Theoretical and Computational Aspects of Mixture Models, with Applications to Empirical Bayes Methods

by

Sile Tao

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

STATISTICS

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

This thesis studies mixture models, in particular the estimation of mixing distributions and their applications to empirical Bayes prediction. The objectives are two-fold: to study the large-sample property of empirical Bayes estimators; to develop algorithms for the nonparametric estimation of mixing distributions as well as methods inspired by the Kiefer-Wolfowitz nonparametric maximum likelihood estimator.

Asymptotic optimality of empirical Bayes estimators is a topic that has been in past studied by various authors, starting from Robbins (1956), and continued by Deely and Zimmer (1976), Robbins (1964), and Rutherford and Krutchkoff (1969). They all worked in somewhat different settings, focusing not only on mixture models but the general empirical Bayes methodology. Moreover, these authors considered exclusively the squared loss in predictions. In this thesis, we establish asymptotic optimality for the empirical Bayes estimators; the results apply not only for the squared loss, but for a large class of convex loss functions. A consistency result of Bayes estimators for mixture models for a large class of convex loss functions is provided under mild conditions. Nowadays, decision problems involving alternative loss functions other than the squared loss are becoming increasingly popular. For instance, Mukherjee, Brown, and Rusmevichientong (2015) have recently applied a parametric empirical Bayes method to the so-called newsvendor problem involving a piecewise linear loss function. The last chapter of this thesis compares their methodology with one that is based on mixture models, and discusses the potential of the latter in this field.

The second part of the thesis is devoted to the estimation of mixing distribution in mixture models. Based on the breakthrough of Koenker and Mizera (2014), see also Dicker and Zhao (2014), Abadie and Kasy (2017), we propose four estimation methods/algorithms. Cutting-Plane Method, which for technical reasons comes last, is in fact an alternative algorithm for the Kiefer-Wolfowitz nonparametric maximum likelihood estimator studied by Koenker and Mizera. However, unlike their algorithm, the Cutting-Plane Method is also applicable in higher-dimensional parameter spaces. The same is true for the remaining three proposed methods. Projected Stochastic Gradient is capable of working in even higher dimensions but its convergence may be slow. Stochastic Average Approximation is generally much faster but in some versions, its estimation target differs from that of Kiefer-Wolfowitz nonparametric maximum likelihood estimator. This is even more true for Constraint Resampling, which is in fact an autonomous and novel estimation method; its properties, as well as those of other proposed methods are assessed via simulations and theoretical results. The penultimate chapter is devoted to facilitate the multivariate data-analytical applications of the developed algorithms. Nonparametric empirical Bayes methods are studied in the presence of explanatory variables. A nonparametric empirical Bayes regression model is later proposed. In contrast to some of the previous approaches, such a regression model has a very simple form and inherits most of theoretical properties of nonparametric empirical Bayes procedures. Unlike methods based on the partial linear model, the parameter estimation procedure is equivalent to solving a convex optimization problem in function space and can be efficiently solved by the proposed algorithms.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Ivan Mizera for his guidance, his unceasing encouragement and the freedom he has given me to find my own research path throughout my entire PhD program. He introduced me to empirical Bayes methods and mathematical optimization, two fascinating fields that continuously stimulate my interests. He has taught me many things from what are good research topics to what is the right order of doing them. I am privileged to be supervised by him and words fail to express my deepest regards towards him.

I would like to express my deepest gratitude to Dr. Keumhee Chough, Dr. Nicolas Guay, Dr. Linglong Kong, Dr. Brendan Pass and Dr. Juxin Liu for being a part of my thesis examining committee and for going over my thesis.

Moreover, I would like to thank to the group studies organized by Dr. Linglong Kong and Dr. Ivan Mizera, where I learned optimization and modern statistics. We had a lot of fun there. I would also like to thank to Training Consultant Center: the experience working there is unique and it reshapes my understanding of statistics.

Furthermore, I would like to thank all my friends. Chenzhe Diao, Chi Dong, Peng Liu, Michelle Michelle, Matthew Stephen, Wei Tu, Dengdeng Yu, Li Zhang and Ning Zhang gave insightful suggestions on the various stages of my research.

Special thanks go out to my parent, Yang Xia and Zhenmin Tao, and my grandmother, Tianmei Huang, for their love and unceasing support.

Table of Contents

Chapter 1

Introduction

This thesis concerns the problem of fitting mixture models and the applications of the mixture methodology arising in the empirical Bayes methodology.

In the first formal estimation of a mixture model, Pearson (1894) studied the ratio of "forehead breath" to body length of 1000 crabs and estimated a two component normal mixture model by the method of moments. He was assuming the underlying distribution of his data has the density

$$f_1\left(y\right)p_1 + f_2\left(y\right)p_2,$$

where f_i are component densities and p_i are component weights of the mixing distribution so that $p_1 + p_2 = 1$ and $p_i \geq 0$ for all i. This is an example of the simplest mixture model which has finite and known number of components. We are left to estimate the support points, the masses of the discrete mixing distribution and the parameters of the components.

A more general version of this scheme may have the number of components unknown; in such a case, we may put masses and support points together to get one mixing probability measure. Identifiability issues may push us to assume the components known up to nuisance parameters which are in turn fully determined by the mixing probability measure. The mixture distributions then have the density

$$\int_{\Theta} f\left(y|\theta\right) dF\left(\theta\right),$$

where Θ is the parameter space and the form of f is assumed to be known – up to θ, which in turn is controlled by a probability measure F, typically considered in quite extensive generality, in a "nonparametric way". It is this type of mixture models that is considered in this thesis.

In this thesis, we study the asymptotic optimality of empirical Bayes estimator for a large class of convex loss functions. Nowadays, there are many decision problems involving loss functions other than the squared loss. For example, recently Mukherjee et al. (2015) have applied a parametric empirical Bayes method to the newsvendor problem where a piecewise linear loss function is used to describe whether a vendor orders too much or too little. However, the majority of empirical Bayes literature consider only the quadratic loss and leave the questions of the large-sample properties under alternative loss functions not answered, e.g. Deely and Zimmer (1976), Robbins (1964) and Rutherford and Krutchkoff (1969) all focused on the asymptotic optimality of empirical Bayes estimator under the squared loss. In this thesis, we prove the asymptotic optimality for a large class of convex loss functions under some mild conditions. Moreover, we provide a consistency result of Bayes decision rule for mixture models for a large class of convex loss functions.

In the second part, we propose optimization algorithms for solving or approximating the Kiefer-Wolfowitz nonparametric maximum likelihood estimator. When the dimension of parameter space is one, we can use the discretization based method proposed by Koenker and Mizera (2014) by restricting the prior distribution on a fine grid and then applying modern interior-point method. Such an approach reduces the computational effort by several orders of magnitude by comparison to prior EM-based methods. However, as the dimension of the parameter space increases, the number of grid points required in the discretization method grows exponentially fast and the problem quickly becomes computationally intractable. Possible cure comes out of an insight that although the primal formulation of the Kiefer-Wolfowitz maximum likelihood estimation problem is infinite-dimensional, the objective function of the dual formulation is finite dimensional. For this reason, we focus on solving the dual problem and we propose four numerical algorithms for the dual problem which aim for solving or approximating Kiefer-Wolfowitz MLE when the dimension of parameter space is relatively high.

Chapter 5 of the thesis is to facilitate the multivariate data-analytic application of the developed algorithms. To this end, we study how to incorporate nonparametric empirical Bayes methods in the presence of explanatory variables and propose a novel regression model, called the nonparametric empirical Bayes regression. In the past, various approaches have been tried to generalize the empirical Bayes framework to regression problems, e.g. Cohen, Greenshtein, and Ritov (2013), Fay III and Herriot (1979), Jiang and Zhang (2010) and Koenker (2015). In contrast to some of the previous approaches, our new model has a very simple form and inherits most of theoretical properties of nonparametric empirical Bayes procedure. Furthermore, unlike the methods based on the partial linear model, the parameter estimation procedure in our proposed regression model is equivalent to solving a convex optimization problem in function space and it can be efficiently solved by the developed algorithms.

The thesis is organized as follows. In Chapter 2, we review empirical Bayes paradigms and discuss both parametric and nonparametric empirical Bayes; we also introduce the basic of convex optimization and some algorithms will be used later on. In Chapter 3, we show certain empirical Bayes procedures under alternative loss functions are asymptotically optimal. In Chapter 4, we propose four alternative algorithms for the Kiefer-Wolfowitz dual problem which aim for solving the maximum likelihood estimation problem when the dimension of parameter space is relatively high. In Chapter 5, we study how to incorporate nonparametric empirical Bayes methods in the presence of explanatory variables and propose the nonparametric empirical Bayes regression model. In Chapter 6, we study the newsvendor problem in the inventory management.

Chapter 2

Preliminaries

2.1 Oracle Predictions in Mixture Models

We are concerned with the problem of estimating $\theta_i \in \mathbb{R}^p$, based on the observations $y_1, ..., y_n$ and

$$y_i \overset{ind}{\sim} f_i(\cdot|\theta_i) \tag{2.1}$$

for $i = 1, ..., n$ and $y_i \in \mathbb{R}^p$. The provision of different error distributions f_i is to allow for inclusion of covariates; without them all $f_i = f$, which will be assumed in what follows, unless the contrary is explicitly specified.

The θ_i's are viewed as drawn independently from a distribution F. The performance of an estimator $\hat{\theta}_i$ is evaluated based on the mean squared loss function

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{\theta}_i - \theta_i\|^2.$$

Suppose the mixing distribution F is known, the problem of multiple prediction fits into the standard Bayesian paradigm: the mathematical model for the pair (y, θ) is identical to a Bayesian model in which the conditional distribution of y given the realized parameter $\Theta = \theta$ is $f(y|\theta)$ and F is the prior distribution on Θ. Given an i.i.d. sample $y_1, ..., y_n$ from the mixture density

$$f(y) = \int_{\mathit{\Theta}} f(y|\theta)\, dF(\theta),$$

we want to find Bayes estimators or decision rules so that the Bayes risk using the squared loss is minimized.

Assume the decision rule $d(y)$ is separable: $d(y) = (d_1(y_1), ..., d_n(y_n))$. It is sufficient to consider univariate Bayes decision problems. Let $\pi(\theta)$ be the density of the prior F, then the posterior density has the form

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}.$$

To obtain the Bayes estimator of θ, we choose a decision rule d to minimize the Bayes risk

$$\int_{\mathcal{Y}} \int_{\mathit{\Theta}} (\theta - d)^2 f(y|\theta)\pi(\theta)\, dy d\theta,$$

which is equivalent to minimize the expected posterior loss

$$\int_{\mathit{\Theta}} (\theta - d)^2 \pi(\theta|y) d\theta,$$

where $\mathcal{Y}$ is the sample space and $\mathit{\Theta}$ is the parametric space. Differentiating with respect to d and taking into account the posterior density integrates to 1, we obtain the Bayes estimator as the posterior mean

$$d(y) = \mathbb{E}(\theta|y) = \int_{\mathit{\Theta}} \theta \frac{f(y|\theta)dF(\theta)}{f(y)},$$

where $f(y) = \int_{\mathit{\Theta}} f(y|\theta)dF(\theta)$.

The following theorem states that the Bayes rule exists in most cases and it is optimal with respect to the selected loss function.

Theorem 2.1. *(Lehmann and Casella, 1998, page 228) Suppose the following assumptions hold for the problem of estimating θ with non-negative loss function $L(\theta, d)$.*

(a) There exists an estimator δ_0 with finite risk.

(b) For almost all y, there exists a value $\delta_F(y)$ minimizing

$$\mathbb{E}\left(L\left(\Theta, \delta(y)\right)|y\right).$$

Then $\delta_F(y)$ is a Bayes estimator.

2.2 Empirical Bayes Methodology Based on Mixture Models

Empirical Bayes methods estimate either the mixing distribution or the Bayes rule directly from the mixture. In this thesis, we start with a method based on estimating mixing distributions. Such a method has been proposed in 1956 by Robbins and then elaborated by Kiefer and Wolfowitz (1956).

Suppose the prior is completely unspecified, the method of Kiefer and Wolfowitz (1956) estimates the unknown prior via maximum likelihood. This amounts to

$$\min_{F \in \mathcal{F}} -\sum_{i=1}^{n} \log \int_{\mathbb{R}^p} f\left(y_i|\theta\right) dF(\theta), \tag{2.2}$$

where $\mathcal{F}$ is the class of all probability distribution functions on $\mathbb{R}^p$. Once the estimator F_n is obtained, we replace F in the Bayes decision rule by its MLE

F_n. For example, under the squared loss, the estimated Bayes decision rule has the form

$$\frac{\int_{\mathbb{R}^p} \theta f\left(y|\theta\right) dF_n(\theta)}{\int_{\mathbb{R}^p} f\left(y|\theta\right) dF_n(\theta)}.$$

For the large-sample properties, Kiefer and Wolfowitz (1956) studied the consistency of Kiefer-Wolfowitz MLE F_n and they showed the MLE F_n converge weakly to the true prior F in a sense that

$$\int \tilde{f} dF_n \to \int \tilde{f} dF$$

for every continuous and bounded real-valued function $\tilde{f}$. The consistency proof of Kiefer and Wolfowitz assumes the finiteness of Kullback-Leibler information which is hard to verify in the mixture model. Pfanzagl (1988) relaxed this assumption by merely requiring the continuity of likelihood function. A modern review of the consistency of Kiefer-Wolfowitz MLE can be found in Chen (2017).

The recent revival interest in empirical Bayes methods for compound decision problem is in computational methods, see, e.g. Brown (2008), Brown and Greenshtein (2009), Cohen et al. (2013), Dicker and Zhao (2014), Efron (2011, 2012, 2013), Greenshtein and Itskov (2014), Jiang and Zhang (2009, 2010). To numerically solve the problem (2.2), some discretization methods have to be used. However, the number of grid points required in the discretization grows exponentially fast as the dimension of the parameter space p increases. The discretization method is not suitable to solve the problem with a large p.

2.3 Alternative Methodologies in Empirical Bayes Prediction

For completeness, the alternative empirical Bayes approaches, such as Tweedie's formula and James-Stein estimator, are reviewed in this section. Readers familiar with empirical Bayes methodology may skip the basic exposition presented here.

2.3.1 Tweedie's Formula

Suppose the observed data y and the parameter of interest η are taken from $\mathbb{R}^p$ with $p \geq 1$. The parameter η has a prior density q and the real-valued likelihood function of η is taken from multivariate exponential family, as defined in DasGupta (2011),

$$f(y|\eta) = \exp\left\{\eta^T T\left(y\right) - A(\eta)\right\} f_0(y),$$

where η is is the natural or canonical parameter of the family, T is a known function and $A(\eta)$ is the cumulant generating function and $f_0(y) = f(y|\eta = 0)$.

Theorem 2.2. *Under the settings above, Bayes rule for η with the squared loss has the form*

$$\mathbb{E}\left(\eta|y\right) = \nabla \log\left(f\left(y\right)/f_0\left(y\right)\right). \tag{2.3}$$

Proof. Write $\lambda\left(y\right) = \log\left(f\left(y\right)/f_0\left(y\right)\right)$. Then

$$\begin{aligned}\nabla\lambda\left(y\right) &= \left(\frac{f_0\left(y\right)}{f\left(y\right)}\right)\left(\frac{\nabla f\left(y\right) f_0\left(y\right) - f\left(y\right)\nabla f_0\left(y\right)}{f_0^2\left(y\right)}\right) \\ &= \frac{\nabla f\left(y\right)}{f\left(y\right)} - \frac{\nabla f_0\left(y\right)}{f_0\left(y\right)}. \end{aligned} \tag{2.4}$$

By Lebesgue's dominated convergence theorem, differentiation under the integral sign is legitimate for the exponential family, so that

$$\begin{aligned}
\nabla f\left(y\right) &= \int \frac{d}{dy} f\left(y|\eta\right) q\left(\eta\right) d\eta \\
&= \int \left\{ e^{\eta^T T(y)-\psi(\eta)} \eta f_0\left(y\right) + e^{\eta^T T(y)-\psi(\eta)} \nabla f_0\left(y\right) \right\} q\left(\eta\right) d\eta \\
&= f_0\left(y\right) \int e^{\eta^T T(y)-\psi(\eta)} \eta q\left(\eta\right) d\eta + \nabla f_0\left(y\right) \int e^{\eta^T T(y)-\psi(\eta)} q\left(\eta\right) d\eta \\
&= f_0\left(y\right) \int e^{\eta^T T(y)-\psi(\eta)} \eta q\left(\eta\right) d\eta + \frac{\nabla f_0\left(y\right)}{f_0\left(y\right)} f\left(y\right). \qquad (2.5)
\end{aligned}$$

Substituting (2.5) into (2.4), we obtain

$$\begin{aligned}
\nabla \lambda\left(y\right) &= \frac{f_0\left(y\right) \int e^{\eta^T T(y)-\psi(\eta)} \eta q\left(\eta\right) d\eta}{f\left(y\right)} \\
&= \mathbb{E}\left(\eta|y\right).
\end{aligned}$$

□

The expression (2.3) is called Tweedie's formula and it was first provided by Robbins (1956). Efron calls such an expression Tweedie's formula because that Robbins "credits personal correspondence with Maurice Kenneth Tweedie for an extraordinary Bayesian estimation formula". In some literature, this formula is also referred to Robbins' formula but in this thesis we follow the terminology of Efron's paper. The formula (2.3) coincides for $p = 1$ with that derived in Efron (2011), who mentions a possibility of multivariate extension. As the latter is not readily available in the literature, we provide a multivariate version here.

According to Tweedie's formula, the Bayes rule depends directly on the marginal distribution which can be estimated by kernel density estimation or

Lindsey's method (Efron, 2011); therefore, it is in principle not necessary to estimate the prior density q.

Example 2.3. Suppose $y|\mu \sim N_p(\mu, \Sigma)$, where the covariance matrix Σ is known. Then the canonical parameter is $\eta = \Sigma^{-1}\mu$ and the function T has the form $T(y) = y$. From Tweedie's formula (2.3), we have

$$\begin{aligned}\mathbb{E}(\eta|y) &= \frac{d\log\left(f(y)/f_0(y)\right)}{dy} \\ &= \frac{d\log(f(y))}{dy} - \frac{d\log(f_0(y))}{dy} \\ &= \frac{d\log(f(y))}{dy} + \Sigma^{-1}y,\end{aligned}$$

where

$$f_0(y) = (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{y^T\Sigma^{-1}y}{2}\right\}.$$

Then Bayes rule for μ would be,

$$\mathbb{E}(\mu|y) = y + \Sigma\frac{d\log(f(y))}{dy}.$$

2.3.2 Parametric Empirical Bayes

Consider the setting defined in Section 2.1. If we can specify the parametric family of F but leave certain hyperparameters unknown and the hyperparameters eventually are estimated from the data, this is called parametric empirical Bayes. The first major work in this area was made by Efron and Morris (1975, 1977). The procedure is first writing out the marginal distribution and then obtain the estimators for all hyperparameters. As soon as the prior is specified, the standard Bayesian follows and we can compute the posterior expectation without any trouble.

In this section, James-Stein estimator is studied as an example of parametric empirical Bayes.

The approach leading to the James-Stein estimator assumes the observed density and the prior are both normal. To illustrate the idea, let us start with the univariate case. Consider $\theta \sim N(0, a)$ with a unknown and the error distribution $f(y|\theta)$ is $N(\theta, 1)$. Since we do not know the value of a in the prior $N(0, a)$, the standard Bayes approach cannot be used directly. However, we can follow empirical Bayes paradigm and extract the information about a from the marginal distribution of y.

It is not hard to see that the marginal distribution of y is again a normal distribution with mean 0 and variance $a + 1$. The Bayes estimator of θ_i is

$$\hat{\theta}_i = \mathbb{E}(\theta_i|y_i) = \left(1 - \frac{1}{\hat{a} + 1}\right) y_i.$$

Using the method of moments, we obtain a for an estimator

$$\hat{a} = \frac{\sum_{i=1}^n y_i^2}{n} - 1.$$

In the empirical Bayes, the unknown term $1/(a + 1)$ is unbiasedly estimated by $(n - 2)/\sum_{i=1}^n y_i^2$. This results is the James-Stein estimator

$$\hat{\theta}_i^{(JS)} = \left(1 - \frac{n-2}{\sum_{i=1}^n y_i^2}\right)_+ y_i,$$

where the notation $(\cdot)_+$ is defined as

$$(x)_+ := \max\{x, 0\}.$$

More generally, assume that $\theta_i \overset{iid}{\sim} N(M, A)$ and $y_i|\theta_i \overset{ind}{\sim} N(\theta_i, \sigma_0^2)$ with $i = 1, ..., n$ and $n \geq 4$, where the hyperparameters M and A are the mean and variance of the prior distribution. The marginal density of y_i is

$$y_i \sim N(M, A + \sigma_0^2)$$

and the posterior density

$$\theta_i|y_i \sim N(M + B(y_i - M), B\sigma_0^2),$$

where

$$B = \frac{A}{A + \sigma_0^2}.$$

Now the Bayes estimator of θ_i is

$$\hat{\theta}_i = M + B(y_i - M).$$

Although the values of A and B are unknown at the beginning, we can obtain the estimators from marginal density. Eventually, the James-Stein estimator acquires the form

$$\hat{\theta}_i^{(JS)} = \overline{y} + \left(1 - \frac{(n-3)\sigma_0^2}{S}\right)_+ (y_i - \overline{y}), \tag{2.6}$$

where $S = \sum_{i=1}^n (y_i - \overline{y})^2$.

The estimator (2.6) shrinks each observed value y_i toward sample mean $\overline{y}$. The amount of shrinkage depends on other observations. This fact might counter our intuition because each observation y_i is taken independently, but in most cases this type of shrinkage will reduce the total squared of error and improve the performance of the estimate of θ_i.

For $n \geq 3$, the James-Stein estimator always has smaller risk than MLE with the squared loss .

Theorem 2.4. *(James and Stein, 1961) For $n \geq 3$, the following is true that*

$$\mathbb{E}\left\{||\hat{\theta}^{(JS)} - \theta||^2\right\} < \mathbb{E}\left\{||\hat{\theta}^{(MLE)} - \theta||^2\right\}$$

for all θ.

2.4 Convex Optimization

For self-containedness, an overview of deterministic and stochastic optimization is provided, which will be used to develop algorithms to solve or approximate Kiefer-Wolfowitz MLE in later chapter. In particular, we discuss cutting-plane method, stochastic gradient method and sample average approximation. Readers familiar with optimization may skip this section.

2.4.1 Deterministic Convex Programming

Convex optimization is a subfield of mathematical optimization that studies the problems of minimizing convex functions over convex sets. The convexity makes optimization easier since the local minimum must be global minimum. Therefore, first-order conditions are sufficient for optimality (Rockafellar, 1993).

Subsections 2.4.1.1 and 2.4.1.2 follow the materials in Boyd and Vandenberghe (2004).

2.4.1.1 Terminology

Basic terminology in convex optimization is introduced here. In deterministic convex programming, we consider problems having the form

$$\min_{x \in \mathbb{R}^n} f_0(x) \tag{2.7}$$

$$\text{subject to} \qquad f_i(x) \leq 0,\ i = 1, ..., m,$$

$$h_i(x) = 0,\ i = 1, ..., m'$$

where the functions $f_0, f_1, ...f_m$ are convex and $h_1, ..., h_{m'}$ are affine. We call the function $f_0 : \mathbb{R}^n \to \mathbb{R}$ objective function or cost function. The inequalities $f_i(x) \leq 0$ are called the inequality constraints and the corresponding functions $f_i : \mathbb{R}^n \to \mathbb{R}$ the inequality constraint functions. The equations $h_i(x) = 0$ are called the equality constraints and $h_i : \mathbb{R}^n \to \mathbb{R}$ are the equality constraint functions. If there is no constraints, i.e., $m = m' = 0$, the problem is called unconstrained.

The set of points for which the objective and all constraint functions are defined is called the domain of the problem:

$$D = \bigcap_{i=1}^{m} \mathrm{dom} f_i \cap \bigcap_{i=1}^{m'} \mathrm{dom} h_i.$$

A point $x \in D$ is called feasible if it satisfies all the inequality constraints $f_i(x) \leq 0,\ i = 1, ..., m$ and all the equality constraints $h_i(x) = 0,\ i = 1, ..., m'$. An optimization problem is said to be feasible if there exists at least one feasible point, and infeasible otherwise. The set of all feasible points is called the feasible set or the constraint set.

