1-\theta^*)^{m-k_0}\phi_0^{*k_{00}}(1-\phi_0^*)^{k_0-k_{00}}\phi_1^{*k_{10}}(1-\phi_1^*)^{m-k_0-k_{10}} \\
= &\sum_{k_0=0}^{m-1} \sum_{k_{00}=max(1,2k_0-m+1)}^{k_0} \sum_{k_{10}=0}^{min(k_{00}-1,m-2k_0+k_{00}-1)} \binom{m}{k_0}\binom{k_0}{k_{00}}\binom{m-k_0}{k_{10}} \\
&\cdot \theta^{*k_0}(1-\theta^*)^{m-k_0}\phi_0^{*k_{00}}(1-\phi_0^*)^{k_0-k_{00}}\phi_1^{*k_{10}}(1-\phi_1^*)^{m-k_0-k_{10}} \\
= &\sum_{k_0=0}^{m-1} \binom{m}{k_0}\theta^{*k_0}(1-\theta^*)^{m-k_0} \sum_{k_{00}=max(1,2k_0-m+1)}^{k_0} \sum_{k_{10}=0}^{min(k_{00}-1,m-2k_0+k_{00}-1)} \\
&\binom{k_0}{k_{00}}\binom{m-k_0}{k_{10}}\phi_0^{*k_{00}}(1-\phi_0^*)^{k_0-k_{00}}\phi_1^{*k_{10}}(1-\phi_1^*)^{m-k_0-k_{10}} \qquad (4.4)
\end{aligned}
$$

Equation (4.4) is simply a mixture of binomial terms $\binom{m}{k_0}\theta^{*k_0}(1-\theta^*)^{m-k_0}$ with coefficients:

$$
\begin{aligned}
c_{k_0} = &\sum_{k_{00}=max(1,2k_0-m+1)}^{k_0} \sum_{k_{10}=0}^{min(k_{00}-1,m-2k_0+k_{00}-1)} \binom{k_0}{k_{00}}\binom{m-k_0}{k_{10}} \\
&\cdot \phi_0^{*k_{00}}(1-\phi_0^*)^{k_0-k_{00}}\phi_1^{*k_{10}}(1-\phi_1^*)^{m-k_0-k_{10}}
\end{aligned}
$$

As shown in [29], for a mixture of consecutive binomial terms, if the coefficients form a log concave sequence, then the mixture is log concave with respect to its parameter.

In this case, that means that if the sequence of coefficients c_{k_0} is log concave, then the function (4.4) is log concave w.r.t. θ^*. Hence, to prove our result, we need only show that the sequence c_{k_0} is log concave, i.e. $c_{k_0}^2 > c_{k_0-1} \cdot c_{k_0+1}\ \forall k_0 \in \{1, ..., m-2\}$.

To remove the min and max statements and simplify the expression for c_{k_0}, the proof will be divided into three cases. The first will cover the case when k_0 is such that $m - 2k_0 \geq 2$, the second will cover the case when $m - 2k_0 \leq -2$ and the third will be the case that $m - 2k_0 \in \{-1, 0, 1\}$.

For the first potential k_0 values, consider the case where $m - 2k_0 \geq 2$. This also means that $m - 2k_0 \geq 0$, $m - 2(k_0 - 1) \geq 0$ and $m - 2(k_0 + 1) \geq 0$. Then, the mixture coefficients will be:

$$c_{k_0} = \sum_{k_{00}=1}^{k_0} \binom{k_0}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \sum_{k_{10}=0}^{k_{00}-1} \binom{m-k_0}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}}$$

I will now show that $c_{k_0}^2 \geq c_{k_0+1} \cdot c_{k_0-1}$. Start by determining the expressions on each side of the inequality. τ will be the sum of the k_{00} value from both c_{k_0} expressions, and τ_1 will be the k_{00} value from the first. λ will be the sum of the k_{10} value from both c_{k_0} expressions, and λ_1 will be the k_{10} value from the first.

$$\begin{aligned}
c_{k_0}^2 =& \left\{ \sum_{k_{00}=1}^{k_0} \binom{k_0}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \sum_{k_{10}=0}^{k_{00}-1} \binom{m-k_0}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}} \right\}^2 \\
=& \sum_{\tau=2}^{2k_0} \sum_{\tau_1=max(1,\tau-k_0)}^{min(\tau-1,k_0)} \binom{k_0}{\tau_1} \binom{k_0}{\tau-\tau_1} \phi_0^{*\tau} (1-\phi_0^*)^{2k_0-\tau} \\
& \left(\sum_{k_{10}=0}^{\tau_1-1} \binom{m-k_0}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}} \right) \left(\sum_{k_{10}=0}^{\tau-\tau_1-1} \binom{m-k_0}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}} \right) \\
=& \sum_{\tau=2}^{2k_0} \sum_{\tau_1=max(1,\tau-k_0)}^{min(\tau-1,k_0)} \binom{k_0}{\tau_1} \binom{k_0}{\tau-\tau_1} \phi_0^{*\tau} (1-\phi_0^*)^{2k_0-\tau} \\
& \sum_{\lambda=0}^{\tau-2} \sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)} \binom{m-k_0}{\lambda_1} \binom{m-k_0}{\lambda-\lambda_1} \phi_1^{*\lambda} (1-\phi_1^*)^{2m-2k_0-\lambda}
\end{aligned}$$

$$
\begin{aligned}
& c_{k_0+1} \cdot c_{k_0-1} \\
=& \left\{ \sum_{k_{00}=1}^{k_0+1} \binom{k_0+1}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0+1-k_{00}} \sum_{k_{10}=0}^{k_{00}-1} \binom{m-k_0-1}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-1-k_{10}} \right\} \\
& \cdot \left\{ \sum_{k_{00}=1}^{k_0-1} \binom{k_0-1}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-1-k_{00}} \sum_{k_{10}=0}^{k_{00}-1} \binom{m-k_0+1}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0+1-k_{10}} \right\} \\
=& \sum_{\tau=2}^{2k_0} \sum_{\tau_1=max(1,\tau-k_0-1)}^{min(k_0-1,\tau-1)} \binom{k_0-1}{\tau_1} \binom{k_0+1}{\tau-\tau_1} \phi_0^{*\tau} (1-\phi_0^*)^{2k_0-\tau} \\
& \left(\sum_{k_{10}=0}^{\tau_1-1} \binom{m-k_0+1}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0+1-k_{10}} \right) \\
& \left(\sum_{k_{10}=0}^{\tau-\tau_1-1} \binom{m-k_0-1}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-1-k_{10}} \right) \\
=& \sum_{\tau=2}^{2k_0} \sum_{\tau_1=max(1,\tau-k_0-1)}^{min(k_0-1,\tau-1)} \binom{k_0-1}{\tau_1} \binom{k_0+1}{\tau-\tau_1} \phi_0^{*\tau} (1-\phi_0^*)^{2k_0-\tau} \\
& \sum_{\lambda=0}^{\tau-2} \sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)} \binom{m-k_0+1}{\lambda_1} \binom{m-k_0-1}{\lambda-\lambda_1} \phi_1^{*\lambda} (1-\phi_1^*)^{2m-2k_0-\lambda}
\end{aligned}
$$

Hence, to show that $c_{k_0}^2 \geq c_{k_0+1} \cdot c_{k_0-1}$, we must show that:

$$\sum_{\tau=2}^{2k_0}\sum_{\lambda=0}^{\tau-2}\phi_0^{*\tau}(1-\phi_0^*)^{2k_0-\tau}\phi_1^{*\lambda}(1-\phi_1^*)^{2m-2k_0-\lambda}$$
$$\sum_{\tau_1=max(1,\tau-k_0)}^{min(\tau-1,k_0)}\binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1}\sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)}\binom{m-k_0}{\lambda_1}\binom{m-k_0}{\lambda-\lambda_1}$$
$$\geq$$
$$\sum_{\tau=2}^{2k_0}\sum_{\lambda=0}^{\tau-2}\phi_0^{*\tau}(1-\phi_0^*)^{2k_0-\tau}\phi_1^{*\lambda}(1-\phi_1^*)^{2m-2k_0-\lambda}$$
$$\sum_{\tau_1=max(1,\tau-k_0-1)}^{min(k_0-1,\tau-1)}\binom{k_0-1}{\tau_1}\binom{k_0+1}{\tau-\tau_1}\sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)}\binom{m-k_0+1}{\lambda_1}\binom{m-k_0-1}{\lambda-\lambda_1}$$

Note that the limits of summation over τ and λ are the same for $c_{k_0}^2$ and $c_{k_0+1}\cdot c_{k_0-1}$, as well as the ϕ_0^* and ϕ_1^* terms. Hence, to show that the inequality holds over the entire summation, it is sufficient to show that it holds over each term in the summations over τ and λ.

