ABSTRACT

Title of dissertation: ESSAYS IN STATISTICAL ANALYSIS: ISOTONIC REGRESSION AND FILTERING

Jinhang Xue
Doctor of Philosophy, 2018

Dissertation directed by: Professor Ilya O. Ryzhov
Department of Decision, Operations,
and Information Technologies

In many real-world applications in optimal information collection and stochastic approximation, statistical estimators are often constructed to learn the true parameter value of some utility functions or underlying signals. Many of these estimators exhibit excellent empirical performance, but full analyses of their consistency are not previously available, thus putting decision-makers in somewhat of a predicament regarding implementation. The goal of this dissertation is to fill this blank of missing consistency proofs.

The first part of this thesis considers the consistency of estimating a monotonic cost function which appears in an optimal learning algorithm that incorporates isotonic regression with a Bayesian policy known as Knowledge Gradient with Discrete Priors (KGDP). Isotonic regression deals with regression problems under order constraints. Previous literature proposed to estimate the cost function by a weighted sum of a pool of candidate curves, each of which is generated by the isotonic regression estimator based on all the previous observations that have been

collected, and the weights are calculated by KGDP. Our primary objective is to establish the consistency of the suggested estimator. Some minor results, regarding with the knowledge gradient algorithm and the isotonic regression estimator under insufficient observations, are also discussed.

The second part of this thesis focuses on the convergence of the bias-adjusted Kalman filter (BAKF). The BAKF algorithm is designed to optimize the statistical estimation of a non-stationary signal that can only be observed with stochastic noise. The algorithm has numerous applications in dynamic programming and signal processing. However, a consistency analysis of the process that approximates the underlying signal has heretofore not been available. We resolve this open issue by showing that the BAKF stepsize satisfies the well-known conditions on almost sure convergence of a stochastic approximation sequence, with only one additional assumption on the convergence rate of the signal compared to those used in the derivation of the original problem.

ESSAYS IN STATISTICAL ANALYSIS: ISOTONIC REGRESSION AND FILTERING

by

Jinhang Xue

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Paul J. Smith, Chair
Professor Ilya O. Ryzhov, Co-Chair
Professor Xin He
Professor Leonid Koralov
Professor Kunpeng Zhang

Dedication

To my family.

Acknowledgments

First and foremost, I owe my deepest gratitude to my advisor, Professor Ilya O. Ryzhov, for his inspiring insights, continuous support and patient guidance throughout my research. The completion of this dissertation would not have been possible without his advisory. It is my great honor to have the opportunity to work with Professor Ryzhov, his enthusiasm, expertise and rigorousness make me a much better researcher and will always influence me in my future career.

I would also like to express my most sincere appreciation to my co-advisor, Professor Paul J. Smith, for the detailed revision to my thesis and his mentorship throughout my years of graduate study. The linear regression course he taught has been very important to my research and internship, and his humorous teaching style enlightened me to improve the interaction with my own students as a teaching assistant.

I am truly grateful to my other committee members, Professor Leonid Koralov, Professor Xin He and Professor Kunpeng Zhang, for taking time to read my thesis and attend my defense. Special thanks to Professor Abram M. Kagan for helping me with the qualification exams, and to Professor Yuan Liao for attending my preliminary candidacy exam and being a good friend.

I really appreciate two other fellow students of Professor Ryzhov, Dr. Ye Chen and Dr. Cheng Jie, for their time in helping me with specific research questions and to have very meaningful discussions with me. I have special thanks to Dr. Sikai Qu for his insights and collaboration in many of our course projects, and to Dr. Xia Li

for her referral on my summer internship.

I thank my fellow soccer teammates, most of which are also doctoral students, for the brotherhood we developed, and for sharing their own academic research that broaden my vision on the application of statistical analysis in many other disciplines.

I would like to thank my dear friends Zhiwei Tan, Bonan Ren, Chen Qian, Luyu Sun, Dongmian Zou, Zhang Zhang, Yiran Li and Huili Liu for the wonderful years we spend together. Additionally I would like to express my appreciation to all the faculty and staff of the department of mathematics in general for their great support and service.

Last but very importantly, I owe my deepest gratitude to my family for their unconditional love and support, without which none of my achievements and research abroad could ever be possible. My deepest condolences reach to my maternal grandfather and grandmother, whose last days I wish I could stand by their side. I could not appreciate more to my father Lian Xue and my mother Jun Wang who always believe in my potentials. Especially, my mother always encourages me with her own experience of Ph.D. research whenever I am puzzled or stressed. Their love to me are beyond words. My tremendous thankfulness to my loving and supportive girlfriend Linxi Gao, for her invigoration and understanding during the past two years of our long-distance. I am also very grateful to my aunt Zheng Chang and uncle Zhiyan Hu for coming overseas to see me, and bring cares and greetings from other family members that have been very supportive to me since my childhood.

Contents

List of Tables

List of Figures

List of Abbreviations

BAKF	Bias-Adjusted Kalman Filter
CSD	Cumulative Sum Diagram
DP	Discrete Priors
DP-R	Discrete Priors with Resampling
DP-R&R	Discrete Priors with Resampling and Regeneration
GCM	Greatest Convex minorant
KGDP	Knowledge Gradient with Discrete Priors
KGDP-R	KGDP with Resampling
KGDP-R&R	KGDP with Resampling and Regeneration
SA	Stochastic Approximation
SLLN	Strong Law of Large Numbers

Chapter 1: Introduction

In this chapter, we will review the development of regression analysis under order restrictions, namely, isotonic regression, with concepts and theoretical results that applied in this thesis. We then introduce one class of Bayesian methods, called discrete priors, and compare it with previously established algorithms in the literature. We shows how this Bayesian framework is incorporated with isotonic regression to construct our weighted sum estimator. An application in transportation science is presented at the end as an illustration to the motivation of our work.

1.1 Literature Review

In the application of regression analysis, issues emerge very often where order constraints have to be exerted on the regression function. Such challenges draw a great attention in literature for the recent decades. We begin with the history of isotonic regression analysis and its interaction with Bayesian methodology.

Isotonic Regression

The usual goal of regression analysis is to estimate the conditional expectation of a dependent variable $\tilde{y}$ given some independent variables $\tilde{x}$. The function

$\mu(x) = \mathbb{E}(\tilde{y}|\tilde{x} = x)$ of x, as the regression of $\tilde{y}$ on $\tilde{x}$, often furnishes the best fit to the distribution of $\tilde{y}$ in the sense of least squares. In many applications, we are able to assert a priori that the dependent response has a monotonic relation with one or more predictors. That is, there are some prespecified structure or order constraints imposed on the response. Structure involving orderings and inequalities is often useful since it is easy to interpret, understand, and explain. For example, in epidemiological studies, it is often of the researcher's interest to evaluate the relationship between the dosage of a particular toxic exposure and the likelihood of an adverse response, controlling for confounding factors. Considering potential regulatory provisions and identifying public health significance, it is critical to efficiently estimate the dose-response function. In such studies, one may typically presume a priori that the occurrence of an adverse response would not be less likely as dose increases, adjusting for key confounding factors such as gender and age. It is well documented that incorporating such ordering restrictions can improve the efficiency of statistical inference procedures and power to detect trend in response (Barlow and Scheuer, 1966; Robertson et al., 1988). More real world applications of such order constraints can be found in operations research (Maxwell and Muckstadt, 1985), biology (Lee, 1996), genetics (Gjuvsland et al., 2013), psychology (Kalish et al., 2016), meteorology (Roth et al., 2015), environmental science (Hussian et al., 2005), signal processing (Acton and Bovik, 1998), economics (Aït-Sahalia and Duarte, 2003), sports (Dawson and Magee, 2001) and many others fields.

Isotonic regression (IR), sometimes called monotonic regression (MR), studies regression problems in which knowledge of the predictors in a certain experiment

determines an ordering, partial or total, of the corresponding values of $\mu(x)$. Here, the word "isotonic" means "order-preserving"; order restrictions on parameters require that the parameter to be isotonic with respect to a partial order on the index set. The theory of order-restricted estimation and testing was developed under a variety of scenarios. The first comprehensive monographs are Barlow et al. (1972) and Robertson et al. (1988), which provided all fundamental theoretical results in estimation and hypothesis testing. In modern application, constraints often crop up either as (partial/total, linear/non-linear, or even implicit) *ordering* on the so-called parameter space(s), or as (linear/non-linear) *inequalities* on contemplated parameters; they may also show up on the observations or experimental outcomes termed the *responses*. Most recently, the field of *constrained* statistical inference (Silvapulle and Sen, 2011), specifically focuses on the constraints in the form of ordering and inequality restraints of diverse types, and extends beyond classical likelihood-based parametric models.

1.1.1 Bayesian Approach

We begin our in-depth review with Bayesian related topics as this is the main research direction of our thesis. In order-restricted inference, Bayesian methods seek to incorporate prior information regarding a collection of parameters to improve the quality of the inference. An early application of Bayesian idea was by Kraft and van Eeden (1964), who studied the problem of estimating a nondecreasing set of binomial parameters (bioassay, in their terminology). Barlow et al. (1972) described

an approach to Bayesian estimation for independent samples from members of an exponential family and give a theorem which yields the mode of the posterior distribution as an isotonic regression. Dykstra and Laud (1981) studied a Bayesian nonparametric approach to estimate an increasing hazard rate. Broffitt (1984) considered a Bayesian approach to some order-restricted problems which are motivated by the so-called graduation techniques in actuarial science. Sedransk et al. (1985) discussed the problem of estimating an ordered set of multinomial probabilities using Bayesian methods. Bacchetti (1989) developed additive isotonic models that generalize linear models by replacing lines with nondecreasing transformations. Bayesian ideas received more attention starting with the 1990s. Lavine and Mockus (1995) provided a theoretical study on a nonparametric Bayes method for isotonic regression. Dunson and Neelon (2003) applied Bayesian inference on order-constrained parameters in generalized linear models. Neelon and Dunson (2004) established a new framework for Bayesian isotonic regression and order-restricted inference with a prior formulated as a latent autoregressive normal process. Dunson (2005) proposed a Bayesian semiparametric approach for inference on an unknown isotonic regression function with count data. Cai and Dunson (2007) suggested a Bayesian approach for addressing multivariate isotonic regression splines motivated by applications to carcinogenicity studies. Brezger and Steiner (2008) incorporated Bayesian p-splines with monotonic regression to deal with the issue of estimating price response functions from store-level scanner data. Wang and Dunson (2011) developed a new class of density regression models that incorporate stochastic-ordering constraints and that offer a prior structure which enables full support to the conditional distribu-

tion of the response given the predictors. For a more general overview of Bayesian perspectives, in the broader concept of constrained statistical inference than as in isotonic regression, we refer readers to Chapter 8 of Silvapulle and Sen (2011).

Among these paper, Dunson's work claims that one of the most appealing advantages of applying prior information is the ease of incorporating complex order restrictions on either regression functions or parameters, but this may not always hold. As in frequentist statistics, we obtain an estimator by solving an optimization problem (with order restrictions) in some parameter space. As long as the space is a convex set, it is easy to incorporate additional constraints. However, in Bayesian statistics, we need a (prior) distribution for the parameters we try to estimate, and there are no standard distributions on the space of isotonic functions. Therefore, in such occasions it maybe more difficult to specify a Bayesian prior distribution while leveraging the order restrictions.

Nevertheless, the advantage of Bayesian analysis is that it equips us with a much richer and deeper uncertainty model than in frequentist statistics, where each model parameter is assigned a point estimate. On the contrary, a Bayesian would provide a whole (posterior) distribution over all the values that the parameter can take, yielding a more detailed profile on how likely the parameter is to take on different values. The quality of the Bayesian modeling process is measured by the degree to which a posterior distribution is more informed than a prior distribution for the unknown parameters of interest.Therefore as concluded in Gill (2014, Sec.1.2), with Bayesian analysis, assertions about unknown model parameters are not expressed in the conventional way as single point estimates along with associated reliability as-

sessed through the standard null hypothesis significance test. Instead the emphasis is on making probabilistic statements using prior and posterior distributions. Such characteristics are particularly desirable because when people try to make decisions subject to unknown parameters, they really need this extra uncertainty to help appraise the probability that a decision is either way too sub-optimal or much worse than our expectation. It also helps to assess, under various scenarios, the deficiency of our model performance if indeed there is a large discrepancy between the true parameter value and our belief.

However, the benefits of Bayesian analysis do not come without a price. Many calculations in Bayesian analysis, including posterior densities and simulation algorithms, suffer extremely heavy numerical complexity and require massive computational resources. For instance, algorithms like Markov chain Monte Carlo (MCMC) or Metropolis-Hastings, let us evaluate and calculate the posterior density, but do not allow for efficient updating.

Remark (The Bayes-frequentist controversy).

According to Carlin and Louis (2008), traditionally frequentists evaluate procedures based on imagining repeated sampling from a particular model (the likelihood), which defines the probability distribution of the observed data conditional on unknown parameters. Properties of the procedure are evaluated in this repeated sampling framework for fixed values of unknown parameters; good procedures perform well over a broad range of parameter values. In contrast, Bayesians require a sampling model and, in addition, a prior distribution on all unknown quantities in the

model. The prior and likelihood are used to compute the conditional distribution (the posterior distribution) of the unknowns given the observed data, from which all statistical inferences arise.

Historically, frequentists have criticized Bayesian procedures for their inability to deal with all but the most basic examples, for over-reliance on computationally convenient priors, and for being too fragile in their dependence on a specific prior (i.e., for a lack of robustness in settings where the data and prior conflict). Bayesians have criticized frequentists for failure to incorporate relevant prior information, inefficiency and inflexibility. Another common Bayesian criticism is that, while frequentist methods do avoid dependence on any single set of prior beliefs, the resulting claims of "objectivity" are often illusory since such methods still require assumptions about the underlying data generating mechanism, such as a simple (often normal) model free from confounding, selection bias, measurement error, etc. Bayesians often remark that the choice of prior distribution is only one assumption that should be explicitly declared and checked in a statistical analysis. Importantly, the Bayesian formalism propagates uncertainty through the analysis enabling a more detailed assessment of the variability in estimated quantities of interest.

1.1.2 Discrete Priors

We now review some Bayesian methods that have been recently developed and are relevant to our work. Chen et al. (2015) proposed a framework based on Discrete Priors (DP) for searching the true function which, for us, is isotonic among finitely

many alternative functions (or "curves"). Discrete priors make Bayesian inference and updating to be more concise and computationally efficient, and allow us to handle more complicated objects (like isotonic functions or any other functions). While other work usually imposes a Bayesian model directly on the isotonic function, in the DP approach, the isotonic functions are deterministic and the Bayesian prior distribution simply randomizes over functions in the so-called *candidate set.*[1] The candidate set includes finitely many isotonic functions, one of which must be the true function. One major limitation of using such a finite candidate set is the risk of inconsistent estimation if in fact none of the candidate curves are near the true curve. To handle this issue, He and Powell (2016) and He (2017) proposed Discrete Priors with Resampling (DP-R). They construct a so-called *pool of curves,* which is a set containing considerably more alternative functions then the candidate set. Then, the curves in the candidate set are updated and replaced by resampling from the pool of curves. Nevertheless, the DP-R still requires that the true function is either contained in the pool of curves or close to one of its alternatives. As this could still be unrealistic for many real world settings, Huang et al. (2018) further relax such assumption by proposing the so-called Discrete Priors with Resampling and Regeneration (DP-R&R) as a modification to DP-R. They update the functions in the pool of curves with a regeneration process which brings better alternative functions (in the sense of mean squared error defined later) into consideration. In their work and throughout the information collection process, none of the functions

[1]Note that, because the isotonic functions are fixed, we can connect our analysis of the estimators (of these isotonic functions) back to frequentist statistics.

in the pool of curves are presumed to be equal to or close to the true curve.

