University of Alberta

ROBUST ACTIVE LEARNING

by

Rui Nie

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistical Machine Learning

Department of Mathematical and Statistical Sciences

Fall 2015
Edmonton, Alberta

Abstract

This dissertation first introduces the concepts of robust active learning (also called optimal experimental design in statistics), and the possible advantages of it over the traditional passive learning method. Then a general regression problem with possibly misspecified models is presented, and divided into three specific problems due to different choices of loss functions and optimizing methods.

After that, the three problems are all solved with a minimax approach but in different ways to get the optimal design densities for the active learning method.

Finally, simulations are used to compare active learning with passive learning results on specific examples, and the experiment results prove that active learning is more robust and advantageous than passive learning in these examples.

Acknowledgements

I wish to express my sincere gratitude to my supervisor, Dr. Wiens, for his continuous and patient guidance, support, help and encouragement throughout the research and writing of my thesis. The knowledge and research methodologies I've gained during the process will be a lifelong treasure.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation for Active Learning

Consider this problem: researchers want to establish a program to automatically detect cancer of an organ from a patient's X-ray picture of that organ. However, although X-ray pictures of the organ are plentiful and easy to get, whether a patient has the cancer and how bad the cancer is have to be verified by experts or even determined over time. Therefore, when researchers collect data to optimize the program, they can find inputs (X-ray pictures) of the program easily and cheaply, but obtaining the output of the program (diagnosis of the cancer) is comparatively hard, expensive and time-consuming.

This kind of problem is very common in practice. To save cost and effort, researchers will choose a sample from the whole dataset, and only look for the outputs for this sample. The criterion for choosing this sample is to get the most effective sample, the one that will generate the largest prediction accuracy with the smallest sample size.

1.2 Active Learning and Experimental Design

In the machine learning field, active learning is defined in contrast to passive learning (Cohn, Ghahramani, and Jordan, 1996). In passive learning, the researcher is treated as a passive recipient of the data to be used for optimizing the program; usually the computer will randomly select a sample from the whole dataset to be the input data; but usually this random sample is not the most effective one. However, in active learning, the researchers can decide which sample to use by themselves; they either rely on some experts' knowledge or construct some models to find the most effective sample. The advantage of active learning over passive learning has been established by many authors in various tasks, such as text classification (Tong and Koller, 2002), information extraction (Scheffer, Decomain, and Wrobel, 2001; Olsson, 2009) and spoken language understanding (Tur, Hakkani-Tür, and Schapire, 2005).

The aim of active learning is to find the best locations of the unlabeled sample points, so that the parameters estimated by the sample can be the most accurate. Many strategies for determining the sample locations are proposed, a review of the general frameworks of those strategies can be seen in Settles (2009). In classification field, for instance, uncertainty sampling method is widely used, where researchers choose sample points about which they are least certain how to label, because those points are the most informative ones. In regression field, however, the "variance reduction" method (Settles, 2009, p. 21) is widely used. It is called optimal experimental design in statistics field (Kiefer, 1959; Fedorov, 1972; Pukelsheim, 1993).

Traditional experimental design methods aim to find sample points that minimize the variability of the estimated parameters or the predicted output, so that these sample points generate the most accurate estimates. There is no unique way to measure the variability of the estimates, so a number of criteria have been proposed. Among them we will talk about the I-optimality and G-optimality in this dissertation. The I-optimality aims to minimize the integration of the variance of the predicted output, and the G-optimality aims to minimize the maximum variance of the predicted output.

However, the traditional experimental design methods only make sense when the model can be assumed to be true, or when the bias caused by the difference between the true model and the misspecified model is small enough to be ignored. Otherwise, we also have to consider the bias part. Therefore, in this dissertation, we define the loss function to be the Integrated Mean Squared Error (IMSE) and the Maximum Mean Squared Error (Max MSE) of the predicted output; they correspond to the traditional I-optimality and G-optimality problem, but use the mean squared error, which contains both variance and bias, to replace variance in the traditional theory.

1.3 Our Contribution

Many literatures have considered the misspecification of the model and improve on the traditional optimal design theory (Kanamori and Shimodaira, 2003; Sugiyama, 2006). The research direction of this thesis is most close to that of Sugiyama (2006), so we make comparisons with that paper. Later in Chapter 2, we can see that the model error term ψ (1.1) causes the occurrence of $\boldsymbol{b}_{n;\psi,w}$, thus causing the bias (2.1). Then when asymptotics of MSE of $\hat{\boldsymbol{\theta}}$ are taken, $\boldsymbol{S}_{\psi,w,p}$ (2.4) also occurs, which is a major component in the squared bias part of the MSE. Sugiyama claimed the error term ψ to be "inaccessible"

(Sugiyama, 2006, p. 151), thus, when he solved for the best design density, he ignored the bias part $S_{\psi,p}$ in (2.6) by bounding $\tau_n = o(1)$ in (1.3). We will improve on that by enlarging the bound and keeping both the bias and variance, and solve the "inaccessible" problem by a minimax method - minimizing the maximum (over ψ) MSE. That will be the first time to explicitly take the model misspecification into account in the process of solving for the best design density, and that is the major contribution of this dissertation.

1.4 Robust Active Learning

The aim of robust active learning is to get an active learning design that is least sensitive to small errors of the model. In practice, the analytic form of the true model is usually unknown, the estimated model more or less has some deviation from the true model, thus a robust design will be useful in such situations. In this thesis we will analyze models that have small deviations from the true model. In order to get a robust design, we will use a minimax approach to optimize the design, that is to first maximize the loss function over the model error, then find a probability design that minimizes the maximized loss.

Wen, Yu, and Greiner (2014) also consider the minimax estimation problems in active learning with misspecified models, but with the loss function maximized over the "reweighting" functions in weighted likelihood estimates which are then optimized, and minimized by the unknown parameters of the model. They assume that the test distribution is uncertain, and learn when the "reweighting" is needed and provide a "reweighting" algorithm. In this thesis, we assume that the test distribution is known, and the "reweighting" function (2.5) is chosen in Chapter 2, the focus is to learn the best training distribution. Our work can be preceded by the studies of Wen et al. (2014).

In our research, the costs of obtaining every training example is considered to be the same, and the sample size is fixed. For research of active learning problem with different costs and (unknown) discriminative power for different learning examples and under budgetary constraints, see Kapoor and Greiner (2005).

1.5 Problem Formulation

This paper looks into the robust active learning problem in regression scenario. Suppose a design space χ; in the previous cancer diagnosis example, χ can be seen as the set that contains all pixel vectors of the X-ray pictures. Suppose $\boldsymbol{x} = (x_1, x_2, ..., x_s)$ is one element in χ; it could be the pixel vector of one of the X-ray pictures in the cancer diagnosis example. Also suppose $\boldsymbol{x}$ has a density of $q(\boldsymbol{x})$ in χ. Let y denote the corresponding output of $\boldsymbol{x}$; it could be the numeric diagnosis result defined by researchers in the cancer diagnosis example. Suppose the regression model is

$$y = E(y|\boldsymbol{x}) + \epsilon,$$

$$E(y|\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta} + \psi(\boldsymbol{x}). \tag{1.1}$$

In the model, ϵ is one of the i.i.d. random errors with mean zero and unknown variance σ_ϵ^2; $\boldsymbol{f}(\boldsymbol{x})$ is an r-dimensional column vector of regressors, each element of which is a function of several functionally independent variables $\boldsymbol{x}$; $\boldsymbol{\theta}$ is an r-dimensional column vector of unknown parameters; and $\psi(\boldsymbol{x})$ is the error term that indicates the difference between the true model $E(y|\boldsymbol{x})$ and the estimated model $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}$. The prime in the function indicates transposition. We assume that the analytic form of $E(y|\boldsymbol{x})$ is unknown, and the researcher estimates $E(y|\boldsymbol{x})$ with $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}$, thus leading to the existence of the error term $\psi(\boldsymbol{x})$.

To ensure the uniqueness of $\boldsymbol{\theta}$ and $\psi(\boldsymbol{x})$, define

$$\boldsymbol{\theta} = arg \min_{\eta} \int_{\chi} (E[y|\boldsymbol{x}] - \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\eta})^2 q(\boldsymbol{x}) d\boldsymbol{x},$$

$$\psi(\boldsymbol{x}) = E[y|\boldsymbol{x}] - \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}.$$

The above two equations lead to the orthogonality requirement

$$\int_{\chi} \boldsymbol{f}(\boldsymbol{x})\psi(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = \boldsymbol{0}. \tag{1.2}$$

We assume that the magnitude of the error term $\psi(\boldsymbol{x})$ is bounded. Otherwise, if the difference between true model and the estimated model is unlimited, the estimated model would be meaningless. Therefore, assume that

$$\int_{\chi} \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} \leq \tau_n^2, \tag{1.3}$$

for a given constant τ_n. This τ_n may or may not depend on n.

The process of optimizing the regression model is:

(1) Take a sample of size n from χ with a design density $p(\boldsymbol{x})$. Suppose the sample is $\{\boldsymbol{x}_i\}_{i=1,2,\ldots,n}$.

(2) Find out the corresponding output $\{y_i\}_{i=1,2,\ldots,n}$ for the sample input $\{\boldsymbol{x}_i\}_{i=1,2,\ldots,n}$.

(3) Optimize the parameter vector $\boldsymbol{\theta}$ in the regression model using the sample $\{\boldsymbol{x}_i, y_i\}_{i=1,2,\ldots,n}$ thus the estimated model that has the greatest prediction accuracy.

In passive learning, the sample is a chosen uniformly at random, so the density $p(\boldsymbol{x}) = q(\boldsymbol{x})$; while in active learning, $p(\boldsymbol{x})$ is designed by the researcher.

The aim of active learning in this problem setting is, with a fixed sample size, to find the best sample distribution $p(\boldsymbol{x})$ that yields the regression model with the greatest prediction accuracy.

In chapters 2, 3 and 4, we discuss three ways to find the best density $p(\boldsymbol{x})$, all with a minimax approach. In Chapter 2, we use a Weighted Least Squares (WLS) method to estimate parameter $\boldsymbol{\theta}$. We define the loss function to be the Integrated Mean Squared Error (IMSE) of the fitted value of $\hat{y}$; first maximize the loss function over the error term $\psi(\boldsymbol{x})$ and then minimize the maximized loss function over density $p(\boldsymbol{x})$, thus to find the best density $p(\boldsymbol{x})$. In Chapter 3, we also use the WLS method to estimate $\boldsymbol{\theta}$, but we define another loss function – Maximum Mean Squared Error (Max MSE) of $\hat{y}$, and use a minimax approach to find the best $p(\boldsymbol{x})$. In Chapter 4, we use the traditional Ordinary Least Squares (OLS) method to estimate $\boldsymbol{\theta}$, and again use IMSE as the loss function, then find $p(\boldsymbol{x})$ by a minimax method.

Chapter 2

Active Learning with WLS Estimation and Loss Function IMSE

2.1 MSE of the WLS Estimate

2.1.1 MSE of the Sample WLS Estimate

In this chapter the unknown parameter vector $\boldsymbol{\theta}$ is estimated by the Weighted Least Squares (WLS) estimate $\hat{\boldsymbol{\theta}}_{WLS}$. This is motivated by Sugiyama (2006), where $\hat{\boldsymbol{\theta}}_{WLS}$ with weight of the sample points – $w_0(\boldsymbol{x}) = q(\boldsymbol{x})/p(\boldsymbol{x})$, is claimed to be an asymptotically unbiased estimate. Later in Section 2.2 we will prove the validity of this claim. Now we first generate the $\hat{\boldsymbol{\theta}}_{WLS}$ and its Mean Squared Error (MSE) with a general weight w.

Suppose that $\boldsymbol{x} = (x_1, x_2, ..., x_r)$ is a design point that is randomly sampled from χ with a design density $p(\boldsymbol{x})$, and there are n design points. Define $\boldsymbol{W}$ to be a $n \times n$ diagonal matrix whose diagonal elements are the positive weights at each of the design points: $w(\boldsymbol{x_1}), w(\boldsymbol{x_2}), ..., w(\boldsymbol{x_n})$. Define

$$\boldsymbol{X} = [\boldsymbol{f}(\boldsymbol{x_1}), \boldsymbol{f}(\boldsymbol{x_2}), ...\boldsymbol{f}(\boldsymbol{x_n})]',$$

$$\boldsymbol{Y} = [y_1, y_2, ..., y_n]',$$

$$\boldsymbol{M}_{n;w} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x_i})w(\boldsymbol{x_i})\boldsymbol{f}'(\boldsymbol{x_i}),$$

$$\boldsymbol{D}_{n;w} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x_i})w^2(\boldsymbol{x_i})\boldsymbol{f}'(\boldsymbol{x_i}),$$

$$\boldsymbol{b}_{n;\psi,w} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x_i})w(\boldsymbol{x_i})\psi(\boldsymbol{x_i}).$$

Obviously $\boldsymbol{M_{n;w}} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$, and is semi-positive definite, since

$$\boldsymbol{c}'\boldsymbol{M}_{n;w}\boldsymbol{c} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{c'}\boldsymbol{f}(\boldsymbol{x_i}))^2 w(\boldsymbol{x_i}) \geq 0,$$

for any r-dimensional nonzero column vector $\boldsymbol{c}$. Assume that $\boldsymbol{M_{n;w}}$ is invertible, which is equivalent to the assumption that $\boldsymbol{X}$ has full rank. Then the weighted least squares estimate is

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{WLS} &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y} \\ &= \boldsymbol{M}_{n;w}^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{f}(\boldsymbol{x_i})w(\boldsymbol{x_i})y_i \\ &= \theta + \boldsymbol{M}_{n;w}^{-1}\boldsymbol{b}_{n;\psi,w} + \boldsymbol{M}_{n;w}^{-1}\cdot\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x_i})w(\boldsymbol{x_i})\epsilon_i.\end{aligned}$$

Because the random errors $\{\epsilon_i\}_{i=1,2,...,n}$ are i.i.d. variables with mean zero and variance σ_ϵ^2 , the expectation of $\hat{\boldsymbol{\theta}}_{WLS}$ over random error is

$$E_\epsilon[\hat{\boldsymbol{\theta}}_{WLS}] = \boldsymbol{\theta} + \boldsymbol{M}_{n;w}^{-1}\boldsymbol{b}_{n;\psi,w}; \tag{2.1}$$

the covariance matrix of $\hat{\boldsymbol{\theta}}_{WLS}$ is

$$\begin{aligned} COV_\epsilon[\hat{\boldsymbol{\theta}}_{WLS}] &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\sigma_\epsilon^2 \\ &= \frac{\sigma_\epsilon^2}{n}\boldsymbol{M}_{n;w}^{-1}\boldsymbol{D}_{n;w}\boldsymbol{M}_{n;w}^{-1}. \end{aligned}$$

Then the MSE of $\hat{\boldsymbol{\theta}}_{WLS}$ conditional on the sample is

$$\begin{aligned} MSE_\epsilon[\hat{\boldsymbol{\theta}}_{WLS}] &= COV_\epsilon[\hat{\boldsymbol{\theta}}_{WLS}] + (E_\epsilon[\hat{\boldsymbol{\theta}}_{WLS}] - \boldsymbol{\theta})(E_\epsilon[\hat{\boldsymbol{\theta}}_{WLS}] - \boldsymbol{\theta})' \\ &= \boldsymbol{M}_{n;w}^{-1}\left\{\frac{\sigma_\epsilon^2}{n}\boldsymbol{D}_{n;w} + \boldsymbol{b}_{n;\psi,w}\boldsymbol{b}'_{n;\psi,w}\right\}\boldsymbol{M}_{n;w}^{-1}. \end{aligned} \tag{2.2}$$

2.1.2 Asymptotics of the MSE

First we introduce the asymptotics of the matrices that appear in the function of the MSE. Define

$$\begin{aligned} \boldsymbol{M}_{w,p} &= \int_\chi \boldsymbol{f}(\boldsymbol{x})w(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}, \\ \boldsymbol{D}_{w,p} &= \int_\chi \boldsymbol{f}(\boldsymbol{x})w^2(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}, \\ \boldsymbol{b}_{\psi,w,p} &= \int_\chi \boldsymbol{f}(\boldsymbol{x})w(\boldsymbol{x})\psi(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}, \end{aligned}$$

$$\begin{aligned} \boldsymbol{S}_{\psi,w,p} &= COV_{\boldsymbol{x}}\big[\boldsymbol{f}(\boldsymbol{x})w(\boldsymbol{x})\psi(\boldsymbol{x})\big] \\ &= \int_\chi (\boldsymbol{f}(\boldsymbol{x})w(\boldsymbol{x})\psi(\boldsymbol{x}) - \boldsymbol{b}_{\psi,w,p})(\boldsymbol{f}(\boldsymbol{x})w(\boldsymbol{x})\psi(\boldsymbol{x}) - \boldsymbol{b}_{\psi,w,p})'p(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_\chi \boldsymbol{f}(\boldsymbol{x})w^2(\boldsymbol{x})\psi^2(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} - \boldsymbol{b}_{\psi,w,p}\boldsymbol{b}'_{\psi,w,p}. \end{aligned}$$

Lemma 2.1: $\boldsymbol{M}_{n;w} = \boldsymbol{M}_{w,p} + O_p\left(\frac{1}{\sqrt{n}}\right)$; $\boldsymbol{D}_{n;w} = \boldsymbol{D}_{w,p} + O_p\left(\frac{1}{\sqrt{n}}\right)$; $\boldsymbol{b}_{n;\psi,w}\boldsymbol{b'}_{n;\psi,w} = \boldsymbol{b}_{\psi,w,p}\boldsymbol{b'}_{\psi,w,p} + \frac{1}{n}\boldsymbol{S}_{\psi,w,p} + o_p\left(\frac{1}{n}\right)$.