The optimal value p^* of the problem is defined as

$$p^* = \inf_x \left\{ f_0(x) : f_i(x) \leq 0, i = 1, ..., m,\ h_i(x) = 0, i = 1, ..., m' \right\}.$$

The optimal value p^* is allowed to take on the extended values $\pm\infty$. If the problem is infeasible, we have $p^* = \infty$. If there is a sequence of feasible points $\{x_k\}$ such that $\lim_{k\to\infty} f_0(x_k) = -\infty$, then $p^* = -\infty$ and such a problem is called unbounded below.

We say x^* is an optimal point or minimizer to the problem (2.7), if x^* is feasible and $f_0(x^*) = p^*$. The set of all optimal points is the optimal set denoted

$$X_{opt} = \left\{ x : f_i(x) \leq 0, i = 1, ..., m,\ h_i(x) = 0, i = 1, ..., m', f_0(x) = p^* \right\}.$$

2.4.1.2 Lagrange Duality in Convex Optimization

Taking the constraints into account, we write the objective function with a weighted sum of the constraints: the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m'} \to \mathbb{R}$ associated with the problem 2.7 is defined to be

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{m'} \nu_i h_i(x).$$

The vectors λ and ν are called Lagrange's multiplier vectors. We define the Lagrange dual function as the minimum value of the Lagrangian over x:

$$g(\lambda, \nu) = \inf_{x\in D} L(x, \lambda, \nu) = \inf_{x\in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{m'} \nu_i h_i(x) \right).$$

If the Lagrangian is unbounded below in x, the dual function takes the value $-\infty$.

We use the curled inequality symbol $\succeq$ (and its strict form $\succ$) to denote componentwise inequality (and strict inequality, respectively) between vectors.

It can be shown (Boyd and Vandenberghe, 2004, page 216) that for $\lambda \succeq 0$ and any ν, the dual function gives the lower bound on the optimal value p^* of the problem (2.7)

$$g(\lambda, \nu) \leq p^*. \tag{2.8}$$

A natural question is: what is the greatest lower bound that can be obtained from the dual function? This leads to the optimization problem

$$\max_{\lambda,\nu} g(\lambda, \nu) \tag{2.9}$$

$$\text{subject to } \lambda \succeq 0,$$

which is called the Lagrange dual problem associated with the problem (2.7); the original problem (2.7) is then called the primal problem.

If we denote the optimal value of the dual as d^*, it can be shown that $d^* \leq p^*$. The equality does not hold in general, but if the primal problem is convex with the equality constraints $Ax = b$, we usually have strong duality $d^* = p^*$ under some mild conditions. One of those is called Slater's condition.

Before we state Slater's condition, we need to introduce some related concepts. The set of all affine combinations of points in some set $C \subset \mathbb{R}^n$ is called the affine hull of C and and denoted $\operatorname{aff}(C)$:

$$\operatorname{aff}(C) := \left\{ \sum_{i=1}^{k} \theta_i x_i : x_1, ..., x_k \in C, \sum_{i=1}^{k} \theta_k = 1 \right\}.$$

We define the relative interior of the set C, denoted relint (C), as its interior relative to aff (C):

$$\mathrm{relint}(C) := \left\{x \in C | B(x,r) \bigcap \mathrm{aff}(C) \subseteq C, \text{ for some } r > 0\right\},$$

where $B\left(x,r\right) = \{y : \|y - x\| \leq r\}$.

Slater's condition says that there exists an $x \in \mathrm{relint}(D)$, such that each inequality is strictly satisfied:

$$f_i(x) < 0,\ i = 1, ..., m, \quad Ax = b.$$

Theorem 2.5. *(Slater's theorem) If the problem is convex of the form*

$$\min f_0(x)$$

$$\textit{subject to } f_i(x) \leq 0, \quad i = 1, ..., m,$$

$$Ax = b,$$

with $f_0, ..., f_m$ convex and satisfies Slater's condition, then strong duality holds.

For any optimization problem with differentiable objective and constraint functions satisfying strong duality, any pair of primal and dual optimal points, say $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu})$, must satisfy the Karush–Kuhn–Tucker (KKT) conditions:

$$f_i(\tilde{x}) \leq 0,\ i = 1, ..., m,$$

$$h_i(\tilde{x}) = 0,\ i = 1, ..., m',$$

$$\tilde{\lambda}_i \geq 0,\ i = 1, ..., m,$$

$$\tilde{\lambda}_i f_i(\tilde{x}) = 0,\ i = 1, ..., m,$$

$$\nabla f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{i=1}^{m'} \tilde{\nu}_i \nabla h_i(\tilde{x}) = 0.$$

Moreover, when the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal.

Theorem 2.6. *If a convex optimization problem with differentiable objective and constraint function satisfies Slater's condition, then the KKT conditions provide necessary and sufficient conditions for optimality: x is optimal if and only if there are (λ, ν) that, together with x, satisfy the KKT conditions.*

2.4.1.3 Cutting-plane Methods

Cutting-plane methods are a class of methods that solve general convex and quasiconvex optimization problems by iteratively refining the feasible set. Usually these methods are "less efficient for problems to which interior-point methods apply" (Boyd and Vandenberghe, 2007), but cutting-plane methods do not require evaluating the objective and all the constraint functions, as well as as their first and second derivatives at each iteration. This make cutting-plane methods attractive for problems with a very large number of constraints. The introduction provided in this subsection is based on the materials of Boyd and Vandenberghe (2007).

The goal of cutting-plane method is to find a point in a convex set $X \subset \mathbb{R}^n$, known as the target set, or, in some cases, to determine that X is empty. In an optimization problem, the target set X can be taken as the set of optimal points for the problem, and our goal is to find an optimal point for the problem.

We do not have direct access to any description of the target set X, otherwise the optimization problem is solved already. Instead, for any query point $x \in \mathbb{R}^n$, we have a piece of information, called *oracle*, which tells us either $x \in X$ (in which case we are done), or returning a separating hyperplane between x and X, *i.e.*, $a \neq 0$ and b such that

$$a^T z \leq b \text{ for } z \in X, \qquad a^T x \geq b.$$

This hyperplane is called a *cutting-plane*, or *cut*, since it cuts or eliminates the halfspace $\left\{z : a^T z > b\right\}$ from our search: no point in this halfspace could be in the target set X.

Now let us discuss the details of how to construct cutting-planes in inequality constrained problems. Consider an inequality constrained problem

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

$$\text{subject to} \qquad f_i(x) \leq 0, \quad i = 1, ..., m,$$

where $f_0, ..., f_m$ are convex. The target set X is the optimal set.

Definition 2.7. (Subgradient) For a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$, g is a subgradient of f at x if

$$f(y) \geq f(x) + g^T (y - x) \qquad \text{for all } y.$$

The set of all subgradients of f at x is called the subdifferential of f at x, denoted as $\partial f(x)$.

To find a cutting-plane for this problem at the query point x, we first check for feasibility. If x is not feasible, then we compute the subgradient g_j of any violated constraint f_j at x (If f_j is differentiable, then $g_j = \nabla f_j(x)$.) and construct a cut as

$$f_j(x) + g_j^T(z - x) \leq 0.$$

The above halfspace defines a cutting-plane, since any optimal point $z \in X$ satisfies the linear inequality: From the definition of subgradient

$$f_j(x) + g_j^T(z - x) \leq f_j(z) \qquad \text{for all } z.$$

Any optimal point must be feasible: $f_j(z) \leq 0$. Therefore, all the optimal points lie on the one side of the hyperplane and this gives a cutting-plane.

Now suppose that the query point x is feasible. Take $g_0 \in \partial f_0(x)$. If $g_0 = 0$, then from the definition of subgradient we have $f_0(x) \leq f_0(z)$ for all z. The query point x must be an optimal and we are done. So we assume $g_0 \neq 0$. In this case we construct a cutting-plane as

$$g_0^T(z - x) \leq 0.$$

Again, we need to justify all the optimal points lie on the one side of the hyperplane: For any $z \in X$, we must have

$$g_0^T(z - x) \leq f_0(z) - f_0(x) \leq 0.$$

Without loss of generality we assume the target set X is contained in a polyhedron $\mathcal{P}_0 = \{z : Cz \leq d\}$ which is known. Now suppose the algorithm does not stop in k steps: none of the query points were announced by the oracle to be in the target set X, then we have k cutting-planes

$$a_i^T z \leq b_i, \quad i = 1, ..., k.$$

From the construction of the algorithm, every point in target set must satisfy these inequalities:

$$X \subset \mathcal{P}_k = \{z : Cz \leq d, a_i^T z \leq b_i, \quad i = 1, ..., k\}.$$

To find the minimizer, we only need to consider points in the *localization polyhedron* $\mathcal{P}_k$.

If $\mathcal{P}_k$ is empty, then the target set X is empty: the problem has no solution and we stop. If it is not, we choose a new query point $x^{(k+1)}$ in $\mathcal{P}_k$. If $x^{(k+1)} \in X$, then we are done. If not, the oracle generates a new cutting-plane and we can update the localization polyhedron by adding the new inequality.

Algorithm 1 Cutting-plane algorithm

```
Ensure: an initial polyhedron $\mathcal{P}_0 = \{z : Cz \leq d\} \supset X$
k ← 0
repeat
    Choose a point $x^{(k+1)}$ in $\mathcal{P}_k$
    if $x^{(k+1)} \in X$ then
        stop
    else
        update $\mathcal{P}_k$ by adding the new inequality
            $\mathcal{P}_{k+1} \leftarrow \mathcal{P}_k \cap \left\{z : a_{k+1}^T z \leq b_{k+1}\right\}$
    end if
    if $\mathcal{P}_{k+1} = \emptyset$ then
        stop
    end if
    $k \leftarrow k + 1$
```

The critical step that how to choose the next query point $x^{(k+1)}$ in inside the current localization polyhedron $\mathcal{P}_k$ is not fully specified. Different ways of choosing query points lead to different cutting-plane algorithms.

There are a large class of cutting-plane methods, called interior point cutting-plane methods, selecting query points in a way that the size of $\mathcal{P}_{k+1}$ is as small as possible, or equivalently, the new cut removes irrelevant points as many as possible from the current polyhedron $\mathcal{P}_k$. When we query the oracle at $x^{(k+1)}$, we do not know which halfspace will be returned; we only know $x^{(k+1)}$ will be in the excluded halfspace. No matter which halfspace is returned by the oracle, we want a good reduction in the size of localization polyhedron. This suggests that we should choose $x^{(k+1)}$ to be some kind of center of $\mathcal{P}_k$. For this choice of query point, we can cut away a good portion of $\mathcal{P}_k$ no matter which halfspace is returned by the oracle. Many modern cutting-plane algorithms are within this class: the center of gravity algorithm, maximum volume ellipsoid cutting-plane method, Chebyshev center cutting-plane method and analytic center cutting-plane method.

In the very first paper of cutting-plane methods of Kelley (1960), the query point $x^{(k)}$ is chosen as the optimal solution to the current polyhedron approximation $\mathcal{P}_k$, instead of all kinds of centers. For certain class of problems, Kelley's method can outperform interior point cutting-plane methods (du Merle, Goffin, and Vial, 1998): If "the optimal point of the current polyhedral approximation of the problem turns out to be optimal for the original problem itself, then [Kelley's algorithm] terminates at once with a provable optimal solution". Other interior point cutting-plane methods are slower by design, since they avoid the optimal point of polyhedral approximation in the next iteration.

An example follows. Consider a convex optimization problem

$$\min_{x\in\mathbb{R}^n} f_0\left(x\right)$$

$$\text{subject to} \qquad f_i\left(x\right) \leq 0, \quad i = 1, ..., m.$$

Apply any basic cutting-plane methods. If the algorithm does not stop at k step and the query point $x^{(k)}$ is infeasible, then the preceding discussion enable us to construct cuts as

$$f_j\left(x^{(k)}\right) + g_j^T\left(z - x^{(k)}\right) \leq 0$$

and the index j can be any violated constraint

$$j \in \left\{i : f_i\left(x\right) > 0, \text{ for some } i = 1, ..., m\right\}.$$

Common choices for the indices are the most violated inequality $\arg\max_i f_i\left(x\right)$, any violated inequality or simply all violated inequalities.

2.4.2 Stochastic Convex Programming

Stochastic programming is mathematical programming with stochastic elements present in objective function or constraints. Generally a stochastic programming problem has the form

$$\min_{x\in X} \mathbb{E} f_0\left(x, Y\right)$$

$$\text{subject to} \qquad \mathbb{E} g_i\left(x, Y\right) \leq 0, \quad i = 1, ..., m,$$

$$\mathbb{E}h_i\left(x, Y\right) = 0, \quad i = 1, ..., m',$$

where X is a nonempty closed subset of $\mathbb{R}^n$, Y is a random variable with probability distribution P which is supported on a set $\mathcal{Y} \subset \mathbb{R}^p$ and all the functions f_0, f_i and h_i map $X \times \mathcal{Y}$ to $\mathbb{R}$. Such a problem is called convex if the following conditions hold: (1) f_0 and g_i are all convex in x for each y, $i = 1, ..., m$; (2) h_i is affine in x for each y, $i = 1, ..., m'$; (3) X is a convex set. Except for very limited cases, the stochastic terms do not have analytical expressions and therefore difficult to evaluate. If the dimension p is relatively large, the discretization based methods will not perform well, since the number of grid points grows exponentially fast. Later on we will discuss two commonly used algorithms in the field: stochastic gradient method and sample average approximation. References for stochastic programming are Kall, Wallace, and Kall (1994), Shapiro, Dentcheva, and Ruszczyński (2009).

In fact, many statistical problems are stochastic programming problems. Consider a Bayes parameter estimation problem under the loss $L\left(\theta, d\right)$, where θ is the parameter of interest and d is the Bayes decision rule for θ. Let $f\left(y|\theta\right)$ be the sampling distribution. We are looking for a decision d minimizing the Bayes risk:

$$\min_d \mathbb{E}_\theta \mathbb{E}_Y L\left(\theta, d\right)$$

or equivalently

$$\min_d \mathbb{E}_{\theta|y}\left(L\left(\theta, d\right)\right).$$

It is well known that under the squared loss $L\left(\theta, d\right) = \left(\theta - d\right)^2$, the Bayes rule is the posterior mean. If the prior distribution is known, the computation of Bayes decision rule is usually straightforward. However, for alternative loss functions other than the squared loss, we usually do not have explicit expression and we need numerical algorithms to solve these problems.

2.4.2.1 Stochastic Gradient Method

Stochastic gradient method is a gradient method replacing the true gradient with a sample approximation in each iteration and the algorithm economizes on the computational cost of gradient at every iteration. For these reasons, stochastic gradient method is widely used in large-scale machine learning problems (Bottou, 2010). This overview is based on a survey of Boyd and Mutapcic (2006).

Consider an unconstrained minimization problem of a convex differentiable objective function $f : \mathbb{R}^n \to \mathbb{R}$. Let $x^{(k)}$ denote the kth iterate, $\alpha_k > 0$ the kth step size. The classical gradient methods use the update

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right), \qquad k \in \mathbb{N}.$$

The idea of stochastic gradient method is the following: If the true gradient $\nabla f\left(x^{(k)}\right) = \mathbb{E}\left(\tilde{g}; x^{(k)}\right)$ for some random vector $\tilde{g}$, we can replace it by a sample approximation. In such a way, the update steps can be done efficiently. With some carefully chosen step sizes, the algorithm converges.

To see an example, let us consider a linear regression problem, which has the form

$$\min_{\beta \in Q} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2,$$

where Q is a nonempty compact convex subset of $\mathbb{R}^p$ and the sample size n can be very large. Due to the appearance of Q, we may not have analytical solution. Another difficulty here is the large sample size n: If we are thinking projected gradient based methods, then in each iteration we need to compute the gradient $g\left(x,y;\beta\right)=\left(-2/n\right)\sum_{i=1}^{n}\left(y_i-x_i^T\beta\right)x_i$. The summation over n terms here is computationally expensive. Instead, we may try some stochastic gradient methods. First, we can interpret the deterministic regression problem as a stochastic optimization problem:

$$\min_{\beta\in Q}\mathbb{E}_{P_n}\left(\left(Y-X^T\beta\right)^2\right),$$

where P_n is the empirical distribution function. Let $\beta^{(k)}$ denote the kth iterate, $\alpha_k>0$ the kth step size and $\xi_1^{(k)},...,\xi_N^{(k)}$ a random sample of (X,Y) with $N\ll n$. Stochastic gradient method uses the update

$$\beta^{(k+1)}=\beta^{(k)}-\frac{\alpha_k}{N}\sum_{i=1}^{N}g\left(\xi_i^{(k)};\beta^{(k)}\right).$$

As we can see, in the update steps, it only uses a small fraction of training data and it can be much more efficient than the gradient descent methods.

In fact, the requirement of f to be differentiable is unnecessary. It can be relaxed by replacing the gradient by its subgradient. Such a method is known as stochastic subgradient method. See Boyd and Mutapcic (2006) for details.

2.4.2.2 Sample Average Approximation

Suppose we have a stochastic optimization problem with expectations involved: either in the objective function or in the constraints (or both). The underlying probability distributions are known but the expectations are usually difficult to evaluate. In that case, it is natural to think about some sampling techniques and replace the true probability distributions with their sample estimates. In the literature on stochastic programming, e.g. Shapiro et al. (2009), Shapiro (2010), this is referred to as the sample average approximation problem, and in machine learning as the empirical mean optimization.

To be precise, let us consider the following stochastic problem:

$$\min_{x\in X} \mathbb{E}\left(f\left(x,Y\right)\right).$$

Here X is a nonempty closed subset of $\mathbb{R}^n$, Y is a random variable with probability distribution P which is supported on a set $\mathcal{Y} \subset \mathbb{R}^p$ and $f : X \times \mathcal{Y} \to \mathbb{R}$. Suppose we have a sample $\xi_1, ..., \xi_N$ of the random variable Y, for example, they can be generated by Monte Carlo sampling techniques, then the sample average approximation problem has the form

$$\min_{x\in X} \frac{1}{N}\sum\nolimits_{i=1}^{N} f\left(x,\xi_i\right).$$

With some regularity conditions, we expect the sample based optimal value $\hat{f}_N$ and optimal solution $\hat{x}_N$ get close to the true ones as the sample size increases. In this way, we solve the stochastic programming problem by randomization techniques.

Chapter 3

Asymptotic Optimality of Empirical Bayes Predictions Based on Mixture Models

In this chapter, we prove the convergence of Bayes risk for certain empirical Bayes procedures and show the consistency of Bayesian decision rules under a large class of convex loss functions.

Nowadays, there are many decision problems involving alternative loss functions other than the squared loss. For example, in robust regression, check loss and Huber's loss are widely used to avoid the dangers posed by outliers. In certain applications, alternative loss functions are selected simply because they have real-world interpretations: for the newsvendor problem in the inventory management, a piecewise linear loss function is used to describe the loss whether a vendor orders too much or too little. However, the majority of empirical Bayes literature only consider the quadratic loss and leave the questions of asymptotic optimality under alternative loss functions not answered.

In a fundamental paper, Robbins (1964) proved an empirical Bayes procedure is asymptotically optimal. The paper gives conditions under which a consistent estimator of the Bayes procedure is asymptotically optimal. However, one of the conditions does not hold for the regular quadratic loss function when the parameter space is unbounded. Deely and Zimmer (1976) recognize this fact and obtain asymptotic optimality property for quadratic loss function under some weaker conditions than Robbins. Martin (2015) generalizes the results of Deely and Zimmer to general loss functions and proves an empirical Bayes procedure is asymptotically optimal, but Martin's result relies on a strong assumption that Lebesgue's dominant theorem holds. In this chapter, we prove the asymptotic optimality for a large class of convex loss functions under some mild conditions.

3.1 Asymptotic Optimality

3.1.1 Notation and Theorem

Suppose the parameter of interest is θ, taking values in $\mathbb{R}$, an observable random variable Y takes values in $\mathcal{Y}$, an action a takes values in set A, a decision rule t maps $\mathcal{Y}$ into A, a loss function $L(\theta, a) \geq 0$, and a prior distribution F on $\mathbb{R}$. It is assumed that Y has a conditional probability density function with respect to some measure μ on $\mathcal{Y}$ denoted by $f(y|\theta)$. The Bayes risk associate with a decision rule t is given by

$$r(F, t) = \int_{\mathbb{R}} \int_{\mathcal{Y}} L(\theta, t(y)) f(y|\theta) \, d\mu(y) \, dF(\theta).$$

First, let us introduce some concepts given by Robbins (1964).

Definition 3.1. (Empirical Bayes procedure) Let $t_F(y)$ denote the Bayes rule with the true prior distribution F. A function $t_n(y) := t_n(y_1, ..., y_n; y)$ based on the past and present observations and taking values in action space is called empirical Bayes decision procedure if

$$t_n(y) \xrightarrow{p} t_F(y) \qquad \forall y.$$

That is, for any y and $\epsilon > 0$,

$$\mathbb{P}\left\{\omega : \left|t_n\left(Y_1(\omega), ..., Y_n(\omega); y\right) - t_F(y)\right| > \epsilon\right\} \to 0, \quad \text{as}\, n \to \infty.$$

Definition 3.2. (Asymptotically optimal) If an empirical Bayes procedure t_n has the property that

$$r(F, t_n) \to r(F, t_F), \quad n \to \infty,$$

then t_n is said to be asymptotically optimal.

In the following, we assume the loss function L is unbounded. For a bounded loss function, asymptotic optimality can be obtained by using Robbins' method.

We make the following assumptions:

(A1) $\sup_{y \in \mathcal{Y}} t_F(y) < \infty$,

(A2) $t_n(y)$ is a consistent estimator of $t_F(y)$,

(A3) There exists a representing function ψ satisfying

$$L(\theta, a) = \psi(|\theta - a|),$$

where ψ is an unbounded increasing convex function on $\mathbb{R}^+ \cup \{0\}$.

By assumption (A1) and (A2), there exists $0 < M, N < \infty$ such that $n > N$ implies $|t_n(y)| < M$. To obtain the asymptotic optimality, we further need

(A4) $\mathbb{E}_F \psi(M + |\theta|) < \infty$.

Note that this assumption is stronger than $\mathbb{E}_F \psi(|\theta|) < \infty$, since ψ is increasing on $\mathbb{R}^+$. Therefore, we only require (A4) holds.

Remark 3.3. Due to a result of Karlin and Rubin (1956), if $f(y|\theta)$ is a member of exponential family and the loss function satisfies (A3), then the Bayes rule $t_F(y)$ is increasing in y. If $\mathcal{Y} = \mathbb{R}$, then (A1) requires $\lim_{y\to\infty} t_F(y) < \infty$.

Theorem 3.4. *Under the assumptions (A1)-(A4), the modified decision rule*

$$\tilde{t}_n(y) = \begin{cases} M & \text{if } t_n(y) > M \\ t_n(y) & \text{otherwise} \\ -M & \text{if } t_n(y) < -M \end{cases}$$

is asymptotically optimal.

Proof. From the construction of $\tilde{t}_n$ and the assumption (A2), we have $\lim_n \tilde{t}_n(y) = t_F(y)$. Write $h(\theta) := \psi(M + |\theta|)$. Then

$$\begin{aligned} L\left(\theta, \tilde{t}_n(y)\right) &= \psi\left(|\tilde{t}_n(y) - \theta|\right) \\ &\le \psi\left(|\tilde{t}_n(y)| + |\theta|\right) \\ &\le h(\theta). \end{aligned}$$

By Lebesgue's dominated convergence theorem, we have

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}_{\mu,F}\left(L\left(\theta,\tilde{t}_n\left(y\right)\right)\right) &= \mathbb{E}_{\mu,F}\left(\lim_{n\to\infty}L\left(\theta,\tilde{t}_n\left(y\right)\right)\right)\\ &= \mathbb{E}_{\mu,F}\left(L\left(\theta,t_F\left(y\right)\right)\right).\end{aligned}$$

The last equality is due to the facts that (a) every convex function is continuous; (b) $t_n\left(y\right)$ is a consistent estimator of $t_F\left(y\right)$. □

For a real number k, let $\lceil k\rceil$ denote the ceiling function which maps k to the least integer.