To the best of our knowledge, a complete consistency analysis of the estimators proposed in these three works (Chen et al., 2015; He and Powell, 2016; Huang et al., 2018) has not been available. Therefore our goal in this thesis is to carry out a theoretical investigation of the consistency property of the estimator of the true function under DP-R&R. Although the statistical model is Bayesian, our analytical techniques are non-Bayesian. We rely on some tools from frequentist statistical analysis to prove the consistency of what we called the "Bayes-inspired estimators." We will discuss more mathematical detail about DP-R&R in Section 1.3.

1.1.3 Other Applications and Extensions

Initially, isotonic regression arose when several means were compared and estimated. Armitage (1955) gave an early insight in testing linear trends in proportions and frequencies. Later, Bartholomew (1959a; 1959b) provided a test of homogeneity for ordered alternatives, where the alternatives are some mean parameters to be estimated, which was further refined and extended in Bartholomew (1961). Numerous hypothesis testing questions under extremely diversified constraints have been explored in Barlow et al. (1972, Ch.3 and 4) and Robertson et al. (1988, Ch.2, 4 and 5), and with application to biomedical science and bioinformatics in Silvapulle and Sen (2011, Ch.3, 4, 5, 6 and 9). For other issues dealing with monotone and unimodal density estimation, convexity, and goodness of fit, see Groeneboom (1985; 1989; 2001b; 2001a) and Durot (2001; 2002; 2010).

When it may not be possible to arrive at a total or simple ordering, especially when the alternatives (or independent predictors) are multivariate or multidimensional vectors, one may consider partial ordering in issues related to monotonic functions (see Brunk et al., 1957; Hansohm and Hu, 2012; Stout, 2015). Estimation subject to order restrictions arises also when estimating different variations of variance in the analysis of variance. A discussion of theory and examples under both complete and partial ordering among the expected mean squares is available in Thompson (1962). Despite the popular normality assumptions on the conditional distribution of the response $\tilde{y}$ given $\tilde{x} = x$ in many regression analysis, the general theory of isotonic regression makes no distributional assumptions of this kind although many cases that are under consideration fall into the realm of exponential families. Another exponential family that often appears in practice is the binomial distribution, which is well investigated in Ayer et al. (1955) and van Eeden (1956; 1957a; 1957b; 1958). If sometimes the value of predictor variable x represents time, then the ordering is in the sense that the conditional expectation of the response is believed to be varying in a given pattern with time. Instead of observing $\tilde{y}$ on a discrete and finite set of x values, a sequence of $\tilde{y}$ values over a block of time is now recorded. Boswell (1966) analyzed such an example with the Poisson process where the rate parameter is a function of time; for a more recent study for count data, see Dunson (2005). There is plenty of work applying isotonic regression under nonparametric and semiparametric scenarios; see for example Mukerjee (1988) and Cheng (2009). Isotonic regression plays an important role in solving the so-called change-point problem; a review in this area can be found in Khodadadi and Asghar-

ian (2008). Isotonic regression has also emerged in the domain of data mining and machine learning (Caruana and Niculescu-Mizil, 2006; Du and Goel, 2018).

Extensions include, but are not limited to, generalized isotonic regression for fitting isotonic models under convex diffrentiable loss functions through recursive partitioning (Luss and Rosset, 2014); generalized monotonic regression using random change points (Holmes and Heard, 2003); nearly-isotonic regression which is formulated as a convex optimization problem (Tibshirani et al., 2011); sensitivity analysis in isotonic regression (Chakravarti, 1993); online version of isotonic regression when data are collected sequentially (Kotłowski et al., 2016); isotonic regression in efficient learning of generalized linear models and single index models (Kakade et al., 2011); isotonic regression via partitioning with L_p metric (Stout, 2013); and smoothing isotonic regression with regularization/penalization (Sysoev and Burdakov, 2018; Wu et al., 2015).

1.2 Preliminaries of Isotonic Regression

In this section, we introduce a number of definitions and consistency theorems that are previously established in literature, as the foundation of our proofs in Chapter 2.

1.2.1 Basic Concepts

As we mentioned previously, Barlow et al. (1972) and Robertson et al. (1988) provide a broad review of the fundamental theory, tests and applications in order-

restricted statistical inference. Based on these two monographs we now introduce necessary definitions, properties and theorems in isotonic regression analysis for this thesis. Recalling that the word "isotonic" means order preserving, we begin with the definition of simple order and partial order according to Barlow et al. (1972, Def. 1.2 in Sec. 1.3).

Definition 1.2.1 (Simple and Partial Order)**.**

Let $\boldsymbol{X} = \{x_1, x_2, \ldots, x_k\}$. A binary relation "$\lesssim$" on $\boldsymbol{X}$ establishes a *simple order* on $\boldsymbol{X}$ if:

(1) it is reflexive: $x \lesssim x$ for $x \in \boldsymbol{X}$;

(2) it is transitive: $x, y, z \in \boldsymbol{X}$, $x \lesssim y$, $y \lesssim z$ imply $x \lesssim z$;

(3) it is antisymmetric: $x, y \in \boldsymbol{X}$, $x \lesssim y$, $y \lesssim x$ imply $x = y$;

(4) every two elements are comparable: $x, y \in \boldsymbol{X}$ implies either $x \lesssim y$ or $y \lesssim x$.

A *partial order* is reflexive, transitive and antisymmetric, but there may be non-comparable elements. Notice that every simple order is a partial order.

Frequently, $x_i, i = 1, 2, ..., k$ are distinct real numbers, in which case we can assume that $x_1 < x_2 < \cdots < x_k$ without loss of generality. Therefore $x_1 < x_2 < \cdots < x_k$ establishes a simple order on $\boldsymbol{X}$. As we only consider alternatives that are always comparable, we assume to use the simple order on $\boldsymbol{X}$ henceforward unless specified.

For $i = 1, 2, \ldots, k$, set $m(x_i)$ be the number of observations recorded at x_i. Let $y_j(x_i)$, $j = 1, 2, \ldots, m(x_i)$, be the corresponding set of measurements of some

quantity. That is, $\forall x_i \in \boldsymbol{X}$, $y_1(x_i), \ldots, y_{m(x_i)}(x_i)$ are observations on a distribution with some unknown mean parameter denoted by $\mu(x_i) = \mu_i$.

Definition 1.2.2 (Isotonic Function)**.**

A real valued function f on $\boldsymbol{X}$ is called *isotonic* if $x, y \in \boldsymbol{X}$ and $x \leq y$ imply $f(x) \leq f(y)$. If $x, y \in \boldsymbol{X}$ and $x \leq y$ imply $f(x) \geq f(y)$, then f is called antitonic with respect to the simple order $x \leq y$.

Definition 1.2.3 (Isotonic Regression)**.**

Let g be a given function on $\boldsymbol{X}$ and w a given positive function on $\boldsymbol{X}$. An isotonic function g^* on $\boldsymbol{X}$ is an *isotonic regression of g with weights w with respect to the simple ordering* $x_1 \leq x_2 \leq \cdots \leq x_k$ if it minimizes the sum

$$\sum_{x \in X} [g(x) - f(x)]^2 w(x) \tag{1.2.1}$$

in the class of isotonic functions f on $\boldsymbol{X}$. When the weight function and the simple ordering are understood, we call g^* simply an isotonic regression of g.

Definition 1.2.4 (Sample Regression)**.**

Suppose $\mu = \mu(x), x \in \boldsymbol{X}$ is some mean function to be estimated. For $i = 1, 2, \ldots, k$, let $m(x_i)$ be the number of observations recorded at x_i. Let $y_j(x_i)$, $j = 1, 2, \ldots, m(x_i)$, be the corresponding set of independent measurements of some quantity. The *sample regression function* $\bar{y}$ is defined by

$$\bar{y}(x_i) = \frac{1}{m(x_i)} \sum_{j=1}^{m(x_i)} y_j(x_i), \quad x_i \in \boldsymbol{X},$$

or the sample average of all the observation values measured at x_i. In the context of isotonic regression the sample regression function $\bar{y}$ can be regarded as a *basic estimate* of the mean function μ.

A simple approach for estimating μ is to solve the ordinary least squares problem

$$\sum_{x_i \in X} \sum_{j=1}^{m(x_i)} [y_j(x_i) - f(x_i)]^2$$

where the values of $f(x_i)$ are decision variables. Since

$$\sum_{j=1}^{m(x_i)} [y_j(x_i) - f(x_i)]^2 = \sum_{j=1}^{m(x_i)} [y_j(x_i) - \bar{y}(x_i)]^2 + m(x_i) [\bar{y}(x_i) - f(x_i)]^2 ,$$

an equivalent problem is to minimize

$$\sum_{x_i \in X} [\bar{y}(x_i) - f(x_i)]^2 m(x_i) \tag{1.2.2}$$

in the class of linear functions f on $\boldsymbol{X}$.

If no restriction were to be placed on μ, its least squares estimate is clearly the function $\bar{y}(x), x \in \boldsymbol{X}$. In the situation where μ is known to be nondecreasing in x, a least squares estimate of μ would be obtained by minimizing the weighted sum of squares (1.2.2) in the class of isotonic functions with respect to the simple order on $\boldsymbol{X}$: functions f such that $x_i \leq x_j$ implies $f(x_i) \leq f(x_j)$.

Definition 1.2.5 (Sample Isotonic Regression)**.**

Let w be a positive function on $\boldsymbol{X}$. The *sample isotonic regression*, or *isotonized sample regression function* with weights w is the isotonic regression $\bar{y}^*$ of $\bar{y}$. That is, it minimizes

$$\sum_{x \in X} \left[\bar{y}(x) - f(x)\right]^2 w(x) \tag{1.2.3}$$

subject to $f(x_i) \leq f(x_j)$ for $x_i \leq x_j$. In practice, we use the number of observations recorded on x as the weight, that is: $w(x) = m(x), x \in \boldsymbol{X}$.

Comparing Definition 1.2.5 and Definition 1.2.3, g^* is the isotonic regression of g and $\bar{y}^*$ is the isotonic regression of $\bar{y}$. We see that $\bar{y}^*$ and $\bar{y}$ are just special cases of g^* and g respectively. In the case of a simple order, estimation of an isotonic regression function is closely related to estimation of a cumulative (sample) regression function. This part has no major impact to our analysis in this thesis. Therefore we direct readers to Barlow et al. (1972, Sec.2.2), Brunk (1970) and Marshall (1970) for more detailed discussions.

Graphical interpretation–greatest convex minorant

A graphical interpretation of the isotonic regression is to plot the cumulative sums

$$G_j = \sum_{i=1}^{j} g(x_i) w(x_i)$$

against the cumulative sums of weights

$$W_j = \sum_{i=1}^{j} w(x_i),$$

where $x_i \in \boldsymbol{X}$ and $j = 1, 2, ..., k$. That is, one plots the points $P_j = (W_j, G_j)$ with $P_0 = (0, 0)$ in the Cartesian plane. These points constitute the *cumulative sum diagram* (CSD) of the given function g with weights w. The slope of the segment joining P_{j-1} to P_j is just $g(x_j)$. The slope of the chord joining P_{i-1} to $P_j (i \leq j)$ represents the weighted average

$$Av\{x_i, x_{i+1}, \ldots, x_j\} = \sum_{r=i}^{j} g(x_r)w(x_r) / \sum_{r=i}^{j} w(x_r). \tag{1.2.4}$$

Barlow et al. (1972) prove that the isotonic regression of g, i.e. g^*, is given by the slope of the *greatest convex minorant* (GCM) of the CSD. The GCM is also commonly known as the "convex envelope" and graphically it is the path along which a taut string lies if it joins P_0 and P_k and is constrained to lie below the CSD. The value of the isotonic regression g^* at a point x_j is just the slope of the GCM at the point P_j^* with horizontal coordinate $\sum_{i=1}^{j} w(x_i)$. If P_j^* is a corner of the graph of the GCM, $g^*(x_j)$ is the slope of the segment extending to the left. An illustration of CSD with its corresponding GCM is given in Figure 1.2.1. The isotonic regression g^* can be easily calculated by the *Pool-Adjacent-Violators* (PAV) algorithm in $O(n)$ steps (Ayer et al., 1955). The graphs of CSD and GCM were investigated long before Barlow et al. (1972). A program developed by Kruskal (1964) provides the scheme of implementing the Pool-Adjacent-Violator algorithm, and an illustration of the computation with a portion of the data from Bhattacharyya and Klotz (1966) can be found in Barlow et al. (1972, Example 1.3 in Sec.1.2) . For other variations of this algorithm we refer readers to Barlow et al. (1972, Sec.2.3). For more detail

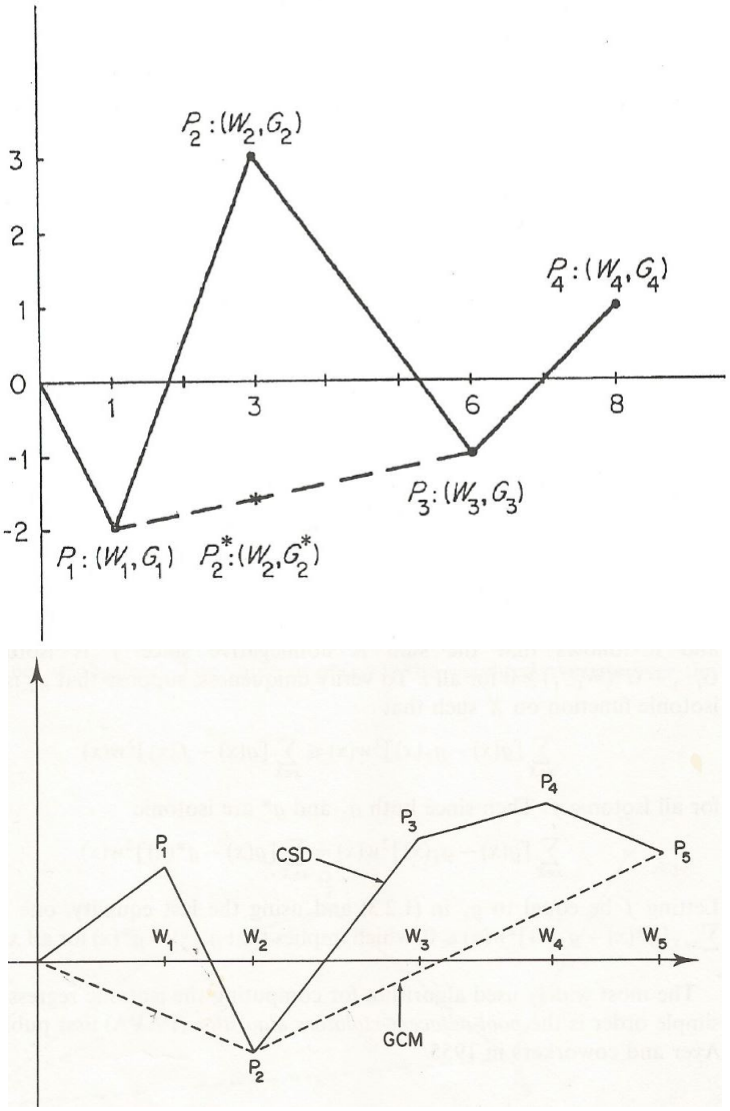

Figure 1.2.1: Examples of CSD and GCM (Barlow et al., 1972; Robertson et al., 1988)

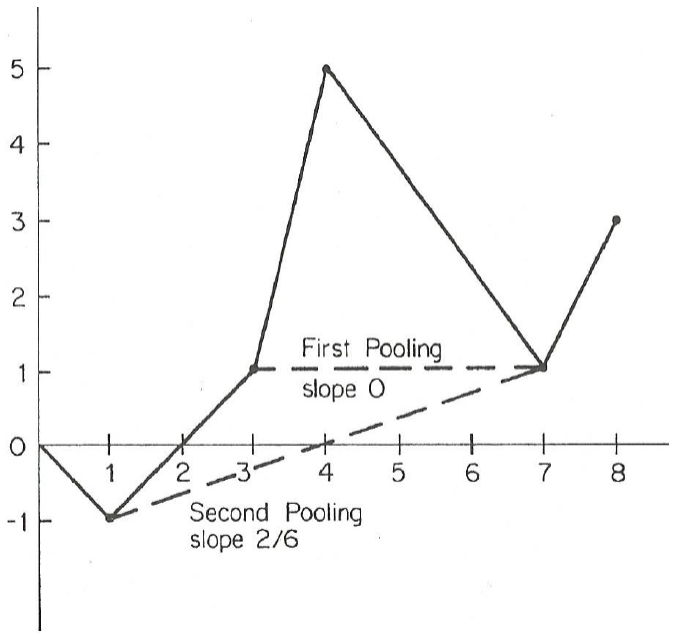

Figure 1.2.2: Graphical interpretation of pooling adjacent violators (Barlow et al., 1972)

and recent developments in GCM, PAV and computationally and algorithmically related topics, see for example Lee et al. (1983), Pardalos and Xue (1999), Anevski and Hössjer (2006), and Luss et al. (2010). We present a simple illustration of this algorithm in Figure 1.2.2.