Proof: *The j,kth element in $\boldsymbol{M}_{n;w}$ is*

$$\left\{\boldsymbol{M}_{n;w}\right\}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} f_j(\boldsymbol{x_i})w(\boldsymbol{x_i})f_k(\boldsymbol{x_i})$$

The expectation of $\left\{\boldsymbol{M}_{n;w}\right\}_{j,k}$ is

$$\begin{aligned} E_{\boldsymbol{x}}\Big[\left\{\boldsymbol{M}_{n;w}\right\}_{j,k}\Big] &= \int_{\mathcal{X}} f_j(\boldsymbol{x})w(\boldsymbol{x})f_k(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\ &= \left\{\boldsymbol{M}_{w,p}\right\}_{j,k} \end{aligned}$$

The variance of $\left\{\boldsymbol{M}_{n;w}\right\}_j, k$ is

$$VAR_{\boldsymbol{x}}\Big[\left\{\boldsymbol{M}_{n;w}\right\}_{j,k}\Big] = \frac{1}{n}Var_{\boldsymbol{x}}\big[f_j(\boldsymbol{x})w(\boldsymbol{x})f_k(\boldsymbol{x})\big]$$

By Theorem 14.4-1 in (Bishop, Fienberg, and Holland, 2007), we have

$$\left\{\boldsymbol{M}_{n;w}\right\}_{j,k} - E_{\boldsymbol{x}}\Big[\left\{\boldsymbol{M}_{n;w}\right\}_{j,k}\Big] = O_p\left(\sqrt{Var_{\boldsymbol{x}}\Big[\left\{\boldsymbol{M}_{n;w}\right\}_{j,k}\Big]}\right),$$

i.e.

$$\left\{\boldsymbol{M}_{n;w}\right\}_{j,k} - \left\{\boldsymbol{M}_{w,p}\right\}_{j,k} = O_p\left(\frac{1}{\sqrt{n}}\right),$$

meaning the set of values $\sqrt{n}\Big\{\left\{\boldsymbol{M}_{n;w}\right\}_{j,k} - \left\{\boldsymbol{M}_{w,p}\right\}_{j,k}\Big\}$ are stochastically bounded. If the above equation applies to every entry of a matrix, then it applies to the whole matrix. Therefore,

$$\boldsymbol{M}_{n;w} - \boldsymbol{M}_{w,p} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Similarly, we can also prove

$$\boldsymbol{D}_{n;w} - \boldsymbol{D}_{w,p} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

It is obvious that

$$E_{\boldsymbol{x}}\left[\boldsymbol{b}_{n;\psi,w}\right] = \boldsymbol{b}_{\psi,w,p},$$
$$COV_{\boldsymbol{x}}\left[\boldsymbol{b}_{n;\psi,w}\right] = \frac{1}{n}\boldsymbol{S}_{\psi,w,p}.$$

By the Weak Law of Large Numbers, we have the following convergence in probability:

$$\boldsymbol{b}_{n;\psi,w} \xrightarrow{P} \boldsymbol{b}_{\psi,w,p}, \tag{2.3}$$

$$n\Big((\boldsymbol{b}_{n;\psi,w} - \boldsymbol{b}_{\psi,w,p})(\boldsymbol{b}_{n;\psi,w} - \boldsymbol{b}_{\psi,w,p})' - \frac{1}{n}\boldsymbol{S}_{\psi,w,p}\Big) \xrightarrow{P} \mathbf{0}.$$

This is equivalent to

$$\boldsymbol{b}_{n;\psi,w}\boldsymbol{b'}_{n;\psi,w} = \boldsymbol{b}_{\psi,w,p}\boldsymbol{b'}_{\psi,w,p} + \frac{1}{n}\boldsymbol{S}_{\psi,w,p} + o_p\left(\frac{1}{n}\right).$$

Thus Lemma 2.1 is proved.

With Lemma 2.1 , we conclude that

$$\begin{aligned} MSE_\epsilon[\hat{\boldsymbol{\theta}}_{WLS}] &= \frac{\sigma_\epsilon^2}{n}\left[\boldsymbol{M}_{w,p} + O_p\left(\frac{1}{\sqrt{n}}\right)\right]^{-1}\left[\boldsymbol{D}_{w,p} + O_p\left(\frac{1}{\sqrt{n}}\right)\right]\left[\boldsymbol{M}_{w,p} + O_p\left(\frac{1}{\sqrt{n}}\right)\right]^{-1} \\ &+ \left[\boldsymbol{M}_{w,p} + O_p\left(\frac{1}{\sqrt{n}}\right)\right]^{-1}\left[\boldsymbol{b}_{\psi,w,p}\boldsymbol{b'}_{\psi,w,p} + \frac{1}{n}\boldsymbol{S}_{\psi,w,p} + o_p\left(\frac{1}{n}\right)\right]\left[\boldsymbol{M}_{w,p} + O_p\left(\frac{1}{\sqrt{n}}\right)\right]^{-1} \\ &= \frac{\sigma_\epsilon^2}{n}\boldsymbol{M}_{w,p}^{-1}\boldsymbol{D}_{w,p}\boldsymbol{M}_{w,p}^{-1} + \boldsymbol{M}_{w,p}^{-1}\left(\boldsymbol{b}_{\psi,w,p}\boldsymbol{b'}_{\psi,w,p} + \frac{1}{n}\boldsymbol{S}_{\psi,w,p}\right)\boldsymbol{M}_{w,p}^{-1} + O_p\left(n^{-3/2}\right). \end{aligned} \tag{2.4}$$

2.2 Weight w_0 and Loss Function IMSE

Assume that $q(\boldsymbol{x}) \neq 0$, $p(\boldsymbol{x}) \neq 0$. Now we choose weight to be

$$w_0(\boldsymbol{x}) = \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}, \tag{2.5}$$

because with $w_0(\boldsymbol{x})$,

$$\begin{aligned}\boldsymbol{b}_{\psi,w_0,p} &= \int_\chi \boldsymbol{f}(\boldsymbol{x}) w_0(\boldsymbol{x}) \psi(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_\chi \boldsymbol{f}(\boldsymbol{x}) \psi(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} \\ &= \boldsymbol{0}\end{aligned}$$

by (1.2); so that by (2.3) $\boldsymbol{b}_{n;\psi,w_0}$ is asymptotically zero. Then $E_\epsilon[\hat{\boldsymbol{\theta}}_{WLS}] = \boldsymbol{\theta}$ asymptotically by (2.1), thus this weight w_0 will make θ_{WLS} asymptotically unbiased.

Then we have

$$\begin{aligned}\boldsymbol{M}_{w_0,p} = \textstyle\int_\chi \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{f}'(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} \overset{def}{=} \boldsymbol{U}, \\ \boldsymbol{D}_{w_0,p} = \textstyle\int_\chi \boldsymbol{f}(\boldsymbol{x}) \frac{q^2(\boldsymbol{x})}{p(\boldsymbol{x})} \boldsymbol{f}'(\boldsymbol{x}) d\boldsymbol{x} \overset{def}{=} \boldsymbol{T}_p, \\ \boldsymbol{S}_{\psi,w_0,p} = \textstyle\int_\chi \boldsymbol{f}(x) \frac{q^2(\boldsymbol{x})}{p(\boldsymbol{x})} \psi^2(\boldsymbol{x}) \boldsymbol{f}'(\boldsymbol{x}) d\boldsymbol{x} \overset{def}{=} \boldsymbol{S}_{\psi,p}.\end{aligned}$$

We assume that if $\boldsymbol{c}'\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ (a.e. $x \in \chi$), then $\boldsymbol{c} = \boldsymbol{0}$. With this assumption, $\boldsymbol{U}$ is positive definite, since for any nonzero column vector $\boldsymbol{c}$,

$\boldsymbol{c}'\boldsymbol{U}\boldsymbol{c} = \int_\chi (\boldsymbol{c}'\boldsymbol{f}(\boldsymbol{x}))^2 q(\boldsymbol{x})d\boldsymbol{x} > 0$. Then by (2.4)

$$\begin{aligned} MSE_\epsilon\Big[\hat{\boldsymbol{\theta}}_{WLS}\Big] &= \frac{\sigma_\epsilon^2}{n}\boldsymbol{U}^{-1}\boldsymbol{T}_p\boldsymbol{U}^{-1} + \frac{1}{n}\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}\boldsymbol{U}^{-1} + O(n^{-3/2}) \\ &= \frac{1}{n}\boldsymbol{U}^{-1}\{\sigma_\epsilon^2\boldsymbol{T}_p + \boldsymbol{S}_{\psi,p}\}\boldsymbol{U}^{-1} + O(n^{-3/2}), \end{aligned} \tag{2.6}$$

The MSE of the estimated $\hat{y}$ is

$$\begin{aligned} MSE_\epsilon(\hat{y}) &= E_\epsilon\bigg[\Big\{\boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}}_{WLS} - E[y|\boldsymbol{x}]\Big\}^2\bigg] \\ &= E_\epsilon\bigg[\Big\{\boldsymbol{f}'(\boldsymbol{x})(\hat{\boldsymbol{\theta}}_{WLS} - \boldsymbol{\theta}) - \psi(\boldsymbol{x})\Big\}^2\bigg] \\ &= \boldsymbol{f}'(\boldsymbol{x})MSE_\epsilon\Big[\hat{\boldsymbol{\theta}}_{WLS}\Big]\boldsymbol{f}(\boldsymbol{x}) + \psi^2(\boldsymbol{x}). \end{aligned} \tag{2.7}$$

The value of $MSE_{\epsilon,\boldsymbol{z}}(\hat{y})$ is dependent on the design points in the random sample. We want to establish a stable loss function that is not affected by the randomness of the sample, so we set the loss function to be the expectation of $MSE_\epsilon(\hat{y})$ integrated over points in design space χ, and call it Integrated Mean Squared Error (IMSE).

$$\begin{aligned} IMSE \quad &= \int_\chi MSE_\epsilon(\hat{y})q(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_\chi \boldsymbol{f}'(\boldsymbol{x})MSE_\epsilon\Big[\hat{\boldsymbol{\theta}}_{WLS}\Big]\boldsymbol{f}(\boldsymbol{x})q(\boldsymbol{x})dx + \int_\chi \psi^2(x)q(\boldsymbol{x})dx \qquad (2.8) \\ &= tr\Big[\boldsymbol{U}\Big\{MSE_\epsilon\Big[\hat{\boldsymbol{\theta}}_{WLS}\Big]\Big\}\Big] + \int_\chi \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{n}tr\big[\boldsymbol{U}^{-1}\{\sigma_\epsilon^2\boldsymbol{T}_p + \boldsymbol{S}_{\psi,p}\}\big] + \int_\chi \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} + O(n^{-3/2})(2.9) \end{aligned}$$

Since $\int_\chi \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}$ does not depend on the design, we concentrate on optimizing the leading term in (2.9). In Section 3.2 of (Sugiyama, 2006), $\tau_n =$

$\tau = o(1)$ is assumed, then $\boldsymbol{T}_p = O(1)$ and $\boldsymbol{S}_{\psi,p} = o(1)$, so that $tr\big[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}\big]$ is also ignored, and the problem becomes minimizing $tr\big[\boldsymbol{U}^{-1}\boldsymbol{T}_p\big]$ only, which is called variance-only approach. To extend and improve on the results by (Sugiyama, 2006), we assume $\tau_n = \tau = O(1)$ now, then the term $tr[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}]$ remains.

To find the best $p(\boldsymbol{x})$, we adopt a minimax approach – first find the maximum of the loss function over $\psi(\boldsymbol{x})$, then find the $p(\boldsymbol{x})$ that can minimize that maximized loss. To be specific, the problem becomes:

$$\min_p \max_\psi tr\big[\boldsymbol{U}^{-1}\{\sigma_\epsilon^2\boldsymbol{T}_p + \boldsymbol{S}_{\psi,p}\}\big], \tag{2.10}$$

with the maximization done subject to (1.2) and (1.3), and minimization subject to the requirement that $p(\boldsymbol{x})$ be a probability density.

2.3 Maximization over ψ

First we deal with the maximization part in (2.10).

$$\begin{aligned}&\max_\psi tr\big[\boldsymbol{U}^{-1}\{\sigma_\epsilon^2\boldsymbol{T}_p + \boldsymbol{S}_{\psi,p}\}\big]\\ &= \max_\psi \big\{\sigma_\epsilon^2 tr[\boldsymbol{U}^{-1}\boldsymbol{T}_p] + tr[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}]\big\}.\end{aligned}$$

Only the second part $tr[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}]$ contains $\psi(\boldsymbol{x})$, so only this part needs to be maximized. Now the maximization problem becomes

$$\begin{aligned}\max_\psi tr[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}] &= \max_\psi tr\left[\boldsymbol{U}^{-1}\int_\chi \boldsymbol{f}(\boldsymbol{x})\frac{q^2(\boldsymbol{x})}{p(\boldsymbol{x})}\psi^2(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})d\boldsymbol{x}\right]\\ &= \max_\psi \int_\chi \boldsymbol{f}'(x)\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}.\end{aligned} \tag{2.11}$$

subject to (1.2) and (1.3).

Assume that

$$\int_\chi \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = t^2, \tag{2.12}$$

then $t^2 \leq \tau^2$ by (1.3). Define

$$h(\boldsymbol{x}) = \psi^2(\boldsymbol{x})q(\boldsymbol{x})/t^2. \tag{2.13}$$

By (2.12), $h(\boldsymbol{x})$ is a density function. Then

$$\begin{aligned} tr[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}] &= t^2 \int_\chi a_p(\boldsymbol{x})h(\boldsymbol{x})d\boldsymbol{x} \\ &\leq t^2 \max_{\boldsymbol{x}} a_p(\boldsymbol{x}) \\ &\leq \tau^2 \max_{\boldsymbol{x}} a_p(\boldsymbol{x}). \end{aligned}$$

From above it can be seen that the maximum of $tr[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}]$ is reached on condition that $t = \tau$, so we let $t^2 = \tau^2$, and thus we have

$$\int_\chi \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = \tau^2. \tag{2.14}$$

We use (2.14) to replace the constraint (1.3) in this maximization section.

Lemma 2.2: *The equation* $\int_\chi a_p(\boldsymbol{x})h(\boldsymbol{x})d\boldsymbol{x} = \max_{\boldsymbol{x}} a_p(\boldsymbol{x})$*, constraint* (1.2) *and* (2.14) *can all be satisfied at the same time when* $h(\boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{x}^*)$*, where* $\delta(\boldsymbol{x})$ *is a delta function, and* $\boldsymbol{x}^* = arg \max_{\boldsymbol{x}} a_p(\boldsymbol{x})$*.*

Proof: *For simplification, we only prove this in the one dimensional case, and when* χ *is a continuous domain. To be specific, suppose that* $\chi = [c, d]$ *(c,* *d are constants, they can be either finite or infinite),*

$$h(x) = \delta(x - x^*) = \lim_{n->\infty} \delta_n(x - x^*), \tag{2.15}$$

where, according to (Arfken and Weber, 2011), the delta sequence can be defined to be

$$\delta_n(x - x^*) = \begin{cases} 0, & c < x < x^* - \frac{1}{2n} \\ n, & x^* - \frac{1}{2n} < x < x^* + \frac{1}{2n} \\ 0, & x^* + \frac{1}{2n} < x < d \end{cases} .$$

Then

$$\begin{aligned} \int_\chi a_p(x)h(x)dx &= \int_c^d a_p(x)\delta(x - x^*)dx \\ &= \lim_{n->\infty} \int_c^d a_p(x)\delta_n(x - x^*)dx \\ &= \max_x a_p(x). \end{aligned}$$

The constraint (2.14) *is satisfied by* (2.15) *and the fact that the delta function is a density function. Constraint* (1.2) *is also satisfied, which is proved below.*

$$\begin{aligned} \int_\chi f(x)\psi(x)q(x)dx &= \int_\chi f(x)\sqrt{h(x)q(x)}dx \\ &= \int_\chi f(x)\sqrt{\delta(x - x^*)q(x)}dx \\ &= \lim_{n->\infty} \int_c^d f(x)\sqrt{\delta_n(x - x^*)q(x)}dx \\ &= \lim_{n->\infty} \int_{x^*-\frac{1}{2n}}^{x^*+\frac{1}{2n}} \sqrt{n}f(x)\sqrt{q(x)}dx \\ &= 0. \end{aligned}$$

Therefore,

$$\int_{\chi} a_p(x)h(x)dx = \max_{x} a_p(x),$$

when $h(x) = \delta(x - x^*)$. Thus Lemma 2.2 is proved.

With Lemma 2.2, we know

$$\max_{\psi} tr\big[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}\big] = \tau^2 \max_{\boldsymbol{x}} a_p(\boldsymbol{x}).$$

Then we get

$$\begin{aligned}
\max_{\psi} tr\big[\boldsymbol{U}^{-1}\{\sigma_\epsilon^2\boldsymbol{T}_p + \boldsymbol{S}_{\psi,p}\}\big] &= \sigma_\epsilon^2 tr\big[\boldsymbol{U}^{-1}\boldsymbol{T}_p\big] + \max_{\psi} tr\big[\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}\big] \\
&= \sigma_\epsilon^2 tr\big[\boldsymbol{U}^{-1}\boldsymbol{T}_p\big] + \tau^2 \max_{\boldsymbol{x}} a_p(\boldsymbol{x}) \\
&= \sigma_\epsilon^2 tr\Big\{\boldsymbol{U}^{-1}\int_{\chi} \boldsymbol{f}(\boldsymbol{x})\frac{q^2(\boldsymbol{x})}{p(\boldsymbol{x})}\boldsymbol{f}'(\boldsymbol{x})d\boldsymbol{x}\Big\} + \tau^2 \max_{\boldsymbol{x}} a_p(\boldsymbol{x}) \\
&= \sigma_\epsilon^2 \int_{\chi} a_p(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} + \tau^2 \max_{\boldsymbol{x}} a_p(\boldsymbol{x}) \\
&= \sigma_\epsilon^2 E_q\big[a_p(\boldsymbol{x})\big] + \tau^2 \max_{\boldsymbol{x}} a_p(\boldsymbol{x}),
\end{aligned}$$

where $a_p(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$. Therefore, we have the following theorem.

Theorem 2.1:

$$\frac{\max_{\psi} tr[\boldsymbol{U}^{-1}\{\sigma_\epsilon^2\boldsymbol{T}_p + \boldsymbol{S}_{\psi,p}\}]}{\sigma_\epsilon^2 + \tau^2} = (1-\nu)E_q\big[a_p(\boldsymbol{x})\big] + \nu \max_{\boldsymbol{x}} a_p(\boldsymbol{x}), \qquad (2.16)$$

where $\nu = \tau^2/(\sigma_\epsilon^2 + \tau^2) \in [0, 1]$ may be chosen by the experimenter, representing the relative concern for errors due to bias rather than to variance.

Now (2.16) needs to be minimized by density $p(\boldsymbol{x})$.

2.4 Minimization over p

In this section we will first work out a general solution solution to the minimization problem, then use a straight line model as an example and work out the exact solution of that example.

2.4.1 General Solution

For the purpose of illustration, before obtaining the general solution for $0 \leq \nu \leq 1$, we first do the minimization for special cases when $\nu = 0$ and 1.

Minimization when $\nu = 0$

When $\nu = 0$, the minimization problem becomes

$$\min_{p} E_q[a_p(\boldsymbol{x})] \quad \textit{subject to} \quad \int_{\chi} p(\boldsymbol{x}) d\boldsymbol{x} = 1. \tag{2.17}$$

Lemma 2.3: *The minimizer of* (2.17) *is*

$$p(\boldsymbol{x}) = \frac{\sqrt{b(\boldsymbol{x})q(\boldsymbol{x})}}{\int_{\chi} \sqrt{b(\boldsymbol{x})q(\boldsymbol{x})} d\boldsymbol{x}}, \tag{2.18}$$

where

$$b(\boldsymbol{x}) = a_p(\boldsymbol{x})p(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})q(\boldsymbol{x}).$$

The minimum value is $(\int_{\chi} \sqrt{b(\boldsymbol{x})q(\boldsymbol{x})})^2$.

