Bayesian Supervised Learning using Full and Limited Support Priors

by Paul Rademacher

B.S. in Electrical and Computer Engineering, May 2010, Cornell University
M.Eng. in Electrical and Computer Engineering, May 2011, Cornell University

A Dissertation Proposal submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial satisfaction of the requirements
for the degree of Doctor of Philosophy

Fall 2019

Dissertation Qualifying Examination directed by

Miloš Doroslovački
Associate Professor of Electrical and Computer Engineering

Dissertation Qualifying Exam Committee:

Miloš Doroslovački
Associate Professor of Engineering and Applied Science
Dissertation Director

Murray Loew
Professor of Engineering and Applied Science
Committee Member

Miguel Lejeune
Professor of Engineering and Applied Science
Committee Member

Mahdi Imani
Assistant Professor of Engineering and Applied Science
Committee Member

Omur Ozel
Assistant Professor of Engineering and Applied Science
Committee Member

Abstract

Bayesian Supervised Learning using Full and Limited Support Priors

Since Bayes put forth his famous theorem in 1763, there has been contention between the practitioners of classic statistical inference and those who favor the Bayesian approach. While some postulate that the imposition of prior knowledge hinders the ability of the data to "speak for itself" and identify the unknown data-generating distribution, it can be argued that Bayesian decision methods are necessary for applications in which the volume of data is limited. With Bayesian learning, a highly informative prior which strongly emphasizes the true data model will lead to low risk, even if the data space is undersampled; however, if the prior is ill-conceived, the design will be poor.

Today, research interest in statistical inference has grown dramatically because of its role in supervised machine learning, which has seen a surge in popularity due to notable results on various classification benchmarks. Many of these advances have been attributed to the increased availability of high-performance computing resources and our resultant ability to train parametric learners with many degrees of freedom. However, the value of this flexibility is unavoidably dependent on the learning operation. Modern machine learning practices are dominated by non-Bayesian methods, which are commonly driven by the empirical risk minimization objective. Training a decision function with this objective often leads to designs that do not generalize well to novel data. This phenomenon is actually *more* pronounced when the set of achievable decision functions is unrestricted. This trend underscores a critical point: *the empirical risk is not actually the objective function that the designer wants to minimize.*

Bayesian methods were historically unpopular due to the computational complexity inherent in integral approximation. However, this complexity is also reflective of the

wide variety of learning approaches afforded under the Bayesian perspective. Different learning objectives with different emphases on the data and on prior knowledge are derived via the user's specification of a prior distribution, which weights the candidate data-generating models. A prior distribution with *minimal* localization will dictate the use of empirical risk minimization. As such, informative prior knowledge should be considered *mandatory* for most practical data-limited applications.

This thesis will develop detailed theory for Bayesian supervised learning, including analyses of the consequences of informative prior selection. First, extensive theory for learning with *full support* priors will be set down; specific attention will be given to the class of Dirichlet distributions, which lead to tractable predictions and enable both informative and non-informative priors. The full support is of specific importance, as it is required to guarantee asymptotically consistent estimation of the data-generating model and thus minimal risk in the limit of training data volume. It is contested that full support is required for truly non-informative priors. It is common in the literature to misrepresent priors as non-informative as long as they are approximately uniform on their support, even if the support is limited - depending on the restriction of the support, such priors can actually be viewed as informative.

Subsequently, the theory will be extended for decision functions based on priors of *limited support*, specifically with support on sets of low *intrinsic dimensionality*. Such degenerate distributions are highly informative and often valuable when learning from high-dimensional data; also, they are necessary in practice for data drawn from continuous sets. The excess risk imposed in the limit of training data volume will be evaluated for different support dimensionalities, and improvement over full-support Bayesian learners will be assessed. The relationship between the dimensionality of a learner's prior and the computational complexity of real-world parametric implementation will be of interest as well.

Practical application of the learning techniques developed in this thesis will focus on

the automation of human-level speech/image recognition capabilities - the data space for such problems is typically large and thus undersampled. It is widely understood that humans extract audiovisual features that significantly compress their sensory input; as such, Bayesian learners with low-dimensionality support priors will be applied to enforce a similar dimensionality reduction and enable optimal parametric learners. Furthermore, the priors will be localized around models with concentrated predictive distributions, reflective of low-risk human performance on such tasks.

Table of Contents

List of Figures

Chapter 1: Background

1.1 From Classical Inference to Machine Learning

The debate as to whether or not Bayesian approaches are suitable for statistical applications such as detection and estimation has a long history [**?**]. For these well-established fields, many believe that the unknown element that statistically characterizes the observed data should not be treated as random, but rather as deterministic. This viewpoint is the foundation for a variety of classical inference methods that have been in use for decades. In estimation theory, the Cramér-Rao lower bound (CRLB) and Rao-Blackwell-Lehmann-Scheffe (RBLS) theorem provide solutions that are optimal (under certain conditions) [**?**]. In detection theory, the Neyman-Pearson theorem and generalized likelihood ratio test (GLRT) provide hypotheses without assigning probabilities to the underlying events [**?**].

This distinction between deterministic and Bayesian methods in these traditional fields has been inherited by one of the fastest-growing scientific disciplines today: machine learning (ML). For *supervised learning*, a set of input/output training observations is generated by an unknown probability function, or "model", and is then used to design a decision function which operates on novel inputs [**?**]. The goal is to make decisions which tend to result in minimal "loss", which is measured by a user-defined function that compares each decision to the corresponding unobserved random element. Two of the most popular sub-topics in supervised learning are regression and classification, which naturally relate to estimation and detection.

Most current supervised learning research is focused on the design of *parametric* learning functions; that is, algorithms whose resultant decision functions can be characterized by a finite quantity of (typically real number) parameters. Of course, the attention on parametric learning is a consequence of the practicalities of real-

world implementation - in our digital world, virtually all machine learning solutions are deployed on computers, and as such, they are bounded by the number of finite representations the computer memory can provide. However, as our computing capabilities continue to expand, so does our capacity to design higher-dimensional parametric learning functions.

Much of the popularity of and focus on machine learning today can be attributed to the resurgence of the multilayer perceptron and the success of deep neural networks (DNN) on classification challenges for speech and image recognition. Notable results include the successful application of deep belief networks to phone recognition on the TIMIT database [**?**] and deep convolutional neural networks (CNN) at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2010 [**?**]. By significantly improving upon the existing performance results at those times, widespread interest in new learning techniques was generated. Although the concept of the neural network has its roots in research as early as the 1950's, it was not until recently that we have had widespread access to computational power sufficient for training such large numbers of parameters on such voluminous collections of high-dimensional data. Many researchers credit the advances in computing and availability of large data sets as critical factors for these machine learning achievements.

Like other historically popular supervised learning algorithms, such as support vector machines (SVM) and decision trees, these deep learning algorithms do not derive from a Bayesian viewpoint. In fact, almost all of the highest profile machine learning approaches today are non-Bayesian.

1.2 The Bayesian Perspective

A common argument for why the unknown data model should not be treated as random is that there are often no environmental factors that suggest it is randomly generated. Although this justification is usually sound, the general application of

Bayesian techniques is motivated by a different perspective; specifically, when the unknown element is treated as random and assigned a probability function, said function can simply be interpreted as a measure of the user's confidence in different data-generating models prior to any data being observed [**?**]. The selection of this prior distribution is fundamental for Bayesian inference.

The success or failure of Bayesian learning methods hinges on how well the prior knowledge imparted by the designer matches reality. If a highly "informative" prior [**?**] is chosen that is localized around the actual data probability distribution, low risk learning functions are possible even with limited training data; however, if the informative prior is poorly designed, a good solution may not be achieved. Conversely, a "non-informative" prior that weights the different models without preference provides a more robust solution for all models, but may underperform relative to learners based on well-selected informative priors.

With large data spaces, designing a sensible prior distribution for Bayesian learning is challenging and often prohibitive. However, the wide range of learning functions afforded under the Bayesian perspective is valuable. In fact, many learning methods based on a deterministic treatment of the data-generating model have equivalents in Bayesian learning. This is prevalent among learning techniques that form point estimates of the data model via maximization, avoiding the potentially intractable integrations that are characteristic of full Bayesian learning methods.

Arguably the most widely used non-Bayesian technique for point estimation is the maximum likelihood (ML) method, with which the conditional distribution of the data given the model, or the "likelihood function", is maximized for the observed training set. The maximizing argument is treated as the true model and then used as a predictive distribution for novel observations [**?**]. The related Bayesian approach for model point estimation is the maximum *a posteriori* (MAP) method, which utilizes a user-designed prior distribution to formulate the posterior distribution of the model

given the training data; the maximizing model is then used as a point estimate. If the prior is uniform over its support (non-informative), then the MAP estimate is equivalent to the ML estimate.

Another example of equivalence between Bayesian and non-Bayesian learning methods is found in generalized squared-error regression applications. Commonly, the regressor's parameters are designed using *empirical risk minimization*, which is arguably the most common optimization metric for the design of classical and modern machine learning algorithms. It can be shown that the solution to this optimization is equivalent to ML estimation when the likelihood function is Gaussian with a mean value that is linearly dependent on the parameters. More interestingly, if the empirical risk objective is augmented with certain types of regularizing terms, the resultant estimate is equivalent to a MAP estimate produced using a Gaussian prior over the regression parameters [**?**].

These examples illustrate that many classical estimation/learning methods, even without specifying a prior distribution, implicitly express a lack of *a priori* model preference. Additionally, they highlight the flexibility of Bayesian learning methods, which can use both informative and non-informative prior distributions to control how decision functions trend with training data and how well they can perform for data-limited applications.

Chapter 2: Motivations

2.1 Learning via Empirical Risk Minimization

The design of a parametric learning algorithm, either Bayesian or non-Bayesian, can be decomposed into two tasks: the specification of how the training data maps to the finite-dimensional parameter space and the specification of how the parameters map to the higher-dimensional space of decision functions. The former is the "learning" procedure; the latter defines the mechanism for prediction. These two design tasks should be performed jointly, as the performance of a learning algorithm depends on both. However, it can be argued that the majority of current ML research directions are focused on the latter.

The notable modern successes of parametric machine learning algorithms, specifically deep learning approaches, and the widespread proliferation of open source code [**?**] have made them an easy choice for many practitioners who are faced with a complex regression or classification application. Additionally, it can be argued that the lack of explainability of many of these methods and their "black-box" treatment has actually contributed to their dominance, rather than hindered it.

Yet like all supervised learning approaches, they do not perform well for all problems. A basis for this principle was put forth by David Wolpert, who coined the "No Free Lunch" theorem (NFLT) [**?**], which demonstrates that for any two learning approaches, there are as many problems for which the first outperforms the second as there are for which the second is superior. This theorem bears a critical query: what exactly about the design of these widely-used non-Bayesian algorithms makes them so effective for certain learning applications? As mentioned, the increased power of computing resources has enabled the training of higher-dimensional parameter spaces than ever before; however, the parameter adaptation operations most commonly

selected, specifically empirical risk minimization, are not notably different from those used historically. Thus, to answer this question, the operations typically used by non-Bayesian ML functions for learning their parameters must be considered.

The failures of non-Bayesian methods are frequently attributed to a phenomenon termed "overfitting"; that is, the failure of the learner to perform well on novel data not used during training. Overfitting is a direct result of empirical risk optimization, which attempts to minimize the aggregate loss assessed on the training samples. Interestingly, creating a design that performs well on the training data is straightforward - some of the simplest learning algorithms can achieve minimal empirical risk. For example, the nearest-neighbor algorithm simply classifies novel data by applying a distance measure, determining the "nearest" training sample, and copying the label; if the training samples are distinct, zero risk will be achieved [**?**].

The overfitting issue underscores an important point: *the empirical risk is not actually the objective function that the designer wants to minimize.* This is illustrated by the performance of nearest-neighbor on novel data, which suffers if the training data volume is insufficient. The desirable learner property is termed "good generalization" by non-Bayesians, but in truth this simply means that it achieves low *expected risk* with respect to the unknown data-generating probability model.

Thus, if parameter training via empirical risk minimization does not guarantee generalization, why do modern non-Bayesian learning functions perform well? The answer must be strictly dependent on their predefined set of achievable functions. Parametric learning approaches, which account for nearly all popular algorithms, implicitly disallow the use of most of the complete function space and hopefully, any functions that lead to overfitting. Although no specific preference is given to any of the achievable functions in this subspace (assuming the empirical risk objective is not augmented with regularizing terms), the selection of this function subspace should be thought of as the imposition of *prior knowledge*. From this perspective,

even learning approaches that are widely deemed non-Bayesian can be seen from a Bayesian viewpoint.

2.2 The Necessity of Prior Knowledge

This new perspective enables a Bayesian interpretation of all parametric learning methods. Again consider maximum likelihood, one of the most classic and well-established "non-Bayesian" learning approaches. Essentially all textbook examples of maximum likelihood establish a finite-dimensional subspace of the set of probability functions and then determine which model maximizes the likelihood function [**?**] - the selection of this subspace is critical to the efficacy of ML and should be viewed as user-imposed prior knowledge. Specifically, the ML estimate is equivalent to a MAP estimate based on a prior distribution that is uniform on a probability function subspace; this probability distribution is *degenerate* and thus, highly informative.

If the user wanted a non-Bayesian approach that truly gave no preference to any data-generating model, the full set of probability distributions should be used. In this case, the ML method would show that the most likely model would be the *empirical distribution* generated from the training data [**?**]. The use of such a point estimate for novel predictions would dictate the use of the empirical risk metric. While it can be shown that empirical risk minimization is equivalent to expected risk minimization in the limit of training data volume, in the practical scenario of limited training data all the overfitting phenomena previously discussed will again be at issue.

This point suggests that the use of prior knowledge is *mandatory* for effective machine learning on most applications - the user must impose structure for the optimization routine to produce a high performance decision function. For non-probabilistic approaches to parametric supervised learning (neural networks, support vector machines, decision trees, etc.), the prior knowledge is imparted via the specific mapping from the parameter space directly to the decision function space. For Bayesian

parametric learning, the prior knowledge maps the parameter space to the full space of probability distributions, which then determines the optimal decision function.

2.3 Parametric Learning and Dimensionality Reduction

The mapping from the learned parameters to the decision function subspace becomes increasingly important in applications where the parameter space is of significantly lower dimensionality than the complete training data space; this is commonly referred to as "dimensionality reduction". While algorithms that learn from low-dimensional data can define enough parameters to guarantee retention of all statistically informative data components, those operating on high-dimensional data will typically have no practical option but to transform the data into compact features.

This is the case for many human recognition tasks of interest for machine learning research. Consider image recognition. A single 8-bit RGB (red, green, blue) pixel has over 16 million possible values, a moderate image size is 256-by-256 pixels, and some image databases can have over tens of millions of labeled images - the amount of digital memory needed to represent such a training set is enormous. Consequently, the data is compressed by a massive factor, forming the learned parameters.

Although dimensionality reduction of training data is generally suboptimal, for high-dimension tasks such as these it is also sensible. Despite the increased availability of training data for modern machine learning algorithms, the data space is still critically undersampled; this is sometimes referred to as the "curse of dimensionality" [**?**]. From this viewpoint, highly informative prior knowledge is a must for low-risk predictions. Significant dimensionality reduction of the training data will reflect a strong prior understanding of the data statistics and, if well conceived, will preserve the underlying information.

Chapter 3: Problem Statement

The goal of the thesis will be to investigate how a user's prior knowledge can be best exploited for the design of parametric decision functions for supervised machine learning. The initial research will focus on applications in which the joint set that an observed/unobserved datum is drawn from is finite; this will enable a simpler analysis of the role of dimensionality reduction in the selection of the finite-dimension function space.

As the prior knowledge used to define a learning algorithm is necessarily human-generated, the techniques developed in this thesis will specifically focus on regression and classification applications that mimic human prediction/recognition tasks. For example, speech recognition and image recognition algorithms are motivated by the desire to automate jobs typically performed by humans. Despite the recent advances in machine learning performance on such tasks, machines have yet to match the capabilities of their human counterparts.

3.1 Data Model and Objective

Consider an observable random element $\mathrm{x} \in \mathcal{X}$ and and unobservable random element $\mathrm{y} \in \mathcal{Y}$ which are jointly distributed according to an unknown probability distribution $\theta \in \Theta = \mathcal{P}(\mathcal{Y} \times \mathcal{X})$, such that $\mathrm{P}_{\mathrm{y,x}\,|\theta} = \theta$. The operator $\mathcal{P}$ returns the set of probability distributions over the argument set. Note that the uppercase PMF notation used throughout this section implies that the random elements are discrete; PDF's are used when x and/or y are continuous random variables/processes.

Also observed is a random sequence of N samples from θ, denoted $\mathrm{D} \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$; an alternative representation that can be used is $\mathrm{D} \equiv (\mathrm{Y}, \mathrm{X})$. The N data pairs are conditionally independent of one another and are identically distributed as

$\mathrm{P}_{\mathrm{D}_n\,|\theta} = \theta$. The samples are also conditionally independent from (y, x). Thus,

$$\mathrm{P}_{\mathrm{y,x,D}\,|\theta}(y, x, D|\theta) = \mathrm{P}_{\mathrm{y,x}\,|\theta}(y, x|\theta) \prod_{n=1}^{N} \mathrm{P}_{\mathrm{D}_n\,|\theta}\left(Y_n, X_n|\theta\right). \tag{3.1}$$

The task in supervised machine learning is to design a decision function $f : \mathcal{D} \mapsto \mathcal{H}^{\mathcal{X}}$ which produces a mapping from the space of the observed random elements to a decision space $\mathcal{H}$. Define the function space $\mathcal{F} = \left\{\mathcal{H}^{\mathcal{X}}\right\}^{\mathcal{D}}$, such that $f \in \mathcal{F}$. The learning functions are non-parametric, and there are no restrictions on the set of achievable functions $\mathcal{F}$.

The metric guiding the design is a loss function $\mathcal{L} : \mathcal{H} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ which penalizes the decision $h \in \mathcal{H}$ based on the value of y. The objective is to minimize the conditional expected loss, or conditional "risk",

$$\begin{aligned} \mathcal{R}_{\Theta}(f; \theta) &= \mathrm{E}_{\mathrm{y,x,D}\,|\theta}\left[\mathcal{L}\left(f(\mathrm{x}; \mathrm{D}), \mathrm{y}\right)\right] \\ &= \mathrm{E}_{\mathrm{D}\,|\theta}\left[\mathrm{E}_{\mathrm{x}\,|\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\mathcal{L}\left(f(\mathrm{x}; \mathrm{D}), \mathrm{y}\right)\right]\right]\right]. \end{aligned} \tag{3.2}$$

where the conditional independence of random element y from the training data D given the model θ is used. As the model θ is not observed, $\mathcal{R}_{\Theta} : \Theta \mapsto {\mathbb{R}_{\geq 0}}^{\mathcal{F}}$ is not a feasible objective function for optimization. This is the fundamental challenge of supervised learning: the true risk objective is unknown and the designer can never be precisely sure how well any decision function performs.

3.1.1 Clairvoyant Decision

It is informative to formulate the optimal decision function assuming the model θ was in fact observed; it will be referred to as the "clairvoyant" function, following terminology used in [**?**]. This clairvoyant decision function $f_{\Theta} : \Theta \mapsto \mathcal{F}$ is represented

by

$$f_\Theta(\theta) = \arg\min_{f \in \mathcal{F}} \mathcal{R}_\Theta(f; \theta) \ . \tag{3.3}$$

For a given set of observations x and D, the function $f_\Theta(\theta) \in \mathcal{F}$ selects the decision $h = \arg\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\Big[\,\mathcal{L}(h, \mathrm{y})\Big]$. Note the conditional independence of y from D in (3.2) - the knowledge of θ renders the training data D uninformative. As such, the range of the clairvoyant function is recast as $f_\Theta : \Theta \mapsto \mathcal{H}^{\mathcal{X}}$ and the decisions are

$$f_\Theta(\mathrm{x}; \theta) = \arg\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\Big[\,\mathcal{L}(h, \mathrm{y})\Big] \ . \tag{3.4}$$

The corresponding clairvoyant risk for a given model θ is

$$\begin{aligned} \mathcal{R}^*_\Theta(\theta) &\equiv \mathcal{R}_\Theta\Big(f_\Theta(\theta); \theta\Big) \\ &= \min_{f \in \mathcal{F}} \mathcal{R}_\Theta(f; \theta) \\ &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\Big[\,\mathcal{L}(h, \mathrm{y})\Big]\right] \ . \end{aligned} \tag{3.5}$$

3.1.2 Bayes Decision

To design an optimal decision function $f \in \mathcal{F}$, an operator must be chosen to remove the dependency of the conditional risk on θ and form an objective function $\mathcal{F} \mapsto \mathbb{R}_{\geq 0}$. One choice is to integrate over Θ; to ensure a non-negative objective value, the weighting function should be non-negative. Also, as scaling the objective function will not change its minimizing argument, the weighting function can be constrained to integrate to one. These are the requirements for a valid probability density function (PDF); as such, the model θ is treated as a random process and a Bayesian approach can be adopted.

Define the PDF $\mathrm{p}_\theta \in \mathcal{P}(\Theta)$. Now the Bayes risk can be formulated as

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_\theta\left[\mathcal{R}_\Theta(f;\theta)\right] \\ &= \mathrm{E}_{\mathrm{y,x,D}}\left[\mathcal{L}(f(\mathrm{x};\mathrm{D}),\mathrm{y})\right] \\ &= \mathrm{E}_\mathrm{D}\left[\mathrm{E}_{\mathrm{x}\,|\,\mathrm{D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\right]\right]\right]\end{aligned} \tag{3.6}$$

and y, x, and D are treated as jointly distributed random elements. Observe that the Bayesian predictive distributions can be represented as $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{D}}\left[\mathrm{P}_{\mathrm{x}\,|\theta}\right]$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}} = \mathrm{E}_{\theta|\,\mathrm{x,D}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$, the expected values of the corresponding clairvoyant distributions with respect to the model posteriors $\mathrm{p}_{\theta|\,\mathrm{D}}$ and $\mathrm{p}_{\theta|\,\mathrm{x,D}}$, respectively.

Finally, express the optimal learning function

$$f^* = \arg\min_{f\in\mathcal{F}} \mathcal{R}(f)\ . \tag{3.7}$$

The decision expressed by the learning function f^* given observed values of x and D is

$$\begin{aligned}f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\mathcal{L}(h,\mathrm{y})\right] \\ &= \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\theta|\,\mathrm{x,D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\mathcal{L}(h,\mathrm{y})\right]\right]\ .\end{aligned} \tag{3.8}$$

Thus, the Bayesian approach uses the model posterior $\mathrm{p}_{\theta|\,\mathrm{x,D}}$ to integrate out the dependency on the model given the observable random elements. The minimum Bayes risk is

$$\begin{aligned}\mathcal{R}^* &\equiv \mathcal{R}(f^*) \\ &= \min_{f\in\mathcal{F}} \mathcal{R}(f) \\ &= \mathrm{E}_\mathrm{D}\left[\mathrm{E}_{\mathrm{x}\,|\,\mathrm{D}}\left[\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D}}\left[\mathcal{L}(h,\mathrm{y})\right]\right]\right]\ .\end{aligned} \tag{3.9}$$

3.2 Applications to Common Loss Functions

In this section, loss functions typical for classification and regression applications, specifically the 0-1 loss function and the squared-error loss function, are adopted. The conditional risk (3.2) is assessed, clairvoyant learners (3.4) are found, and the clairvoyant risk (3.5) is expressed.

3.2.1 Regression: the Squared-Error Loss

The squared-error (SE) loss function is arguably the most commonly used loss function for regression, or in fact for any estimation problem. This can be attributed to its quadratic form, which enables a closed-form expression of the minimizing estimation function.

It is assumed that the unobserved random element y is a scalar random variable; that is, $\mathcal{Y} \subseteq \mathbb{R}$. Additionally, the learning function's estimate is allowed to assume real numbers; thus, $\mathcal{H} = \mathbb{R} \supseteq \mathcal{Y}$.

The loss function is defined as

$$\mathcal{L}(h, y) = (h - y)^2 \, . \tag{3.10}$$

Substituting the squared-error loss into (3.2), the conditional squared-error risk is

$$\begin{aligned} \mathcal{R}_\Theta(f; \theta) &= \mathrm{E}_{\mathrm{x}\,|\theta} \left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta} \left[\mathrm{E}_{\mathrm{D}\,|\theta} \left[\Big(f(\mathrm{x}; \mathrm{D}) - \mathrm{y} \Big)^2 \right] \right] \right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta} \left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta} \right] + \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\theta} \left[\Big(f(\mathrm{x}; \mathrm{D}) - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta} \Big)^2 \right] , \end{aligned} \tag{3.11}$$

a sum of two terms. The first term is the expected conditional variance of the true predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$. The second term is the expected squared bias between the Bayesian estimate and the true mean $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$.

3.2.1.1 Clairvoyant Estimation

To find the clairvoyant estimator, the squared-error loss is substituted into (3.4); note that the objective function is quadratic over the argument $h \in \mathcal{H} = \mathbb{R}$. It is easily shown that the function over h is positive-definite; as such, the minimizing decision h is the sole stationary point. Setting the first derivative of the function to zero, the clairvoyant estimate is the expected value of y given the model θ and the observed value x, such that

$$\begin{aligned} f_\Theta(\mathrm{x};\theta) &= \arg\min_{h\in\mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[(h-\mathrm{y})^2\right] \\ &= \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\ . \end{aligned} \tag{3.12}$$

Substituting the loss and clairvoyant function into (3.5), the resulting clairvoyant risk is

$$\begin{aligned} \mathcal{R}^*_\Theta(\theta) &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[(\mathrm{y}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta})^2\right]\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]\ . \end{aligned} \tag{3.13}$$

Observe that the general conditional risk (3.11) can be represented as $\mathcal{R}_\Theta(f;\theta) = \mathcal{R}^*_\Theta(\theta) + \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\left(f(\mathrm{x};\mathrm{D}) - f_\Theta(\mathrm{x};\theta)\right)^2\right]$. The first summand is equal to the clairvoyant squared-error; the second term is dependent on the difference between the general estimate and the clairvoyant estimate. Figure 3.1 displays the clairvoyant risk for models $\theta(\cdot, x)$ independent of x.