1.2.2 Properties of Isotonic Regression Estimates

We state theorems on the uniqueness, existence and consistency of the isotonic regression estimator g^* from Barlow et al. (1972, Sec.1.3, 2.1, 2.6), which are the most important properties for our work.

Theorem 1.2.1 (Uniqueness)**.**

An isotonic regression g^* of g with weights w is an isotonic function on $\boldsymbol{X}$ with

respect to a simple order and satisfies

$$\sum_{x \in \boldsymbol{X}} [g(x) - g^*(x)] [g^*(x) - f(x)] w(x) \geq 0$$

and

$$\sum_{x \in \boldsymbol{X}} [g(x) - f(x)]^2 w(x) \geq \sum_{x \in \boldsymbol{X}} [g(x) - g^*(x)]^2 w(x) + \sum_{x \in \boldsymbol{X}} [g^*(x) - f(x)]^2 w(x)$$

for every isotonic function f on $\boldsymbol{X}$.

Conversely, if an isotonic function u satisfies

$$\sum_{x \in \boldsymbol{X}} [g(x) - u(x)] [u(x) - f(x)] w(x) \geq 0$$

for every isotonic function f on $\boldsymbol{X}$ then u is an isotonic regression of g with weights w. There is at most one such isotonic function.

Theorem 1.2.2 (Necessary and Sufficient Condition)**.**

An isotonic function u on $\boldsymbol{X}$ is the isotonic regression of g with weights w if and only if

$$\sum_{x \in \boldsymbol{X}} [g(x) - u(x)] u(x) w(x) = 0$$

and

$$\sum_{x \in \boldsymbol{X}} [g(x) - u(x)] f(x) w(x) \leq 0$$

for all isotonic function f. The isotonic regression g^* of g also satisfies

$$\sum_{x \in \boldsymbol{X}} g(x)w(x) = \sum_{x \in \boldsymbol{X}} g^*(x)w(x)$$

For a slightly different version of this theorem, see Barlow et al. (1972, Thm.1.8). The existence of an isotonic regression is proved as a corollary to the following Theorem 1.2.3, which is also used later to prove consistency.

Theorem 1.2.3 (Existence)**.**

If g_1 and g_2 are isotonic functions on $\boldsymbol{X}$ such that $g_1(x) \leq g(x) \leq g_2(x)$ for $x \in \boldsymbol{X}$, and if g^* is an isotonic regression of g, then also $g_1(x) \leq g^*(x) \leq g_2(x)$ for $x \in \boldsymbol{X}$. In particular, if a and b are constants such that $a \leq g(x) \leq b$ for $x \in \boldsymbol{X}$, then also $a \leq g(x) \leq b$ for $x \in \boldsymbol{X}$. Hence, an isotonic regression of g exists.

The consistency of isotonic regression estimator is the foundation of our theoretical work in this thesis. The following Theorem 1.2.4 of Barlow et al. (1972, Thm.2.1) states that if $\bar{y}^*$ is the isotonic regression of the sample mean $\bar{y}$, then $\bar{y}^*$ is better then $\bar{y}$ in terms of least squares.

Theorem 1.2.4 (Isotonic Regression of an Estimator is Better Than the Estimator Itself)**.**

Let μ be an unknown function on $\boldsymbol{X}$, known to be isotonic with respect to the simple order on $\boldsymbol{X}$. Let $w(x), x \in \boldsymbol{X}$ be a set of positive weights. Let g be an estimate of

μ. Let g^* be the isotonic regression of g with weights w. Then

$$\sum_{x\in \boldsymbol{X}} [\mu(x) - g^*(x)]^2 w(x) \leq \sum_{x\in \boldsymbol{X}} [\mu(x) - g(x)]^2 w(x).$$

We can interpret the squared difference $[\mu(x) - g(x)]^2$ as a "loss" sustained if $g(x)$ is taken as estimate of the unknown true $\mu(x)$. In this sense Theorem 1.2.4 indicates that if μ is isotonic, the (weighted) average loss is not greater when the estimate g is replaced by its isotonic regression g^*.

The main theorem of consistency (Barlow et al., 1972, Thm.2.2) further shows that if an estimator (or a basic estimator like $\bar{y}$) of μ is consistent (or strongly consistent), then its isotonic regression $\bar{y}^*$ is also a consistent (strongly consistent) estimator of μ. Consider a set $\boldsymbol{X}$, not necessarily finite, endowed with a partial order. Let μ be an isotonic function on $\boldsymbol{X}$. Let $\{\boldsymbol{X}_n\}$ be an expanding class of finite subsets of $\boldsymbol{X}$: $\boldsymbol{X}_n \subset \boldsymbol{X}_{n+1} \cdots \subset \boldsymbol{X}$; set $\boldsymbol{X}' = \cup_n \boldsymbol{X}_n$. (In certain applications, $\boldsymbol{X}_n$ may coincide with $\boldsymbol{X}$ for each n.) Denote by μ_n a restriction of μ to $\boldsymbol{X}_n$: $\mu_n(x) = \mu(x)$ for $x \in \boldsymbol{X}_n$. Let $\tilde{g}_n$ be an estimator of μ, $n = 1, 2, \ldots$. By applying Theorem 1.2.4, Barlow et al. (1972, Thm.2.2) state the following main consistency results.

Theorem 1.2.5 (Consistency)**.**

For $n = 1, 2, \ldots$, let $w_n(x), x \in \boldsymbol{X}_n$ (or $\tilde{w}_n(x), x \in \boldsymbol{X}_n$) be positive reals (or positive random variables). Let $\{\tilde{g}_n, \boldsymbol{X}_n\}$ be a consistent (strongly consistent) sequence of estimators of μ at each $x \in \boldsymbol{X}' = \cup_n \boldsymbol{X}_n$. Let μ be isotonic on $\boldsymbol{X}$. Denote by $\tilde{g}_n^*$ the isotonic regression of $\tilde{g}_n$ on $\boldsymbol{X}_n$ with weights w_n (or $\tilde{w}_n$), $n = 1, 2, \ldots$. Then $\{\tilde{g}_n^*, \boldsymbol{X}_n\}$ also is a consistent (strongly consistent) sequence of estimators of μ at

each $x \in \boldsymbol{X}'$.

The above two theorems deal with consistency when μ is isotonic on $\boldsymbol{X}$. If μ is not isotonic, define μ^* as the isotonic regression of μ, and consider estimating the (unknown) isotonic regression function μ^*. It should be noted that μ^* is not defined until a weight function w on $\boldsymbol{X}$ is specified. Once the weight w on $\boldsymbol{X}$ is clear, one may take the isotonic regression of $\bar{y}$ with weights w as estimator of μ^*. Then by using the strong law of large numbers, Barlow et al. (1972, Thm. 2.15) prove that $\bar{y}^*$ converges to μ^* almost surely as $\min_{x_i \in \boldsymbol{X}} m(x_i) \to \infty$.

Theorem 1.2.6 (Consistency of Sample Isotonic Regression Estimator)**.**
The sample isotonic regression function $\bar{y}^*$ *with weights* $w(x), x \in \boldsymbol{X}$*, converges with probability* 1 *to the isotonic regression* μ^* *at each* $x \in \boldsymbol{X}$ *as*

$$\min_{x \in \boldsymbol{X}} m(x) \to \infty$$

where $m(x)$ *is the number of observations recorded at* x*.*

The first theorem of this kind to our knowledge was given by Ayer et al. (1955). A generalization appears in Brunk (1955), and both are subsumed by Brunk (1958, Thm. 6.2), of which a corrected version appears in Brunk (1970, Thm. 4.1 in Sec.1.3). For an extension of these results to generalized isotonic regression, see Robertson and Wright (1975).

We now state one more theorem from Robertson et al. (1988, Thm.1.3.5) that will be used in the proof of a lemma later. Suppose g and w are functions defined

on $\boldsymbol{X}$, set

$$Av(\boldsymbol{A}) = \frac{\sum_{x \in A} g(x)w(x)}{\sum_{x \in A} w(x)}$$

for those $\boldsymbol{A}$ such that $\boldsymbol{A} \neq \emptyset$ and $\boldsymbol{A} \subset \boldsymbol{X}$. That is: $Av(\boldsymbol{A})$ is a special average of $g(x)$ over a nonempty subset $\boldsymbol{A}$ of $\boldsymbol{X}$ with weights w. Let $[g^* = c]$ denote the set $\{x \in \boldsymbol{X} : g^*(x) = c\}$. While $Av(\boldsymbol{A})$ depends on g, this will not be made explicit in the notation.

Theorem 1.2.7.

If c is any real number and if the subset $[g^* = c]$ of $\boldsymbol{X}$ on which g^* takes the value c is nonempty, then

$$c = Av([g^* = c]).$$

Note that the definition of $Av(\boldsymbol{A})$ is very similar to the expression 1.2.4.

1.3 Paradigm of Discrete Prior

In this section, we introduce the definition of the candidate set, the pool of curves and the information collection process, as well as two major algorithms: resampling and regeneration, all of which lead to the construction of our weighted sum estimator.

1.3.1 General Setup and Notations

We will make modifications to the corresponding definition and notations in Section 1.2.1 to maintain a consistent style. We begin with a slight modification to

the definition of $\boldsymbol{X}$ in the previous context and in the Definition 1.2.1.

Definition 1.3.1 (Set of Choices)**.**

Set

$$\boldsymbol{X} = \{x_1, x_2, ..., x_M\}$$

as a collection of M different *choices* where $x_1, x_2, ..., x_M$ are M distinct real numbers. Without loss of generality for the purpose of our research needs, we may further specify that

$$x_1 = 1, x_2 = 2, ..., x_M = M.$$

That is: $\boldsymbol{X} = \{1, 2, ..., M\}$. From now on we use this $\boldsymbol{X} = \{1, 2, ..., M\}$ for the rest of this thesis unless otherwise specified.

Definition 1.3.2. Define an unknown function

$$\mu(x_i) = \mu_i,\ x_i \in \boldsymbol{X},$$

which is assumed to be isotonic with respect to the simple order $x_1 < x_2 < ... < x_M$, that is, $1 < 2 < ... < M$, on the set $\boldsymbol{X}$. We call $\mu(x_i) = \mu_i$ the *true function*, or the *true curve*. For simplicity, we may write $\mu(x), x \in \boldsymbol{X}$ when the context is clear.

In many practical applications, such a true function μ is considered as a utility function (Chen et al., 2015) or a cost function (Huang et al., 2018) either to be maximized or minimized over the set of alternatives. The main goal of this thesis, however, is to study the consistency of the proposed estimator of this true

function through an information collection process with a budget of $\boldsymbol{N}$ sequential measurements.

In each experiment, we only query one choice of $x \in \boldsymbol{X}$ and obtain a single noisy measurement $\hat{y}(x)$ of $\mu(x)$. At the nth experiment (or call it iteration or measurement), we measure x^{n-1} and choose to measure $x^n \in \boldsymbol{X}$ in the next round of experiment according to some exogenous decision-making rule. Note that x^n with superscript index n indicates that it is the one element in $\boldsymbol{X}$ we query at the time $(n+1)$ experiment, and it should not be confused with the notation in Definition 1.3.1. We assume in the $(n+1)$th experiment, the inherent independent sequence of noise, W^{n+1}, is normally distributed with mean zero and variance σ^2 that is known to us. Thus, for $n = 0, 1, ..., \boldsymbol{N}-1$, the independent sample measurement $\hat{y}^{n+1}(x^n)$ for the alternative x^n is of the form

$$\hat{y}^{n+1}(x^n) = \mu_{x^n} + W^{n+1}$$

where $W^{n+1} \overset{iid}{\sim} N(0, \sigma^2)$. Notice that, as pointed out in Chen et al. (2015), the decision x is indexed by superscript n and the measurement $\hat{y}$ is indexed by superscript $(n+1)$ for the purpose of emphasizing the fact that $\hat{y}^{n+1}$ is an unknown stochastic value when the measurement decision is made at time n. The value of $\hat{y}^{n+1}$ will only be observed at the time $(n+1)$ experiment after the time n measurement has been performed. We formally state the filtration in the following definition.

Definition 1.3.3 (Filtration Up To Time n)**.**

Denote by $\mathcal{F}^n$ the sigma algebra generated by the history of decisions and

Time	0	1	2	...	$n-1$	n	$n+1$
Measurement		$\hat{y}^1(x^0)$	$\hat{y}^2(x^1)$	...	$\hat{y}^{n-1}(x^{n-2})$	$\hat{y}^n(x^{n-1})$	$\hat{y}^{n+1}(x^n)$
Decision	x^0	x^1	x^2	...	x^{n-1}	x^n	x^{n+1}

Table 1.1: Information Collection

measurements up to time n. That is:

$$\mathcal{F}^n = \sigma(x^0, \hat{y}^1(x^0), x^1, \hat{y}^2(x^1), ..., x^{n-1}, \hat{y}^n(x^{n-1})).$$

An illustration of the information collection process can be found in Table 1.1.

Definition 1.3.4 (Information Collection)**.**

We define the following three quantities that are related to information collection:

(1) Let

$$\widehat{\boldsymbol{H}}^n = \{\hat{y}^1(x^0), \hat{y}^2(x^1), ..., \hat{y}^n(x^{n-1}), x^i \in \boldsymbol{X}, i = 0, 1, 2, ..., n-1\}$$

be the history of independent observations up to time n, i.e., the first n information collection measurements.

(2) Let

$$w^n = w^n(x), x \in \boldsymbol{X} = \{1, 2, ..., M\}$$

be the number of measurements taken on the alternative x up to time n. Then $\sum_{x \in \boldsymbol{X}} w^n(x) = n$. Throughout this thesis, we assume that $w^n(x) \to \infty$ as $n \to \infty$ for each $x \in \boldsymbol{X}$.

(3) Define the sample mean function of each x up to time n as

$$\bar{y}^n = \bar{y}^n(x) = \frac{1}{w^n(x)} \sum_{j=1}^{w^n(x)} \hat{y}_j(x),\ x \in \boldsymbol{X}.$$

Notice that, for the sake of conciseness, we slightly sacrifice the clarity in notation by abusing the usage of subscript of $\hat{y}$. Here each $\hat{y}_j(x), j = 1, 2, ..., w^n(x)$ is a measurement taken on that particular x and it may not be the jth observation in the sequence of information collection $\widehat{\boldsymbol{H}}^n$.

Remark. This w^n is in the same role of the function m in Definition 1.2.4. Again this is due to the fact that in real applications we use the number of measurements taken on each alternative as the weight function in sample isotonic regression. We use $x \in \boldsymbol{X}$ when generically referring to an arbitrary element of $\boldsymbol{X}$, and use $x^i \in \boldsymbol{X}$ to additionally emphasize its order in the sequence of measurements.