Proof: *By Cauchy-Schwarz Inequality, we have*

$$\int_{\chi} \frac{b(\boldsymbol{x})q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x} \cdot \int_{\chi} p(\boldsymbol{x}) d\boldsymbol{x} \geq (\int_{\chi} \sqrt{b(\boldsymbol{x})q(\boldsymbol{x})} d\boldsymbol{x})^2. \tag{2.19}$$

Since $\int_\chi \frac{b(\boldsymbol{x})q(\boldsymbol{x})}{p(\boldsymbol{x})}dx = E_q[a_p(\boldsymbol{x})]$ *and* $\int_\chi p(\boldsymbol{x})d\boldsymbol{x} = 1$, (2.19) *becomes*

$$E_q[a_p(\boldsymbol{x})] \geq (\int_\chi \sqrt{b(\boldsymbol{x})q(\boldsymbol{x})}d\boldsymbol{x})^2,$$

equality is reached if and only if $\frac{b(\boldsymbol{x})q(\boldsymbol{x})}{p(\boldsymbol{x})} = \lambda p(\boldsymbol{x})$ *(*λ *is a constant); then since* $p(\boldsymbol{x})$ *is a density, we get* (2.18)*, and Lemma 2.3 is proved.*

Notice that since $\nu = 0$, only the first part in (2.16) remains, so the problem becomes the same as the variance-only problem in Section 3.2 in (Sugiyama, 2006), and the results are the same, too.

Minimization when $\nu = 1$

When $\nu = 1$, the minimization problem becomes

$$\min_p\{\max_x a_p(\boldsymbol{x})\} \;\; \textit{subject to} \;\; \int_\chi p(\boldsymbol{x})d\boldsymbol{x} = 1. \tag{2.20}$$

Lemma 2.4: *The minimizer of* (2.20) *is*

$$p(\boldsymbol{x}) = \frac{b(\boldsymbol{x})}{\int_\chi b(\boldsymbol{x})d\boldsymbol{x}}. \tag{2.21}$$

The minimum value is r*, which is the dimension of the vector* $\boldsymbol{f}(\boldsymbol{x})$*.*

Proof:

$$\max_{\boldsymbol{x}} a_p(\boldsymbol{x}) \geq E_p[a_p(\boldsymbol{x})] = \int_\chi \frac{b(\boldsymbol{x})}{p(\boldsymbol{x})}p(\boldsymbol{x})dx = \int_\chi b(\boldsymbol{x})d\boldsymbol{x}.$$

When (2.21) *holds,*

$$a_p(\boldsymbol{x}) = \frac{b(\boldsymbol{x})}{p(\boldsymbol{x})} = \int_\chi b(\boldsymbol{x})d\boldsymbol{x}.$$

Thus (2.21) *is proved to be the minimizer. Then*

$$\begin{aligned}\min_{p}\{\max_{\boldsymbol{x}} a_p(\boldsymbol{x})\} &= \int_{\chi} b(\boldsymbol{x})d\boldsymbol{x} = \int_{\chi} \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} \\ &= tr\bigg\{\int_{\chi} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\cdot \boldsymbol{U}^{-1}\bigg\} \\ &= tr\big[\boldsymbol{U}\boldsymbol{U}^{-1}\big] = tr\boldsymbol{I}_r = r.\end{aligned} \tag{2.22}$$

Therefore, Lemma 2.4 is proved.

Minimization when $0 \le \nu \le 1$

In this part we will obtain a general solution when $0 \le \nu \le 1$. Note that the solutions in Section 2.4.1.1 and 2.4.1.2, when $\nu = 0$ and $\nu = 1$, are special cases of the solution in this section.

It is convenient to first minimize (2.16) over $a_p(\boldsymbol{x})$ then recover $p(\boldsymbol{x})$. To simplify notation, define $a(\boldsymbol{x}) = a_p(\boldsymbol{x})$ and then $p(\boldsymbol{x}) = b(\boldsymbol{x})/a(\boldsymbol{x})$.

We will minimize (2.16) in two steps: first fix $max_{\boldsymbol{x}} a(\boldsymbol{x})$ and minimize $E_q[a(\boldsymbol{x})]$, then minimize the whole thing over m.

Define

$$A_m = \bigg\{a(\cdot) \mid \max_{\boldsymbol{x}} a(\boldsymbol{x}) = m \; and \int_{\chi} \frac{b(\boldsymbol{x})}{a(\boldsymbol{x})}d\boldsymbol{x} = 1\bigg\}.$$

First we need to find the minimizer of $E_q[a(\boldsymbol{x})]$ in class A_m.

Ignore the constraint $\max_{\boldsymbol{x}} a(\boldsymbol{x}) = m$ for the moment, first minimize $E_q[a(\boldsymbol{x})]$ subject to $\int_{\chi} \frac{b(\boldsymbol{x})}{a(\boldsymbol{x})}d\boldsymbol{x} = 1$; the problem is equivalent to minimizing

$$\int_{\chi} \Big[a(\boldsymbol{x})q(\boldsymbol{x}) + \lambda\frac{b(\boldsymbol{x})}{a(\boldsymbol{x})}\Big]d\boldsymbol{x}, \tag{2.23}$$

where λ is a Lagrange multiplier.

To minimize (2.23), it is sufficient to minimize the integrand in (2.23). The minimizer of the integrand is found to be $\sqrt{\lambda}\sqrt{\frac{b(\boldsymbol{x})}{q(\boldsymbol{x})}}$ ($\lambda \geq 0$). If this minimizer is to satisfy the constraint $\max_{\boldsymbol{x}} a(\boldsymbol{x}) = m$, one choice is to truncate it whenever it gets over m. (Here m has to be no greater than the maximum value of the minimizer.)

Define

$$\begin{aligned} a_{\text{-}}(\boldsymbol{x}) &= c_m\sqrt{\frac{b(\boldsymbol{x})}{q(\boldsymbol{x})}} = c_m\sqrt{\boldsymbol{f}(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}, \\ a_m(\boldsymbol{x}) &= min(a_{\text{-}}(\boldsymbol{x}), m), \end{aligned} \tag{2.24}$$

where c_m is chosen so that

$$\int_{\chi} \frac{b(\boldsymbol{x})}{a_m(\boldsymbol{x})} d\boldsymbol{x} = 1; \tag{2.25}$$

and where $m \in [r, \max_{\boldsymbol{x}} a_{\text{-}}(\boldsymbol{x})]$ (by (2.22)). It is obvious that $a_m(\cdot) \in A_m$. In the following, Lemma 2.5 proves that such a c_m exists, so that such construction method is valid; then Theorem 2.2 shows the general solution to the minimization problem.

Lemma 2.5: *There exists at least one* $c_m \in (0, +\infty)$ *such that* (2.25) *is satisfied.*

Proof:

(1) Assume that

$$\begin{aligned} a_{\text{-}}(\boldsymbol{x}) &= c\sqrt{\frac{b(\boldsymbol{x})}{q(\boldsymbol{x})}} = c\sqrt{\boldsymbol{f}(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})} \quad (c > 0), \\ a_m(\boldsymbol{x}) &= min(a_{\text{-}}(\boldsymbol{x}), m). \end{aligned}$$

This c here is not relevant to m. We make the assumption in Section 2.2 that if $\boldsymbol{c}'\boldsymbol{f}(\boldsymbol{x}) = 0$ (a.e. $x \in \chi$), then $\boldsymbol{c} = \boldsymbol{0}$, and thus $\boldsymbol{U}$ is positive definite. Therefore, $\boldsymbol{f}(\boldsymbol{x}) \neq \boldsymbol{0}$, and $\boldsymbol{f}(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x}) > 0$. From the definition of $a_m(\boldsymbol{x})$ we know that

$$\frac{b(\boldsymbol{x})}{a_m(\boldsymbol{x})} = max\left(\frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{m}, \frac{\sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}}{c}\right)q(\boldsymbol{x}).$$

Define

$$S(c) = \left\{x \in \chi \mid \sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})} < m/c\right\},$$

then on $S(c)$ we have $\frac{b(\boldsymbol{x})}{a_m(\boldsymbol{x})} = \frac{\sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}}{c}q(\boldsymbol{x})$. Therefore,

$$\int_\chi \frac{b(\boldsymbol{x})}{a_m(\boldsymbol{x})}d\boldsymbol{x} \geq \int_{S(c)} \frac{b(\boldsymbol{x})}{a_m(\boldsymbol{x})}d\boldsymbol{x} = \frac{1}{c}\int_{S(c)} \sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}q(\boldsymbol{x})dx.$$

When $c \to 0+$, $S(c)$ will become larger and larger and approach χ, so $\int_{S(c)} \sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}q(\boldsymbol{x})dx$ will also increase and approach $E_{\boldsymbol{x}}\left[\sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}\right] > 0$. Also when $c \to 0+$, $1/c \to +\infty$, thus

$$\frac{1}{c}\int_{S(c)} \sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}q(\boldsymbol{x})dx \to +\infty,$$

and so

$$\int_\chi \frac{b(\boldsymbol{x})}{a_m(\boldsymbol{x})}d\boldsymbol{x} \to +\infty > 0.$$

(2) When $c \to +\infty$, $a_m(x) = m$,

$$\int_\chi \frac{b(x)}{a_m(x)}dx - 1 = \int_\chi \frac{b(x)}{m}dx - 1 \leq 0 \text{ (\textit{since} } m \geq \int_\chi b(x)dx \text{ \textit{by} (2.22))}.$$

So there exists at least one root of c in $(0,\ +\infty)$ that satisfies (2.25).

Theorem 2.2: *(1) The minimizer of $E_q[a(\boldsymbol{x})]$ in class A_m is a_m.*

(2) Put the minimizer $a_m(\boldsymbol{x})$ into (2.16), *it will become*

$$\alpha(m) \coloneqq (1-\nu)E_q[a_m(\boldsymbol{x})] + \nu \cdot m. \tag{2.26}$$

Minimize (2.26) *over $m \in [r, \max_x a_-(x)]$. If m_ν is the minimizer of* (2.26) *then a_{m_ν} will be the minimizer of* (2.16), *and $p_\nu = \frac{b(x)}{a_{m_\nu}(x)}$ will be the corresponding optimal design density.*

Proof: *(1) Let $a(\cdot)$ be any other member of A_m. We have*

$$\int_\chi \frac{b(\boldsymbol{x})}{a(\boldsymbol{x})} d\boldsymbol{x} = \int_\chi \frac{b(\boldsymbol{x})}{a_m(\boldsymbol{x})} d\boldsymbol{x} = 1$$

Then

$$
\begin{aligned}
&E_q[a(\boldsymbol{x})] - E_q[a_m(\boldsymbol{x})] \\
&= \int_\chi a(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} - \int_\chi a_m(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} + c_m^2\bigg(\int_\chi \frac{b(\boldsymbol{x})}{a(\boldsymbol{x})}d\boldsymbol{x} - \int_\chi \frac{b(\boldsymbol{x})}{a_m(\boldsymbol{x})}d\boldsymbol{x}\bigg) \\
&= \int_\chi a(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} - \int_\chi a_m(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} + \bigg(\int_\chi \frac{{a_-}^2(\boldsymbol{x})q(\boldsymbol{x})}{a(\boldsymbol{x})}d\boldsymbol{x} - \int_\chi \frac{{a_-}^2(\boldsymbol{x})q(\boldsymbol{x})}{a_m(\boldsymbol{x})}d\boldsymbol{x}\bigg) \quad \textit{by } (2.24) \\
&= \int_{a_-(\boldsymbol{x})\le m} \bigg\{a(\boldsymbol{x})q(\boldsymbol{x}) - a_-(\boldsymbol{x})q(\boldsymbol{x}) + \frac{{a_-}^2(\boldsymbol{x})q(\boldsymbol{x})}{a(\boldsymbol{x})} - \frac{{a_-}^2(\boldsymbol{x})q(\boldsymbol{x})}{a_-(\boldsymbol{x})}\bigg\}d\boldsymbol{x} \\
&\quad + \int_{a_-(\boldsymbol{x})>m} \bigg\{a(\boldsymbol{x})q(\boldsymbol{x}) - mq(\boldsymbol{x}) + \frac{{a_-}^2(\boldsymbol{x})q(\boldsymbol{x})}{a(\boldsymbol{x})} - \frac{{a_-}^2(\boldsymbol{x})q(\boldsymbol{x})}{m}\bigg\}d\boldsymbol{x} \\
&\ge \int_{a_-(\boldsymbol{x})\le m} \bigg\{a(\boldsymbol{x})q(\boldsymbol{x}) - a_-(\boldsymbol{x})q(\boldsymbol{x}) + \frac{{a_-}^2(\boldsymbol{x})q(\boldsymbol{x})}{a(\boldsymbol{x})} - \frac{{a_-}^2(\boldsymbol{x})q(\boldsymbol{x})}{a_-(\boldsymbol{x})}\bigg\}d\boldsymbol{x} \\
&\quad + \int_{a_-(\boldsymbol{x})>m} \bigg\{a(\boldsymbol{x})q(\boldsymbol{x}) - mq(\boldsymbol{x}) + \frac{m^2q(\boldsymbol{x})}{a(\boldsymbol{x})} - \frac{m^2q(\boldsymbol{x})}{m}\bigg\}d\boldsymbol{x} \\
&= \int_{a_-(\boldsymbol{x})\le m} \frac{q(\boldsymbol{x})}{a(\boldsymbol{x})}[a_-(\boldsymbol{x}) - a(\boldsymbol{x})]^2 d\boldsymbol{x} + \int_{a_-(\boldsymbol{x})>m} \frac{q(\boldsymbol{x})}{a(\boldsymbol{x})}[m - a(\boldsymbol{x})]^2 d\boldsymbol{x} \\
&= \int_\chi \frac{q(\boldsymbol{x})}{a(\boldsymbol{x})}[a_m(\boldsymbol{x}) - a(\boldsymbol{x})]^2 d\boldsymbol{x} \\
&\ge 0,
\end{aligned}
$$

with equality iff $a(\boldsymbol{x}) \equiv a_m(\boldsymbol{x})$*. Thus* a_m *is proved to be the minimizer of* $E_q[a(x)]$ *in class* A_m*.*

(2) If the minimizer of (2.26) *is* m_ν*, and the corresponding minimizer in class* A_{m_ν} *is* a_{m_ν}*, assume that* $a_0(\boldsymbol{x})$ *is any other one of* $a(\boldsymbol{x})$ *that is different from* $a_{m_\nu}(\boldsymbol{x})$*. Assume that* $\max_{\boldsymbol{x}} a_0(\boldsymbol{x}) = m_0$*, and that the minimizer in class*

A_{m_0} is a_{m_0}. Then we have

$$\begin{aligned}
&(1-\nu)E_q[a_{m_\nu}(\boldsymbol{x})] + \nu \max_{\boldsymbol{x}} a_{m_\nu}(\boldsymbol{x}) \\
&= \alpha(m_\nu) \\
&\leq \alpha(m_0) && \textit{since } m_\nu \textit{ is the minimizer of } \alpha(m) \\
&= (1-\nu)E_q[a_{m_0}(\boldsymbol{x})] + \nu \cdot m_0 \\
&\leq (1-\nu)E_q[a_0(\boldsymbol{x})] + \nu \cdot m_0 && \textit{since } a_{m_0} \textit{ is the minimizer in class } A_{m_0} \\
&= (1-\nu)E_q[a_0(\boldsymbol{x})] + \nu \max_{\boldsymbol{x}} a_0(\boldsymbol{x}).
\end{aligned}$$

Therefore, a_{m_ν} is proved to be the minimizer of (2.16), then $p_\nu = \frac{b(x)}{a_{m_\nu}(x)}$ is the corresponding optimal design density.

2.4.2 Straight Line Example

In this section we specify a straight line example and minimize $\alpha(m)$ numerically. Assume that the dimension of $\boldsymbol{x}$ is 1, and the design space for x is $\chi = (-\infty, \infty)$. Also, assume that $q(x)$ is a normal density with mean 0 and standard deviation σ_q, and that $\boldsymbol{f}(x) = [1, x]'$, thus $r = 2$. Then

$$\boldsymbol{U} = \int_\chi \boldsymbol{f}(x)\boldsymbol{f}'(x)q(x)dx = \begin{bmatrix} 1 & 0 \\ 0 & \sigma_q^2 \end{bmatrix},$$

$$b(x) = \boldsymbol{f}'(x)\boldsymbol{U}^{-1}\boldsymbol{f}(x)q(x) = (1 + \frac{x^2}{\sigma_q^2})q(x),$$

$$a_\_(x) = c_m\sqrt{\boldsymbol{f}'(x)U^{-1}\boldsymbol{f}(x)} = c_m\sqrt{1 + \frac{x^2}{\sigma_q^2}},$$

$$a_m(x) = \min(a_\_(x), m),$$

$$a_-(x) < m \Rightarrow -\sigma_q\sqrt{\frac{m^2}{c_m^2} - 1} < x < \sigma_q\sqrt{\frac{m^2}{c_m^2} - 1}.$$

Notice that under this setting, the passive learning method will make the maximum of the loss function (2.16) to be infinity, since when $p(x) = q(x)$, $a_p(x) = 1 + x^2$, and so $\max_x a_p(x) = +\infty$. But with our active learning method, the maximum of the loss function is bounded. This in a way shows that our active learning method is more robust and advantageous than passive learning.

Since $\max_x a_-(x) \in [2, +\infty)$, we restrict $m \geq 2$. Regarding the value of c_m,

1. When $c_m \leq 0$, it is impossible, because $a_m(x) > 0$.
2. When $0 < c_m \leq m$, c_m is defined by

$$\int_{a_-(x)<m} \frac{b(x)}{a_-(x)}dx + \int_{a_-(x)\geq m} \frac{b(x)}{m}dx = 1,$$

i.e.

$$\int_{-\sigma_q\sqrt{\frac{m^2}{c_m^2}-1}}^{\sigma_q\sqrt{\frac{m^2}{c_m^2}-1}} \frac{b(x)}{a_-(x)}dx + \int_{\sigma_q\sqrt{\frac{m^2}{c_m^2}-1}}^{\infty} \frac{b(x)}{m}dx + \int_{-\infty}^{-\sigma_q\sqrt{\frac{m^2}{c_m^2}-1}} \frac{b(x)}{m}dx = 1.$$

Because the integrands are even functions, the above is equivalent to

$$2\int_0^{\sigma_q\sqrt{\frac{m^2}{c_m^2}-1}} \frac{b(x)}{a_-(x)}dx + 2\int_{\sigma_q\sqrt{\frac{m^2}{c_m^2}-1}}^{\infty} \frac{b(x)}{m}dx = 1,$$

i.e.

$$\frac{2}{c_m}\int_0^{\sigma_q\sqrt{\frac{m^2}{c_m^2}-1}} \sqrt{1 + \frac{x^2}{\sigma_q^2}}q(x)dx + \frac{2}{m}\int_{\sigma_q\sqrt{\frac{m^2}{c_m^2}-1}}^{\infty} (1 + \frac{x^2}{\sigma_q^2})q(x)dx = 1. \quad (2.27)$$

3. When $c_m > m$, obviously $a_m(x) = m$. Then by the constraint (2.25) we get $a_m(x) = m = \int_\chi b(x)dx = 2$. It always has the same solution as when $c_m = m$.