To prove our result, it is hence sufficient to show that the following inequality holds $\forall 2 \leq \tau \leq 2k_0$, $0 \leq \lambda \leq \tau - 2$:

$$\sum_{\tau_1=max(1,\tau-k_0)}^{min(\tau-1,k_0)}\binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1}\sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)}\binom{m-k_0}{\lambda_1}\binom{m-k_0}{\lambda-\lambda_1} \tag{4.5}$$
$$\geq \sum_{\tau_1=max(1,\tau-k_0-1)}^{min(k_0-1,\tau-1)}\binom{k_0-1}{\tau_1}\binom{k_0+1}{\tau-\tau_1}\sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)}\binom{m-k_0+1}{\lambda_1}\binom{m-k_0-1}{\lambda-\lambda_1}$$

By Lemma 4, we know that this inequality holds.

So, the series of coefficients c_{k_0} is log concave when $k_0 \leq \frac{m-2}{2}$.

Next, consider the case where $k_0 \geq \frac{m+2}{2}$. This is equivalent to looking at the terms c_{m-k_0} when $k_0 \leq \frac{m-2}{2}$. So, assume that $k_0 \leq \frac{m-2}{2}$ and then we can change the order of

summation and do variable substitution:

$$
\begin{aligned}
&c_{m-k_0} \\
&= \sum_{k_{00}=m-2k_0+1}^{m-k_0} \sum_{k_{10}=0}^{-m+2k_0+k_{00}-1} \binom{m-k_0}{k_{00}} \binom{k_0}{k_{10}} \phi_0^{*k_{00}} (1-\phi_0^*)^{m-k_0-k_{00}} \phi_1^{*k_{10}} (1-\phi_1^*)^{k_0-k_{10}} \\
&= \sum_{k_{10}=0}^{k_0-1} \sum_{k_{00}=m-2k_0+k_{10}+1}^{m-k_0} \binom{m-k_0}{k_{00}} \binom{k_0}{k_{10}} \phi_0^{*k_{00}} (1-\phi_0^*)^{m-k_0-k_{00}} \phi_1^{*k_{10}} (1-\phi_1^*)^{k_0-k_{10}} \\
&= \sum_{k_{11}=m-2k_0+1}^{m-k_0} \sum_{k_{01}=2k_0-m}^{4k_0-2m+k_{11}-1} \binom{m-k_0}{k_0-k_{01}} \binom{k_0}{m-k_0-k_{11}} \\
&\qquad \cdot \phi_0^{*k_0-k_{01}} (1-\phi_0^*)^{m-2k_0+k_{01}} \phi_1^{*m-k_0-k_{11}} (1-\phi_1^*)^{2k_0-m+k_{11}} \\
&= \sum_{k_{11}=1}^{k_0} \sum_{k_{01}=2k_0-m}^{2k_0-m+k_{11}-1} \binom{m-k_0}{k_0-k_{01}} \binom{k_0}{k_0-k_{11}} \phi_0^{*k_0-k_{01}} (1-\phi_0^*)^{m-2k_0+k_{01}} \phi_1^{*k_0-k_{11}} (1-\phi_1^*)^{k_{11}} \\
&= \sum_{k_{11}=1}^{k_0} \sum_{k_{01}=0}^{k_{11}-1} \binom{m-k_0}{k_{01}} \binom{k_0}{k_{11}} \phi_0^{*m-k_0-k_{01}} (1-\phi_0^*)^{k_{01}} \phi_1^{*k_0-k_{11}} (1-\phi_1^*)^{k_{11}}
\end{aligned}
$$

Note here that once we have rewritten c_{m-k_0} as above, the limits of the summations and the binomial terms are identical to those found in the expansion of c_{k_0}.

In the part of the proof for $k_0 \leq \frac{m-2}{2}$, we saw that when we multiply the coefficients together, the ϕ_0^* and ϕ_1^* terms in $c_{k_0}^2$ and $c_{k_0+1} \cdot c_{k_0-1}$ cancel out. Hence, in this case, once those terms are removed, we will be left with an expression identical to Inequality (4.5), which we already know holds. Hence, when $k_0 \geq \frac{m+2}{2}$, then c_{k_0} is log concave.

The final case to consider is where $k_0 \in \{\frac{m-1}{2}, \frac{m}{2}, \frac{m+1}{2}\}$. This case is quite simple. Consider, for example, $k_0 = \frac{m}{2}$.

Then:

$$
\begin{aligned}
c_{k_0} &= \sum_{k_{00}=1}^{k_0} \sum_{k_{10}=0}^{k_{00}-1} \binom{k_0}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-k_{00}} \binom{m-k_0}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-k_{10}} \\
c_{k_0-1} &= \sum_{k_{00}=1}^{k_0-1} \sum_{k_{10}=0}^{k_{00}-1} \binom{k_0-1}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0-1-k_{00}} \binom{m-k_0+1}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0+1-k_{10}} \\
c_{k_0+1} &= \sum_{k_{00}=1}^{k_0+1} \sum_{k_{10}=0}^{k_{00}-3} \binom{k_0+1}{k_{00}} \phi_0^{*k_{00}} (1-\phi_0^*)^{k_0+1-k_{00}} \binom{m-k_0-1}{k_{10}} \phi_1^{*k_{10}} (1-\phi_1^*)^{m-k_0-1-k_{10}}
\end{aligned}
$$

In this case, however, we see that the expressions for c_{k_0} and c_{k_0-1} are identical to those that we saw in the derivation for the case when $k_0 \leq \frac{m-2}{2}$. The expression for c_{k_0+1} is similar, except that the upper bound of summation over k_{10} is lower. Hence, we have the same expression for $c_{k_0}^2$, but $c_{k_0+1} \cdot c_{k_0-1}$ will be less. Hence, $c_{k_0}^2 \geq c_{k_0+1} \cdot c_{k_0-1}$ still holds. The same is true for the cases that $k_0 = \frac{m-1}{2}$ and $k_0 = \frac{m+1}{2}$.

So, the series c_{k_0} is log concave. By the result in [29], the mixture $\sum_{k_0=0}^{m-1} c_{k_0} \binom{m}{k_0} \theta^{*k_0}(1-\theta^*)^{m-k_0}$ is log concave w.r.t. θ^*.

This concludes the proof of Theorem 7.

□

Now, I will show that the Lemma does indeed hold. Recall Lemma 4:

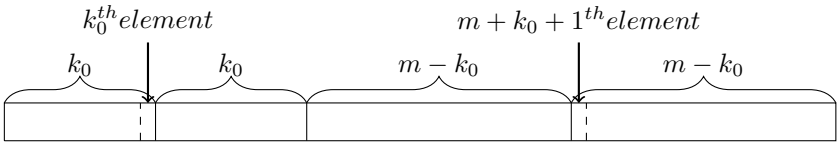
Lemma 4. *Suppose $m \in \mathbb{N}$. Then, $\forall k_0, \tau, \lambda$ s.t. $1 \leq k_0 \leq \frac{m-2}{2}$, $2 \leq \tau \leq 2k_0$, $0 \leq \lambda \leq \tau - 2$:*

$$\begin{aligned} &\sum_{\tau_1=max(1,\tau-k_0)}^{min(\tau-1,k_0)} \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1} \sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)} \binom{m-k_0}{\lambda_1}\binom{m-k_0}{\lambda-\lambda_1} \\ &\geq \sum_{\tau_1=max(1,\tau-k_0-1)}^{min(k_0-1,\tau-1)} \binom{k_0-1}{\tau_1}\binom{k_0+1}{\tau-\tau_1} \sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)} \binom{m-k_0+1}{\lambda_1}\binom{m-k_0-1}{\lambda-\lambda_1} \end{aligned} \tag{4.6}$$

Proof. First consider the case where $1 > \tau - k_0$, and this becomes:

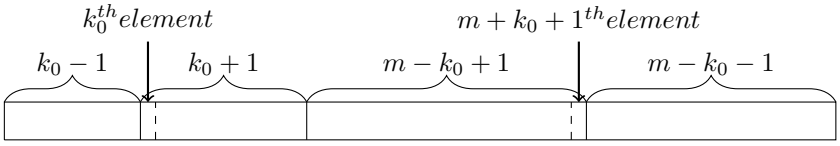
$$\begin{aligned} &\sum_{\tau_1=1}^{\tau-1} \sum_{\lambda_1=max\{0,\tau_1+\lambda+1-\tau\}}^{min\{\tau_1-1,\lambda\}} \binom{k_0}{\tau_1} \cdot \binom{k_0}{\tau-\tau_1} \cdot \binom{m-k_0}{\lambda_1} \cdot \binom{m-k_0}{\lambda-\lambda_1} \\ &\geq \sum_{\tau_1=1}^{\tau-1} \sum_{\lambda_1=max\{0,\tau_1+\lambda+1-\tau\}}^{min\{\tau_1-1,\lambda\}} \binom{k_0-1}{\tau_1} \cdot \binom{k_0+1}{\tau-\tau_1} \cdot \binom{m-k_0+1}{\lambda_1} \cdot \binom{m-k_0-1}{\lambda-\lambda_1} \end{aligned} \tag{4.7}$$

To show that inequality (4.7) holds, first consider the left side of the inequality. This amounts to the number of combinations that can be made from $2m$ elements such that, if we consider the elements divided up into two sections of size k_0 and two sections of size $m - k_0$, the following conditions must hold:



- 1.1: from the first section of k_0 elements, at least 1 is chosen
- 1.2: from the second section of k_0 elements, at least 1 is chosen
- 1.3: between the first and second sections of k_0 elements, τ are chosen
- 1.4: from the first section of $m - k_0$ elements, fewer elements are chosen than from the first section of k_0 elements
- 1.5: from the second section of $m - k_0$ elements, fewer elements are chosen than from the second section of k_0 elements
- 1.6: between the first and second sections of $m - k_0$ elements, λ are chosen

Similarly, the right side of inequality (4.7) amounts to the number of combinations that can be made from $2m$ elements such that, if we consider the elements divided up into a sections of size $k_0 - 1, k_0 + 1, m - k_0 + 1$ and $m - k_0 - 1$, the following conditions must hold:



- 2.1: from the section of $k_0 - 1$ elements, at least 1 is chosen
- 2.2: from the section of $k_0 + 1$ elements, at least 1 is chosen
- 2.3: between the sections of $k_0 + 1$ and $k_0 - 1$ elements, τ are chosen
- 2.4: from the section of $m - k_0 + 1$ elements, fewer elements are chosen than from the section of $k_0 - 1$ elements
- 2.5: from the section of $m - k_0 - 1$ elements, fewer elements are chosen than from the section of $k_0 + 1$ elements

- 2.6: between the sections of $m - k_0 + 1$ and $m - k_0 - 1$ elements, λ are chosen

So, showing that inequality (4.7) holds is equivalent to showing that the number of combinations of $2m$ elements that satisfy the first set of conditions but not the second set is greater than the number of combinations that satisfy the second set but not the first set.

Start by considering the combinations that satisfy the first set of conditions but not the second. By satisfying the first set of conditions, conditions $2.2, 2.3, 2.5$ and 2.6 are also necessarily satisfied. Also, note that if condition 2.1 is not satisfied, then neither is condition 2.4. So, we only need to count the number of combinations that satisfy the first set of conditions but fail condition 2.4.

Since the first set of conditions are satisfied, we know that the number of elements in the first section of $m-k_0$, denoted n_{m-k_0}, is less than the number of elements in the first section of k_0, n_{k_0}. If $n_{m-k_0} = n_{k_0} - 1$, then condition 2.4 will fail if either the $m + k_0 + 1^{th}$ element is selected or if the k_0^{th} element is selected and the $m + k_0 + 1^{th}$ is not. If $n_{m-k_0} = n_{k_0} - 2$, then condition 2.4 will fail if both the k_0^{th} and $m + k_0 + 1^{th}$ elements are selected.

Hence, the number of combinations that satisfy the first set of conditions but not the second is:

$$\sum_{\tau_1=1}^{\lambda} \Bigg[\binom{k_0}{\tau_1} \binom{k_0}{\tau - \tau_1} \binom{m-k_0}{\tau_1 - 1} \binom{m-k_0-1}{\lambda - \tau_1} + \binom{k_0 - 1}{\tau_1} \binom{k_0}{\tau - \tau_1 - 1} \binom{m-k_0}{\tau_1 - 1} \binom{m-k_0-1}{\lambda - \tau_1} \Bigg] + \sum_{\tau_1=1}^{\lambda+1} \binom{k_0 - 1}{\tau_1 - 1} \binom{k_0}{\tau - \tau_1} \binom{m-k_0}{\tau_1 - 1} \binom{m-k_0-1}{\lambda - \tau_1 + 1}$$

Next, consider the number of combinations that satisfy the second set of conditions but not the first. By satisfying the second set of conditions, conditions $1.1, 1.3, 1.4$ and 1.6 are also necessarily satisfied. Also note that if condition 1.2 is not satisfied, then neither is condition 1.5. So, we only need to count the number of combinations that satisfy the second set of conditions but fail condition 1.5.

Since the second set of conditions are satisfied, we know that the number of elements in the section of $m - k_0 - 1$, denoted n_{m-k_0-1}, is less than the number of elements in the section of $k_0 + 1$, n_{k_0+1}. If $n_{m-k_0-1} = n_{k_0+1} - 1$, then condition 1.5 will fail if either the

$m + k_0 + 1^{th}$ element is selected or if the k_0^{th} element is selected and the $m + k_0 + 1^{th}$ is not. If $n_{m-k_0-1} = n_{k_0+1} - 2$, then condition 1.5 will fail if both the k_0^{th} and $m + k_0 + 1^{th}$ elements are selected.

So, the number of combinations that satisfy the second set of conditions but not the first is:

$$\sum_{\tau_1=1}^{\lambda} \Bigg[\binom{k_0-1}{\tau-\tau_1}\binom{k_0+1}{\tau_1}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1} + \binom{k_0-1}{\tau-\tau_1-1}\binom{k_0}{\tau_1}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1}\Bigg] + \sum_{\tau_1=1}^{\lambda+1} \binom{k_0-1}{\tau-\tau_1}\binom{k_0}{\tau_1-1}\binom{m-k_0}{\lambda-\tau_1+1}\binom{m-k_0-1}{\tau_1-1}$$

Showing that Inequality 4.7 holds is equivalent to showing that the following holds:

$$\begin{aligned} &\sum_{\tau_1=1}^{\lambda} \Bigg[\binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1}\binom{m-k_0}{\tau_1-1}\binom{m-k_0-1}{\lambda-\tau_1} + \binom{k_0-1}{\tau_1}\binom{k_0}{\tau-\tau_1-1}\binom{m-k_0}{\tau_1-1}\binom{m-k_0-1}{\lambda-\tau_1}\Bigg] + \sum_{\tau_1=1}^{\lambda+1} \binom{k_0-1}{\tau_1-1}\binom{k_0}{\tau-\tau_1}\binom{m-k_0}{\tau_1-1}\binom{m-k_0-1}{\lambda-\tau_1+1} \\ \geq &\sum_{\tau_1=1}^{\lambda} \Bigg[\binom{k_0-1}{\tau-\tau_1}\binom{k_0+1}{\tau_1}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1} + \binom{k_0-1}{\tau-\tau_1-1}\binom{k_0}{\tau_1}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1}\Bigg] + \sum_{\tau_1=1}^{\lambda+1} \binom{k_0-1}{\tau-\tau_1}\binom{k_0}{\tau_1-1}\binom{m-k_0}{\lambda-\tau_1+1}\binom{m-k_0-1}{\tau_1-1} \end{aligned} \tag{4.8}$$

To show that Inequality (4.8) holds, and hence that Inequality (4.7) holds, it is sufficient to break it into three inequalities and show that they each hold:

$$\sum_{\tau_1=1}^{\lambda} \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1}\binom{m-k_0}{\tau_1-1}\binom{m-k_0-1}{\lambda-\tau_1} \geq \sum_{\tau_1=1}^{\lambda} \binom{k_0-1}{\tau-\tau_1}\binom{k_0+1}{\tau_1}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1} \tag{4.9}$$

$$\sum_{\tau_1=1}^{\lambda} \binom{k_0-1}{\tau_1}\binom{k_0}{\tau-\tau_1-1}\binom{m-k_0}{\tau_1-1}\binom{m-k_0-1}{\lambda-\tau_1} \geq \sum_{\tau_1=1}^{\lambda} \binom{k_0-1}{\tau-\tau_1-1}\binom{k_0}{\tau_1}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1} \tag{4.10}$$

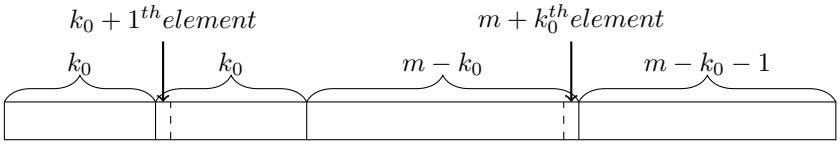
$$\sum_{\tau_1=1}^{\lambda+1} \binom{k_0-1}{\tau_1-1}\binom{k_0}{\tau-\tau_1}\binom{m-k_0}{\tau_1-1}\binom{m-k_0-1}{\lambda-\tau_1+1} \geq \sum_{\tau_1=1}^{\lambda+1} \binom{k_0-1}{\tau-\tau_1}\binom{k_0}{\tau_1-1}\binom{m-k_0}{\lambda-\tau_1+1}\binom{m-k_0-1}{\tau_1-1} \tag{4.11}$$

Note that, for Inequalities (4.9), (4.10) and (4.11), there are certain values of τ_1 for which the term on the left is less than the corresponding term on the right.