1.3.2 Candidate Set and Pool of Curves

The concept of a candidate set is first introduced in Chen et al. (2015) to overcome the computational difficulty of Bayesian inference on functions. Later, He and Powell (2016) extend the idea with a much larger set, the pool of curves, such that the curves in the candidate set can be updated by the curves in the pool of curves with some resampling criterion. Huang et al. (2018) allows the functions in the pool of curves to be updated and extended with some regeneration mechanism. Since both of these two sets will be updated during the information collection process, naturally their notation will include the time index n.

Definition 1.3.5 (Candidate Set and Pool of Curves)**.**

Denote by $\mathcal{C}^n$ and $\mathcal{P}^n$ respectively the *candidate set* and the *pool of curves* up to the nth information collection experiment. Set

$$\mathcal{C}^n = \{f_1^n, f_2^n, ..., f_L^n\}.$$

We use $\mathcal{C}$ and $\mathcal{P}$ when we refer to the candidate set and pool of curves generically.

Throughout the entire information collection process, the number of elements in the candidate set is fixed at $\|\mathcal{C}^n\| = L$ for all n. However, we allow the L candidates in $\mathcal{C}^n$ to be replaced by other better (in terms of some criterion) functions. The superscript n indicates the current number of iterations and the subscripts $i, i = 1, 2, ..., L$ only indicate that f_i^n is the ith candidate in $\mathcal{C}^n$.

As will be elaborated more in Section 1.3.3, we generate new candidate functions by isotonic regression and add them into $\mathcal{P}^n$ as extensions. Because we may not able to add a function into $\mathcal{P}^n$ at each time the regeneration procedure is triggered, we distinguish the sequence of isotonic regression estimators from the actual sequence of functions that are being added into the pool of curves in the following definition using the notations in Definition 1.3.4:

Definition 1.3.6 (Sample Isotonic Regression Based on the Measurement History $\widehat{\boldsymbol{H}}^n$)**.**

On the set of alternatives $\boldsymbol{X}$, define $\{f^n\}_{n=1}^{\infty}$ as the sequence of sample isotonic regression estimators of $\bar{y}^n$ with weight w^n according to the measurement history $\widehat{\boldsymbol{H}}^n$.

Let $\left\{f^{N_q}\right\}_{q=1}^{\infty}$ be the sequence of functions that is added into the pool of curves. That is, $\{N_q\}_{q=1}^{\infty}$ is the subsequence of time indices at which a function in $\{f^n\}_{n=1}^{\infty}$ is added into the pool of curves. In addition, for two functions $h, g \in \{f^n\}_{n=1}^{\infty}$ that are added into the pool of curves with h added later than g, we write N^h and N^g as the moment of time (or order of the experiment) when h and g are being added respectively. Then in this case, $N^h > N^g$.

Now we may write

$$\mathcal{P}^n = \{f^{N_1}, f^{N_2}, ..., f^{N_{l(n)}}\}$$

with cardinality $\|\mathcal{P}^n\| = l(n)$. Because we only add at most one function at a time and never remove functions from the pool of curves, for the time indices $n > N^h > N^g$ we have

$$\mathcal{P}^{N^g} \subseteq \mathcal{P}^{N^h} \subseteq \mathcal{P}^n.$$

For example, up to the 100th iteration, we could have

$$\mathcal{P}^{100} = \{f^{N_1}, f^{N_2}, f^{N_3}\} = \{f^7, f^{25}, f^{90}\}.$$

Thus in this particular example, the cardinality of the current pool of curves $\mathcal{P}^{100}$ is $l(n) = l(100) = 3$, with $N_1 = 7, N_2 = 25, N_{l(n)} = N_{l(100)} = N_3 = 90$. That is: in the first 100 iterations we add in total three candidates at the 7th, 25th and 90th iteration.

We end this subsection with the definition of three types of the Mean Squared

Error (MSE) used in the regeneration and resampling procedures and later in the consistency proofs. In general they are the MSE between two functions, between a function and the sample mean, and between a function with the measurement history up to the time this function is added into the pool of curves.

Definition 1.3.7 (Mean Squared Errors (MSE))**.**

For any two functions $h, g \in \{f^n\}$ that are added into $\mathcal{P}$ at the time N^h and N^g respectively, denote by $\widehat{\boldsymbol{H}}^{N^h} = \{\hat{y}^1, \hat{y}^2, ..., \hat{y}^{N^h}\}$ the measurement history up to time N^h. We define the following mean squared errors using the measurement history $\widehat{\boldsymbol{H}}^{N^h}$ and $\widehat{\boldsymbol{H}}^n$ accordingly:

(1) Define the MSE between h and $\widehat{\boldsymbol{H}}^{N^h}$ as

$$\widehat{\Xi}(h, \widehat{\boldsymbol{H}}^{N^h}) = \frac{1}{N^h} \sum_{x \in \boldsymbol{X}} \sum_{1 \le j \le w^{N^h}(x)} [h(x) - \hat{y}_j(x)]^2.$$

(2) Define the MSE between h and g as

$$\Xi(h, g) = \frac{1}{M} \sum_{x \in \boldsymbol{X}} [h(x) - g(x)]^2.$$

Note that $\Xi(h, g)$ defined in this way is indeed the square of an L_2-norm up to a scale factor $1/M$. Additionally the sample mean up to time N^h, $\bar{y}^{N^h} = \bar{y}^{N^h}(x), x \in \boldsymbol{X}$, is also a finite function whose M values on $\boldsymbol{X}$ can be viewed as an M-dimensional vector. Thus in the same fashion we write:

(3) The MSE between the function h and $\bar{y}^{N^h}$ as

$$\Xi(h, \bar{y}^{N^h}) = \frac{1}{M} \sum_{x \in \boldsymbol{X}} [h(x) - \bar{y}^{N^h}(x)]^2.$$

Similarly, for an arbitrary n, the MSE between the function h with $\bar{y}^n$ based on $\widehat{\boldsymbol{H}}^n$ is defined as

$$\Xi(h, \bar{y}^n) = \frac{1}{M} \sum_{x \in \boldsymbol{X}} [h(x) - \bar{y}^n(x)]^2.$$

1.3.3 Regeneration Procedure

We begin with the formulation of the prior probabilities of the candidates in $\mathcal{C}^n$. Suppose for each candidate $f_l^n \in \mathcal{C}^n$ there is a corresponding current probability $p_l^n, l = 1, 2, ..., L$. That is: p_l^n is the current prior probability that the l-th candidate, f_l^n, is the true curve μ at the current experiment time n. In this sense the functions in the candidate set $\mathcal{C}^n$ are considered as the *discrete priors* of μ with probability $p_l^n, l = 1, 2, ..., L$. Conditioning on f_l^n being the true curve and on deciding to measure x^n in the next iteration, the $(n+1)$th observation is given by $\hat{y}^{n+1} \sim N(f_l^n(x^n), \sigma^2)$, and the likelihood of f_l^n is given by

$$\mathcal{L}(\hat{y}^{n+1} | \mu = f_l^n) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{[\hat{y}^{n+1} - f_l^n(x^n)]^2}{2\sigma^2} \right\}.$$

Applying Bayes Theorem, the posterior probability of f_l^n being the true curve, p_l^{n+1}, is proportional to the prior p_l^n multiplied by the above likelihood:

$$\begin{aligned}
p_l^{n+1} &= \mathbb{P}(\mu = f_l^n | \mathcal{F}^{n+1}) \\
&= \mathbb{P}(\mu = f_l^n | \hat{y}^{n+1}(x^n), x^n, ..., \hat{y}^1(x^0), x^0) \\
&\propto \mathcal{L}(\hat{y}^{n+1} | \mu = f_l^n, x^n, ..., \hat{y}^1(x^0), x^0) \cdot \mathbb{P}(\mu = f_l^n | x^n, ..., \hat{y}^1(x^0), x^0) \\
&= \mathcal{L}(\hat{y}^{n+1} | \mu = f_l^n, x^n) \cdot \mathbb{P}(\mu = f_l^n | x^n, \mathcal{F}^n) \\
&= \mathcal{L}(\hat{y}^{n+1} | \mu = f_l^n, x^n) \cdot \mathbb{P}(\mu = f_l^n | \mathcal{F}^n) \\
&= \mathcal{L}(\hat{y}^{n+1} | \mu = f_l^n) \cdot p_l^n \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left[\hat{y}^{n+1} - f_l^n(x^n)\right]^2}{2\sigma^2} \right\} p_l^n.
\end{aligned}$$

Since $p_l^{n+1}, l = 1, 2, ..., L$ should sum to 1, after dividing by the normalizing factor

$$\sum_{l'=1}^{L} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left[\hat{y}^{n+1} - f_{l'}^n(x^n)\right]^2}{2\sigma^2} \right\} p_{l'}^n,$$

the updating equation for p_l is given by

$$p_l^{n+1} = \frac{\exp\left\{ -\left[\hat{y}^{n+1} - f_l^n(x^n)\right]^2 / (2\sigma^2) \right\} p_l^n}{\sum_{l'=1}^{L} \exp\left\{ -\left[\hat{y}^{n+1} - f_{l'}^n(x^n)\right]^2 / (2\sigma^2) \right\} p_{l'}^n}. \tag{1.3.1}$$

As will be discussed in the next section, when resampling happens we are not able to use (1.3.1). Instead we need to use the posterior probability derived according to the entire measurement history up to time $(n+1)$ if the resampling is triggered at time $(n+1)$. In that case, conditional on f_l^{n+1} being the true curve,

the likelihood of f_l^{n+1} based on all of the previous observations is given by

$$\begin{aligned}&\mathcal{L}\left(\widehat{\boldsymbol{H}}^{n+1}=\{\hat{y}^1,\hat{y}^2,...,\hat{y}^{n+1}\}\,\middle|\mu=f_l^{n+1}\right)\\&=\prod_{i=0}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left[\hat{y}^{i+1}-f_l^{n+1}(x^i)\right]^2}{2\sigma^2}\right\}.\end{aligned}$$

Then the corresponding posterior p_l^{n+1} is updated by

$$\begin{aligned}p_l^{n+1}&=\frac{\prod_{i=0}^{n}\exp\left\{-\left[\hat{y}^{i+1}-f_l^{n+1}(x^i)\right]^2/(2\sigma^2)\right\}p_l^0}{\sum_{l'=1}^{L}\prod_{i=0}^{n}\exp\left\{-\left[\hat{y}^{i+1}-f_{l'}^{n+1}(x^i)\right]^2/(2\sigma^2)\right\}p_{l'}^0}\\&=\frac{\prod_{i=0}^{n}\exp\left\{-\left[\hat{y}^{i+1}-f_l^{n+1}(x^i)\right]^2/(2\sigma^2)\right\}}{\sum_{l'=1}^{L}\prod_{i=0}^{n}\exp\left\{-\left[\hat{y}^{i+1}-f_{l'}^{n+1}(x^i)\right]^2/(2\sigma^2)\right\}}\end{aligned}\tag{1.3.2}$$

where $p_{l'}^0, l'=1,2,...,L$ are assumed to have a uniform prior distribution, i.e., $p_1^0=p_2^0=...=p_L^0=1/L$.

The idea of regenerating new candidates to be added into $\mathcal{P}$ was first proposed in Huang et al. (2018). The regeneration is triggered every n^{reg} iterations and the procedure can be summarized in Algorithm 1.

Again even with such regeneration procedure, we do not assume that the true curve μ would ever be included in $\mathcal{P}$.

1.3.4 Resampling Procedure

The resampling method was first introduced by He and Powell (2016) and then slightly simplified by Huang et al. (2018). The resampling is triggered either every n^{res} iterations, or over a certain percentage of the candidates in $\mathcal{C}^n$ satisfying

Algorithm 1: Regeneration Procedure

Suppose at the nth iteration, the regeneration is triggered after recording the observation $\hat{y}^n$. Then regenerate the curve f^n as the sample isotonic regression estimator of $\bar{y}^n$ with weight w^n based on the measurement history $\widehat{\boldsymbol{H}}^n$. Notice that f^n is an element of $\{f^n\}_{n=1}^{\infty}$ in Definition 1.3.6.

for *each curve* $f^{N_j}, j = 1, 2, ..., l(n)$, *in the pool* $\mathcal{P}^n$ **do**

    Calculate the MSE $\Xi(f^n, f^{N_j})$

endfor

if *all the MSE* $\Xi(f^n, f^{N_j}) \geq \epsilon_{reg}^n$ *for some positive threshold value* ϵ_{reg}^n **then**

    Insert the curve f_n into the pool, i.e., $\mathcal{P}^{n+1} = \mathcal{P}^n \bigcup \{f^n\}$.

endif

Reduce the threshold to $\epsilon_{reg}^{n+1} = \epsilon_{reg}^n \cdot \gamma$ where $\gamma \in (0, 1)$. The initial threshold is $\epsilon_{reg}^0 = \epsilon_{reg}$ so that $\epsilon_{reg}^n = \epsilon_{reg} \cdot \gamma^n$, where γ^n is the nth power of γ.

Return the updated pool $\mathcal{P}^{n+1}$

$p_l^n \leq \epsilon^{res}$ where $\epsilon^{res} \in (0,1)$ is some threshold value. Combining the process in these two works with our own modification for the purpose of later proof, we present the modified resampling procedure in Algorithm 2. Note that in the Step 4 of Algorithm 2, the weighted sampling without replacement is not a simple process. We refer to the Algorithm 1 in He and Powell (2016) for more computation details.

As in the case of regeneration, we do not assume that the true curve μ would ever be included in $\mathcal{C}$. Notice that in each round of iteration the resampling procedure may be triggered more than one time since it could take several rounds of resampling such that a certain percentage, say 25% of the (or $L/4$ that many) candidates in $\mathcal{C}^n$ are all satisfying $p_l^n \leq \epsilon^{res}$.

1.3.5 Estimator of the True Curve

Definition 1.3.8. Denote by $\bar{f}^n$ the estimator of the true curve μ on $\boldsymbol{X}$ up to the nth information collection experiment. Let $\mathbb{L} = \{1, 2, ..., L\}$ be the index set for $\mathcal{C}$. Then $\bar{f}^n$ is formulated as a weighted sum of all the candidate functions in $\mathcal{C}^n$:

$$\bar{f}^n(x) = \sum_{l \in \mathbb{L}} p_l^n f_l^n(x),\ x \in \boldsymbol{X}.$$

In this thesis we study the asymptotic behavior of this estimator and prove the strong consistency of $\bar{f}^n$ as an estimator of μ.

Algorithm 2: Resampling Procedure

Let $\boldsymbol{N}$ be the information collection budget which is the maximum number of experiments we can afford.

for $n = 0$ to $\boldsymbol{N} - 1$ **do**

Step 1 Take the measurement $\hat{y}^{n+1}$ and update each p_l^n for $f_l^n \in \mathcal{C}^n$ by (1.3.1) i.e.,

$$p_l^{n+1} = \frac{\exp\left\{-\left[\hat{y}^{n+1} - f_l^n(x^n)\right]^2 / (2\sigma^2)\right\} p_l^n}{\sum_{l'=1}^{L} \exp\left\{-\left[\hat{y}^{n+1} - f_{l'}^n(x^n)\right]^2 / (2\sigma^2)\right\} p_{l'}^n}$$

Step 2 Remove curves with $p_l^{n+1} \leq \epsilon^{res}$ from $\mathcal{C}^n$. Let K^{del} be the number of deleted curves.

If $K^{del} = 0$ then remove one curve with the smallest p_l^{n+1} from $\mathcal{C}^n$ and set $K^{del} = 1$.