In conclusion of 1, 2 and 3, we can constrain c_m to be $0 < c_m \leq m$, and find c_m by (2.27) once m is known. There exists at least one root of c_m in $(0, m]$ because, when $c \to 0+$,

$$\int_\chi \frac{b(x)}{a_m(x)}dx - 1 \to +\infty,$$

and when $c = m$,

$$\int_\chi \frac{b(x)}{a_m(x)}dx - 1 = \int_\chi \frac{b(x)}{m} - 1 \leq 0.$$

Program Structure

To work out the minimizer numerically, the structure of our program is as follows. In the program, we set $\sigma_q = 1$.

(1) Express c_m by a function of m, by solving (2.27) using a nonlinear root finding function, with constraint $0 < c_m \leq m$.

(2) Substitute c_m in (2.24) by the function of m, so as to express $a_m(x)$ by a function of m. Put that $a_m(x)$ into (2.26), then we can express $\alpha(m)$ by a function of m. Minimize $\alpha(m)$, find the minimizer m^* and the corresponding $\alpha(m^*)$ by a nonlinear function minimizer, subject to the constraint that $m \geq 2$.

(3) Obtain the best $p(x)$ by $p(x) = b(x)/a_{m^*}(x)$.

Program Result

In the following, Table 2.1 shows the numerical solutions of m and $\alpha(m)$ when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$, respectively; Figure 2.1 shows the plots of

$p(x)$, $q(x)$ and $w(x) = 0.2w_0(x)$ when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$, respectively. Note that when the weights $w_0(x)$ are multiplied by a positive constant, the value of the loss function in (2.10) will also be multiplied by the same constant, but the optimization problem will not change. So taking $w(x) = 0.2w_0(x)$ here will not affect the optimization result.

ν	0	0.2	0.4	0.6	0.8	1
m	∞	2.2018	2.0285	2.0001	2.0001	2.0001
$\alpha(m)$	1.8348	1.9668	1.9980	2.0000	2.0001	2.0001

Table 2.1: Solutions of m and $\alpha(m)$ when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$, respectively.

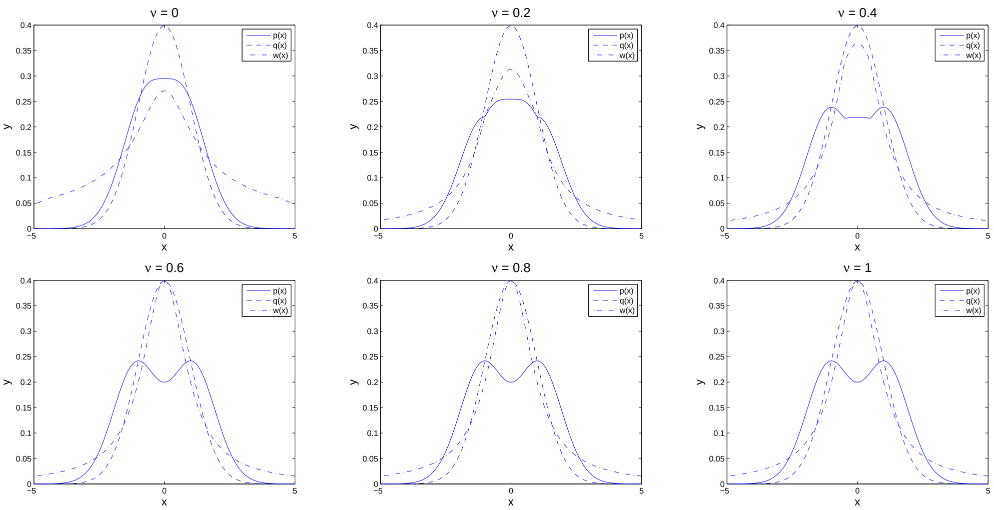

Figure 2.1: Plots of $p(x)$, $q(x)$ and $w(x) = 0.2w_0(x)$ when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$, respectively.

Chapter 3

Active Learning with WLS Estimation and Loss Function Max MSE

3.1 Loss Function Max MSE

In this chapter, we adopt another loss function. We set the loss function to be the maximum of the MSE of $\hat{y}$ in (2.7), and call it Max MSE. Again, we will use a minimax approach to find the best $p(\boldsymbol{x})$. In (2.7), the term $\psi^2(\boldsymbol{x})$ does not depend on $p(\boldsymbol{x})$, so we again concentrate on the leading term of (2.7). We again assume that $\tau_n = \tau = O(1)$. To be specific, the minimax problem is

$$\min_{p} \max_{\psi, \boldsymbol{t}} \boldsymbol{f}'(\boldsymbol{t}) MSE_{\epsilon}\left[\hat{\boldsymbol{\theta}}_{WLS}\right] \boldsymbol{f}(\boldsymbol{t}). \tag{3.1}$$

The $MSE_{\epsilon}\left[\hat{\boldsymbol{\theta}}_{WLS}\right]$ is affected by the randomness of the sample. In order to make the MSE stable, and to make it easy to solve for the best $p(\boldsymbol{x})$, we can replace the $MSE_\epsilon\Big[\hat{\boldsymbol{\theta}}_{WLS}\Big]$ with its asymptotic value in (2.6), so that the minimax problem becomes

$$\min_{p}\max_{\psi,\boldsymbol{t}} \boldsymbol{f}'(\boldsymbol{t})\boldsymbol{U}^{-1}\big\{\sigma_\epsilon^2\boldsymbol{T}_p+\boldsymbol{S}_{\psi,p}\big\}\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{t}), \tag{3.2}$$

with the maximization subject to (1.2) and (1.3).

3.2 Maximization over t and ψ

Assume that dimensions of vector $\boldsymbol{x}$ and $\boldsymbol{t}$ are both 1; $t,\ x\in\chi=[-1,1]$; $\boldsymbol{f}(x)=[1,x]'$, $\boldsymbol{f}(t)=[1,t]'$; both p and q are symmetric densities. Notice that we have changed the assumption of χ from $(-\infty,+\infty)$ in Section 2.4.2 to $[-1,1]$ in this section, because the loss functions of these two sections are different. If t is not bounded, in this section the loss function will go to infinity as $t\to\infty$.

With the above assumptions, we have

$$\begin{aligned}
&\max_{\psi,t}\boldsymbol{f}'(t)\boldsymbol{U}^{-1}\big\{\sigma_\epsilon^2\boldsymbol{T}_p+\boldsymbol{S}_{\psi,p}\big\}\boldsymbol{U}^{-1}\boldsymbol{f}(t)\\
&=\max_{\psi}\max_{t}\left\{\sigma_\epsilon^2\boldsymbol{f}'(t)\boldsymbol{U}^{-1}\boldsymbol{T}_p\boldsymbol{U}^{-1}\boldsymbol{f}(t)+\boldsymbol{f}'(t)\boldsymbol{U}^{-1}\boldsymbol{S}_{\psi,p}\boldsymbol{U}^{-1}\boldsymbol{f}(t)\right\}\\
&=\max_{\psi}\max_{t}\left\{\sigma_\epsilon^2\int_\chi[\boldsymbol{f}'(t)\boldsymbol{U}^{-1}\boldsymbol{f}(x)]^2\frac{q^2(x)}{p(x)}dx+\int_\chi[\boldsymbol{f}'(t)\boldsymbol{U}^{-1}\boldsymbol{f}(x)]^2\frac{q^2(x)}{p(x)}\psi^2(x)dx\right\}\\
&=\max_{\psi}\max_{t}\left\{\sigma_\epsilon^2\int_{-1}^{1}(1+\frac{1}{\sigma_q^4}t^2x^2)\frac{q^2(x)}{p(x)}dx+\int_{-1}^{1}(1+\frac{1}{\sigma_q^4}t^2x^2)\frac{q^2(x)}{p(x)}\psi^2(x)dx\right\},
\end{aligned} \tag{3.3}$$

where $\sigma_q^2=\int_{-1}^{1}x^2q(x)dx$. Since $t\in[-1,1]$, (3.3) is maximized with $t=$

-1 *or* 1. So (3.3) becomes

$$\max_{\psi}\left\{\sigma_\epsilon^2\int_{-1}^{1}(1+\frac{1}{\sigma_q^4}x^2)\frac{q^2(x)}{p(x)}dx+\int_{-1}^{1}(1+\frac{1}{\sigma_q^4}x^2)\frac{q^2(x)}{p(x)}\psi^2(x)dx\right\} \tag{3.4}$$

Similarly with Section 2.3, denote

$$a_p(x)=(1+\frac{1}{\sigma_q^4}x^2)\frac{q(x)}{p(x)},$$

$$h(x)=\psi^2(x)q(x)/\tau^2,$$

then (3.4) becomes

$$\sigma_\epsilon^2 E_q[a_p(x)]+\tau_n^2\max_h\int_{-1}^{1}a_p(x)h(x)dx. \tag{3.5}$$

Then similarly with Claim 2.3 and 2.4, it can be proved that (3.5) is maximized by

$$h(x)=\delta(x-x^*),$$

where $\delta(x)$ is a delta function and $x^*=arg\max_x a_p(x)$. The maximum of (3.5) is

$$\begin{aligned}&\sigma_\epsilon^2 E_q[a_p(x)]+\tau^2\max_x a_p(x)\\ &=(1-\nu)E_q[a_p(\boldsymbol{x})]+\nu\max_{\boldsymbol{x}} a_p(\boldsymbol{x})\end{aligned} \tag{3.6}$$

where $\nu=\tau^2/(\sigma_\epsilon^2+\tau^2)\in[0,1]$.

3.3 Minimization over p

Under the assumptions in Section 3.2, the form of the minimization problem (3.6) in this chapter is the same with the minimization problem (2.16) in Chapter 2, so the forms of general solutions are also identical.

When it comes to the exact example, we also work on a straight line example here, and inherit all the assumptions and notations in Chapter 2, except for two differences:

(1) The domain χ has changed to $[-1, 1]$.

(2) Since the domain has changed, we also change the assumption of $q(x)$ from the standard normal density in Chapter 2 to

$$q(x) = \frac{3}{4}(1 - x^2). \tag{3.7}$$

Now $\sigma_q^2 = \frac{1}{5}$.

Because of the above changes, the steps to determine the values of c_m have changed, too.

1. When $c_m \leq 0$, it is impossible, because $a_m(x) > 0$.

2. When $0 < c_m \leq \frac{m}{\sqrt{26}}$, $a_-(x) \leq m$ always holds, so c_m is defined by

$$\int_\chi \frac{b(x)}{a_-(x)} dx = 1,$$

i.e.

$$\frac{1}{c_m} \int_{-1}^{1} \sqrt{1 + 25x^2} q(x) dx = 1. \tag{3.8}$$

3. When $\frac{m}{\sqrt{26}} < c_m \leq m$, c_m is defined by

$$\int_{a_{-}(x)<m} \frac{b(x)}{a_{-}(x)}dx + \int_{a_{-}(x)\geq m} \frac{b(x)}{m}dx = 1,$$

i.e.

$$\int_{-\frac{1}{5}\sqrt{\frac{m^2}{c_m^2}-1}}^{\frac{1}{5}\sqrt{\frac{m^2}{c_m^2}-1}} \frac{b(x)}{a_{-}(x)}dx + \int_{\frac{1}{5}\sqrt{\frac{m^2}{c_m^2}-1}}^{1} \frac{b(x)}{m}dx + \int_{-1}^{-\frac{1}{5}\sqrt{\frac{m^2}{c_m^2}-1}} \frac{b(x)}{m}dx = 1.$$

Because the integrands are even functions, the above is equivalent to

$$2\int_{0}^{\frac{1}{5}\sqrt{\frac{m^2}{c_m^2}-1}} \frac{b(x)}{a_{-}(x)}dx + 2\int_{\frac{1}{5}\sqrt{\frac{m^2}{c_m^2}-1}}^{1} \frac{b(x)}{m}dx = 1,$$

i.e.

$$\frac{2}{c_m}\int_{0}^{\frac{1}{5}\sqrt{\frac{m^2}{c_m^2}-1}} \sqrt{1+25x^2}q(x)dx + \frac{2}{m}\int_{\frac{1}{5}\sqrt{\frac{m^2}{c_m^2}-1}}^{1} (1+25x^2)q(x)dx = 1. \quad (3.9)$$

4. When $c_m > m$, obviously $a_m(x) = m$. Then we get $a_m(x) = m = \int_{\chi} b(x)dx = 6$. It always has the same solution as when $c_m = m$.

In conclusion of 1, 2, 3 and 4, we can first constrain c_m to be $0 < c_m \leq m$, and then find c_m by (3.8) if $0 < c_m \leq \frac{m}{\sqrt{26}}$, and by (3.9) if $\frac{m}{\sqrt{26}} < c_m \leq m$.

Other parts of the solution are the same with those in Section 2.4.2.

Program Structure

The program structure is also similar to that of Chapter 2.

(1) Define an equation to be (3.8) if $0 < c_m \leq \frac{m}{\sqrt{26}}$, and (3.9) if $\frac{m}{\sqrt{26}} < c_m \leq m$. Obtain a function of m to get c_m by solving the above equation using a

nonlinear root finding function, subject to $0 < c_m \leq m$.

(2) With the c_m, we can express $a_m(x)$ by a function of m and x. Put that $a_m(x)$ into (3.6), find the minimizer m^* and the corresponding $\alpha(m^*)$ by a nonlinear function minimizer, subject to the constraint that $m \geq 6$. (Since $\min_p \max_x a_p(x) = 6$.)

(3) Obtain the best $p(x)$ by $p(x) = b(x)/a_{m^*}(x)$.

Program Result

In the following, Table 3.1 shows the numerical solutions of m and $\alpha(m)$ when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$, respectively; Figure 3.1 shows the plots of $p(x)$, $q(x)$ and $w(x) = 0.3w_0(x)$ when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$, respectively. Again, taking $w(x) = 0.3w_0(x)$ here will not change the optimization result.

ν	0	0.2	0.4	0.6	0.8	1
m	11.2484	6.9736	6.4365	6.1582	6.0088	6.0000
$\alpha(m)$	4.9472	5.4916	5.7800	5.9420	5.9993	6.0000

Table 3.1: Solutions of m and $\alpha(m)$ when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$, respectively.

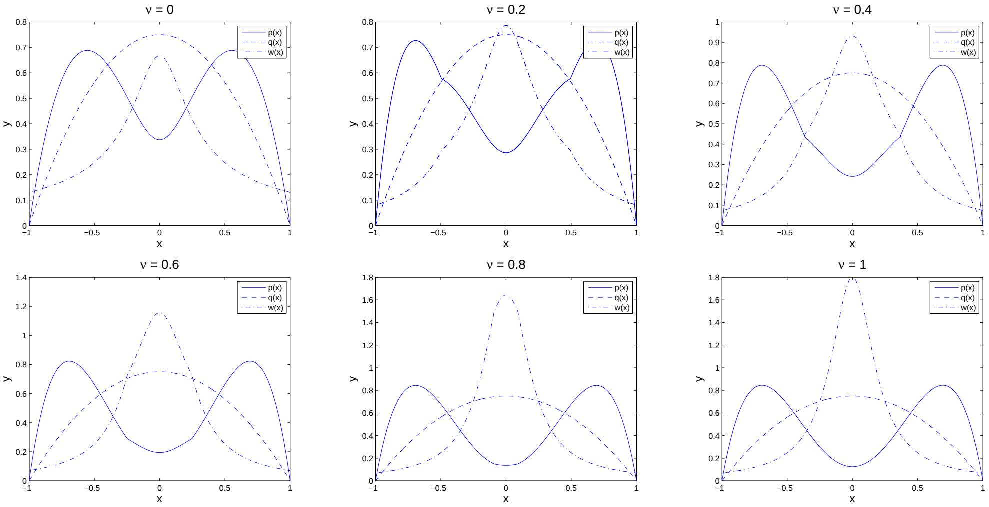

Figure 3.1: Plots of $p(x)$, $q(x)$ and $w(x) = 0.3w_0(x)$ when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$, respectively.

Chapter 4

Active Learning with OLS Estimation and Loss Function IMSE

4.1 Problem Formulation

In this chapter, we explore the solution when the parameter vector $\boldsymbol{\theta}$ is estimated by the most commonly used Ordinary Least Squares (OLS) estimate. Again, we use IMSE as the loss function in this chapter.

With OLS, the weight $w \equiv 1$. Unlike the WLS estimate, we will not get an asymptotically unbiased estimate this time. Thus we assume $\tau_n = \tau/\sqrt{n}$, then $\tau_n^2 = O(n^{-1})$, so that $\boldsymbol{S}_{\psi,w,p} = O(n^{-1})$ can be ignored, which will simplify the problem a bit. To be specific, when $\boldsymbol{b}_{\psi,w,p} \neq \boldsymbol{0}$, by (2.4) and (2.8),

$$
\begin{aligned}
IMSE \quad & = \frac{1}{n} tr\Big[\boldsymbol{U}\boldsymbol{M}_{w,p}^{-1}\Big\{\sigma_\epsilon^2 \boldsymbol{D}_{w,p} + \boldsymbol{S}_{\psi,w,p} + n\boldsymbol{b}_{\psi,w,p}\boldsymbol{b}'_{\psi,w,p}\Big\}\boldsymbol{M}_{w,p}^{-1}\Big] \\
& + \int_\chi \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} + O(n^{-3/2}) \\
& = \frac{1}{n} tr\Big[\boldsymbol{U}\boldsymbol{M}_{w,p}^{-1}\Big\{\sigma_\epsilon^2 \boldsymbol{D}_{w,p} + \boldsymbol{S}_{\psi,w,p} + \boldsymbol{b}_{\sqrt{n}\psi,w,p}\boldsymbol{b}'_{\sqrt{n}\psi,w,p}\Big\}\boldsymbol{M}_{w,p}^{-1}\Big] \\
& + \int_\chi \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} + O(n^{-3/2}) \qquad (4.1) \\
& = \frac{1}{n} tr\Big[\boldsymbol{U}\boldsymbol{M}_{w,p}^{-1}\Big\{\sigma_\epsilon^2 \boldsymbol{D}_{w,p} + \boldsymbol{S}_{\psi,w,p}\Big\}\boldsymbol{M}_{w,p}^{-1} + \boldsymbol{b}'_{\sqrt{n}\psi,w,p}\boldsymbol{M}_{w,p}^{-1}\boldsymbol{U}\boldsymbol{M}_{w,p}^{-1}\boldsymbol{b}_{\sqrt{n}\psi,w,p}\Big] \\
& + \int_\chi \psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} + O(n^{-3/2}). \qquad (4.2)
\end{aligned}
$$

Since $\tau_n^2 = O(n^{-1})$ implies that

$$
\boldsymbol{S}_{\psi,w,p} = \int_\chi \boldsymbol{f}(\boldsymbol{x})w^2(\boldsymbol{x})\psi^2(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} - \boldsymbol{b}_{\psi,w,p}\boldsymbol{b}'_{\psi,w,p} = O(n^{-1}),
$$

and since

$$
\boldsymbol{D}_{w,p} = \int_\chi \boldsymbol{f}(\boldsymbol{x})w^2(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = O(1),
$$

$$
\begin{aligned}
\boldsymbol{b}_{\sqrt{n}\psi,w,p}\boldsymbol{b}'_{\sqrt{n}\psi,w,p} & = \Big(\int_\chi \boldsymbol{f}(\boldsymbol{x})w(\boldsymbol{x})\sqrt{n}\psi(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\Big)\Big(\int_\chi \boldsymbol{f}(\boldsymbol{x})w(\boldsymbol{x})\sqrt{n}\psi(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\Big)' \\
& = O(1),
\end{aligned}
$$

from (4.1), we can see that it is reasonable to ignore $\boldsymbol{S}_{\psi,w,p}$.