3.2.1.2 Bayesian Estimation

To find the optimal estimator, the squared-error loss is substituted into (3.8). Again, the function over h is positive-definite; as such, the minimizing decision h is the sole stationary point. Setting the first derivative of the function to zero, the optimal

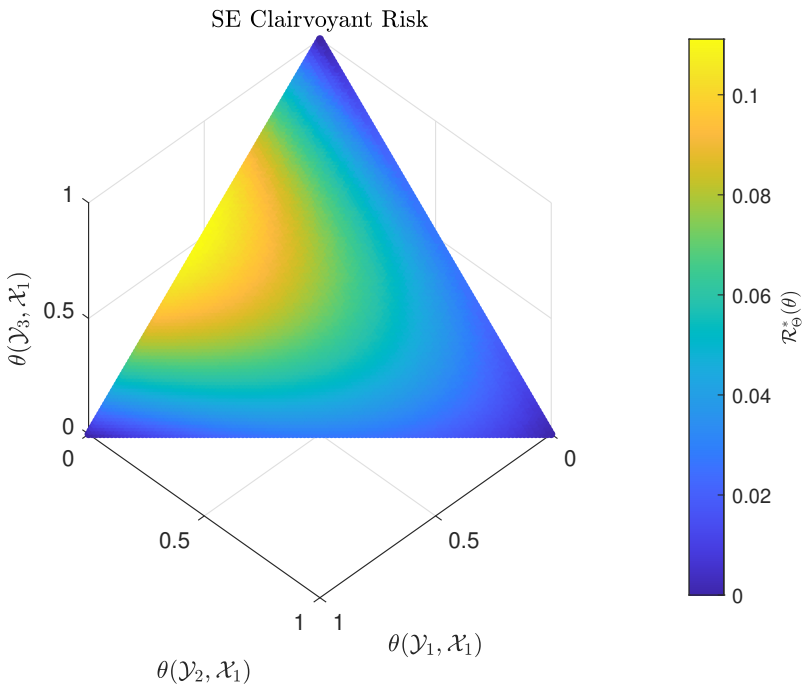

Figure 3.1: Clairvoyant Risk for the Squared-Error Loss Function, constant $\theta(\cdot, x)$

estimate is the expected value of y given the training data and the observed value x, such that

$$\begin{aligned} f^*(\mathrm{x}; \mathrm{D}) &= \arg\min_{h \in \mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \left[(h - \mathrm{y})^2 \right] \\ &= \mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}} \left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta} \right] . \end{aligned} \tag{3.14}$$

An interesting form for the optimal estimator is $f^*(\mathrm{x}; \mathrm{D}) = \mathrm{E}_{\theta|\,\mathrm{x},\mathrm{D}} \left[f_\Theta(\mathrm{x}; \theta) \right]$. Substituting the squared-error loss into the second line of (3.8), the optimal Bayes estimator is the conditional expected value of the clairvoyant estimate with respect to the model posterior distribution.

The Bayes squared-error risk for a general learning function is

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y},\mathrm{x}\,|\,\theta}\left[\Big(f(\mathrm{x};\mathrm{D})-\mathrm{y}\Big)^2\right]\right]\right] \\ &= \mathrm{E}_{\theta}\left[\mathcal{R}^*_{\Theta}(\theta)\right] + \mathrm{E}_{\mathrm{x},\mathrm{D},\theta}\left[\Big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\Big)^2\right] \\ &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] + \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\Big(f(\mathrm{x};\mathrm{D})-\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\Big)^2\right] .\end{aligned} \tag{3.15}$$

Substituting the optimal estimator (3.14) into Equation (3.15), the minimum Bayes risk is the expected conditional variance

$$\begin{aligned}\mathcal{R}^* &= \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\right] \\ &= \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mathrm{C}_{\theta|\,\mathrm{x},\mathrm{D}}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]\right] .\end{aligned} \tag{3.16}$$

The first term is the irreducible risk. The second term is the expected variance of the clairvoyant estimate $f_{\Theta}(\mathrm{x};\theta) = \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$ with respect to the model posterior PDF $\mathrm{p}_{\theta|\,\mathrm{x},\mathrm{D}}$.

3.2.2 Classification: the 0-1 Loss

In this section, the developed framework is applied to a common machine learning task: classification. In classification problems, the set $\mathcal{Y}$ is countable and typically finite. Furthermore, the hypothesis space is usually identical to the unobserved variable space; that is, $\mathcal{H} = \mathcal{Y}$. The 0-1 loss function is the most widely used for these problems; it is represented as

$$\mathcal{L}(h, y) = 1 - \delta[h, y] . \tag{3.17}$$

Applying the 0-1 loss, the conditional risk (3.2) for a general classifier is

$$\begin{aligned}\mathcal{R}_{\Theta}(f;\theta) &= 1 - \mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y},\mathrm{x}\,|\,\theta}\left[\delta\Big[f(\mathrm{x};\mathrm{D}),\mathrm{y}\Big]\right]\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\theta\Big(f(x;\mathrm{D}),x\Big)\right] .\end{aligned} \tag{3.18}$$

3.2.2.1 Clairvoyant Hypothesis

To find the clairvoyant classifier, the 0-1 loss is substituted into (3.4); given an observation x, the optimum hypothesis is simply the value y that maximizes the conditional model $\tilde{\theta}(x)$,

$$\begin{aligned} f_{\Theta}(\mathrm{x};\theta) &= \arg\min_{h\in\mathcal{Y}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[1-\delta[h,y]\right] \\ &= \arg\max_{h\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}(h|\,\mathrm{x},\theta) \\ &= \arg\max_{y\in\mathcal{Y}} \theta(y,\mathrm{x})\ . \end{aligned} \tag{3.19}$$

Substituting the 0-1 loss and clairvoyant hypothesis into (3.5), the resulting clairvoyant risk is

$$\begin{aligned} \mathcal{R}^{*}_{\Theta}(\theta) &= 1-\mathrm{E}_{\mathrm{x}\,|\theta}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}(y|\,\mathrm{x},\theta)\right] \\ &= 1-\sum_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \theta(y,x)\ . \end{aligned} \tag{3.20}$$

Figure 3.2 displays the clairvoyant risk for models $\theta(\cdot,x)$ independent of x. Intuitively, the models that are more concentrated lead to lower probability of error.

3.2.2.2 Bayesian Hypothesis

To determine the optimal Bayesian classification function, the 0-1 loss from Equation (3.17) is substituted into Equation (3.8) to find

$$\begin{aligned} f^{*}(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathcal{Y}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[1-\delta[h,\mathrm{y}]\right] \\ &= \arg\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x},\mathrm{D})\ . \end{aligned} \tag{3.21}$$

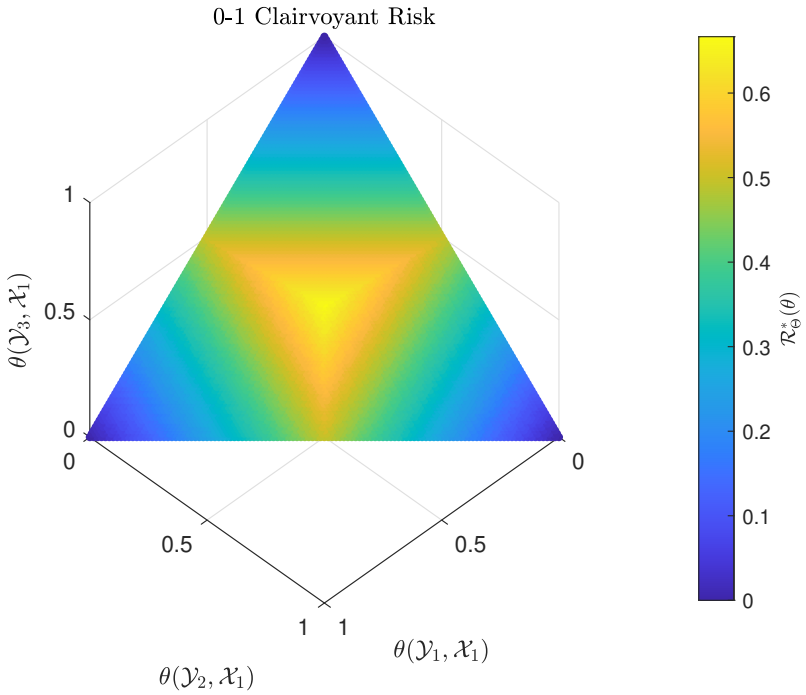

Figure 3.2: Clairvoyant Risk for the 0–1 Loss Function, constant $\theta(\cdot, x)$

The optimal classifier chooses the value $y \in \mathcal{Y}$ that maximizes the Bayesian predictive PMF for the observed values of x and D.

Using the 0-1 loss, the minimum Bayes probability of error (3.9) is

$$\mathcal{R}(f) = 1 - \mathrm{E}_{\mathrm{x,D}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}\left(f(\mathrm{x};\mathrm{D})|\,\mathrm{x},\mathrm{D}\right)\right] \ . \tag{3.22}$$

Substituting the optimal learner (3.21) into the general risk (3.22), the minimum probability of error is

$$\mathcal{R}^* = 1 - \mathrm{E}_{\mathrm{x,D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\right] \ . \tag{3.23}$$

3.3 Application to Human Recognition Tasks

Human recognition abilities are certainly remarkable and can in part be attributed to the immeasurable volume of "training data" that we consume during our lives.

Figure 3.3: Randomly generated RGB image

However, individual humans must be said to have a type of "prior knowledge" even before being born - all healthy human beings pre-process our multi-sensory input in essentially the same way and assess similarities between observations using the same metrics.

This pre-processing undoubtedly exploits the highly structured nature of our sensory records. Consider visual recognition. Not only do we process a fraction of the electromagnetic spectrum, but the matter which our received energy reflects off of is highly structured. To underscore this point, consider the uniformly randomly generated RGB image in Figure 3.3, which could no human would ever mistake for an image recorded from the visual spectrum. Indeed, the space of sensible RGB images for visual recognition only comprises a minuscule fraction of the complete RGB image space.

It may also be said that all humans possess a mechanism for "dimensionality reduction" - our mental records of our observations are not complete, but rather are represented by a compact set of descriptors. Understanding these "feature extraction" procedures performed by our sensory organs has been a research interest for those seeking to improve the automation of human recognition.

Another important perspective regarding how humans make predictions is that the

Figure 3.4: Handwritten digit samples from the MNIST Database

labels we use are almost always *extrinsic* to our observations. That is, the categories we use to describe things are of our own design - in fact, much of our internal classification procedure can be thought of as *unsupervised learning*. To exemplify this, consider speech recognition. Spoken "words" are nothing more than fluctuations of air pressure, but we use man-made descriptors to efficiently represent the distinct patterns we emit. In image recognition on ImageNet [**?**], different images are labeled as "container ships" or "lifeboats"; this distinction is man-made as well.

From this point of view, it should be unsurprising that humans outperform machines on these recognition tasks - we define the classes ourselves, ensuring joint class-data probability distributions that lead to minimal risk. With sufficient prior knowledge, parametric decision functions that meet these levels of performance should be realizable. To illustrate this point, consider the handwritten digit samples [**?**] shown in Figure 3.4. Undoubtedly, most human observers will have no problem accurately classifying all of these samples.

Chapter 4: Related Work

4.1 Statistical Learning

Supervised learning has a long history predating the widespread acceptance of the newly popular phrase "machine learning". Such a broad technical area inescapably spans a variety of fields - learning theory is of theoretical interest in applied mathematics and statistics, and it has found practical interest for applications in electrical engineering and computer science.

Many important results in supervised learning theory come from a statistical foundation. Fundamental questions about the existence of asymptotically consistent estimators have found their answers in the works of researchers including Stone [**?**], who first studied the consistency of the k-nearest neighbor classification rule. In 1971, Vapnik and Chervonenkis set down theory on empirical risk minimization [**?**], providing bounds on the risk achievable by a classifier when it is restricted to a specific group of functions. Although works such as these are not typically considered "Bayesian", they do implicitly underscore the importance of restricting the design of the decision function to a limited function space, especially when the empirical risk objective drives the design.

Non-Bayesian treatments of statistical inference such as these commonly provide results in the form of asymptotic trends for large sample sizes or as risk lower bounds [**?**]. However, not as much is offered in the way of practical solutions for limited volumes of data - this is where the Bayesian perspective is most valuable. Since Bayes put forth his theorem in 1763 [**?**], interest in a Bayesian approach to statistical inference has waned and waxed. The varying popularity of these methods can be attributed in part to a philosophical belief that prior assumptions degraded the information inferred from the data alone; additionally, the computational burden of Bayesian methods (via

approximations of integrals) made them unattractive compared to classical methods [**?**].

While modern advances in computing have made Bayesian techniques more applicable to a variety of problems, the computational challenges inherent in Bayesian inference persist. Many modern research directions on Bayesian learning attempt to move away from the most tractable (and most presumptuous) methods, such as those that assume Gaussian data distributions, while still providing computational methods for practical implementation. Such methods will likely have relevance for this thesis, especially for Bayesian learners based on priors with low-dimensional support sets.

One of the more popular approaches is to use a *hierarchical prior*, which defines the data-generating model in terms of other unknown "latent variables", or hyperparameters, and then remove the dependency on the model with a single procedure. An example is the expectation-maximization (EM) algorithm [**?**], which performs an alternating optimization by iteratively evaluating expectations of the data likelihood function and then forming maximum likelihood estimates of the model hyperparameters; after convergence, the point estimates of the hyperparameters are used to formulate the model posterior and inferences are made as usual. This algorithm is widely used for Gaussian mixture models (GMM's) in clustering applications.

Use of the EM algorithm requires the model posterior given the observed data. When evaluation of the model posterior is intractable, different methods must be used - a significant portion of research on Bayesian methods focuses on overcoming this difficulty. Approximate inference methods are one such class of solutions. With variational approximation, the EM algorithm can be approximated by forming an estimate of the posterior from a class of distributions that enables a tractable solution [**?**]. Another alternative is the expectation propagation algorithm [**?**], which estimates the posterior by performing an optimization on the Kullback-Leiber divergence metric. Non-deterministic methods are also popular for approximate inference, using random sampling to approximate expectations; some notable approaches are

importance sampling [?] and Markov chain Monte Carlo (MCMC) methods, including the Metropolis-Hastings algorithm [?]. These approximation methods will be used as necessary for Bayesian learning with limited-support priors.

Use of the Dirichlet distribution for Bayesian inference is typically found in non-parametric learning by way of the continuous Dirichlet process [?]. The most common use is for unsupervised learning applications; specifically, clustering with infinite numbers of clusters [?]. For supervised learning, however, most current research on Bayesian learning from data drawn from continuous sets uses Gaussian processes [?, ?] to weight decision functions rather than the data-generating probability distributions.

4.2 Feature Extraction for Speech/Image Recognition

Existing research on human recognition is extensive; some related works focusing on feature extraction may be used to guide design Bayesian methods for this thesis.

In speech recognition, feature design is motivated by investigations suggesting that the human auditory system performs a real-time spectral analysis [?] of its input. Furthermore, the frequency information is processed on a logarithmic scale [?]; this perspective has motivated the widespread adoption of features like the Mel frequency cepstral coefficients (MFCC's) [?], which attempt to compress signals into a lower-dimensional space for word/phoneme classification. Similarly, for image recognition, work by Campbell [?, ?] uses a Fourier-based signal decomposition to isolate informative signal components; for example, visual images contain most of their energy in low-frequency Fourier coefficients.

Both of these human recognition applications highlight the inclination of our sensory processing systems to decompose information in a variety of ways and at a variety of scales. This observation underscores the benefit of using features based on a multiresolution signal decomposition. In particular, the family of wavelet representations [?] is popular for these learning applications; a wide range of interesting

research directions, including multiscale edge detection and curvelet approximation, have provided features for efficient signal description.

Our understanding of certain human sensory processes has led to specialized signal transformations that overcome undesirable effects in more standard feature extraction. For example, in optical character recognition (OCR) (and in fact for most human recognition tasks), the translation dependency of a feature can significantly raise the probability of error. Reconciling our desire for features that describe spatial locality with the need for translation-invariant representation is challenging - even popular features such as those based on multiresolution wavelet transformations [**?**] do not provide both. Focused research on shift-invariant features for two-dimensional data [**?**] aims to ensure this property and provide learning algorithms with features that well approximate those utilized in our own cognition.

Chapter 5: Approach

Having provided the perspective that *all* machine learning algorithms either implicitly or explicitly use prior knowledge, this thesis will focus on a strictly Bayesian approach to parametric supervised learning. Many existing treatments of Bayesian machine learning address only parametric learners, which reduce the data dimensionality, and yet fail to sufficiently consider the implications of how the mapping between the finite set of parameters and the probability model is selected. As such, the framework set down previously will be used; it is general enough to accommodate learning designs based on any prior distribution, even those with full support over potentially infinite-dimensional function spaces. Designs using both full support and limited support prior distributions will be considered for application to human recognition tasks.

5.1 Model and Data Transformations

To simplify the analyses, the model θ and the training data D will be frequently interpreted with alternate representations. The model θ is split into conditional and marginal distributions using a bijective transformation. The training data is compressed into the empirical distribution, a sufficient statistic for the model; this empirical distribution is decomposed into "conditional" and "marginal" forms as well.

5.1.1 Marginal and Conditional Distributions of θ

As only y is unobservable, it will be useful to decompose the model distribution as $\theta \equiv (\theta', \tilde{\theta})$. First, introduce the marginal distribution $\theta' \equiv \sum_{y \in \mathcal{Y}} \theta(y, \cdot) \in \mathcal{P}(\mathcal{X})$; note that the summation is replaced by an integral when y is a continuous random variable. Next, introduce the conditional distributions $\tilde{\theta} \in \mathcal{P}(\mathcal{Y})^{\mathcal{X}}$ defined as $\tilde{\theta}(x) \equiv \theta(\cdot, x)/\theta'(x)$.

This decomposition enables the clairvoyant distributions to be represented as $\mathrm{P}_{\mathrm{x}\,|\theta} = \mathrm{P}_{\mathrm{x}\,|\theta'} = \theta'$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\tilde{\theta}} = \tilde{\theta}(\mathrm{x})$; these distributions will be of recurring importance.

5.1.2 Sufficient Statistic: the Empirical PMF

For countable sets $\mathcal{Y}$ and $\mathcal{X}$, the distribution of D conditioned on the model can be formulated as

$$\begin{aligned} \mathrm{P}_{\mathrm{D}\,|\theta}\left(D|\theta\right) &= \prod_{n=1}^{N} \mathrm{P}_{\mathrm{D}_n\,|\theta}\left(D_n|\theta\right) \\ &= \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{N}(y,x;D)} \ , \end{aligned} \tag{5.1}$$

where the dependency on the training data D is expressed though a transform function $\bar{N} : \mathcal{D} \mapsto \bar{\mathcal{N}}$, where the range is

$$\bar{\mathcal{N}} = \left\{ \bar{n} \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \bar{n}(y,x) = N \right\} \tag{5.2}$$

and the function is defined as

$$\begin{aligned} \bar{N}(y,x;D) &= \sum_{n=1}^{N} \delta\Big[(y,x), D_n\Big] \\ &= \sum_{n=1}^{N} \delta\left[y, Y_n\right] \delta\left[x, X_n\right] \ . \end{aligned} \tag{5.3}$$

This function counts the number of occurrences of the pair (y,x) in the training set D.

Note that $\mathrm{P}_{\mathrm{D}\,|\theta}$ depends on the training data D only through the transform $\bar{N}$; $\bar{N}(\mathrm{D})$ is thus a sufficient statistic [**?**] for the model θ. Consequently, other distributions of interest P_{D}, $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}}$, and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$ will also depend on D via $\bar{N}(\mathrm{D})$. As such, it is useful to define a new random process $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D}) \in \bar{\mathcal{N}}$.

Frequently, the corresponding distributions $\mathrm{P}_{\bar{\mathrm{n}}}$, $\mathrm{P}_{\mathrm{x}\,|\,\bar{\mathrm{n}}}$, and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ will be used to find

the optimal decision functions and the minimum risk. Note that $\mathcal{M}\left(\bar{N}(D)\right) \mathrm{P}_{\mathrm{D}\,|\,\theta}(D|\theta) = \mathrm{P}_{\bar{\mathrm{n}}\,|\,\theta}\left(\bar{N}(D)|\theta\right)$, where $\mathcal{M}$ is the multinomial operator. Also note that $\mathrm{P}_{\mathrm{x}\,|\,\mathrm{D}}(D) = \mathrm{P}_{\mathrm{x}\,|\,\bar{\mathrm{n}}}\left(\bar{N}(D)\right)$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(x, D) = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\left(x, \bar{N}(D)\right)$.

The cardinality of the random process's domain is $|\bar{\mathcal{N}}| = \mathcal{M}\left(\{N, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1\}\right)$; this can be shown using the stars-and-bars method [**?**]. The cardinality of original set is $|\,\mathcal{D}\,| = \left(|\,\mathcal{Y}\,||\,\mathcal{X}\,|\right)^N$; thus, $|\bar{\mathcal{N}}| \leq |\,\mathcal{D}\,|$ and the sufficient statistic compactly represents the valuable information in the training data. Also, observe that the set $\{\bar{n}/N : \bar{n} \in \bar{\mathcal{N}}\} \subset \Theta$ and thus the empirical distribution $\bar{N}(\mathrm{D})/N$ assumes one of a finite number of the elements from Θ.

Conditioned on the model θ, the PMF of $\bar{\mathrm{n}}$ is a multinomial distribution [**?**],

$$\begin{aligned}
\mathrm{P}_{\bar{\mathrm{n}}\,|\,\theta}(\bar{n}|\theta) &= \sum_{D:\bar{N}(D)=\bar{n}} \mathrm{P}_{\mathrm{D}\,|\,\theta}(D|\theta) \\
&= \left|\{D : \bar{N}(D) = \bar{n}\}\right| \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{n}(y,x)} \\
&= \mathcal{M}(\bar{n}) \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{n}(y,x)} \\
&= \mathrm{Multi}\left(\bar{n}; N, \theta\right),
\end{aligned} \tag{5.4}$$

where the multinomial operator $\mathcal{M}$ is used.

Also, observe that the maximum likelihood estimate of θ given the training statistic [**?**] is the empirical distribution,

$$\theta_{\mathrm{ML}}\left(\bar{n}\right) = \arg\max_{\theta\in\Theta} \mathrm{P}_{\bar{\mathrm{n}}\,|\,\theta}(\bar{n}|\theta) = \frac{\bar{n}}{N}\,. \tag{5.5}$$

5.1.3 Marginal and Conditional Distributions of $\bar{\mathrm{n}}$

Also of interest are the marginal and conditional distributions of the joint training data.

As before, the dependency on the training data can be simplified using a sufficient

statistic. Introduce the "marginalized" random process n' over the set $\mathcal{X}$, defined as $\mathrm{n}' \equiv \sum_{y \in \mathcal{Y}} \bar{\mathrm{n}}(y, \cdot) \in \mathcal{N}'$, where

$$\mathcal{N}' = \left\{ n' \in {\mathbb{Z}_{\geq 0}}^{\mathcal{X}} : \sum_{x \in \mathcal{X}} n'(x) = N \right\} . \tag{5.6}$$

By the aggregation property of Multinomial random processes [**?**], the aggregation conditioned on the model θ is distributed as $\mathrm{n}' | \theta \sim \mathrm{Multi}(N, \theta')$.

Also of interest is the distribution of $\bar{\mathrm{n}}$ conditioned on its aggregation n'. Using the multinomial distribution properties proven in Appendix A.2, it can be shown that when conditioned on the model θ as well, the PMF of $\bar{\mathrm{n}}$ is

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}} | \, \mathrm{n}', \theta}(\bar{n} | n', \theta) &= \prod_{x \in \mathcal{X}} \left[\mathcal{M}\left(\bar{n}(\cdot, x)\right) \prod_{y \in \mathcal{Y}} \tilde{\theta}(y; x)^{\bar{n}(y,x)} \right] \\ &= \prod_{x \in \mathcal{X}} \mathrm{Multi}\left(\bar{n}(\cdot, x); n'(x), \tilde{\theta}(x) \right) , \end{aligned} \tag{5.7}$$

over the domain $\left\{ \bar{n} \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y} \times \mathcal{X}} : \sum_{y \in \mathcal{Y}} \bar{n}(y, \cdot) = n' \right\}$. Observe that conditioning on the aggregation renders the function segments $\bar{\mathrm{n}}(\cdot, x)$ independent of one another and that they are also Multinomial, such that $\bar{\mathrm{n}}(\cdot, x) | \, \mathrm{n}'(x), \theta \sim \mathrm{Multi}\left(\mathrm{n}'(x), \tilde{\theta}(x) \right)$. Furthermore, the dependency on θ is expressed through the conditional model $\tilde{\theta}$.

5.2 Full Support Priors

Using this framework, an analysis of regression and classification algorithms based on full support priors (both informative and non-informative) is of interest. In the literature, typically only one type of probability distribution with full support is discussed: the Dirichlet distribution. Not only does the Dirichlet distribution have full support, but it can also be parameterized in different ways to achieve both informative and non-informative priors. Additionally, it is a *conjugate prior* for likelihood functions in the exponential family [**?**] and is thus an attractive option for supervised learning

with independently and identically distributed training data.

The Dirichlet PDF for the model random process $\theta \in \Theta$ is [**?**]

$$\begin{aligned} \mathrm{p}_\theta(\theta) &= \beta(\alpha)^{-1} \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\alpha(y,x)-1} \\ &= \mathrm{Dir}\left(\theta;\alpha\right), \end{aligned} \tag{5.8}$$

where the user-selected PDF parameters $\alpha : \mathcal{Y}\times\mathcal{X} \mapsto \mathbb{R}^+$ are introduced and β is the generalized beta function.

The parameter α controls around which models θ the PDF concentrates and how strongly. For convenience, introduce the concentration parameter $\alpha_0 \equiv \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \alpha(y,x)$.

The first and second joint moments of the model are

$$\mu_\theta = \frac{\alpha}{\alpha_0} \tag{5.9}$$

and

$$\mathrm{E}_\theta\left[\theta(y,x)\theta(y',x')\right] = \frac{\alpha(y,x)\alpha(y',x') + \alpha(y,x)\delta[y,y']\delta[x,x']}{\alpha_0(\alpha_0+1)}\,. \tag{5.10}$$

Observe that $\mathrm{P}_{\mathrm{y,x}} = \mu_\theta = \alpha/\alpha_0$. The covariance is

$$\begin{aligned} \Sigma_\theta(y,x,y',x') &= \mathrm{E}_\theta\left[\Big(\theta(y,x)-\mu_\theta(y,x)\Big)\Big(\theta(y',x')-\mu_\theta(y',x')\Big)\right] \\ &= \frac{\mu_\theta(y,x)\delta[y,y']\delta[x,x'] - \mu_\theta(y,x)\mu_\theta(y',x')}{\alpha_0+1}\,. \end{aligned} \tag{5.11}$$

Also, for $\alpha(y,x) > 1$, the maximizing value of the distribution is

$$\theta_{\max} = \arg\max_{\theta\in\Theta} \mathrm{p}_\theta(\theta) = \frac{\alpha-1}{\alpha_0 - |\,\mathcal{Y}\,||\,\mathcal{X}\,|}\,. \tag{5.12}$$

This can be easily shown by maximizing the logarithm of the distribution using the

method of Lagrange multipliers.