Corollary 3.5. *(k-th power absolute distance loss) Consider the loss function $L\left(\theta,a\right)=\left|\theta-a\right|^k$, where $k\geq 1$. Assume the following conditions hold:*

(1) $\sup_{y\in\mathcal{Y}}t_F\left(y\right)<\infty$,

(2) $t_n\left(y\right)$ is a consistent estimator of $t_F\left(y\right)$,

(3) $\mathbb{E}_F\left(\left|\theta\right|^{\lceil k\rceil}\right)<\infty$.

Then the modified Bayes rule $\tilde{t}_n\left(y\right)$ is asymptotically optimal.

Proof. To apply Theorem 3.4, we only need to justify (A4): $\mathbb{E}_F\left(\left(M+\left|\theta\right|\right)^k\right)<\infty$. Without loss of generality, we can assume $M\geq 1$ and then

$$\mathbb{E}_F\left(\left(M+\left|\theta\right|\right)^k\right)\leq\mathbb{E}_F\left(\left(M+\left|\theta\right|\right)^{\lceil k\rceil}\right).$$

It is sufficient to show $\mathbb{E}_F\left(\left(M+\left|\theta\right|\right)^{\lceil k\rceil}\right)$ is finite. Using binomial expansion, we have

$$\mathbb{E}_F\left(\left(M+\left|\theta\right|\right)^{\lceil k\rceil}\right) = \sum_{m=0}^{\lceil k\rceil}\begin{pmatrix}\lceil k\rceil\\ m\end{pmatrix}M^{\lceil k\rceil-m}\mathbb{E}_F\left(\left|\theta\right|^m\right).$$

By Jensen's inequality, if the pth-moment $\mathbb{E}\left(|Z|^p\right)$ is finite, then all lower moments must be finite: For $m < p$,

$$\left(\mathbb{E}\left(|Z|^m\right)\right)^{p/m} \leq \mathbb{E}\left(\left(|Z|^m\right)^{p/m}\right) = \mathbb{E}\left(|Z|^p\right).$$

(Because $|z|^p$ is convex for $p \geq 1$.) From assumption (3), we know $\mathbb{E}_F\left(|\theta|^{\lceil k \rceil}\right)$ is finite, so all lower moments must be finite and the assumption (5) in Theorem 3.4 holds. □

Example 3.6. (Check loss) For $\tau \in (0,1)$, pinball loss is represented by

$$\psi(r) = \begin{cases} -(1-\tau) r & \text{if } r < 0 \\ \tau r & \text{if } r > 0 \end{cases}.$$

Assume the following conditions hold:

(1) $\sup_{y \in \mathcal{Y}} t_F(y) < \infty$,

(2) $t_n(y)$ is a consistent estimator of $t_F(y)$,

(3) $\mathbb{E}_F(|\theta|) < \infty$.

Then the modified Bayes rule $\tilde{t}_n(y)$ is asymptotically optimal.

Example 3.7. (Huber loss) Huber loss is represented by

$$\psi(r) = \begin{cases} (r)^2/2 & \text{if } |r| < c \\ c\left(|r| - c/2\right) & \text{if } |r| > c \end{cases},$$

for some constant $c > 0$. Assume the following conditions hold:

(1) $\sup_{y \in \mathcal{Y}} t_F(y) < \infty$,

(2) $t_n(y)$ is a consistent estimator of $t_F(y)$,

(3) $\mathbb{E}_F\left(|\theta|\right) < \infty$.

Then the modified Bayes rule $\tilde{t}_n\left(y\right)$ is asymptotically optimal.

3.1.1.1 The Squared Loss Function

For the regular squared loss function $L\left(\theta, a\right) = \left(\theta - a\right)^2$, a bound of $|t_n\left(y\right)|$ can be found explicitly. We make the following assumptions:

(B1) f and F are unimodal and symmetric,

(B2) $\mathbb{E}\left(Y^2\right) < \infty$ and $\mathbb{E}_F\left(\theta^2\right) < \infty$,

(B3) $t_n\left(y\right)$ is a consistent estimator of $t_F\left(y\right)$.

The proof of the following result is very similar to Deely and Zimmer (1976).

Theorem 3.8. *Under the squared loss and the assumptions above, the modified decision rule of t_n:*

$$\tilde{t}_n\left(y\right) = \begin{cases} |y| + |\bar{y}_n| & \textit{if } \ t_n\left(y\right) > |y| + |\bar{y}_n| \\ t_n\left(y\right) & \textit{otherwise} \\ -|y| - |\bar{y}_n| & \textit{if } \ t_n\left(y\right) < -|y| - |\bar{y}_n| \end{cases}$$

is asymptotically optimal.

Proof. Due to Verbeek (1973), if the assumption (B1) holds and θ is a location parameter, then

$$|\mathbb{E}\left(\theta|y\right)| \leq |y| + |\lambda|,$$

where $\lambda = \mathbb{E}\left(Y\right)$. By assumption (B1), (B3) and the law of large numbers, we have

$$\tilde{t}_n\left(y\right) \xrightarrow{p} \mathbb{E}\left(\theta|y\right).$$

Let

$$h_n\left(y,\theta\right) := \left(|y| + |\bar{y}_n|\right)^2 + 2|\theta|\left(|y| + |\bar{y}_n|\right) + \theta^2,$$

then

$$L\left(\theta,\tilde{t}_n\right) = \left(\tilde{t}_n\left(y\right) - \theta\right)^2 \leq h_n\left(y,\theta\right).$$

Now

$$h_n\left(y,\theta\right) \overset{p}{\to} \left(|y| + |\lambda|\right)^2 + 2|\theta|\left(|y| + |\lambda|\right) + \theta^2$$

and

$$\begin{aligned}\mathbb{E}_{\mu,F}\left(h_n\left(y,\theta\right)\right) &= \mathbb{E}\left(Y^2\right) + 2\mathbb{E}\left(|Y|\right)\mathbb{E}\left(\bar{Y}_n\right) + \mathbb{E}\left(\bar{Y}_n^2\right) \\ &\quad +2\mathbb{E}\left(|Y||\theta|\right) + 2\mathbb{E}\left(|\bar{Y}_n|\right)\mathbb{E}\left(|\theta|\right) + \mathbb{E}\left(\theta^2\right).\end{aligned}$$

Under assumption (B2),

$$\lim_{n\to\infty}\mathbb{E}_{\mu,F}\left(h_n\left(y,\theta\right)\right) = \mathbb{E}_{\mu,F}\left(\lim_{n\to\infty}h_n\left(y,\theta\right)\right) < \infty.$$

Finally, by the generalized Lebesgue dominated convergence theorem, the result follows. □

3.1.2 Estimators Based on Kiefer-Wolfowitz MLE

We turn now to Kiefer-Wolfowitz MLE based estimators, that is, $t_n\left(y\right) := t_{F_n}\left(y\right)$, where F_n is the Kiefer-Wolfowitz MLE. Suppose all the regularity conditions in Kiefer and Wolfowitz (1956) hold, we have $F_n \Rightarrow F$. Unfortunately weak convergence in general does not guarantee the consistency of $t_n\left(y\right)$. However, for certain types of loss functions, weak convergence is sufficient to give the consistency. In those cases, the consistency assumption can be replaced by the list of regularity conditions in Kiefer and Wolfowitz (1956).

Theorem 3.9. *(Squared loss) For the squared loss function $L(\theta, a) = (\theta - a)^2$ and $y|\theta \sim N(\theta, 1)$, if all the regularity conditions in Kiefer and Wolfowitz (1956) hold, then the estimated Bayes rule $t_n(y) = \mathbb{E}_{F_n}(\theta|y)$ is a consistent estimator of $t_F(y)$.*

Proof. Under the squared loss, the Bayes estimator has the form

$$\mathbb{E}_{F_n}(\theta|y) \quad = \quad \left(\int_{\mathbb{R}} \theta\varphi(y-\theta)\, dF_n(\theta)\right) / f(y).$$

By L'Hôpital's rule, we have $\lim_{\theta\to\infty} \theta\varphi(y-\theta) = \lim_{\theta\to-\infty} \theta\varphi(y-\theta) = 0$ and then $\theta\varphi(y-\theta)$ is continuous and bounded function of θ. From the definition of weak convergence, the result follows. □

For convex loss functions with bounded derivatives, Kiefer-Wolfowitz MLE based estimator $t_n(y)$ is consistent.

Theorem 3.10. *(Huber, 2011, page 54) Let $\gamma(a;\theta) = \partial L(\theta, a)/\partial a$ be a monotone increasing, but not necessarily continuous, function in a that takes values of both signs and F the true prior distribution function. Then the estimator T of location, defined by*

$$\int_{\mathbb{R}} \gamma\left(\theta - T\left(\tilde{F}\right)\right) d\tilde{F}(\theta) = 0,$$

is weakly continuous at F if and only if γ is bounded and $T(F)$ is unique.

Theorem 3.11. *(Consistency) For convex loss functions with bounded derivatives, if all the regularity conditions in Kiefer and Wolfowitz (1956) hold, then* *the Kiefer-Wolfowitz MLE based estimator $t_n(y)$ is a consistent estimator of $t_F(y)$.*

Proof. The second derivative of a convex function is always nonnegative, so its first derivative is an increasing function. The result is a consequence of Theorem 3.10. □

For Kiefer-Wolfowitz MLE based estimator, the assumption (A2) can be replaced by:

(A2′) $\partial L(\theta, a)/\partial a$ is bounded.

Theorem 3.12. *(Asymptotic optimality) Under the assumptions (A1), (A2′), (A3) and (A4), the modified Kiefer-Wolfowitz MLE based estimator $\tilde{t}_n(y)$ is asymptotically optimal.*

Proof. A consequence of Theorem 3.4 and 3.11. □

3.2 Consistency of Expected Posterior Loss and Minimizers

In this section, we study large sample theories for the expected posterior loss and the minimizers. More importantly, one result below tells us when the decision rule is an empirical Bayes procedure, which is a key assumption of Theorem 3.4 in the previous section.

Let F_n be a consistent MLE, that is, F_n converge weakly to the true prior distribution F. For each observed y, the decision rule induced by F_n is

$$t_n(y) \quad = \quad \arg\min_t \int_\Theta \int_\mathcal{Y} L(\theta, t) f(y|\theta)\, d\mu(y)\, dF_n(\theta)$$

$$= \quad \arg\min_{t} \frac{\int_{\Theta} L(\theta, t) f(y|\theta) \, dF_n(\theta)}{\int_{\Theta} f(y|\theta) \, dF_n(\theta)}.$$

For convenience, let us write

$$g(t; F_n) = \frac{\int_{\Theta} L(\theta, t) f(y|\theta) \, dF_n(\theta)}{\int_{\Theta} f(y|\theta) \, dF_n(\theta)}$$

and

$$g(t; F) = \frac{\int_{\Theta} L(\theta, t) f(y|\theta) \, dF(\theta)}{\int_{\Theta} f(y|\theta) \, dF(\theta)}.$$

We want to show for each observed y, the estimated expected posterior loss converge to the true one:

$$g(t_n; F_n) \to g(t; F), \text{ as } n \to \infty.$$

We make the following assumptions:

(A1) The integrand $L(\theta, a) f(y|\theta)$ is continuous and bounded in θ,

(A2) $\int_{\Theta} f(y|\theta) \, dF(\theta) > 0$ for all $y \in \mathcal{Y}$,

(A3) $L(\theta, a)$ is convex in a.

Theorem 3.13. *(Consistency of the expected posterior loss) Under the assumptions (A1)-(A2), we have*

$$g(t_n; F_n) \overset{a.s.}{\to} g(t; F).$$

Proof. A consequence of the definition of weak convergence of measures. □

Theorem 3.14. *(Consistency of the minimizers) Under the assumptions (A1)-*

(A3), for each observed y, the decision t_n is an empirical Bayes procedure, i.e.

$$t_n\left(y\right) \overset{a.s.}{\rightarrow} t_F\left(y\right).$$

Proof. Under the assumption (A3), the estimated expected posterior loss $g\left(t; F_n\right)$ is convex in t. As a consequence of the convexity lemma (See, e.g. Rockafellar (1970, page 266) and Pollard (1991)), the pointwise convergence of the objective functions is sufficient to give the consistency of minimizers. Under the assumptions (A1)-(A2), $g\left(t; F_n\right) \overset{a.s.}{\rightarrow} g\left(t; F\right)$ and the result follows. □

We will see when $f\left(y|\theta\right)$ is a member of one parameter exponential family, the assumption (A1) is usually true. In the examples below, we consider real-valued parameters and the k-th power absolute distance loss, i.e. $L\left(\theta, a\right) = |\theta - a|^k$, where $k \geq 1$. Then the assumption (A3) automatically holds.

Example 3.15. (Binomial distribution) Consider $y|\theta \sim Bin(n, \theta)$. The parameter space $\Theta = [0, 1]$ is compact. Then the assumption (A1) is trivially true.

Example 3.16. (Normal distribution with a known variance) Consider $y|\theta \sim N\left(\theta, 1\right)$ and the parameter space $\Theta = \mathbb{R}$. The assumption (A2) clearly hold. Furthermore, the integrand $I\left(\theta\right) := L\left(\theta, a\right) f\left(y|\theta\right)$ is the product of two continuous functions, hence, it is continuous. For the boundedness, using L'Hôpital's rule, we have $\lim_{\theta\rightarrow\infty} I\left(\theta\right) = \lim_{\theta\rightarrow-\infty} I\left(\theta\right) = 0$. Therefore, the assumption (A1) is true.

Remark 3.17. Using the same arguments above, we are able to show the assumption (A1) holds for Poisson distribution, exponential distribution, Gamma distribution with known shape parameter and Weibull distribution with known shape parameter as well.

Remark 3.18. For Gamma and Weibull distributions, there are more than one way of parametrizations. Specifically, for Gamma distribution with a known the shape parameter α, we choose

$$f\left(y|\theta\right)=\frac{\theta^{\alpha}}{\Gamma\left(\alpha\right)}y^{\alpha-1}e^{-\theta y}.$$

For Weibull distribution, we use

$$f\left(y|\theta\right)=\theta\alpha\left(\theta y\right)^{\alpha-1}e^{-\left(\theta y\right)^{\alpha}},$$

where the shape parameter α is known. If a different parametrization is used, the assumption (A1) may not hold: For example, consider a Gamma distribution with the shape parameter 1 can also be parametrized as

$$f\left(y|\tilde{\theta}\right)=\frac{1}{\tilde{\theta}}e^{-\frac{y}{\tilde{\theta}}}.$$

Then the integrand $I\left(\tilde{\theta}\right)$ goes to infinity as $\tilde{\theta}\to\infty$.

Chapter 4

Algorithms for Kiefer-Wolfowitz Dual Problem

As we already discussed in Chapter 2, the Kiefer-Wolfowitz maximum likelihood estimation problem can be formulated as a convex optimization problem in function space. When the dimension of parameter space is one, we can use the discretization based method proposed by Koenker and Mizera (2014) by restricting the prior F on a fine grid and then applying interior-point methods. Such an approach reduces the computational effort by "several orders of magnitude" by comparison to prior EM-based methods. The number of grid points required in the discretization method is about the square root of the sample size (Dicker and Zhao, 2014) but it grows exponentially fast as the dimension of the parameter space increases and the problem quickly becomes computationally intractable. Possible cure comes out of an insight that although the primal formulation of the maximum likelihood estimation problem is infinite-dimensional, the objective function of the dual formulation is finite dimensional. For this reason, in this chapter we focus on solving the

Kiefer-Wolfowitz dual problem and we propose four alternative algorithms for the dual problem which aim for solving the maximum likelihood estimation problem when the dimension of parameter space is relatively high.

In Section 4.1, we study the primal and dual formulations for the maximum likelihood estimation problem. In Section 4.2, we explore the geometric properties of the maximum likelihood estimation problem which play important roles when we develop optimization algorithms. In Section 4.3, we find out the dual problem is equivalent to a stochastic programming problem and develop a new projected stochastic gradient method to solve it. In Section 4.4 and 4.5, the stochastic problem is solved by using sampling based methods: the sample average approximation and a bootstrap aggregating like algorithm. In Section 4.6, we describe a cutting-plane method for solving the dual problem.

4.1 Primal and Dual Formulations of Kiefer-Wolfowitz Problem

In this section, we study the primal and dual formulations for Kiefer-Wolfowitz maximum likelihood estimation problem.

Recall that the Kiefer-Wolfowitz primal problem has the form

$$\min_{F\in\mathcal{F}} -\sum_{i=1}^{n} \log \int_{\mathbb{R}^p} f\left(y_i|\theta\right) dF(\theta), \tag{4.1}$$

where $\mathcal{F}$ is the class of all probability distribution functions on $\mathbb{R}$. The following result characterizes the relationship between the primal (2.2) and the dual formulation.

Theorem 4.1. *(Koenker and Mizera, 2014) The solution $\hat{F}_n$ of the primal* *problem exists and is an atomic probability measure with no more than* n *atoms. The locations* $\hat{\theta}_j$ *and the masses* $\hat{f}_j$ *at these locations can be found via the dual characterization: the solution* $\hat{v}$ *of*

$$\max_{v \in \mathbb{R}^n_+} \sum_{i=1}^{n} \log v_i$$

$$\textit{subject to} \quad \sum_{i=1}^{n} v_i f\left(y_i | \theta\right) \leq n, \qquad \textit{for all}\, \theta \in \mathbb{R}^p$$

satisfies the extremal equations

$$\sum_j f\left(y_i | \hat{\theta}_j\right) \hat{f}_j = \frac{1}{\hat{v}_i}, \quad i = 1, ..., n$$

and $\hat{\theta}_j$ *are exactly those* θ *where the dual constraint is active.*

Although the primal formulation of the Kiefer-Wolfowitz maximum likelihood estimation problem is infinite-dimensional, if we are working on its dual problem, we are back to $n-$dimensional Euclidean space (with infinitely many constraints). This result gives potential to develop efficient algorithms for the problem with large p.

4.2 Geometric Properties of Kiefer-Wolfowitz Maximum Likelihood Estimation Problem

Before proceeding to any concrete algorithms, we study the location of support set supp$(\hat{F})$ in the primal problem and the location of $\hat{v}$ in the dual problem, which play important roles when we develop optimization algorithms. In this section, we give a proof that for $p \geq 1$ all the support points of Kiefer-

Wolfowitz MLE lie in the convex hull of the observed points. Also, we provide an alternative proof that the solution of the dual problem lies on the boundary of the feasible set. In contrast to the previous proof of Lindsay (1983) based on the results of convex geometry, we establish the result by using the KKT conditions in semi-infinite programming.

The following result is first mentioned in Lindsay (1981) with no proof. To establish the result, we need to introduce some notation from Lindsay (1983). Let $f_F(y) = \int_\Theta f(y|\theta)\, dF(\theta)$, $f_\theta = (f(y_1|\theta), ..., f(y_n|\theta))$, $f_F = (f_F(y_1), ..., f_F(y_n))$. We write $\phi(x) = \sum_{i=1}^n \log x_i$ and define the gradient function of ϕ at f_{F_0} towards $f_{F_1}$to be

$$\begin{aligned} \Phi(f_{F_1}; f_{F_0}) &= \lim_{\epsilon \to 0} \epsilon^{-1} \{\phi((1-\epsilon) f_{F_0} + \epsilon f_{F_1}) - \phi(f_{F_0})\} \qquad (4.2) \\ &= \sum_{i=1}^n \{f_{F_1}(y_i) - f_{F_0}(y_i)\} / f_{F_0}(y_i), \end{aligned}$$

where $\epsilon \in (0,1)$. We write

$$D(\theta; F) := \Phi(f_\theta; f_F) = \sum_{i=1}^n \frac{L_i(\theta)}{\mathcal{L}_i(F)} - n,$$

where $L_i(\theta) = f(y_i|\theta)$ and $\mathcal{L}_i(F) = f_F(y_i)$.

Definition 4.2. (Convex hull) The convex hull of a set C, denoted $\operatorname{conv}(C)$, is the set of all convex combinations of points $x_1, ..., x_n$ in C:

$$\operatorname{conv}(C) = \left\{ \sum_{i=1}^n \alpha_i x_i : x_i \in C, \alpha_i \geq 0, i = 1, ..., n, \sum_{i=1}^n \alpha_i = 1 \right\}.$$

When we study the convergence of the algorithms, we fix the sample size n and let the number of iterations tend to infinity. For convenience, we usually denote Kiefer-Wolfowitz MLE $\hat{F}_n$ as $\hat{F}$.

Theorem 4.3. *For $p \geq 1$, if $y_i|\theta_i \overset{ind}{\sim} N_p\left(\theta_i, \Sigma\right)$, then all the support points of Kiefer-Wolfowitz MLE $\hat{F}$ lie in the convex hull of the observed points $\{y_i; 1 \leq i \leq n\}$.*

Proof. Let $\theta^* := \theta^*(\hat{F}) \in \mathbb{R}^p$ be a mode of the gradient function $D(\theta; \hat{F})$. By Theorem 19 in Lindsay (1995), the support points of $\hat{F}$ are necessarily modes. Then for the mode θ^*, we have

$$\nabla D(\theta^*; \hat{F}) = 0. \tag{4.3}$$

Since

$$\nabla L_i\left(\theta\right)/L_i\left(\theta\right) = -\Sigma^{-1}\left(y_i - \theta\right), \qquad \text{for each } i,$$

equation (4.3) can be rewritten as

$$\sum_{i=1}^{n} \frac{\nabla L_i\left(\theta^*\right)}{\mathcal{L}_i(\hat{F})} = \sum_{i=1}^{n} \frac{-\Sigma^{-1}\left(y_i - \theta^*\right) L_i\left(\theta^*\right)}{\mathcal{L}_i(\hat{F})} = 0. \tag{4.4}$$

By Theorem 19 in Lindsay (1995), we should have

$$\sum_{i=1}^{n} \frac{L_i\left(\theta^*\right)}{\mathcal{L}_i(\hat{F})} = n. \tag{4.5}$$

Combining (4.4) and (4.5), we obtain

$$\theta^* = \sum_{i=1}^{n} \alpha_i y_i,$$

where $\alpha_i := \alpha_i(\theta^*(\hat{F}), \hat{F}) = L_i\left(\theta^*\right)/(n\mathcal{L}_i(\hat{F}))$ and $\sum_{i=1}^{n} \alpha_i = 1$. Solving this equation for θ^* gives the result. □

Remark 4.4. As a consequence of Theorem 4.3, for $p = 1$, all the support points of $\hat{F}$ lie in the interval

$$[\min \{y_i; 1 \leq i \leq n\}, \max \{y_i; 1 \leq i \leq n\}].$$

Theorem 4.5. *(Reduced dual problem) For $p \geq 1$, if $y_i|\theta_i \overset{ind}{\sim} N_p(\theta_i, \Sigma)$, then Kiefer-Wolfowitz dual problem is equivalent to the reduced problem*

$$\min_{v \in Q} - \sum_{i=1}^{n} \log v_i \qquad \textit{subject to} \sum_{i=1}^{n} v_i L_i(\theta) \leq n, \quad \textit{for all}\, \theta \in \textit{conv}(y), \quad (4.6)$$

where Q is a n-dimensional box $[0, u_1] \times ... \times [0, u_n]$ with

$$u_i = 1 / \inf_{\theta \in conv(y)} L_i(\theta).$$

Proof. First we show that the feasible set of the reduced problem

$$\min_{v \in \mathbb{R}_+^n} - \sum_{i=1}^{n} \log v_i \qquad \text{subject to} \sum_{i=1}^{n} v_i L_i(\theta) \leq n, \quad \text{for all}\, \theta \in \text{conv}(y)$$

contains the feasible set of the original dual problem. One direction is clear: If v^* is the the solution of Kiefer-Wolfowitz dual problem, then it is the solution of the reduced problem. Consider the other direction: If v^* is the the solution of the reduced problem, the extremal equations

$$\sum_j L_i\left(\theta_j^*\right) f_j^* = \frac{1}{v_i^*}, \qquad i = 1, ..., n$$

indicate that for each i, v_i^* does not depend on the θ out of supp$(\hat{F})$, hence it does not depend on the θ out of conv(y) by Theorem 4.3.