Now, consider (4.9):

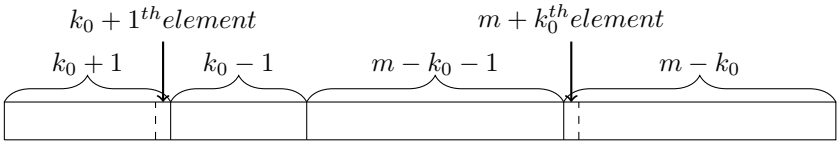
$$\sum_{\tau_1=1}^{\lambda} \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1}\binom{m-k_0}{\tau_1-1}\binom{m-k_0-1}{\lambda-\tau_1} \geq \sum_{\tau_1=1}^{\lambda} \binom{k_0-1}{\tau-\tau_1}\binom{k_0+1}{\tau_1}\binom{m-k_0-1}{\tau_1-1}\binom{m-k_0}{\lambda-\tau_1}$$

Using the same approach, we can see that the left side of the inequality consists of the number of combinations of $2m-1$ elements, such that, if we consider the elements divided up into sections of size $k_0, k_0, m-k_0$ and $m-k_0-1$:



- 3.1: in the first section of k_0 elements, at least 1 and at most λ elements are chosen
- 3.2: between the first and second sections of k_0 elements, τ are chosen
- 3.3: the section of $m - k_0$ elements has one less selected than the first section of k_0 elements
- 3.4: between the sections of $m - k_0$ and $m - k_0 - 1$ elements, $\lambda - 1$ are chosen

Similarly, the right side of the inequality consists of the number of combinations of $2m - 1$ elements, such that, if we consider the elements divided up into sections of size $k_0 + 1, k_0 - 1, m - k_0 - 1$ and $m - k_0$:



- 4.1: in the section of $k_0 + 1$ elements, at least 1 and at most λ elements are chosen
- 4.2: between the sections of $k_0 + 1$ and $k_0 - 1$ elements, τ are chosen
- 4.3: the section of $m - k_0 - 1$ elements has one less selected than the section of $k_0 + 1$ elements
- 4.4: between the sections of $m - k_0 - 1$ and $m - k_0$ elements, $\lambda - 1$ are chosen

Now, consider combinations that satisfy the third set of conditions but not the fourth. By satisfying the third set of conditions, then conditions 4.2 and 4.4 must also be satisfied. Also, note that if condition 4.1 is not satisfied, then neither is condition 4.3. So, we need only consider combinations that satisfy the third set of conditions but fail condition 4.3.

This can occur if either the $k_0 + 1^{th}$ element is selected, or if the $k_0 + 1^{th}$ element is not selected but the $m + k_0^{th}$ element is.

$$\sum_{\tau_1=1}^{\lambda} \binom{k_0}{\tau_1} \cdot \binom{k_0-1}{k-\tau_1-1} \cdot \binom{m-k_0}{\tau_1-1} \cdot \binom{m-k_0-1}{\lambda-\tau_1}$$
$$+ \sum_{\tau_1=1}^{\lambda-1} \binom{k_0}{\tau_1+1} \cdot \binom{k_0-1}{k-\tau_1-1} \cdot \binom{m-k_0-1}{\tau_1-1} \cdot \binom{m-k_0-1}{\lambda-\tau_1-1}$$

Now, consider combinations that satisfy the fourth set of conditions but not the third. By satisfying the fourth set of conditions, then conditions 3.2 and 3.4 must also be satisfied. Also, note that if condition 3.1 is not satisfied, then neither is condition 3.3. So, we need only consider combinations that satisfy the fourth set of conditions but fail condition 3.3.

This can occur if either the $k_0 + 1^{th}$ element is selected, or if the $k_0 + 1^{th}$ element is not selected but the $m + k_0^{th}$ element is.

$$\sum_{\tau_1=1}^{\lambda} \binom{k_0}{\tau_1-1} \cdot \binom{k_0-1}{\tau-\tau_1} \cdot \binom{m-k_0-1}{\tau_1-1} \cdot \binom{m-k_0}{\lambda-\tau_1}$$
$$+ \sum_{\tau_1=1}^{\lambda-1} \binom{k_0}{\tau_1} \cdot \binom{k_0-1}{\tau-\tau_1} \cdot \binom{m-k_0-1}{\tau_1-1} \cdot \binom{m-k_0-1}{\lambda-\tau_1-1}$$

So, showing that Inequality (4.9) holds is equivalent to showing that the following inequality holds:

$$\sum_{\tau_1=1}^{\lambda} \binom{k_0}{\tau_1} \cdot \binom{k_0-1}{\tau-\tau_1-1} \cdot \binom{m-k_0}{\tau_1-1} \cdot \binom{m-k_0-1}{\lambda-\tau_1}$$
$$+ \sum_{\tau_1=1}^{\lambda-1} \binom{k_0}{\tau_1+1} \cdot \binom{k_0-1}{\tau-\tau_1-1} \cdot \binom{m-k_0-1}{\tau_1-1} \cdot \binom{m-k_0-1}{\lambda-\tau_1-1}$$
$$\geq$$
$$\sum_{\tau_1=1}^{\lambda} \binom{k_0}{\tau_1-1} \cdot \binom{k_0-1}{\tau-\tau_1} \cdot \binom{m-k_0-1}{\tau_1-1} \cdot \binom{m-k_0}{\lambda-\tau_1}$$
$$+ \sum_{\tau_1=1}^{\lambda-1} \binom{k_0}{\tau_1} \cdot \binom{k_0-1}{\tau-\tau_1} \cdot \binom{m-k_0-1}{\tau_1-1} \cdot \binom{m-k_0-1}{\lambda-\tau_1-1}$$

Hence, to show that Inequality (4.9) holds, it is sufficient to show that the following two results hold:

$$\begin{aligned}&\sum_{\tau_1=1}^{\lambda}\binom{k_0}{\tau_1}\cdot\binom{k_0-1}{\tau-\tau_1-1}\cdot\binom{m-k_0}{\tau_1-1}\cdot\binom{m-k_0-1}{\lambda-\tau_1}\\&\quad\geq\sum_{\tau_1=1}^{\lambda}\binom{k_0}{\tau_1-1}\cdot\binom{k_0-1}{\tau-\tau_1}\cdot\binom{m-k_0-1}{\tau_1-1}\cdot\binom{m-k_0}{\lambda-\tau_1}\end{aligned}\tag{4.12}$$

$$\begin{aligned}&\sum_{\tau_1=1}^{\lambda-1}\binom{k_0}{\tau_1+1}\cdot\binom{k_0-1}{\tau-\tau_1-1}\cdot\binom{m-k_0-1}{\tau_1-1}\cdot\binom{m-k_0-1}{\lambda-\tau_1-1}\\&\quad\geq\sum_{\tau_1=1}^{\lambda-1}\binom{k_0}{\tau_1}\cdot\binom{k_0-1}{\tau-\tau_1}\cdot\binom{m-k_0-1}{\tau_1-1}\cdot\binom{m-k_0-1}{\lambda-\tau_1-1}\end{aligned}\tag{4.13}$$