Of specific interest is how p_θ changes as the concentration parameter approaches its limiting values. For $\alpha_0 \to \infty$, the PDF concentrates at its mean, resulting in

$$\mathrm{p}_\theta(\theta) \to \delta\left(\theta - \frac{\alpha}{\alpha_0}\right) . \tag{5.13}$$

Conversely, for $\alpha_0 \to 0$, the PDF tends toward

$$\mathrm{p}_\theta(\theta) \to \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \frac{\alpha(y,x)}{\alpha_0} \delta\Big(\theta - \delta[\cdot, y]\delta[\cdot, x]\Big) , \tag{5.14}$$

which distributes its weight among the $|\,\mathcal{Y}\,||\,\mathcal{X}\,|$ models with an ℓ_0 norm satisfying $\|\theta\|_0 = 1$. Note that the Dirac delta for these formulas is defined on the set Θ, such that $\int_\Theta \delta(\theta)\mathrm{d}\theta = 1$.

These trends are demonstrated with Figure 5.1. The cardinalities $|\,\mathcal{Y}\,| = 3$ and $|\,\mathcal{X}\,| = 1$ are chosen to enable visualization, despite the implication that x is deterministic; these cardinalities will be used for many subsequent figures as well. Note that for $\alpha_0 = 2.99$, $\alpha < 1$ and the PDF values at the boundaries of the domain tend to infinity; this is not captured by the plot color scale.

5.2.1 Uniform Prior

When the parameterizing function is $\alpha(y, x) = 1$, the distribution becomes a uniform PDF and is represented as

$$\mathrm{p}_\theta = \Big(|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1\Big)! . \tag{5.15}$$

Note that the concentration parameter is $\alpha_0 = |\,\mathcal{Y}\,||\,\mathcal{X}\,|$ and $\mathrm{P}_{\mathrm{y,x}} = \Big(|\,\mathcal{Y}\,||\,\mathcal{X}\,|\Big)^{-1}$ is also uniform.

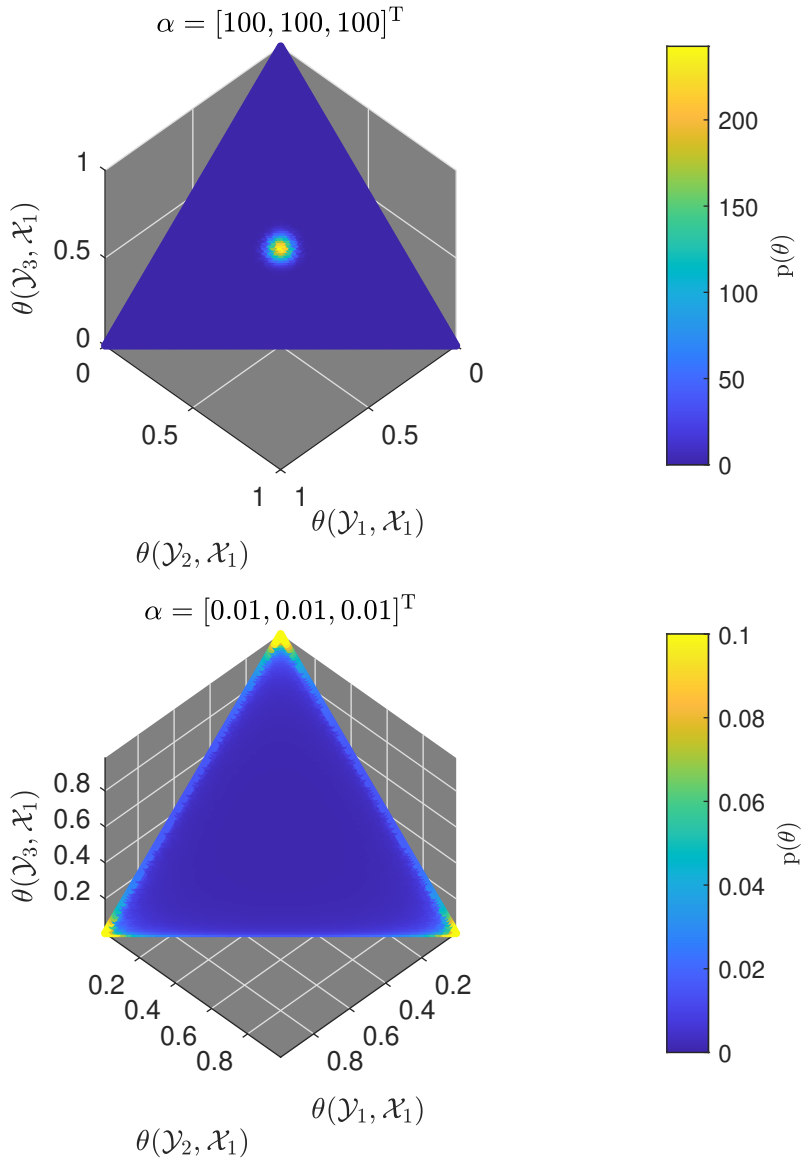

Figure 5.1: Model prior PDF for different concentrations α_0

5.2.2 Marginal and Conditional Distributions

The marginal distribution Θ' and the conditional distribution $\tilde{\Theta}$ will also be of interest. By the aggregation property [**?**], Θ' is a Dirichlet random process parameterized by $\alpha' : \mathcal{X} \mapsto \mathbb{R}^+$, where $\alpha' \equiv \sum_{y \in \mathcal{Y}} \alpha(y, \cdot)$. Note that $\mathrm{P}_{\mathrm{x}} = \mu_{\Theta'} = \alpha'/\alpha_0$.

Also of interest is the distribution of the predictive model $\tilde{\Theta}$ conditioned on the marginal Θ'. As demonstrated in Appendix A.1, these random processes are jointly distributed as

$$\begin{aligned} \mathrm{p}_{\tilde{\Theta}|\Theta'}\left(\tilde{\theta}|\theta'\right) &= \prod_{x \in \mathcal{X}} \left[\beta\Big(\alpha(\cdot, x)\Big)^{-1} \prod_{y \in \mathcal{Y}} \tilde{\theta}(y; x)^{\alpha(y,x)-1} \right] \\ &= \prod_{x \in \mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x); \alpha(\cdot, x)\right), \end{aligned} \tag{5.16}$$

a product of Dirichlet distributions defined on $\tilde{\theta} \in \mathcal{P}(\mathcal{Y})^{\mathcal{X}}$. As shown, the processes $\tilde{\Theta}(x)$ are Dirichlet with parameterizing functions $\alpha(\cdot, x)$, independent of one another, and independent of the marginal distribution Θ'. Observe that the values $\alpha'(x)$ represent the concentration parameters for the individual Dirichlet processes; also, note that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \mu_{\tilde{\Theta}(\mathrm{x})} = \alpha(\cdot, \mathrm{x})/\alpha'(\mathrm{x})$.

5.3 Limited Support Priors

Analyses will also be performed for decision functions derived from a variety of priors with limited-dimensional support, such that $\dim(\Theta) < |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1$. The implications of a degenerate prior probability distribution will be explored, with focus on limited training data volume benefits and issues. The asymptotic risk achieved by such learners will be of specific interest as well; while full support priors guarantee asymptotically consistent estimation of the true data-generating model, degenerate priors cannot ensure that the model is identified. Furthermore, the relationship between support dimensionality and the computational complexity of implementation will be explored -

with fewer degrees of freedom needed to characterize the support, optimal decision functions should be realizable with fewer parameters.

Investigation into Bayesian learning with low-dimensional priors will have a special focus on applicability to human recognition tasks. In consideration of the insights into human recognition discussed in Section 3, special priors will be defined to reflect our own capabilities.

The first motivating insight is that the data processed during human recognition is highly structured and seems to have an *intrinsic dimensionality* that is much lower than that of the full dimensional space that they reside in. With this in mind, a potential restriction of the model support is

$$\left\|\theta'\right\|_0 = \sum_{x\in\mathcal{X}} \Big(1 - \delta\big[\theta'(x), 0\big]\Big) \leq M_{\mathcal{X}} < |\,\mathcal{X}\,|\,, \tag{5.17}$$

which only considers marginal models with restricted ℓ_0 norms. Figure 5.2 shows how such a restriction for $|\,\mathcal{X}\,| = 3$ lowers the intrinsic dimensionality of the prior support from 2-dimensional to 1-dimensional. Note that fewer parameters are needed to describe points in this subspace relative to a full support prior.

The second insight into human recognition is that our internal "pre-processing" greatly reduces the complexity of our sensory input. As such, machine learning functions should be able to perform feature extraction and effect significant dimensionality reduction without any loss of performance. This suggests that the prior distribution's support should be limited to a subspace that guarantees a low-dimensional *sufficient statistic*.

To this effect, consider a transform function $T : \mathcal{X} \mapsto \mathcal{T}$ which maps to a lower-cardinality range $|\,\mathcal{T}\,| \leq M_{\mathcal{X}} < |\,\mathcal{X}\,|$. The general function can be defined via its partitioning of the domain $\mathcal{X}$ using the group of sets

$$\mathcal{X}_{\mathrm{s}}(t) = \Big\{x \in \mathcal{X} : T(x) = t\Big\}\,, \quad t \in \mathcal{T}\,, \tag{5.18}$$

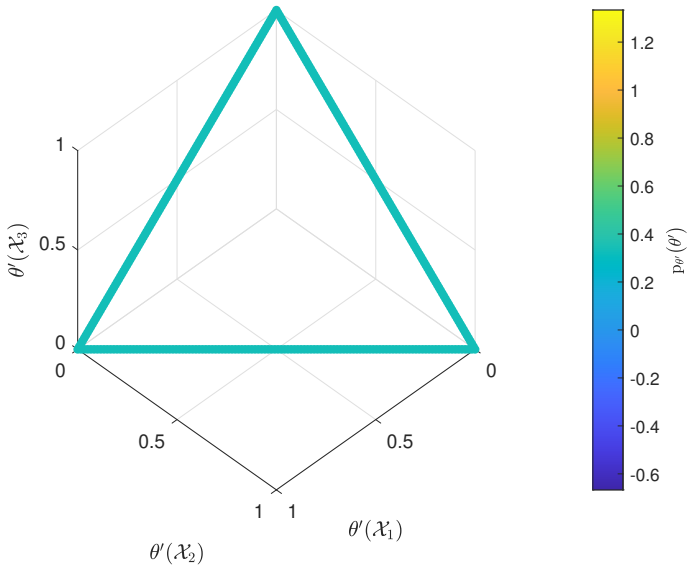

Figure 5.2: Marginal model prior for $\left\|\theta'\right\|_0 \leq 2$

satisfying $\mathcal{X} = \bigcup_{t\in\mathcal{T}} \mathcal{X}_{\mathrm{s}}(t)$.

Defining the transformed data random element $\mathrm{t} \equiv T(\mathrm{x})$, it can be shown that observation of the transform value refines the conditional data PMF as

$$\mathrm{P}_{\mathrm{x}\,|\theta',\mathrm{t}}(x|\theta',t) = \frac{\theta'(x)}{\sum_{x'\in\mathcal{X}_{\mathrm{s}}(t)}\theta'(x')}\ , \tag{5.19}$$

that is, the distribution $\mathrm{P}_{\mathrm{x}\,|\theta'} = \theta'$ normalized over the restricted domain defined by $T(x) = t$.

For the transformed element to be a sufficient statistic, it must preserve all the original information from the observed data for inferences regarding θ'. Formally, the requirement is that the observation x is conditionally independent of the model θ' given the statistic t [?], which for this framework implies

$$\frac{\theta'(x)}{\sum_{x'\in\mathcal{X}_{\mathrm{s}}(t)}\theta'(x')} = g(x;t)\ , \quad \forall x\in\mathcal{X},\ t\in\mathcal{T}\ . \tag{5.20}$$

Each of these conditions imposes a constraint on the set of models θ', lowering the dimensionality of the space on which any prior distribution may be defined. For

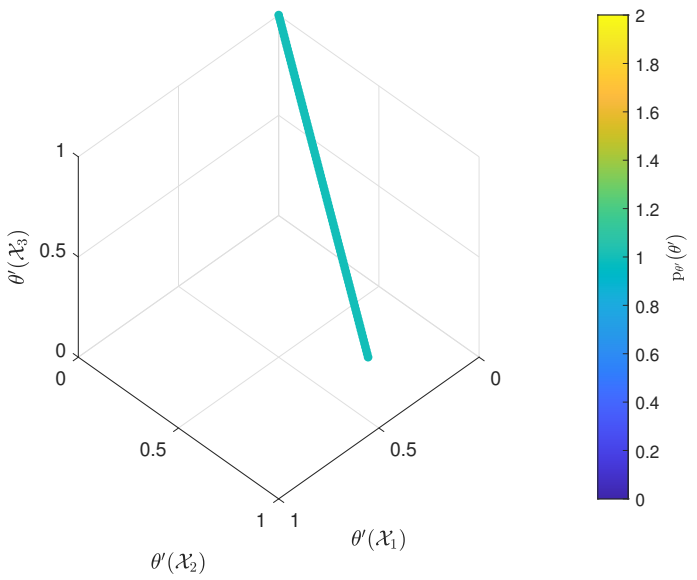

Figure 5.3: A marginal model prior for $|\mathcal{T}| = 2$

example, consider the case $|\mathcal{X}| = 3$ and $|\mathcal{T}| = 2$; Figure 5.3 illustrates a limited support prior distribution that would satisfy the requirements for such a sufficient statistic. This demonstrates that in Bayesian learning, limited dimensional prior distributions can lead to the existence of data sufficient statistics. Furthermore, the degree of information compression provided by the statistic is inherently linked to the dimensionality of the prior's support. As such, the thesis will use limited-support prior distributions to form Bayesian predictive distributions that are suitable for human recognition applications. A primary research focus will be on how the user's intuition of sensible transformations T drives the selection of the prior distribution support.

Another insight into human recognition performance is that we ourselves have defined the labels and their statistical relationship to our sensory observations, creating a joint distribution θ that leads to low clairvoyant risk $\mathcal{R}^*_\Theta(\theta)$. Casting this intuition into the learning framework used for this thesis, it is expected that the true predictive models $\tilde{\theta}(x)$ will be highly definitive, or in a sense, "sparse". Different metrics will be used to assess how certain a predictive model is. An obvious measure from information theory is conditional entropy (Figure 5.4), which could be restricted to low values.

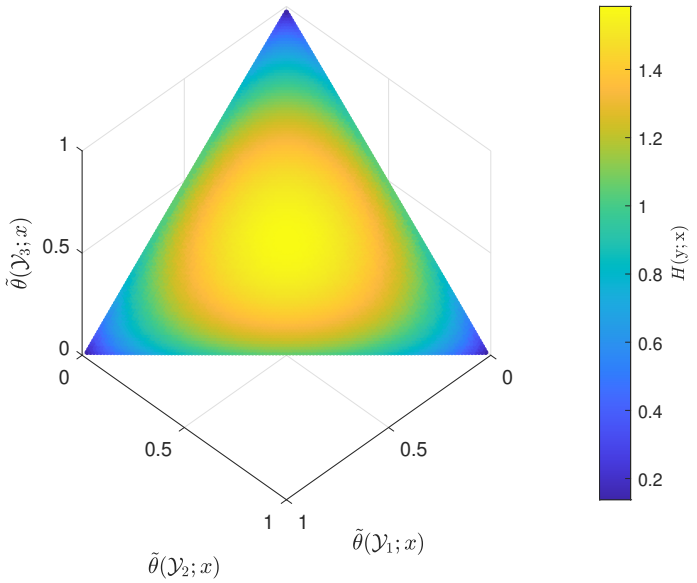

Figure 5.4: Conditional entropy for $\tilde{\theta}(x)$

Another class of possible metrics are the L^p-norms. Of particular interest is the L^∞ norm, $\left\|\tilde{\theta}(x)\right\|_\infty = \max_{y\in\mathcal{Y}} \left|\tilde{\theta}(y;x)\right|$, as it is equivalent to the 0–1 clairvoyant risk after an affine transformation. By providing a lower bound $\left\|\tilde{\theta}(x)\right\|_\infty \geq \rho > 1/|\,\mathcal{Y}\,|$ on the L^∞ norms allowed, the prior support can be limited. While this restriction does limit the prior's support, it does not lower its dimensionality. With this in mind, another restriction that may be used is an upper bound on the L^0-norm, such that

$$\left\|\tilde{\theta}(x)\right\|_0 = \sum_{y\in\mathcal{Y}} \left(1 - \delta\left[\tilde{\theta}(y;x), 0\right]\right) \leq M_{\mathcal{Y}} < |\,\mathcal{Y}\,| \,. \tag{5.21}$$

If a sufficiently low value of $M_{\mathcal{Y}}$ is used, this condition alone may suffice; if it is used in conjunction with the L^∞ condition, the intrinsic dimensionality of the prior support can be kept arbitrarily low while ensuring only models mapping to low clairvoyant risk are considered. Figure 5.6 shows a possible restriction.

The most extreme restriction of this type would be to limit the predictive model $\tilde{\theta}(x)$ prior support to the $|\,\mathcal{Y}\,|$ PMF's satisfying $\left\|\tilde{\theta}(x)\right\|_\infty = 1$. By limiting the support to a countable set, the computational complexity of implementing decision functions

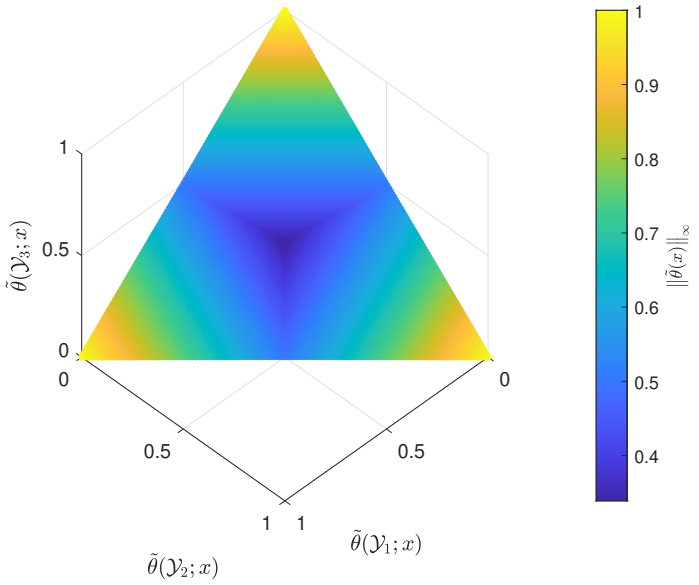

Figure 5.5: L^∞-norm of $\tilde{\theta}(x)$

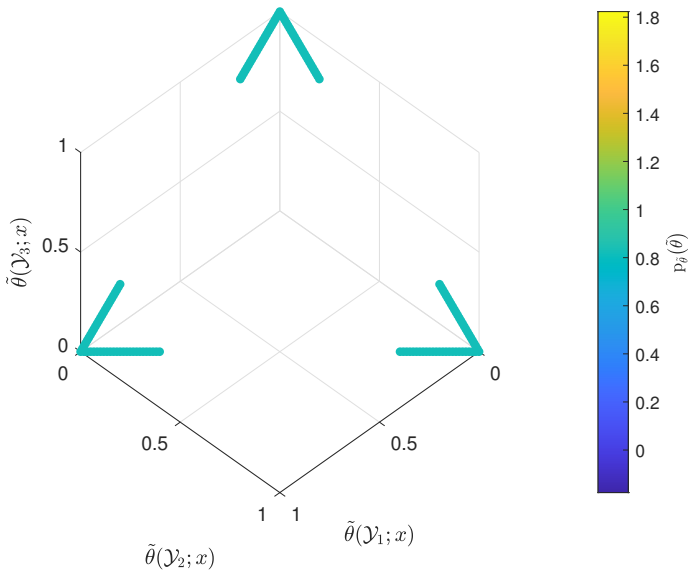

Figure 5.6: Limited support for $\tilde{\theta}(x)$, $\left\|\tilde{\theta}(x)\right\|_0 \leq 2$, $\left\|\tilde{\theta}(x)\right\|_\infty \geq 0.8$

based on such a prior will be minimized. While such functions have the potential to operate effectively on simpler problems, they will also undoubtedly be subject to the lack of adaptability inherent in learning with extremely informative prior distributions.

Chapter 6: Preliminary Results

The work to date has focused on providing optimal decision functions derived from Dirichlet prior distributions. Risk analysis of the resultant Bayesian estimators and classifiers for regression and classification, respectively, has been performed. Results accepted for publication include *Predictive Distribution Estimation for Bayesian Machine Learning using a Dirichlet Prior*, presented at the 53rd Annual Asilomar Conference on Signals, Systems, and Computers and *Bayesian Learning for Classification using a Uniform Dirichlet Prior*, presented at the 7th IEEE Global Conference on Signal and Information Processing.

This section details optimal decision functions when the sets $\mathcal{Y}$ and $\mathcal{X}$ have a finite number of elements and the model θ is characterized by a Dirichlet distribution. The data marginal distribution $\mathrm{P}_{\bar{\mathrm{n}}}$ and the Bayesian predictive PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ are provided and used for regression and classification applications via the squared-error and 0–1 loss functions, respectively.

6.1 Probability Distributions

6.1.1 Training Set PMF, $\mathrm{P}_{\bar{\mathrm{n}}}$

Next, the distribution of the sufficient statistic $\bar{\mathrm{n}}$ will be represented. As a Dirichlet distribution characterizes the parameters of the multinomial distribution $\mathrm{P}_{\mathrm{D}\,|\,\theta}$, the marginal PMF of $\bar{\mathrm{n}}$ is a Dirichlet-Multinomial distribution [**?**] parameterized by α,

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) &= \mathcal{M}(\bar{n})\beta(\alpha)^{-1}\beta(\alpha+\bar{n}) \\ &= \mathrm{DM}\left(\bar{n}; N, \alpha\right). \end{aligned} \tag{6.1}$$

The first and second joint moments of $\bar{\mathrm{n}}$ are

$$\mu_{\bar{\mathrm{n}}} = N\frac{\alpha}{\alpha_0} = N\mu_\theta \tag{6.2}$$

and

$$\begin{aligned}&\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}(y,x)\bar{\mathrm{n}}(y',x')\right] \\ &\quad= \frac{N}{\alpha_0+1}\Big((\alpha_0+N)\mu_\theta(y,x)\delta[y,y']\delta[x,x'] + \alpha_0(N-1)\mu_\theta(y,x)\mu_\theta(y',x')\Big)\,.\end{aligned} \tag{6.3}$$

The covariance function is

$$\begin{aligned}\Sigma_{\bar{\mathrm{n}}}(y,x,y',x') &= \frac{N(\alpha_0+N)}{\alpha_0+1}\Big(\mu_\theta(y,x)\delta[y,y']\delta[x,x'] - \mu_\theta(y,x)\mu_\theta(y',x')\Big) \\ &= N(\alpha_0+N)\Sigma_\theta(y,x,y',x')\,.\end{aligned} \tag{6.4}$$

Again, the data PMF's for minimal and maximal concentration α_0 are relevant. For $\alpha_0 \to \infty$, the model PDF p_θ concentrates at its mean and thus $\bar{\mathrm{n}}$ is characterized by a multinomial distribution,

$$\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) \to \mathcal{M}(\bar{n}) \prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\left(\frac{\alpha(y,x)}{\alpha_0}\right)^{\bar{n}(y,x)} \tag{6.5}$$

Conversely, for $\alpha_0 \to 0$, the PMF tends toward

$$\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) \to \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}\frac{\alpha(y,x)}{\alpha_0}\delta\Big[\bar{n}, N\delta[\cdot,y]\delta[\cdot,x]\Big]\,. \tag{6.6}$$

Figure 6.1 displays the distribution of $\bar{\mathrm{n}}$ for $N = 10$ and different model concentrations α_0. Observe that for large α_0, the distribution approaches a multinomial distribution $\bar{\mathrm{n}} \sim \mathrm{Multi}(N, \alpha/\alpha_0)$. Figure 6.2 shows how a specific model prior influences the data PMF differently for different N. Observe that as the number of training samples increases, the PMF $\mathrm{P}_{\bar{\mathrm{n}}}$ tends toward $\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) \approx N^{1-|\mathcal{Y}||\mathcal{X}|}\,\mathrm{p}_\theta(\bar{n}/N)$; this can be

proven using Gautschi's inequality [**?**].

Uniform Prior For the uniform distribution, $\alpha(y, x) = 1$,

$$\mathrm{P}_{\bar{\mathrm{n}}} = |\bar{\mathcal{N}}|^{-1} = \mathcal{M}\left(\{N, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1\}\right)^{-1}. \tag{6.7}$$

The distribution of $\bar{\mathrm{n}}$ is uniform over the set $\bar{\mathcal{N}}$.

6.1.1.1 Marginal and Conditional Distributions

It is also useful to express the marginal and conditional distributions for the training data given the Dirichlet prior. Recall that $\mathrm{n}'|\theta \sim \mathrm{Multi}(N, \theta')$; by the aggregation property of Dirichlet-Multinomial functions [**?**], the random process is distributed as $\mathrm{n}' \sim \mathrm{DM}(N, \alpha')$.

Also of interest is the distribution of $\bar{\mathrm{n}}$ conditioned on its aggregation n'. Using the Dirichlet-Multinomial properties presented in Appendix A.3, it can be shown that

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}}|\,\mathrm{n}'}(\bar{n}|n') &= \prod_{x\in\mathcal{X}} \left[\mathcal{M}\left(\bar{n}(\cdot, x)\right)\beta\Big(\alpha(\cdot, x)\Big)^{-1}\beta\Big(\alpha(\cdot, x) + \bar{n}(\cdot, x)\Big)\right] \\ &= \prod_{x\in\mathcal{X}} \mathrm{DM}\left(\bar{n}(\cdot, x); n'(x), \alpha(\cdot, x)\right) \end{aligned} \tag{6.8}$$

over the domain $\left\{\bar{n} \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}} \bar{n}(y, \cdot) = n'\right\}$. Observe that conditioning on the aggregation renders the function segments $\bar{\mathrm{n}}(\cdot, x)$ independent of one another and that they are also Dirichlet-Multinomial, such that $\bar{\mathrm{n}}(\cdot, x)|\,\mathrm{n}'(x) \sim \mathrm{DM}\left(\mathrm{n}'(x), \alpha(\cdot, x)\right)$.

6.1.2 Predictive PMF, $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$

As shown in Equation (3.8), the decision selected by the optimally designed function depends on $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$, the distribution of the unobserved y conditioned on all observable random elements. This PMF will be expressed next.