Step 3 Calculate the mean squared errors $\Xi(h, \bar{y}^n)$ for each $h \in \mathcal{P}^n$ based on $\widehat{\boldsymbol{H}}^n$. Select K curves ($K > K^{del}$) with the smallest MSEs.

Step 4 From the K curves, using weighted (p_l^n) sampling without replacement, select K^{del} curves to be resampled into $\mathcal{C}^n$.

Step 5 Again update the probabilities p_l^{n+1} of each candidate in $\mathcal{C}^n$ by (1.3.2), i.e.,

$$p_l^{n+1} = \frac{\prod_{i=0}^{n} \exp\left\{-\left[\hat{y}^{i+1} - f_l^{n+1}(x^i)\right]^2 / (2\sigma^2)\right\}}{\sum_{l'=1}^{L} \prod_{i=0}^{n} \exp\left\{-\left[\hat{y}^{i+1} - f_{l'}^{n+1}(x^i)\right]^2 / (2\sigma^2)\right\}}$$

Here we do not use (1.3.1) because p^n values are not available for the K^{del} curves that are newly added into $\mathcal{C}^n$.

end for

Return the updated $\mathcal{C}^{n+1}$ and $p_l^{n+1}, l = 1, 2, ..., L$.

1.4 Motivating Example

Our research topic originates from the field of supply chain and logistics, specifically, the vehicle allocation problem in mega-cities. In the recent decade, the prosperity of e-commerce generates enormous numbers of business-to-consumer (B2C) customers in major large cities worldwide. Such a dramatic change of business model greatly challenges the city logistics, due to the immense demands with high volatility based on the location of customers in different regions, and the need to make deliveries subject to limited shipping capacities and resources.

Huang et al. (2018) studied the vehicle allocation problem in Beijing and it can be generally described as follows. Suppose that a logistic company owns M homogeneous vehicles and the entire urban area its delivery service covers can be divided into R regions. In each region r, on a daily basis, set $\boldsymbol{L}_r$ as the random vector of all the customer locations and $\boldsymbol{D}_r$ as the random vector of the delivery quantities. Then the pair $(\boldsymbol{L}_r(\omega), \boldsymbol{D}_r(\omega))$ is a realization of the everyday task that the company faces. Now, a natural question to ask is how to allocate the M vehicles to the R regions such that the total expected operational cost is minimized. The cost function can be different for each choice of region r. An illustration of two different cost functions for two regions r and r' is given in Figure 1.4.1.

If the true operational cost function is known or easy to evaluate, we essentially face a deterministic resource allocation problem which can be solved using standard optimization algorithms. However in practice the accurate assessment for the cost of each set of vehicle allocation plans could be computationally expensive and subject

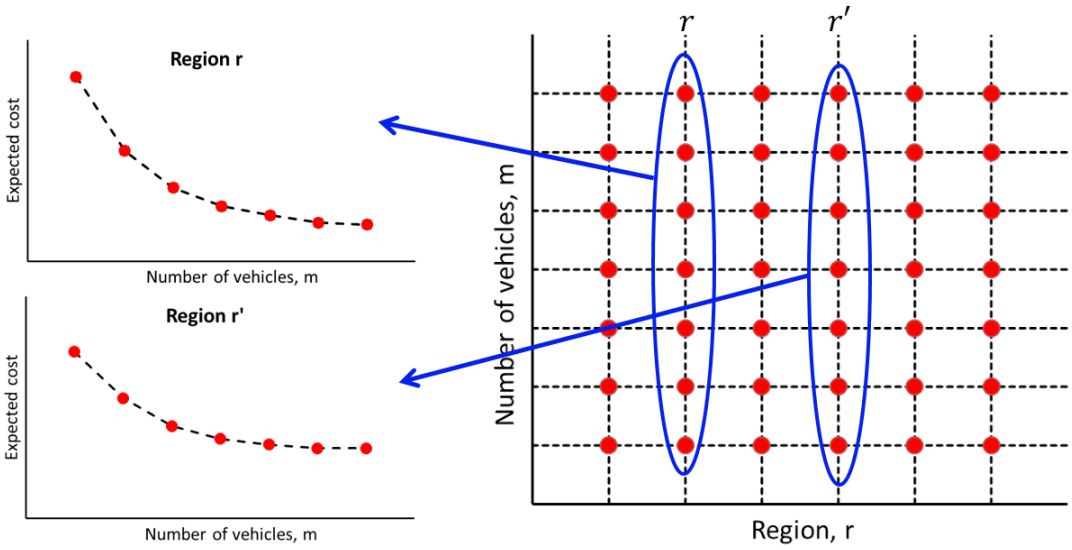

Figure 1.4.1: Different expected cost functions of two regions in (Huang et al., 2018)

to some limited experimental budget. Therefore we need an efficient way to collect information regarding with the true cost function, and such a topic falls into the realm of *optimal information collection and learning,* which is beyond the primary scope of this thesis. We discuss it in Appendix A.

To relate this motivating example back to Section 1.3, $\boldsymbol{X}$ is the set of all possible vehicle allocations within one region. Recall that the cost function is assumed to be monotonic and piecewise linear, and that we only observe sample points for costs on discrete integer values, namely, the number of vehicles assigned to a region. The goal is to estimate the true cost function μ by a series of measurements $(\widehat{\boldsymbol{H}}^n)$ taken on each allocation. Although the cost function could be depending on regions, the theoretical derivations are very similar for each different region. Therefore we only study the problem within one region and we do not use a subscript r for any of the quantities defined in the previous sections.

Chapter 2: Consistency of Weighted Estimator in DP-R&R

In this chapter we present a consistency analysis of the estimator $\bar{f}^n$. Section 2.1 shows that, through the regeneration procedure, we are able to add infinitely many functions to the pool of curves and they can be arbitrarily close to the true curve. Section 2.2 argues that the curves (with the smallest MSE) that we resample into the candidate set are in fact getting closer to the true curve during the information collection process. Section 2.3 demonstrates that for functions that are distanced away from the true curve, we are able to remove them by the resampling procedure so that eventually all such inferior functions will be deleted and substituted by functions that can be arbitrarily close to the true curve. Finally in Section 2.4 we establish the strong consistency of the estimator $\bar{f}^n$ for the true function μ. As a supplement to the consistency proof, in Section B we discuss the limiting behavior of isotonic regression estimators where not every alternative is measured infinitely many times.

2.1 Cardinality of Pool of Curves

Given a fixed sample path ω, as n goes to infinity the limiting cardinality of $\mathcal{P}^n$ is either finite or infinite. We investigate these two cases below, beginning with a

lemma which shows that the event where an isotonic regression estimator coincides with the true function is a measure zero set.

Lemma 2.1.1. *Suppose a function* $h = f^{N^h} \in \{f^n\}_{n=1}^{\infty}$ *is an isotonic regression estimator of* $\bar{y}^{N^h}(x)$*. For any subset* $\boldsymbol{X}' \subseteq \boldsymbol{X}$*, define*

$$\widetilde{B} = \{\omega : h(x) = \mu(x) \text{ for all } x \in \boldsymbol{X}' \subseteq \boldsymbol{X}\},$$

which is the event that h *partially coincides with* μ *on* $\boldsymbol{X}$*. Then* $\mathbb{P}(\widetilde{B}) = 0$*.*

Proof. First, notice that the isotonic regression estimator h is the optimal solution to a quadratic programming problem in equation (1.2.3). It is shown by Theorem 1.2.1 that the isotonic regression estimator is unique. We then derive our argument based on the relationship between the isotonic regression estimator and the Greatest Convex Minorant (GCM) of the Cumulative Sum Diagram (CSD) of $\bar{y}^{N^h}(x), x = 1, 2, \ldots, M$. In general, the isotonic regression h partitions $\boldsymbol{X}$ into sets on which h is constant, that is, into *level sets* for $\bar{y}^{N^h}$, called *solution blocks* by Barlow et al. (1972). On each of these solution blocks the value of h is the weighted average of the values of $\bar{y}^{N^h}$ over the block, using weights w^{N^h} defined in the Definition 1.3.4. In other words, the solution blocks are the sets of consecutive elements of $\boldsymbol{X}$ on each of which h assumes a particular value. There are at most finitely many solution blocks since $\boldsymbol{X}$ is a finite set.

Let $E_i, i \in D = \{1, 2, ..., d\}$ be the d solution blocks of $\bar{y}^{N^h}$ on $\boldsymbol{X}$. Then $\boldsymbol{X} = \bigcup_{i \in D} E_i$ where $E_i \cap E_j = \emptyset, i \neq j$. Suppose $\mu_1, \mu_2, ..., \mu_d$ are the d distinct values of $\mu(x)$ on $\boldsymbol{X}$ and $d \leq M$. Then there exists a subset of indices $D' \subseteq D$ such that

$\boldsymbol{X}' \subseteq \bigcup_{j \in D'} E_j$ and $h(x) = \mu(x) = \mu_j, \forall x \in E_j, j \in D'$. Denote by $\xi = \xi(x) = \xi_x$ the function we need to estimate, then in our case $\xi(x) = \bar{y}^{N^h}(x), x \in \boldsymbol{X}$. It follows from Theorem 1.2.7 that

$$\mu_j = \frac{\sum_{x \in E_j} \xi_x \cdot w^{N^h}(x)}{\sum_{x \in E_j} w^{N^h}(x)}, \qquad j \in D'. \tag{2.1.1}$$

Let $d' = \|D'\|$. Then the vector $(\xi_1, \xi_2, \ldots, \xi_M)^t$ loses d' degrees of freedom due to (2.1.1) above. Because the sample means $\bar{y}^{N^h}(x)$ are normally distributed and independent due to the exogenousness of the sequence of decisions x^n, the vector $\bar{y}^{N^h}$ is multivariate normal. Denote by $\boldsymbol{\eta}$ an M-dimensional multivariate normal random variable with mean equal to the value of the true curve and with some known covariance matrix Σ, i.e., $\boldsymbol{\eta} \sim N_M(\boldsymbol{\mu}, \Sigma)$ where

$$\boldsymbol{\mu} = (\mu(1), \mu(2), \ldots, \mu(M))^t = (\mu_1, \mu_2, ..., \mu_M)^t.$$

Define

$$\mathscr{E} = \left\{ (\xi_1, \xi_2, \ldots, \xi_M)^t \in \mathbb{R}^M \,|\, h = \underset{f \text{ isotonic on } \boldsymbol{X}}{\arg\min} \sum_{x=1}^{M} \left[f(x) - \xi_x \right]^2 w^{N^h}(x), \forall x \in \boldsymbol{X} \right\},$$

then $dim(\mathscr{E}) = M - d' < M$, since $1 \leq d' \leq d \leq M$. That is, $\mathscr{E}$ lives in a lower-dimensional subspace of $\mathbb{R}^M$. In order to show $\mathbb{P}(\widetilde{B}) = 0$, it is sufficient to show that $\mathbb{P}(\boldsymbol{\eta} \in \mathscr{E}) = 0$. Letting λ denote the Lebesgue measure on a M-dimensional Euclidean space, it follows that $\lambda(\mathscr{E}) = 0$. Since the induced measure of $\boldsymbol{\eta}$ is absolutely continuous with respect to Lebesgue measure, $\lambda(\mathscr{E}) = 0$ implies

$\mathbb{P}(\boldsymbol{\eta} \in \mathscr{E}) = 0$. Hence $\mathbb{P}(\widetilde{B}) = \mathbb{P}(\boldsymbol{\eta} \in \mathscr{E}) = 0$. $\square$

Theorem 2.1.1 (*Finite Cardinality*)**.**

For a fixed sample path ω such that $\lim_{n\to\infty} \|\mathcal{P}^n(\omega)\| < \infty$, we have:

i) *The true curve $\mu \in \lim_{n\to\infty} \mathcal{P}^n(\omega)$. That is, there exists a large enough N such that for all $n \geq N$ we have $\mu \in \mathcal{P}^n(\omega)$.*

ii) *Let $A = \{\omega : \lim_{n\to\infty} \|\mathcal{P}^n(\omega)\| < \infty\}$. Then $\mathbb{P}(A) = 0$.*

Proof. For simplicity, we use $\mathcal{P}^n(\omega) = \mathcal{P}^n$ once the sample path ω is fixed. Because we assume the cardinality of the limiting set is finite, we define the finite set $\mathcal{P}^\infty = \left\{f_{[1]}, f_{[2]}, \ldots, f_{[\bar{L}]}\right\}$ as the limiting set of $\mathcal{P}^n$. That is, there exists N large enough such that

$$\mathcal{P}^n = \mathcal{P}^\infty = \left\{f_{[1]}, f_{[2]}, \ldots, f_{[\bar{L}]}\right\}$$

for all $n \geq N$.

We proceed by contradiction. Suppose $\mu \notin \mathcal{P}^\infty$, set $\delta = \min_{l\in I} \Xi\left(f_{[l]}, \mu\right)$ and $I = \left\{1, 2, \ldots \bar{L}\right\}$. Recall $\epsilon^n_{reg} = \gamma^n \cdot \epsilon_{reg}$ where ϵ_{reg} is a pre-specified fixed threshold and $\gamma \in (0, 1)$. Then set $N_\gamma(\delta, \epsilon_{reg}) = \ln(\delta/\epsilon_{reg}) / \ln(\gamma)$ such that, for $n > N_\gamma(\delta, \epsilon_{reg})$, we have

$$n > \frac{\ln(\delta/\epsilon_{reg})}{\ln(\gamma)}$$

$$
\begin{aligned}
\Longrightarrow \quad n \cdot \ln(\gamma) &< \ln(\delta/\epsilon_{reg}) \\
\Longrightarrow \quad \gamma^n &< \delta/\epsilon_{reg} \\
\Longrightarrow \quad \gamma^n \cdot \epsilon_{reg} &< \delta \\
\Longrightarrow \quad \sqrt{\epsilon_{reg}^n} &< \sqrt{\delta},
\end{aligned}
$$

which indicates $0 < \sqrt{\delta} - \sqrt{\epsilon_{reg}^n}$. Recall that f^n is the isotonic regression estimator of $\bar{y}^n(x)$ based on the measurement history up to the nth step. Moreover, $\bar{y}^n(x)$ is a strongly consistent estimator of $\mu(x)$ at each $x \in \boldsymbol{X}$. Thus by Theorem 1.2.5 and Theorem 1.2.6, f^n is also a strongly consistent estimator of $\mu(x)$ at each $x \in \boldsymbol{X}$. The strong consistency of f^n guarantees that for any ε satisfying $0 < \varepsilon < \sqrt{\delta} - \sqrt{\epsilon_{reg}^n}$, there exists N_ε such that for $n \geq N_\varepsilon$, we have $\sqrt{\Xi(f^n, \mu)} < \varepsilon$. For such ε and

$$
N_{\varepsilon,\gamma} = \max\{N_\gamma(\delta, \epsilon_{reg}), N_\varepsilon\},
$$

we have for all $l \in I$ and $n \geq N_{\varepsilon,\gamma}$:

$$
\begin{aligned}
\sqrt{\Xi\left(f^n, f_{[l]}\right)} &= \sqrt{\frac{1}{M}\sum_{x=1}^{M}\left[f^n(x) - f_{[l]}(x)\right]^2} \\
&= M^{-\frac{1}{2}}\|f^n - f_{[l]}\|_2 \\
&= M^{-\frac{1}{2}}\|(f^n - \mu) + \left(\mu - f_{[l]}\right)\|_2 \\
&\geq M^{-\frac{1}{2}}\{\|\mu - f_{[l]}\|_2 - \|f^n - \mu\|_2\} \\
&= \sqrt{\Xi\left(f_{[l]}, \mu\right)} - \sqrt{\Xi(f^n, \mu)}
\end{aligned}
$$

$$\geq \sqrt{\delta} - \varepsilon$$

$$> \sqrt{\epsilon_{reg}^n}.$$

Therefore, $\Xi\left(f^n, f_{[l]}\right) > \epsilon_{reg}^n$ for all $l \in I$ and $n \geq N_{\varepsilon,\gamma}$. That is, we find at least one function $f_{N_{\varepsilon,\gamma}} \neq f_{[l]}$ for all $l \in I$ such that $\min_{l \in I} \Xi\left(f_{N_{\varepsilon,\gamma}}, f_{[l]}\right) > \epsilon_{reg}^n$. Therefore we should add $f_{N_{\varepsilon,\gamma}}$ into the set $\mathcal{P}^\infty$. However, this is a contradiction to the fact that $\mathcal{P}^\infty$ is the finite and fixed limiting set of $\mathcal{P}^n$.