From the extremal equations and the discussion above, we observe that

$$\sum_j L_i\left(\theta_j^*\right) f_j^* \geq \min_j L_i\left(\theta_j^*\right) \geq \inf_{\theta \in \operatorname{conv}(y)} L_i(\theta)$$

and then

$$v_i^* \leq 1 / \inf_{\theta \in \operatorname{conv}(y)} L_i(\theta), \quad i = 1, ..., n.$$

□

Next we provide a novel proof to show that the solution of the dual problem $\hat{v}$ lies on the boundary of the feasible set. This approach uses standard KKT arguments in *semi-infinite programming,* which is different from Lindsay (1983). A short overview of semi-infinite programming is provided first. Some references are Hettich and Kortanek (1993), López and Still (2007) and Shapiro (2009).

Semi-infinite programming is an optimization problem in finitely many variable $x = (x_1, ..., x_n) \in \mathbb{R}^n$ on a feasible set described by infinitely many constraints:

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{subject to } g(x,t) \leq 0,\ \forall t \in T,$$

where T is an infinite *index set.* The semi-infinite programming is called *convex* if the objective function $f(x)$ is convex and, for every index $t \in T$, the constraint function $g(\cdot, t)$ is convex. For a feasible $\bar{x}$, denote the active index set as

$$T_a(\bar{x}) = \{t \in T : g(\bar{x}, t) = 0\}.$$

Assume the objective function $f(x)$ is continuously differentiable on $\mathbb{R}^n$ and the index set T is compact. As in finite convex programming, the KKT conditions are necessary and sufficient for optimality López and Still (2007).

Remark 4.6. The requirement of compactness of T is satisfied by Theorem 4.5.

Theorem 4.7. *The solution of the dual problem lies on the boundary of the feasible set.*

Proof. For convenience, let us write $L_i(\theta) = f(y_i|\theta)$ and $L(\theta) = (L_1(\theta), ..., L_n(\theta))$. First, we show Slater's condition holds: Let $v_i = 1/(2L_i(\theta))$, then $L(\theta)^T v = n/2 < n$ for all $\theta \in \operatorname{conv}(y)$. Now let v^* be the minimizer of Problem 4.6. By Lemma 4 and Theorem 2(b) of López and Still (2007), there exist multipliers $\alpha_1^*, ..., \alpha_k^*, \beta^* \geq 0$ and indices $\theta_1, ..., \theta_k \in T_a(v^*)$ with $k \leq n$ such that

$$-\frac{1}{v_i^*} - \beta_i^* + \sum_{j=1}^{k} \alpha_j^* L_i(\theta_j) = 0, \qquad i = 1, ..., n.$$

The KKT conditions are:

$$L(\theta)^T v^* \leq n, \qquad \forall \theta \in \operatorname{conv}(y)$$

$$-v^* \leq 0,$$

$$\alpha_j^* \geq 0, \qquad j = 1, ..., k$$

$$\beta^* \geq 0,$$

$$\alpha_j^* \left(L(\theta_j)^T v^* - n \right) = 0, \qquad j = 1, ..., k \tag{4.7}$$

$$\beta_i^* v_i^* = 0, \qquad i = 1, ..., n$$

$$-\frac{1}{v_i^*} - \beta_i^* + \sum_{j=1}^{k} \alpha_j^* L_i(\theta_j) = 0, \qquad i = 1, ..., n. \tag{4.8}$$

Since $v_i^* \neq 0$, the equality $\beta_i^* v_i^* = 0$ implies $\beta_i^* = 0$ for all i. Then from

(4.8) we have $\sum_{j=1}^{k} \alpha_j^* L_i(\theta_j) = 1/v_i^*$ for all i. This suggests there is at least one $\alpha_j^* \neq 0$. The condition of complementary slackness (4.7) therefore implies $L(\theta_j)^T v^* - n = 0$ for some θ_j. That is, the solution lies on at least one constraint line. □

4.3 Projected Stochastic Gradient Methods

In this section, we study a novel projected stochastic gradient method and apply it to solve Kiefer-Wolfowitz dual problems. The main appeal of this algorithm is that in each iteration, the random direction one moves towards only depends on finitely many randomly selected constraints, which makes it useful for large scale programming problems. More importantly, we establish a connection between semi-infinite and stochastic programming based on the work of Tadić, Meyn, and Tempo (2006) so that an optimization problem with infinitely many constraints can be reformulated as a problem with one stochastic constraint. In Sections 4.4 and 4.5, we shall see that some of the results in this section contribute in the development of various sampling methods.

Tadić et al. (2006) establish a result that a large class of semi-infinite programming problem can be reformulated as a constrained stochastic programming problem by introducing a penalty function. Let $\mathcal{B}^p$ denote the class of Borel-measurable sets on $\mathbb{R}^p$ and μ a probability measure on $\mathcal{B}^p$ satisfying $\mu(A) > 0$ for any non-empty open set $A \subset \mathbb{R}^p$. A continuous and differentiable function $h : \mathbb{R} \to \mathbb{R}_+$ with positive support:

$$h(t) = \begin{cases} 0 & \text{for all } t \in (-\infty, 0] \\ > 0 & \text{for all } t \in (0, \infty). \end{cases}$$

To make the problem convex, we further require h to be convex and non-decreasing. Let Θ be an $\mathbb{R}^p$-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$ whose probability measure is μ:

$$P(\Theta \in B) = \mu(B), \quad B \in \mathcal{B}^p.$$

Theorem 4.8. *(Tadić et al., 2006, Corollary 2) Let $g : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ be a Borel-measurable function such that*

$$g(v, \cdot) \text{ is continuous on } \mathbb{R}^p \text{ for each } v \in \mathbb{R}^n.$$

Under the assumptions above, for a continuous objective function $f : \mathbb{R}^n \to \mathbb{R}$, the semi-infinite programming

$$\min_{v \in \mathbb{R}^n} f(v)$$

$$\text{subject to} \qquad g(v, \xi) \leq 0, \qquad \forall \xi \in \mathbb{R}^p$$

is equivalent to the constrained stochastic optimization problem

$$\min_{v \in \mathbb{R}^n} f(v)$$

$$\text{subject to} \qquad \mathbb{E}(h(g(v, \Theta))) \leq 0.$$

We apply the result of Tadić et al. (2006) on the dual problem. Recall from Theorem 4.5 that Q is a n-dimensional box

$$[0, u_1] \times ... \times [0, u_n],$$

where

$$u_i = 1/ \inf_{\theta \in \mathrm{conv}(y)} L_i(\theta).$$

Theorem 4.9. *Kiefer-Wolfowitz dual problem is equivalent to a stochastic programming problem*

$$\min_{v \in Q} - \sum_{i=1}^{n} \log(v_i) \tag{4.9}$$

$$\textit{subject to} \qquad \psi(v) = \mathbb{E}(h(g(v, \Theta))) = 0,$$

where $g(v, \theta) = L(\theta)^T v - n$.

Proof. The function $g(v, \cdot)$ is continuous on $\mathbb{R}^p$ for each v. Then the result is a consequence of Theorem 4.5 and Theorem 4.8. □

Let $\Pi_Q(x)$ denote projection of a point $x \in \mathbb{R}^n$ onto the set Q. The stochastic programming problem (4.9) could be then solved using a projected stochastic gradient method:

$$v^{(k+1)} = \Pi_Q(v^{(k)} - \gamma^{(k+1)} \nabla(- \sum_{i=1}^{n} \log(v_i^{(k)}) + \delta^{(k+1)} \psi(v^{(k)})), \qquad k \geq 0 \tag{4.10}$$

where $\left\{\delta^{(k)}\right\}_{k \geq 1}$ is an increasing sequence of positive reals such that $\lim_{k \to \infty} \delta^{(k)} = \infty$, $\left\{\gamma^{(k)}\right\}_{k \geq 1}$ is sequence of positive reals. By Lebesgue's dominated convergence theorem, we have

$$\nabla \psi(v) = \nabla \mathbb{E}(h(g(v, \Theta))) = \mathbb{E}(h'(g(v, \Theta)) \nabla_v g(v, \Theta)).$$

For many problems, the gradient $\nabla \psi$ can not be computed explicitly. This motivate us thinking of Monte-Carlo sampling methods and approximating the expectation by its sample average.

The idea, known as stochastic gradient method, has already been explored in Subsection 2.4.2.1. Let $\xi_1^{(k)}, ..., \xi_N^{(k)}$ be a sequence of i.i.d. uniform random variables on $\operatorname{conv}(y)$ in kth iteration. The update step in projected stochastic gradient method has the form

$$v^{(k)} = \Pi_Q(v^{(k)} - \gamma^{(k+1)}(-\frac{1}{v^{(k)}} + \frac{\delta^{(k+1)}}{N}\sum_{i=1}^{N} h'(g(v^{(k)}, \xi_i^{(k)}))L(\xi_i^{(k)}))),$$

where $k \geq 0$.

With a carefully chosen penalty function h and step sizes γ, δ and other mild conditions, we prove the algorithm eventually converges to the optimal solution. A general convergence result is provided below.

4.3.0.1 Convergence Analysis

The convergence of the previously mentioned projected stochastic gradient method is studied here. First we provide a convergence result in a general setup.

Suppose a continuous function $h : \mathbb{R} \to \mathbb{R}_+$has support $(0, \infty)$:

$$h(t) = \begin{cases} 0 & \text{for all } t \in (-\infty, 0] \\ > 0 & \text{for all } t \in (0, \infty) \end{cases}$$

and h is differentiable. Suppose $g(\cdot, y)$ is differentiable and convex for each $y \in \mathbb{R}^p$. Assume a predetermined set $Q \subset \mathbb{R}^n$ is compact and convex.

Consider a semi-infinite programming problem

$$\min_{x \in Q} f(x)$$

$$\text{subject to} \qquad g(x,y) \leq 0, \quad \forall y \in \mathbb{R}^p$$

or equivalently a stochastic programming problem

$$\min_{x \in Q} f(x)$$

$$\text{subject to} \qquad \psi(x) = \mathbb{E}\left(h\left(g\left(x,Y\right)\right)\right) = 0.$$

Denote the standard Euclidean norm as $\|\cdot\|$. For an integer $n \geq 1$ and $z \in \mathbb{R}^n$, $\rho > 0$, the associated closed balls are defined as

$$B_\rho^n(z) = \left\{z' \in \mathbb{R}^n : \|z - z'\| \leq \rho\right\}.$$

Let D be the feasible set $D = \{x \in \mathbb{R}^n : g(x,y) \leq 0, \forall y \in \mathbb{R}^p\}$ and $\left\{\delta^{(k)}\right\}_{k\geq 1}$ an increasing sequence of positive reals satisfying $\lim_{k\to\infty} \delta^{(k)} = \infty$. The projected stochastic gradient method generates iteratively the sequence $\left\{X^{(k)}\right\}_{k\geq 0}$ via

$$\begin{aligned} X^{(k+1)} &= \Pi_Q(X^{(k)} - \gamma^{(k+1)}(\nabla f\left(X^{(k)}\right) \\ &\quad +\delta^{(k+1)} h'\left(g\left(X^{(k)}, Y^{(k+1)}\right)\right) \nabla_x g\left(X^{(k)}, Y^{(k+1)}\right))), \end{aligned} \tag{4.11}$$

where $k \geq 0$.

Definition 4.10. (Lipschitz continuity) A real-valued function $f(x)$ is called Lipschitz continuous if there exists a real constant $K \geq 0$ such that, for all x

and y in the domain, $|f(x) - f(y)| \leq K\|x-y\|$. A real-valued function $f(x)$ is called locally Lipschitz continuous if for every x in the domain there exists a neighborhood U of x such that f restricted to U is Lipschitz continuous.

The following assumptions are required in the analysis of the algorithm:

Assumption 1. $\gamma^{(k)}, \delta^{(k)} > 0$ for $k \geq 1$, $\sum_{k=1}^{\infty} \gamma^{(k)} = \infty$, $\sum_{k=1}^{\infty} \left(\gamma^{(k)}\right)^2 < \infty$, $\sum_{k=1}^{\infty} \gamma^{(k)}\delta^{(k)} = \infty$ and $\sum_{k=1}^{\infty} \left(\gamma^{(k)}\delta^{(k)}\right)^2 < \infty$.

Assumption 2. f is convex and ∇f is locally Lipschitz continuous.

Assumption 3. h is convex and nondecreasing; $g(\cdot, y)$ is convex for each $y \in \mathbb{R}^p$. For all $\rho \in [1, \infty)$, there exists a Borel-measurable function $\phi_\rho : \mathbb{R}^p \to [1, \infty)$ and such that

$$\int \phi_\rho^4(y)\, \mu(dy) < \infty,$$

and for all $x, x', x'' \in B_\rho^n$, $y \in \mathbb{R}^p$,

$$\begin{aligned}
\max\{|h(g(x,y))|, |h'(g(x,y))|, \|\nabla_x g(x,y)\|\} &\leq \phi_\rho(y), \\
|h'(g(x',y)) - h'(g(x'',y))| &\leq \phi_\rho(y)\|x' - x''\|, \\
\|\nabla_x g(x',y) - \nabla_x g(x'',y)\| &\leq \phi_\rho(y)\|x' - x''\|.
\end{aligned}$$

Assumption 4. $D \cap Q \neq \emptyset$, $\eta^* := \inf_{x \in D \cap Q} f(x) > -\infty$ and the set of optimizers $D^* := \{x \in D : f(x) = \eta^*\}$ is non-empty.

Assumption 1 holds when the step sizes γ, δ are carefully chosen. Assumption 2 corresponds to the properties of the objective function f. Assumption 3 holds when the penalty function h is carefully selected; the function ψ is well-defined, finite and differentiable and $\nabla\psi$ is locally Lipschitz continuous. Assumption 4 ensures the optimization problem is well-defined and it has non-trivial solutions.

Using the method of Tadić et al. (2006), we are able to establish the convergence result of the algorithm.

Let $f_k(x) = f(x) + \delta_{k+1}\psi(x)$ for $x \in \mathbb{R}^n$, $k \geq 1$. Denote

$$\begin{aligned}
\kappa_{k+1} &= \gamma_{k+1}\delta_{k+1}\left(\nabla\psi(X_k) - h'\left(g\left(X_k, Y_{k+1}\right)\right)\nabla_x g\left(X_k, Y_{k+1}\right)\right) \\
\epsilon_{1,k+1} &= 2\left(X_k - \Pi_{D^*}(X_k)\right)^T \kappa_{k+1} \\
\epsilon_{2,k+1} &= \|Z_{k+1} - X_k\|^2 \\
\epsilon_{k+1} &= \epsilon_{1,k+1} + \epsilon_{2,k+1} \\
Z_{k+1} &= X_k - \gamma_{k+1}\left(\nabla f(X_k) + \delta_{k+1}h'\left(g\left(X_k, Y_{k+1}\right)\right)\nabla_x g\left(X_k, Y_{k+1}\right)\right).
\end{aligned}$$

So we have

$$Z_{k+1} = X_k - \gamma_{k+1}\nabla f_k(X_k) + \kappa_{k+1} \tag{4.12}$$

and

$$X_{k+1} = \Pi_Q(Z_{k+1}).$$

Definition 4.11. (Nonexpansive) Let B be a Banach space and C a nonempty bounded closed and convex subset of B. A mapping $T : C \to B$ is said to be nonexpansive if

$$\|Tx - Ty\| \leq \|x - y\|, \quad x, y \in C.$$

Since $f_k(x)$ is convex and $\Pi_{D^*}(\cdot)$, $\Pi_Q(\cdot)$ are nonexpansive, for any ω we have

$$\begin{aligned}
\left(X_k - \Pi_{D^*}(X_k)\right)^T \nabla f_k(X_k) &\geq f_k(X_k) - f_k\left(\Pi_{D^*}(X_k)\right) \\
&= f_k(X_k) - \eta^* \qquad (4.13)
\end{aligned}$$

and

$$\begin{aligned}\|X_{k+1}-\Pi_{D^*}\left(X_{k+1}\right)\| &\leq \|X_{k+1}-\Pi_{D^*}\left(X_k\right)\| \\ &= \|\Pi_Q\left(Z_{k+1}\right)-\Pi_Q\left(\Pi_{D^*}\left(X_k\right)\right)\| \\ &\leq \|Z_{k+1}-\Pi_{D^*}\left(X_k\right)\|, \qquad k\geq 0. \qquad (4.14)\end{aligned}$$

Then for any ω and all $k \geq 0$, (4.14) yields

$$\begin{aligned}\|X_{k+1}-\Pi_{D^*}\left(X_{k+1}\right)\|^2 &\leq \|Z_{k+1}-\Pi_{D^*}\left(X_k\right)\|^2 \\ &= \|\left(X_k-\Pi_{D^*}\left(X_k\right)\right)+\left(Z_{k+1}-X_k\right)\|^2 \\ &= \|X_k-\Pi_{D^*}\left(X_k\right)\|^2 \\ &\quad +2\left(X_k-\Pi_{D^*}\left(X_k\right)\right)^T\left(Z_{k+1}-X_k\right)+\|Z_{k+1}-X_k\|^2 \\ &= \|X_k-\Pi_{D^*}\left(X_k\right)\|^2-2\gamma_{k+1}\left(X_k-\Pi_{D^*}\left(X_k\right)\right)^T\nabla f_k\left(X_k\right) \\ &\quad +\epsilon_{k+1}.\end{aligned}$$

Lemma 4.12. *Suppose Assumptions 1-3 hold. Then* $\lim_{k\to\infty}\|Z_{k+1}-X_k\|=0$ *and* $\lim_{k\to\infty}\|X_{k+1}-X_k\|=0$ *almost surely on the event* $\left\{\sup_{k\geq 0}\|X_k\|<\infty\right\}$.

Proof. Let $\rho\in[1,\infty)$ and $K_\rho\in[\rho,\infty)$ denotes an upper bound of $\|\Pi_{D^*}\left(\cdot\right)\|$, $\|\nabla\psi\|$ and $\|\nabla f\|$ on the set Q. From Assumption 3,

$$\begin{aligned}\|\kappa_{k+1}\| \quad 1_{\{\|X_k\|\leq\rho\}} &\leq 2K_\rho\gamma_{k+1}\delta_{k+1}\phi_\rho^2\left(Y_{k+1}\right), \\ |\epsilon_{1,k+1}|1_{\{\|X_k\|\leq\rho\}} &\leq 4K_\rho\|\kappa_{k+1}\|1_{\{\|X_k\|\leq\rho\}}, \\ |\epsilon_{2,k+1}|1_{\{\|X_k\|\leq\rho\}} &\leq 2K_\rho^2\gamma_{k+1}^2+2\|\kappa_{k+1}\|^2 1_{\{\|X_k\|\leq\rho\}}\end{aligned}$$

for $k \geq 0$. This implies that

$$\begin{aligned}
\mathbb{E}\left(\sum_{k=1}^{\infty}\|\kappa_k\|^2 1_{\{\|X_k\|\leq\rho\}}\right) &\leq 4K_\rho^2\sum_{k=1}^{\infty}\gamma_k^2\delta_k^2\mathbb{E}\left(\phi_\rho^2\left(Y_k\right)\right) < \infty, \\
\mathbb{E}\left(\sum_{k=1}^{\infty}|\epsilon_{1,k}|^2 1_{\{\|X_k\|\leq\rho\}}\right) &\leq 16K_\rho^2\mathbb{E}\left(\sum_{k=1}^{\infty}\|\kappa_k\|^2 1_{\{\|X_k\|\leq\rho\}}\right) < \infty, \\
\mathbb{E}\left(\sum_{k=1}^{\infty}|\epsilon_{2,k}|^2 1_{\{\|X_k\|\leq\rho\}}\right) &\leq 2\mathbb{E}\left(\sum_{k=1}^{\infty}\|\kappa_k\|^2 1_{\{\|X_k\|\leq\rho\}}\right) \\
&\quad +4K_\rho^2\sum_{k=1}^{\infty}\gamma_{k+1}^2\left(1+\delta_{k+1}^2\right) < \infty.
\end{aligned}$$

Note that

$$Z_{k+1} - X_k = -\gamma_{k+1}\left(\nabla f\left(X_k\right) + \delta_{k+1}\nabla\psi\left(X_k\right)\right) + \kappa_{k+1}.$$

From Assumptions 2 and 3, we have

$$\|Z_{k+1} - X_k\| \leq K_\rho\gamma_{k+1}\left(1+\delta_{k+1}\right) + \|\kappa_{k+1}\|, \quad \forall\omega. \tag{4.15}$$

Owing to Assumption 1 and (4.15), we obtain

$$\lim_{k\to\infty}\|Z_{k+1} - X_k\| = 0, \qquad \text{w.p.1.} \tag{4.16}$$

For arbitrary sample ω,

$$\begin{aligned}
\|X_{k+1} - X_k\| &= \|\Pi_Q\left(Z_{k+1}\right) - X_k\| \\
&\leq \|Z_{k+1} - X_k\|.
\end{aligned}$$

Using (4.16), we have the desired result. □

Let $d(x, S)$ denote the distance of a point x from set S, defined by

$$d(x, S) = \inf \{d(x, s) : s \in S\}.$$

Theorem 4.13. *Let $\left\{X^{(k)}\right\}_{k\geq 0}$ be generated by (4.11), and assume assumptions 1-4 hold. Then*

$$\mathbb{P}\left(\lim_{k\to\infty} d\left(X^{(k)}, D^*\right) = 0\right) = 1.$$

Proof. By Lemma 4.12 and the similar arguments in Theorem 4 of Tadić et al. (2006), we have the result. □

Next we show the general projected stochastic gradient algorithm can be applied to our problem.

Theorem 4.14. *For Kiefer-Wolfowitz dual problem, let $\left\{X^{(k)}\right\}_{k\geq 0}$ be generated by (4.11). Let $h(t) = (\max\{0, t\})^2$ and γ, δ satisfy the assumption 1, then $v^{(k)}$ converge to the optimal solution v^* almost surely as $k \to \infty$.*

Proof. To apply Theorem 4.13, we need to justify the assumptions 1-4. Assumption 1 and 4 immediately hold. Assumption 2 is true due to the fact that any continuously differentiable function is locally Lipschitz. Assumption 3 follows the comments in Tadić et al. (2006, page 7-8), since h is chosen to be piecewise quadratic and $g(\cdot, y)$ is linear for any y. □

4.3.0.2 Stopping Criteria

Following the suggestion of Ermoliev (1983), for an objective function f_0, we terminate the algorithm if there is no improvement in the objective function after M iterations

$$\left| f_0\left(v^{(k)}\right) - \frac{1}{M+1} \sum_{j=k-M}^{k} f_0\left(v^{(j)}\right) \right| \leq \epsilon,$$

where M is a relatively large prespecified integer and ϵ is a predetermined threshold.

4.3.0.3 Numerical Experiment

We compute the MLE via the projected stochastic gradient method for $p = 2$ and $n = 100$. Suppose $y_i|\theta_i \overset{ind}{\sim} N_p\left(\theta_i, \Sigma\right)$ and $\theta_i \overset{iid}{\sim} N\left(10\mathbf{1}_p, S\right)$, where $i = 1, ..., n$, $\Sigma_{jk} = 0.9^{|j-k|}$, $\mathbf{1}_p = (1, ..., 1) \in \mathbb{R}^p$, $S_{jk} = 0.8^{|j-k|}$ and $j, k \in \{1, 2\}$. For the penalty function and step sizes, we choose $h\left(t\right) = \left(\max\left\{0, t\right\}\right)^2$, $\gamma^{(k)} = k^{-1/2}$ and $\delta^{(k)} = 1$.

In Figure 1, we observe that the convergence of the algorithm can be very slow: When the algorithm starts, a lot of time it moves to the wrong directions so that the objective value can increase. After hundreds of iterations, it gradually converges. When it gets close to the true value, the convergence slows down.