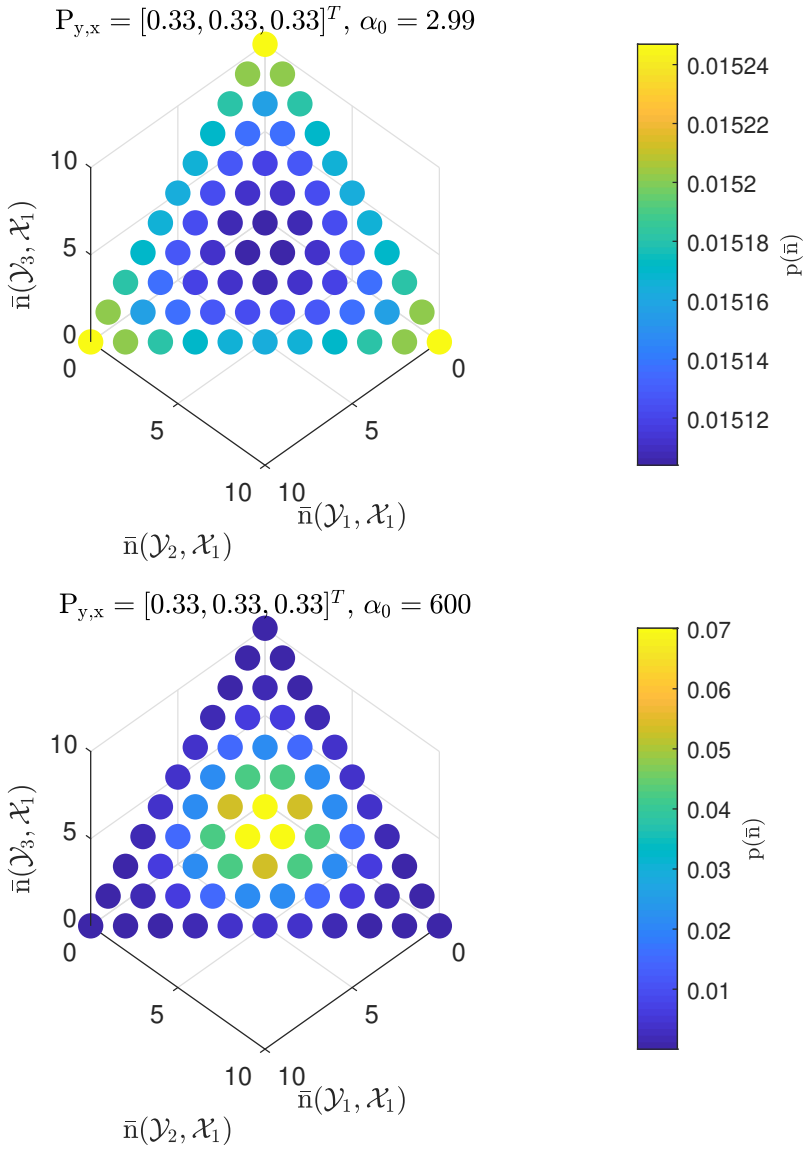

Figure 6.1: $\mathrm{P}(\bar{\mathrm{n}})$ for different prior concentrations α_0

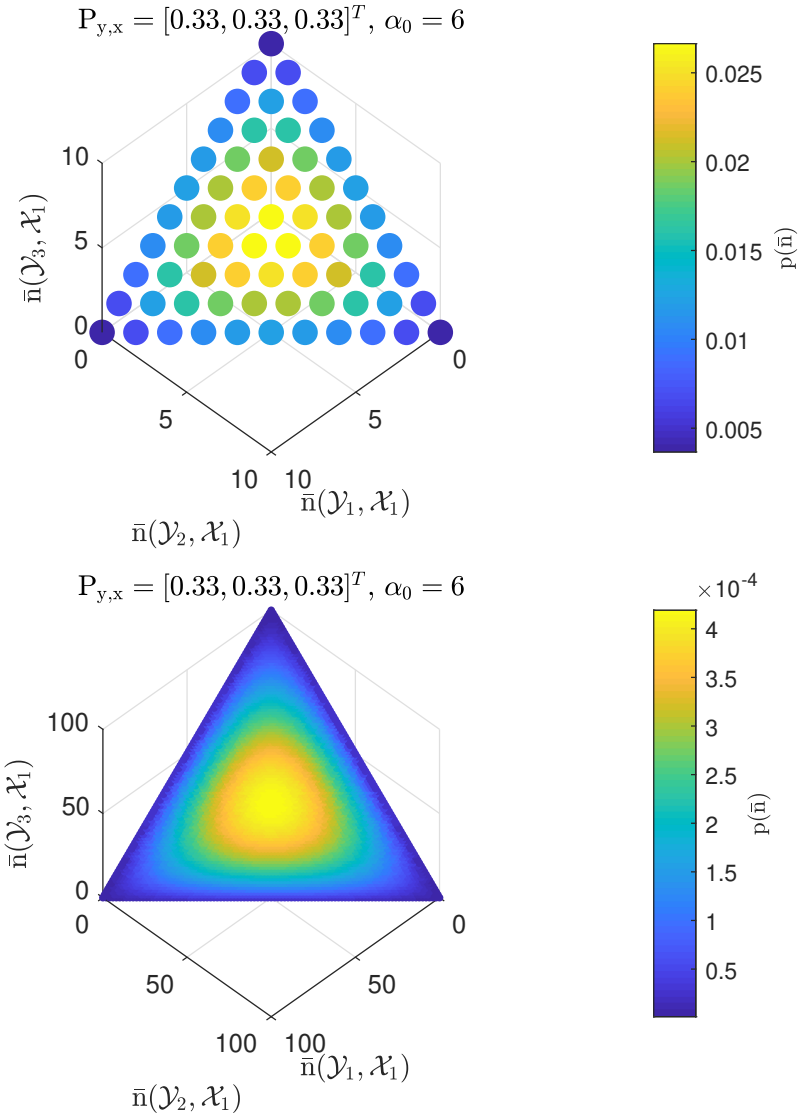

Figure 6.2: P($\bar{\mathrm{n}}$) for different training set sizes N

The Bayesian distributions $\mathrm{P}_{\mathrm{x}\,|\,\bar{\mathrm{n}}}$ and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ can also be found from the posterior distributions $\mathrm{p}_{\theta'|\,\bar{\mathrm{n}}}$ and $\mathrm{p}_{\tilde{\theta}|\,\mathrm{x},\bar{\mathrm{n}}}$, respectively. As the Dirichlet assumption renders θ' and $\tilde{\theta}$ independent, it can be shown that θ' is conditionally independent of $\bar{\mathrm{n}}$ given n'. Furthermore, the Dirichlet distribution $\mathrm{p}_{\theta'}$ is the conjugate prior for $\mathrm{P}_{\mathrm{n}'|\theta'}$. As a result, $\theta'|\,\mathrm{n}' \sim \mathrm{Dir}(\alpha' + \mathrm{n}')$ and

$$\begin{aligned} \mathrm{P}_{\mathrm{x}\,|\,\bar{\mathrm{n}}}(\bar{n}) = \mu_{\theta'|\,\bar{\mathrm{n}}}(\bar{n}) &= \mu_{\theta'|\,\mathrm{n}'}\left(\sum_y \bar{n}(y,\cdot)\right) \\ &= \frac{\alpha' + \sum_y \bar{n}(y,\cdot)}{\alpha_0 + N}\,, \end{aligned} \tag{6.9}$$

where the dependency on $\bar{\mathrm{n}}$ is expressed only through the marginal random process n'.

The posterior $\mathrm{p}_{\tilde{\theta}|\,\mathrm{x},\bar{\mathrm{n}}}$ can be simplified by noting that the independence of θ' and $\tilde{\theta}$ implies $\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\mathrm{x}} = \mathrm{E}_{\tilde{\theta}}\left[\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\tilde{\theta}}\right] = \mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'}$. Consequently, $\tilde{\theta}$ is conditionally independent of x given $\bar{\mathrm{n}}$. Thus, as $\mathrm{p}_{\tilde{\theta}}$ is a conjugate prior for $\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\tilde{\theta}}$ the posterior distribution is

$$\begin{aligned} \mathrm{p}_{\tilde{\theta}|\,\bar{\mathrm{n}},\mathrm{x}}\left(\tilde{\theta}|\bar{n},x\right) = \mathrm{p}_{\tilde{\theta}|\,\bar{\mathrm{n}}}\left(\tilde{\theta}|\bar{n}\right) &= \prod_{x'\in\mathcal{X}} \mathrm{p}_{\tilde{\theta}(x')|\,\bar{\mathrm{n}}(\cdot,x')}\left(\tilde{\theta}(x')|\bar{n}(\cdot,x')\right) \\ &= \prod_{x'\in\mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x');\alpha(\cdot,x') + \bar{n}(\cdot,x')\right) \end{aligned} \tag{6.10}$$

and the distinct model conditional PMF's are independent from one another. Observe that when the conditioning is performed using the sufficient statistic, the independent conditional models $\tilde{\theta}(x)$ are only dependent on their corresponding subset of the empirical PMF, $\bar{\mathrm{n}}(\cdot,x)$.

The concentration parameter for the predictive model posterior increases proportionately with increasing volumes of training data; consequently, as $n'(x) \to \infty$, the posterior converges to $\mathrm{p}_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}} \to \delta\left(\cdot - \bar{\mathrm{n}}(\cdot,x)/\,\mathrm{n}'\right)$. Thus, as more data is collected, the model can be more positively identified and used to formulate minimum risk decisions. Conversely, as $\alpha'(x) \to \infty$, the prior model certainty is stronger, and the posterior tends toward $\mathrm{p}_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}} \to \delta(\cdot - \alpha(\cdot,x)/\alpha'(x))$, independent of the training data.

Figure 6.3 shows the influence of the training data on the model distribution; after conditioning on the training data (via $\bar{\mathrm{n}}$), the PDF concentration shifts away from the models favored by the prior knowledge and towards other models that better account for the observations.

The Bayes predictive PMF can thus be expressed as

$$\begin{aligned} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(x,\bar{n}) &= \mu_{\tilde{\theta}(x)|\,\mathrm{x},\bar{\mathrm{n}}}(x,\bar{n}) = \mu_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}(\cdot,x)}\Big(\bar{n}(\cdot,x)\Big) \\ &\equiv \frac{\alpha(\cdot,x)+\bar{n}(\cdot,x)}{\alpha'(x)+n'(x)} \\ &= \left(\frac{\alpha'(x)}{\alpha'(x)+n'(x)}\right)\frac{\alpha(\cdot,x)}{\alpha'(x)} + \left(\frac{n'(x)}{\alpha'(x)+n'(x)}\right)\frac{\bar{n}(\cdot,x)}{n'(x)}\,. \end{aligned} \tag{6.11}$$

A consequence of the Dirichlet prior is that the predictive PMF for a given value of x depends only on the corresponding training data $\bar{\mathrm{n}}(\cdot,\mathrm{x})$, such that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(x,\bar{n}) = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}(\cdot,\mathrm{x})}\Big(x,\bar{n}(\cdot,x)\Big)$. This is intuitive considering the independence of the conditional models $\tilde{\theta}(x)$ of one another.

The last representation provided for $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ interprets the distribution as a convex combination of two conditional distributions. The first distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$ is independent of the training data and based on the prior knowledge implied via the model PDF parameter; the second distribution is the conditional empirical PMF and depends only on the data.

The weighting factors $\alpha'(x)$ and $n'(x)$ are the concentration of the conditional prior $\tilde{\theta}(x)$ and the number of training samples satisfying $X_n = \mathrm{x}$. As $n'(x)/\alpha'(x) \to 0$, the PMF tends toward the conditional distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$, which depends only on the model parameter α. As $n'(x)/\alpha'(x) \to \infty$, $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ tends towards the empirical conditional distribution.

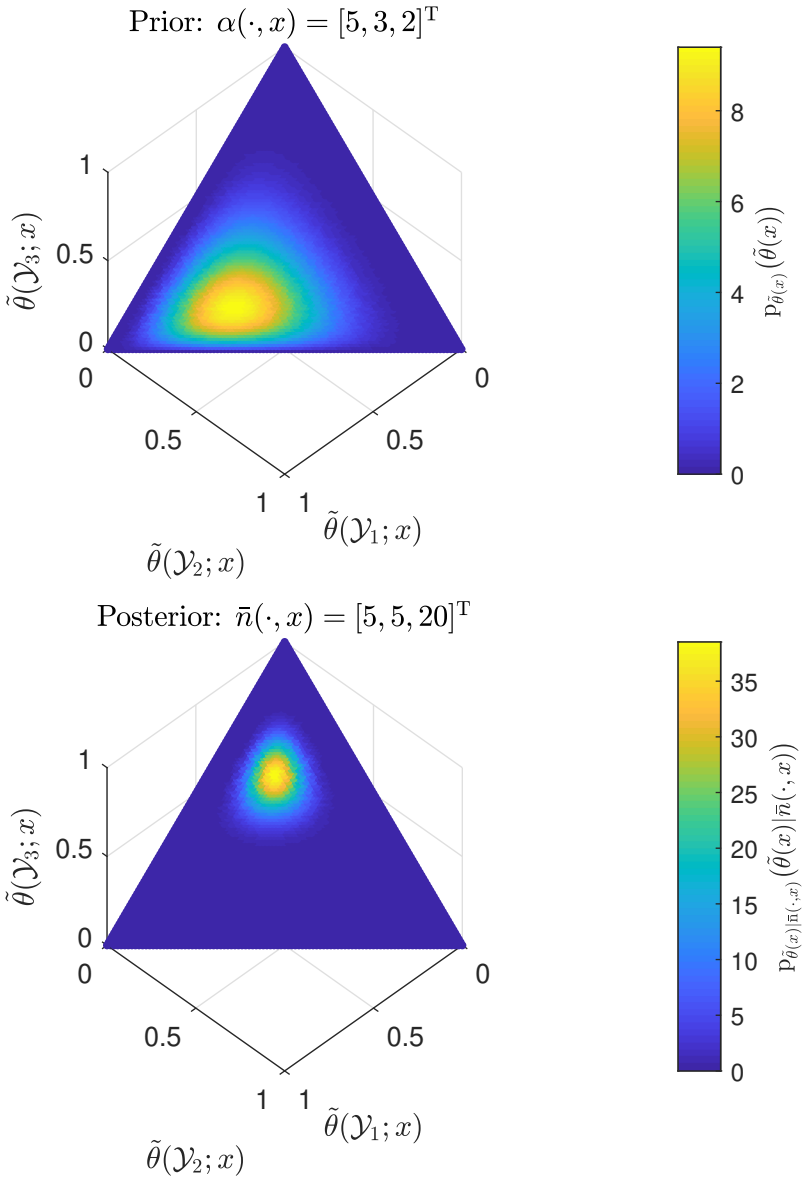

Figure 6.3: Model PDF, prior and posterior

6.2 Model Estimation Perspective

It is informative to treat the distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ as an estimate of the unknown conditional PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \equiv \tilde{\theta}(\mathrm{x})$ and investigate the effects of informative prior knowledge. For a given x and corresponding number of training samples $\mathrm{n}'(\mathrm{x})$, the expected value of the estimate conditioned on the true model θ is

$$\begin{aligned} \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] &= \mathrm{E}_{\bar{\mathrm{n}}(\cdot,\mathrm{x})|\,\mathrm{n}'(\mathrm{x}),\tilde{\theta}(\mathrm{x})}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}(\cdot,\mathrm{x})}\right] \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} + \left(\frac{\mathrm{n}'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)\tilde{\theta}(\mathrm{x})\ , \end{aligned} \tag{6.12}$$

where the properties of a multinomial distribution conditioned on its aggregation have been used. The result is a convex combination of the conditional data-independent distribution $\alpha(\cdot, x)/\alpha'(x)$ and the true conditional distribution $\tilde{\theta}(\mathrm{x})$. The convex coefficients are dependent on the "marginal" values α' and n'; note that as the number of matching training samples $\mathrm{n}'(x)$ increases relative to $\alpha'(x)$, the estimate tends towards the true conditional PMF.

To aid characterization of the estimator, define the random process $\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta) \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \in \mathbb{R}^{\mathcal{Y}}$. For a given x and corresponding number of training samples $\mathrm{n}'(\mathrm{x})$, the bias of the conditional PMF estimate is

$$\begin{aligned} \mathrm{Bias}(\mathrm{x}, \mathrm{n}', \theta) &= \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta)\right] \\ &= \frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\left(\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} - \tilde{\theta}(\mathrm{x})\right) \end{aligned} \tag{6.13}$$

and its covariance function is

$$\begin{aligned}\mathrm{Cov}(y, y'; \mathrm{x}, \mathrm{n}', \theta) &= \mathrm{C}_{\bar{\mathrm{n}} \,|\, \mathrm{n}', \theta} \Big[\mathrm{P}_{\mathrm{y} \,|\, \mathrm{x}, \bar{\mathrm{n}}}(\cdot \,|\, \mathrm{x}, \bar{\mathrm{n}}) \Big](y, y') \\ &= \frac{\Sigma_{\bar{\mathrm{n}}(\cdot,\mathrm{x}) \,|\, \mathrm{n}'(\mathrm{x}), \tilde{\theta}(\mathrm{x})}(y, y')}{\Big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\Big)^2} \\ &= \frac{\mathrm{n}'(\mathrm{x})}{\Big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\Big)^2} \Big(\tilde{\theta}(y; \mathrm{x}) \delta[y, y'] - \tilde{\theta}(y; \mathrm{x}) \tilde{\theta}(y'; \mathrm{x}) \Big) \ ,\end{aligned} \tag{6.14}$$

where the properties of multinomial random processes have been used. Note that the bias is proportionate to the difference between the true conditional model and the data-independent estimate. The scaling factor tends from one to zero as $\mathrm{n}'(x)/\alpha'(x)$ tends from zero to infinity; as such, more informative priors (large $\alpha'(x)$) will lead to PMF estimates that are prone to bias. Conversely, the variance of the PMF estimate tends to zero as $\alpha'(x) \to \infty$.

Combining the estimator bias and variance, the conditional second moments of $\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta)$ are

$$\begin{aligned}\mathcal{E}(y, y'; \mathrm{x}, \mathrm{n}', \theta) &= \mathrm{E}_{\bar{\mathrm{n}} \,|\, \mathrm{n}', \theta} \Big[\Delta(y; \mathrm{x}, \bar{\mathrm{n}}, \theta) \Delta(y'; \mathrm{x}, \bar{\mathrm{n}}, \theta) \Big] \\ &= \mathrm{Bias}(y; \mathrm{x}, \mathrm{n}', \theta) \mathrm{Bias}(y'; \mathrm{x}, \mathrm{n}', \theta) + \mathrm{Cov}(y, y'; \mathrm{x}, \mathrm{n}', \theta) \ .\end{aligned} \tag{6.15}$$

As $\mathrm{n}'(x) \to \infty$, this function tends to zero and thus the underlying model $\tilde{\theta}(x)$ is determined precisely. A more practical case is estimation with a finite volume of training data. Specification of the Dirichlet model prior can be interpreted as providing a distribution estimate $\alpha(\cdot, x)/\alpha'(x)$ and a confidence level $\alpha'(x)$. Higher confidence reduces error due to the variance of the estimator, but increases the error due to bias between the true model and its estimate; low confidence renders the estimate unbiased, but maximizes the estimator variance.

Also of interest, the conditional expectation of $\mathcal{E}(\cdot,\cdot;\mathrm{x},\mathrm{n}',\theta)$ is

$$\begin{aligned}
&\mathrm{E}_{\mathrm{x},\mathrm{n}'\,|\,\theta}\left[\mathcal{E}(y,y';\mathrm{x},\mathrm{n}',\theta)\right] \qquad (6.16)\\
&\quad=\mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta}\left[\left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)^2\right]\left(\frac{\alpha(y,\mathrm{x})}{\alpha'(\mathrm{x})}-\tilde{\theta}(y;\mathrm{x})\right)\left(\frac{\alpha(y',\mathrm{x})}{\alpha'(\mathrm{x})}-\tilde{\theta}(y';\mathrm{x})\right)\right]\\
&\qquad+\mathrm{E}_{\mathrm{x}\,|\,\theta}\left[\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta}\left[\frac{\mathrm{n}'(\mathrm{x})}{\left(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\right)^2}\right]\left(\tilde{\theta}(y;\mathrm{x})\delta[y,y']-\tilde{\theta}(y;\mathrm{x})\tilde{\theta}(y';\mathrm{x})\right)\right]\ .
\end{aligned}$$

To exemplify how the model estimate $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ approximates $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$, consider a scenario with $|\,\mathcal{Y}\,|=10$. The data-independent PMF $\alpha(\cdot,x)/\alpha'(x)$ and true model $\tilde{\theta}(x)$ are shown in Figure 6.4 - note the significant mismatch.

Figures 6.5 and 6.6 show how the bias and variance of the estimate change for different values of $\mathrm{n}'(x)$ and $\alpha'(x)$. The plot markers represent the conditional mean of the estimator, $\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y|\,\mathrm{x},\bar{\mathrm{n}})\right]$; the upper and lower error bars represent the square-root of the expected squared deviation above and below the conditional mean, respectively. Each individual plot heading provides the error $\sqrt{\sum_{y\in\mathcal{Y}}\mathcal{E}(y,y;\mathrm{x},\mathrm{n}',\theta)}$ to assess the quality of the PMF estimate.

Observe that for $\mathrm{n}'(x)=1$, the high variance of the $\alpha'(x)=0.1$ estimate (favoring the empirical PMF) renders it worse than the $\alpha_0=10$ estimate; in fact, the variance is so high that the error exceeds that of the data-independent estimate $\alpha(\cdot,x)/\alpha'(x)$ (Figure 6.4). Conversely, for $\mathrm{n}'(x)=10$, the confidence of the $\alpha'(x)=10$ estimate leads to high bias, and the $\alpha'(x)=0.1$ estimate is superior. For $\mathrm{n}'(x)=100$, both the $\alpha'(x)=0.1$ and $\alpha'(x)=10$ estimates begin converging to the true distribution - this is guaranteed due to the full support of the Dirichlet prior.

6.3 Applications to Common Loss Functions

In this section, the Dirichlet prior is applied to the regression and classification applications. Optimal learners f^* are found, the corresponding minimum Bayes risk

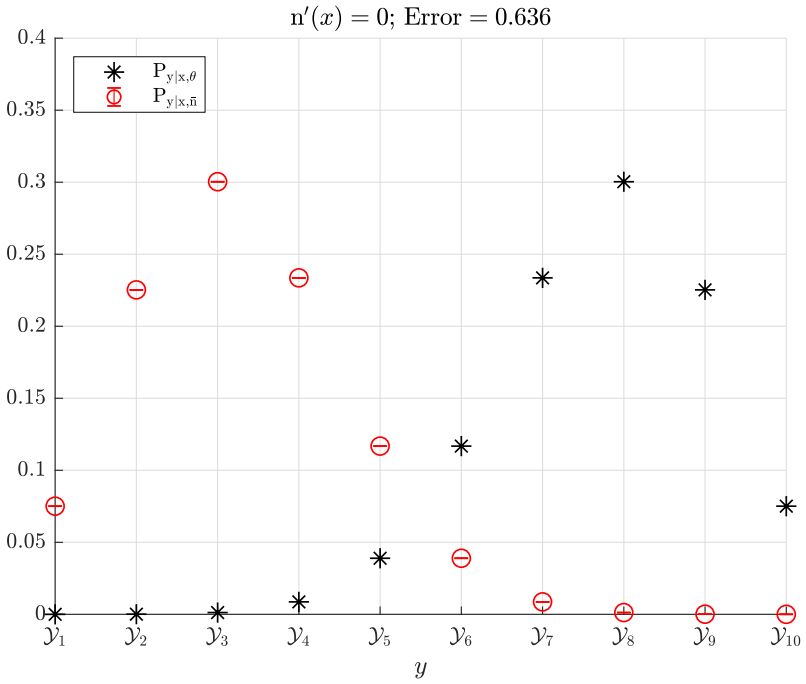

Figure 6.4: Model θ estimate, no training data

$\mathcal{R}^*$ is assessed, and the conditional risk $\mathcal{R}_\Theta(f^*;\theta)$ is analyzed.