To show ii), define $B = \{\omega : \mu \in \mathcal{P}^\infty(\omega)\}$. By the result in i) we have for $\forall \omega \in A$, it implies that $\omega \in B$. Thus

$$\begin{aligned}
\mathbb{P}(A) &\leq \mathbb{P}(B) \\
&= \mathbb{P}\left(\mu(x) = f_{[l']}(x) \in \mathcal{P}^{N'}, \forall x \in \boldsymbol{X}, \text{ for some } l' \in I \text{ and } N' \text{ large enough}\right) \\
&= \mathbb{P}\left(f_{[l']}(x) \text{ takes exactly the same value as } \mu(x) \text{ for all } x \in \boldsymbol{X}\right) \qquad (2.1.2)
\end{aligned}$$

We need to show that (2.1.2) evaluates to zero. Equivalently, we can show that $\mathbb{P}(\boldsymbol{\eta} \in \mathscr{C}) = 0$ where $\boldsymbol{\eta}$ is as defined in Lemma 2.1.1 and $\mathscr{C}$ is defined as:

$$\mathscr{C} = \left\{ (\xi_1, \xi_2, \ldots, \xi_M)^t \in \mathbb{R}^M \,|\, \mu = \underset{f \text{ isotonic on } \boldsymbol{X}}{\arg\min} \sum_{x=1}^{M} \left[f(x) - \xi_x\right]^2 w^{N^h}(x), \forall x \in \boldsymbol{X} \right\}.$$

We see that $\mathscr{C}$ is just a special case of $\mathscr{E}$ where $h = f_{[l']}$, $\boldsymbol{X}' = \boldsymbol{X}$ and $d' = d = M$. Therefore, applying Lemma 2.1.1 implies $dim(\mathscr{C}) = M - M = 0$. It follows that $\mathbb{P}(\boldsymbol{\eta} \in \mathscr{C}) = 0$, and hence $\mathbb{P}(A) = 0$. □

Theorem 2.1.1 shows that, first, if the pool of curves were to have finite number

of curves, then it would contain the true curve; second, in fact the cardinality of the pool of curves is infinite almost surely. The following corollary states that we are able to add infinitely many curves into $\mathcal{P}$ that are arbitrarily close to the true curve.

Corollary 2.1.1 (*Infinite Cardinality*)**.**
For a fixed sample path ω such that $\lim_{n\to\infty}\|\mathcal{P}^n(\omega)\| = \infty$, $\left\{f^{N_q}\right\}_{q=1}^{\infty}$ is the sequence that are being added into the pool of curves. Then $f^{N_q}(x)$ converges to $\mu(x)$ almost surely on $\boldsymbol{X}$.

Proof. Again followed by Theorem 1.2.6, the sequence of isotonic regression estimator $\{f^n\}_{n=1}^{\infty}$ converge to $\mu(x)$ almost surely and uniformly (since $\boldsymbol{X}$ is a finite set) on $\boldsymbol{X}$ as $n \to \infty$. The sequence $\left\{f^{N_q}\right\}_{q=1}^{\infty}$ is just a subsequence of $\{f^n\}_{n=1}^{\infty}$. Therefore, the strong consistency follows trivially from the strong consistency of $\{f^n\}_{n=1}^{\infty}$. □

2.2 Resampling from $\mathcal{P}^n$ into $\mathcal{C}^n$

Motivation

According to the version of the resampling algorithm in Huang et al. (2018), the criterion for adding functions from $\mathcal{P}^n$ to $\mathcal{C}^n$ is that we pick a certain number of functions that have the smallest MSE with respect to the measurement history up to the moment this function is added into $\mathcal{P}^n$. We have shown that, almost surely, the sequence of functions being added into $\mathcal{P}^n$ converges to the true curve μ. Then the criterion requires that the functions that are getting closer and closer to μ should

also have smaller MSE with respect to the most recent measurement history. That is, if we choose an arbitrary ε and $h, g \in \mathcal{P}^n$ such that $\Xi(h, \mu) < \varepsilon < \Xi(g, \mu)$, and $N^g < N^h \leq n$, then we hope to see that this implies $\widehat{\Xi}\left(h, \widehat{\boldsymbol{H}}^n\right) < \widehat{\Xi}\left(g, \widehat{\boldsymbol{H}}^n\right)$. In other words, h is a better candidate than g in the sense of MSE based on the latest updated measurement history $\widehat{\boldsymbol{H}}^n$. Furthermore, to ensure that we add "good" functions from $\mathcal{P}$ to $\mathcal{C}$, we need the curves with the smallest MSE also to be within ε of the true curve μ.

Modification to the Resampling Algorithm

We first modify the MSE used at Step 2 of the Resampling Procedure in Huang et al. (2018, Algorithm 2). In the nth experiment when calculating the MSE of a function h in $\mathcal{P}^n$ based on the measurement history up to time n, instead of using

$$\widehat{\Xi}\left(h, \widehat{\boldsymbol{H}}^n\right) = \frac{1}{n} \sum_{x \in \boldsymbol{X}} \sum_{j=1}^{w^n(x)} [\hat{y}_j(x) - h(x)]^2,$$

we now refer to the MSE between h and the sample mean with respect to $\widehat{\boldsymbol{H}}^n$:

$$\Xi\left(h, \bar{y}^n\right) = \frac{1}{M} \sum_{x \in \boldsymbol{X}} [h(x) - \bar{y}^n(x)]^2.$$

This modification is reflected in Step 3 of Algorithm 2. In addition, recall that in each round we denote by K^{del} the number of deleted curves and by K the number of curves with the smallest MSE where $K > K^{del}$.

Theorem 2.2.1 (Functions with Smallest MSE)**.**

Define $f^n_{N_{[1]}}, f^n_{N_{[2]}}, ..., f^n_{N_{[K]}} \in \mathcal{P}^n$ *as the functions with the* K *smallest* $\Xi\left(f, \bar{y}^n\right), f \in \mathcal{P}^n$*, and sort them in ascending order, i.e.,* $\Xi\left(f^n_{N_{[i]}}, \bar{y}^n\right) < \Xi\left(f^n_{N_{[j]}}, \bar{y}^n\right)$ *for* $i < j$*. Then the following holds almost surely:*

i) For any $\varepsilon > 0$ *and any* $g = f^{N^g} \in \mathcal{P}^n$ *with* $\Xi(g, \mu) > 9\varepsilon/4$*, there exists a time* $T > N^g$ *such that for all* N_s *and* N_t *such that* $N_s > N_t > T > N^g$ *we have* $\Xi(g, \bar{y}^{N_s}) > \varepsilon > \Xi(f^{N_t}, \bar{y}^{N_s})$ *and* $\Xi(f^{N_t}, \mu) < \varepsilon$*.*

ii) For the previous ε *and time* T*, we have* $\Xi\left(f^n_{N_{[i]}}, \mu\right) < \varepsilon, i = 1, 2, ..., K$*, for* $\forall n > T$*.*

Proof. Recall that $l(n)$ is the number of curves in $\mathcal{P}^n$ and $N_{l(n)}$ indexes the most recent added element of $\mathcal{P}^n$. It is always true that either $N_{l(n)} = n$ or $N_{l(n)} < n$. One should keep in mind that it is possible to have

$$\mathcal{P}^{N_{l(n)}} = \mathcal{P}^{N_{l(n)}+1} = ... = \mathcal{P}^{N_{l(n)}+q'} = \mathcal{P}^n$$

for some $q' \in \boldsymbol{Z}^+$, which corresponds to the situation that there is no function being added at the $(N_{l(n)} + 1)$th, ..., $(N_{l(n)} + q')$th experiment. Therefore we first consider the case when n is the very first time that the pool of curves contains all of the functions $f^{N_1}, f^{N_2}, ..., f^{N_{l(n)}}$. That is: the case with $N_{l(n)} = n$.

For statement i), set $d_0 = \Xi(g, \mu)$. For any $\varepsilon < 4d_0/9$, there is $\Xi(g, \mu) = d_0 > 9\varepsilon/4$, or equivalently, $\sqrt{\Xi(g, \mu)} > 3\sqrt{\varepsilon}/2$. For this particular ε, because the sequence $\left\{f^{N_q}\right\}_{q=1}^{\infty}$ also converges to the true curve μ almost surely, there exists a time $T_1 > N^g$ such that for any time $N_s > N_t > T_1 > N^g$, we have $\Xi(f^{N_t}, \mu) < \varepsilon/4$ and

$\Xi(f^{N_s}, \mu) < \varepsilon/4$. Note that the sample mean $\bar{y}^n(x)$ is also an M-dimensional discrete function. Thus by the strong law of large numbers (SLLN), assuming $w^n(x) \to \infty$ as $n \to \infty$, we have $\bar{y}^n(x) \xrightarrow{a.s.} \mu(x), \forall x \in \boldsymbol{X}$. Therefore there exists a time $T_2 > N^g$ such that for any time $N_s > N_t > T_2 > N^g$, we have $\Xi(\bar{y}^{N_t}, \mu) < \varepsilon/4$ and $\Xi(\bar{y}^{N_s}, \mu) < \varepsilon/4$. Equivalently, $-\sqrt{\Xi(\bar{y}^{N_t}, \mu)} > -\sqrt{\varepsilon}/2$ and $-\sqrt{\Xi(\bar{y}^{N_s}, \mu)} > -\sqrt{\varepsilon}/2$. Thus, there exists $T = \max(T_1, T_2)$ such that for any time $N_s > N_t > T > N^g$, we have the following two inequalities based on the triangle inequality of an L_2-norm:

$$\begin{aligned}
\sqrt{\Xi(g, \bar{y}^{N_s})} &\geq \sqrt{\Xi(g, \mu)} - \sqrt{\Xi(\bar{y}^{N_s}, \mu)} \\
&> \frac{3\sqrt{\varepsilon}}{2} + \left(-\frac{\sqrt{\varepsilon}}{2}\right) \\
&= \sqrt{\varepsilon}
\end{aligned}$$

indicating $\Xi(g, \bar{y}^{N_s}) > \varepsilon$ and at the same time

$$\begin{aligned}
0 &\leq \sqrt{\Xi(f^{N_t}, \bar{y}^{N_s})} \\
&\leq \sqrt{\Xi(f^{N_t}, \mu)} + \sqrt{\Xi(\bar{y}^{N_s}, \mu)} \\
&< \frac{\sqrt{\varepsilon}}{2} + \frac{\sqrt{\varepsilon}}{2} \\
&= \sqrt{\varepsilon}
\end{aligned}$$

indicating $\Xi(f^{N_t}, \bar{y}^{N_s}) < \varepsilon$.

Therefore we have shown that there exists a time $T > N^g$ such that for all $N_s > N_t > T > N^g$ we have $\Xi(g, \bar{y}^{N_s}) > \varepsilon > \Xi(f^{N_t}, \bar{y}^{N_s})$ and $\Xi(f^{N_t}, \mu) < \varepsilon/4 < \varepsilon$. Statement i) guarantees that any function that is sufficiently far from the true μ

will eventually be a worse candidate than functions from the large pool.

For statement ii), it is shown in statement i) that $\Xi(g, \bar{y}^{N_s}) > \varepsilon > \Xi(f^{N_t}, \bar{y}^{N_s})$ for any $N_s > N_t > T > N^g$ and function $g = f^{N^g}$ with $\Xi(g, \mu) > \varepsilon$. This indicates any function g that is far enough from the true μ will always have a bigger MSE with respect to the latest sample mean than those functions that can be arbitrarily close to the true μ. Remember that we assume $N_{l(n)} = n$, so consider $N_{l(n)} = n = N_s > N_t > T$, meaning that the current pool of curves is

$$\mathcal{P}^n = \mathcal{P}^{N_s} = \{f^{N_1}, f^{N_2}, ..., f^T, ..., f^{N_t}, ..., f^{N_s}\}.$$

Because we assumed that $w^n(x) \to \infty$, we may then choose $N_s > T$ large enough such that at least K terms are added between times T and N_s. Following that we have two cases:

(1) $f^{N_s}_{N_{[i]}} \in \left\{f^T, ..., f^{N_t}, ..., f^{N_s}\right\} \backslash \{f^T\}$ and $\Xi\left(f^{N_s}_{N_{[i]}}, \bar{y}^{N_s}\right) < \varepsilon$ for all $i = 1, 2, ..., K$. That is, all of the K functions with the smallest MSE are added after time T.

(2) There is at least one such function, say $f^{N_s}_{N_{[j]}}, j \in \{1, 2, ..., K\}$, satisfies $f^{N_s}_{N_{[j]}} \in \left\{f^{N_1}, f^{N_2}, ..., f^T\right\}$ and $\Xi\left(f^{N_s}_{N_{[j]}}, \bar{y}^{N_s}\right) < \varepsilon$. That is at least one of the K functions with the smallest MSE is added before time T.

For case (1), there exists N_{t_i} with $T < N_{t_i} \leq N_s, i = 1, 2, ..., K$ such that $\Xi\left(f^{N_{t_i}}, \bar{y}^{N_s}\right)$ is the ith smallest. Therefore, setting $f^{N_s}_{N_{[i]}} = f^{N_{t_i}}, i = 1, 2, ..., K$, we find that all of the K functions $f^{N_s}_{N_{[i]}}$ in $\mathcal{P}^{N_s} = \mathcal{P}^n$ satisfy $\Xi\left(f^n_{N_{[i]}}, \mu\right) < \varepsilon, i = 1, 2, ..., S$. For case (2), we claim that such $f^{N_s}_{N_{[j]}}$ satisfies $\Xi\left(f^{N_s}_{N_{[j]}}, \mu\right) < \varepsilon$. Otherwise, if $\Xi\left(f^{N_s}_{N_{[j]}}, \mu\right) > \varepsilon$,

from statement i) it must be true that $\Xi\left(f_{N_{[j]}}^{N_s}, \bar{y}^{N_s}\right) > \varepsilon$. This is a contradiction to the condition of $f_{N_{[j]}}^{N_s}$ in case (2).

We now consider the second situation when $N_{l(n)} < n$. For simplicity, set $N_{l(n)} = N_t < n$. So N_t is now the last function being added into $\mathcal{P}^n$, but the functions $f^{N_t+1}, f^{N_t+2}, ..., f^n$ are not in $\mathcal{P}^n$. By a similar argument as in statement i), for all $\varepsilon > 0$ we are able to find a large enough time T' satisfying:

- A function $f^{T'}$ is added into $\mathcal{P}^n$ at time T';
- $T' < N_t$, i.e., T' is not the last time we add a function into $\mathcal{P}^n$;
- For $g = f^{N^g}$ and all time points $n > m > T' > N^g$, we have:

  1) f^m can be arbitrarily close to μ, that is,

$$\Xi(f^m, \mu) < \frac{\varepsilon}{4} < \varepsilon$$

  2) f^m can be arbitrarily close to the sample mean, that is,

$$\begin{aligned} 0 &\leq \sqrt{\Xi(f^m, \bar{y}^n)} \\ &\leq \sqrt{\Xi(f^m, \mu)} + \sqrt{\Xi(\bar{y}^n, \mu)} \\ &< \frac{\sqrt{\varepsilon}}{2} + \frac{\sqrt{\varepsilon}}{2} \\ &= \sqrt{\varepsilon} \end{aligned}$$

  which implies $\Xi(f^m, \bar{y}^n) < \varepsilon$.