Overall to show that Inequality (4.7) holds, it would be sufficient to show that the following set of inequalities hold:

$$\begin{aligned}&\sum_{\tau_1=1}^{\lambda}\binom{k_0}{\tau_1}\cdot\binom{k_0-1}{\tau-\tau_1-1}\cdot\binom{m-k_0}{\tau_1-1}\cdot\binom{m-k_0-1}{\lambda-\tau_1}\\&\quad\geq\sum_{\tau_1=1}^{\lambda}\binom{k_0}{\tau_1-1}\cdot\binom{k_0-1}{\tau-\tau_1}\cdot\binom{m-k_0-1}{\tau_1-1}\cdot\binom{m-k_0}{\lambda-\tau_1}\end{aligned}\tag{4.12 revisited}$$

$$\begin{aligned}&\sum_{\tau_1=1}^{\lambda-1}\binom{k_0}{\tau_1+1}\cdot\binom{k_0-1}{\tau-\tau_1-1}\cdot\binom{m-k_0-1}{\tau_1-1}\cdot\binom{m-k_0-1}{\lambda-\tau_1-1}\\&\quad\geq\sum_{\tau_1=1}^{\lambda-1}\binom{k_0}{\tau_1}\cdot\binom{k_0-1}{\tau-\tau_1}\cdot\binom{m-k_0-1}{\tau_1-1}\cdot\binom{m-k_0-1}{\lambda-\tau_1-1}\end{aligned}\tag{4.13 revisited}$$

$$\sum_{\tau_1=1}^{\lambda} \binom{k_0-1}{\tau_1} \binom{k_0}{\tau-\tau_1-1} \binom{m-k_0}{\tau_1-1} \binom{m-k_0-1}{\lambda-\tau_1} \geq \sum_{\tau_1=1}^{\lambda} \binom{k_0-1}{\tau-\tau_1-1} \binom{k_0}{\tau_1} \binom{m-k_0}{\lambda-\tau_1} \binom{m-k_0-1}{\tau_1-1} \tag{4.10 revisited}$$

$$\sum_{\tau_1=1}^{\lambda+1} \binom{k_0-1}{\tau_1-1} \binom{k_0}{\tau-\tau_1} \binom{m-k_0}{\tau_1-1} \binom{m-k_0-1}{\lambda-\tau_1+1} \geq \sum_{\tau_1=1}^{\lambda+1} \binom{k_0-1}{\tau-\tau_1} \binom{k_0}{\tau_1-1} \binom{m-k_0}{\lambda-\tau_1+1} \binom{m-k_0-1}{\tau_1-1} \tag{4.11 revisited}$$

Next, consider Inequality (4.12). Using binomial decomposition on the $\binom{m-k_0}{x}$ terms, we can say that to show that Inequality (4.12) holds, it is sufficient to show that the following inequalities hold:

$$\sum_{\tau_1=1}^{\lambda} \binom{k_0}{\tau_1} \binom{k_0-1}{\tau-\tau_1-1} \binom{m-k_0-1}{\tau_1-1} \binom{m-k_0-1}{\lambda-\tau_1} \geq \sum_{\tau_1=1}^{\lambda} \binom{k_0}{\tau_1-1} \binom{k_0-1}{\tau-\tau_1} \binom{m-k_0-1}{\tau_1-1} \binom{m-k_0-1}{\lambda-\tau_1} \tag{4.14}$$

$$\sum_{\tau_1=2}^{\lambda} \binom{k_0}{\tau_1} \binom{k_0-1}{\tau-\tau_1-1} \binom{m-k_0-1}{\tau_1-2} \binom{m-k_0-1}{\lambda-\tau_1} \geq \sum_{\tau_1=1}^{\lambda-1} \binom{k_0}{\tau_1-1} \binom{k_0-1}{\tau-\tau_1} \binom{m-k_0-1}{\tau_1-1} \binom{m-k_0-1}{\lambda-\tau_1-1} \tag{4.15}$$

Similarly for Inequality (4.10), we can do the variable substitution $\tau_1' = \lambda - \tau_1 + 1$ on the left side and then do binomial decomposition on the $\binom{k_0}{x}$ terms. Then, showing that the following two inequalities hold is sufficient to show that Inequality (4.10) holds.

$$\sum_{\tau_1=1}^{\lambda}\binom{k_0-1}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-3}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1} \geq \sum_{\tau_1=1}^{\lambda}\binom{k_0-1}{\tau-\tau_1-1}\binom{k_0-1}{\tau_1-1}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1} \tag{4.16}$$

$$\sum_{\tau_1=1}^{\lambda}\binom{k_0-1}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-2}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1} \geq \sum_{\tau_1=1}^{\lambda}\binom{k_0-1}{\tau-\tau_1-1}\binom{k_0-1}{\tau_1}\binom{m-k_0}{\lambda-\tau_1}\binom{m-k_0-1}{\tau_1-1} \tag{4.17}$$

For Inequality (4.11), do the variable substitution $\tau_1' = \lambda - \tau_1 + 2$ on the left side and then do binomial decomposition on the $\binom{k_0}{x}$ terms. Then, showing that the following two inequalities hold is sufficient to show that Inequality (4.11) holds.

$$\sum_{\tau_1=1}^{\lambda+1}\binom{k_0-1}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-3}\binom{m-k_0}{\lambda-\tau_1+1}\binom{m-k_0-1}{\tau_1-1} \geq \sum_{\tau_1=2}^{\lambda+1}\binom{k_0-1}{\tau-\tau_1}\binom{k_0-1}{\tau_1-2}\binom{m-k_0}{\lambda-\tau_1+1}\binom{m-k_0-1}{\tau_1-1} \tag{4.18}$$

$$\sum_{\tau_1=1}^{\lambda+1}\binom{k_0-1}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-2}\binom{m-k_0}{\lambda-\tau_1+1}\binom{m-k_0-1}{\tau_1-1} \geq \sum_{\tau_1=1}^{\lambda+1}\binom{k_0-1}{\tau-\tau_1}\binom{k_0-1}{\tau_1-1}\binom{m-k_0}{\lambda-\tau_1+1}\binom{m-k_0-1}{\tau_1-1} \tag{4.19}$$

So in total, I have shown that in order to prove that Inequality (4.7) holds, it is sufficient to show that Inequalities (4.13), (4.14), (4.15), (4.16), (4.17), (4.18) and (4.19) hold.

Each of these inequalities has the nice property that the sum of the first and last terms on the left is greater than the sum of the first and last terms on the right. The same can be said of the sums of the second and second from last terms, and so on until we reach the center. I will prove this next.

Start by considering Inequality (4.13). If we use a change of variable $\tau_1' = \lambda - \tau_1$ on the left side of the inequality, then this yields the following inequality:

$$\sum_{\tau_1=1}^{\lambda-1} \binom{k_0}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-1}\binom{m-k_0-1}{\lambda-\tau_1-1}\binom{m-k_0-1}{\tau_1-1}$$
$$\geq \sum_{\tau_1=1}^{\lambda-1} \binom{k_0}{\tau_1}\binom{k_0-1}{\tau-\tau_1}\binom{m-k_0-1}{\lambda-\tau_1-1}\binom{m-k_0-1}{\tau_1-1}$$
$$\iff$$
$$\sum_{\tau_1=1}^{\lambda}\left[\binom{k_0}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-1} - \binom{k_0}{\tau_1}\binom{k_0-1}{\tau-\tau_1}\right]\binom{m-k_0-1}{\lambda-\tau_1-1}\binom{m-k_0-1}{\tau_1-1} \geq 0$$

So, this is a weighted sum of terms of the form $\binom{k_0}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-1} - \binom{k_0}{\tau_1}\binom{k_0-1}{\tau-\tau_1}$, with coefficients of the form $\binom{m-k_0-1}{\lambda-\tau_1-1}\binom{m-k_0-1}{\tau_1-1}$.

Start by noting the well known property that consecutive binomial coefficient (ex. $(\binom{a}{0}, \binom{a}{2}, ..., \binom{a}{a})$) form a log concave sequence, and products of log concave sequences are also log concave. Hence, the summation coefficients $\binom{m-k_0-1}{\lambda-\tau_1-1}\binom{m-k_0-1}{\tau_1-1}$ form a log concave sequence in τ_1.

Furthermore, note that the coefficients are symmetric around $\tau_1 = \frac{\lambda}{2}$:

$$\binom{m-k_0-1}{\lambda-(\frac{\lambda}{2}+n)-1}\binom{m-k_0-1}{(\frac{\lambda}{2}+n)-1} = \binom{m-k_0-1}{(\frac{\lambda}{2}-n)-1}\binom{m-k_0-1}{\lambda-(\frac{\lambda}{2}-n)-1}$$

Where $\frac{\lambda}{2} \pm n$ are natural numbers.

Hence, the coefficients $\binom{m-k_0-1}{\lambda-\tau_1-1}\binom{m-k_0-1}{\tau_1-1}$ form a log concave sequence that is symmetric around $\frac{\lambda}{2}$, which must therefore be its maxima. Hence, the coefficients form a sequence that is non-decreasing until $\frac{\lambda}{2}$, and is non-increasing in a symmetric manner thereafter.