It is informative to substitute the Bayes predictive distribution using the Dirichlet prior (6.11) into Equation (3.8), expressing the decision for a given input x and training set D as

$$
\begin{aligned}
f^*(\mathrm{x};\mathrm{D}) &= \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\Big[\mathcal{L}(h,\mathrm{y})\Big] \qquad (6.17)\\
&= \arg\min_{h\in\mathcal{H}} \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+N'(\mathrm{x};\mathrm{D})}\right)\sum_{y\in\mathcal{Y}}\frac{\alpha(y,\mathrm{x})}{\alpha'(\mathrm{x})}\,\mathcal{L}(h,y) + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x})+N'(\mathrm{x};\mathrm{D})}\right)\sum_{y\in\mathcal{Y}}\frac{\bar{N}(y,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})}\,\mathcal{L}(h,y)\\
&= \arg\min_{h\in\mathcal{H}} \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+N'(\mathrm{x};\mathrm{D})}\right)\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}\Big[\mathcal{L}(h,\mathrm{y})\Big] + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x})+N'(\mathrm{x};\mathrm{D})}\right)\frac{\sum_{n=1}^{N}\delta\Big[\mathrm{x},\mathrm{X}_n\Big]\,\mathcal{L}\left(h,\mathrm{Y}_n\right)}{\sum_{n=1}^{N}\delta\Big[\mathrm{x},\mathrm{X}_n\Big]}\,.
\end{aligned}
$$

The metric to be minimized can be represented as a convex combination of two expected losses. The first expected loss is evaluated with respect to the conditional distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$, which reflects the prior knowledge imparted by the model parameter α. The second term is a conditional empirical risk, or the average

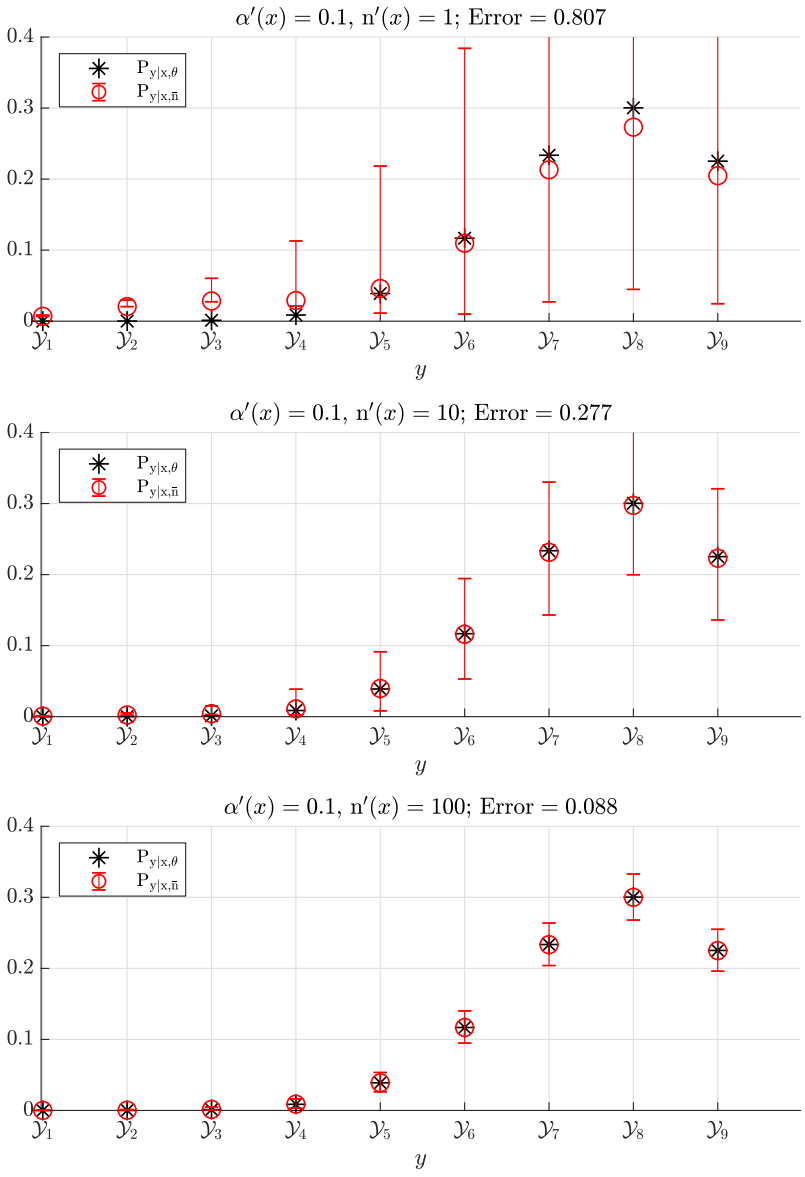

Figure 6.5: Model θ estimates, $\alpha_0 = 0.1$

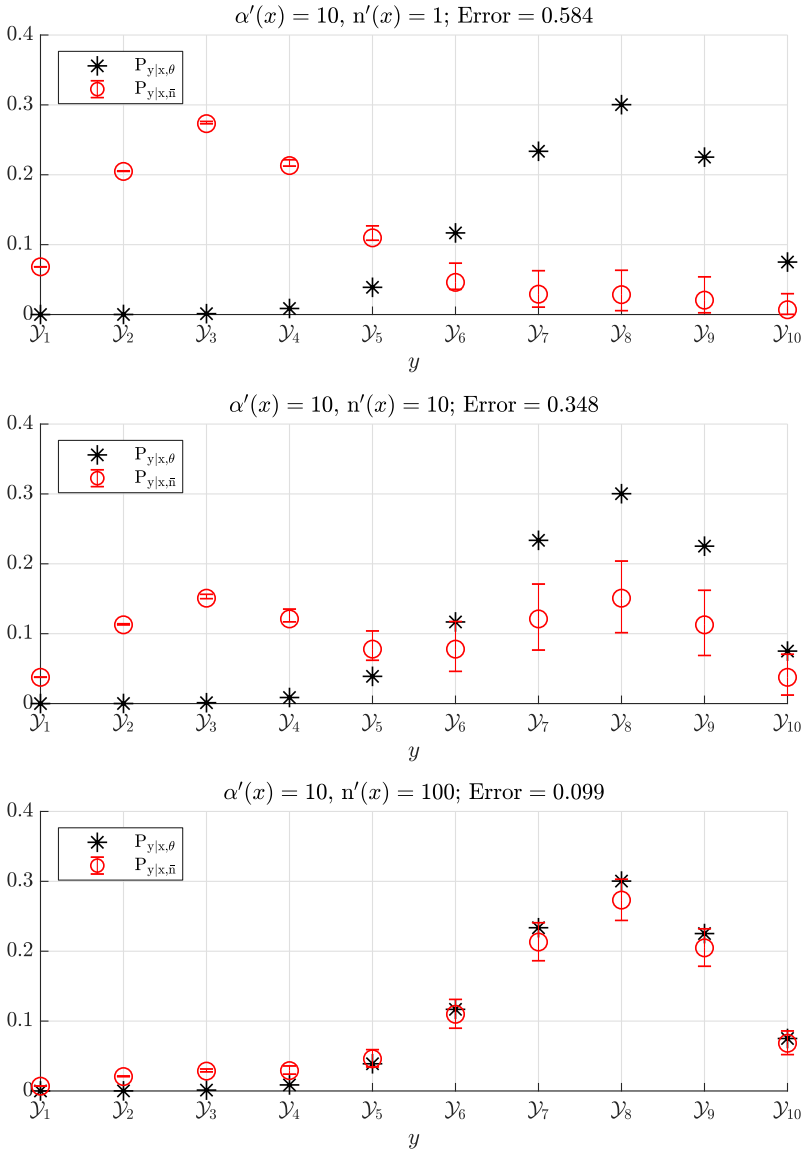

Figure 6.6: Model θ estimates, $\alpha_0 = 10$

loss among samples Y_n whose corresponding values X_n match the observed value x. The convex weights are inherited from the conditional distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$; thus, for a given observation x, the model prior parameter $\alpha'(\mathrm{x})$ and the number of matching training samples $N'(\mathrm{x};\mathrm{D})$ dictate which of the two expectations is emphasized.

6.3.1 Regression: the Squared-Error Loss

The elements of the finite cardinality set $\mathcal{Y}$ are real numbers, such that $\mathcal{Y} \subset \mathbb{R}$. Again, $\mathcal{H} = \mathbb{R} \supset \mathcal{Y}$.

6.3.1.1 Bayesian Estimation

Substituting the Bayes predictive distribution for a Dirichlet prior into (3.14), the optimal Bayesian estimate is

$$\begin{aligned} f^*(\mathrm{x};\mathrm{D}) &= \mu_{\mathrm{y}\,|\,\mathrm{x,D}} \qquad (6.18)\\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right) \sum_{y\in\mathcal{Y}} y \frac{\alpha(y,\mathrm{x})}{\alpha'(\mathrm{x})} + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right) \sum_{y\in\mathcal{Y}} y \frac{\bar{N}(y,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right) \mu_{\mathrm{y}\,|\,\mathrm{x}} + \left(\frac{N'(\mathrm{x};\mathrm{D})}{\alpha'(\mathrm{x}) + N'(\mathrm{x};\mathrm{D})}\right) \frac{\sum_{n=1}^{N} \delta\Big[\mathrm{x}, \mathrm{X}_n\Big]\, \mathrm{Y}_n}{N'(\mathrm{x};\mathrm{D})} \,. \end{aligned}$$

The optimal estimate is interpreted as a convex combination of two separate estimates - the expected value of y conditioned on the observed x and the mean of the training values Y_n which have a value X_n matching the observed value x. The weighting factors are the same as those of $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$; thus, stronger prior information (larger $\alpha'(\mathrm{x})$) provides more weight to the estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}}$, and more voluminous training data puts emphasis on the empirical conditional mean.

The minimum Bayes squared-error is $\mathcal{R}^* = \mathrm{E}_{\mathrm{x,D}}\Big[\Sigma_{\mathrm{y}\,|\,\mathrm{x,D}}\Big]$. Using the sufficient statistic $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D})$, the minimum risk can also be represented as $\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\Big[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\Big]$; as such,

the expectations are performed over $\bar{\mathrm{n}}$. Decompose the conditional variance as

$$\Sigma_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} = \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} \tag{6.19}$$

and assess the expected values of these terms separately using distributions derived from the Dirichlet prior. The first term is simply

$$\begin{aligned} \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\Big[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2]\Big] = \mathrm{E}_{\mathrm{y}}[\mathrm{y}^2] &= \sum_{y\in\mathcal{Y}} y^2 \left(\sum_{x\in\mathcal{X}} \frac{\alpha(y,x)}{\alpha_0}\right) \\ &= \mathrm{E}_{\mathrm{x}}\Big[\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2]\Big] = \sum_{x\in\mathcal{X}} \frac{\alpha'(x)}{\alpha_0} \sum_{y\in\mathcal{Y}} y^2 \frac{\alpha(y,x)}{\alpha'(x)}\ , \end{aligned} \tag{6.20}$$

where the different functions of α are represented by the PMF's of y and x. Next, find

$$\begin{aligned} \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{x}}\left[\frac{\Big(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{y\in\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\Big)^2}{\alpha'(\mathrm{x})\Big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\Big)^2}\right]\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\bar{\mathrm{n}}}\left[\frac{\alpha_0\Big(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{y\in\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\Big)^2}{\alpha'(\mathrm{x})\Big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\Big)(\alpha_0+N)}\right]\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\mathrm{E}_{\mathrm{n}'}\left[\frac{\alpha_0\,\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'}\left[\Big(\alpha'(\mathrm{x})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \sum_{y\in\mathcal{Y}} y\bar{\mathrm{n}}(y,\mathrm{x})\Big)^2\right]}{\alpha'(\mathrm{x})\Big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})\Big)(\alpha_0+N)}\right]\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\,\mathrm{E}_{\mathrm{n}'}\Big[\,\mathrm{n}'(\mathrm{x})\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2] + \Big(\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x}) + 1\Big)\alpha'(\mathrm{x})\mu^2_{\mathrm{y}\,|\,\mathrm{x}}\Big]}{\alpha'(\mathrm{x})\Big(\alpha'(\mathrm{x}) + 1\Big)(\alpha_0+N)}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{N\,\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x}}[\mathrm{y}^2] + \Big(\alpha_0\alpha'(\mathrm{x}) + N\alpha'(\mathrm{x}) + \alpha_0\Big)\mu^2_{\mathrm{y}\,|\,\mathrm{x}}}{\Big(\alpha'(\mathrm{x}) + 1\Big)(\alpha_0+N)}\right]\ . \end{aligned} \tag{6.21}$$

The above formulation exploits the statistical characterization of the aggregation, $\mathrm{n}' \sim \mathrm{DM}(N,\alpha')$; also used is the property that the Dirichlet-Multinomial random process $\bar{\mathrm{n}}$ conditioned on its aggregation n' yields independent conditional DM functions $\bar{\mathrm{n}}(\cdot,x)|\,\mathrm{n}'(x) \sim \mathrm{DM}\Big(\,\mathrm{n}'(x), \alpha(\cdot,x)\Big)$.

Finally, combine the two formulas to represent the minimum Bayes risk,

$$\begin{aligned}\mathcal{R}^* &= \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}[\mathrm{y}^2] - \mu^2_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\alpha_0\alpha'(\mathrm{x}) + N\alpha'(\mathrm{x}) + \alpha_0}{\Big(\alpha'(\mathrm{x})+1\Big)(\alpha_0+N)}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] \\ &= \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + (\alpha_0+N)^{-1}}{\mathrm{P}_{\mathrm{x}}(\mathrm{x}) + \alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right] .\end{aligned} \tag{6.22}$$

The minimum risk is the expected value of the scaled conditional variance with respect to $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$. The expectation is taken with respect to the prior marginal distribution $\mathrm{P}_{\mathrm{x}} = \alpha'/\alpha_0$.

The scaling factor for each term $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ depends on the marginal P_{x}, as well as on the prior concentration α_0 and the number of training samples N. Observe that with no training data ($N = 0$), the scaling factor becomes unity and the risk is $\mathcal{R}^* = \mathrm{E}_{\mathrm{x}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$. Conversely, as $N \to \infty$, the Bayes risk is $\mathcal{R}^* \to \mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x})}{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+\alpha_0^{-1}}\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$; note that this is equivalent to the irreducible risk $\mathrm{E}_{\theta}\left[\mathcal{R}^*_{\Theta}(\theta)\right] = \mathrm{E}_{\mathrm{x},\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right]$. Also, as the model concentration parameter $\alpha_0 \to 0$, the risk tends to zero (for $N > 0$); as $\alpha_0 \to \infty$, the risk tends toward $\mathrm{E}_{\mathrm{x}}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x}}\right]$.

To illustrate these trends, explicitly define the sets $\mathcal{Y} = \{i/M_{\mathrm{y}} : i = 0,\ldots,M_{\mathrm{y}}-1\}$ and $\mathcal{X} = \{i/M_{\mathrm{x}} : i = 0,\ldots,M_{\mathrm{x}}-1\}$. Assume that the conditional variance $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ is independent of x; in this case, the squared-error becomes the conditional variance scaled by a factor dependent on the marginal distribution P_{x}, such that $\mathcal{R}^* = \Sigma_{\mathrm{y}\,|\,\mathrm{x}}\,\mathrm{E}_{\mathrm{x}}\left[\frac{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+(\alpha_0+N)^{-1}}{\mathrm{P}_{\mathrm{x}}(\mathrm{x})+\alpha_0^{-1}}\right]$. Figures 6.7 and 6.8 display how the risk changes with N and α_0 when $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$ and P_{x} are fixed.

It may not seem intuitive for the risk to decrease when α_0 is smaller – the variance of the model θ increases and the prior knowledge is less definitive. This is a result of the Dirichlet PDF weight shifting towards the $|\,\mathcal{Y}\,||\,\mathcal{X}\,|$ models which have ℓ_0 norms satisfying $\|\theta\|_0 = 1$. Although these PMF's are maximally separated (and uncorrelated), they all have zero variance. The optimal learner (6.18) will simply use

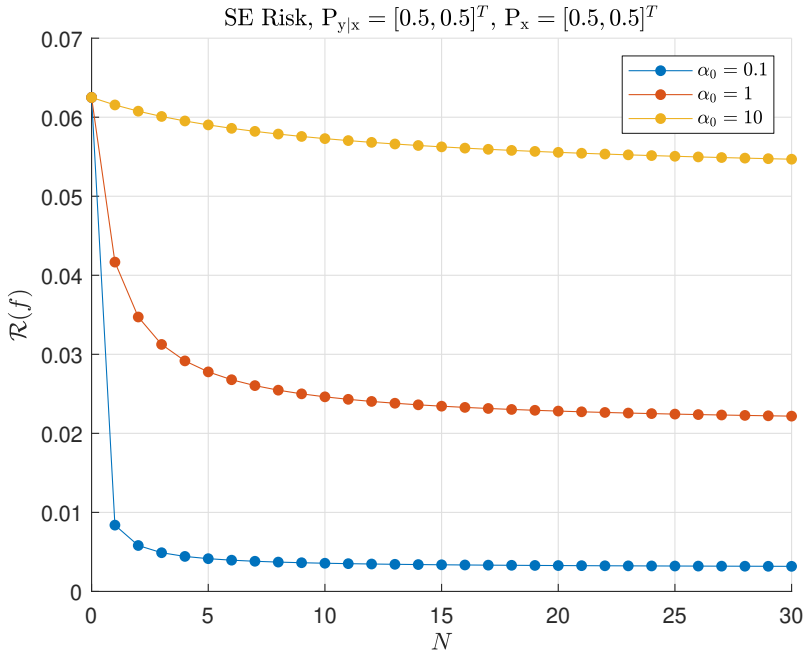

Figure 6.7: Minimum SE Risk for different training set sizes N

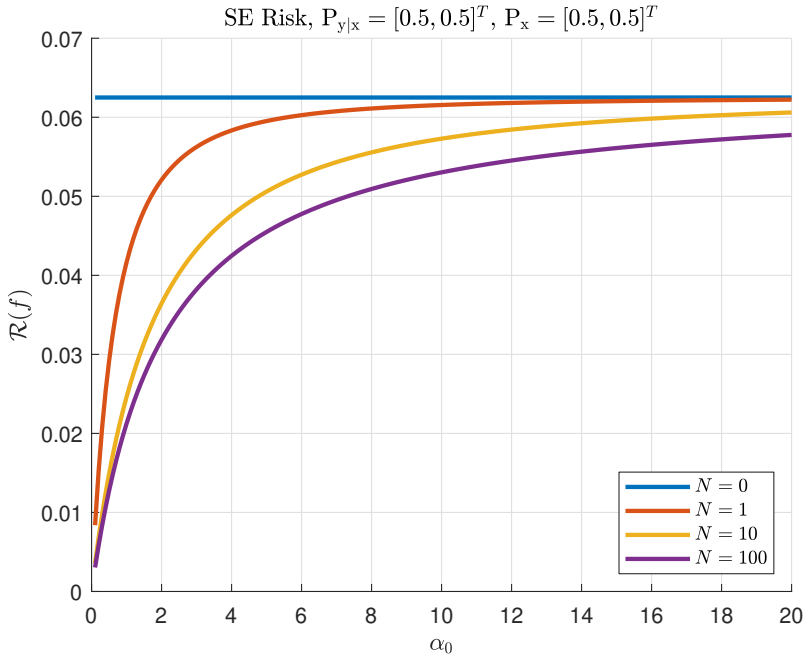

Figure 6.8: Minimum SE Risk for different prior concentrations α_0

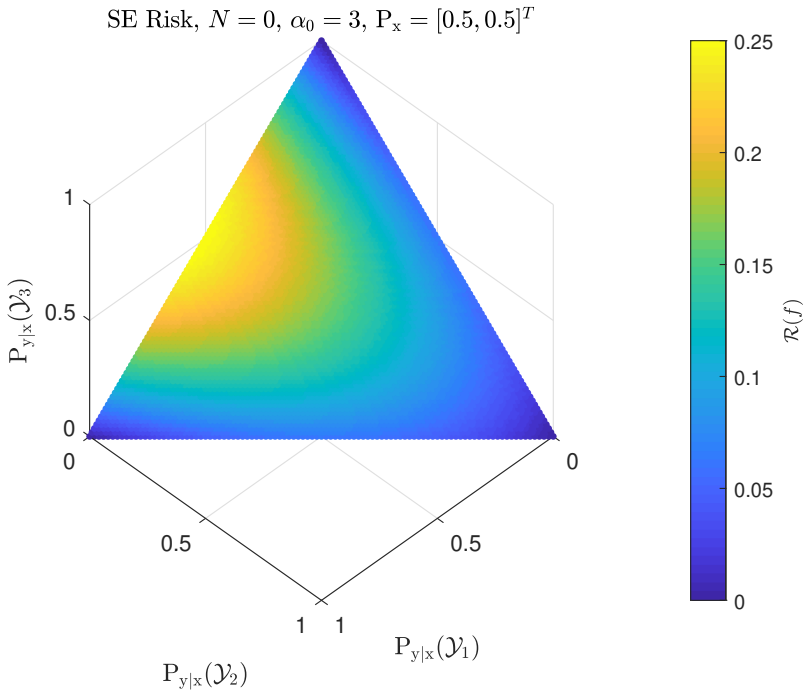

Figure 6.9: Minimum SE Risk for different PMF's $\mathrm{P_{y\,|\,x}}$

the empirical distribution supplied via the training data - this allows exact identification of θ with a single training pair.

It is also informative to visualize how the minimum squared-error changes for a fixed volume of training data N and a fixed prior concentration α_0. First, consider how the risk changes with the conditional PMF $\mathrm{P_{y\,|\,x}}$. Figure 6.9 demonstrates how the squared-error tends towards zero for PMFs that have ℓ_0-norm equal to one.

Next, consider the effect of the marginal distribution $\mathrm{P_x}$. Figure 6.10 demonstrates how the risk changes with this marginal PMF. Observe that the risk is maximal at the distributions satisfying $\| \mathrm{P_x} \|_0 = 1$; the scaling factor for the conditional variance $\Sigma_{\mathrm{y\,|\,x}}$ becomes $\frac{1+(\alpha_0+N)^{-1}}{1+\alpha_0^{-1}}$. Conversely, for $\mathrm{P_x} = 1/|\,\mathcal{X}\,|$ the scaling factor becomes $\frac{|\,\mathcal{X}\,|^{-1}+(\alpha_0+N)^{-1}}{|\,\mathcal{X}\,|^{-1}+\alpha_0^{-1}}$ and the risk is minimal. Figures 6.11 and 6.12 show how different marginals $\mathrm{P_x}$ affect the risk as a function of N and α_0, respectively.

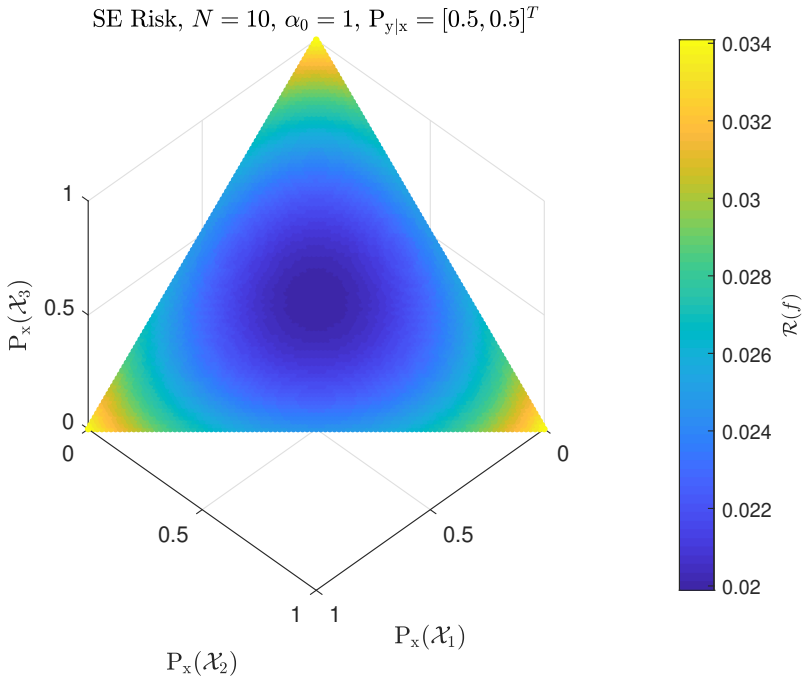

Figure 6.10: Minimum SE Risk for different PMF's P_{x}

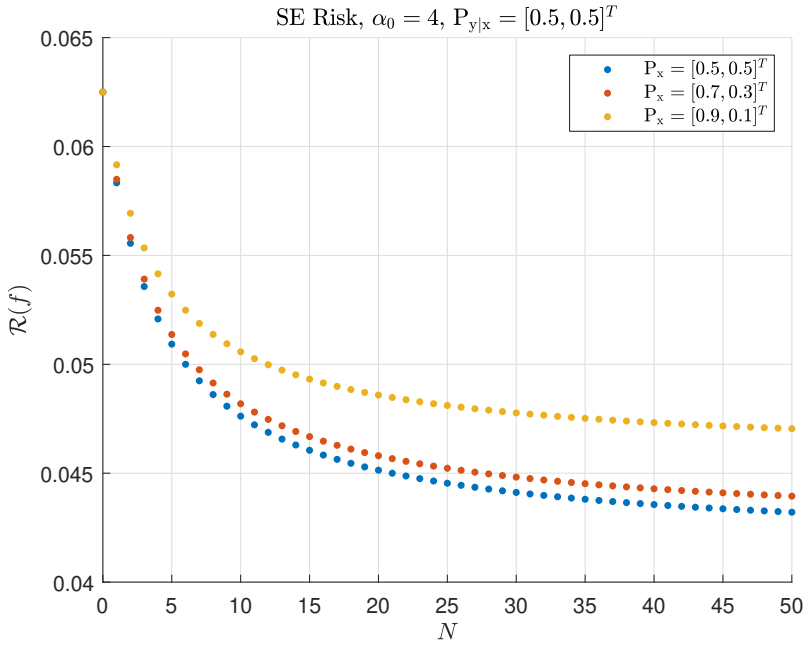

Figure 6.11: Minimum SE Risk for different training set sizes N

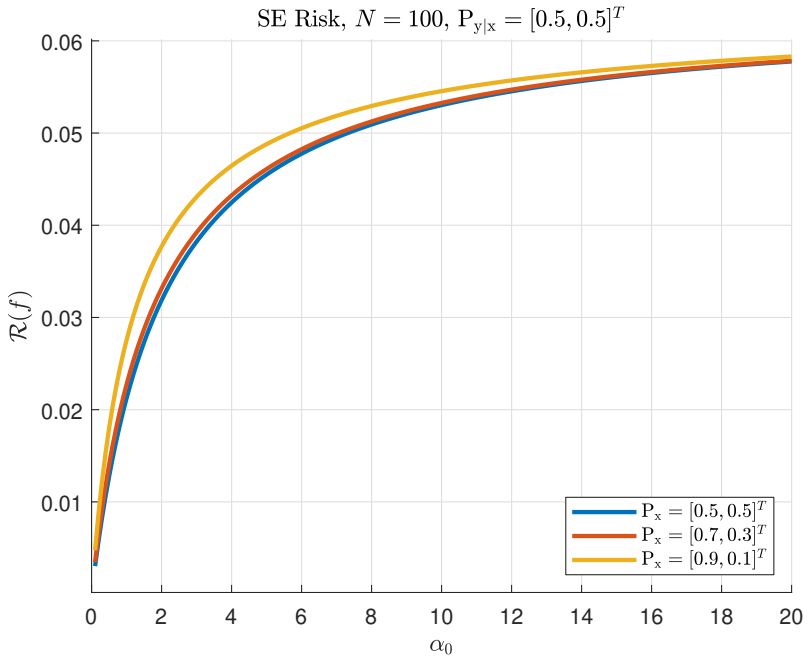

Figure 6.12: Minimum SE Risk for different prior concentrations α_0

Uniform Prior The optimal estimator for a uniform prior is

$$f^*(\mathrm{x};\mathrm{D}) = \left(\frac{|\,\mathcal{Y}\,|}{N'(\mathrm{x};\mathrm{D}) + |\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y + \left(\frac{N'(\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D}) + |\,\mathcal{Y}\,|}\right)\sum_{y\in\mathcal{Y}} y\frac{\bar{N}(y,\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D})} \tag{6.23}$$

$$= \left(\frac{|\,\mathcal{Y}\,|}{N'(\mathrm{x};\mathrm{D}) + |\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y + \left(\frac{N'(\mathrm{x};\mathrm{D})}{N'(\mathrm{x};\mathrm{D}) + |\,\mathcal{Y}\,|}\right)\frac{\sum_{n=1}^{N}\delta\Big[\,\mathrm{x},\mathrm{X}_n\,\Big]\,\mathrm{Y}_n}{N'(\mathrm{x};\mathrm{D})}\,.$$

Now, the model prior contribution to the weighting factors depends on the cardinality $|\,\mathcal{Y}\,|$, and the prior expectation is simply the average of the elements of $\mathcal{Y}$.

For the uniform model prior, the risk reduces to

$$\mathcal{R}^* = \frac{1 + \Big(N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|\Big)^{-1}}{1 + |\,\mathcal{Y}\,|^{-1}}\left[\left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y^2\right) - \left(\frac{1}{|\,\mathcal{Y}\,|}\sum_{y\in\mathcal{Y}} y\right)^2\right]\,. \tag{6.24}$$

Since all possible values of x are equally probable and the conditional probability $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}}$ is uniform and independent of x, the risk simply becomes the variance of the set $\mathcal{Y}$ scaled by a factor dependent on $|\,\mathcal{Y}\,|$ and on $N/|\,\mathcal{X}\,|$. Without training data ($N = 0$),

the scaling is unity; as $N/|\,\mathcal{X}\,| \to \infty$, the scaling factor is $\left(1 + |\,\mathcal{Y}\,|^{-1}\right)^{-1}$.