When $w = 1$,

$$
\begin{aligned}
\boldsymbol{M}_{1,p} = \boldsymbol{D}_{1,p} = \textstyle\int_\chi \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \overset{\text{def}}{=} \boldsymbol{V}_p, \\
\boldsymbol{b}_{\sqrt{n}\psi,1,p} = \textstyle\int_\chi \boldsymbol{f}(\boldsymbol{x})\sqrt{n}\psi(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \overset{\text{def}}{=} \boldsymbol{b}_{\sqrt{n}\psi,p},
\end{aligned}
$$

Put them into (4.2), and remember that $\boldsymbol{S}_{\psi,w,p}$ is ignored, so the optimization problem becomes

$$\min_{p}\left\{\sigma_{\epsilon}^{2}tr\left[\boldsymbol{U}\boldsymbol{V}_{p}^{-1}\right]+\max_{\psi}\left\{\boldsymbol{b}'_{\sqrt{n}\psi,p}\boldsymbol{V}_{p}^{-1}\boldsymbol{U}\boldsymbol{V}_{p}^{-1}\boldsymbol{b}_{\sqrt{n}\psi,p}+\int_{\chi}n\psi^{2}(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\right\}\right\}, \tag{4.3}$$

with the maximization done subject to (1.2) and (1.3), and minimization subject to the constraint that $p(\boldsymbol{x})$ be a probability density.

4.2 Maximization over ψ

In the following, we will solve this minimax problem, the approach we use first appeared in (Wiens, 1992), but requires modifications, which are made here.

Define

$$\boldsymbol{H}_{p}=\boldsymbol{V}_{p}\boldsymbol{U}^{-1}\boldsymbol{V}_{p},$$

$$\boldsymbol{K}_{p}=\int_{\chi}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})\tfrac{p^{2}(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x},$$

$$\boldsymbol{G}_{p}=\boldsymbol{K}_{p}-\boldsymbol{H}_{p}=\int_{\chi}\left[(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\boldsymbol{I}_{r}-\boldsymbol{V}_{p}\boldsymbol{U}^{-1})\boldsymbol{f}(\boldsymbol{x})\right]\left[(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\boldsymbol{I}_{r}-\boldsymbol{V}_{p}\boldsymbol{U}^{-1})\boldsymbol{f}(\boldsymbol{x})\right]'q(\boldsymbol{x})d\boldsymbol{x}.$$

The matrix $\boldsymbol{G}_p$ is clearly positive semi-definite. Assume that $\boldsymbol{G}_p$ is positive definite. If not, we could first perturb it to make it non-singular, and then pass to the limit (Heo, Schmuland, and Wiens, 2001).

Define

$$\boldsymbol{r}(x)=(\tau/\sqrt{n})\boldsymbol{G}_{p}^{-1/2}\Big(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\boldsymbol{I}_{r}-\boldsymbol{V}_{p}\boldsymbol{U}^{-1}\Big)\boldsymbol{f}(\boldsymbol{x}).$$

Denote Ψ as the class of all $\psi(x)$ defined in (1.2) and (1.3).

Lemma 4.1: *The class* $\Psi_0=\{\psi_{\boldsymbol{\beta}}(\boldsymbol{x})=\boldsymbol{r}'(\boldsymbol{x})\boldsymbol{\beta}\mid\|\boldsymbol{\beta}\|=1\}$ *is a sub-class of* Ψ.

Proof:

$$\begin{aligned}
&\int_\chi \boldsymbol{f}(\boldsymbol{x})\psi_{\boldsymbol{\beta}}(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} \\
=&\int_\chi \boldsymbol{f}(\boldsymbol{x})\boldsymbol{r}'(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\cdot\boldsymbol{\beta} \\
=&\int_\chi \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})q(\boldsymbol{x})\Big(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\boldsymbol{I}_r-\boldsymbol{V}_p\boldsymbol{U}^{-1}\Big)'d\boldsymbol{x}\cdot\boldsymbol{G}_p^{-1/2}\boldsymbol{\beta} \\
=&\Big(\boldsymbol{V}_p-\boldsymbol{V}_p\Big)\boldsymbol{G}_p^{-1/2}\boldsymbol{\beta} \\
=&\ 0.
\end{aligned}$$

So (1.2) *is satisfied.*

$$\begin{aligned}
&\int_\chi \psi^2_{\boldsymbol{\beta}}(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} \\
&=\boldsymbol{\beta}'\int_\chi \boldsymbol{r}'(\boldsymbol{x})\boldsymbol{r}(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\boldsymbol{\beta} \\
&=\frac{\tau^2}{n}\boldsymbol{\beta}'\boldsymbol{G}_p^{-1/2}\cdot\int_\chi\Big[(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\boldsymbol{I}_r-\boldsymbol{V}_p\boldsymbol{U}^{-1})\boldsymbol{f}(\boldsymbol{x})\Big]\Big[(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\boldsymbol{I}_r-\boldsymbol{V}_p\boldsymbol{U}^{-1})\boldsymbol{f}(\boldsymbol{x})\Big]'q(\boldsymbol{x})d\boldsymbol{x}\cdot\boldsymbol{G}_p^{-1/2}\boldsymbol{\beta} \\
&=\frac{\tau^2}{n}\boldsymbol{\beta}'\boldsymbol{G}_p^{-1/2}\boldsymbol{G}_p\boldsymbol{G}_p^{-1/2}\boldsymbol{\beta} \\
&=\frac{\tau^2}{n}=\tau_n^2.
\end{aligned} \tag{4.4}$$

So (1.3) *is also satisfied. Thus Lemma 4.1 is proved.*

Lemma 4.2: *The* ψ *that maximizes* (4.2) *is in* Ψ_0.

Proof:

$$\begin{aligned}
&\int_{\chi} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{r}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\
=&\frac{\tau}{\sqrt{n}}\int_{\chi} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})\Big(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\boldsymbol{I}_r - \boldsymbol{V}_p\boldsymbol{U}^{-1}\Big)p(\boldsymbol{x})d\boldsymbol{x}\cdot\boldsymbol{G}_p^{-1/2} \\
=&\frac{\tau}{\sqrt{n}}\Big(\boldsymbol{K}_p - \boldsymbol{V}_p\boldsymbol{U}^{-1}\boldsymbol{V}_p\Big)\cdot\boldsymbol{G}_p^{-1/2} \qquad (4.5)\\
=&\frac{\tau}{\sqrt{n}}\boldsymbol{G}_p\boldsymbol{G}_p^{-1/2} \\
=&\frac{\tau}{\sqrt{n}}\boldsymbol{G}_p^{1/2}
\end{aligned}$$

Let $\psi \in \Psi$ *be arbitrary and set* $\boldsymbol{\beta}_* = \frac{\boldsymbol{G}_p^{-1/2}\boldsymbol{b}_{\sqrt{n}\psi,p}}{\left\|\boldsymbol{G}_p^{-1/2}\boldsymbol{b}_{\sqrt{n}\psi,p}\right\|}$. *By* (4.5),

$$\begin{aligned}
&\boldsymbol{b}'_{\sqrt{n}\psi_{\beta_*},p}\boldsymbol{V}_p^{-1}\boldsymbol{U}\boldsymbol{V}_p^{-1}\boldsymbol{b}_{\sqrt{n}\psi_{\boldsymbol{\beta}_*},p} \\
=&\int_{\chi}\boldsymbol{f}'(\boldsymbol{x})\sqrt{n}\psi_{\boldsymbol{\beta}_*}(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\cdot\boldsymbol{V}_p^{-1}\boldsymbol{U}\boldsymbol{V}_p^{-1}\cdot\int_{\chi}\boldsymbol{f}(\boldsymbol{x})\sqrt{n}\psi_{\boldsymbol{\beta}_*}(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\
=&\,n\boldsymbol{\beta}'_*\cdot\int_{\chi}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{r}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\cdot\boldsymbol{V}_p^{-1}\boldsymbol{U}\boldsymbol{V}_p^{-1}\cdot\int_{\chi}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{r}'(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\cdot\boldsymbol{\beta}_* \qquad (4.6)\\
=&\,n\cdot\frac{\tau^2}{n}\boldsymbol{\beta}'_*\boldsymbol{G}_p^{1/2}\boldsymbol{H}_p^{-1}\boldsymbol{G}_p^{1/2}\boldsymbol{\beta}_* \\
=&\,\tau^2\frac{\boldsymbol{b}'_{\sqrt{n}\psi,p}\boldsymbol{H}_p^{-1}\boldsymbol{b}_{\sqrt{n}\psi,p}}{\left\|\boldsymbol{G}_p^{-1/2}\boldsymbol{b}_{\sqrt{n}\psi,p}\right\|^2}
\end{aligned}$$

Therefore, by (4.6) *and* (4.4), *the IMSE* (4.2) *evaluated at* $\psi_{\boldsymbol{\beta}_*}$ *gives*

$$IMSE_{|\psi_{\boldsymbol{\beta}_*}} = \frac{1}{n}\Big\{\sigma_\epsilon^2 tr\big[\boldsymbol{U}\boldsymbol{V}_p^{-1}\big] + \tau^2\frac{\boldsymbol{b}'_{\sqrt{n}\psi,p}\boldsymbol{H}_p^{-1}\boldsymbol{b}_{\sqrt{n}\psi,p}}{\left\|\boldsymbol{G}_p^{-1/2}\boldsymbol{b}_{\sqrt{n}\psi,p}\right\|^2} + \tau^2\Big\}. \qquad (4.7)$$

Also, we have

$$\begin{aligned}
&\int_{\chi} \boldsymbol{r}(\boldsymbol{x})\psi(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} \\
&= \frac{\tau}{\sqrt{n}}\boldsymbol{G}_p^{-1/2}\int_{\chi}\Big(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\boldsymbol{I}_r - \boldsymbol{V}_p\boldsymbol{U}^{-1}\Big)\boldsymbol{f}(\boldsymbol{x})\psi(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} \\
&= \frac{\tau}{\sqrt{n}}\boldsymbol{G}_p^{-1/2}\cdot\frac{1}{\sqrt{n}}\Big(\boldsymbol{b}_{\sqrt{n}\psi,p} - \sqrt{n}\boldsymbol{V}_p\boldsymbol{U}^{-1}\int_{\chi}\boldsymbol{f}(\boldsymbol{x})\psi(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\Big) \\
&= \frac{\tau}{n}\boldsymbol{G}_p^{-1/2}\boldsymbol{b}_{\sqrt{n}\psi,p},
\end{aligned}$$

by the constraint (1.2). *Then by* (1.3), (4.4) *and Cauchy-Schwarz Inequality,*

$$\begin{aligned}
\frac{\tau^2}{n} &\geq \sqrt{\int_{\chi}\psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}}\sqrt{\int_{\chi}\psi^2_{\boldsymbol{\beta}_*}(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}} \\
&\geq \Big|\int_{\chi}\psi(\boldsymbol{x})\psi_{\boldsymbol{\beta}_*}(\boldsymbol{x})d\boldsymbol{x}\Big| \\
&= \Big|\int_{\chi}\psi(\boldsymbol{x})\boldsymbol{r}'(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\cdot\boldsymbol{\beta}_*\Big| \\
&= \frac{\tau}{n}\left\|\boldsymbol{G}_p^{-1/2}\boldsymbol{b}_{\sqrt{n}\psi,p}\right\|,
\end{aligned}$$

so that

$$\left\|\boldsymbol{G}_p^{-1/2}\boldsymbol{b}_{\sqrt{n}\psi,p}\right\| \leq \tau.$$

Therefore,

$$\begin{aligned}
IMSE_{|\psi_{\boldsymbol{\beta}_*}} &= \frac{1}{n}\Big\{\sigma_\epsilon^2 tr\big[\boldsymbol{U}\boldsymbol{V}_p^{-1}\big] + \tau^2\frac{\boldsymbol{b}'_{\sqrt{n}\psi,p}\boldsymbol{H}_p^{-1}\boldsymbol{b}_{\sqrt{n}\psi,p}}{\left\|\boldsymbol{G}_p^{-1/2}\boldsymbol{b}_{\sqrt{n}\psi,p}\right\|^2} + \tau^2\Big\} \\
&\geq \frac{1}{n}\Big\{\sigma_\epsilon^2 tr\big[\boldsymbol{U}\boldsymbol{V}_p^{-1}\big] + \tau^2\frac{\boldsymbol{b}'_{\sqrt{n}\psi,p}\boldsymbol{H}_p^{-1}\boldsymbol{b}_{\sqrt{n}\psi,p}}{\tau^2} + \int_{\chi}n\psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}\Big\} \\
&= IMSE_{|\psi}.
\end{aligned}$$

Since $\psi \in \Psi$ *is arbitrary and* $\psi_{\beta_*} \in \Psi_0$, *we know that* Ψ_0 *contains the* ψ *that* *maximizes* (4.2)*, thus Lemma 4.2 is proved.*

Evaluating (4.2) at $\psi_{\boldsymbol{\beta}}$ for arbitrary β gives

$$IMSE_{|\psi_\beta} = \frac{\sigma_\epsilon^2}{n} tr\big[\boldsymbol{U}\boldsymbol{V}_p^{-1}\big] + \frac{\tau^2}{n}\boldsymbol{\beta}'\Big(\boldsymbol{G}_p^{1/2}\boldsymbol{H}_p\boldsymbol{G}_p^{1/2} + \boldsymbol{I}_r\Big)\boldsymbol{\beta};$$

now maximizing over β yields the result that $\max_\psi IMSE$ is $(\sigma_\epsilon^2+\tau^2)/n$ times

$$L_\nu(p) = (1-\nu)tr\big[\boldsymbol{U}\boldsymbol{V}_p^{-1}\big] + \nu ch_{max}\big[\boldsymbol{K}_p\boldsymbol{H}_p^{-1}\big],$$

where $\nu = \tau^2/(\sigma_\epsilon^2+\tau^2)$ and ch_{max} denotes the maximum eigenvalue.

Now what is left is to find a density $p(x)$ that minimize $L_\nu(p)$, and that requires specification of models. In the following, we will discuss the solution of the above problem with a straight line model.

4.3 Minimization over p with a Straight Line Model

Assume that $\boldsymbol{f}(x) = [1, x]'$, $\chi = (-\infty, \infty)$, both $p(x)$ and $q(x)$ are symmetric, and the variance of $q(x)$ is σ_q^2. Then

$$L_\nu(p) = (1-\nu)\bigg(1+\frac{\sigma_q^2}{\int_\chi x^2 p(x)dx}\bigg) + \nu\ max\bigg(\int_X \frac{p^2(x)}{q(x)}dx,\ \frac{\int_X x^2\frac{p^2(x)}{q(x)}dx}{(\int_\chi x^2p(x)dx)^2}\sigma_q^2\bigg), \tag{4.8}$$

where $\int_\chi \frac{p^2(x)}{q(x)}dx$ and $\frac{\int_\chi x^2\frac{p^2(x)}{q(x)}dx}{(\int_\chi x^2p(x)dx)^2}\sigma_q^2$ are the two eigenvalues of $\boldsymbol{K}_p\boldsymbol{H}_p^{-1}$.

Then we solve the problem with the approach appeared in (Daemi and Wiens, 2013). Denote the above two eigenvalues as $E_1(p)$ and $E_2(p)$. Suppose we choose the first eigenvalue $E_1(p)$ in (4.8) and find the minimizing density

$p_1(x)$ of

$$L_{\nu,1}(p) = (1-\nu)\left(1 + \frac{\sigma_q^2}{\int_{\mathcal{X}} x^2 p(x)dx}\right) + \nu \int_X \frac{p^2(x)}{q(x)}dx$$

Lemma 4.3: *If $L_{\nu,1}(p_1) \geq L_{\nu,2}(p_1)$, then $p_1(x)$ is the minimax solution of $L_\nu(p)$ (4.8).*

Proof:

$$\begin{aligned} L_\nu(p) = {} & \max\left\{L_{\nu,1}(p), L_{\nu,2}(p)\right\} \geq L_{\nu,1}(p) \\ & \geq L_{\nu,1}(p_1) = \max\left\{L_{\nu,1}(p_1), L_{\nu,1}(p_1)\right\} \\ & = L_\nu(p_1). \end{aligned}$$

Thus Lemma 4.3 is proved.

Similarly, if we choose the second eigenvalue $E_2(p)$ in (4.8) and find the minimizing $p_2(x)$ of

$$L_{\nu,2}(p) = (1-\nu)\left(1 + \frac{\sigma_q^2}{\int_{\mathcal{X}} x^2 p(x)dx}\right) + \nu \, \frac{\int_{\mathcal{X}} x^2 \frac{p^2(x)}{q(x)}dx}{(\int_{\mathcal{X}} x^2 p(x)dx)^2}\sigma_q^2.$$

Then if we can prove $L_{\nu,2}(p_2) \geq L_{\nu,1}(p_2)$, then $p_2(x)$ is the minimax solution of $L_\nu(p)$ (4.8).

Therefore, in the following, we will solve the minimization problem in three steps.

Notice that the numerical solution is dependent on $q(x)$, so in the following we assume $q(x)$ is the standard normal density.

Step 1. Assume that at the minimizing $p(x)$, $L_{\nu,1}(p) \geq L_{\nu,2}(p)$, i.e., $E_1(p) \geq E_2(p)$. Find the minimizing $p(x)$, and then check if the above assumption truly holds. If so, then this $p(x)$ is the solution of the minimax problem (4.3); if not, continue to Step 2.

Step 2. Assume that at the minimizing $p(x)$, $L_{\nu,2}(p) \geq L_{\nu,1}(p)$, i.e., $E_2(p) \geq E_1(p)$. Find the minimizing $p(x)$, and then check if the above assumption truly holds. If so, then this $p(x)$ is the solution of the minimax problem (4.3); if not, continue to Step 3.

Step 3. If both Step 1 and Step 2 fail, (we can see from Sections 4.3.1 and 4.3.2 that this is truly the case here), then a general construction technique can solve the problem and that is practiced in the following Step 3 section.