4.4 Sample Average Approximation

In this section, we apply sample average approximation methods to solve KW dual problems and establish the convergence result. From the previous discussion, we already see that the Kiefer-Wolfowitz dual problem is equivalent to a stochastic programming problem: Infinitely many constraints can be summarized as one stochastic constraint. The underlying distribution μ

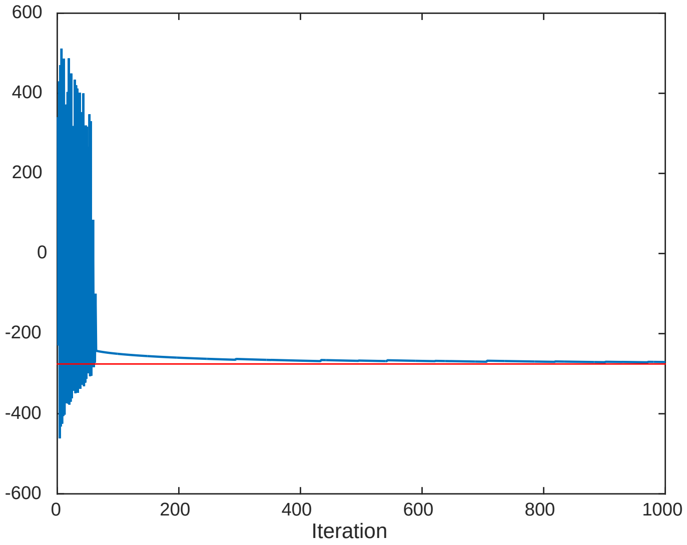

Figure 1: Convergence of the objective values using the projected stochastic gradient method. The red horizontal line corresponds to the objective value using the interior-point method with a fine grid.

is chosen by us and therefore it is known. The evaluation of the integration is rather difficult. It is possible to conduct some sampling schemes from the distribution μ and solve the corresponding deterministic problem. The superiority of this approach is that (1) The approximated problem only involves finitely many constraints and it is implementable; (2) It preserves the convexity of the problem: If the stochastic problem is convex, then its corresponding sample average approximation problem is convex as well. Therefore, sample average approximation problem can be then efficiently solved by deterministic convex algorithms.

Recall that the dual problem can be reformulated as a stochastic programming problem

$$\min_{v \in Q} - \sum_{i=1}^{n} \log\left(v_i\right)$$

$$\text{subject to} \qquad \mathbb{E}\left(h\left(g\left(v, \Theta\right)\right)\right) = 0. \tag{4.17}$$

Suppose we have a random sample $\xi_1, ..., \xi_N$ of N realizations of random variable Θ. The equality constraint (4.17) can be approximated by $\sum_{i=1}^{N} h\left(g\left(v, \xi_i\right)\right)/N = 0$. Since the penalty function is always nonnegative, we replace the equality in the original sample average approximation problem by inequality and make the problem convex. The sample average approximation of Kiefer-Wolfowitz dual problem then has the form

$$\min_{v \in Q} - \sum_{i=1}^{n} \log\left(v_i\right) \tag{4.18}$$

$$\text{subject to} \qquad \frac{1}{N} \sum_{i=1}^{N} h\left(g\left(v, \xi_i\right)\right) \leq 0.$$

The following results further simplify the problem and show that the optimal solution of the sample average approximation problem $\hat{v}_N$ can be obtained by solving Kiefer-Wolfowitz dual problem with randomly chosen N constraints.

Theorem 4.15. *The sample average approximation problem (4.18) is equivalent to*

$$\min_{v \in Q} - \sum_{i=1}^{n} \log\left(v_i\right)$$

$$\textit{subject to} \qquad g\left(v, \xi_i\right) \leq 0, \qquad i = 1, ..., N.$$

Proof. Note that for a given random sample $\xi_1, ..., \xi_N$, the two sets

$$A_N = \left\{ v \in \mathbb{R}_+^n : \sum_{i=1}^{N} h\left(g\left(v, \xi_i\right)\right)/N \leq 0 \right\}$$

and

$$B_N = \left\{ v \in \mathbb{R}_+^n : g\left(v, \xi_i\right) \leq 0,\ i = 1, ..., N \right\}$$

are equivalent. The result follows. □

4.4.0.1 Asymptotic Properties

Consistency results are provided here. It should be pointed out that when the convergence of the sample average approximation algorithms is studied, the size of randomly selected constraints N goes to infinity but the sample size of data n is kept fixed.

Let f^* and v^* denote the optimal value and the optimal solution of Kiefer-Wolfowitz dual problem; $\hat{f}_N$ and $\hat{v}_N$ the optimal value and the optimal solution of the sample average approximation problem (4.18).

Consider a general convex stochastic problem taking the form

$$\min_{v \in Q} f\left(v\right) = \mathbb{E}F\left(v, \Theta\right),$$

$$\text{subject to} \qquad u\left(v\right) = \mathbb{E}G\left(v, \Theta\right) \leq 0,$$

where Q is a nonempty closed subset of $\mathbb{R}^n$; $F\left(\cdot, \xi\right)$ and $G\left(\cdot, \xi\right)$ are convex. Its sample average approximation problem has the form

$$\min_{v \in Q} \frac{1}{N} \sum_{i=1}^{N} F\left(v, \xi_i\right),$$

$$\text{subject to} \qquad u_N\left(v\right) = \frac{1}{N} \sum_{i=1}^{N} G\left(v, \xi_i\right) \leq 0.$$

Denote the feasible sets for the stochastic problem and its sample average approximation problem by

$$V = \left\{v \in Q : u\left(v\right) \leq 0\right\}$$

and

$$V_N = \{v \in Q : u_N(v) \leq 0\}.$$

Write the deviation of the set A from the set B as

$$\mathbb{D}(A, B) := \sup_{x \in A} \operatorname{dist}(x, B),$$

where $\operatorname{dist}(x, B) := \inf_{x' \in B} \|x - x'\|$.

Definition 4.16. (Semi-continuity) Suppose X is a topological space, x_0 is a point in X and $f : X \to \overline{\mathbb{R}}$ is an extended real-valued function. We say that f is upper semi-continuous at x_0 if for every $\epsilon > 0$ there exists a neighborhood U of x_0 such that $f(x) \leq f(x_0) + \epsilon$ for all $x \in U$ when $f(x_0) > -\infty$, and $f(x)$ tends to $-\infty$ as x tends towards x_0 when $f(x_0) = -\infty$. We say f is lower semi-continuous at x_0 if for every $\epsilon > 0$ there exists a neighborhood U of x_0 such that $f(x) \geq f(x_0) - \epsilon$ for all x in U when $f(x_0) < \infty$, and $f(x)$ tends to ∞ as x tends towards x_0 when $f(x_0) = \infty$.

Remark 4.17. A function is continuous at x_0 if and only if it is upper and lower semi-continuous there.

The consistency of the algorithm for Kiefer-Wolfowitz dual problem is a consequence of the following result.

Theorem 4.18. *(Shapiro et al., 2009, Theorem 5.5) Suppose that: (i) the function F is random lower semicontinuous, (ii) for almost every ξ the function $F(\cdot, \xi)$ is convex, (iii) the set V is closed and convex, (iv) the expected value function u is lower semicontinuous and there exists a point $\bar{v} \in V$ such that $f(v) < \infty$ for all v in a neighborhood of $\bar{v}$, (v) the set S of optimal solutions of the true problem is nonempty and bounded, and (vi) the law of* *large numbers holds pointwise. In addition, we assume (a) If $v_N \in V_N$ and v_N converge w.p. 1 to a point v, then $v \in V$, (b) For some point $v \in S$ there exists a sequence $v_N \in V_N$ such that $v_N \to v$ w.p. 1. Then $\hat{f}_N \to f^*$ and $\mathbb{D}\left(\hat{S}_N, S\right) \to 0$ w.p. 1 as $N \to \infty$.*

Corollary 4.19. *(Consistency of Dual) For Kiefer-Wolfowitz dual problem, $\hat{f}_N \to f^*$ and $\hat{v}_N \to v^*$ almost surely.*

Proof. The result is a consequence of Theorem 4.18. The conditions (i) , (ii), (iii), (v) and (vi) in Theorem 4.18 can be verified directly. The second part of the condition (iv) is also straight. We only need to justify the first part of the condition (iv): the expected value function $U(v) := \mathbb{E}h(g(v, \Theta))$ is lower semicontinuous. In fact, U is continuous on Q. Let $\{v_t\}$ be a sequence such that $v_t \to v_0$. Now we consider the sequence of measurable functions $f_t(\theta) = h(g(v_t, \theta))$. From the construction, the function $h(g(\cdot, \theta))$ is continuous for every θ. Now

$$\lim_{t\to\infty} U(v_t) = \lim_{t\to\infty} \int_{\text{conv}(y)} f_t(\theta)\, \mu d(\theta)$$

is bounded above by some constant by the extreme value theorem. By Lebesgue's dominated convergence theorem, we have

$$\begin{aligned} \lim_{t\to\infty} \int_{\text{conv}(y)} f_t(\theta)\, \mu d(\theta) &= \int_{\text{conv}(y)} \lim_{t\to\infty} f_t(\theta)\, \mu d(\theta) \\ &= \int_{\text{conv}(y)} h(g(v_0, \theta))\, \mu d(\theta) \\ &= U(v_0). \end{aligned}$$

Thus

$$\lim_{t\to\infty} U(v_t) = U(v_0),$$

which means U is continuous. The condition (a) in Theorem 4.18 is a consequence of Theorem 2 in Jennrich (1969). For the condition (b), chose $v_N = v^*$ for all N and the result follows. □

Before we give a consistency result for the primal problems, we introduce some concepts and notation on the weak topology.

Let Ω be an arbitrary metric space, let $\mathcal{B}$ be the Borel-σ-algebra and consider probability measures F_n and F defined on $\mathcal{B}$. Let $\mathfrak{R}$ be the space of all probability measures on $(\Omega, \mathcal{B})$.

Definition 4.20. (Weak convergence) We say F_n converge weakly to F if $\int \psi dF_n \to \int \psi dF$ for all bounded and continuous functions ψ and we write $F_n \Rightarrow F$.

The weak(-star) topology in $\mathfrak{R}$ is the weakest topology such that, for every bounded continuous function ψ, the map

$$F \to \int \psi dF$$

from $\mathfrak{R}$ into $\mathbb{R}$ is continuous.

Lemma 4.21. *(Huber, 2011, Lemma 2.1) A linear functional T is weakly continuous on $\mathfrak{R}$ if and only if it can be represented in the form*

$$T(F) = \int \psi(\theta)\, dF(\theta)$$

for some bounded and continuous function ψ.

Let F^* denote the optimal solution of Kiefer-Wolfowitz primal problem; F_N the optimal solution of the primal problem restricted on the randomly generated grid $\{\xi_1, ..., \xi_N\}$. Without loss of generality, we take $n = 1$ and define a linear functional by

$$T(F) = \int_{\mathbb{R}^p} f(y|\theta)\, dF(\theta).$$

We further assume $f(y|\theta)$ is a member of exponential family so that $f(y|\cdot)$ is continuous and bounded.

Theorem 4.22. *(Consistency of Primal) If T is one-to-one, then F_N converge weakly to F^* almost surely:*

$$\mathbb{P}\left(\left\{\omega : \lim_{N\to\infty} F_N(\cdot)(\omega) \Rightarrow F^*\right\}\right) = 1.$$

Proof. By Lemma 4.21, the linear functional T is weakly continuous at F^*. By Theorem 4.19, $\hat{v}_N \to v^*$ almost surely. From one of the KKT conditions, we have

$$T(F_N) = 1/\hat{v}_N.$$

Then $T(F_N) \to T(F^*)$ almost surely. Under the assumption, the result follows by applying continuous mapping theorem. □

Theorem 4.23. *(Consistency of Decision Rules) Consider a problem of multiple prediction discussed in Chapter 2. Let the loss function $\rho : (\theta, q) \mapsto \rho(\theta, q) \in \mathbb{R}$ defined for $\theta \in \mathbb{R}^p$ and $q \in \mathcal{Q}$, where $\mathcal{Q}$ is a nonempty subset of $\mathbb{R}^p$. Let $\rho(\cdot, q)$ be a closed proper convex function which attains its infimum at an unique point and let $\hat{q}_N := q(\hat{F}_N)$ be Bayes decision rule of θ with a sample average approximation estimator $\hat{F}_N$ of Kiefer-Wolfowitz MLE. Assume the condition of Theorem 4.22 holds. Then $\hat{q}_N \to q^*$ almost surely.*

Proof. By Theorem 4.22, we have $F_N \Rightarrow F$ almost surely. Denote

$$dQ := f\left(y|\theta\right)/f\left(y\right)dF$$

and

$$dQ_N := f\left(y|\theta\right)/f\left(y\right)dF_N.$$

Then $Q_N \Rightarrow Q$ almost surely. Recall that the Bayes decision problem in the setting of Kiefer-Wolfowitz maximum likelihood procedure is to solve

$$\min_{q\in Q}\int_{\text{conv}(y)}\rho\left(\theta,q\right)dQ_N\left(\theta\right).$$

By Corollary 27.2.2. of Rockafellar (1970, page 266), it is sufficient to show the pointwise convergence of $\int_{\text{conv}(y)}\rho\left(\theta,q\right)dQ_N\left(\theta\right)$. By continuity of convex functions and the definition of weak convergence, the result follows. □

4.4.0.2 Theoretical Sample Size Estimation

Now we try to answer the question that how many random constraints N are required to obtain a good solution for sample average approximation problems. We show that the probability that an optimal solution to sample average approximation is an ϵ-optimal solution to the original problem goes to 1 exponentially fast as the size of randomly selected constraints N increases. In this subsection, we choose the sampling distribution to be multivariate normal, but the method can deal with the general sampling distributions.

To formulate the result, we need to introduce some notation. Given $\epsilon > 0$, define the relaxed feasible region as

$$S^{\epsilon} := \left\{ v \in Q : \mathbb{E}\left(h\left(g\left(v, \Theta\right)\right)\right) \leq \epsilon \right\}.$$

Then S^0 represents the feasible region of Kiefer-Wolfowitz dual problem. Let $\xi_1, ..., \xi_N$ be a sample of size N of Θ. Correspondingly we define

$$S_N^{\epsilon} := \left\{ v \in Q : \frac{1}{N} \sum_{i=1}^{N} h\left(g\left(v, \xi_i\right)\right) \leq \epsilon \right\}.$$

Wang and Ahmed (2008) show that, under proper conditions, the probability that the feasible set of the sample average approximation problem is "sandwiched" between $S^{-\epsilon}$ and S^{ϵ} goes to one exponentially fast:

$$\mathbb{P}\left(S^{-\epsilon} \subset S_N^0 \subset S^{\epsilon}\right) \geq 1 - Me^{-\beta\epsilon^2 N},$$

for some constants $M, \beta > 0$.

In the dual problem, given any sample $\xi_1, ..., \xi_N$, we always have $S^{-\epsilon} \subset S_N^0$, so we are more interested to evaluate the probability $\mathbb{P}\left(S_N^0 \subset S^{\epsilon}\right)$. A similar result can be obtained from the work of Wang and Ahmed (2008).

Let D be the diameter of the set Q, i.e., $D = \max_{v_1, v_2 \in Q} \|v_1 - v_2\|$ and

$$\Phi_p = \left(\max_{v \in Q} h'((2\pi)^{-\frac{p}{2}} \sum_{i=1}^{n} v_i - n)\right) \left(n^{1/2} \left(2\pi\right)^{-p/2}\right).$$

Define

$$\nu\left(\epsilon, p\right) := \left(4\Phi_p/\epsilon + 1\right)^{-1}$$

and

$$\beta\left(\epsilon,p\right):=\min_{v\in Q}\frac{\left(\epsilon-2\Phi_p\nu\left(\epsilon,p\right)\right)^2}{2\mathrm{Var}\left(h\left(g\left(v,\Theta\right)\right)\right)}.$$

We want to establish an analogue result of Proposition 2 of Wang and Ahmed (2008). It is straight to verify that the assumptions (C1) to (C3) and (C5) in Wang and Ahmed (2008) hold. For (C4), it suffices to show the derivative of $h\left(g\left(\cdot,\xi\right)\right)$ is bounded for any $v\in Q$. Especially, if this upper bound is independent of ξ, then we are done.

Lemma 4.24. *For any $\xi\in\Omega$, we have*

$$\Big|h\left(g\left(v_1,\xi\right)\right)-h\left(g\left(v_2,\xi\right)\right)\Big|\leq\Phi_p\|v_1-v_2\|,\quad\forall v_1,v_2\in Q$$

where $\Phi_p=\left(\max_{v\in Q}h'\left(\left(2\pi\right)^{-\frac{p}{2}}\sum_{i=1}^n v_i-n\right)\right)\left(n^{1/2}\left(2\pi\right)^{-p/2}\right)$.

Proof. Note that

$$\begin{aligned}\max_{v\in Q}\left\|\frac{\partial}{\partial v}h\left(g\left(v,\xi\right)\right)\right\| &= \max_{v\in Q}h'\left(g\left(v,\xi\right)\right)\|L\left(\xi\right)\| \\ &\leq \max_{v\in Q}h'\left(g\left(v,\xi\right)\right)\left(n^{1/2}\left(2\pi\right)^{-p/2}\right).\end{aligned}$$

Under the assumptions, h' is a non-decreasing function and then

$$\begin{aligned}h'\left(g\left(v,\xi\right)\right) &= h'\left(L\left(\xi\right)^Tv-n\right) \\ &\leq h'\Big(\left(2\pi\right)^{-\frac{p}{2}}\sum_{i=1}^n v_i-n\Big).\end{aligned}$$

Choosing $\Phi_p=\left(\max_{v\in Q}h'\left(\left(2\pi\right)^{-\frac{p}{2}}\sum_{i=1}^n v_i-n\right)\right)\left(n^{1/2}\left(2\pi\right)^{-p/2}\right)$, we have the result. □

Next we give the result of convergence rate and sample size determination for Kiefer-Wolfowitz dual problem. Given $\nu > 0$, define a finite subset of Q_ν of Q such that for any $v \in Q$, there exists $v' \in Q_\nu$satisfying $\|v - v'\| \leq \nu$. Then $|Q_\nu| \leq (D/\nu)^n$. Define

$$\nu(\epsilon, p) := (4\Phi_p/\epsilon + 1)^{-1}$$

and

$$\beta(\epsilon, p) := \min_{v \in Q} \frac{(\epsilon - 2\Phi_p \nu(\epsilon, p))^2}{2\mathrm{Var}(h(g(v, \Theta)))}.$$

Theorem 4.25. *For Kiefer-Wolfowitz dual problem, given $\epsilon > 0$, we have (1) convergence rates:*

$$\mathbb{P}\left(S_N^0 \subset S^\epsilon\right) \geq 1 - \left(\frac{D}{\nu(\epsilon, p)}\right)^n e^{-N\beta(\epsilon, p)};$$

(2) estimate of the sample size for $\mathbb{P}(S_N^0 \subset S^\epsilon) \geq 1 - \alpha$ to hold:

$$N \geq \frac{1}{\beta(\epsilon, p)} \log\left(\frac{1}{\alpha}\left(\frac{D}{\nu(\epsilon, p)}\right)^n\right).$$

Proof. For convenience, let us write $\tilde{g}(v) = \mathbb{E}(h(g(v, \Theta)))$ and $\tilde{g}_N(v) = \sum_{i=1}^N h(g(v, \xi_i))/N$. Denote the large deviations rate function as

$$I(u) := \sup_{s \in \mathbb{R}} \left\{su - \log \mathbb{E}\left(e^{sZ}\right)\right\}.$$

Note that

$$\begin{aligned} \mathbb{P}\left(S_N^0 \subset S^\epsilon\right) &\geq 1 - \mathbb{P}(\exists v \in Q \text{ s.t. } \tilde{g}_N(v) - \tilde{g}(v) > \epsilon) \\ &\geq 1 - \mathbb{P}(\exists v \in Q_\nu \text{ s.t. } \tilde{g}_N(v) - \tilde{g}(v) > \epsilon - 2\Phi_p \nu(\epsilon, p)) \end{aligned}$$

$$
\begin{aligned}
&\geq 1 - \sum_{v \in Q_{\nu(\epsilon,p)}} e^{-NI_v(\epsilon - 2\Phi_p \nu(\epsilon,p))} \\
&\geq 1 - |Q_{\nu(\epsilon,p)}| e^{-Nb(\epsilon,p)} \\
&\geq 1 - \left(\frac{D}{\nu(\epsilon,p)} \right)^n e^{-Nb(\epsilon,p)},
\end{aligned}
$$

where $b(\epsilon,p) := \min_{v \in Q} I_v(\epsilon - 2\Phi_p \nu(\epsilon,p))$. By Assumptions (C3) and (C5),

$$b(\epsilon,p) \geq \min_{v \in Q} \frac{(\epsilon - 2\Phi_p \nu(\epsilon,p))^2}{2\text{Var}(h(g(v,\Theta)))}$$

To get $\mathbb{P}(S_N^0 \subset S^\epsilon) \geq 1 - \alpha$, it is sufficient to set

$$\left(\frac{D}{\nu(\epsilon,p)} \right)^n e^{-N\beta(\epsilon)} \leq \alpha$$

and the result follows. □

Definition 4.26. (ϵ-optimality) A feasible point $\bar{v}$ is called ϵ-optimal of the problem, if for any feasible point v and an objective function f_0,

$$f_0(v) \geq f_0(\bar{v}) - \epsilon \qquad \text{for some } \epsilon \geq 0.$$

The above result reveals that: (1) The probability that an optimal solution to sample average approximation is an ϵ-optimal solution to the original problem goes to 1 exponentially fast as the size of randomly selected constraints N increases; (2) Even with exponential convergence, if $\beta(\epsilon,p)$ is small, the convergence can be slow and the estimated sample size can be large.

Remark 4.27. For certain cases, the parameters D, Φ_p and $\nu(\epsilon,p)$ are easy to evaluate. According to Theorem 4.5, the set Q can be chosen as a n-

dimensional box $[0, u_1] \times ... \times [0, u_n]$, where

$$u_i = 1 / \inf_{\theta \in \mathrm{conv}(y)} L_i(\theta).$$

If the penalty function is $h(t) = (\max\{0, t\})^2$, then we obtain $D = \|u\|_2$ and

$$\Phi_p = 2\Big((2\pi)^{-\frac{p}{2}} \sum_{i=1}^{n} u_i - n\Big) \Big(n^{1/2} (2\pi)^{-p/2}\Big).$$

4.4.0.3 Variance Reduction Method

The sample average approximation estimator $\hat{f}_N$ can be viewed a sample statistic based on $\xi_1, ..., \xi_N$. The risk of $\hat{f}_N$ comes from two parts: bias $\mathbb{E}\left(\hat{f}_N\right) - f^*$ and variance. Due to the results in Theorem 4.19, for large sample size N, the bias of $\hat{f}_N$ is negligible. Now the only concern is its variance. One possible way of reducing the variance is to solve sample average approximation problem M times on the different samples and take the average of their solutions. This is known as batch means method in numerical optimization, e.g. Norkin, Pflug, and Ruszczyński (1998), Mak, Morton, and Wood (1999) and Linderoth, Shapiro, and Wright (2006).

Let $\hat{f}_N^1, ..., \hat{f}_N^M$ and $\hat{v}_N^1, ..., \hat{v}_N^M$ be the computed optimal values and the optimal solutions of sample average approximation problems. Now consider the average of those sample average approximation estimators

$$\hat{f}_{N,M} = \frac{1}{M} \sum_{m=1}^{M} \hat{f}_N^m$$

and

$$\hat{v}_{N,M} = \frac{1}{M} \sum_{m=1}^{M} \hat{v}_N^m.$$

Theorem 4.28. *(Consistency) For Kiefer-Wolfowitz dual problem and a fixed M, $\hat{f}_{N,M} - f^* \to 0$ almost surely, as $N \to \infty$.*

Proof. By the triangule inequality,

$$\hat{f}_{N,M} - f^* = \left|\hat{f}_{N,M} - f^*\right| \leq \frac{1}{M}\sum_{m=1}^{M}\left|\hat{f}_N^m - f^*\right| = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{f}_N^m - f^*\right).$$

The result follows from Corollary 4.19 for each m. □

Theorem 4.29. *For fixed N and M, when the M batches $\left\{\xi^{1,m}, \xi^{2,m}, ..., \xi^{N,m}\right\}_{m=1}^{M}$ are i.i.d., $\hat{f}_{N,M}$ and $\hat{v}_{N,M}$ has smaller risk than $\hat{f}_N$ and $\hat{v}_N$.*

Proof. Under the assumption, $\hat{f}_N^1, ..., \hat{f}_N^M$ are independent and have the same distribution. Observe that

$$\begin{aligned}\mathbb{E}\left[\left(\hat{f}_{N,M} - f^*\right)^2\right] &= \left(\mathbb{E}\left(\hat{f}_{N,M}\right) - f^*\right)^2 + \mathrm{Var}\left(\hat{f}_{N,M}\right) \\ &= \left(\mathbb{E}\left(\hat{f}_N\right) - f^*\right)^2 + \frac{1}{M}\mathrm{Var}\left(\hat{f}_N\right) \\ &\leq \mathbb{E}\left[\left(\hat{f}_N - f^*\right)^2\right].\end{aligned}$$

Applying the same arguments on $\hat{v}_{N,M}$, we prove the second part. □

4.4.0.4 Testing for Optimality

Recall that from Chapter 2, Kiefer-Wolfowitz primal problem has unique solution $\hat{F}_n$ that are supported on at most n points. Hence, for finite sample size N, sample average approximation estimator $\hat{v}_N$ is infeasible almost surely, so is $\hat{v}_{N,M}$. If $\hat{v}_{N,M}$ is close to the boundary of the feasible set: $\mathbb{E}\left(h\left(g\left(\hat{v}_{N,M}, \Theta\right)\right)\right) \leq \epsilon$ for some prespecified threshold $\epsilon > 0$, then we conclude that $\hat{v}_{N,M}$ is close to

v^*. In practice, given a sample $\xi_1, ..., \xi_{N'}$, it suffices to check

$$\frac{1}{N'} \sum_{i=1}^{N'} h\left(g\left(\hat{v}_{N,M}, \xi_i\right)\right) \leq \epsilon.$$

4.5 Constraints Resampling

Inspired by Breiman (1996), we propose a bagging-like algorithm to approximate Kiefer-Wolfowitz MLE. In contrast to Breiman's algorithm, the repeated samples are taken from the constraints of the optimization problem instead of the training data set. Like sample average approximation methods, our algorithm preserves the convexity of the problem as well.