To visualize the performance, use the explicit sets $\mathcal{Y}$ and $\mathcal{X}$ defined earlier. The conditional variance becomes

$$\Sigma_{\mathrm{y}\,|\,\mathrm{x}} = \frac{|\,\mathcal{Y}\,|^2 - 1}{12|\,\mathcal{Y}\,|^2} = \frac{1 - |\,\mathcal{Y}\,|^{-2}}{12} \tag{6.25}$$

and the minimum risk is expressed as

$$\begin{aligned} \mathcal{R}^* &= \frac{\left(1 - |\,\mathcal{Y}\,|^{-1}\right)\left(1 + \left(N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|\right)^{-1}\right)}{12} \\ &= \left(\frac{|\,\mathcal{Y}\,|}{N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|}\right)\frac{1 - |\,\mathcal{Y}\,|^{-2}}{12} + \left(\frac{N/|\,\mathcal{X}\,|}{N/|\,\mathcal{X}\,| + |\,\mathcal{Y}\,|}\right)\frac{1 - |\,\mathcal{Y}\,|^{-1}}{12}\,. \end{aligned} \tag{6.26}$$

Interestingly, the minimum squared-error for the uniform prior can be represented as a convex combination of two separate risk values with weighting factors dependent on $|\,\mathcal{Y}\,|$ and $N/|\,\mathcal{X}\,|$. Thus, for a uniform prior, the risk depends on the number of elements in $\mathcal{Y}$ and the number of training samples "per element of $\mathcal{X}$". Note the relationship of these weighting factors to those of the conditional PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$, which depend on $\alpha'(\mathrm{x})$ and on $N'(\mathrm{x};\mathrm{D})$. For the uniform prior, $\alpha'(\mathrm{x}) = |\,\mathcal{Y}\,|$ and $\mathrm{E}_{\mathrm{D}}\left[N'(\mathrm{D})\right] = N/|\,\mathcal{X}\,|$.

The first risk is the conditional variance $\Sigma_{\mathrm{y}\,|\,\mathrm{x}}$ - this is intuitively satisfying as the corresponding weight becomes unity when $N = 0$. The second risk is the squared-error with infinite training data. Note that the reduction of the risk between these two extreme cases is modest and that the attenuating factor increases towards unity for applications with more possible outcomes. Figure 6.13 illustrates the difference between these cases.

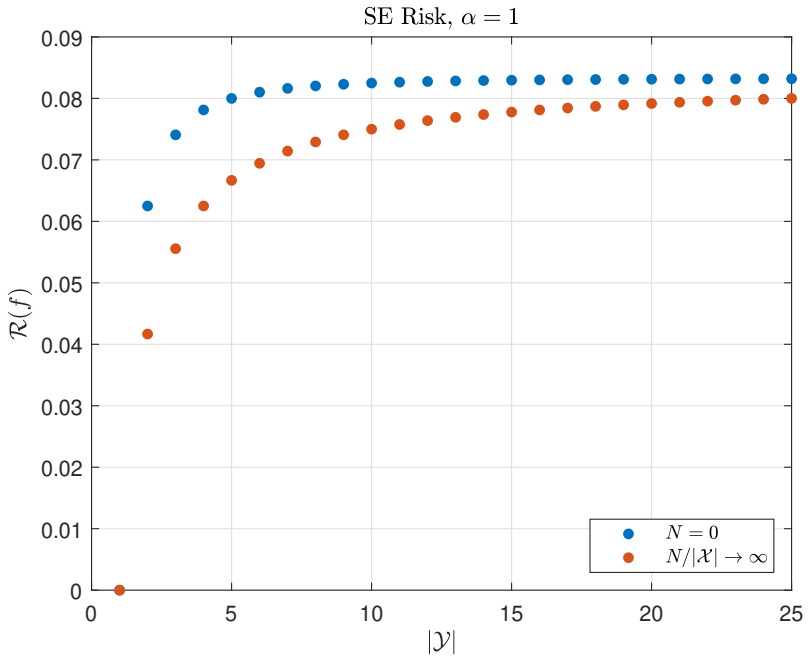

Figure 6.13: Minimum SE Risk, Uniform Prior, zero and infinite training data

6.3.1.2 Conditional Squared-Error for a Dirichlet-based Estimator

Having derived the optimal estimator based on a Dirichlet model prior, it is informative to consider the conditional risk $\mathcal{R}_\Theta(f^*;\theta)$ and analyze how different prior parametrizations α influence the squared-error for different models θ. Starting from the conditional squared-error risk (3.11) and substituting the Bayesian estimator (3.14), the formula simplifies to

$$\begin{aligned}\mathcal{R}_\Theta(f^*;\theta) &= \mathcal{R}^*_\Theta(\theta) + \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\left(f^*(\mathrm{x};\mathrm{D}) - f_\Theta(\mathrm{x};\theta)\right)^2\right] \\ &= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] + \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x,D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right].\end{aligned} \tag{6.27}$$

Defining the excess conditional risk $\mathcal{R}_{\Theta,\mathrm{ex}}(f;\theta) \equiv \mathcal{R}_\Theta(f;\theta) - \mathcal{R}^*_\Theta(\theta)$, the second term above is $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) = \mathrm{E}_{\mathrm{x,D}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x,D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right]$, the average squared bias between the Bayesian predictive mean and the true predictive mean.

Evaluation of the excess risk for an estimator based on the Dirichlet prior will

be performed using the sufficient statistic $\bar{\mathrm{n}}$ in place of the training set D. Using the random process $\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta) \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \in \mathbb{R}^{\mathcal{Y}}$ introduced in 6.2, the term is expressed as

$$\begin{aligned}
\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) &= \mathrm{E}_{\mathrm{x},\mathrm{D}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right] \qquad (6.28)\\
&= \sum_{y\in\mathcal{Y}} y \sum_{y'\in\mathcal{Y}} y'\,\mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}\,|\theta}\left[\Delta(y;\mathrm{x},\bar{\mathrm{n}},\theta)\Delta(y';\mathrm{x},\bar{\mathrm{n}},\theta)\right]\\
&= \sum_{y\in\mathcal{Y}} y \sum_{y'\in\mathcal{Y}} y'\,\mathrm{E}_{\mathrm{x},\mathrm{n}'\,|\theta}\left[\mathcal{E}(y,y';\mathrm{x},\mathrm{n}',\theta)\right]\\
&= \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\,\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left[\frac{\mathrm{n}'(\mathrm{x})}{\left(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\right)^2}\right]\right]\\
&\qquad + \mathrm{E}_{\mathrm{x}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2 \mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left[\left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)^2\right]\right] ,
\end{aligned}$$

where the function $\mathcal{E}$ is defined in (6.15).

The excess conditional risk can thus be represented as the conditional expectation (with respect to $\mathrm{P}_{\mathrm{x}\,|\theta}$) of a sum of two random functions of x. The first function measures the additional variance beyond that of the clairvoyant estimator (i.e. the clairvoyant squared-error); like the clairvoyant risk, it depends on $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}$, the conditional variance of the clairvoyant estimate for a given observation of x. The second function is dependent on the squared bias between the clairvoyant estimate $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}$ and the data-independent estimate $\mu_{\mathrm{y}\,|\,\mathrm{x}}$. This term alone is influenced by the data-independent Bayes predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$.

The two second-order (in terms of y) terms are scaled by factors dependent on the prior concentration $\alpha'(x)$ and on $\theta'(x)$ and N via conditional expectations with respect to $\mathrm{n}'(x)$. Note that by the aggregation property of multinomial distributions, the random variable $\mathrm{n}'(x)|\theta'(x) \sim \mathrm{Bi}\left(N, \theta'(x)\right)$. Closed-forms have not been found for the function expectations of binomial random variables above.

It is informative to consider the trends in the conditional squared-error risk (6.28)

for different volumes of training data N and for different selections of α. First consider how the excess risk changes with the training volume N. For $N = 0$, it is evident that the excess risk is $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{x}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right]$, the expected squared bias between the clairvoyant and data-independent estimators. As N tends to infinity, the binomial distribution controlling the scaling factors concentrates at $\mathrm{n}'(x) \approx N\theta'(x)$; as such, the two expectations of interest tend to zero and thus $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to 0$. This desirable property of the estimator is a consequence of the full support of the Dirichlet prior, ensuring that the model posterior concentrates at the empirical PMF.

Another interesting point regarding the dependency of the excess conditional risk on N is that, depending on the learner parameterization, there may be a local maximum. Consider the trivial case of $|\,\mathcal{X}\,| = 1$ - treating N as a real number, there would be a maximum at

$$N \equiv \alpha'(x)\left(1 - 2\alpha'(x)\frac{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2}{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}\right) . \tag{6.29}$$

Note that as the squared-difference between the prior mean and true mean increases, the maximizing value decreases (even below zero). Thus, the worse the prior estimate, the more likely the excess squared-error will decrease monotonically with N. Conversely, if the prior estimate is accurate, a local maximum may occur and additional training data may (temporarily) compromise the estimator performance. Also consider the effect of prior concentration; informative priors with sufficiently high $\alpha'(x)$ will not have the local maxima.

The excess risk at this potentially non-integral value would be

$$\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{x}\,|\theta}\left[\frac{\frac{1}{\alpha'(\mathrm{x})} - \frac{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2}{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}}{4\left(1 - \frac{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2}{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}\right)}\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] . \tag{6.30}$$

Figures 6.14 and 6.15 exemplify the excess conditional squared-error as a function

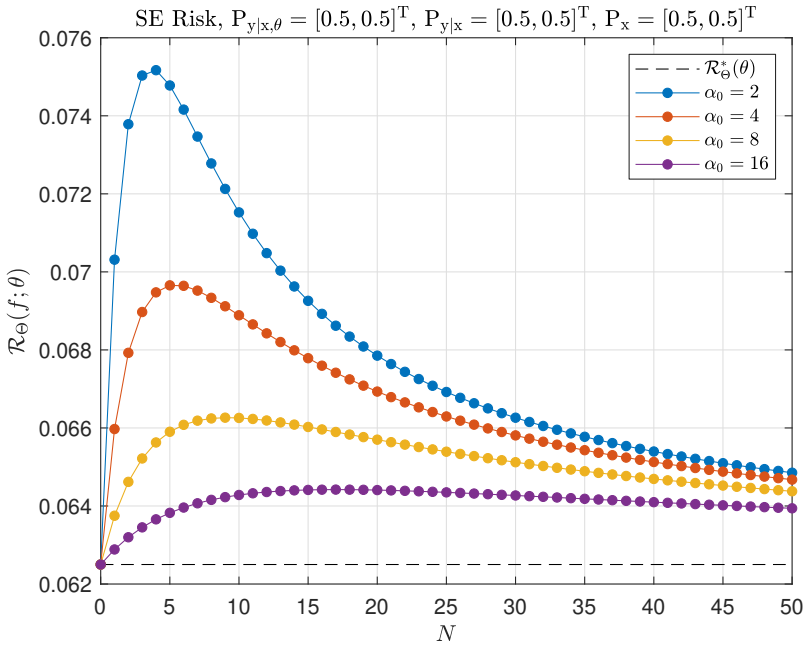

Figure 6.14: Conditional SE Risk versus N, unbiased Dirichlet estimators of varying concentration

of N for estimators based on Dirichlet priors of varying concentration $\alpha'(x)$. The former shows local maxima for an unbiased estimator; note that higher concentration results in superior performance. The latter uses biased estimators and as such, learners based on low concentration achieve lower risk.

Next consider the effects of the Dirichlet prior parameters. The analysis will interpret the Dirichlet parameters as the conditional prior distributions $\alpha(\cdot, x)/\alpha'(x)$ and their concentrations $\alpha'(x)$.

First consider the conditional prior PMF's $\alpha(\cdot, x)/\alpha'(x)$; as shown, they manifest themselves in the risk through the squared estimator bias. It is clear that regardless of how the values $\alpha'(x)$ are chosen, the best selections for these conditional priors must have first moments matching those of the corresponding clairvoyant predictive distributions $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ for each $x \in \mathcal{X}$. Such estimators are unbiased; as a result, the excess conditional risk is equivalent to the first term in (6.28), measuring additional variance due to model uncertainty.

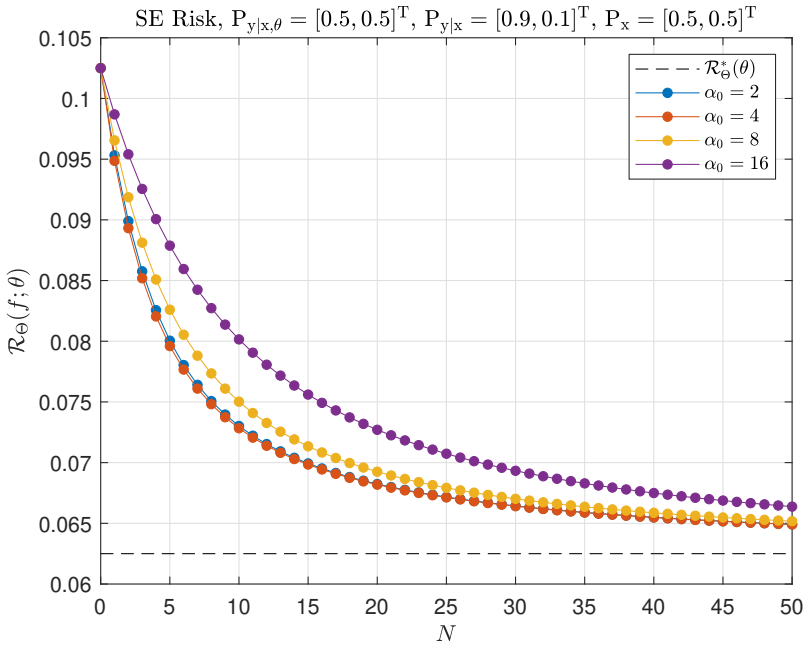

Figure 6.15: Conditional SE Risk versus N, biased Dirichlet estimators of varying concentration

The other user-selected Dirichlet parameters $\alpha'(x)$ are the concentration parameters for the corresponding conditional distributions; they control important bias/variance trade-offs via the two scaling factors in (6.28). First, consider the asymptotic trends.

Consider how the excess risk tends as the priors become maximally concentrated. As the parameters $\alpha'(x) \to \infty$, the excess risk tends to $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{x}\,|\theta}\left[\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right]$, the expected conditional squared-error between the means of the Bayesian predictive PMF and the clairvoyant predictive PMF. This is intuitive given that the estimator tends toward a data-independent solution; analogous to the discussion in Section 6.2, the estimator may be biased, but will have no variance due to the training data statistics.

Conversely, if concentrations $\alpha'(x) \to 0$ are chosen, the Bayesian estimate tends to

the empirical mean, independent of $\alpha(\cdot, \mathrm{x})/\alpha'(\mathrm{x})$, and the excess risk tends to

$$\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \to \mathrm{E}_{\mathrm{x}\,|\theta}\left[\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}\sum_{n=1}^{N}\binom{N}{n}\theta'(\mathrm{x})^n\big(1-\theta'(\mathrm{x})\big)^{N-n}\frac{1}{n}\right] + \mathrm{E}_{\mathrm{x}\,|\theta}\left[\big(1-\theta'(\mathrm{x})\big)^{N}\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2\right]. \tag{6.31}$$

Note that the first term's scaling factor is proportionate to the first inverse moment of a positive binomial random variable [**?**]. The second term's scaling factor tends to $\mathrm{P}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left(0|\theta'(x)\right)$, the probability that no training samples are observed matching the value x. As N increases, this term tends to zero, the risk due to the prior estimate bias decreases, and the excess risk becomes a function of θ only.

Of further interest are the values $\alpha'(x)$ that minimize the excess squared-error for a given prior conditional distribution $\alpha(\cdot, x)/\alpha'(x)$. With the asymptotic values of the excess risk known, all that remains is to determine any local minima. Since the excess risk is a sum of $|\,\mathcal{X}\,|$ terms of identical form, each dependent on its own concentration $\alpha'(x)$, only one component needs to be minimized.

Calculating the first derivative with respect to $\alpha'(x)$, it can be shown that for $N > 0$ and $\theta'(x) > 0$, only one stationary point exists, at

$$\alpha'(x) \equiv \frac{\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta}}{\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^2}. \tag{6.32}$$

Calculation of the second derivative confirms that this value is a local minimum. Furthermore, the excess risk evaluated at these values is

$$\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) = \mathrm{E}_{\mathrm{x}\,|\theta}\left[\mathrm{E}_{\mathrm{n}'(\mathrm{x})|\theta'(\mathrm{x})}\left[\frac{1}{\mathrm{n}'(\mathrm{x})\Sigma^{-1}_{\mathrm{y}\,|\,\mathrm{x},\theta}+\left(\mu_{\mathrm{y}\,|\,\mathrm{x}}-\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}\right)^{-2}}\right]\right], \tag{6.33}$$

which can be easily shown to be less than both the asymptotic values for $\alpha'(x) \to 0$ and $\alpha'(x) \to \infty$. Thus, the concentration values (6.32) provide the minimum excess

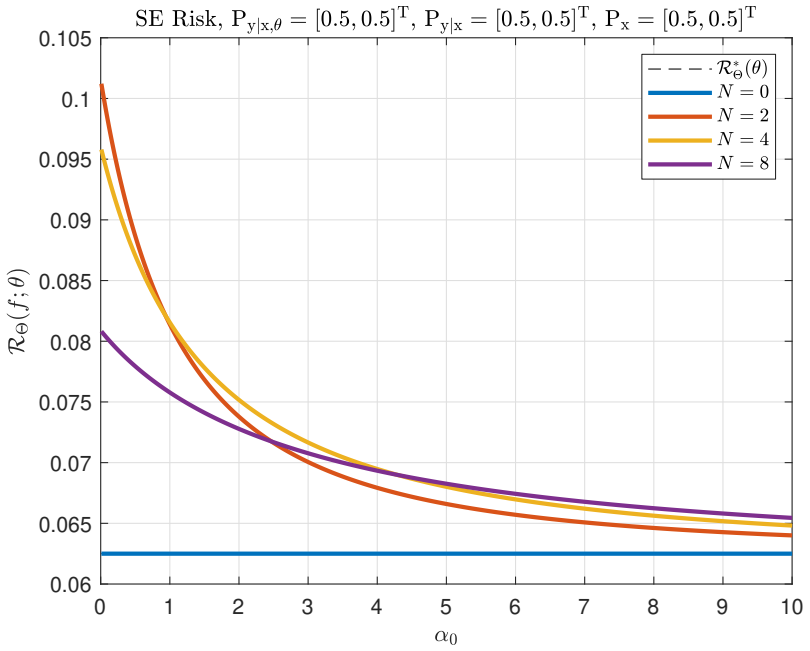

Figure 6.16: Conditional SE Risk versus $\alpha'(x)$, unbiased Dirichlet estimator using varying training set volumes

risk for the given prior conditional distributions.

Note that the minimizing concentration values $\alpha'(x)$ are inversely proportional to the squared-bias of the prior conditional mean. This is sensible; the better the match between the true and prior predictive distributions, the more confidence should be expressed. Also, low concentrations are preferable when the model has low conditional variance; these models can be quickly identified with learners prioritizing the empirical PMF estimate over the prior estimate. Additionally, note that these values $\alpha'(x)$ do not depend on the training volume N.

Figures 6.16 and 6.17 show how the excess conditional squared-error trends as a function of the Dirichlet learner concentration. Note that the latter is based on a biased prior estimate and thus the optimal Dirichlet concentration value is lower.

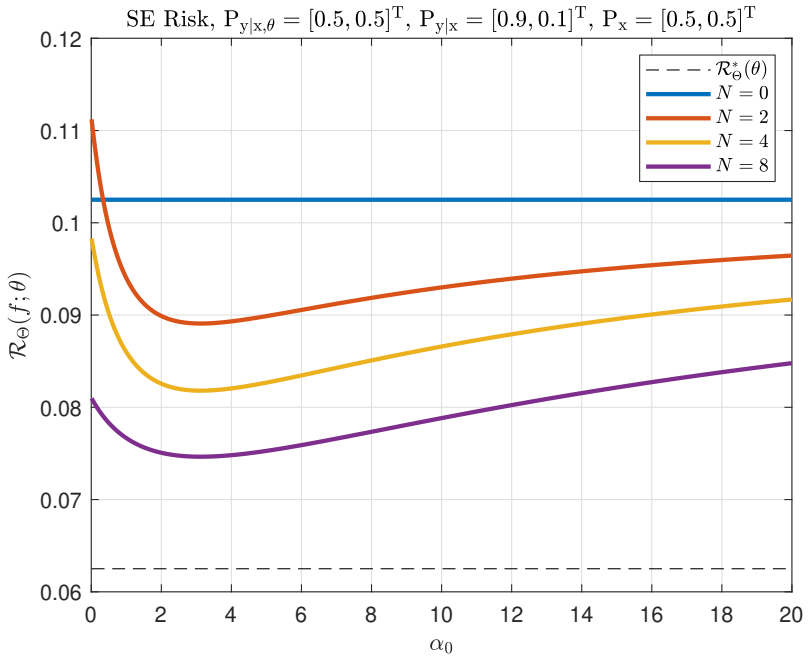

Figure 6.17: Conditional SE Risk versus $\alpha'(x)$, biased Dirichlet estimator using varying training set volumes

6.3.2 Classification: the 0-1 Loss

In this section, 0-1 loss classifiers based on the Dirichlet prior distribution are derived and their performance is assessed.

6.3.2.1 Bayesian Classification

To determine the optimal learning function, the 0-1 loss from Equation (3.17) is substituted into Equation (6.17) and Equation (3.8) to find

$$\begin{aligned} f^*(x; D) &= \arg\max_{y \in \mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|x, D) \\ &= \arg\max_{y \in \mathcal{Y}} \frac{\alpha(y, x) + \bar{N}(y, x; D)}{\alpha'(x) + N'(x; D)} \\ &= \arg\max_{y \in \mathcal{Y}} \Big(\alpha(y, x) + \bar{N}(y, x; D)\Big) \ . \end{aligned} \tag{6.34}$$

Using the Dirichlet prior, different classes are "scored" by counting the number of training samples with a value of X_n matching that of x and combining with the prior parameters $\alpha(\cdot, \mathrm{x})$.

Evaluating the minimum risk (3.23) using the distributions derived from the Dirichlet prior, the Bayes minimum probability of error is

$$\begin{aligned} \mathcal{R}^* &= 1 - \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x},\mathrm{D})\right] \\ &= 1 - \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\frac{\max_{y\in\mathcal{Y}}\left(\alpha(y,\mathrm{x}) + \bar{\mathrm{n}}(y,\mathrm{x})\right)}{\alpha'(\mathrm{x}) + \mathrm{n}'(\mathrm{x})}\right] \\ &= 1 - \sum_{x\in\mathcal{X}} \frac{\mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}}\left(\alpha(y,x) + \bar{\mathrm{n}}(y,x)\right)\right]}{\alpha_0 + N} \ . \end{aligned} \tag{6.35}$$

Figures 6.18 and 6.19 plot the minimum Bayes probability of error against training data volume N and prior concentration α_0, respectively. Note that for $N = 0$, the Bayes risk is $\mathcal{R}^* = 1 - \sum_{x\in\mathcal{X}} \frac{\max_{y\in\mathcal{Y}} \alpha(y,x)}{\alpha_0}$. Additionally, consider the risk for maximal/minimal values of the Dirichlet concentration. For $\alpha_0 \to 0$ (and $N > 1$), the risk is $\mathcal{R}^* = 0$; conversely, for $\alpha_0 \to \infty$, the risk tends to $\mathcal{R}^* \to 1 - \sum_{x\in\mathcal{X}} \frac{\max_{y\in\mathcal{Y}} \alpha(y,x)}{\alpha_0}$. These trends can be visualized in Figures 6.20 and 6.21.

Uniform Prior When the uniform prior is used, the Bayes classifier simplifies to

$$f^*(x; D) = \arg\max_{y\in\mathcal{Y}} \bar{N}(y, x; D) \ , \tag{6.36}$$

a conditional majority decision which chooses the class from $\mathcal{Y}$ most often represented among training set samples D with a matching input value x. This is intuitive, as the model PDF parameter α imparts no confidence as to which classes may be most likely.