Next, note that the summation term $\binom{k_0}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-1} - \binom{k_0}{\tau_1}\binom{k_0-1}{\tau-\tau_1}$ has a symmetry about $\tau_1 = \frac{\lambda+1}{2}$ where the expressions on one side of $\frac{\lambda+1}{2}$ are the negative of those on the other. That is to say that the summation term corresponding to $\frac{\lambda+1}{2} + n$ is the negative of the one corresponding to $\frac{\lambda+1}{2} - n$:

$$\begin{aligned}
&\binom{k_0}{\lambda - (\frac{\lambda+1}{2}+n)+1}\binom{k_0-1}{\tau-\lambda+(\frac{\lambda+1}{2}+n)-1} - \binom{k_0}{(\frac{\lambda+1}{2}+n)}\binom{k_0-1}{\tau-(\frac{\lambda+1}{2}+n)} \\
&= \binom{k_0}{(\frac{\lambda+1}{2}-n)}\binom{k_0-1}{\tau-(\frac{\lambda+1}{2}-n)} - \binom{k_0}{\lambda-(\frac{\lambda+1}{2}-n)+1}\binom{k_0-1}{\tau-\lambda+(\frac{\lambda+1}{2}-n)-1} \\
&= -\left[\binom{k_0}{\lambda - (\frac{\lambda+1}{2}-n)+1}\binom{k_0-1}{\tau-\lambda+(\frac{\lambda+1}{2}-n)-1} - \binom{k_0}{(\frac{\lambda+1}{2}-n)}\binom{k_0-1}{\tau-(\frac{\lambda+1}{2}-n)}\right]
\end{aligned}$$

Where $\frac{\lambda+1}{2} \pm n$ are natural numbers.

Furthermore we can show that the term is non-negative for $\tau_1 < \frac{\lambda+1}{2}$. We therefore have a sum where any negative term has a corresponding positive term of equal magnitude with a coefficient that is at least as great.

Start by using the following identity, where $b \leq a$:

$$\begin{aligned}
\binom{a-1}{b} &= \frac{(a-1)!}{b! \cdot (a-b-1)!} \\
&= \frac{a! \cdot (a-b)}{a \cdot b! \cdot (a-b)!} \\
&= \frac{a-b}{a} \cdot \binom{a}{b}
\end{aligned}$$

Then, we can rewrite the terms as:

$$\begin{aligned}
&\binom{k_0}{\lambda-\tau_1+1}\binom{k_0-1}{\tau-\lambda+\tau_1-1} - \binom{k_0}{\tau_1}\binom{k_0-1}{\tau-\tau_1} \\
&= \binom{k_0}{\lambda-\tau_1+1}\binom{k_0}{\tau-\lambda+\tau_1-1}\frac{k_0-\tau+\lambda-\tau_1+1}{k_0} - \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1}\frac{k_0-\tau+\tau_1}{k_0}
\end{aligned}$$

Then, note that whenever $\tau_1 \leq \frac{\lambda+1}{2}$, the fraction in the first term will be at least as great as that in the second:

$$\begin{gathered} k_0 - \tau + \lambda - \tau_1 + 1 \geq k_0 - \tau + \tau_1 \\ \iff \\ \frac{\lambda+1}{2} \geq \tau_1 \end{gathered}$$

Next, I show that if $\tau_1 \leq \frac{\lambda+1}{2}$, then $\binom{k_0}{\lambda-\tau_1+1}\binom{k_0}{\tau-\lambda+\tau_1-1} \geq \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1}$.

By the result from Sagan's paper [33], we have that $\binom{n}{j}\binom{n}{l} \leq \binom{n}{j+1}\binom{n}{l-1}$ for $0 \leq j < l \leq n$. By applying this recursively, we get that $\binom{n}{j}\binom{n}{l} \leq \binom{n}{j+x}\binom{n}{l-x}$ with $x \geq 0$ and $j + x \leq l$.

If it holds that $\tau_1 + (\tau - \lambda - 1) \leq \tau - \tau_1$, then we can use this result with $n = k_0, j = \tau_1, l = \tau - \tau_1$ and $x = \tau - \lambda - 1$. This yields:

$$\begin{gathered} \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1} \leq \binom{k_0}{\tau_1 + (\tau - \lambda - 1)}\binom{k_0}{\tau - \tau_1 - (\tau - \lambda - 1)} \\ \iff \\ \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1} \leq \binom{k_0}{\lambda - \tau_1 + 1}\binom{k_0}{\tau - \lambda + \tau_1 - 1} \end{gathered}$$

However, the condition $\tau_1 + (\tau - \lambda - 1) \leq \tau - \tau_1$ is equivalent to $\tau_1 \leq \frac{\lambda+1}{2}$.

Hence, whenever $\tau_1 \leq \frac{\lambda+1}{2}$, then the summation term must be non-negative.

Finally, note that since τ_1 is summed from 1 to λ, $\frac{\lambda+1}{2}$ is the midpoint of the summation. As such, in inequality (4.13), every negative term in the summation will have a corresponding positive term with weight greater than or equal to its weight.

Inequalities (4.14), (4.15), (4.16), (4.17), (4.18) and (4.19) are proven using the exact same technique.

Hence, I have shown that Inequality 4.3 holds whenever $1 > \tau - k_0$.

For the case that $1 < \tau - k_0$, the exact same approach and techniques are used. Again, we use the combinatorial meaning of the inequality to find terms that are on one side but not the other. We then use binomial decomposition, and end up with a set of sufficient inequalities that all hold over pairs of symmetric points.

The final case is where $1 = \tau - k_0$. In this case, Inequality (4.3) becomes:

$$\sum_{\tau_1=1}^{\tau-1} \binom{k_0}{\tau_1}\binom{k_0}{\tau-\tau_1} \sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)} \binom{m-k_0}{\lambda_1}\binom{m-k_0}{\lambda-\lambda_1} \tag{4.20}$$
$$\geq \sum_{\tau_1=1}^{\tau-2} \binom{k_0-1}{\tau_1}\binom{k_0+1}{\tau-\tau_1} \sum_{\lambda_1=max(0,\lambda-\tau+\tau_1+1)}^{min(\tau_1-1,\lambda)} \binom{m-k_0+1}{\lambda_1}\binom{m-k_0-1}{\lambda-\lambda_1}$$

This inequality is easy to verify, since the left side is equal to the left side in Inequality (4.7), and the right side is less than the right side in Inequality (4.7).

Hence, since we already know that Inequality (4.7) holds, Inequality (4.3) also holds.

□

4.6 Conjecture on Minimum and Results

Throughout the work detailed in this Chapter so far, we have been examining Equation (4.2), which is the probability of selecting a sample that yields a classifier within ϵ of the optimal classifier, given the optimal parameterization of the underlying distribution. I showed how this function is mostly continuous, with surfaces of discontinuity over the parameter space $[0, 1]^3$. Furthermore, I proved that the minimum of this function must be achieved on either one of those surfaces of discontinuity, or on the boundary of the cube.

Beyond this, I have done some work towards showing where, precisely, on the discontinuity surfaces or the boundary surfaces such minima could occur. Because the proofs are not yet complete I have not included them here. However, I have strong reasons to believe that the minimum must be achieved at one of a finite list of points. Hence, we know that the underlying distribution with the lowest probability of selecting a sample that yields an adequately accurate classifier must have one of several optimal model parameterizations.

Then, for a given m, by simply assessing the probability of choosing a good sample as defined in Equation (4.2) at each of these candidate minima, we can find the minimum probability of generating a sufficiently accurate classifier using a sample of size m. We can then be certain that, regardless of what the actual underlying distribution is, the probability of choosing a sample that yields a satisfactory classifier is at least as high as it is at the minimum. Finally, we can iteratively increase m until the probability of selecting

a sufficiently accurate classifier is at least $1 - \delta$ at each of the candidate minima. We will then have found the smallest possible integer m that provides confidence above our threshold $1 - \delta$ that we will meet our accuracy target.

Due to the symmetry of the objective function (4.2), as described earlier in this Chapter, we can restrict ourselves only to Regions A and B, since Regions C and D will achieve the same minimum. Then, I believe the minimum of Equation (4.2) will be achieved at one of the points identified in Conjecture 1.