Breiman's bagging algorithm is briefly review first. Consider a regression problem. Suppose we make prediction $\hat{f}\left(x\right)$ at input x on the training data $\left\{\left(x_i, y_i\right); 1 \leq i \leq n\right\}$. In the same spirit of the variance reduction method in sample average approximation, Bagging averages the prediction over a collection of bootstrap samples, so that its variance is reduced. For each bootstrap sample from the training data set, we fit a regression model $\hat{f}^{*b}\left(x\right)$. The bagging estimate is defined by

$$\hat{f}_{bag}\left(x\right) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}\left(x\right),$$

where B is a prespecified positive integer, which represents the number of bootstrap samples used. It can be showed that for any model or estimator $\hat{f}$, the bagging version always has a lower mean-squared error than $\hat{f}$.

Now we return to the estimation problem of Kiefer-Wolfowitz MLE. Recall that from Section 4.4, by introducing the penalty function h, a Kiefer-

Wolfowitz dual problem can be reformulated as a stochastic programming problem

$$\min_{v \in Q} - \sum_{i=1}^{n} \log(v_i)$$
$$\text{subject to} \qquad \mathbb{E}(h(g(v,\Theta))) \leq 0.$$

Let $\left\{\xi^{1,m}, ..., \xi^{N,m}\right\}$ be a batch that independently drawn from a probability distribution P and $\xi^{i,m} \overset{iid}{\sim} \mu(\cdot)$ for any m. Let F_N be nonparametric empirical Bayes estimator based on a batch with size N. Under the squared loss, the Bayes decision rule is denoted as

$$d(y; F_N) = \mathbb{E}_{F_N}(\theta | y).$$

In below, we use superscript m to emphasize the dependence of decision rule and batch sample. Our aggregated estimator is then defined as

$$d_A(y) = \mathbb{E}_P(d(y; F_N))$$

and its empirical estimate

$$\hat{d}_A(y) = \frac{1}{M} \sum_{m=1}^{M} d^m(y; F_N).$$

The average error e in $d(y; F_N)$ is

$$e = \mathbb{E}_P \mathbb{E}_{Y,\theta} ||\theta - d(Y; F_N)||^2.$$

Define the error in the aggregated decision rule d_A to be

$$e_A = \mathbb{E}_{Y,\theta} \|\theta - d_A(Y)\|^2.$$

The following result is an analogue to Breiman (1996). It shows that making predictions on multiple samples has smaller risk than do it once.

Theorem 4.30. *$e_A \leq e$.*

Proof. Using the inequality $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$ gives

$$\begin{aligned} e &= \mathbb{E}\|\theta\|^2 - 2\mathbb{E}(\theta d_A(Y)) + \mathbb{E}_{Y,\theta}\mathbb{E}_P \|d(Y; F_N)\|^2 \\ &\geq \mathbb{E}_{Y,\theta} \|\theta - d_A(Y)\|^2 = e_A. \end{aligned}$$

□

Remark 4.31. If the number of repeated samples is one, our constraints resampling algorithm is equivalent to sample average approximation (without variance reduction).

Remark 4.32. Sampling without replacement is used in our algorithm. The classical bagging algorithm involves sampling with replacement. If we sample N constraints with replacement and use them for our optimization problem where only u of them are unique, then it is equivalent to simply using u unique constraints. Hence, sampling without replacement is sufficient.

4.6 Cutting-Plane Methods

In this section we discuss how to apply Kelley's cutting-plane method to solve Kiefer-Wolfowitz dual problems. Recall that in Kelley's method, the query point is chosen to be the optimal solution to the current polyhedron approximation. If the optimal point of the current polyhedral approximation of the problem is optimal for the original problem then Kelley's algorithm terminates. This is indeed the case for the dual problem, since the solution of Kiefer-Wolfowitz dual problem lies on the boundary of the feasible set. Therefore, Kelley's method can be an efficient alternative algorithm for the dual problems.

The appropriate choice of the target set is crucial to the success of the cutting-plane methods. From the convergence analysis in Theorem 4.34, the number of iterations is controlled by the ratio of the sizes of initial polyhedron and target set. Quite often the set of the optimal solutions is chosen to be the target set (Boyd and Vandenberghe, 2007). If that is the case, the target set is just a singleton, since the dual problem is strictly convex. We can expect using a sequence of polyhedrons outer-approximate to a point (In practice, this can be a ball centered at the dual solution with a small radius.) can take very long. On the other hand, if the feasible set is selected to be the target set, the sizes of initial polyhedron and target set can not be very different.

We choose the target set based on the following simple result. Denote the feasible set of Kiefer-Wolfowitz dual problem

$$A = \bigcap_{\theta \in \text{conv}(y)} \left\{ v \in \mathbb{R}^n_{\geq 0} : \sum_{i=1}^{n} v_i L_i(\theta) - n \leq 0 \right\}$$

and a simpler set with finitely many constraints

$$B = \bigcap_{\theta \in O} \left\{ v \in \mathbb{R}^n_{\geq 0} : \sum_{i=1}^{n} v_i L_i(\theta) - n \leq 0 \right\},$$

where $|O| < \infty$. From the construction, $A \subset B$. For convenience, we name the minimization problems

$$\min_{v \in \mathbb{R}^n_{\geq 0}} - \sum_{i=1}^{n} \log v_i$$

over the set A and B as Problem A and Problem B respectively. Let p^*_A and p^*_B be the optimal values to Problems A and B.

Fact 4.33. *If $v^*_B \in A$, then Problems A and B are equivalent.*

Proof. It is clear that $p^*_B \leq p^*_A$. If the optimal point $v^*_B \in A$, then the two problems have the same optimal point and the optimal value, therefore Problems A and B are equivalent. □

From this simple observation, we take the set A as the target set: If the algorithm does not stop in k step, then all the query points are infeasible points and we can always construct cutting-planes. Otherwise, the query point need to be on the boundary of A and it is the optimal point to Problem A.

Next we discuss how to construct cutting-planes. Let O_0 be the initial set of grid points and define a polyhedron as

$$\mathcal{P}_0 = \bigcap_{\theta \in O_0} \left\{ v \in \mathbb{R}^n_{\geq 0} : \sum_{i=1}^{n} v_i L_i(\theta) - n \leq 0 \right\}.$$

Certainly, we have the target set A is contained in $\mathcal{P}_0$. For convenience, we write the constraint function $g : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ as

$$g(v, \theta) = \sum_{i=1}^{n} L_i(\theta) v_i - n.$$

If the algorithm does not stop in k step, the query point is selected to be the

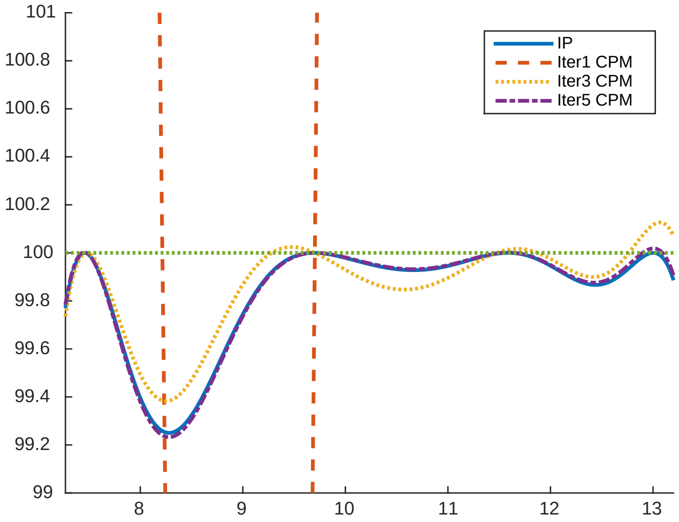

Figure 2: Plot of the constraint function $\tilde{g}(\theta) := L(\theta)^T v$ *when the dimension of parameter space is one. Blue curve represents the constraint function using interior-point method with an equally spaced fine grid, indicated as IP; the other curves correspond the constraint function using the cutting-plane method in different iterations, indicated as Iteri CPM.*

minimizer of the problem restricted on $\mathcal{P}_k$:

$$v^{(k)} = \arg\min_{v \in \mathcal{P}_k} - \sum_{i=1}^{n} \log v_i, \quad k = 0, 1, 2, \ldots$$

This problem is a finite convex programming which can be efficiently solved by standard techniques in convex programming, for example, interior-point methods. Also, the query point $v^{(k)}$ must be infeasible, otherwise the algorithm stops. Then there exists at least one violated constraint

$$g\left(v^{(k)}, \theta\right) > 0, \qquad \text{for } \theta \in \operatorname{conv}(y).$$

To locate those violated constraints, we chose O_k to be the set of local maxima of $g\left(v^{(k)}, \cdot\right)$ such that $g\left(v^{(k)}, \theta\right) > 0$. In such a way, we made a compromise between convergence speed and computational cost: If only slightly violated constraints are added, the algorithm can be slow. If too many constraints are added, the localization polyhedrons become complicated and the computational cost of obtaining query points increases. At last, we construct cuts

$$g\left(v^{(k)},\theta\right)+L\left(\theta\right)^{T}\left(v-v^{(k)}\right)\leq 0,\qquad \theta\in O_{k}$$

which can be further reduced to

$$L\left(\theta\right)^{T}v-n\leq 0,\qquad \theta\in O_{k}. \tag{4.19}$$

Algorithm 2 Kelley's cutting-plane method for Kiefer-Wolfowitz dual problem

```
Ensure: Initial polyhedron $\mathcal{P}_0$ that constraints the feasible set $A$
$k \leftarrow 0$
while not converged do
    Obtain query point $v^{(k)} \in \mathcal{P}_k$:
        $v^{(k)} = \arg\min_{v\in\mathcal{P}_k} -\sum_{i=1}^{n} \log v_i$
    Collect the set of violated constraints $I_k$
    Construct cutting-planes $L(\theta_j)^T v - n \leq 0,\ j \in I_k$
    Update localization polyhedron
        $\mathcal{P}_{k+1} \leftarrow \mathcal{P}_k \cap \left\{v : L(\theta_j)^T v - n \leq 0,\ j \in I_k\right\}$
    $k \leftarrow k+1$
end while
```

A convergence analysis of the cutting-plane method is provided below. Assume the target set A contains a ball B_r with radius r and $\mathcal{P}_0$ is contained in a ball B_R with radius R. At each step of the cutting-plane method the volume of $\mathcal{P}_k$ is reduced at least by a factor $\gamma < 1$.

Theorem 4.34. *The cutting-plane algorithm terminates at finite steps.*

Proof. Suppose the algorithm does not stop in k steps. Then

$$\operatorname{vol}(B_r) \leq \operatorname{vol}(\mathcal{P}_k) \leq \gamma^k \operatorname{vol}(\mathcal{P}_0) \leq \gamma^k \operatorname{vol}(B_R),$$

where the function $\operatorname{vol}_n(B_R)$ computes the volume of a Euclidean ball of radius R in $n-$dimensional Euclidean space

$$\operatorname{vol}_n(B_R) = \frac{\pi^{n/2}}{\Gamma(n/2+1)} R^n$$

and Γ is Euler's gamma function. Solving this inequality, we obtain

$$k \leq \frac{n \log(R/r)}{\log(1/\gamma)}.$$

□

4.6.0.1 Cuts Adding

One of the challenging of using cutting-plane methods here is adding cuts. In finite programming, this will not be an issue: Because the number of constraints is finite, given an infeasible query point $v^{(k)}$, we can simply compute $L(\theta_i)^T v - n$ for finitely many i and select those violated constraints. If the number of constraints is infinite, adding cuts eventually becomes a highly nonconvex programming problem:

$$\max_{\theta \in \operatorname{conv}(y)} L(\theta)^T v^{(k)}. \tag{4.20}$$

The objective function is a linear combination of Gaussian densities centered at observed data. General nonconvex problems are very difficult to solve globally: We either use some heuristic optimization methods, which are fast but do not guarantee a global solution or choose some global optimization methods often slow. For small p, the divide and conquer algorithms, such as branch and bound algorithms, can be used and they guarantee to find a global solution. For relatively large p, we need some heuristic approaches, for example, conducting gradient methods multiple times from different initial starting points.

4.6.0.2 Stopping Criteria

From the design of the algorithm, if a query point $v^{(k)}$ is feasible, then it must be the minimizer and the algorithm stops. However, such a feasibility test in semi-infinite programming is equivalent to solving a nonconvex programming problem and it is very difficult in general (López and Still, 2007). One of the possibilities is to reformulate the dual problem as a stochastic programming problem and justify the boundary condition. The details of the reformulation is given in Section 4.3.

Fact 4.35. *Suppose a continuous function $h : \mathbb{R} \to \mathbb{R}_+$ has support $(0, \infty)$:*

$$h(t) = \begin{cases} 0 & \textit{for all } \ t \in (-\infty, 0] \\ > 0 & \textit{for all } \ t \in (0, \infty) \end{cases}$$

and h is differentiable. Let $\xi_1, ..., \xi_N$ be a sequence of i.i.d. random variables defined in Section 4.4 and $\tilde{v}$ any point on the boundary of the feasible set for

Kiefer-Wolfowitz dual problem. Then

$$\frac{1}{N}\sum_{i=1}^{N} h\left(g\left(\tilde{v},\xi_i\right)\right) \to 0 \quad \textit{almost surely.}$$

Proof. A consequence of Theorem 4.9 and the law of large numbers. □

Remark 4.36. The algorithm terminates if the query point $v^{(k)}$ lies on the boundary of the feasible set. Since all the query points are infeasible, if $v^{(k)}$ is on the boundary, then it is the minimizer.

4.7 Computational Concerns

4.7.1 Minimum Volume Ellipsoid

In the stochastic approaches, we use random samples from a probability distribution defined on the convex hull $\mathrm{conv}\,(y)$ in $\mathbb{R}^p$. However, for $p > 3$, computing the convex hull of a finite set can be difficult: Even if the vertices of a convex polytope are given, construction of its faces is a non-trivial task (Avis, Bremner, and Seidel, 1997).

Instead, we may consider enlarge our search space by computing a minimum volume ellipsoid covering the finite set $\{y_1, ..., y_n\}$, which can be formulated as a convex programming problem (Vandenberghe and Boyd, 1998). Once the ellipsoid is found, sampling can be efficiently done: For example, if uniform sampling is considered, we can first generate points within the unit hypersphere and then rescale the points to the ellipsoid.

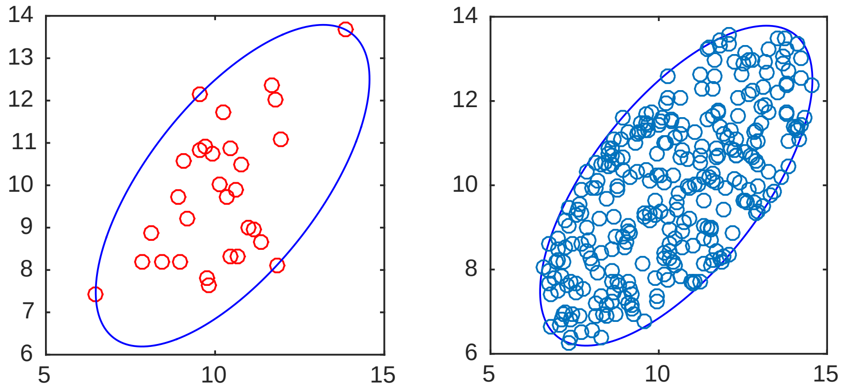

Figure 3: The left panel indicates the observed data and the corresponding minimum volume ellipsoid; the right panel shows a randomly generated dense grid within the ellipsoid.

4.7.2 Curse of Dimensionality

4.7.2.1 Gigantic Norm

Due to the nature of the statistical model and Kiefer-Wolfowitz MLE problem, we show that without extra assumptions, it is very difficult for any optimization algorithm handling the problems with large p. To illustrate the idea, we assume the Gaussian noise has mean zero and covariance $\sigma^2 I_p$.

Fact 4.37. *Let v^* be the optimal solution of Kiefer-Wolfowitz dual problem, then $\|v^*\| \geq n^{1/2}\,(2\pi)^{-p/4}\,\sigma^{p/2}$.*

Proof. We know v^* lies on the boundary of the dual feasible set. Let $r\,(\theta)$ denote the distance from the origin to the hyperplane $L\,(\theta)^T\,v - n = 0$. Then

$\|v^*\|$ should be greater than $\inf_\theta r(\theta)$. Observe that

$$r(\theta) = \frac{n}{\|L(\theta)\|} \geq n^{1/2} (2\pi)^{p/2} \sigma^p, \qquad \forall \theta \in \mathbb{R}^p,$$

since each $L_i(\theta)$ is bounded above by $(2\pi)^{-p/2} \sigma^{-p}$. The result follows. □

For fixed n, if p increases and σ is relatively large, then the optimal point v^* moves away from the origin exponentially fast. For example, $n = 100$, $\sigma^2 = 3$ and $p = 20$, then $\|v^*\|$ is at least 5.66×10^{13}.

4.7.2.2 Objective Function

The objective function of Kiefer-Wolfowitz dual problem has the form

$$-\sum_{i=1}^{n} \log v_i.$$

From the discussion in the previous section, we know that for fixed n and relatively large σ, the norm of the optimal solution can be very large and this leads to a so-called vanishing gradient problem for any gradient based algorithms: The gradient $-(1/v_1, ..., 1/v_n)$ will be so small that the product of step size and the gradient stops changing its value in the update. In experiments, if the logarithm objective function is used and p goes beyond 10, the solutions found by gradient based algorithms are interior points of the dual feasible set but not on the boundary. A simple way to fix this problem is to apply a monotone transformation on the objective function. If an exponential function is applied, we will have the geometric mean objective function

$$-\exp\left(\frac{1}{n}\sum_{i=1}^{n} \log v_i\right) = -\left(\Pi_{i=1}^{n} v_i\right)^{1/n}.$$

Previously, with $(\Sigma_{noise})_{ij} = 3 \times 0.8^{|i-j|}$, the algorithm can only work up to $p = 8$. This simple change can make the algorithms able to deal the cases p up to 11 with the same covariance matrix. Certainly, if the magnitude of covariance gets smaller, the algorithm can handle higher value of p. For example, with $(\Sigma_{noise})_{ij} = 0.8^{|i-j|}$ and the modified objective function, the algorithm are able to handle the cases p up to 15.

4.8 Simulations

In this section, we study the performance of our proposed methods through numerical experiments. To evaluate the performance, we calculate MSE for each method

$$\frac{1}{n}\sum_{i=1}^{n} \|\theta_i - \hat{\theta}_i\|_2^2.$$

The tables below report MSE of the cutting-plane method denoted CPM, sample average approximation estimator using variance reduction techniques denoted SAA, constraints resampling method denoted CR compared to the naïve Bayes methods proposed by Dicker and Zhao (2014) denoted naiveBayes and the naïve estimator, $d\left(y_i\right) = y_i$ for all i, denoted naive. As a reference, we also provide the performance of some oracle Bayes estimators: The Bayes estimator with a known prior denoted as Oracle and one with the marginal of the true prior denoted as OM.

Due to the slow convergence, stochastic gradient methods are not implemented here. Although in practice, heuristic choice of step-size and early stopping criteria can speed the algorithms up.

In all the experiments, we choose $p \in \{5, 10\}$ and $n = 100$. For sample average approximation and constraints resampling, we choose the sample size around $\sqrt{np}$ and the ensemble size 10. The results are based on 30 replications.

4.8.1 Experiment 1

We study a simple setup in favor of the naïve Bayes method. Assume the Gaussian noises ϵ_i have mean zero and covariance I_p. Three types of priors are considered:

Case 1: (Normal prior) We let $\theta_i \overset{iid}{\sim} N_p(\mu_s, \Sigma_s)$ with $\mu_s = (10, ..., 10)$ and $(\Sigma_s)_{ij} = 0.9^{|i-j|}$, where $i, j \in \{1, ..., n\}$.

Case 2: (Two-point uniform prior) We generate θ_i from two-point uniform distribution with support $\{5 \times \mathbf{1}_p, 10 \times \mathbf{1}_p\}$.

Case 3: (Multivariate t prior) We generate t_i from multivariate t distribution with mean $\mu_t = (10, ..., 10)$, correlation matrix $(\Sigma_t) = 0.9^{|i-j|}$ and df $= 3$. To have the similar signal-to-noise ratio compared to the previous experiments, we scale t_i by $1/\sqrt{3}$ and shift them by 10 units in every dimension

$$\theta_i \leftarrow (10, ..., 10) + 1/\sqrt{3} \cdot t_i, \qquad t_i \overset{iid}{\sim} t_{\text{df}}(\mu_t, \Sigma_t).$$

p	Case	Oracle	OM	CPM	SAA	CR	naiveBayes	naive
5	1	1.44	2.72	3.69	2.64	**2.16**	2.74	4.98
	2	0.00	0.65	1.00	2.06	1.30	**0.93**	4.99
	3	1.06	2.27	2.30	2.64	**2.08**	2.31	4.99
10	1	4.51	5.38	9.95	8.54	6.29	**5.52**	10.52
	2	0.56	1.53	5.75	6.66	4.32	**2.15**	10.12
	3	2.97	5.45	10.87	11.45	6.49	**4.74**	10.07

Table 1: MSE of various methods in Experiment 1

4.8.2 Experiment 2

We let the Gaussian noises ϵ_i have mean zero and covariance $\Sigma_{noise} = 3 \times 0.8^{|i-j|}$. The three types of signals are generated in the same ways as in Experiment 1.

p	Case	Oracle	OM	CPM	SAA	CR	naiveBayes	naive
5	1	3.90	3.98	5.97	5.90	**5.21**	9.17	14.68
	2	4.07	7.16	7.92	8.46	**7.61**	11.11	14.64
	3	4.00	5.00	5.28	6.25	**5.26**	9.13	15.06
10	1	8.07	8.12	22.99	23.28	**13.90**	18.97	30.16
	2	7.75	13.65	22.79	25.16	**17.72**	22.08	29.57
	3	6.23	8.99	20.89	21.62	**12.65**	17.44	29.26

Table 2: MSE of various methods in Experiment 2

From the experiments above, we observe that (1) If the magnitudes in the covariance matrix of the signals are relatively small so that the independence assumption in the naïve Bayes method likely holds and the naïve Bayes method is likely to have good performance. If the magnitudes in the covariance matrix of the signals are large, then the constraints resampling method outperforms the other methods. (2) It is interesting to see in Experiment 1: For $p = 5$, even in a setup in favor of the naïve Bayes method, our proposed methods are very competitive with the naïve Bayes method if not better. (3) For small p, the cutting-plane method is competitive with constraints resampling; for relatively large p, due to the difficulty of finding the global solutions of nonconvex problems, the performance of the cutting-plane method drops down quickly.

To sum up the strategy of choosing algorithms, if $p \geq 20$ and the magnitude of covariance is relatively large, the naïve Bayes method needs to be used; if

$p \geq 20$ and the magnitude of covariance is small, both naïve Bayes method and constraints resampling can be tried; if $p < 20$, constraints resampling is recommended.