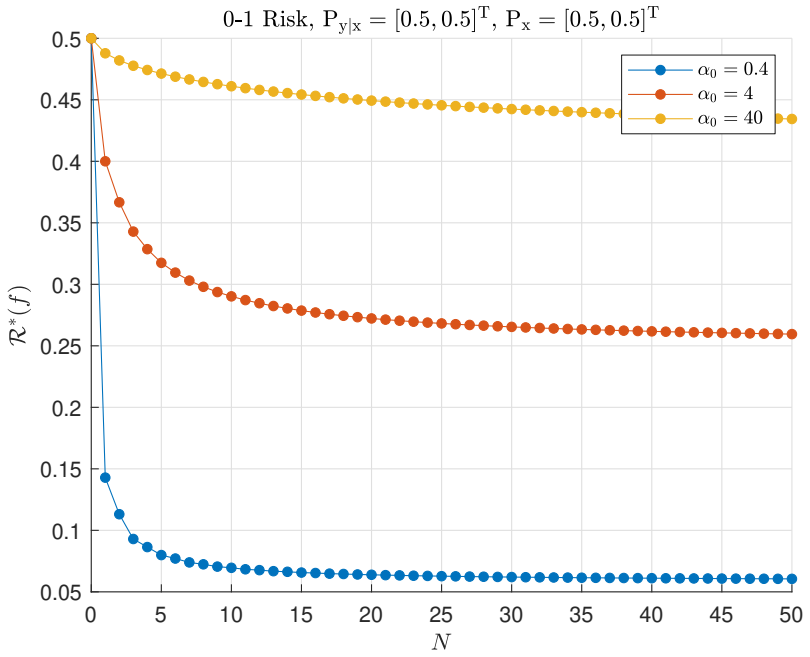

Figure 6.18: Minimum 0-1 Risk for different training data volumes N

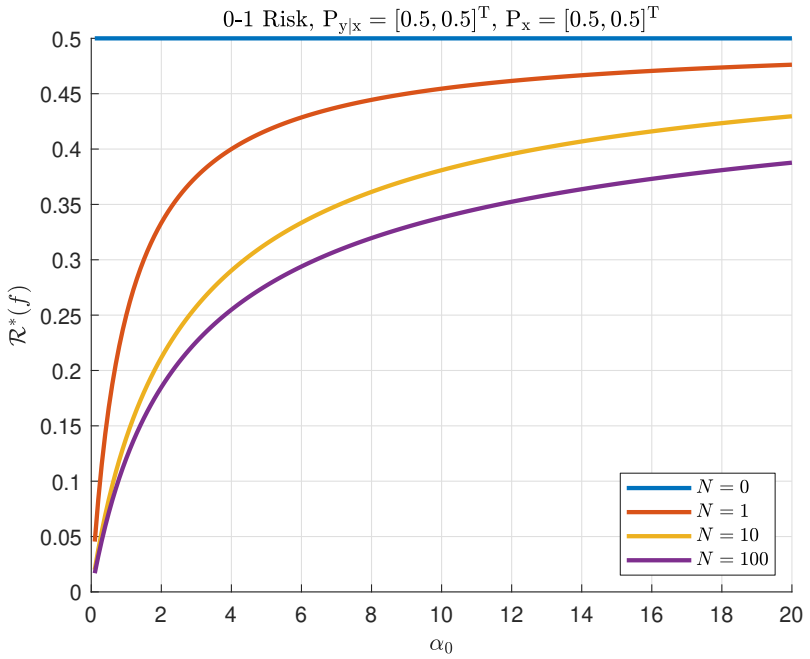

Figure 6.19: Minimum 0-1 Risk for different prior concentrations α_0

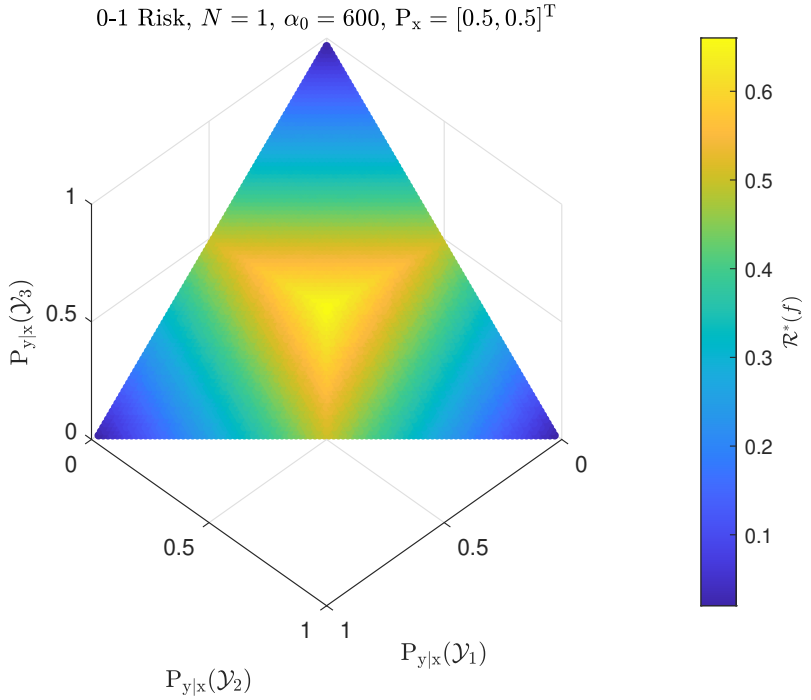

Figure 6.20: Minimum 0-1 Risk for different prior means $\mathrm{P_{y\,|\,x}}$

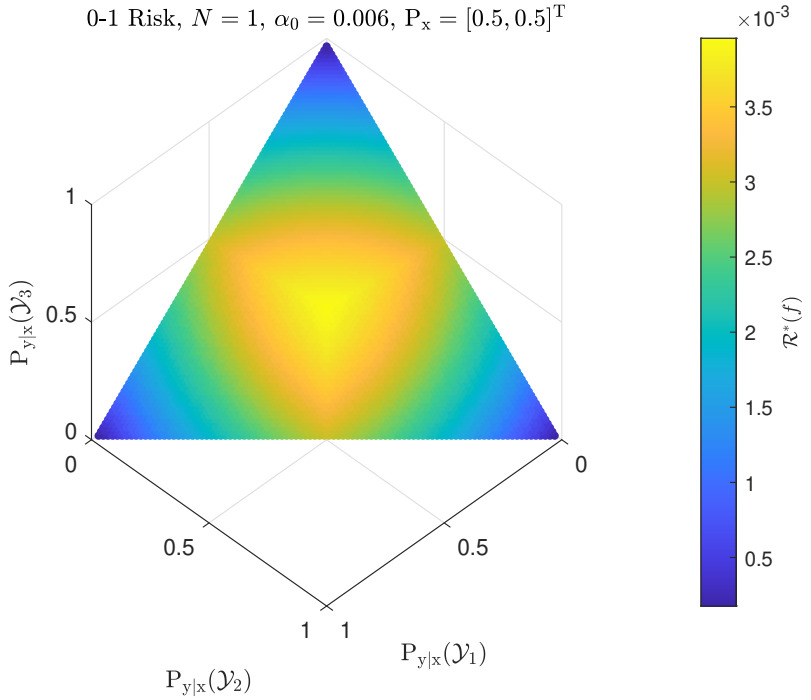

Figure 6.21: Minimum 0-1 Risk for different prior means $\mathrm{P_{y\,|\,x}}$

Using the uniform prior, the minimum Bayes 0-1 risk is

$$\begin{aligned}\mathcal{R}^* &= 1 - \mathrm{E}_{\mathrm{x,D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\right] \\ &= 1 - \frac{1 + |\,\mathcal{X}\,|^{-1}\sum_{x\in\mathcal{X}} \mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)\right]}{|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|}\,.\end{aligned} \tag{6.37}$$

The expectation operates on the maximum value from a subset of a uniform Dirichlet-Multinomial random process. Via the Dirichlet-Multinomial aggregation property [**?**], a consequence of the the uniform PMF $\mathrm{P}_{\bar{\mathrm{n}}}$ is that the individual segments $\bar{\mathrm{n}}(\cdot, x)$ are identically distributed; thus, the expectation will be same for every value x.

To evaluate this expectation, new random variables $\bar{\mathrm{n}}_{\max}(x) \equiv \max_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)$ are introduced and characterized by their identical PMF. To this end, the probability of the event $\mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) = \mathrm{P}\left(\cup_{y\in\mathcal{Y}}\{\bar{\mathrm{n}}(y,x) \geq n\}\right)$ will be determined. As the distribution of $\bar{\mathrm{n}}$ is uniform, the event probability is proportionate to the cardinality of the set $\cup_{y\in\mathcal{Y}}\{\bar{n} : \bar{n}(y,x) \geq n\}$. Using the inclusion-exclusion principle [**?**], the cardinality is represented as

$$\begin{aligned}&\left|\cup_{y\in\mathcal{Y}}\{\bar{n} : \bar{n}(y,x) \geq n\}\right| \\ &= \begin{cases} \binom{N+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} & \text{if } n < 0, \\ \sum_{m=1}^{|\,\mathcal{Y}\,|}\binom{|\,\mathcal{Y}\,|}{m}(-1)^{m-1}\binom{N-mn+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} H\left(\left\lfloor\frac{N}{m}\right\rfloor - n\right) & \text{if } 0 \leq n \leq N, \\ 0 & \text{if } n > N, \end{cases}\end{aligned} \tag{6.38}$$

where $H : \mathbb{Z} \mapsto \{0,1\}$ is the discrete Heaviside step function. For $n < 0$, the cardinality is equivalent to $|\bar{\mathcal{N}}|$.

For $0 \leq n < N$, the cardinality is an alternating binomial summation where the m^{th} term accounts for the different intersections of m of the $|\,\mathcal{Y}\,|$ individual sets $\{\bar{n} : \bar{n}(y,x) \geq n\}$. Observe that the cardinality of the intersections is only dependent on the number of contributing sets m and not on which sets intersect.

Furthermore, note the dependency of the intersection cardinalities on the argument n. The step function contributes such that if $n > \left\lfloor \frac{N}{m} \right\rfloor$, only up to $m-1$ individual sets will intersect. The binomial coefficient $\mathcal{M}\left(\{N - mn, |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1\}\right)$ provides the intersection cardinality for a given m; note the similarity to the cardinality $|\bar{\mathcal{N}}|$ - the only difference is the number of points characterizing the $|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1$ dimensional region.

The probability of interest can thus be expressed as

$$\begin{aligned} \mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) &= \binom{N + |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1}^{-1} \left| \cup_{y \in \mathcal{Y}} \{\bar{n} : \bar{n}(y, x) \geq n\} \right| \qquad (6.39) \\ &= \begin{cases} 1 & \text{if } n < 0, \\ \sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m} (-1)^{m-1} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right) H\left(\left\lfloor \frac{N}{m} \right\rfloor - n\right) & \text{if } 0 \leq n \leq N, \\ 0 & \text{if } n > N. \end{cases} \end{aligned}$$

As the PMF of $\bar{\mathrm{n}}_{\max}(x)$ has support on $n \in [0, \ldots, N]$, the expectation over $\bar{\mathrm{n}}$ is evaluated as

$$\begin{aligned} \mathrm{E}_{\bar{\mathrm{n}}}\left[\,\bar{\mathrm{n}}_{\max}(x)\right] &= \sum_{n=0}^{N} n \Big(\mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) - \mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n+1\right) \Big) \qquad (6.40) \\ &= -1 + \sum_{n=0}^{N} \mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) \\ &= -1 + \sum_{m=1}^{|\mathcal{Y}|} \binom{|\,\mathcal{Y}\,|}{m} (-1)^{m-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right) \end{aligned}$$

and the minimum 0-1 risk is

$$\mathcal{R}^* = 1 - \frac{\sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m} (-1)^{m-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right)}{|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|}\,. \qquad (6.41)$$

It is informative to express the risk for minimal and maximal volumes of training

data. Using the binomial summation identity

$$\sum_{m=0}^{M} \binom{M}{m} (-1)^m g(m) = 0 \,, \tag{6.42}$$

where g is a polynomial function of degree less than M [?], it can be shown that for $N = 0$, the minimum risk is $\mathcal{R}^* = 1 - |\,\mathcal{Y}\,|^{-1}$. This is sensible, as the classes are equiprobable with $\mathrm{P}_{\mathrm{y}} = |\,\mathcal{Y}\,|^{-1}$.

To find the risk for $N \to \infty$, note that

$$\begin{aligned} &\lim_{N\to\infty} \left(|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|\right)^{-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \left(1 - \frac{mn}{N+l}\right) \\ &\quad = \lim_{N/m\to\infty} \frac{|\,\mathcal{X}\,|}{m} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \left(1 - \frac{mn}{N}\right)^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \frac{m}{N} \\ &\quad = \frac{|\,\mathcal{X}\,|}{m} \int_0^1 (1-t)^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \mathrm{d}t \\ &\quad = \frac{1}{m|\,\mathcal{Y}\,|} \,. \end{aligned} \tag{6.43}$$

The irreducible 0–1 risk for the uniform prior tends towards

$$\begin{aligned} \mathcal{R}^* &\to 1 - |\,\mathcal{Y}\,|^{-1} \sum_{m=1}^{|\,\mathcal{Y}\,|} \binom{|\,\mathcal{Y}\,|}{m} (-1)^{m-1} m^{-1} \\ &= 1 - |\,\mathcal{Y}\,|^{-1} \sum_{m=1}^{|\,\mathcal{Y}\,|} m^{-1} \,, \end{aligned} \tag{6.44}$$

providing a lower bound for the achievable 0–1 Bayes risk. The above formulation has made use of the alternating summation identity from [?] to display the risk with a form including the $|\,\mathcal{Y}\,|^{\mathrm{th}}$ harmonic number $H_{|\,\mathcal{Y}\,|} \equiv \sum_{m=1}^{|\,\mathcal{Y}\,|} m^{-1}$. Observe that the irreducible risk does not depend on the cardinality $|\,\mathcal{X}\,|$.

Figure 6.22 demonstrates how the minimum 0-1 risk decreases with training volume N; observe that the risk is more severe for sequences corresponding to higher $|\,\mathcal{Y}\,|$. It is sensible that the probability of error should increase when more classes

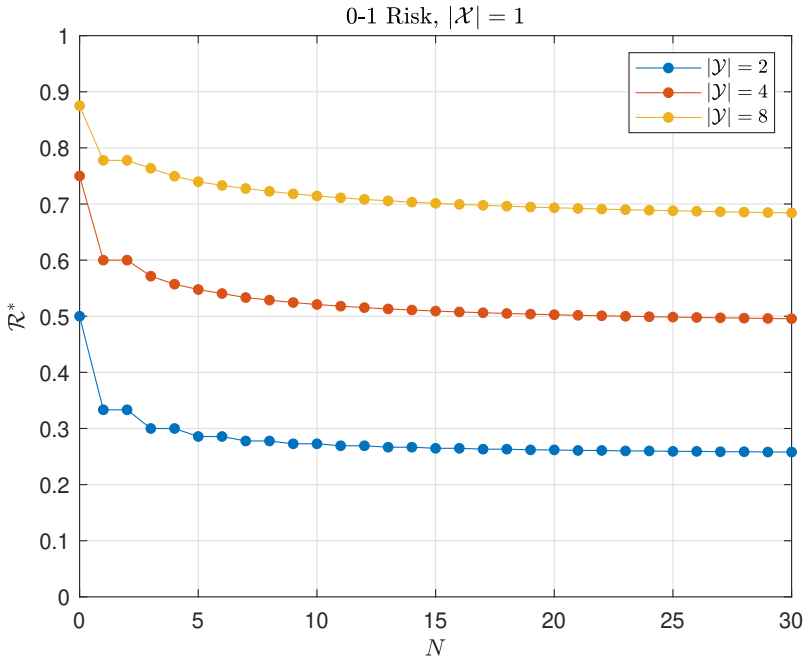

Figure 6.22: Minimum 0-1 Risk vs training set volume N

are considered. Figure 6.23 illustrates the minimum risk with multiple sequences for different cardinalities $|\mathcal{X}|$. Note that risk increases with $|\mathcal{X}|$. Considering $\mathrm{E_D}\left[N'(\mathrm{D})\right] = \mu_{\mathrm{n}'} = N/|\mathcal{X}|$, this should be intuitive - each conditional empirical distribution $\bar{N}(\cdot, x; D)/N'(x; D)$ is forced to approximate $\tilde{\theta}(x)$ with less data.

Further insight into how $|\mathcal{X}|$ affects the risk can be acquired by plotting the risk as a function of $N/|\mathcal{X}|$. In Figure 6.24, it is shown that the minimal risk can be approximated by a function dependent only on $N/|\mathcal{X}|$; of the series plotted, only the series for $|\mathcal{X}| = 1$ shows non-negligible deviation from the others.

It is also informative to graph the $N = 0$ and $N \to \infty$ minimum risk as a function of $|\mathcal{Y}|$; both formulas are independent of $|\mathcal{X}|$. Figure 6.25 displays these bounds; note the margin in the probability of error between the optimal $N = 0$ and $N \to \infty$ classifiers. For $|\mathcal{Y}| = 2$ binary classification, both sequences are at their minimum, and infinite training data provides a reduction in expected probability of error from 0.5 to 0.25. As $|\mathcal{Y}|$ increases, the classification risk for both the $N = 0$ and $N \to \infty$ cases tend to unity and the error reduction for $N \to \infty$ decreases.

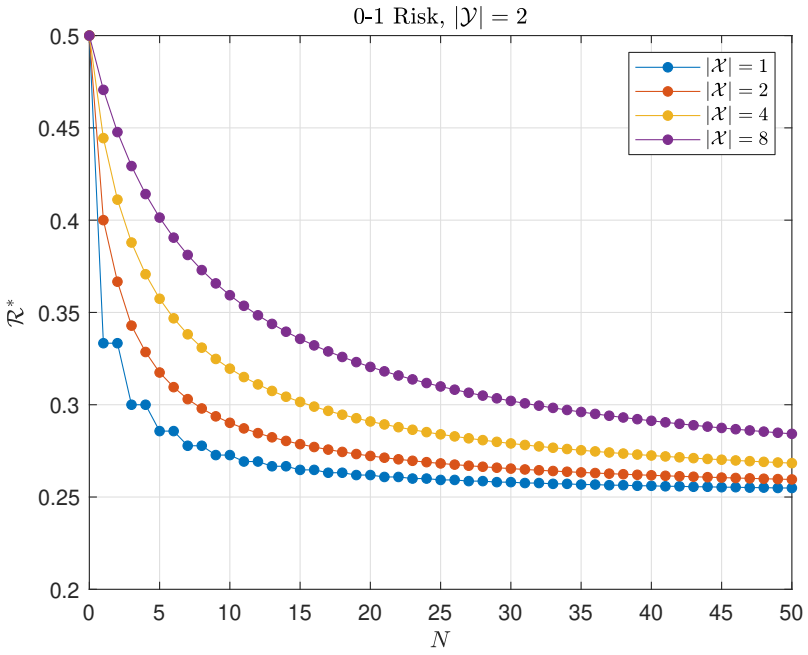

Figure 6.23: Minimum 0-1 Risk vs training set volume N

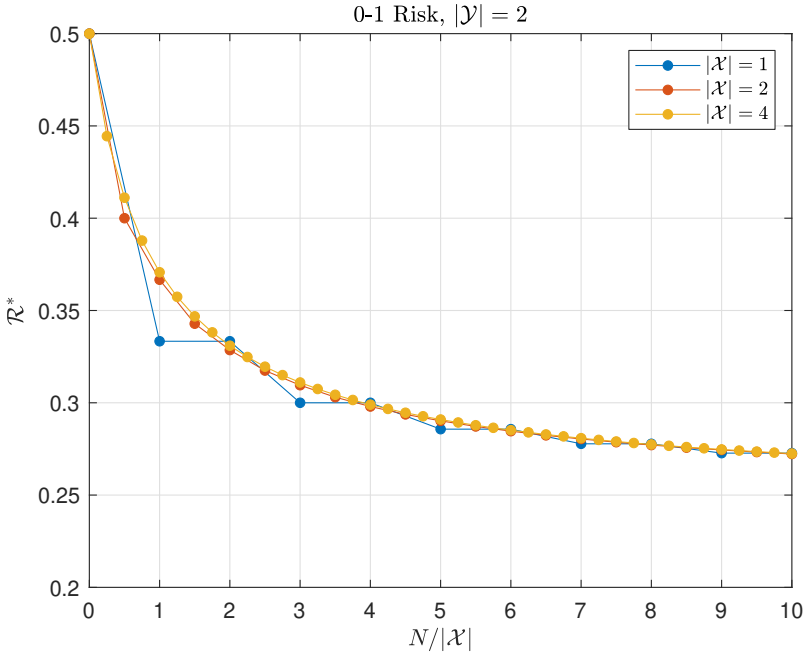

Figure 6.24: Minimum 0-1 Risk vs $N/|\mathcal{X}|$

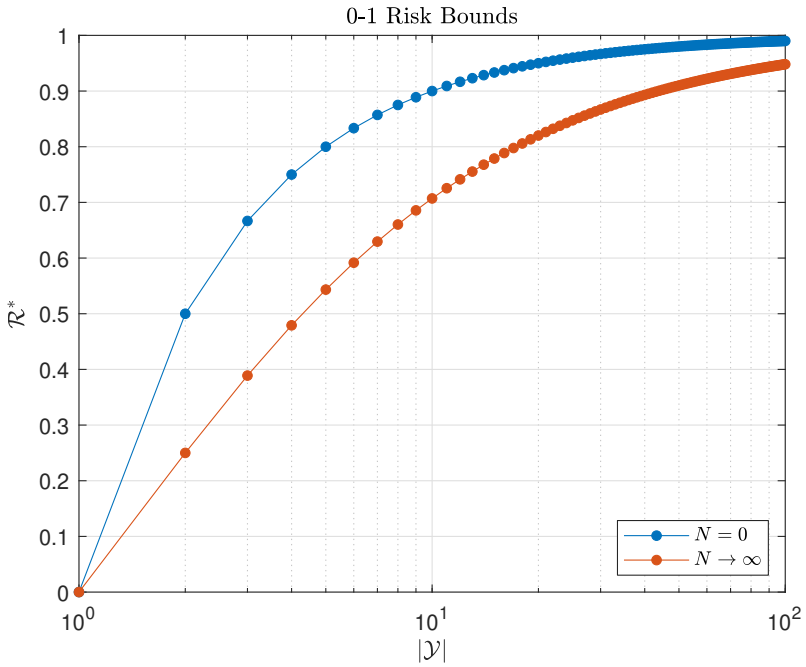

Figure 6.25: Minimum 0-1 Risk vs $|\mathcal{Y}|$

6.3.2.2 Conditional Probability of Error for a Dirichlet-based Classifier

Substituting the optimal Dirichlet-based classifier into the formula for the conditional probability of error 3.18, the risk is

$$\mathcal{R}_\Theta(f;\theta) = 1 - \sum_{x\in\mathcal{X}} \theta'(x)\, \mathrm{E}_{\bar{\mathrm{n}}\,|\theta}\left[\tilde{\theta}\Big(\arg\max_{y\in\mathcal{Y}}\Big(\bar{\mathrm{n}}(y,x)+\alpha(y,x)\Big);x\Big)\right]. \quad (6.45)$$

Figures 6.26 and 6.27 show how the conditional risk trends for classifiers based on well-matched and poorly-matched informative Dirichlet priors, respectively. Note that the well-matched prior does better with higher prior concentrations α_0; this is reflective of the fact that the maximizing arguments $y \in \mathcal{Y}$ of both the true model $\tilde{\theta}(x)$ and the prior mean $\alpha(\cdot,x)/\alpha'(x)$ are the same.

Also, it is important to consider how a given classifier performs for varying models θ. Figures 6.28 and 6.29 demonstrate the excess conditional probability of error achieved by the conditional majority decision (based on a non-informative Dirichlet

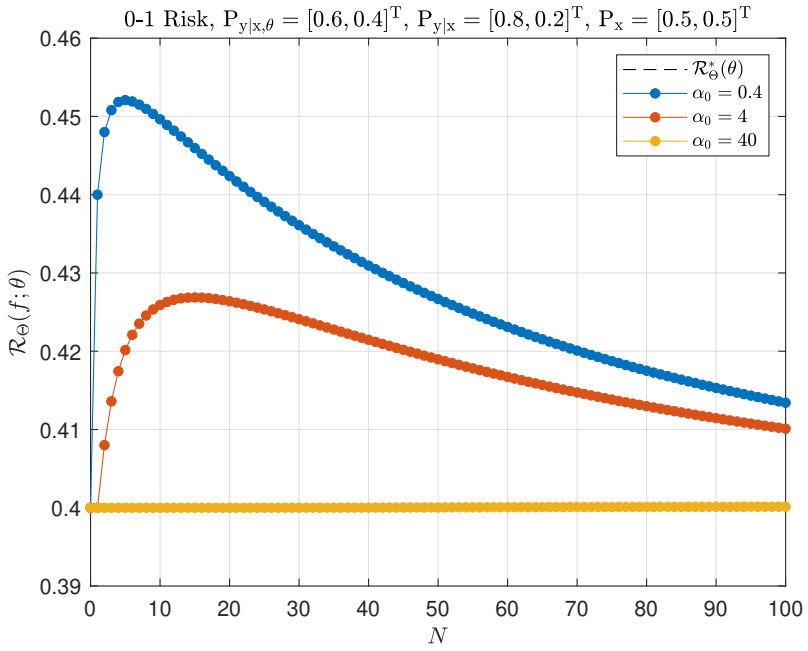

Figure 6.26: Excess conditional probability of error, well-matched informative Dirichlet-based classifier

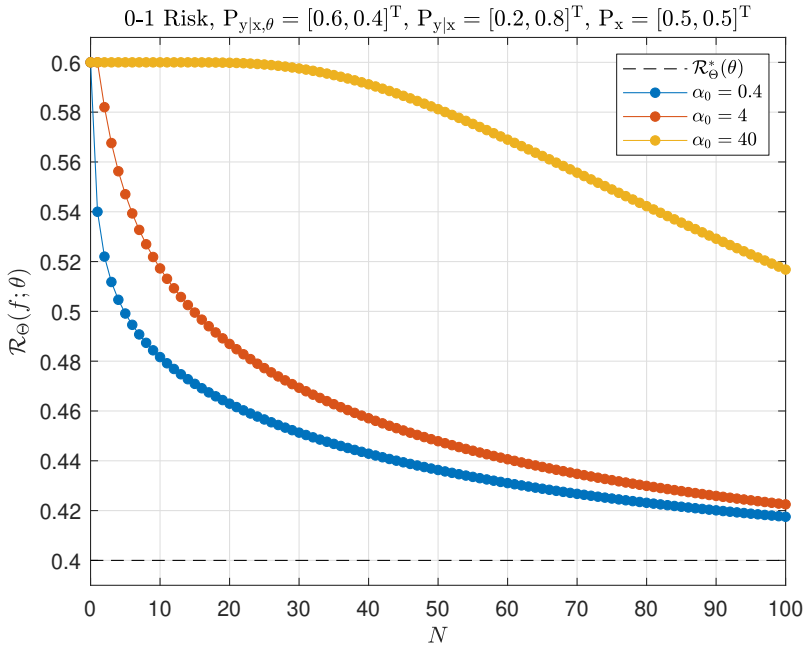

Figure 6.27: Excess conditional probability of error, poorly-matched informative Dirichlet-based classifier

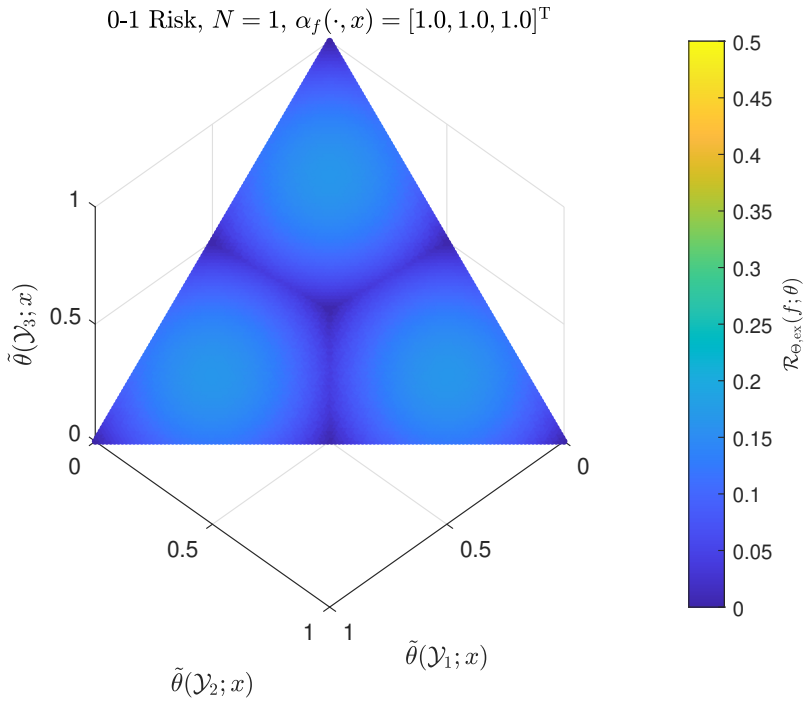

Figure 6.28: Excess conditional probability of error, conditional majority decision

prior) and by a classifier derived from an informative Dirichlet prior, respectively. Note that while the former has fewer models in which the error is critically high, the latter has more models in which the clairvoyant risk $\mathcal{R}^*_\Theta(\theta)$ is achieved. This is a fundamental trade-off between Bayesian learners based on non-informative versus informative priors.