4.3.1 Step 1

First assume $E_1(p) \geq E_2(p)$ at the minimizing $p(x)$, then we only need to minimize

$$L_{\nu,1}(p) = (1-\nu)\left(1 + \frac{\sigma_q^2}{\int_\chi x^2 p(x)dx}\right) + \nu \int_X \frac{p^2(x)}{q(x)}dx. \tag{4.9}$$

First fix the $\int_\chi x^2 p(x)dx$, let it equal σ_p^2, then the first term of $L_{\nu,1}(p)$ is fixed. The problem of minimizing $L_{\nu,1}(p)$ becomes

$$\min_p \int_\chi \frac{p^2(x)}{q(x)}dx$$

subject to

$$\int_\chi p(x)dx = 1,$$
$$\int_\chi x^2 p(x) = \sigma_p^2.$$

Use Lagrange Multiplier and the above problem is equivalent to minimizing

$$\min_{p} \int_{\chi} \left\{ \frac{p^2(x)}{q(x)} - \lambda_1 p(x) - \lambda_2 x^2 p(x) \right\} dx.$$

It is sufficient to only minimize the integrand pointwise over $p(x) \geq 0$. Since the integrand is a quadratic function of $p(x)$ and the coefficient of $p^2(x)$ is positive, the critical point

$$p(x) = \frac{\lambda_1 + \lambda_2 x^2}{2} q(x)$$

is the minimum point.

The density $p(x)$ has to be non-negative. If the critical point is negative, then $p(x)$ is increasing in the non-negative domain, so in this situation $p(x) = 0$ will be the minimizer. Thus the minimizer is in this form:

$$p(x) = (a + bx^2)^+ q(x).$$

Here the notation "$(a + bx^2)^+$" means:

$$(a + bx^2)^+ = \begin{cases} a + bx^2 & if \ a + bx^2 > 0 \\ 0 & if \ a + bx^2 \leq 0 \end{cases}$$

We substitute $p(x)$ in the $L_{\nu,1}(p)$ (4.9) with the above form, then use the non-linear minimizer to find the values of a and b that minimize $L_{\nu,1}(p)$. Note that $p(x)$ should be a density function, so every time before running the minimizer, we divide a and b by the integration of $p(x)$. This is to make sure the values of a and b we get always make $p(x)$ have an integration of 1. Finally, with the solved a and b, we go back to check if the assumption $E_1(p) \geq E_2(p)$ holds.

The first plot in Figure 4.1 shows the values of $L_{\nu,1}(p)$ at the minimizing $p(x)$ against values of ν. The second figure shows the values of a and b against values of ν. The third plot shows the two eigenvalues $E_1(p)$ and $E_2(p)$ at the minimizing $p(x)$ against different values of ν. From the third plot, we can see that at some values of ν (such as when $\nu = 0.5,\ 0.6$), the second eigenvalue is bigger than the first eigenvalue, and that is against our assumption that $E_1(p) \geq E_2(p)$, so the method in Step 1 fails at those values of ν.

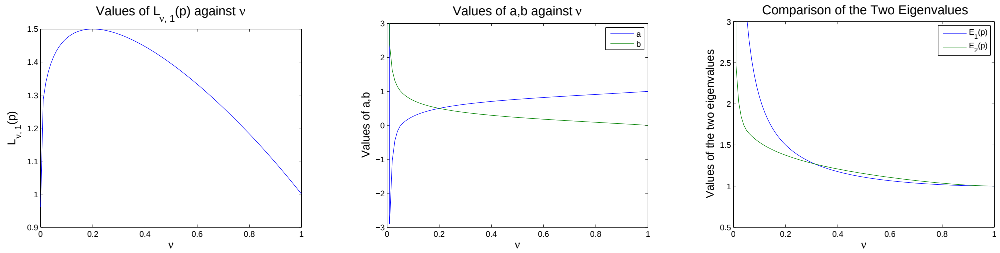

Figure 4.1: Values of $L_{\nu,1}(p)$, a and b, and two eigenvalues against ν in Step 1.

4.3.2 Step 2

Assume that $E_1(p) < E_2(p)$ at the minimizing $p(x)$, now the loss function $L_\nu(p)$ (4.8) becomes

$$L_{\nu,2}(p) = (1-\nu)\left(1 + \frac{\sigma_q^2}{\int_\chi x^2 p(x)dx}\right) + \nu\ \frac{\int_X x^2 \frac{p^2(x)}{q(x)}dx}{(\int_\chi x^2 p(x)dx)^2}\sigma_q^2. \tag{4.10}$$

Similarly with Step 1, we use Lagrange Multiplier and finally find that the minimizer is in the form

$$p(x) = \left(\frac{a + bx^2}{x^2}\right)^+ q(x). \tag{4.11}$$

Again use similar numerical method to minimize $L_{\nu,2}(p)$, and the results are shown in Figure 4.2. The first plot in Figure 4.2 shows the values of $L_{\nu,2}(p)$ at the minimizing $p(x)$ against values of ν. The second figure shows the values of a and b against values of ν. The third plot shows the two eigenvalues $E_1(p)$ and $E_2(p)$ at the minimizing $p(x)$ against different values of ν. The third plot contradicts the assumption that $E_1(p) < E_2(p)$. So the method in Step 2 also fails.

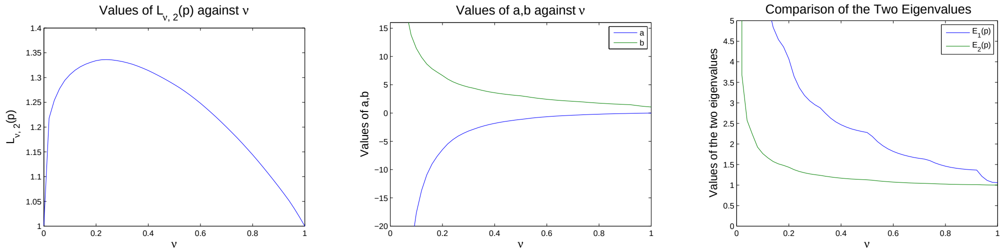

Figure 4.2: Values of $L_{\nu,2}(p)$, a and b, and two eigenvalues against ν in Step 2.

4.3.3 Step 3

1. Fix σ_p^2, find minimizer density $p(x)$ subject to $E_1(p) \geq E_2(p)$. That is

$$\min_{p} \int_{\chi} \frac{p^2(x)}{q(x)} dx,$$

subject to

$$\int_{\chi} p(x)dx = 1,$$

$$\int_{\chi} x^2 p(x)dx = \sigma_p^2,$$

$$\int_{\chi} \left(\frac{p^2(x)}{q(x)} - \frac{\int_{\chi} x^2 \frac{p^2(x)}{q(x)} dx}{\sigma_p^4} \sigma_q^2 - \delta^2 \right) dx = 0,$$

where δ is a slack variable. By using Lagrange Multiplier, the above is equivalent to

$$\min_p \int_\chi \left\{ \frac{p^2(x)}{q(x)} - \lambda_1 p(x) - \lambda_2 x^2 p(x) - \lambda_3 \left(\frac{p^2(x)}{q(x)} - \frac{x^2 p^2(x)}{q(x)} \frac{\sigma_q^2}{\sigma_p^4} \right) \right\} dx, \quad (4.12)$$

where $\lambda_3 \geq 0$. It is sufficient to minimize the integrand pointwise.

The critical point of the integrand is

$$p(x) = \frac{\lambda_1 + \lambda_2 x^2}{2(1-\lambda_3) + 2\lambda_3 \frac{\sigma_q^2}{\sigma_p^4} x^2} q(x). \quad (4.13)$$

2. Fix σ_p^2, find minimizer density $p(x)$ subject to $E_1(p) < E_2(p)$.

Similarly we find that the problem is equivalent to

$$\min_p \int_\chi \left\{ \frac{x^2 p^2(x)}{q(x)} \cdot \frac{\sigma_q^2}{\sigma_p^4} - \lambda_1' p(x) - \lambda_2' x^2 p(x) + \lambda_3' \left(\frac{p^2(x)}{q(x)} - \frac{x^2 p^2(x)}{q(x)} \cdot \frac{\sigma_q^2}{\sigma_p^4} \right) \right\} dx, \quad (4.14)$$

where $\lambda_3' \geq 0$, and the critical point of the integrand is

$$p(x) = \frac{\lambda_1' + \lambda_2' x^2}{2\lambda_3' + 2(1-\lambda_3') \frac{\sigma_q^2}{\sigma_p^4} x^2} q(x). \quad (4.15)$$

3. Now we know that $p(x)$ in (4.13) and (4.15) are critical points of (4.12) and (4.14) respectively, but critical points can be either minimum points or maximum points. We first assume that they are both minimum points, after working out final results, we will come back to prove this assumption. After assuming that they are minimum points, $p(x)$ in (4.12) and (4.14) are both

minimizers, we notice that they are both in the following form:

$$p(x) = \left(\frac{a + bx^2}{c + dx^2}\right)^+ q(x)$$

To avoid over-parameterization, we can divide a, b, c, d by one of them so that there are only three parameters left. It seems better to divide them all by a or b because then cases in Step 1 and Step 2 could be included. (When $c = 1$, $d = 0$, these are cases in Step 1; when $c = 0$, $d = 1$, these are cases in Step 2.) Here we divide them all by a. (In fact experiments show that either dividing by a or by b produces the same result.) Therefore,

$$p(x) = \left(\frac{1 + b'x^2}{c' + d'x^2}\right)^+ q(x). \tag{4.16}$$

We substitute $p(x)$ in the $L_\nu(p)$ (4.8) with the above form, then use the nonlinear minimizer to find the values of b', c' and d' that minimize $L_\nu(p)$. Again, to make sure $p(x)$ is a density, every time before running the minimizer, we multiply c' and d' by the integration of $p(x)$. Results are shown in Table 4.1, Figure 4.3 and 4.4.

From the first plot in Figure 4.4, we can see that when $\nu = 0$, $p(x)$ is similar to a delta function, but with double peaks. That is because, when $\nu = 0$, the minimization problem is equivalent to

$$\max_p \int_\chi x^2 p(x)dx \;\; \textit{subject to} \;\; \int_\chi p(x)dx = 1.$$

The best $p(x)$ to this problem would be a density that concentrates mass at the two points where x^2 reaches maximum. Since $\chi = (-\infty, +\infty)$, theoretically, $p(x)$ has to be a density that concentrates mass at $-\infty$ and $+\infty$. But since

$p(x)$ is restricted to be in the form of (4.16), $|x| = 4.5740$ might be the largest value $p(x)$ can concentrate at. In fact, with this $p(x)$, $L_0(p) = 1.0478$, as shown in Table 4.1, which is very close to the ideal situation when $L_0(p) = 1$, so this result can be accepted.

From the last plot in Figure 4.4, we can see that when $\nu = 1$, $p(x) = q(x)$.

ν	0	0.2	0.4	0.6	0.8	1
$L_\nu(p)$	1.0478	1.5	1.4473	1.3345	1.1836	1
b'	-75.24	1.0	0.5224	0.4482	0.4034	1
c'	5.685e5	2.0	1.467	1.282	1.136	1
d'	-27177.0	0	0.03467	0.1221	0.2281	1

Table 4.1: Values of $L_\nu(p)$ and $b',\ c',\ d'$ at $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.$

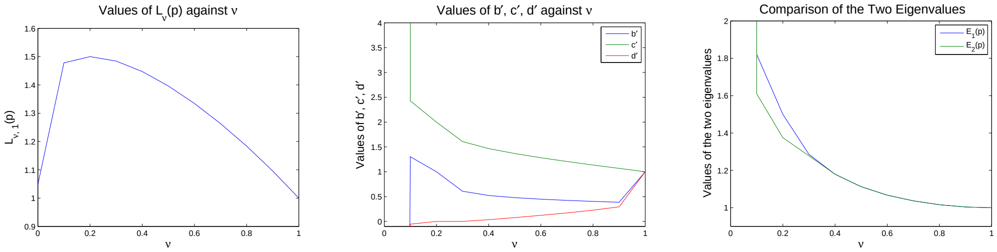

Figure 4.3: Values of $L_\nu(p)$, values of $b',\ c',\ d'$, and two eigenvalues against ν in Step 3. Note that these plots are not truly continuous plots; only points at $\nu = 0, 0.1, 0.2, ..., 1$ are experimental results, then these points are connected.

Finally, in order to prove that $p(x)$ in (4.13) and (4.15) are minimizer of (4.12) and (4.14) respectively, we need to show that the coefficient of $p^2(x)$ in the integrand of (4.12) and (4.14):

$$(1 - \lambda_3) + \lambda_3 \frac{\sigma_q^2}{\sigma_p^4} x^2$$

and

$$\lambda_3' + (1 - \lambda_3') \frac{\sigma_q^2}{\sigma_p^4} x^2$$

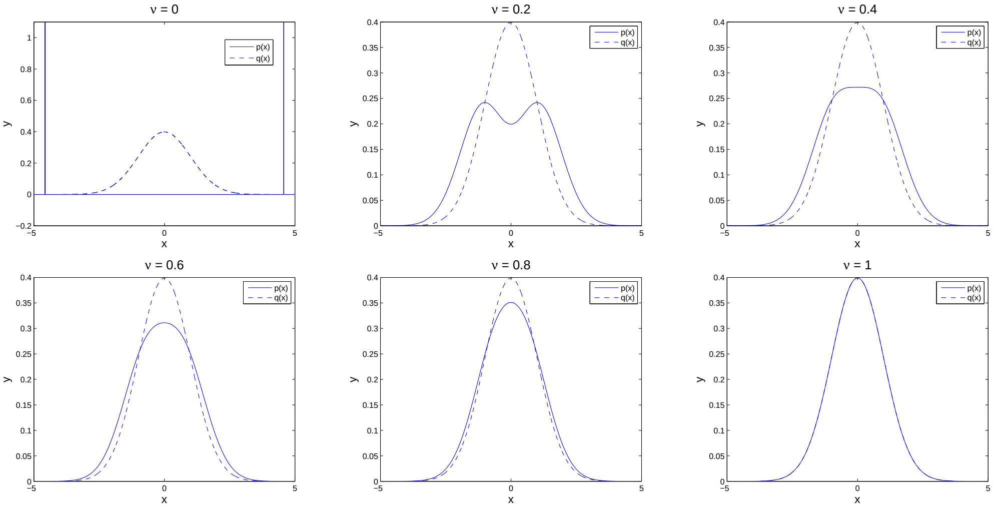

Figure 4.4: Plots of p(x) when $\nu = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$ respectively in Step 3.

are both positive.

Since $\lambda_3 \geq 0$, $\lambda_3' \geq 0$, if we can prove that $1 - \lambda_3$ has the same sign as λ_3 and that $1 - \lambda_3'$ has the same sign as λ_3', then $1 - \lambda_3$ and $1 - \lambda_3'$ will be proved to be positive, too. That is equivalent to show that c' and d' in (4.16) have the same sign. The final results show that, except when $\nu = 0$, which we have already talked about above, c' and d' are always positive, thus c' and d' can be seen as always having the same sign, then we prove that the $p(x)$ we get are truly minimizers.

Chapter 5

Simulations and Comparisons

In this chapter, we apply different forms of error term $\psi(\boldsymbol{x})$ to the problem, and compare the loss of our active learning method with the loss of traditional passive learning method. Because the solutions in this thesis are restricted to model assumptions specified in Chapters 2, 3 and 4 (for the exact assumptions see Sections 2.4.2, 3.2 and 4.3), and it is hard to find real world examples that fit the assumptions, so we only test our solutions on simulations.

We assume that $\psi_i(x) = \tau_n r_i(x)$ ($i \in \{1,\ 2,\ 3\}$; $\tau_n = \tau$ in Chapters 2 and 3, $\tau_n = \tau/\sqrt{n}$ in Chapter 4), and choose the three different forms of $r_i(x)$:

$$r_1(x) = \frac{\frac{x^2}{\sigma_q^2} - 1}{\sqrt{\int_\chi (\frac{x^2}{\sigma_q^2} - 1)^2 q(x) dx}},$$

$$r_2(x) = \frac{\frac{x^3}{E_q[x^4]} - \frac{x}{\sigma_q^2}}{\sqrt{\int_\chi (\frac{x^3}{E_q[x^4]} - \frac{x}{\sigma_q^2})^2 q(x) dx}},$$

$$r_3(x) = \frac{\frac{x^4}{E_q[x^4]} - \frac{x^2}{\sigma_q^2}}{\sqrt{\int_\chi (\frac{x^4}{E_q[x^4]} - \frac{x^2}{\sigma_q^2})^2 q(x) dx}},$$

to represent three different forms of model error. All of these choices of $\psi(x)$

meet the constraint (1.2) and (1.3).

Note that the value of ν, which represents the relative magnitude of model error to the sum of model error and sampling error, is unknown to the experimenter, thus the assumptions of ν by the experimenter might be different from the true value of ν. Let ν_1 represent the assumptions of ν by the experimenter, and $p_{\nu_1}(x)$ represent the design density constructed under the assumptions. Let ν_2 represent the true value of ν. In this chapter, we choose $\nu = 0.2,\ 0.5,\ 0.8$, which is equivalent to $\tau_n = 0.5\sigma_\epsilon,\ \sigma_\epsilon,\ 2\sigma_\epsilon$, to represent situations when the magnitude of model error is less, equal or greater than the magnitude of sampling error. We will compute loss functions with the three different choices of ν_1, ν_2 and $\psi(x)$ respectively, and compare the performances of active learning with that of passive learning.

In Chapters 2, 3 and 4, we use different loss functions and methods to solve the problems, and give different examples to get numerical results. Therefore, in the following three sections, we do simulations following the assumptions and settings of each of the three chapters, and show the results respectively. Each section is divided into two parts. In the first part we still use the asymptotic loss function defined in Chapters 2, 3 and 4, and compare the asymptotic results of passive learning and active learning. In the second part, we carry out experiments and make comparisons based on finite samples. Notice that in the first part, the loss functions we define depend on ν, the relative importance of σ_ϵ and τ_n, not on the true values of σ_ϵ and τ_n, so we do not make assumptions on their true values. However, in the second part, the loss functions not only depend on ν but also on σ_ϵ and τ_n (only knowing one of them is enough), so here we uniformly assume that $\sigma_\epsilon = 1$ in the second part.