Chapter 5

Nonparametric Empirical Bayes Regression

In this chapter we propose a new regression model, called the nonparametric empirical Bayes regression, so that it incorporates nonparametric empirical Bayes methods in the presence of explanatory variables.

In the past, various approaches have been tried to generalize the empirical Bayes framework to regression problems, e.g. Cohen et al. (2013), Fay III and Herriot (1979), Jiang and Zhang (2010), and Koenker (2015). In contrast to some of the previous approaches, our new model has a very simple form and inherits most of theoretical properties of nonparametric empirical Bayes procedure. Furthermore, unlike the methods based on the partial linear model, the parameter estimation procedure in our proposed regression model is equivalent to solving a convex optimization problem in function space and it can be efficiently solved by optimization techniques proposed in Chapter 4.

In Section 5.1, an introduction of Bayes linear regression is given. In Section 5.2, we describe nonparametric empirical Bayes regression and its theoretical properties. In Section 5.3, the famous baseball data set is analyzed using our proposed method.

5.1 Bayes Linear Regression

Bayes linear regression model was first proposed in the landmark paper by Lindley and Smith (1972). In this approach, regression coefficients are viewed as a random variable following a prior distribution. To be precise, we consider a regression problem with the input pairs $(x_1, y_1), ..., (x_n, y_n)$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Moreover, we assume

$$y|\theta \sim N(X\theta, \Sigma) \quad \text{and} \quad \theta \sim N(\mu_0, S_0),$$

where X is a $n \times p$ fixed design matrix and Σ is known. The posterior distribution of θ is

$$\theta|y \sim N(Dd, D),$$

where

$$D = X^T \Sigma^{-1} X + S_0^{-1}$$

and

$$d = X^T \Sigma^{-1} y + S_0^{-1} \mu_0.$$

Thus, under the squared loss, the Bayes decision rule for θ is $\mathbb{E}(\theta|y) = Dd$ and the estimated model has the form

$$X\mathbb{E}(\theta|d) = X(Dd).$$

5.2 Nonparametric Empirical Bayes Regression

The assumption of a known prior in Bayes linear regression can be relaxed by using nonparametric empirical Bayes procedure. This can be viewed as a generalization of nonparametric empirical Bayes model that handles explanatory variables.

To illustrate the idea, let us consider the regression problem in Section 5.1 without specifying the prior distribution. Suppose the input pairs are $(x_1, y_1), ..., (x_n, y_n)$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Furthermore, we assume $y_i|\theta_i \overset{ind}{\sim} N\left(x_i^T \theta_i, 1\right)$ and $\theta_i \overset{iid}{\sim} F(\cdot)$, where X is a $n \times p$ design matrix with ith row equal x_i. The prior distribution function F can be estimated via Kiefer-Wolfowitz maximum likelihood procedure:

$$\tilde{F} := \tilde{F}_n = \arg\min_{F \in \mathcal{F}} - \sum_{i=1}^{n} \int_{\mathbb{R}} \varphi\left(y_i - x_i^T \theta\right) dF(\theta), \qquad (5.1)$$

where $\mathcal{F}$ is the class of all probability distributions and φ is the standard normal density function. The problem (5.1) is again strictly convex, so the solution $\tilde{F}_n$ exists and it is unique. The corresponding dual problem has the form:

$$\min_{v \in \mathbb{R}_+^n} - \sum_{i=1}^{n} \log v_i \qquad (5.2)$$

subject to

$$\sum_{i=1}^{n} \varphi\left(y_i - x_i^T \theta\right) v_i \leq n \quad \text{for all}\, \theta \in \mathbb{R}^p.$$

5.2.1 Theoretical Properties

The primal and the dual problems here possess similar forms to their counterparts in Kiefer and Wolfowitz's nonparametric maximum likelihood problem, so we expect the MLE $\tilde{F}$ enjoys most of the properties that $\hat{F}$ has. In fact, most of the results in this section can be proved in similar ways.

To list the results, we use similar notation in Chapter 4. Let

$$f_F\left(y,x\right)=\int_{\mathbb{R}}\varphi\left(y-x^T\theta\right)dF\left(\theta\right),$$

$$f_\theta=\left(\varphi\left(y_1-x_1^T\theta\right),...,\varphi\left(y_n-x_n^T\theta\right)\right)$$

and

$$f_F=\left(f_F\left(y_1,x_1\right),...,f_F\left(y_n,x_n\right)\right).$$

We can write the gradient function defined in (4.2) as

$$D(\theta;F)=\sum_{i=1}^{n}\frac{L_i\left(\theta\right)}{\mathcal{L}_i(F)}-n,$$

where $L_i\left(\theta\right)=\varphi\left(y_i-x_i^T\theta\right)$ and $\mathcal{L}_i\left(F\right)=\int_{\mathbb{R}}\varphi\left(y_i-x_i^T\theta\right)dF\left(\theta\right)$. We further write $\mathcal{L}(F)=(\mathcal{L}_1(F),...,\mathcal{L}_n(F))$.

Theorem 5.1. *Under the regularity conditions in Kiefer and Wolfowitz (1956), $\tilde{F}_n\Rightarrow F$.*

Proof. See Kiefer and Wolfowitz (1956). □

Theorem 5.2. *The following three statements are equivalent:*

1. *$\tilde{F}$ maximizes $\mathcal{L}(F)$.*
2. *$\tilde{F}$ minimizes $\sup_\theta D(\theta;F)$.*

3. $sup_\theta D(\theta; \tilde{F}) = 0$.

Proof. See the proof of Theorem 19 in Lindsay (1995). □

Theorem 5.3. *The MLE $\tilde{F}_n$ has no more than n points of support.*

Proof. See the proof of Theorem 4.1. □

Theorem 5.4. *The solution of the dual problem (5.2) lies on the boundary of the feasible set.*

Proof. See the proof of Theorem 4.7. □

Let $\theta^* := \theta^*(\tilde{F}) \in \mathbb{R}^p$ be a mode of the gradient function $D(\theta; \tilde{F})$. Then $\theta^* \in \text{supp}(\tilde{F})$. Let $\Lambda := \Lambda(\theta^*(\tilde{F}), \tilde{F})$ be a $p \times p$ diagonal matrix with $(\Lambda)_{ii} = L_i\left(\theta^*\right)/\mathcal{L}_i(\tilde{F})$.

Theorem 5.5. *If $X^T \Lambda X$ is invertible, then θ^* is equivalent to a weighted least squares estimator.*

Proof. By Theorem 5.2, the support points of $\tilde{F}$ are necessarily modes. Then for the mode θ^*, we have

$$\nabla D(\theta^*; \tilde{F}) = 0. \tag{5.3}$$

Since

$$\nabla L_i\left(\theta\right)/L_i\left(\theta\right) = \left(y_i - x_i^T\theta\right)x_i^T, \qquad \text{for each } i,$$

the equation (5.3) can be rewritten as

$$\sum_{i=1}^{n} \frac{\nabla L_i\left(\theta^*\right)}{\mathcal{L}_i(\tilde{F})} = \sum_{i=1}^{n} \frac{L_i\left(\theta^*\right)}{\mathcal{L}_i(\tilde{F})}\left(y_i - x_i^T\theta^*\right)x_i^T = 0. \tag{5.4}$$

Under the assumption, we obtain

$$\theta^* = \left(X^T \Lambda X\right)^{-1} X^T \Lambda y.$$

□

5.2.2 Nonparametric Empirical Bayes Personalized Regression

One natural generalization of the ideas of Bayes linear regression in nonparametric empirical Bayes setting is to fit the model:

$$x_i \mathbb{E}_{\tilde{F}}\left(\theta \right| \{x_i, y_i; 1 \leq i \leq n\}) = x_i \frac{\int_{\mathbb{R}} \theta \Pi_{i=1}^{n} \varphi\left(y_i - x_i^T \theta\right) d\tilde{F}\left(\theta\right)}{\int_{\mathbb{R}} \Pi_{i=1}^{n} \varphi\left(y_i - x_i^T \theta\right) d\tilde{F}\left(\theta\right)}, \quad i = 1, ..., n.$$

In the spirit of nonparametric empirical Bayes methods, we can enhance the flexibility of this model by fitting one regression line for one subject. The regression coefficients for each subject is jointly determined by all the training data via the maximum likelihood procedure. We call this type of models as nonparametric empirical Bayes personalized regression.

Let the decision rule $d\left(y\right)$ be separable: $d\left(y\right) = \left(d_1\left(y_1\right), ..., d_n\left(y_n\right)\right)$. It is sufficient to consider univariate Bayes estimators. Under the squared loss, the Bayes rule d is the posterior mean. Hence, the model for any subject has the form

$$x \mathbb{E}_{\tilde{F}}\left(\theta | x, y\right) \quad = \quad x \frac{\int_{\mathbb{R}} \theta \varphi\left(y - x^T \theta\right) d\tilde{F}\left(\theta\right)}{\int_{\mathbb{R}} \varphi\left(y - x^T \theta\right) d\tilde{F}\left(\theta\right)}. \tag{5.5}$$

The other way of viewing it is to rewrite the regression model (5.5) as

$$\sum_{j=1}^{J} w_j x \theta_j,$$

where $\theta_j \in \text{supp}(\tilde{F})$, $J = |\text{supp}(\tilde{F})|$, $\{\tilde{f}_j; 1 \leq j \leq J\}$ are the corresponding mass of $\tilde{F}$ and

$$w_j = \frac{\varphi\left(y - x^T\theta_j\right)\tilde{f}_j}{\sum_{j=1}^J \varphi\left(y - x^T\theta_j\right)\tilde{f}_j}, \qquad \sum_{j=1}^J w_j = 1.$$

By Theorem 5.5, any element in $\operatorname{supp}(\tilde{F})$ is equivalent to a weighted least squares estimator and the model (5.5) can be understood as a convex combination of weighted least squares regressions.

5.2.3 Implementation

The maximum likelihood procedure (5.1) is equivalent to solving a convex optimization problem in the space of probability distribution functions on $\mathbb{R}^p$. All the methods developed in Chapter 4 can be directly applied.

5.2.3.1 Maximum Volume Inscribed Ellipsoid

If the stochastic methods in Chapter 4 are considered, some sampling schemes need to be developed. For Kiefer-Wolfowitz MLE problem, the support of $\hat{F}$ is located within the convex hull of the observed data. So we can define a probability distribution on a set that contains the convex hull. However, we do not have this property in the regression setting. Instead, we can impose a requirement that all the predicted values are between $l = \min\{y_i\}$ and $u = \max\{y_i\}$:

$$l\mathbf{1}_n \leq X\theta \leq u\mathbf{1}_n$$

and define the probability distribution on the polytope described by a set of linear inequalities. It is known that sampling from a convex polytope is difficult (Kannan, Lovász, and Simonovits, 1997), especially when the dimension p is

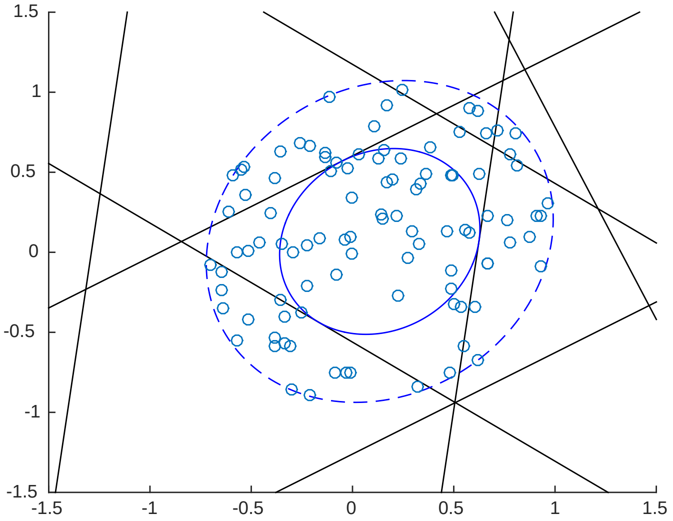

Figure 4: The maximum volume inscribed ellipsoid (solid) that lies inside the convex polytope and the corresponding enlarged ellipsoid (dashed). The solid black lines represent the constraints $l \leq x_i^T \theta \leq u$ for $i = 1, ..., n$.

large. On the other hand, sampling from an ellipsoid can be done efficiently, so we turn to an alternative problem of finding the ellipsoid with the maximum volume that lies inside the convex polytope. This problem can be formulated as a convex programming problem (Vandenberghe and Boyd, 1998) and efficiently solved.

However, the maximum volume inscribed ellipsoid is only a subset of the convex polytope and it does not necessarily contain all the support points of $\hat{F}$. To fix this problem, we can enlarge the ellipsoid and obtain a dilated one with the same center: Suppose the maximum volume inscribed ellipsoid has the form $\theta^T A\theta = 1$ and the enlarged and dilated ellipsoid is $\theta^T (\delta A)\, \theta = 1$, where $0 < \delta < 1$. Choosing $\delta = 1/2$ provides us satisfied results in our experiments.

5.3 Baseball Batting Average Prediction

In this section, we study the Baseball data, originally analyzed by Brown (2008); Cohen et al. (2013); Jiang and Zhang (2010) and Koenker and Mizera (2014) and apply our proposed nonparametric empirical Bayes regression model to predict the batting average in the second half season.

The data consists of batting records of each major league player in 2005. For each player, we are given the number of at bats N_{1i} and the number of hits H_{1i} in the first half of season. In addition, for every player it is known whether he is a pitcher or a batter. The goal is to use midseason batting averages $R_{1i} = H_{1i}/N_{1i}$ for $i = 1, ..., n_1$ to predict second half averages $R_{2i} = H_{2i}/N_{2i}$ for $i = 1, ..., n_2$. Let S_1 and S_2 denote the set of players in the first half of season and the second half of season respectively. All $n_1 = |S_1| = 567$ players with more than ten bats in the the first half season are used to predict performance of $n_2 = |S_1 \cap S_2| = 499$ players who also has more than ten bats in the second half season.

A reasonable model for the data is $H_{ti} \sim Bin\left(N_{ti}, p_i\right)$ conditional on N_{ti}, where $t = 1, 2$, $i = 1, ..., n_t$, and p_i is the batting probability of the ith player. Brown (2008) suggested a transformation to induce approximate normality

$$Y_{ti} = \arcsin\left(\sqrt{\frac{H_{ti} + 1/4}{N_{ti} + 1/2}}\right) \approx N\left(\theta_i, \sigma_{ti}^2\right),$$

where $\theta_i = \arcsin\sqrt{p_i}$ and $\sigma_{ti}^2 = 1/4N_{ti}$.

For the evaluation, the naïve estimator $\tilde{Y}_{2i} = Y_{1i}$ is used as a benchmark. Adopting the notation in Brown (2008), we compute the normalized total sum of prediction errors

$$TSE = \frac{\sum_{i=1}^{n_2} \left[\left(Y_{2i} - \hat{Y}_{2i} \right)^2 - \sigma_{2i}^2 \right]}{\sum_{i=1}^{n_2} \left[\left(Y_{2i} - \tilde{Y}_{2i} \right)^2 - \sigma_{2i}^2 \right]}, \tag{5.6}$$

where $\hat{Y}_{2i}$ denotes the predictions for the second half of season using various methods. Let z_i denote a vector of covariates which consists of an indicator variable of whether the player is a pitcher and the number of at bats and their interaction. We consider a linear regression model

$$Y_i = z_i^T \theta_i + \epsilon_i, \qquad \text{where } \epsilon_i \sim N\left(0, \sigma_i^2\right), \theta_i \overset{iid}{\sim} F\left(\cdot\right).$$

The table below reports TSE (5.6) of nonparametric empirical Bayes regression denoted NPEBReg, compared to seven procedures considered in Jiang and Zhang (2010) and Koenker and Mizera (2014). The first two of them LSE and WLSE are regression based methods; the third estimator EBJS is a James-Stein version of WLSE; the fourth and the fifth estimators are the EM implementations of the GMLEB estimator proposed by Jiang and Zhang; the remaining two are the interior-point implementations of the GMLEB estimator proposed by Koenker and Mizera.

The result of NPEBReg is slightly inferior to EBJS and WGMLEBEM but its performance is better than the other five methods. Possibly this is because the true prior has an approximate normal shape. In contrast to the GMLEB based methods, the parameter estimation for NPEBReg is equivalent to solving a convex problem and this can be done efficiently with the methods discussed in Chapter 4.

LSE	WLSE	EBJS	GMLEBEM	WGMLEBEM	GMLEBIP	WGMLEBIP	NPEBReg
0.240	0.204	**0.171**	0.178	0.177	0.194	0.206	0.178

Table 3: TSE in baseball batting average prediction

Chapter 6

Nonparametric Empirical Bayes Prediction under Quantile Loss

Inspired by Mukherjee et al. (2015), in this chapter we provide a case study for the problem of multiple prediction with the absolute loss. We show our Kiefer-Wolfowitz MLE based method outperforms their method in the experiments and a real data example.

Today there are many decision problems involving alternative loss functions other than the squared loss. For example, in robust regression, check loss and Huber's loss are widely used to avoid the dangers posed by outliers. In certain applications, alternative loss functions are selected simply because they have real-world interpretations. This choice of alternative loss functions can immediately make some empirical Bayes methods not applicable. For example, when the regular squared loss is chosen and the sampling distribution is a member of exponential family, the Tweedie's formula expresses Bayes rule in terms of the marginal distribution. However, such functional relations between the quantity of interest and the marginal distribution are quite rare and for many widely used loss functions the expressions do not exist. There is no such a limitation for Kiefer-Wolfowitz MLE based empirical Bayes method, since the estimated prior distribution is obtained from maximum likelihood procedure and independent from the loss function.

In Section 6.1, we introduce notation and setup of the problem. In Section 6.2, we describe an approach using Kiefer-Wolfowitz MLE based empirical Bayes method and prove consistency and asymptotic optimality of the nonparametric empirical Bayes estimators. In Section 6.3, we repeat the experiments designed by Mukherjee et al. (2015). A real data example is provided in Section 6.4, we use the proposed nonparametric empirical Bayes method to predict the monthly demand for a manufacturing company.

6.1 Basic Setup

Motivated by Mukherjee et al. (2015), we consider the following prediction problem: Suppose we have n products indexed by i, and for each i, the observed historical demand X_i and the unobserved future demand Y_i are distributed according to a normal distribution with an unknown mean θ_i,

$$X_i = \theta_i + \sqrt{\nu_{p,i}}\epsilon_{1,i} \qquad \text{for } i = 1, 2, ..., n$$

$$Y_i = \theta_i + \sqrt{\nu_{f,i}}\epsilon_{2,i} \qquad \text{for } i = 1, 2, ..., n,$$

where the noise $\{\epsilon_{j,i} : j = 1, 2; i = 1, ..., n\}$ are i.i.d. from a standard normal distribution, and the past and future variances $\nu_{p,i}$, $\nu_{f,i}$ are known for all i. These can also be written as $X|\theta \sim N(\theta, \Sigma_p)$ and $Y|\theta \sim N(\theta, \Sigma_f)$ where Σ_p and Σ_f are n dimensional diagonal matrices with ith entries $\nu_{p,i}$ and $\nu_{f,i}$

respectively. The objective is to compute an estimate or decision rule $\hat{q} = \{\hat{q}_i(X) : 1 \leq i \leq n\}$ based on the past data X such that $\hat{q}$ optimally predicts Y under the loss

$$\frac{1}{n}\sum_{i=1}^{n} l_i(\theta_i, q_i),$$

where for each $X = x$, the associated predictive loss is given by

$$l_i(\theta_i, q_i(x)) = \mathbb{E}_{Y_i}\left[b_i(Y_i - q_i(x))^+ + h_i(q_i(x) - Y_i)^+\right] \tag{6.1}$$

with $b_i, h_i > 0$ for all i, where the notation $()^+$ is defined as

$$(x)^+ := \max\{x, 0\}.$$

Remark 6.1. If $b_i + h_i = 1$, the loss function (6.1) corresponds to the regular check function.

6.1.1 The Newsvendor Problem

One motivation for this piecewise linear loss function is the newsvendor problem in the inventory management. The problem considers a vendor who sells a large amount of products. Based on the observed demand X in the previous period, the vendor needs to determine the stocking quantity $\hat{q}_i$ of each product in the next period. There is an obvious tradeoff between ordering too much and inventory is left over at the end of the period versus ordering too little and sales are lost. Suppose each unit of inventory incurs a holding cost $h_i > 0$ and each unit of lost sale incurs a cost of $b_i > 0$, the vendor's loss function is given by (6.1). Usually the lost sales cost is much higher than the inventory cost, which leads to a highly asymmetric loss function. This problem

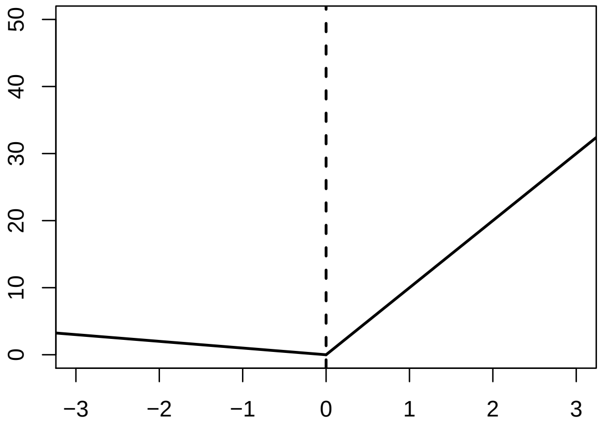

Figure 5: Piecewise linear loss with $b = 10$ and $h = 1$

of determining optimal stock level can be traced back to Edgeworth (1888) who applied the central limit theorem to determine the optimal cash reserves to satisfy random withdrawals from depositors and the modern formulation relates to the work of Arrow, Harris, and Marschak (1951).

To have some understanding of the asymmetric piecewise linear loss, we write the residual as $r = y - q$ and consider the following function of r

$$f(r) = br^{+} + h(-r)^{+}.$$

Suppose the lost sales cost of this product is $b = \$10$ per unit and its holding cost is $h = \$1$ per unit. Let us consider two scenarios: If the vendor orders too little and holds 3 units of potentially unsold product in one period, the cost that he possibly incurs is \$30. If he orders too much and holds 3 units of unsold product, then he will lose \$3.

It is worth noting that the original newsvendor problem does not belong to any typical prediction type of problems because of lacking independent variables, therefore, regression and supervised learning techniques can not be applied here. Traditionally, the problem is solved using Bayes approach: the mean θ is treated as a random variable following a known prior distribution. For computational easiness, conjugate priors are usually chosen.

6.2 Nonparametric Empirical Bayes Methodology

In the standard Bayesian approach, a good prior is not always easy to obtain: For example, historical data is not available or expensive to obtain. To avoid this problem, Mukherjee et al. (2015) recently propose a parametric empirical Bayes method by studying normal priors and estimating hyperparameters using shrinkage methods. However, if there is no extra information, employing normal priors may not always be appropriate. Here we propose a nonparametric empirical Bayes approach which does not assume any explicit form of prior distribution. Moreover, our proposed method does not involve tuning parameters and the main part can be reformulated as a convex problem due to the work of Koenker and Mizera (2014). It reduces the computational effort by several order of magnitude by comparison to the method of Mukherjee et al. (2015).

Our nonparametric empirical Bayes method can be thought as a two-step process: We need to first derive the Bayes decision rule under the check loss function for a general prior and then compute an estimate of the prior. Once the estimated prior is obtained, we plug it into the Bayes decision rule and obtain the predictions.

6.2.1 Univariate Bayes Estimator and Consistency Results

In this subsection, Bayes decision rule under the check loss for a general prior and some consistency results will be studied.

Again we require the decision rule $d\left(x\right)$ to be separable: $d\left(x\right)=\left(d_1\left(x_1\right),...,d_n\left(x_n\right)\right)$. Then it is sufficient to consider the univariate prediction problem: Suppose the past sales $X|\theta\sim N(\theta,\nu_p)$, the future demand $Y|\theta\sim N(\theta,\nu_f)$ and $\theta\sim F\left(\cdot\right)$.