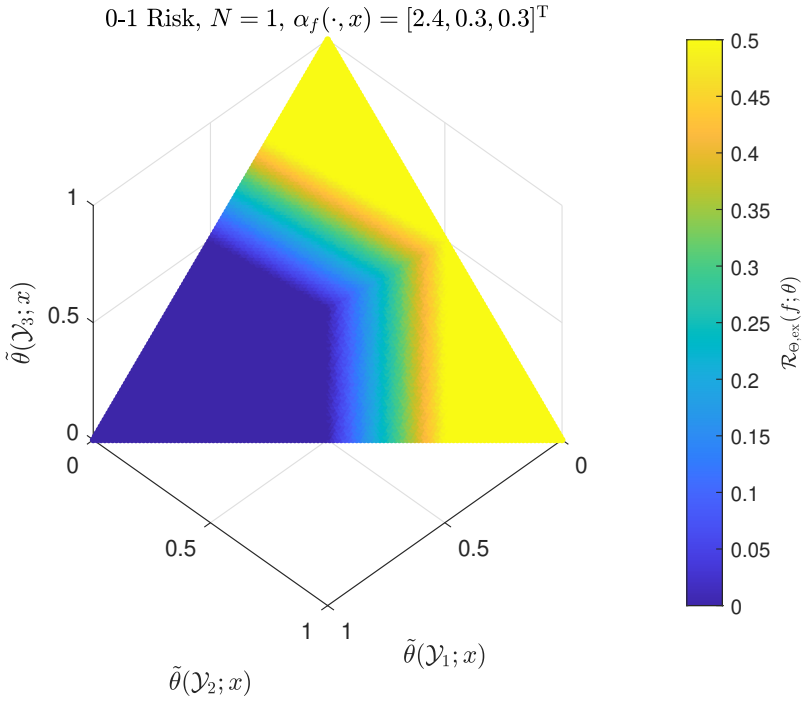

Figure 6.29: Excess conditional probability of error, informative Dirichlet-based classifier

Chapter 7: Plan for Completion

To complete the thesis, additional work is required in some key areas. Most importantly, additional theory needs to be set down for Bayesian learning with prior distributions of limited support. Subsequently, specific low-dimensionality priors will be used to design decision functions for human recognition tasks, exploiting insights into which models are sensible. Additionally, the established framework will be extended for uncountably infinite sets and joint decisions based on multiple simultaneous observations.

7.1 Perform additional research into Bayesian learning with limited-support priors

The primary assertion of this proposal is that all machine learning either explicitly or implicitly depends on prior knowledge imparted by the designer. It is suggested that highly informative prior distributions are required to achieve near-minimal risk with Bayesian approaches when training data is limited; furthermore, it is hypothesized that the specificity of the prior should be expressed using support of relatively low intrinsic dimensionality. These designs enable high performance learning (if the prior support is well chosen) while also keeping the computational complexity of the resulting algorithms sufficiently low, which is a primary concern for practical implementation.

As such, one of the main thrusts for continued investigation is to analyze the consequences of limited-dimension prior distributions with $\dim(\Theta) = M - 1 < |\,\mathcal{Y}\,||\,\mathcal{X}\,| - 1$. Initial research will continue with the assumption that the sets $\mathcal{X}$ and $\mathcal{Y}$ are finite, promoting simplicity and explainable results. Trade studies will be performed to assess how the reduced support dimensionality $M - 1$ and mismatch from the true model θ affect the achievable risk. Additional focus will be put on asymptotic trends with training data volume N - unlike predictive distributions based

on full-support priors, those derived from limited-support priors will certainly not be asymptotically consistent estimators of θ in general and thus the clairvoyant risk may never be met.

A class of low-dimensional priors that will be of specific interest is the class of mixture models satisfying

$$\theta \equiv \sum_{m=1}^{M} \phi_m h_m \ , \tag{7.1}$$

such that the data-generating model θ is a convex combination of distributions $h_m \in \Theta$. In such a case, the convex coefficients ϕ are referred to as "hyperparameters" - an obvious choice for statistical characterization of these coefficients is, again, the Dirichlet PDF. It will be shown that if the mixture distributions h_m have disjoint support, such that

$$h_i(y,x) \cdot h_j(y,x) = 0, \quad \forall\ (y,x) \in \mathcal{Y} \times \mathcal{X}, \quad \forall\ i,j\ , \tag{7.2}$$

then the training data likelihood function $\mathrm{P}_{\mathrm{D}\,|\phi}$ is of exponential form and thus a hyperparameter Dirichlet PDF p_{ϕ} is a conjugate prior.

7.2 Assess performance of low-dimensional support Bayesian learners on human recognition applications

Having developed Bayesian inference methods based on limited-dimensionality priors, the goal will be to successfully demonstrate performance on human recognition tasks. Simulated binary classification data mimicking audiovisual recordings will be used first for proof-of-concept; data drawn from the set $\mathcal{X} = \{0,1\}^K$ will be used, first with singular K and then with a multi-index K to approximate audio records and images, respectively. To demonstrate a property characteristic of such data, distinct patterns will be generated and subjected to random translations and rotations; Figure 7.1 shows example image data.

This problem clearly exhibits the human-recognition task properties discussed in

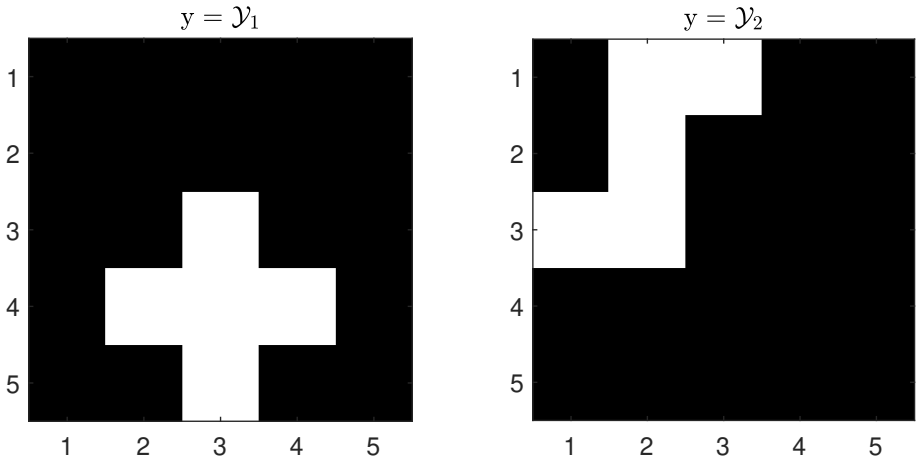

Figure 7.1: Simulated character recognition data

Section 5. First, note that the marginal model $\theta' : \mathcal{X} \mapsto \mathbb{R}_{\geq 0}$ is clearly non-zero for only select values x = x; for the data shown, only $9 + 2(9) = 27$ images x are ever observed, compared to the $2^{25} = 33554432$ possible binary images. Thus, marginal priors with restricted norms $\left\|\theta'\right\|_0 \ll |\,\mathcal{X}\,|$will be used. Second, it should be clear that none of the images x that can be generated when y = $\mathcal{Y}_1$ would ever occur when y = $\mathcal{Y}_2$, and vice versa. As such, the conditional models satisfy $\left\|\tilde{\theta}(x)\right\|_\infty = 1$ and the clairvoyant risk for this problem should be zero.

By using appropriate limited-support prior distributions, it is expected that low probability of error will be realizable even with limited training data and no *a priori* understanding of the patterns used. After successful demonstration, the problem size will be increased, such that training data volume becomes more critical and prior knowledge more important. The decision functions will subsequently be tested on real-world data such as the optical character recognition data [**?**] exemplified in Figure 3.4.

7.3 Generalize existing results for uncountably infinite sets

The preliminary work assumes that the data pairs (y, x) are drawn from a finite set $\mathcal{Y} \times \mathcal{X}$. As mentioned, essentially all modern machine learning algorithms are deployed on digital processors and thus this characterization of the data is not presumptuous - even if the cardinalities of the sets are immense, the sets will undoubtedly be finite if the data is to be stored using a digital representation. However, as many digital records of interest for regression and classification are made by sampling continuous physical phenomena (acoustic fluctuations, electromagnetic waves, etc.), a complete learning theory must include an extension to continuous spaces. An analogous case can be found in harmonic analysis - certainly our understanding of the discrete Fourier transform has benefited from study of the continuous Fourier transform [**?**].

With this consideration, the initial work will be generalized for Euclidean spaces $\mathcal{Y}$ and $\mathcal{X}$ (starting with $\mathcal{Y} = \mathcal{X} = \mathbb{R}$) and the joint distribution θ will be treated as a probability density function. As the set of PDF's $\mathcal{P}(\Theta)$ is infinite-dimensional, an explicit prior distribution p_{θ} cannot be defined over this space. Instead, the model θ is treated as a random process. The empirical distribution is now defined as $\bar{\mathrm{n}} \equiv \sum_{n=1}^{N} \delta\left(\cdot - \mathrm{D}_n\right)$ on its Euclidean domain.

For the preliminary work on learning with full-support Dirichlet priors, generalization efforts have already been started, treating the model as a continuous Dirichlet process $\theta \sim \mathrm{DP}(\alpha)$ [**?**]. Inheriting all the desirable properties of the Dirichlet distribution, it is straightforward to formulate the posterior model process $\theta | \, \bar{\mathrm{n}} \sim \mathrm{DP}(\alpha + \bar{\mathrm{n}})$ and the Bayesian predictive PDF for novel observations. To enable analyses similar to those performed for finite data sets, the concepts of the multinomial process and Dirichlet-multinomial process will be introduced, inheriting properties from their corresponding PMF's over finite domains.

After using the Dirichlet process to investigate Bayesian learning with "full support"

prior knowledge, the limited-support prior results from the finite set case will be generalized as well. Although typically not described as such, this is the type of problem setup assumed by most existing Bayesian parametric learning methods. As before, the continuous model θ will be treated as a mixture of a finite number of distributions h_m and the finite-dimensional "support" of the model will be characterized by hyperparameters ϕ.

7.4 Expand framework for semi-supervised joint decisions based on multiple observations

In the preliminary work, Bayesian decisions are made using the model posterior $\mathrm{p}_{\tilde{\theta}|\,\mathrm{x,D}}$. Under the Dirichlet prior assumption, it has been shown that the predictive models $\tilde{\theta}(x)$ are conditionally independent of the novel observation x given the training data D. In general this will not be the case. As such, the use of x to refine our statistical understanding of the model distribution can be viewed as a *semi-supervised* learning procedure. Semi-supervised learning is of interest in machine learning research - labeled data can be costly to obtain, while unlabeled data is in abundance.

The current results define the decision function and assess the risk for a single decision $h \in \mathcal{H}$ based on a single observation $\mathrm{x} = x$. To enable a more meaningful investigation of semi-supervised learning potential, the existing work will be generalized for L decisions $(h_1, \ldots, h_L) \in \mathcal{H}^L$ based on observations $(\mathrm{x}_1, \ldots, \mathrm{x}_L) \in \mathcal{X}^L$, each corresponding to a distinct unobserved random element $(\mathrm{y}_1, \ldots, \mathrm{y}_L) \in \mathcal{Y}^L$. The loss function will generalize as the sum of the losses incurred by each decision, $\mathcal{L}(h, y) = \sum_{l=1}^{L} \mathcal{L}_0(h_l, y_l)$. This setup also enables a direct comparison of the resultant Bayesian decision functions with non-probabilistic learners that are designed with empirical risk minimization over a training set and are assessed with empirical risk on a reserved test set.

It is straightforward to show that using a Dirichlet prior, the model will again be

conditionally independent of the novel observations and thus the decisions will be made independently using the Bayesian predictive distribution $\mu_{\tilde{\theta}|\,\mathrm{D}}$. Consequently, each decision h_l will be dependent solely on the observation x_l.

Moving away from Dirichlet priors to distributions with statistical dependency between θ' and $\tilde{\theta}$, each specific decision h_l will be dependent on all the novel observations $(\mathrm{x}_1, \ldots, \mathrm{x}_L)$, not just the corresponding observation. How the risk trends with the number of novel observations L will be of interest; it can be shown that the novel observations refine the posterior distribution of θ' and consequently, the posterior of the predictive model $\tilde{\theta}$.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53:370–418, 1763.

[3] Matthew J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London, 01 2003.

[4] Richard Ernest Bellman. *Dynamic Programming*. Dover Publications, Inc., New York, NY, USA, 2003.

[5] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2000.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[7] George E.P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, 1973.

[8] Richard A. Brualdi. *Introductory Combinatorics*. Pearson, fifth edition, 2010.

[9] F. W. Campbell and J. J. Kulikowski. Orientational selectivity of the human visual system. *The Journal of Physiology*, 187(2):437–445, 1966.

[10] F. W. Campbell and J. G. Robson. Application of fourier analysis to the visibility of gratings. *The Journal of Physiology*, 197(3):551–566, 1968.

[11] S. Del Marco, P. Heller, and J. Weiss. An m-band, 2-dimensional translation-invariant wavelet transform and applications. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1077–1080 vol.2, May 1995.

[12] 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 (Methodological)*, 39(1):1–38, 1977.

[13] Luc Devroye, László Győrfi, and Gábor Lugosi. *A Probabilisitic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[14] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2 of *Probability and Mathematical Statistics*. John Wiley & Sons, New York, New York, second edition, 1971.

[15] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[16] James L. Flanagan. *Speech Analysis, Synthesis and Perception*. Communication and Cybernetics. Springer-Verlag, second edition, 1972.

[17] Samuel J. Gershman and David M. Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56, 2012.

[18] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics*. Addison-Wesley, Reading, Massachusetts, second edition, 1994.

[19] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[20] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Probability and Statistics. John Wiley & Sons, 1997.

[21] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume 1 of *Signal Processing Series*. Prentice-Hall, Upper Saddle River, New Jersey, 1993.

[22] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory*, volume 2 of *Signal Processing Series*. Prentice-Hall, Upper Saddle River, New Jersey, 1998.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[24] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[25] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

[26] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[27] S. G. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., Orlando, FL, USA, 3rd edition, 2008.

[28] Edward Meeds and Simon Osindero. An alternative infinite mixture of gaussian process experts. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 883–890. MIT Press, 2006.

[29] Thomas P. Minka. Expectation propagation for approximate bayesian inference. *CoRR*, abs/1301.2294, 2013.

[30] Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton. Deep belief networks for phone recognition. *Science*, 4, 07 2010.

[31] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, fourth edition, 2002.

[32] J. W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247, Sep. 1993.

[33] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[34] C. Radhakrishna Rao. Maximum likelihood estimation for the multinomial distribution. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 18(1/2):139–148, 1957.

[35] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[36] Steven Roman. The logarithmic binomial formula. *American Mathematical Monthly*, 99(7), 1992.

[37] Frederick F. Stephan. The expected value and variance of the reciprocal and other negative powers of a positive bernoullian variate. *The Annals of Mathematical Statistics*, 16(1):50–61, 1945.

[38] Charles J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–620, 07 1977.

[39] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Elsevier, 2015.

[40] V. Vapnik and Alexei Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Theory of Probability and its Applications*, volume 16, 1971.

[41] J. G. Wendel. Note on the gamma function. *The American Mathematical Monthly*, 55(9):563–564, 1948.

[42] David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Comput.*, 8(7):1341–1390, October 1996.

Appendix A:

A.1 Dirichlet random process conditioned on its aggregation

This section details an important property of Dirichlet distributed random processes. The following development first considers Dirichlet random processes over a countable domain and then generalizes for continuous-domain Dirichlet processes.

First, define the PDF of a Dirichlet aggregation [?]. Let the random process $\theta \in \Theta = \mathcal{P}(\mathcal{Y})$ be Dirichlet over the countable set $\mathcal{Y}$ with parameterizing function $\alpha \in \mathbb{R}^{+\mathcal{Y}}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$; the aggregation $\theta' \in \mathcal{P}(\mathcal{Z})$, $\theta'(z) \equiv \sum_{y \in \mathcal{S}(z)} \theta(y)$ is thus also Dirichlet and has a parameterizing function $\alpha' \in \mathbb{R}^{+\mathcal{Z}}$, $\alpha'(z) \equiv \sum_{y \in \mathcal{S}(z)} \alpha(y)$.

The PDF of the original random process θ conditioned on its aggregation θ' can be formulated as

$$\begin{aligned} \mathrm{p}_{\theta|\theta'}(\theta|\theta') &= \frac{\beta(\alpha') \prod_{y \in \mathcal{Y}} \theta(y)^{\alpha(y)-1}}{\beta(\alpha) \prod_{z \in \mathcal{Z}} \theta'(z)^{\alpha'(z)-1}} \\ &= \prod_{z \in \mathcal{Z}} \left[\beta\Big(\big\{\alpha(y) : y \in \mathcal{S}(z)\big\}\Big)^{-1} \frac{\prod_{y \in \mathcal{S}(z)} \theta(y)^{\alpha(y)-1}}{\theta'(z)^{\alpha'(z)-1}} \right] \\ &= \prod_{z \in \mathcal{Z}} \left[\frac{\theta'(z)^{1-|\,\mathcal{S}(z)|}}{\beta\Big(\big\{\alpha(y) : y \in \mathcal{S}(z)\big\}\Big)} \prod_{y \in \mathcal{S}(z)} \left(\frac{\theta(y)}{\theta'(z)}\right)^{\alpha(y)-1} \right], \end{aligned} \tag{A.1}$$

which is defined for $\Big\{\theta \in {\mathbb{R}_{\geq 0}}^{\mathcal{Y}} : \sum_{y \in \mathcal{S}(z)} \theta(y) = \theta'(z), \quad \forall z \in \mathcal{Z}\Big\}$.

Observe that the partitioned segments are conditionally independent; introduce subscript notation to refer to the function segment $\theta_z = \Big\{\theta(y) : y \in \mathcal{S}(z)\Big\}$. The PDF $\mathrm{p}_{\theta|\theta'}$ can now be decomposed as $\mathrm{p}_{\theta|\theta'}(\ldots, \theta_z, \ldots|\theta') = \prod_{z \in \mathcal{Z}} \mathrm{p}_{\theta_z|\theta'(z)}\Big(\theta_z|\theta'(z)\Big)$

Next, normalize the segments of θ to form $\tilde{\theta} = \Big(\ldots, \tilde{\theta}_z, \ldots\Big)$, where $\tilde{\theta}_z \equiv \theta_z/\theta'(z)$, and formulate the conditional PDF

$$\begin{aligned} \mathrm{p}_{\tilde{\theta}|\theta'}\Big(\tilde{\theta}|\theta'\Big) &= \prod_{z \in \mathcal{Z}} \left[\frac{\prod_{y \in \mathcal{S}(z)} \tilde{\theta}_z(y)^{\alpha(y)-1}}{\beta\Big(\big\{\alpha(y) : y \in \mathcal{S}(z)\big\}\Big)} \right] \\ &= \prod_{z \in \mathcal{Z}} \mathrm{Dir}\Big(\tilde{\theta}_z; \big\{\alpha(y) : y \in \mathcal{S}(z)\big\}\Big)\,. \end{aligned} \tag{A.2}$$

which is defined for $\tilde{\theta} \in \prod_{z \in \mathcal{Z}} \Big\{\tilde{\theta}_z \in \mathcal{P}\Big(\mathcal{S}(z)\Big)\Big\}$. Thus after conditioning, the normalized segments $\tilde{\theta}_z$ are Dirichlet distributed, independent of one another, and independent of the aggregation θ'. This principle holds for continuous-domain Dirichlet processes θ as well - the segments $\tilde{\theta}_z$ are now continuous-domain Dirichlet processes.

A.2 Multinomial Distribution Properties

A.2.1 Aggregation

A characteristic of a Multinomial random process is that its aggregations are also Multinomial [**?**]. Consider a random process $\bar{\mathrm{n}} \sim \mathrm{Multi}(N, \theta)$ over the set $\mathcal{Y}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$; the transformed random process $\mathrm{n}'(z) \equiv \sum_{y \in \mathcal{S}(z)} \bar{\mathrm{n}}(y)$ is distributed as $\mathrm{n}' \sim \mathrm{Multi}(N, \theta')$ with parameterizing function $\theta'(z) = \sum_{y \in \mathcal{S}(z)} \theta(y)$.

To prove this principle, define the subset $\tilde{\mathcal{N}} = \left\{ \bar{n} \in \mathcal{N} : \sum_{y \in \mathcal{S}(z)} \bar{n}(y) = n'(z), \quad \forall z \in \mathcal{Z} \right\} \subseteq \bar{\mathcal{N}}$, where the original random process $\bar{\mathrm{n}} \in \bar{\mathcal{N}}$. Next, observe

$$\begin{aligned} \mathrm{P}_{\mathrm{n}'}(n') &= \sum_{\bar{n} \in \tilde{\mathcal{N}}} \mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) \\ &= \mathcal{M}(n') \prod_{z \in \mathcal{Z}} \sum_{n'(z) = \sum_{y \in \mathcal{S}(z)} \bar{n}(y)} \mathcal{M}\left(\{ \bar{n}(y) : y \in \mathcal{S}(z) \} \right) \prod_{y \in \mathcal{S}(z)} \theta(y)^{\bar{n}(y)} \\ &= \mathcal{M}(n') \prod_{z \in \mathcal{Z}} \theta'(z)^{n'(z)} = \mathrm{Multi}(n'; N, \theta') \ , \end{aligned} \tag{A.3}$$

where the multinomial theorem [**?**] has been used.

A.2.2 Conditioned on its Aggregation

If the multinomial random process $\bar{\mathrm{n}}$ is conditioned on its aggregation over the partition $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$, the distinct segments $\bar{\mathrm{n}}(y)$, $y \in \mathcal{S}(z)$ become independent multinomial random processes,

$$\begin{aligned} \mathrm{P}_{\bar{\mathrm{n}} | \, \mathrm{n}'}(\bar{n} | n') &= \frac{\mathcal{M}(\bar{n}) \prod_{y \in \mathcal{Y}} \theta(y)^{\bar{n}(y)}}{\mathcal{M}(n') \prod_{z \in \mathcal{Z}} \theta'(z)^{n'(z)}} \\ &= \prod_{z \in \mathcal{Z}} \left[\mathcal{M}\left(\left\{ \bar{n}(y) : y \in \mathcal{S}(z) \right\} \right) \prod_{y \in \mathcal{S}(z)} \left(\frac{\theta(y)}{\theta'(z)} \right)^{\bar{n}(y)} \right] \\ &= \prod_{z \in \mathcal{Z}} \left[\mathcal{M}\left(\left\{ \bar{n}(y) : y \in \mathcal{S}(z) \right\} \right) \prod_{y \in \mathcal{S}(z)} \tilde{\theta}_z(y)^{\bar{n}(y)} \right] \\ &= \prod_{z \in \mathcal{Z}} \mathrm{Multi}\left(\left\{ \bar{n}(y) : y \in \mathcal{S}(z) \right\}; n'(z), \left\{ \tilde{\theta}_z(y) : y \in \mathcal{S}(z) \right\} \right) \ , \end{aligned} \tag{A.4}$$

on the domain $\left\{ \bar{n} \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}} : \sum_{y \in \mathcal{S}(z)} \bar{n}(y) = n'(z), \quad \forall z \in \mathcal{Z} \right\}$. Observe that the segment over the set $\mathcal{S}(z)$ sums to $\mathrm{n}'(z)$ and is parameterized by the normalized segment $\tilde{\theta}_z \equiv \left\{ \theta(y) / \theta'(z) : y \in \mathcal{S}(z) \right\}$.

A.3 Dirichlet-Multinomial random process conditioned on its aggregation

A defining characteristic of a Dirichlet-Multinomial random process is that its aggregations are also Dirichlet-Multinomial [**?**]. Consider a DM random process $\bar{\mathrm{n}} \sim \mathrm{DM}(N, \alpha)$ over the set $\mathcal{Y}$. Define an arbitrary partition of $\mathcal{Y}$: $\{\ldots, \mathcal{S}(z), \ldots\}$, $z \in \mathcal{Z}$; the transformed random process $\mathrm{n}'(z) \equiv \sum_{y \in \mathcal{S}(z)} \bar{\mathrm{n}}(y)$ is necessarily Dirichlet-Multinomial with parameterizing function $\alpha'(z) = \sum_{y \in \mathcal{S}(z)} \alpha(y)$.

It can be shown that conditioned on the aggregation n', the segments $\Big\{\bar{\mathrm{n}}(y) : y \in \mathcal{S}(z)\Big\}$ of the original random process become independent Dirichlet-Multinomial random processes, such that

$$\begin{aligned}
\mathrm{P}_{\bar{\mathrm{n}}|\,\mathrm{n}'}(\bar{n}|n') &= \frac{\mathcal{M}(\bar{n})\beta(\alpha)^{-1}\beta(\alpha+\bar{n})}{\mathcal{M}(n')\beta(\alpha')^{-1}\beta(\alpha'+n')} \\
&= \left(\prod_{z\in\mathcal{Z}} \frac{\Gamma\Big(\alpha'(z)+n'(z)\Big)}{n'(z)!\Gamma\Big(\alpha'(z)\Big)}\right)^{-1} \left(\prod_{y\in\mathcal{Y}} \frac{\Gamma\Big(\alpha(y)+\bar{n}(y)\Big)}{\bar{n}(y)!\Gamma\Big(\alpha(y)\Big)}\right) \\
&= \prod_{z\in\mathcal{Z}} \left[\frac{n'(z)!\Gamma\Big(\alpha'(z)\Big)}{\Gamma\Big(\alpha'(z)+n'(z)\Big)} \prod_{y\in\mathcal{S}(z)} \frac{\Gamma\Big(\alpha(y)+\bar{n}(y)\Big)}{\bar{n}(y)!\Gamma\Big(\alpha(y)\Big)}\right] \\
&= \prod_{z\in\mathcal{Z}} \mathrm{DM}\left(\Big\{\bar{n}(y) : y \in \mathcal{S}(z)\Big\}; n'(z), \Big\{\alpha(y) : y \in \mathcal{S}(z)\Big\}\right),
\end{aligned} \tag{A.5}$$

on the domain $\Big\{\bar{n} \in {\mathbb{Z}_{\geq 0}}^{\mathcal{Y}} : \sum_{y\in\mathcal{S}(z)} \bar{n}(y) = n'(z), \quad \forall z \in \mathcal{Z}\Big\}$.