3) g is far enough from the sample mean, that is,

$$\begin{aligned}\sqrt{\Xi(g,\bar{y}^n)} &\geq \sqrt{\Xi(g,\mu)} - \sqrt{\Xi(\bar{y}^n,\mu)} \\ &> \frac{3\sqrt{\varepsilon}}{2} + (-\frac{\sqrt{\varepsilon}}{2}) \\ &= \sqrt{\varepsilon}\end{aligned}$$

which implies $\Xi(g,\bar{y}^n) > \varepsilon$.

Therefore we have shown that there exists another time $T' > N^g$ such that for all $n > m > T' > N^g$ we have $\Xi(g,\bar{y}^n) > \varepsilon > \Xi(f^m,\bar{y}^n)$ and $\Xi(f^m,\mu) < \varepsilon$.

In this case, however, not all such f^m are in $\mathcal{P}^n$ because they may not all be added during the regeneration process. That is, assuming $N_{l(n)} = N_t$, we write

$$\mathcal{P}^n = \{f^{N_1}, f^{N_2}, ..., f^{N_u}, f^{T'}, f^{N_{u+1}}, ..., f^{N_t}\},$$

where $N_u < T' < N_{u+1}$, then $f^m \in \mathcal{P}^n$ only for $m \in \{N_{u+1}, ..., N_t\}$ and $f^m \notin \mathcal{P}^n$ for $m \in \{N_t + 1, N_t + 2, ..., n\}$. This means that the K isotonic regression estimators $f^n_{N_{[i]}}$ with the smallest $\Xi\left(f^n_{N_{[i]}}, \bar{y}^n\right)$ may not be added into $\mathcal{P}^n$. In fact, they could all be in the set $\left\{f^{N_t+1}, f^{N_t+2}, ..., f^n\right\}$. Nevertheless, we are still able to identify the first K smallest $\Xi(f^m,\bar{y}^n)$ where $f^m \in \left\{f^{N_{u+1}}, ..., f^{N_t}\right\} \subseteq \mathcal{P}^n$, such that all these f^m still satisfy $0 \leq \Xi(f^m,\bar{y}^n) < \varepsilon$ and $\Xi(f^m,\mu) < \varepsilon$.

Again, pick N_t large enough such that the cardinality of the set $\left\{f^{N_{u+1}}, ..., f^{N_t}\right\}$ satisfies $\| \left\{f^{N_{u+1}}, ..., f^{N_t}\right\} \| \geq K$. Then by similar argument as in the previous case (1) and case (2), there exists time T' and K times $N_{t_i}, i = 1, 2, ..., K$, with either

$N_{t_i} > T'$ or $N_{t_i} < T'$, such that $\Xi\left(f^{N_{t_i}}, \bar{y}^n\right) < \varepsilon$ is the ith smallest among all functions in $\mathcal{P}^n$ and also $\Xi\left(f^{N_{t_i}}, \mu\right) < \varepsilon$, $i = 1, 2, ..., K$. Hence, setting $f^n_{N_{[i]}} = f^{N_{t_i}}, i = 1, 2, ..., K$, we have found all of the K functions $f^n_{N_{[i]}}$ in $\mathcal{P}^n$ and shown that they all satisfy $\Xi\left(f^n_{N_{[i]}}, \mu\right) < \varepsilon, i = 1, 2, ..., K$.

This conclude the proof of the second statement and of Theorem 2.2.1. □

Remember in Algorithm 2, the resampling procedure can be triggered multiple times within each iteration, we then have the following corollary with regard to the sequence of the last function being resampled from $\mathcal{C}$ to $\mathcal{P}$.

Corollary 2.2.1. *Consider a subsequence* $\left\{g^{N_r}\right\}_{r=1}^{\infty} \subseteq \left\{f^{N_q}\right\}_{q=1}^{\infty}$ *where* g^{N_r} *is the last function being resampled into the candidate set from the pool of curves within each iteration. Then* $g^{N_r}(x) \xrightarrow{a.s.} \mu(x), \forall x \in \boldsymbol{X}$, *as* $r \to \infty$.

Proof. Remember that whenever the resampling is triggered, we only sample from the K functions in the pool of curves that have the smallest MSE with respect to the latest sample mean. Therefore, by Theorem 2.2.1, $\forall \varepsilon > 0$, there exists $T'' = max(T, T')$ such that for $f^n_{N_{[1]}}, f^n_{N_{[2]}}, ..., f^n_{N_{[K]}} \in \mathcal{P}^n$, we have $\Xi\left(f^n_{N_{[i]}}, \mu\right) < \varepsilon, i = 1, 2, ..., K$, for $\forall n > T''$. Therefore set $f^n_{N_{max}} = \arg\max_{f^n_{N_{[i]}}}\{\Xi\left(f^n_{N_{[i]}}, \mu\right), i = 1, 2, ..., K\}$, then $0 \leq \Xi(f^n_{N_{max}}, \mu) < \varepsilon$ implies $\Xi(f^n_{N_{max}}, \mu) \xrightarrow{a.s.} 0$. Since r is related to n, set $r = r(n)$ and assume $r(n) \to \infty$ as $n \to \infty$, we have $g^{N_r} = g^{N_{r(n)}} \in \mathcal{P}^n$ for some n. Then $0 \leq \Xi(g^{N_r}, \mu) = \Xi(g^{N_{r(n)}}, \mu) \leq \Xi(f^n_{N_{max}}, \mu)$ for some n. Thus it follows that $\Xi(g^{N_r}, \mu) = \Xi\left(g^{N_{r(n)}}, \mu\right) \xrightarrow{a.s.} 0$. Hence $g^{N_r}(x) \xrightarrow{a.s.} \mu(x), \forall x \in \boldsymbol{X}$, as $r \to \infty$. □

2.3 Improvement of the Candidate Set

2.3.1 Candidate Categorization and a Potential Issue

Recall that in Step 2 of our resampling procedure any candidate $f_l^n \in \mathcal{C}^n$ with $p_l^n < \epsilon^{res}$, where ϵ^{res} is the pre-specified threshold, will be removed and replaced by an updated curve f_l^{n+1}. If no such f_l^n satisfies this condition, we then remove the function $f_{l'}^n$ where $p_{l'}^n = \arg\min_{l \in \mathcal{C}^n} p_l^n$. This makes sure that we are guaranteed to be able to remove at least one function in each experiment so that we can always replace functions in $\mathcal{C}^n$ from the ones in $\mathcal{P}^n$ at each iteration. This is crucial to our analysis because, due to the features of $\mathcal{P}^n$ that have been explored previously, this criterion allows us to keep replacing "bad" functions in $\mathcal{C}$ with "good" functions as defined below:

Definition 2.3.1 (Good and Bad Functions)**.**
Let us state the criterion of categorizing a "good" function and a "bad" function with respect to $\varepsilon > 0$:

1) A function g is called a *bad*, or *inferior*, function if for a fixed $\varepsilon > 0$, $\Xi(g, \mu) > \varepsilon$.

2) A function g is called a *good*, or *superior*, function if for a fixed $\varepsilon > 0$, $\Xi(g, \mu) \leq \varepsilon$.

The general interpretation of a bad function g is that for a given small value ε, among the candidates in $\mathcal{C}$, g is too far away from the true curve compared to other candidates. Thus g is not considered as a promising candidate and we wish

to replace it by a better function that is closer to the true curve, and hence the word "good" is used to describe the superiority of such a function over the original inferior one that needs to be replaced.

Potential Issue

For notational simplicity in this subsection, we consider an arbitrary element, written as $g_l = f_l^n$, in the candidate set $\mathcal{C}^n$. The joint likelihood of g_l based on the measurement history up to n is :

$$\mathcal{L}\left(\widehat{\boldsymbol{H}}^n|\mu = g_l\right) = \left(\sqrt{2\pi}\sigma\right)^{-n} \cdot \prod_{i=0}^{n-1} \exp\left\{-\frac{[\hat{y}^{i+1} - g_l(x^i)]^2}{2\sigma^2}\right\}, x^i \in \boldsymbol{X} = \{1, 2, ..., M\}.$$

Rewrite the MSE between g_l and $\widehat{\boldsymbol{H}}^n$ as:

$$\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^n\right) \quad = \quad \frac{1}{n}\sum_{i=0}^{n-1}[\hat{y}^{i+1} - g_l(x^i)]^2, x^i \in \boldsymbol{X}.$$

We can see that the relation between $\mathcal{L}\left(\widehat{\boldsymbol{H}}^n|\mu = g_l\right)$ and $\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^n\right)$ is:

$$\mathcal{L}\left(\widehat{\boldsymbol{H}}^n|\mu = g_l\right) = \left(\sqrt{2\pi}\sigma\right)^{-n} \cdot \exp\left\{-\frac{n}{2\sigma^2} \cdot \widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^n\right)\right\}.$$

Using this relation and equation (1.3.2), we may rewrite p_l^n in terms of MSE as:

$$p_l^n = \frac{\exp\left\{-(n/(2\sigma^2)) \cdot \widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^n\right)\right\}}{\sum_{l'=1}^{L} \exp\left\{-(n/(2\sigma^2)) \cdot \widehat{\Xi}\left(g_{l'}, \widehat{\boldsymbol{H}}^n\right)\right\}}.$$

We can see that, first, if $\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^n\right)$ has a relatively large value, it implies that g_l is likely (but not definitely) to be an inferior candidate. Second, a relatively large $\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^n\right)$ tends to give a small p_l^n, which could be smaller than the threshold ϵ^{res} leading to the removal of g_l. Nonetheless, it is possible that a bad function g_l happens to have small $\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^n\right)$ which might induce $p_l^n > \epsilon^{res}$, causing the potential problem of not being able to remove the inferior g_l. We would like to eliminate this situation by showing that there exists a time $\widetilde{T}$ and a certain subsequence $\{n_k\}_{k=1}^{\infty}$, along which $\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^{n_k}\right)$ is greater than $\widehat{\Xi}\left(g^{n_k}, \widehat{\boldsymbol{H}}^{n_k}\right)$ after time $\widetilde{T}$. Asymptotically, in that case there will be at least one term in the denominator of $p_l^{n_k}$ of lower order compared to the numerator, which eventually would result in $p_l^{n_k} < \epsilon^{res}$ and the removal of g_l.

2.3.2 Limits of the Ratio $w^n(x)/n$

We start by clarifying some of the notations. We keep $\{f^n\}_{n=1}^{\infty}$ as the sequence of isotonic regression estimators based on $\widehat{\boldsymbol{H}}^n$, and still let $\left\{f^{N_q}\right\}_{q=1}^{\infty} \subseteq \{f^n\}_{n=1}^{\infty}$ be the subsequence that is added into the pool of curves. Then there exists another subsequence (as mentioned in the Corollary 2.2.1), denoted by $\left\{g^{N_r}\right\}_{r=1}^{\infty} \subseteq \left\{f^{N_q}\right\}_{q=1}^{\infty}$ such that g^{N_r} is the last function being resampled into the candidate set out of the K functions in the pool of curves that have the smallest MSE. In fact, the index N_r coincides with n because in each iteration we resample at least one new function from $\mathcal{P}$ to $\mathcal{C}$ due to our modification to the resampling procedure. In other words, for each time index n, there exists a q such that $g^n = f^{N_q}$. Thus, we may write

$\left\{g^{N_r}\right\}_{r=1}^{\infty} = \{g^n\}_{n=1}^{\infty}$, and we use $\{g^n\}_{n=1}^{\infty}$ for the rest of this thesis. Now we are ready to study the limit of the ratio.

Recall that $w^n(x), x \in \boldsymbol{X} = \{1, 2, ..., M\}$ is the number of measurements taken on x up to time n and $\sum_{x\in\boldsymbol{X}} w^n(x) = n$. Each $w^n(x)/n$ is bounded within the interval $[0, 1]$ and $\sum_{x\in\boldsymbol{X}}(w^n(x)/n) = 1$. Therefore, there exists a subsequence $\{g^{n_k}\}_{k=1}^{\infty} \subseteq \{g^n\}_{n=1}^{\infty}$ along which the ratio converges uniformly to a limiting function $\rho(x)$. That is:

$$\lim_{k\to\infty} \frac{w^{n_k}(x)}{n_k} = \rho(x), \forall x \in \boldsymbol{X}.$$

Notice that $\rho(x) \in [0, 1]$ and $\sum_{x\in\boldsymbol{X}} \rho(x) = 1$. Therefore there exists at least one $x_0 \in \boldsymbol{X}$ such that $\rho(x_0) > 0$.

2.3.3 Limiting Behavior of Two Mean Squared Errors

In this section, we consider an inferior function g_l in the candidate set and the previous subsequence $\{g^{n_k}\}_{k=1}^{\infty}$ mentioned in Section 2.3.2. We need to study the limiting behavior of $\widehat{\Xi}\left(g^{n_k}, \widehat{\boldsymbol{H}}^{n_k}\right)$ and $\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^{n_k}\right)$ as mentioned in Section 2.3.1.

2.3.3.1 The Limit of MSE with g^{n_k}

Let us decompose the MSE $\widehat{\Xi}\left(g^{n_k}, \widehat{\boldsymbol{H}}^{n_k}\right)$ as:

$$\begin{aligned}
\widehat{\Xi}\left(g^{n_k}, \widehat{\boldsymbol{H}}^{n_k}\right) &= \frac{1}{n_k} \sum_{x\in\boldsymbol{X}} \sum_{j=1}^{w^{n_k}(x)} \left[\hat{y}_j(x) - g^{n_k}(x)\right]^2 \\
&= \frac{1}{n_k} \sum_{x\in\boldsymbol{X}} \left[\overline{y}^{n_k}(x) - g^{n_k}(x)\right]^2 w^{n_k}(x) + \frac{1}{n_k} \sum_{x\in\boldsymbol{X}} \sum_{j=1}^{w^{n_k}(x)} \left[\hat{y}_j(x) - \overline{y}^{n_k}(x)\right]^2
\end{aligned}$$

$$+\frac{1}{n_k}\sum_{x\in\boldsymbol{X}}\sum_{j=1}^{w^{n_k}(x)}\left[\overline{y}^{n_k}(x)-g^{n_k}(x)\right]\left[\hat{y}_j(x)-\overline{y}^{n_k}(x)\right]$$

$$:= S_1^{n_k}+S_2^{n_k}$$

where the cross term

$$\begin{aligned}
&\frac{1}{n_k}\sum_{x\in\boldsymbol{X}}\sum_{j=1}^{w^{n_k}(x)}\left[\overline{y}^{n_k}(x)-g^{n_k}(x)\right]\left[\hat{y}_j(x)-\overline{y}^{n_k}(x)\right]\\
&=\frac{1}{n_k}\sum_{x\in\boldsymbol{X}}\left[\overline{y}^{n_k}(x)-g^{n_k}(x)\right]\sum_{j=1}^{w^{n_k}(x)}\left[\hat{y}_j(x)-\overline{y}^{n_k}(x)\right]\\
&=\frac{1}{n_k}\sum_{x\in\boldsymbol{X}}\left[\overline{y}^{n_k}(x)-g^{n_k}(x)\right]\cdot 0\\
&=0.
\end{aligned}$$