5.1 Simulation Results of Chapter 2

5.1.1 The Asymptotic Results

The optimization problem in Chapter 2 is (2.10), so in this section we define the loss function to be

$$tr\big[\boldsymbol{U}^{-1}\{\sigma_\epsilon^2\boldsymbol{T}_p+\boldsymbol{S}_{\psi,p}\}\big].$$

According to assumptions of the example in Chapter 2, we have

$$\boldsymbol{U}=\boldsymbol{I}_2, \tag{5.1}$$

$$\boldsymbol{T}_p=\begin{bmatrix}\int_{-\infty}^{\infty}\frac{q^2(x)}{p(x)}dx & 0\\ 0 & \int_{-\infty}^{\infty}x^2\frac{q^2(x)}{p(x)}dx\end{bmatrix},$$

$$\boldsymbol{S}_{\psi,p}=\begin{bmatrix}\int_{-\infty}^{\infty}\psi^2(x)\frac{q^2(x)}{p(x)}dx & 0\\ 0 & \int_{-\infty}^{\infty}x^2\psi^2(x)\frac{q^2(x)}{p(x)}dx\end{bmatrix}.$$

Then

$$\begin{aligned}&tr\big[\boldsymbol{U}^{-1}\{\sigma_\epsilon^2\boldsymbol{T}_p+\boldsymbol{S}_{\psi,p}\}\big]\\&=\int_{-\infty}^{\infty}\Big(\sigma_\epsilon^2+\psi^2(x)\Big)\Big(1+x^2\Big)\frac{q^2(x)}{p(x)}dx\\&=(\sigma_\epsilon^2+\tau^2)\int_{-\infty}^{\infty}\Big((1-\nu)+\nu r^2(x)\Big)\Big(1+x^2\Big)\frac{q^2(x)}{p(x)}dx.\end{aligned}$$

Therefore, we can change the definition of the loss function to be

$$Loss_1(p_{\nu_1},\nu_2,r_i)=\int_{-\infty}^{\infty}\Big((1-\nu_2)+\nu_2r_i^2(x)\Big)\Big(1+x^2\Big)\frac{q^2(x)}{p_{\nu_1}(x)}dx.\ \ (i\in\{1,\ 2,\ 3\}) \tag{5.2}$$

We compute (5.2) with the three $p_{\nu_1}(x)$ to get the losses of active learning method. Then we replace $p_{\nu_1}(x)$ with $q(x)$ in the above computation to get the losses of passive learning method. The results are shown below. In Figure 5.1, we compare plots of the design densities for active learning and passive learning with the three different choices of ν_1. Table 5.1 shows the losses of active learning and passive learning with different ν_1, ν_2, and $r_i(x)$. The table shows that active learning always has lower loss than passive learning, thus it is more advantageous in these cases.

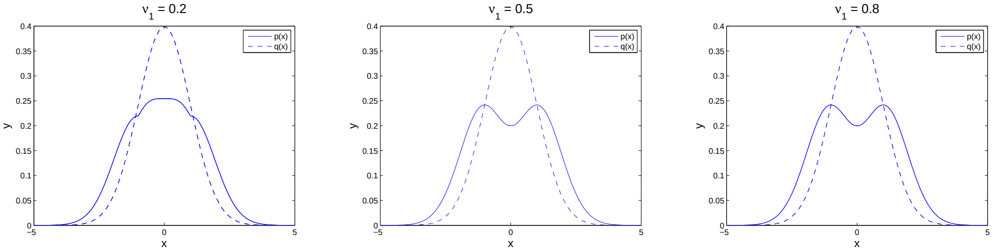

Figure 5.1: Plots of $p_{\nu_1}(x)$ and $q(x)$ when $\nu_1 = 0.2,\ 0.5,\ 0.8$ respectively.

	r_1			r_2			r_3		
ν_2	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
$Loss_1(p_{0.2}, \nu_2, r_i)$	1.945	2.001	2.057	1.956	2.027	2.098	1.966	2.054	2.141
$Loss_1(p_{0.5}, \nu_2, r_i)$	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
$Loss_1(p_{0.8}, \nu_2, r_i)$	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
$Loss_1(q, \nu_2, r_i)$	2.800	4.000	5.200	3.200	5.000	6.800	3.943	6.867	9.771

Table 5.1: Comparisons of active learning and passive learning results.

5.1.2 Results on Finite Samples

The steps of the experiments are as follows.

1. Randomly choose $n = 30$ points from $\chi = (-\infty, +\infty)$ with design density $p(x)$ ($p(x) = q(x)$ for passive learning), denote them as $\{x_i\}_{i=1,2,\ldots,n}$. They form the inputs of the training sample.

2. Find out the corresponding outputs of the inputs in Step 1, so that we have a training sample $\{x_i, y_i\}_{i=1,2,...,n.}$.

3. Use the training sample in Step 2 to compute the estimated parameter vector $\hat{\boldsymbol{\theta}}$, using the WLS method introduced in Chapter 2.

4. Repeat Step 1 to Step 3 for $L = 100$ times, then we have $L = 100$ estimates of the parameter vector, denote them as $\{\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\}_{i=1,2,...,L.}$.

5. Compute the estimated IMSE, variance and the square of the bias of $\hat{y}$, make comparisons of the results when ν_1, ν_2, $r_i(x)$ are different, and make comparisons of active learning and passive learning results.

In Step 5, we mention the IMSE, variance and the square of the bias of $\hat{y}$, now we give definitions of them. In (2.8), since the second term is not affected by $p(x)$, we concentrate on the leading term and define the loss function to be

$$\begin{aligned}
\widehat{IMSE} &= \int_{\chi} \boldsymbol{f}'(x)\widehat{MSE_{\epsilon}\big[\hat{\boldsymbol{\theta}}\big]}\,\boldsymbol{f}(x)q(x)dx \\
&= \int_{\chi} \boldsymbol{f}'(x)\frac{1}{L}\sum_{i=1}^{L}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})'\boldsymbol{f}(x)q(x)dx \\
&= \frac{1}{L}\sum_{i=1}^{L}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})'\left(\int_{\chi} \boldsymbol{f}(x)\boldsymbol{f}'(x)q(x)dx\right)(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) \\
&= \frac{1}{L}\sum_{i=1}^{L}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})'\boldsymbol{U}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) \\
&= \frac{1}{L}\sum_{i=1}^{L}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})'(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) \;\; by \text{ (5.1).}
\end{aligned}$$

Similarly, we define estimated variance and squared bias to be

$$\begin{aligned}\widehat{VAR} &= \int_{\chi} \boldsymbol{f}'(x)\widehat{COV_{\epsilon}\big[\hat{\boldsymbol{\theta}}\big]}\boldsymbol{f}(x)q(x)dx \\ &= \int_{\chi} \boldsymbol{f}'(x)\frac{1}{L}\sum_{i=1}^{L}(\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}})'\boldsymbol{f}(x)q(x)dx \\ &= \frac{1}{L}\sum_{i=1}^{L}(\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}),\end{aligned}$$

$$\begin{aligned}\widehat{SQB} &= \int_{\chi} \boldsymbol{f}'(x)(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})'\boldsymbol{f}(x)q(x)dx \\ &= (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}),\end{aligned}$$

where $\bar{\boldsymbol{\theta}} = \frac{1}{L}\sum_{i=1}^{L}\boldsymbol{\theta_i}$.

The experiment results are in Figure 5.2 and Tables 5.2, 5.3 and 5.4. In Figure 5.2, we show plots of different design densities, and the locations of their corresponding sample input points if the sample size $n = 20$. In the experiment, the sample points are taken randomly with the design density, here in the plots we deliberately take sample points to be the 1st to 20th 21-quantile of the design density, because these locations are the expected locations of the random sample points, so they are most representative.

In Tables 5.2, 5.3 and 5.4, the estimated IMSEs of the passive learning method are always greater than those of the active learning methods, and their differences are apparent, since most differences are greater than three times of their standard deviations. Then look at the variance and bias part separately. It is clear that active learning shows advantage in reducing the variance. When $\nu_2 = 0.2$, their differences in the bias part are small, but when ν_2 increases, which means the magnitude of model error increases (since $\sigma_\epsilon = 1$ in our experiments), the advantage of active learning in reducing the bias part becomes apparent, too.

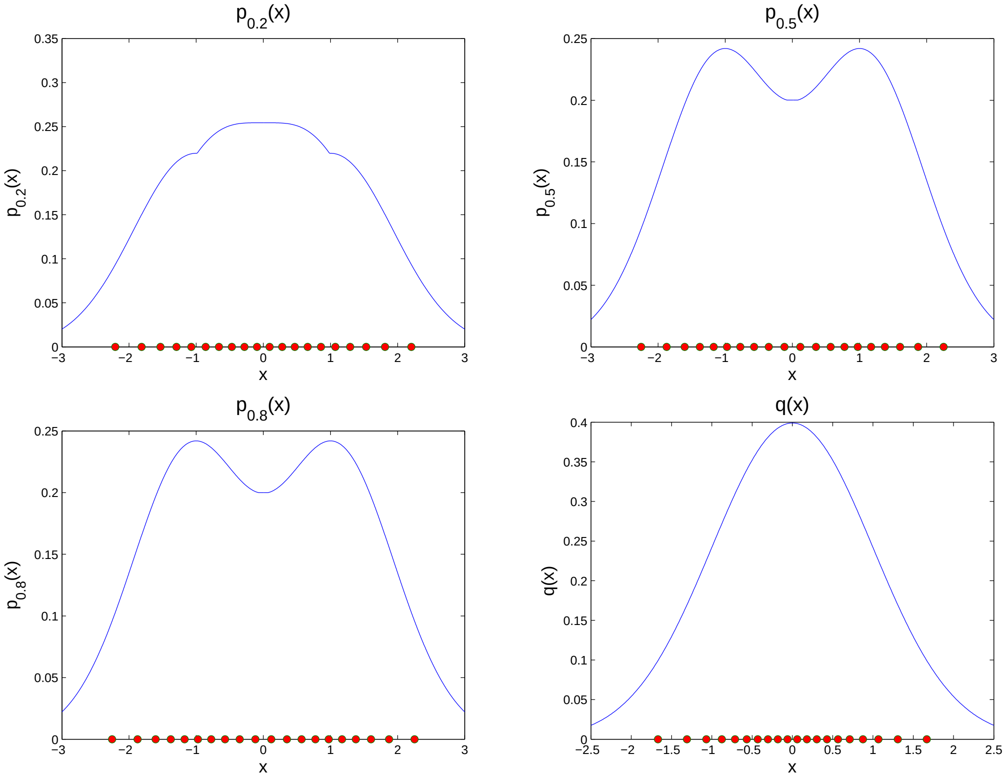

Figure 5.2: Plots of different design densities with $n = 20$ sample input points on the x-axis.

5.2 Simulation Results of Chapter 3

5.2.1 The Asymptotic Results

In Chapter 3, the loss function, Max MSE, is the function in (3.4):

$$\int_{-1}^{1} \left[\sigma_\epsilon^2 + \psi^2(x)\right]\left[1 + \frac{1}{\sigma_q^4}x^2\right]\frac{q^2(x)}{p(x)}dx$$
$$= \left(\sigma_\epsilon^2 + \tau^2\right)\int_{-1}^{1}\left(1 - \nu + \nu r^2(x)\right)\left(1 + \frac{1}{\sigma_q^4}x^2\right)\frac{q^2(x)}{p(x)}dx.$$

Table 5.2: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{IMSE}$ when $\nu_2 = 0.2$ (standard deviations of $\widehat{IMSE}$ in parentheses) (Section 5.1.2)

		r_1			r_2			r_3	
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$
$p_{0.2}$	0.098	0	0.098 (0.0108)	0.096	0.003	0.100 (0.0103)	0.077	0	0.077 (0.0090)
$p_{0.5}$	o.095	0.003	0.098 (0.0107)	0.078	0	0.078 (0.0074)	0.094	0	0.095 (0.0111)
$p_{0.8}$	0.076	0	0.076 (0.0084)	0.079	0	0.079 (0.0068)	0.094	0	0.094 (0.0104)
q	0.124	0.004	0.128 (0.0154)	0.119	0.002	0.122 (0.0142)	0.121	0	0.121 (0.0177)

Table 5.3: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{IMSE}$ when $\nu_2 = 0.5$ (standard deviations of $\widehat{IMSE}$ in parentheses) (Section 5.1.2)

		r_1			r_2			r_3	
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$
$p_{0.2}$	0.141	0	0.142 (0.0136)	0.136	0.002	0.138 (0.0149)	0.140	0.002	0.143 (0.0161)
$p_{0.5}$	0.159	0	0.160 (0.0170)	0.113	0	0.113 (0.0105)	0.158	0	0.158 (0.0152)
$p_{0.8}$	0.137	0	0.137 (0.0138)	0.135	0.001	0.136 (0.0148)	0.156	0	0.156 (0.0104)
q	0.264	0.003	0.267 (0.0379)	0.217	0.013	0.231 (0.0293)	0.259	0.004	0.263 (0.0520)

Table 5.4: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{IMSE}$ when $\nu_2 = 0.8$ (standard deviations of $\widehat{IMSE}$ in parentheses) (Section 5.1.2)

		r_1			r_2			r_3	
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$
$p_{0.2}$	0.310	0	0.310 (0.0319)	0.383	0.002	0.385 (0.0398)	0.358	0.005	0.363 (0.0542)
$p_{0.5}$	0.425	0.002	0.160 (0.0390)	0.350	0	0.350 (0.0447)	0.383	0.001	0.385 (0.0510)
$p_{0.8}$	0.368	0	0.137 (0.0402)	0.325	0.002	0.327 (0.0323)	0.353	0.002	0.355 (0.0524)
q	0.767	0.022	0.789 (0.1065)	0.757	0.045	0.802 (0.1105)	0.779	0.016	0.795 (0.1906)

Therefore, we choose loss function in this section to be:

$$Loss_2(p_{\nu_1}, \nu_2, r_i) = \int_{-1}^{1} \Big(1 - \nu_2 + \nu_2 r_i^2(x)\Big)\Big(1 + \frac{1}{\sigma_q^4} x^2\Big) \frac{q^2(x)}{p_{\nu_1}(x)} dx.$$

Then we adopt the assumptions in Chapter 3, compute and compare simulation results of active learning versus passive learning. Figure 5.3 shows the design density for active learning and passive learning. Table 5.5 compares the loss of active learning with that of passive learning, it proves the advantage of active learning over passive learning in these cases.

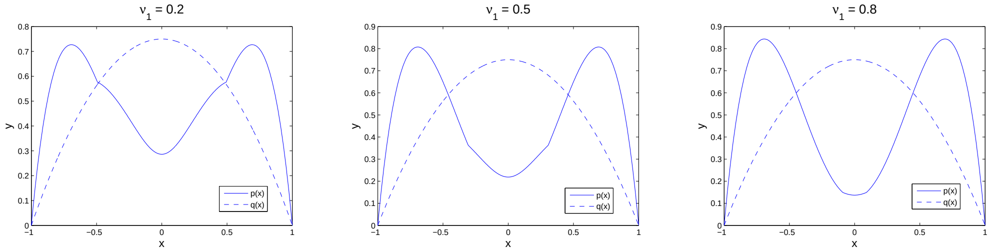

Figure 5.3: Plots of p(x) and q(x) when $\nu_1 = 0.2,\ 0.5,\ 0.8$ respectively.

		r_1			r_2			r_3	
ν_2	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
$Loss_2(p_{0.2}, \nu_2, \psi_i)$	5.270	5.256	5.409	5.492	5.458	5.841	5.715	5.66	6.273
$Loss_2(p_{0.5}, \nu_2, \psi_i)$	5.514	5.581	5.625	5.577	5.744	5.855	5.640	5.907	6.085
$Loss_2(p_{0.8}, \nu_2, \psi_i)$	5.963	5.971	5.971	5.965	5.984	5.985	5.967	5.998	5.999
$Loss_2(q, \nu_2, \psi_i)$	7.333	6.198	7.188	9.333	6.496	8.971	11.333	6.793	10.753

Table 5.5: Comparisons of active learning and passive learning results.

5.2.2 Results on Finite Samples

The steps of experiment in this section are similar to those in Section 5.1.2, but the loss function changes. By (3.1), we define loss function in this section to be

$$\widehat{MaxMSE} = \max_t \boldsymbol{f}'(t)\frac{1}{L}\sum_{i=1}^{L}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})'\boldsymbol{f}(t). \tag{5.3}$$

Denote the maximizer t in (5.3) as t_0, then we define the variance and squared bias part to be

$$\widehat{VAR} = f'(t_0)\frac{1}{L}\sum_{i=1}^{L}(\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}})'\boldsymbol{f}(t_0),$$

$$\widehat{SQB} = f'(t_0)(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})'\boldsymbol{f}(t_0).$$

Similarly with Section 5.1.2, results are shown in Figure 5.4, Tables 5.6, 5.7 and 5.8. In this part, the losses of the passive learning are still greater than the those of active learning, but in some cases their differences are not big, only about one standard deviation. But this part shows that our active learning densities greatly reduce the bias, especially when $\nu_2 = 0.8$.

Table 5.6: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{MaxMSE}$ when $\nu_2 = 0.2$ (standard deviations of $\widehat{MaxMSE}$ in parentheses)

		r_1			r_2			r_3	
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$
$p_{0.2}$	0.259	0	0.259 (0.0348)	0.210	0.028	0.238 (0.0123)	0.247	0.002	0.249 (0.0402)
$p_{0.5}$	0.202	0.010	0.213 (0.0357)	0.272	0.057	0.330 (0.0507)	0.207	0	0.207 (0.0322)
$p_{0.8}$	0.262	0.024	0.285 (0.0519)	0.277	0.008	0.285 (0.0403)	0.240	0.002	0.243 (0.0386)
q	0.301	0.022	0.323 (0.0473)	0.253	0.268	0.521 (0.0691)	0.339	0.011	0.350 (0.0570)

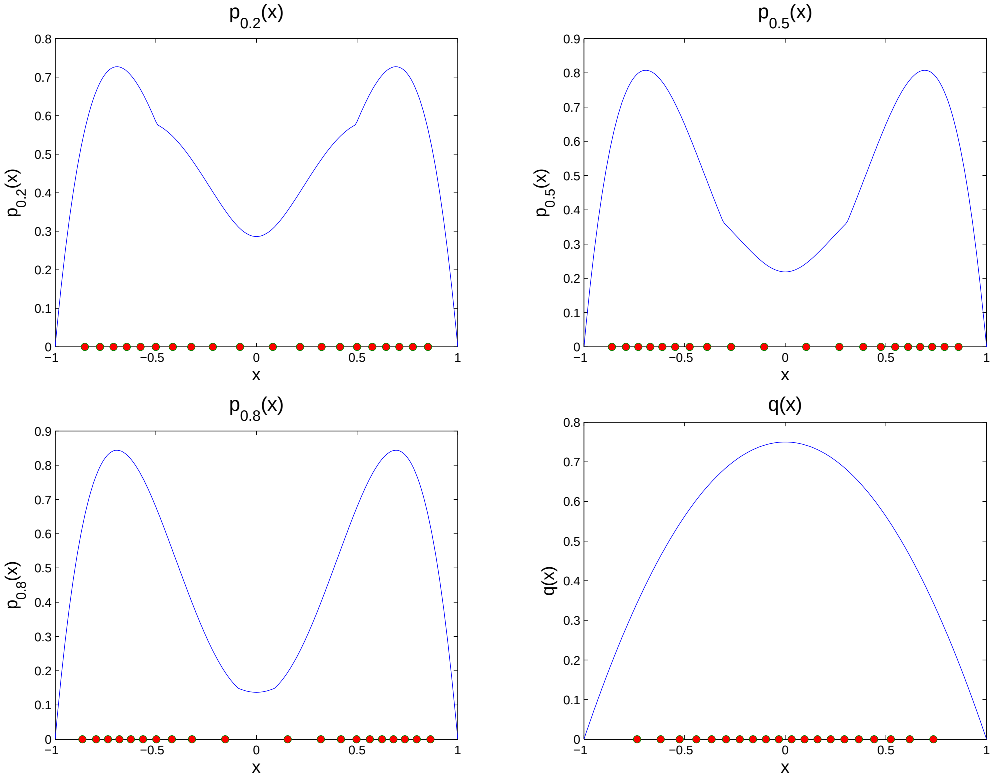

Figure 5.4: Plots of different design densities with $n = 20$ sample points on the x-axis.