In the following, we assume all the regularity conditions in Kiefer and Wolfowitz (1956) hold. As a consequence, we have Kiefer-Wolfowitz MLE F_n converge weakly to the true prior distribution F: $F_n\Rightarrow F$. Let Q and Q_n be the posterior distributions depending on F and F_n respectively. Then

$$dQ\left(\theta\right)=\nu_p^{-1/2}\varphi\left(\nu_p^{-1/2}\left(x-\theta\right)\right)/f_X\left(x\right)dF\left(\theta\right)$$

and

$$dQ_n\left(\theta\right)=\nu_p^{-1/2}\varphi\left(\nu_p^{-1/2}\left(x-\theta\right)\right)/f_X\left(x\right)dF_n\left(\theta\right).$$

Since $\varphi\left(\nu_p^{-1/2}\left(x-\theta\right)\right)$ is continuous and bounded in θ, we have $Q_n\Rightarrow Q$.

Denote the marginals as

$$g\left(z\right)=\nu_f^{-1/2}\int_{\mathbb{R}}\varphi\left(\nu_f^{-1/2}\left(z-\theta\right)\right)dQ\left(\theta\right)$$

and

$$g_n\left(z\right)=\nu_f^{-1/2}\int_{\mathbb{R}}\varphi\left(\nu_f^{-1/2}\left(z-\theta\right)\right)dQ_n\left(\theta\right).$$

Denote the marginal distributions as $G(z)$ and $G_n(z)$.

Lemma 6.2. *$G_n \Rightarrow G$.*

Proof. Since $Q_n \Rightarrow Q$, we have the estimated densities g_n pointwise converge to g with respect to Lebesgue measure. By Scheffé's theorem (Billingsley, 1986), the result follows. □

With the notation above, the expected posterior loss can be written as

$$\int_{\mathbb{R}} b(z-q)^{+} + h(q-z)^{+} \, dG(z).$$

Theorem 6.3. *Consider the univariate prediction problem that the past $X|\theta \sim N(\theta, \nu_p)$, the future $Y|\theta \sim N(\theta, \nu_f)$ and $\theta \sim F(\cdot)$. Under the check loss*

$$\mathbb{E}_{Y|\theta}\left[b(Y-q)^{+} + h(q-Y)^{+}\right],$$

the Bayes decision rule for x is the $b/(b+h)$th quantile of Z

$$q = \inf\{z : G(z) \geq b/(b+h)\}.$$

Proof. The proof is analogue to the standard Bayes estimation under the absolute loss. Therefore, it is removed. □

Remark 6.4. From Theorem 6.3, the ratio b/h is essential for the prediction.

Remark 6.5. Note that the estimated density $g_n(z)$ is continuous. Then its distribution function G_n is strictly increasing and the inverse G_n^{-1} exists. As a consequence, Kiefer-Wolfowitz MLE based estimator has the form

$$\hat{q}_n = G_n^{-1}(b/(b+h)).$$

The following results are the direct consequences of Chapter 3.

Theorem 6.6. *(Consistency) For the newsvendor problem, Kiefer-Wolfowitz MLE based estimator $\hat{q}_n$ converge to the true Bayes estimator $\hat{q}$ almost surely.*

Proof. By Lemma 6.2, we have $G_n \Rightarrow G$. For the newsvendor problem, the loss function is convex and it has monotone bounded derivative. By Theorem 3.11, the result follows. □

Theorem 6.7. *(Asymptotic optimality) For the newsvendor problem, assume (1) $\sup_x q(x) < \infty$ and (2) $\mathbb{E}_F|\theta| < \infty$, then the modified Bayes rule $\tilde{q}_n$ as defined in Chapter 3 is asymptotically optimal.*

Proof. A consequence of Theorem 3.12. □

6.2.2 Implementation

Computing the Kiefer-Wolfowitz MLE can be reformulated as a convex problem and therefore be efficiently solved by modern interior-point methods (Koenker and Mizera, 2014). To obtain the predictions, we need to compute the $b/(b+h)$th quantile for G_n for every x. Note that G_n has a finite mixture Gaussian density and its quantiles can be efficiently found by solving the equation

$$G_n(z) - b/(b+h) = 0 \tag{6.2}$$

using any one-dimensional root finding algorithm. In implementation, we set the searching interval as $\left[\min x - 3\max\sqrt{\nu_f}, \max x + 3\max\sqrt{\nu_f}\right]$. Moreover, as we discussed in Remark 6.5, the solution to (6.2) is unique, so the result returned by the algorithm is the $b/(b+h)$th quantile for G_n.

6.3 Simulations

We repeat the experiments in Mukherjee et al. (2015) and set $\nu_{f,i} = 1$ and $b_i + h_i = 1$ for all $i = 1, ..., n$. To evaluate the performance, for each replication we calculate the empirical risk

$$\frac{1}{n} \sum_{i=1}^{n} \left[b_i \left(y_i - q_i \left(x_i \right) \right)^+ + h_i \left(q_i \left(x_i \right) - y_i \right)^+ \right]. \tag{6.3}$$

Experiment 1. (Two-point support) We study a homoskedastic model with $\nu_{p,i} = 1/3$ for all $i = 1, ..., n$. We consider two different values for θ_i: $1/\sqrt{3}$ and $-3\sqrt{3}$. The loss is defined as: when $\theta_i = 1/\sqrt{3}$, $b_i = 0.51$ and when $\theta_i = -3\sqrt{3}$, $b_i = 0.99$. We consider two different choices of n: when $n = 20$, there are 18 replicates of the (θ_i, b_i) pair of $\left(1/\sqrt{3}, 0, 51\right)$ and 2 replicates of $\left(-3\sqrt{3}, 0.99\right)$. For $n = 100$, there are 90 replicates of the (θ_i, b_i) pair of $\left(1/\sqrt{3}, 0, 51\right)$ and 10 replicates of the latter. The results are based on 50 replications.

Experiment 2. (Standard normal) We consider a homoskedastic model with $\nu_{p,i} = \nu_p$ for all $i = 1, ..., n$. We vary p to numerically test the performance of the methods. We let θ_i i.i.d. from $N(0,1)$ and b_i i.i.d. from uniform $[0.51, 0.99]$. The results are based on 20 replications.

Experiment 3. (Heteroskedastic Models) Several heteroskedastic models are also studied here and the results are based on 20 replications.

Case 1: θ are i.i.d. from Uniform $(0,1)$ and $\nu_{p,i}$ are i.i.d. from Uniform $(0.1, 1/3)$.

Case 2: θ are i.i.d. from N $(0,1)$ and $\nu_{p,i}$ are i.i.d. from Uniform $(0.1, 1/3)$.

Case 3: We consider the dependence between $\nu_{p,i}$ and θ: $\nu_{p,i}$ are i.i.d. from Uniform $(0.1, 1/3)$ and $\theta_i = 5\nu_{p,i}$.

Case 4: The uniform distribution in Case 3 is replaces by Inv-$\chi^2\left(0.1, 1/3\right)$ and $\theta_i = 5\nu_{p,i}$.

Case 5: $\nu_{p,i}$ are i.i.d. from the 2-point distribution $2^{-1}\left(\delta_{0.1} + \delta_{0.5}\right)$ and the θ_i are drawn conditioned on the past variances:

$$(\theta_i|\nu_{p,i} = 0.1) \sim N\left(0, 0.1\right) \qquad \text{and} \qquad (\theta_i|\nu_{p,i} = 0.5) \sim N\left(0, 0.5\right).$$

There are two groups in the data.

Case 6: We assess the sensitivity in the performance of the estimators to the Gaussian noise assumption. $\nu_{p,i}$ are i.i.d. from $\text{Uniform}\left(0.1, 1/3\right)$ and $\theta_i = 5\nu_{p,i}$. The past observations are generated independently from

$$X_i \sim \text{Uniform}\left(\theta_i - \sqrt{3\nu_{p,i}}, \theta_i + \sqrt{3\nu_{p,i}}\right) \quad \text{for } i = 1, ..., n.$$

The tables below report the empirical risk , its standard deviation (in round brackets) and CPU time (in squared brackets) of nonparametric empirical Bayes estimator denoted NPEB, compared to the shrinkage methods proposed by Mukherjee et al. (2015) denoted Zero and GrandMean. As a reference, we also provide the performance of the Bayes decision rule when the prior distribution is completely known denoted as Oracle.

We observe that (1) In all the three experiments, NPEB performs better than the shrinkage methods and several orders of magnitude faster; (2) It is a surprise that in Experiment 2, when the prior is standard normal which is a setup favoring the shrinkage methods, NPEB still has lower empirical risk. This is possibly because NPEB methods has a faster convergence rate respect to the sample size; (3) Zero and GrandMean methods tend to have very similar performance, because both methods try to shrink the estimators to the centers of priors, which are zeros by design.

	n=20	n = 100
Oracle	0.356(0.02)[0.06]	0.360(0.02)[0.03]
NPEB	**0.360**(0.06)[0.15]	**0.362**(0.02)[0.60]
Zero	0.395(0.06)[1144]	0.409(0.03)[5779]
GrandMean	0.396(0.06)[1144]	0.409(0.03)[5779]

Table 4: Empirical risk of NPEB estimator compared to the shrinkage methods in Experiment 1

n	ν_p/ν_f	Oracle	NPEB	Zero	GrandMean
20	1/1	0.36(0.06)[0.11]	**0.38**(0.08)[0.27]	0.46(0.10)[1192]	0.46(0.10)[1192]
	1/2	0.33(0.08)[0.11]	**0.34**(0.08)[0.24]	0.37(0.10)[984]	0.37(0.10)[984]
	1/3	0.32(0.07)[0.10]	**0.32**(0.09)[0.23]	0.34(0.08)[856]	0.34(0.08)[856]
	1/4	0.30(0.07)[0.09]	**0.30**(0.08)[0.17]	0.33(0.06)[541]	0.33(0.06)[541]
	1/5	0.30(0.08)[0.08]	**0.31**(0.08)[0.17]	0.34(0.10)[533]	0.34(0.10)[533]
	1/6	0.32(0.06)[0.08]	**0.32**(0.07)[0.17]	0.35(0.06)[490]	0.35(0.06)[490]
100	1/1	0.35(0.02)[0.36]	**0.36**(0.02)[0.78]	0.43(0.04)[6062]	0.43(0.04)[6062]
	1/2	0.33(0.02)[0.35]	**0.34**(0.03)[0.74]	0.38(0.03)[5433]	0.38(0.03)[5433]
	1/3	0.32(0.02)[0.33]	**0.32**(0.02)[0.65]	0.37(0.02)[4302]	0.37(0.02)[4302]
	1/4	0.30(0.03)[0.26]	**0.30**(0.03)[0.55]	0.34(0.03)[3038]	0.34(0.03)[3038]
	1/5	0.29(0.03)[0.27]	**0.30**(0.03)[0.53]	0.32(0.03)[2762]	0.32(0.03)[2762]
	1/6	0.30(0.03)[0.27]	**0.30**(0.03)[0.53]	0.33(0.03)[2615]	0.33(0.03)[2615]

Table 5: Empirical risk of NPEB estimator compared to the shrinkage methods in Experiment 2

6.4 Forecasts for Product Demand

In this section, we conduct numerical experiments based on demand data from a manufacturing company (Zhao, 2017). We apply our proposed non-

n	Case	Oracle	NPEB	Zero	GrandMean
20	1	0.31(0.07)[0.11]	**0.31**(0.07)[0.72]	0.35(0.09)[945]	0.35(0.09)[945]
	2	0.30(0.07)[0.11]	**0.30**(0.07)[0.72]	0.33(0.07)[803]	0.33(0.07)[803]
	3	0.28(0.05)[0.11]	**0.28**(0.05)[0.68]	0.32(0.08)[1035]	0.32(0.08)[1035]
	4	0.31(0.07)[0.08]	**0.31**(0.07)[0.52]	0.35(0.07)[617]	0.35(0.07)[617]
	5	0.28(0.04)[0.07]	**0.29**(0.05)[0.51]	0.35(0.06)[539]	0.35(0.06)[539]
	6	0.30(0.04)[0.08]	**0.31**(0.04)[0.50]	0.35(0.05)[680]	0.35(0.05)[680]
100	1	0.29(0.02)[0.32]	**0.29**(0.02)[2.05]	0.34(0.03)[5063]	0.34(0.03)[5063]
	2	0.32(0.03)[0.34]	**0.32**(0.03)[2.26]	0.36(0.03)[4285]	0.36(0.03)[4285]
	3	0.29(0.03)[0.31]	**0.29**(0.03)[2.05]	0.34(0.03)[5252]	0.34(0.03)[5252]
	4	0.29(0.03)[0.25]	**0.29**(0.03)[1.49]	0.33(0.04)[2974]	0.33(0.04)[2974]
	5	0.30(0.02)[0.26]	**0.31**(0.02)[1.52]	0.36(0.03)[3015]	0.36(0.03)[3015]
	6	0.30(0.02)[0.24]	**0.30**(0.02)[1.48]	0.35(0.03)[3592]	0.35(0.03)[3592]

Table 6: Empirical risk of NPEB estimator compared to the shrinkage methods in Experiment 3

parametric empirical Bayes method to predict the monthly demand for each product.

In 2016, the company provides 1578 products within 26 product categories. To simplify our analysis, we will look at the category having the highest transaction times and make predictions for the products having at least 10^4 demand in a given month. Assume the variances $\nu_{p,i}$ and $\nu_{f,i}$ for each product are given and they can be estimated using the monthly data from the year 2012 - 2015. The lost sales cost b_i and the inventory cost h_i are not available in this data set. Typically, the lost sales cost is much higher than the inventory cost. Recall the fact 6.4 that only the ratio b_i/h_i matters in the predictions and we assume $b_i = 1$ and $h_i = 0.1$ for all i.

For the evaluation, we take the naïve estimator $q_{naive,i}(x_i) = x_i$ as a benchmark and compute the ratio of the sum of prediction errors under the check loss

$$RSPE\left(q\right)=\frac{SPE\left(q\right)}{SPE\left(q_{naive}\right)},$$

where

$$SPE\left(q\right)=\sum_{i=1}^{n}\left[b_{i}\left(y_{i}-q_{i}\left(x_{i}\right)\right)^{+}+h_{i}\left(q_{i}\left(x_{i}\right)-y_{i}\right)^{+}\right].$$

Given the demands in the month m of 2016, we make prediction for each of products for the month $(m+1)$, where $m=1,...,11$. The table reports the average of the ratio of the sum of the prediction errors RSPE and its standard deviation (in brackets) of nonparametric empirical Bayes estimator denoted NPEB, compared to the shrinkage method governed by grand mean centric priors proposed by Mukherjee et al. (2015) denoted MBR, the James-Stein estimator denoted JS and a naïve grand mean estimator $q_{mean,i}\left(x_{i}\right)=\bar{x}$ denoted MEAN.

We find NPEB outperforms the other methods in this data. This is possibly because the method does not put shape constraint on the prior distribution compared to MBR and JS so that it is more flexible.

NPEB	MBR	JS	MEAN
0.63(0.16)	1.47(0.42)	0.96(0.03)	3.88(0.76)

Table 7: The ratio of the sum of prediction errors RSPE in product demand forecasts for NPEB, compared to MBR, JS and MEAN

Chapter 7

Conclusions and Future Work

This thesis studies the mixture models, in particular the estimation of mixing distributions and their applications to empirical Bayes prediction. In this thesis, we establish the asymptotic optimality for the empirical Bayes estimators; the results apply not only for the squared loss, but for a large class of convex loss functions. A consistency result of Bayes estimators for mixture models for a large class of convex loss functions is provided under mild conditions. In the case study of the newsvendor problem involving quantile loss, our experiments support that our proposed nonparametric empirical Bayes estimator outperforms the shrinkage methods proposed by Mukherjee et al. (2015). Additionally, the proposed nonparametric empirical Bayes estimator is several orders of magnitude faster. In particular, when the prior is standard normal, a setup favoring the shrinkage methods, our method still has lower empirical risk.

The second part of the thesis is devoted to the estimation of mixing distribution in mixture models. We propose four estimation methods/algorithms for computing or approximating the Kiefer-Wolfowitz MLE which are capable of working in higher dimensions parameter space. Projected Stochastic Gradient is capable of working in higher dimensions but its convergence may be slow. Stochastic Average Approximation is generally much faster but in some versions, its estimation target differs from that of Kiefer-Wolfowitz nonparametric maximum likelihood estimator. This is even more true for Constraint Resampling, which is in fact an autonomous and novel estimation method; its properties, as well as those of other proposed methods are assessed via simulations and theoretical results. Cutting-Plane Method, an algorithm can work with problems a very large number of constraints, is considered at last. The experiments show that Constraint Resampling performs in general well and able to solve the problem with the dimension of the parameter space at lest ten.

The penultimate chapter is devoted to facilitate the multivariate data-analytical applications of the developed algorithms. Nonparametric empirical Bayes methods are studied in the presence of explanatory variables. A nonparametric empirical Bayes regression model is later proposed. In contrast to some of the previous approaches, such a regression model has a very simple form and inherits most of theoretical properties of nonparametric empirical Bayes procedures. Unlike methods based on the partial linear model, the parameter estimation procedure is equivalent to solving a convex optimization problem in function space and can be efficiently solved by the proposed algorithms.

Regarding the future work of nonparametric empirical Bayes methods, there are several aspects to be mentioned. One drawback of nonparametric empirical Bayes methods in applied data analysis is that the decision rule is defined only at the ith training data point. In many applications, such as supervised learning problems, the ultimate goal is to generalized the decision rule to new data not represented in the training set. A possible solution is to induce a regression tree whose predictions are as close as possible to the decision rules.

One can also think backward and apply nonparametric empirical Bayes methods to regression tree based models. Let us start with a regression tree model. Given a split node, we find the prediction by minimizing the sum of squares over each region and the prediction is simply the average of the data falling into this region. For the response variables in each region, we can think them are the summations of signals from a certain unknown distribution and Gaussian noises (with a known variance, which can be estimated via cross-validation). Then nonparametric empirical Bayes methods can be applied. Now for each split, instead of the simple averages, we solve a sequence of convex optimization problems which are independent from each other. This can be done efficiently via parallel programming. This approach can then be extended to more sophisticated tree models, such as random forest and boosting.

The other direction is to apply nonparametric empirical Bayes methods on ensemble methods, such as stacking. Given M fitted candidate models, stacking method chooses to find the optimal weights by minimizing the cross-validation error. If we want to apply nonparametric empirical Bayes methods, we can think the candidate models are randomly generated. The estimated weights can be found by solving Kiefer-Wolfowitz maximum likelihood estimation problem restricting on the class of distribution functions with no more than M support points.

Bibliography

Abadie, A. and M. Kasy (2017). The risk of machine learning. *arXiv preprint arXiv:1703.10935*.

Arrow, K. J., T. Harris, and J. Marschak (1951). Optimal inventory policy. *Econometrica: Journal of the Econometric Society*, 250–272.

Avis, D., D. Bremner, and R. Seidel (1997). How good are convex hull algorithms? *Computational Geometry 7*(5-6), 265–301.

Billingsley, P. (1986). *Probability and Measure.* John Wiley & Sons.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer.

Boyd, S. and A. Mutapcic (2006). Subgradient methods. *Lecture notes of EE364b, Stanford University, Winter Quarter 2007.*

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization.* Cambridge University Press.

Boyd, S. and L. Vandenberghe (2007). Localization and cutting-plane methods. *From Stanford EE 364b lecture notes*.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, 113–152.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 1685–1704.

Chen, J. (2017). Consistency of the MLE under mixture models. *Statistical Science 32*(1), 47–63.

Cohen, N., E. Greenshtein, and Y. Ritov (2013). Empirical Bayes in the presence of explanatory variables. *Statistica Sinica*, 333–357.

DasGupta, A. (2011). *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*. Springer Science & Business Media.

Deely, J. and W. Zimmer (1976). Asymptotic optimality of the empirical Bayes procedure. *The Annals of Statistics*, 576–580.

Dicker, L. H. and S. D. Zhao (2014). Nonparametric empirical Bayes and maximum likelihood estimation for high-dimensional data analysis. *arXiv preprint arXiv:1407.2635*.

du Merle, O., J.-L. Goffin, and J.-P. Vial (1998). On improvements to the analytic center cutting plane method. *Computational Optimization and Applications 11*(1), 37–52.

Edgeworth, F. Y. (1888). The mathematical theory of banking. *Journal of the Royal Statistical Society 51*(1), 113–127.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Efron, B. and C. Morris (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association 70*(350), 311–319.

Ermoliev, Y. (1983). Stochastic quasigradient methods and their application to system optimization. *Stochastics: An International Journal of Probability and Stochastic Processes 9*(1-2), 1–36.

Fay III, R. E. and R. A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association 74*(366a), 269–277.

Greenshtein, E. and T. Itskov (2014). Deconvolution, convex optimization, non-parametric empirical Bayes and treatment of non-response. *arXiv preprint arXiv:1406.5840*.

Hettich, R. and K. O. Kortanek (1993). Semi-infinite programming: theory, methods, and applications. *SIAM Review 35*(3), 380–429.

Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pp. 1248–1251. Springer.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 361–379.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics 40*(2), 633–643.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics 37*(4), 1647–1684.

Jiang, W. and C.-H. Zhang (2010). Empirical Bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pp. 263–273. Institute of Mathematical Statistics.

Kall, P., S. W. Wallace, and P. Kall (1994). *Stochastic Programming.* Springer.

Kannan, R., L. Lovász, and M. Simonovits (1997). Random walks and an o*(n5) volume algorithm for convex bodies. *Random Structures and Algorithms 11*(1), 1–50.

Karlin, S. and H. Rubin (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics*, 272–299.

Kelley, Jr, J. E. (1960). The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics 8*(4), 703–712.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Koenker, R. (2015). Adaptive estimation of regression parameters for the gaussian scale mixture model. In *Empirical Economic and Financial Research*, pp. 373–378. Springer.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association 109*(506), 674–685.

Lehmann, E. and G. Casella (1998). *Theory of Point Estimation, Springer-Verlag.*

Linderoth, J., A. Shapiro, and S. Wright (2006). The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research 142*(1), 215–241.

Lindley, D. V. and A. F. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–41.

Lindsay, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution. In *Statistical Distributions in Scientific Work*, pp. 95–109. Springer.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics 11*(1), 86–94.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pp. i–163. JSTOR.

López, M. and G. Still (2007). Semi-infinite programming. *European Journal of Operational Research 180*(2), 491–518.

Mak, W.-K., D. P. Morton, and R. K. Wood (1999). Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters 24*(1), 47–56.

Martin, R. (2015). Asymptotically optimal nonparametric empirical Bayes via predictive recursion. *Communications in Statistics Theory and Methods 44*(2), 286–299.

Mukherjee, G., L. D. Brown, and P. Rusmevichientong (2015). Efficient empirical Bayes prediction under check loss using asymptotic risk estimates. *arXiv preprint arXiv:1511.00028*.

Norkin, V. I., G. C. Pflug, and A. Ruszczyński (1998). A branch and bound method for stochastic global optimization. *Mathematical Programming 83*(1-3), 425–450.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A 185*, 71–110.

Pfanzagl, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *Journal of Statistical Planning and Inference 19*(2), 137–158.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*(2), 186–199.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1*, 157–163.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics 35*(1), 1–20.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Rockafellar, R. T. (1993). Lagrange multipliers and optimality. *SIAM Review 35*(2), 183–238.

Rutherford, J. and R. Krutchkoff (1969). Some empirical Bayes techniques in point estimation. *Biometrika 56*(1), 133–137.

Shapiro, A. (2009). Semi-infinite programming, duality, discretization and optimality conditions. *Optimization 58*(2), 133–161.

Shapiro, A. (2010). Computational complexity of stochastic programming: Monte Carlo sampling approach. In *Proceedings of the International Congress of Mathematicians*, pp. 2979–2995.

Shapiro, A., D. Dentcheva, and A. Ruszczyński (2009). *Lectures on stochastic programming: modeling and theory*. SIAM.

Tadić, V. B., S. P. Meyn, and R. Tempo (2006). Randomized algorithms for semi-infinite programming problems. In *Probabilistic and Randomized Methods for Design under Uncertainty*, pp. 243–261. Springer.

Vandenberghe, L. and S. Boyd (1998). Connections between semi-infinite and semidefinite programming. In *Semi-infinite Programming*, pp. 277–294. Springer.

Verbeek, A. (1973). Bounds of the posterior mean of a location parameter. *Unpublished*.

Wang, W. and S. Ahmed (2008). Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters 36*(5), 515–519.

Zhao, F. (2017). Forecasts for product demand. `https://www.kaggle.com/felixzhao/productdemandforecasting`. Accessed: 2017-10-14.