Conjecture 1. *Suppose we create a naïve Bayes classifier using an i.i.d. sample of m observations from the distribution $\mathcal{D}$ and ERM with misclassification loss. Furthermore, suppose that we would like our classifier to have misclassification error over $\mathcal{D}$ that is within ϵ of that of the optimal naïve Bayes classifier for $\mathcal{D}$, for some small threshold $\epsilon > 0$. Then, the distribution $\mathcal{D}$ that is least likely to yield a sample that will generate a model that meets our accuracy target will have an optimal naïve Bayes parameterization $(\theta^*, \phi_0^*, \phi_1^*)$ from the following list:*

- $(\frac{1}{2} - 2\epsilon, \frac{1}{2}, \frac{1}{2})$
- $(\frac{1}{2} - \epsilon, \frac{1}{2}, \frac{1}{2})$
- $(\frac{1}{2} - \frac{2}{3}\epsilon, \frac{1}{2}, \frac{1}{2})$
- $(\frac{1}{2}, \frac{1}{2} + 2\epsilon, \frac{1}{2} - 2\epsilon)$
- $(\frac{1}{2}, \frac{1}{2} + \epsilon, \frac{1}{2} - \epsilon)$
- $(\frac{1}{2}, 4\epsilon, 0)$
- $(\frac{1}{2} - 2\epsilon, 0, \frac{4\epsilon}{1+4\epsilon})$
- $(2\epsilon, 1, 0)$
- $(1 - 2\epsilon, 1, 0)$
- $(0, 0, 2\epsilon)$
- $(0, 0, 1 - 2\epsilon)$

I've also done an investigation into what the implications of this result will be, once I can fully validate the conjecture. These results give us a concrete way to measure what

sample size is needed to reach a desired accuracy threshold with the desired confidence. Or, alternatively, it could tell us, if we know that we have a specific number of observations in our training set, what is the probability that we will reach our desired accuracy.

Start by assuming that we have a fixed sample size m and a parameter $\epsilon > 0$ that represents how much misclassification in addition to that of the optimal classifier that we are willing to tolerate. Then, determine the minimum probability that we will achieve this accuracy, over the space of distributions. This is done by evaluating Equation (4.2) over the list of points in Conjecture 1, and finding the minimum. Then, we will know that no matter what the underlying distribution actually is, the probability of selecting a sample that yields a sufficiently accurate classifier is at least as great as the minimum we found.

Table 4.1 uses $\epsilon = 0.05$, and shows, for different sample sizes m, the probability of selecting a model with misclassification error within ϵ of the optimal naïve Bayes classifier. It shows this probability at each of the candidate minima, and from those finds the minimum. This minimum is the overall guaranteed probability that we have of obtaining a sufficiently accurate classifier. It could also be thought of as the minimum value that δ can take before m is too small to guarantee that our learned classifier is within ϵ of optimal with probability $1 - \delta$.

Table 4.1: Probability of generating classifier within 0.05 of optimal by sample size for candidate minima parameterizations of the optimal distribution

Point: $(\theta^*, \phi_0^*, \phi_1^*)$	$m = 10$	$m = 50$	$m = 100$	$m = 200$
$(\frac{1}{2} - 2\epsilon, \frac{1}{2}, \frac{1}{2})$	0.52	0.75	0.87	0.96
$(\frac{1}{2} - \epsilon, \frac{1}{2}, \frac{1}{2})$	0.80	0.86	0.92	0.96
$(\frac{1}{2} - \frac{2}{3}\epsilon, \frac{1}{2}, \frac{1}{2})$	0.84	0.85	0.89	0.93
$(\frac{1}{2}, \frac{1}{2} + 2\epsilon, \frac{1}{2} - 2\epsilon)$	0.52	0.75	0.87	0.96
$(\frac{1}{2}, \frac{1}{2} + \epsilon, \frac{1}{2} - \epsilon)$	0.80	0.89	0.94	0.97
$(\frac{1}{2}, 4\epsilon, 0)$	0.62	0.79	0.87	0.94
$(\frac{1}{2} - 2\epsilon, 0, \frac{4\epsilon}{1+4\epsilon})$	0.62	0.79	0.87	0.94
$(2\epsilon, 1, 0)$	1.0	1.0	1.0	1.0
$(1 - 2\epsilon, 1, 0)$	1.0	1.0	1.0	1.0
$(0, 0, 2\epsilon)$	1.0	1.0	1.0	1.0
$(0, 0, 1 - 2\epsilon)$	1.0	1.0	1.0	1.0
MINIMUM	0.52	0.75	0.87	0.93

So, if we have only 10 observations, then regardless of the underlying distribution, we

know that the probability that the sample chosen will yield a classifier with misclassification error within 0.05 of the optimal classifier is at least 52%. If the sample size jumps to 200 observations, then we know that with probability at least 92%, we will reach our accuracy target.

Hence, as shown in Table 4.1, for a given sample size, we can find the probability of choosing a sample that yields our desired accuracy at each point. One interesting point to note is that based on my empirical results, it seems that there are some points that usually yield lower confidence of high accuracy than others. For example, the last four points all seem to generate sufficiently accurate classifiers with near certain probability for all of the sample sizes that I tested. This is because these are the near-deterministic cases, when each feature is perfectly correlated to the class. So, as long as our sample contains observations from each class, we are guaranteed to find the optimal classifier, and this will happen with near-certain probability unless the sample is very very small (<10 observations). So, though it is possible in certain cases that these points could be the true minimum, in practice they can safely be ignored unless the sample is tiny.

We can also use this result to determine what is the minimum sample size needed to guarantee that we meet our accuracy and confidence thresholds. Note that an upper bound on this sample complexity can also be calculated using the Theorem of PAC Learning of Finite Hypothesis Classes, which shows that $m \leq \lceil \frac{2log(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$. In Table 4.2, for selected values of δ and ϵ, I show the exact number of observations needed as calculated using the candidate minima and the upper bound as calculated using the Theorem from learning theory. This table also shows what percent my exact value is of the upper bound. It will be presented as: exact sample complexity/learning theory bound on sample complexity (exact as % of bound).

Table 4.2: Exact sample complexity/learning theory bound (exact as % of bound) for selected ϵ and δ values

	$\epsilon = \mathbf{0.10}$	$\epsilon = \mathbf{0.05}$	$\epsilon = \mathbf{0.02}$
$\delta = \mathbf{0.10}$	27/381 (7%)	137/1,523 (9%)	867/9,515 (9%)
$\delta = \mathbf{0.05}$	69/441 (16%)	303/1,763 (17%)	1,813/11,021 (16%)
$\delta = \mathbf{0.01}$	195/581 (34%)	809/2,323 (35%)	4,693/14,515 (32%)

The results in Table 4.2 show that indeed, the learning theory results, which are known to be loose upper bounds, can in fact be significantly higher than the precise amount of data needed based on Conjecture 1. It would be interesting to see, if I am able to generalize my

result to multiple features, whether the difference would be so significant once the model becomes somewhat more complex.

Also, based on empirical tests, the precise sample complexity becomes larger as a percent of the learning theory bound as the confidence parameter δ becomes more restrictive.

Though these results are based on a conjecture, they show that there is promise in pursuing exact sample complexity bounds for specific models, since they can yield great reductions from the known loose upper bounds. As this is a conjecture, my results are not complete, and it is possible that this list is non-exhaustive. Nonetheless, there is sound reasoning behind it, and I hope to prove its validity in upcoming work, as well as extend this result to include more features and other models, though this would involve reformulating the approach to the problem.

Chapter 5

Conclusion

This thesis provides insights into the naïve Bayes classifier. I proved some important results related to the optima of the likelihood in the case of unsupervised learning, and outlined a framework to determine exactly how many observations are required to learn a near-optimal classifier with high probability using maximum likelihood in the case of supervised learning.

Chapter 3 provides the first characterization of the stationary points of the likelihood function of the naïve Bayes model in unsupervised learning. I showed that global optimality is generally attained for problems with up to three features unless special conditions are met. These conditions can be used in practice to detect whether maximum likelihood techniques such as gradient ascent and expectation maximization could be stuck in a spurious local optimum. I also showed that for settings with any number of features, all the stationary points in the interior of the parameter space possess marginal distributions that match the empirical marginals of the training data. Hence, even if a stationary point is not globally optimal, it still has some nice structure.

This work can be extended in several directions. First of all would be to incorporate cases where the parameters may be 0 or 1, and thus local optima may not be stationary with respect to every parameter. It would also be interesting to apply this comprehensive characterization of stationary points to more than three features. The challenge is that first-order conditions lead to polynomial equations with an increasing degree. Nonetheless, I have performed over 200 experiments by randomly generating distributions with 4-10 features, and using gradient ascent to maximize the likelihood. All of this experimentation supports my conjecture that the characterization described in Theorem 3 should extend to any number of features.

Another important direction for investigation is the quantification of suboptimality of spurious local optima. While we have so far been concerned only with the naïve Bayes model, I hope to extend the characterization of stationary points to arbitrary Bayesian networks with latent variables.