Table 5.7: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{MaxMSE}$ when $\nu_2 = 0.5$ (standard deviations of $\widehat{MaxMSE}$ in parentheses)

		r_1			r_2			r_3	
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$
$p_{0.2}$	0.291	0.013	0.304 (0.0451)	0.338	0.102	0.440 (0.0645)	0.283	0.024	0.308 (0.0505)
$p_{0.5}$	0.382	0	0.382 (0.0519)	0.359	0.068	0.427 (0.0671)	0.288	0.009	0.297 (0.0447)
$p_{0.8}$	0.423	0.004	0.428 (0.0332)	0.395	0.095	0.490 (0.0784)	0.333	0.021	0.354 (0.0615)
q	0.510	0.042	0.552 (0.0546)	0.459	1.126	1.585 (0.1782)	0.439	0.064	0.503 (0.0711)

Table 5.8: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{MaxMSE}$ when $\nu_2 = 0.8$ (standard deviations of $\widehat{MaxMSE}$ in parentheses)

	r_1			r_2			r_3		
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{MaxMSE}$
$p_{0.2}$	0.771	0.046	0.817 (0.0997)	0.958	0.568	1.525 (0.2258)	0.628	0.091	0.720 (0.1173)
$p_{0.5}$	1.012	0.003	1.014 (0.1590)	0.758	0.265	1.023 (0.1247)	0.727	0.044	0.771 (0.1135)
$p_{0.8}$	1.000	0.079	1.079 (0.1722)	0.919	0.393	1.312 (0.1474)	0.714	0.085	0.799 (0.1223)
q	1.570	0.228	1.798 (0.2445)	1.150	4.531	5.680 (0.5220)	0.645	0.246	0.891 (0.1216)

5.3 Simulation Results of Chapter 4

5.3.1 The Asymptotic Results

The loss function in Chapter 4 is (4.3). Because with our choices of $\psi(x)$, $\int_\chi n\psi^2(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = \tau^2$ always, so we only concentrate on the leading term of (4.3). Because

$$\begin{aligned}&\sigma_\epsilon^2 tr[\boldsymbol{U}\boldsymbol{V}_p^{-1}] + \boldsymbol{b}'_{\sqrt{n}\psi,p}\boldsymbol{V}_p^{-1}\boldsymbol{U}\boldsymbol{V}_p^{-1}\boldsymbol{b}_{\sqrt{n}\psi,p}\\ &= \Big(\sigma_\epsilon^2+\tau^2\Big)\Big((1-\nu)tr[\boldsymbol{U}\boldsymbol{V}_p^{-1}] + \nu\boldsymbol{b}'_{\sqrt{n}r,p}\boldsymbol{V}_p^{-1}\boldsymbol{U}\boldsymbol{V}_p^{-1}\boldsymbol{b}_{\sqrt{n}r,p}\Big),\end{aligned}$$

we set the loss function to be

$$Loss_3(p_{\nu_1},\nu_2,r_i) = (1-\nu_2)tr[\boldsymbol{U}\boldsymbol{V}_{p_{\nu_1}}^{-1}] + \nu_2\boldsymbol{b}'_{\sqrt{n}r_i,p_{\nu_1}}\boldsymbol{V}_{p_{\nu_1}}^{-1}\boldsymbol{U}\boldsymbol{V}_{p_{\nu_1}}^{-1}\boldsymbol{b}_{\sqrt{n}r_i,p_{\nu_1}}.$$

With assumptions in Chapter 4, we have

$$\boldsymbol{U} = \begin{bmatrix} 1 & 0 \\ 0 & \sigma_q^2 \end{bmatrix},$$

$$\boldsymbol{V}_p = \begin{bmatrix} 1 & 0 \\ 0 & \sigma_p^2 \end{bmatrix},$$

$$tr[\boldsymbol{U}\boldsymbol{V}_p^{-1}] = 1 + \frac{\sigma_q^2}{\sigma_p^2},$$

$$\boldsymbol{V}_p^{-1}\boldsymbol{U}\boldsymbol{V}_p^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{\sigma_q^2}{\sigma_p^4} \end{bmatrix},$$

$$\boldsymbol{b}_{\sqrt{n}r,p} = \begin{bmatrix} \int_{-\infty}^{\infty} \sqrt{n}r(x)p(x)dx \\ \int_{-\infty}^{\infty} x\sqrt{n}r(x)p(x)dx \end{bmatrix}.$$

Since the $\psi(x)$ we choose are either odd functions or even functions, thus specifically, (1) When $\psi(x)$ is an even function, the second element in $\boldsymbol{b}_{\sqrt{n}\psi,p}$ is 0, so

$$\boldsymbol{b}'_{\sqrt{n}r,p}\boldsymbol{V}_p^{-1}\boldsymbol{U}\boldsymbol{V}_p^{-1}\boldsymbol{b}_{\sqrt{n}r,p} = \left(\int_{-\infty}^{\infty} \sqrt{n}r(x)p(x)dx\right)^2.$$

In this case,

$$Loss_3(p_{\nu_1}, \nu_2, r_i) = (1-\nu_2)\left(1 + \frac{\sigma_q^2}{\sigma_{p_{\nu_1}}^2}\right) + \nu_2\left(\int_{-\infty}^{\infty} \sqrt{n}r_i(x)p_{\nu_1}(x)dx\right)^2.$$

(2) When $\psi(x)$ is an odd function, the first element in $\boldsymbol{b}_{\sqrt{n}\psi,p}$ is 0, so

$$\boldsymbol{b}'_{\sqrt{n}r,p}\boldsymbol{V}_p^{-1}\boldsymbol{U}\boldsymbol{V}_p^{-1}\boldsymbol{b}_{\sqrt{n}r,p} = \frac{\sigma_q^2}{\sigma_p^4}\cdot\left(\int_{-\infty}^{\infty} x\sqrt{n}r(x)p(x)dx\right)^2.$$

In this case,

$$Loss_3(p_{\nu_1}, \nu_2, r_i) = (1-\nu_2)\left(1 + \frac{\sigma_q^2}{\sigma_{p_{\nu_1}}^2}\right) + \nu_2 \cdot \frac{\sigma_q^2}{\sigma_{p_{\nu_1}}^4} \cdot \left(\int_{-\infty}^{\infty} x\sqrt{n}r_i(x)p_{\nu_1}(x)dx\right)^2.$$

Similarly with Section 5.1 and 5.2, we compute $Loss_3(p_{\nu_1}, \nu_2, r_i)$ and compared results of active learning and passive learning. Figure 5.5 shows plots of design density of active learning and passive learning, Table 5.9 shows the comparisons of the loss. We can see that when $\nu_2 = 0.8$ and ν_1 is different from ν_2, most of the time the loss of active learning is a little higher than that of passive learning. But in other cases active learning has lower loss than passive learning.

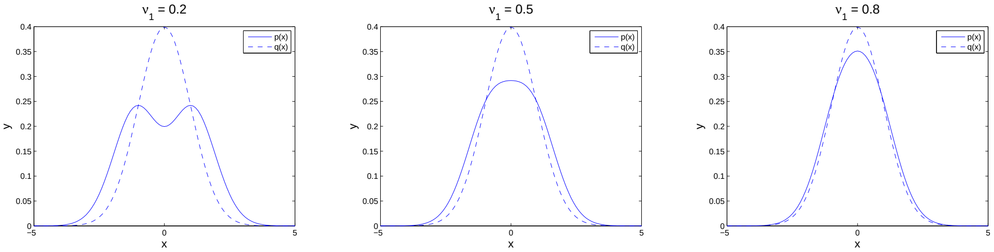

Figure 5.5: Plots of p(x) and q(x) when $\nu_1 = 0.2,\ 0.5,\ 0.8$ respectively.

		ψ_1			ψ_2			ψ_3	
ν_2	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
$Loss_3(p_{0.2}, \nu_2, \psi_i)$	1.300	1.000	0.700	1.275	0.938	0.600	1.243	0.857	0.471
$Loss_3(p_{0.5}, \nu_2, \psi_i)$	1.366	0.895	0.425	1.366	0.895	0.424	1.351	0.857	0.363
$Loss_3(p_{0.8}, \nu_2, \psi_i)$	1.486	0.934	0.383	1.486	0.934	0.382	1.484	0.928	0.373
$Loss_2(q, \nu_2, \psi_i)$	1.600	1.000	0.400	1.600	1.000	0.400	1.600	1.000	0.400

Table 5.9: Comparisons of active learning and passive learning results.

5.3.2 Results on Finite Samples

In this section, the loss function is the same with that in Section 5.1.2, the experiments results are shown below. Remember that in this problem, we bound the magnitude of model error to be $O(\frac{1}{n})$, and thus ignore the bias part, so from the tables below we can see that the biases are all very small. Also, although the losses in passive learning are still greater than those of active learning in most cases, they are pretty close, in some cases when $\nu_2 = 0.8$, passive learning even has smaller loss. This might be because the magnitude of the model error is too small.

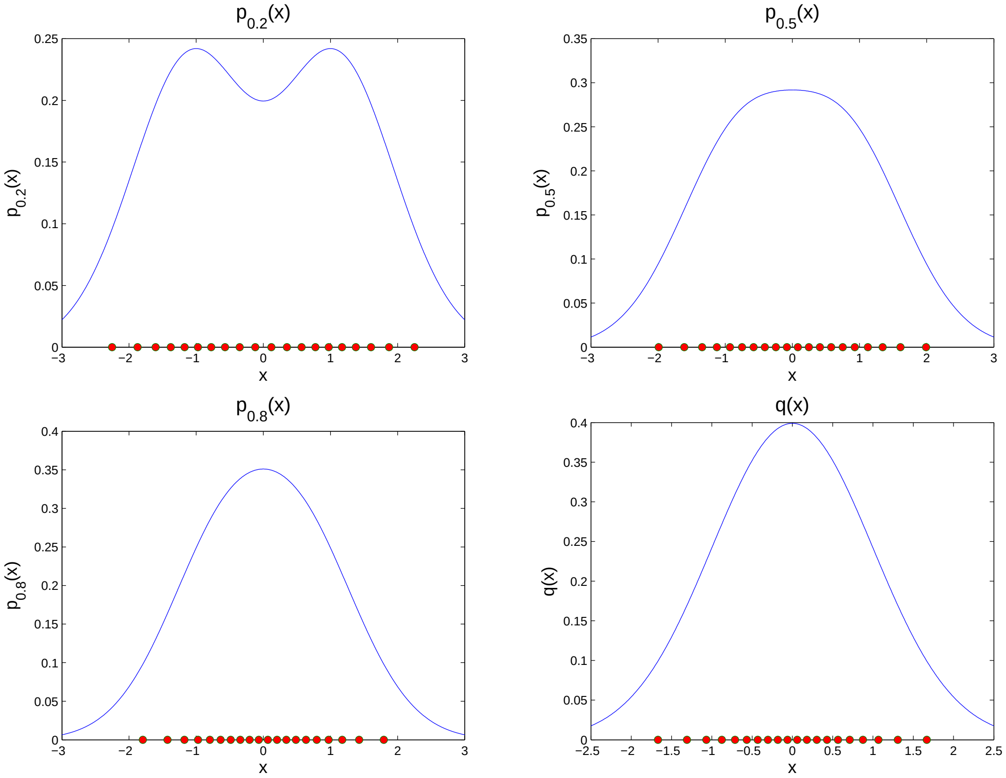

Figure 5.6: Plots of different design densities with $n = 20$ sample points on the x-axis.

In conclusion of Sections 5.1, 5.2 and 5.3, the results prove that, under our model assumptions, active learning is more advantageous than passive learning. Also, it proves the effectiveness of our active learning methods, not only in reducing variance but also in reducing bias of the estimates.

Table 5.10: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{IMSE}$ when $\nu_2 = 0.2$ (standard deviations of $\widehat{IMSE}$ in parentheses)

		r_1			r_2			r_3	
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$
$p_{0.2}$	0.061	0.005	0.065 (0.0067)	0.050	0.006	0.056 (0.0074)	0.061	0	0.062 (0.0077)
$p_{0.5}$	0.069	0	0.070 (0.0076)	0.061	0.001	0.062 (0.0058)	0.062	0	0.063 (0.0069)
$p_{0.8}$	0.062	0	0.062 (0.0063)	0.064	0	0.064 (0.0059)	0.076	0.001	0.077 (0.0086)
q	0.075	0	0.076 (0.0071)	0.077	0	0.077 (0.0080)	0.080	0	0.080 (0.0107)

Table 5.11: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{IMSE}$ when $\nu_2 = 0.5$ (standard deviations of $\widehat{IMSE}$ in parentheses)

		r_1			r_2			r_3	
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$
$p_{0.2}$	0.063	0.014	0.076 (0.0072)	0.069	0.011	0.080 (0.0077)	0.079	0.003	0.082 (0.0094)
$p_{0.5}$	0.060	0.003	0.063 (0.0053)	0.072	0.003	0.075 (0.0074)	0.091	0.002	0.093 (0.0093)
$p_{0.8}$	0.071	0	0.072 (0.0074)	0.072	0	0.072 (0.0087)	0.072	0	0.072 (0.0076)
q	0.085	0	0.085 (0.0104)	0.091	0	0.094 (0.0111)	0.085	0	0.086 (0.0101)

Table 5.12: Comparative values of $\widehat{VAR}$, $\widehat{SQB}$ and $\widehat{IMSE}$ when $\nu_2 = 0.8$ (standard deviations of $\widehat{IMSE}$ in parentheses)

		r_1			r_2			r_3	
p_{ν_1}	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$	$\widehat{VAR}$	$\widehat{SQB}$	$\widehat{IMSE}$
$p_{0.2}$	0.072	0.040	0.112 (0.0131)	0.090	0.040	0.131 (0.0123)	0.128	0.019	0.148 (0.0200)
$p_{0.5}$	0.087	0.014	0.101 (0.0094)	0.086	0.013	0.099 (0.0107)	0.135	0.006	0.140 (0.0195)
$p_{0.8}$	0.110	0	0.110 (0.0100)	0.082	0.002	0.084 (0.0088)	0.096	0	0.096 (0.0128)
q	0.096	0	0.096 (0.0107)	0.108	0.002	0.108 (0.0129)	0.100	0	0.101 (0.0108)

Chapter 6

Conclusion

This thesis is inspired by the possible advantages of active learning over passive learning. The research object of this thesis is the active learning problem in regression field. There are different kinds of methods to solve this kind of problem, such as uncertainty sampling method. The kind of methods we use are optimal experimental design methods, which are widely used in solving active learning problem in regression field. But we have improved the traditional experimental design methods, mainly in that we try to reduce the mean squared error of the estimation, which include both the variance and the bias caused by the model error, while traditional methods just focus on variance reduction.

The assumptions of our research are as follows.

(1) The underlying distribution density of the design space – $q(\boldsymbol{x})$, is known. This $q(\boldsymbol{x})$ is considered to be the density of the testing set of the supervised learning problem. In the cancer diagnosis example we talk about in Section 1.1, this means that the distribution of all the X-ray pictures are known.

(2) The methods to use the sample to estimate the unknown parameters in

the regression model are clear. We have chosen two methods – WLS method with weight $w(\boldsymbol{x}) = q(\boldsymbol{x})/p(\boldsymbol{x})$ and OLS method.

We know that "how well the estimations of the unknown parameters are" is related not only to the estimation methods, but also to how good the sample is. In assumptions (2) we have made clear the estimation methods, so the goals of our research are as follows.

(1) How to find the best sample. (The "best" here means that this sample will result in the "best estimation" of the unknown parameters, and the measurement of the "best estimation" is defined by the loss functions.) That is, if the sample (the training set) is chosen at random with a density $p(\boldsymbol{x})$ (we call it design density) from the design space, we aim to find the best density $p(\boldsymbol{x})$. In the cancer diagnosis example, if the researcher randomly choose n X-ray pictures with a density $p(\boldsymbol{x})$ to be the sample, the aim is to find the best $p(\boldsymbol{x})$.

(2) Compare how good the estimation is when the design density is the best $p(\boldsymbol{x})$ we find and when the design density is $q(\boldsymbol{x})$, that is to compare active learning and passive learning results, and see if active learning truly has advantage over passive learning.

The contributions of this thesis are:

(1) We outline three forms of active learning problems in regression fields and find their solutions. However, our solutions have restrictions, they only apply to models with the same assumptions with ours (for the exact assumptions see Sections 2.4.2, 3.2 and 4.3).

(2) By simulation we prove that our active learning solutions to the three problems are better than passive learning. But again, this statement is only suitable for models with our settings.

(3) Although our solutions are restricted to specific models, in the process

of solving for the best design density, we show an example of how to take not only the variance, but also the bias of the estimation, which caused by the model error, into consideration. This is our major improvement compared with Sugiyama's work (Sugiyama, 2006), in which the bias is ignored by bounding the magnitude of the error.

References

Arfken, G. B., and Weber, H. J. (2011). *Mathematical methods for physicists: a comprehensive guide*. Academic press.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of artificial intelligence research*.

Daemi, M., and Wiens, D. P. (2013). Techniques for the construction of robust regression designs. *Canadian Journal of Statistics*, *41*(4), 679–695.

Fedorov, V. V. (1972). *Theory of optimal experiments*. Elsevier.

Heo, G., Schmuland, B., and Wiens, D. P. (2001). Restricted minimax robust designs for misspecified regression models. *Canadian Journal of Statistics*, *29*(1), 117–128.

Kanamori, T., and Shimodaira, H. (2003). Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference*, *116*(1), 149–162.

Kapoor, A., and Greiner, R. (2005). *Learning and classifying under hard budgets*. Springer.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 272–319.

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing.

Pukelsheim, F. (1993). *Optimal design of experiments* (Vol. 50). siam.

Scheffer, T., Decomain, C., and Wrobel, S. (2001). Active hidden markov models for information extraction. In *Advances in intelligent data analysis* (pp. 309–318). Springer.

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Sugiyama, M. (2006). Active learning in approximately linear regression based on conditional expectation of generalization error. *The Journal of Machine Learning Research*, *7*, 141–166.

Tong, S., and Koller, D. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, *2*, 45–66.

Tur, G., Hakkani-Tür, D., and Schapire, R. E. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, *45*(2), 171–186.

Wen, J., Yu, C.-N., and Greiner, R. (2014). Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *Proceedings of the 31st international conference on machine learning (icml-14)* (pp. 631–639).

Wiens, D. P. (1992). Minimax designs for approximately linear regression. *Journal of Statistical Planning and Inference*, *31*(3), 353–371.