Chapter 4 introduces an algorithm to determine exactly how many observations are needed to ensure with high probability that a classifier that is close to optimal is chosen. Though many previous results give loose bounds on sample complexity, or describe its order with respect to the VC dimension and accuracy and confidence thresholds, this is the first proposal of a method to determine the exact number of points needed using a specific model and algorithm. I also showed that the probability of selecting a sample that yields a sufficiently accurate classifier is monotonic or log concave with respect to one of the model parameters. Hence, I showed that the distributions that are most likely to yield samples that don't meet our accuracy threshold must be on the border of the parameter space or one of the surfaces of discontinuity.

There are several natural continuations for this work. First of all, I would like to fill in several gaps in the proof, so that I can validate the list of points in Conjecture 1. Once this is complete, I would like to do a thorough study of different sample complexity bounds, and compare the exact bound obtained using our method with the loose bounds obtained through learning theory.

Another significant line of study is to see how this result could be extended to include more features, and eventually to other Bayesian networks. However, due to the structure of the argument made, this would require a fundamental reformulation of the problem.

References

[1] Carlos Améndola, Mathias Drton, and Bernd Sturmfels. Maximum likelihood estimates for gaussian mixtures are transcendental. In *International Conference on Mathematical Aspects of Computer and Information Sciences*, pages 579–590. Springer, 2015.

[2] Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

[3] Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[4] András Antos and Gábor Lugosi. Strong minimax lower bounds for learning. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, COLT '96, pages 303–309, New York, NY, USA, 1996. ACM.

[5] P. L. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Trans. Inf. Theor.*, 44(2):525–536, September 2006.

[6] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

[7] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[9] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, October 1989.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[11] M. Augustine Cauchy. Methode generale pour la resolution des systemes dequations simultanees. *Comptes Rendus Hebd. Seances Acad. Sci*, 25, 1847.

[12] David Maxwell Chickering and David Heckerman. Fast learning from sparse data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 109–115. Morgan Kaufmann Publishers Inc., 1999.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[14] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[15] Andrzej Ehrenfeucht and David Haussler. A general lower bound on the number of examples needed for learning. *Inf. Comput.*, 82(3):247–261, September 1989.

[16] Susana Eyheramendy, David D. Lewis, and David Madigan. On the naive bayes model for text categorization, 2003.

[17] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[19] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*. AMS, 2003.

[20] David Haussler, Michael Kearns, and Robert E. Schapire. Bounds on the sample complexity of bayesian learning using information theory and the vc dimension. *Machine Learning*, 14, 1994.

[21] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

[22] Yifen Huang and Tom M Mitchell. Text clustering with extended user feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 413–420. ACM, 2006.

[23] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, pages 4116–4124, 2016.

[24] Sham Machandranath Kakade. On the sample complexity of reinforcement learning. 2003.

[25] John Langford. Quantitatively tight sample complexity bounds. 2002.

[26] David Mcallester. Simplified pac-bayesian margin bounds. In *COLT*, pages 203–215, 2003.

[27] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[28] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[29] X. Mu. Log-concavity of a Mixture of Beta Distributions. *ArXiv e-prints*, December 2013.

[30] Vivek Narayanan, Ishan Arora, and Arjun Bhatia. *Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model*, pages 194–201. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[31] Pascal Poupart. Cs485: Machine learning, assignment 4: Sample complexity, 2012.

[32] Sivan Sabato, Nathan Srebro, and Naftali Tishby. Tight sample complexity of large-margin learning. *CoRR*, abs/1011.5053, 2010.

[33] B. Sagan. Unimodality and the reflection principle. *ArXiv Mathematics e-prints*, December 1997.

[34] Karl-Michael Schneider. A comparison of event models for naive bayes anti-spam e-mail filtering. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 307–314, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[35] Eran Segal, Ben Taskar, Audrey Gasch, Nir Friedman, and Daphne Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17(suppl 1):S243–S252, 2001.

[36] CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[37] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

APPENDICES

Appendix A

Additional Proofs by Collaborators

These two proofs were completed by my collaborator George Trimponias at Huawei.

Theorem 5. *In the consolidated continuous area of Region A, the objective function* (4.2) *is monotonic decreasing w.r.t.* θ^**. That is to say, the probability of choosing a sample that yields a classifier that is accurate within* ϵ *of optimal is monotonic decreasing w.r.t.* θ^**. Under the same conditions in Region D,* (4.2) *is monotonic increasing w.r.t.* θ^**.*

Proof. Assume we are in the consolidated continuous area of Region A. By symmetry, if the result holds in Region A, it will hold in Region D.

Assume first that we are in a continuous area where θ^* is such that the only subcase satisfying $err < \epsilon$ is subcase 1, where $k_{00} < k_{10}, k_{01} < k_{11}$ and $err = 0$. The proof follows the exact same logic for any of the consolidated continuous areas.

Then we have:

$$P(err < \epsilon) = \sum_{k_0=0}^{m} \sum_{k_{00}=0}^{k_0} \sum_{k_{10}=0}^{m-k_0} P(k_0|m, \theta^*) P(k_{00}|k_0, \phi_0^*) P(k_{10}|m - k_0, \phi_1^*) I(k_{00} < k_{10} \wedge k_{01} < k_{11})$$

Hence, we are finding the probability of choosing a sample with $k_{00} < k_{10} \wedge k_{01} < k_{11}$. Note that θ^* denotes the probability of $C = 0$ in the underlying distribution. As θ^* gets smaller, then the probability of having $C = 0$ in our sample will become smaller. So, k_{00}

and k_{01} will tend to become smaller in most samples. Since we know that there are m total observations, then by elimination, k_{10} and k_{11} will tend to become larger in most samples. Hence, as θ^* gets smaller, $P(k_{00} < k_{10} \wedge k_{01} < k_{11})$ will become larger.

Therefore, in the consolidated continuous area of Region A, the objective function (4.2) is monotonic decreasing w.r.t. θ^*.

□

Theorem 6. *In the consolidated continuous area of Region B, the objective function* (4.2) *is monotonic increasing w.r.t.* ϕ_0^* *and monotonic decreasing w.r.t.* ϕ_1^**. That is to say, the probability of choosing a sample that yields a classifier that is accurate within* ϵ *of optimal is monotonic increasing w.r.t.* ϕ_0^* *and monotonic decreasing w.r.t.* ϕ_1^**. Under the same conditions in Region C,* (4.2) *is monotonic decreasing w.r.t.* ϕ_0^* *and monotonic increasing w.r.t.* ϕ_1^**.*

Proof. Assume we are in the consolidated continuous area of Region B. By symmetry, if the result holds in Region B, it will hold in Region C.

Assume first that we are in a continuous area where θ^* is such that the only subcase satisfying $err < \epsilon$ is subcase 3, where $k_{00} > k_{10}, k_{01} < k_{11}$ and $err = 0$. The proof follows the exact same logic for any of the consolidated continuous areas.

Then we have:

$$P(err < \epsilon) =$$
$$\sum_{k_0=0}^{m} \sum_{k_{00}=0}^{k_0} \sum_{k_{10}=0}^{m-k_0} P(k_0|m, \theta^*) P(k_{00}|k_0, \phi_0^*) P(k_{10}|m - k_0, \phi_1^*) I(k_{00} > k_{10} \wedge k_{01} < k_{11})$$

Hence, we are finding the probability of choosing a sample with $k_{00} > k_{10} \wedge k_{01} < k_{11}$. Note that ϕ_0^* denotes $P(F = 0|C = 0)$ in the underlying distribution. As ϕ_0^* gets larger, then the probability of having $F = 0, C = 0$ in our sample will become larger. So, k_{00} will tend to become larger in most samples. Since this means that $1 - \phi_0^*$ is getting smaller, we will similarly have that $P(F = 1|C = 0)$ will become smaller. Hence, k_{01} will tend to become smaller in most samples. Therefore, as ϕ_0^* gets larger, $P(k_{00} > k_{10} \wedge k_{01} < k_{11})$ will become larger.

Similarly, ϕ_1^* denotes $P(F = 0|C = 1)$ in the underlying distribution. As ϕ_1^* gets smaller, then the probability of having $F = 0, C = 1$ in our sample will become smaller.

So, k_{10} will tend to become smaller in most samples. Since this means that $1-\phi_1^*$ is getting larger, we will similarly have that $P(F=1|C=1)$ will become larger. Hence, k_{11} will tend to become larger in most samples. Therefore, as ϕ_1^* gets smaller, $P(k_{00} > k_{10} \wedge k_{01} < k_{11})$ will become larger.

Therefore, in the consolidated continuous area of Region B, the objective function (4.2) is monotonic increasing w.r.t. ϕ_0^* and monotonic decreasing w.r.t. ϕ_1^*.

□