First consider $S_2^{n_k}$. Denote by $s^2_{n_k}(x)$ the sample variance for the observations taken on x. Then:

$$\begin{aligned}
S_2^{n_k}&=\frac{1}{n_k}\sum_{x\in\boldsymbol{X}}\sum_{j=1}^{w^{n_k}(x)}\left[\hat{y}_j(x)-\overline{y}^{n_k}(x)\right]^2\\
&=\sum_{x\in\boldsymbol{X}}\frac{w^{n_k}(x)}{n_k}\cdot\frac{w^{n_k}(x)-1}{w^{n_k}(x)}\left\{\frac{1}{w^{n_k}(x)-1}\sum_{j=1}^{w^{n_k}(x)}\left[\hat{y}_j(x)-\overline{y}^{n_k}(x)\right]^2\right\}\\
&=\sum_{x\in\boldsymbol{X}}\frac{w^{n_k}(x)}{n_k}\cdot\left[\frac{w^{n_k}(x)-1}{w^{n_k}(x)}\cdot s^2_{n_k}(x)\right]. \qquad (2.3.1)
\end{aligned}$$

For the right hand side of (2.3.1), as $k\to\infty$ we have $s^2_{n_k}(x)\xrightarrow{a.s.}\sigma^2(x)$ and $(w^{n_k}(x)-1)/w^{n_k}(x)\xrightarrow{a.s.}1$ by the strong law of large numbers. Therefore

$$\frac{w^{n_k}(x)-1}{w^{n_k}(x)}\cdot s^2_{n_k}(x)\xrightarrow{a.s.}\sigma^2(x)$$

as $k \to \infty$. Thus the right hand side of (2.3.1) converges to the weighted sum of group sample variances. Since we assume uniform group variance with $\sigma^2(x) = \sigma^2, \forall x \in \boldsymbol{X}$, then the right hand side of equation (2.3.1) is σ^2. That is,

$$\begin{aligned} S_2^{n_k} &= \sum_{x \in \boldsymbol{X}} \frac{w^{n_k}(x)}{n_k} \cdot \left[\frac{w^{n_k}(x) - 1}{w^{n_k}(x)} \cdot s_{n_k}^2(x) \right] \\ &\xrightarrow{a.s.} \sum_{x \in \boldsymbol{X}} \rho(x) \cdot \sigma^2(x) \\ &= \sigma^2 \sum_{x \in \boldsymbol{X}} \rho(x) \\ &= \sigma^2. \end{aligned}$$

For $S_1^{n_k}$, first by Corollary 2.2.1, we have $g^{n_k} \xrightarrow{a.s.} \mu(x), \forall x \in \boldsymbol{X}$. Together with the strong consistency of the sample mean $\overline{y}^{n_k}(x)$, for any $\varepsilon > 0$, there exists $T^* > T''$ such that when $n_k > T^*$ we have $\Xi\left(\overline{y}^{n_k}(x), \mu(x)\right) < \varepsilon/(4M)$ and $\Xi\left(g^{n_k}, \mu(x)\right) < \varepsilon/(4M)$. Then by the triangle inequality of L_2-norm, we have

$$\begin{aligned} \sqrt{\Xi\left(\overline{y}^{n_k}(x), g^{n_k}(x)\right)} &< \sqrt{\Xi\left(\overline{y}^{n_k}(x), \mu(x)\right)} + \sqrt{\Xi\left(g^{n_k}, \mu(x)\right)} \\ &< \frac{1}{2}\sqrt{\frac{\varepsilon}{M}} + \frac{1}{2}\sqrt{\frac{\varepsilon}{M}} \\ &= \sqrt{\frac{\varepsilon}{M}} \end{aligned}$$

$$\implies \quad \Xi\left(\overline{y}^{n_k}(x), g^{n_k}(x)\right) \quad < \quad \varepsilon/M \ \cdot$$

Thus,

$$S_1^{n_k} = \sum_{x \in \boldsymbol{X}} \left[\overline{y}^{n_k}(x) - g^{n_k}(x)\right]^2 \frac{w^{n_k}(x)}{n_k}$$

$$
\begin{aligned}
&< \sum_{x \in \boldsymbol{X}} \left[\overline{y}^{n_k}(x) - g^{n_k}(x)\right]^2 \\
&= M \cdot \Xi\left(\overline{y}^{n_k}(x), g^{n_k}(x)\right) \\
&< M \cdot \frac{\varepsilon}{M} \\
&= \varepsilon.
\end{aligned}
$$

That is, $S_1^{n_k}$ converges to zero almost surely. Hence, we have

$$
\begin{aligned}
\lim_{k \to \infty} \widehat{\Xi}\left(g^{n_k}, \widehat{\boldsymbol{H}}^{n_k}\right) &= \lim_{k \to \infty} \left(S_1^{n_k} + S_2^{n_k}\right) \\
&= \lim_{k \to \infty} S_1^{n_k} + \lim_{k \to \infty} S_2^{n_k} \\
&= 0 + \sigma^2 \\
&= \sigma^2.
\end{aligned}
$$

2.3.3.2 The Limit of MSE with g_l

Let us again decompose the MSE $\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^{n_k}\right)$ as:

$$
\begin{aligned}
\widehat{\Xi}\left(g_l, \widehat{\boldsymbol{H}}^{n_k}\right) &= \frac{1}{n_k} \sum_{x \in \boldsymbol{X}} \sum_{j=1}^{w^{n_k}(x)} \left[\hat{y}_j(x) - g_l(x)\right]^2 \\
&= \frac{1}{n_k} \sum_{x \in \boldsymbol{X}} \left[\overline{y}^{n_k}(x) - g_l(x)\right]^2 w^{n_k}(x) + \frac{1}{n_k} \sum_{x \in \boldsymbol{X}} \sum_{j=1}^{w^{n_k}(x)} \left[\hat{y}_j(x) - \overline{y}^{n_k}(x)\right]^2 \\
&:= S_{1,l}^{n_k} + S_2^{n_k}
\end{aligned}
$$

then we still have $S_2^{n_k} \xrightarrow{a.s.} \sigma^2$ as $k \to \infty$.

For $S_{1,l}^{n_k}$, we have:

$$\begin{aligned} S_{1,l}^{n_k} &= \frac{1}{n_k} \sum_{x \in \boldsymbol{X}} \left[\overline{y}^{n_k}(x) - g_l(x)\right]^2 w^{n_k}(x) \\ &= \sum_{x \in \boldsymbol{X}} \left[\overline{y}^{n_k}(x) - g_l(x)\right]^2 \cdot \frac{w^{n_k}(x)}{n_k} \end{aligned} \tag{2.3.2}$$

For the right hand side of equation (2.3.2), by the discussion in Section 2.3.2 we have:

1) $\lim_{k \to \infty} w^{n_k}(x)/n_k = \rho(x) \in [0, 1]$.

2) There exists at least one $x_0 \in \boldsymbol{X}$ such that $\rho(x_0) > 0$.

3) The SLLN gives $\overline{y}^{n_k}(x) \xrightarrow{a.s.} \mu(x), \forall x \in \boldsymbol{X}$, which implies

$$\left[\overline{y}^{n_k}(x) - g_l(x)\right]^2 \xrightarrow{a.s.} \left[\mu(x) - g_l(x)\right]^2, \forall x \in \boldsymbol{X}.$$

Since $S_{1,l}^{n_k}$ is a finite sum, the limit of $S_{1,l}^{n_k}$ exists. So, we can define

$$L_1^{'} := \lim_{k \to \infty} S_{1,l}^{n_k} = \sum_{x \in \boldsymbol{X}} \left[\mu(x) - g_l(x)\right]^2 \cdot \rho(x).$$

Due to Lemma 2.1.1, we know that for an isotonic regression estimator $g_l(x)$, $\mathbb{P}(g_l(x) = \mu(x), \forall x \in \boldsymbol{X}) = 0$. Together with the above properties 1) and 2) of $\rho(x)$ also stated in Section 2.3.2, we know that $L_1^{'} > 0$ almost surely.

Hence, we have

$$\begin{aligned}\lim_{k\to\infty}\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right) &= \lim_{k\to\infty}\left(S_{1,l}^{n_k}+S_2^{n_k}\right)\\ &= L_1^{'}+\sigma^2\\ &> \sigma^2.\end{aligned}$$

2.3.4 Removing Inferior Functions

Theorem 2.3.1 (Removal of Bad Functions)**.**

Suppose that an inferior function g_l with respect to ε is in $\mathcal{C}^n$ for some n. Then there exists a time N greater than n at which g_l will be removed from $\mathcal{C}^N$ and replaced by a good (superior) function.

Proof. Because $\lim_{k\to\infty}\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)=\sigma^2$ and $\lim_{k\to\infty}\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)=L_1^{'}+\sigma^2>\sigma^2$, for any $\varepsilon<\frac{L_1^{'}}{2}$, there exists $T_1>T^*$ such that for $k>T_1$, $\left|\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)-\sigma^2\right|<\varepsilon$. There also exists $T_1^{\prime}>0$ such that for $k>T_1^{\prime}$, $\left|\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)-\left(L_1^{'}+\sigma^2\right)\right|<\varepsilon$. Therefore, taking $T_l=\max(T_1,T_1^{\prime})$, for $k>T_l$ we have:

$$\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)<\sigma^2+\varepsilon<\sigma^2+\frac{L_1^{'}}{2}<\sigma^2+L_1^{'}-\varepsilon<\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right).$$

That is,

$$\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)-\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)>0$$

for $k>T_l$. We write $\mathcal{C}^{n_k}\backslash\{g_l,g^{n_k}\}$ as the n_kth candidate set with g_l and g^{n_k} deleted,

and let

$$S_{\mathcal{C}^{n_k}\setminus\{g_l,g^{n_k}\}} = \sum_{g_{l'}\in\mathcal{C}^{n_k}\setminus\{g_l,g^{n_k}\}} \exp\left\{-\frac{n_k}{2\sigma^2}\cdot\widehat{\Xi}\left(g_{l'},\widehat{\boldsymbol{H}}^{n_k}\right)\right\}.$$

We may rewrite the probability $p_l^{n_k}$ of g_l as:

$$\begin{aligned}
&p_l^{n_k}\\
&= \frac{\exp\left\{-(n_k/(2\sigma^2))\cdot\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)\right\}}{\sum_{l'}^{L}\exp\left\{-(n_k/(2\sigma^2))\cdot\widehat{\Xi}\left(g_{l'},\widehat{\boldsymbol{H}}^{n_k}\right)\right\}}\\
&= \frac{\exp\left\{-(n_k/(2\sigma^2))\cdot\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)\right\}}{\exp\left\{-(n_k/(2\sigma^2))\cdot\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)\right\}+\exp\left\{-(n_k/(2\sigma^2))\cdot\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)\right\}+S_{\mathcal{C}^{n_k}/\{g_l,g^{n_k}\}}}\\
&\leq \frac{\exp\left\{-(n_k/(2\sigma^2))\cdot\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)\right\}}{\exp\left\{-(n_k/(2\sigma^2))\cdot\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)\right\}+\exp\left\{-(n_k/(2\sigma^2))\cdot\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)\right\}}\\
&= \frac{1}{1+\exp\left\{(n_k/(2\sigma^2))\cdot\left[\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)-\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)\right]\right\}}.
\end{aligned}$$

The condition $p_l^{n_k} < \epsilon^{res}$ suffices to remove g_l which equivalently requires that for $k > T_l$, we have:

$$\begin{aligned}
\frac{1}{1+\exp\left\{(n_k/(2\sigma^2))\cdot\left[\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)-\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)\right]\right\}} &< \epsilon^{res}\\
\frac{2\sigma^2\cdot\ln\left(1/\epsilon^{res}-1\right)}{\widehat{\Xi}\left(g_l,\widehat{\boldsymbol{H}}^{n_k}\right)-\widehat{\Xi}\left(g^{n_k},\widehat{\boldsymbol{H}}^{n_k}\right)} &< n_k
\end{aligned} \tag{2.3.3}$$

Notice that the left hand side of equation (2.3.3) can be positive if we request $\epsilon^{res} < \frac{1}{2}$, which is usually imposed by researchers applying DP-R. Therefore there exists T_l' such that inequality (2.3.3) holds for all $k > T_l'$.

Therefore take $\widetilde{T}_l = \max(T_l, T_l')$. For $k > \widetilde{T}_l$, we have $p_l^{n_k} < \epsilon^{res}$. Hence, we found a time $\widetilde{T}_l$ and a subsequence $\{n_k\}_{k=1}^{\infty}$, along which the "bad" function g_l could

be removed after time $\widetilde{T}_l$. Finally, suppose the cardinality of the candidate set is $\|\mathcal{C}\| = L$. Then there are at most L inferior functions $g_{l'}, l' = 1, 2, ..., L$. Thus for each bad $g_{l'}$, we can find a time $\widetilde{T}_{l'}$ after which $p_{l'}^{n_k} < \epsilon^{res}$ so that we are able to remove $g_{l'}$. Take $\widetilde{T}_{max} = \max_{l'=1,2,...,L}(\widetilde{T}_{l'})$. Then after $\widetilde{T}_{max}$, all the bad functions in the candidate set can be removed. Furthermore, by Theorem 2.2.1, all of the removed bad functions will be replaced by good functions within ε of μ. □

We are now ready to establish the consistency of the estimator of the true function specified in Definition 1.3.8.

2.4 Consistency of the Weighted Estimator

Recall from Definition 1.3.8 that the iteration index $n = 0, 1, 2, ..., \boldsymbol{N} - 1$, $\{f_l^n\}_{n=1}^{\infty}$ is an evolving sequence of functions that represents the l-th element in the candidate set. We write the nth candidate set $\mathcal{C}^n = \{f_1^n, f_2^n, ..., f_L^n\}$ and still use p_l^n as the probability of f_l^n. The estimator of the true function μ after time is:

$$\bar{f}^n(x) = \sum_{l \in \mathbb{L}} p_l^n f_l^n(x),\ x \in \boldsymbol{X}$$

where $\mathbb{L} = \{1, 2, ..., L\}$ is the index set.

Theorem 2.4.1 (Consistency)**.**

The function $\bar{f}^n(x)$ is a strongly consistent estimator of the true function $\mu(x), \forall x \in \boldsymbol{X}$.

Proof. By Theorem 2.3.1 and Theorem 2.2.1, for any $\varepsilon > 0$, there exists a time N_{all}

such that when $n > N_{all}$, for $\forall l \in \mathbb{L}$ and $\forall x \in \boldsymbol{X}$ we have almost surely

$$\Xi(f_l^n(x), \mu(x)) \quad < \quad \frac{\varepsilon^2}{M},$$

which indicates

$$\begin{aligned} 0 &\leq \frac{1}{M} \left[f_l^n(x) - \mu(x)\right]^2 \\ &\leq \frac{1}{M} \sum_{x \in \boldsymbol{X}} \left[f_l^n(x) - \mu(x)\right]^2 \\ &= \Xi(f_l^n(x), \mu(x)) \\ &< \frac{\varepsilon^2}{M}, \end{aligned}$$

and

$$\begin{aligned} & 0 \leq \quad \frac{1}{M} \left[f_l^n(x) - \mu(x)\right]^2 < \frac{\varepsilon^2}{M} \\ \Longrightarrow \quad & 0 \leq \quad \left|f_l^n(x) - \mu(x)\right| < \varepsilon. \end{aligned}$$

Therefore, $f_l^n(x)$ converges to $\mu(x)$ almost surely for $\forall l \in \mathbb{L}$ and $\forall x \in \boldsymbol{X}$. This implies that for $\forall l \in \mathbb{L}$, $\forall x \in \boldsymbol{X}$ and $n > N_{all}$ we have:

$$\begin{aligned} \left|\bar{f}^n(x) - \mu(x)\right| &= \left|\sum_{l \in \mathbb{L}} p_l^n f_l^n(x) - \mu(x) \cdot 1\right| \\ &= \left|\sum_{l \in \mathbb{L}} p_l^n f_l^n(x) - \sum_{l \in \mathbb{L}} p_l^n \mu(x)\right| \\ &\leq \sum_{l \in \mathbb{L}} p_l^n \left|f_l^n(x) - \mu(x)\right| \end{aligned}$$

$$< \varepsilon \cdot \sum_{l \in \mathbb{L}} p_l^n$$

$$= \varepsilon$$

Hence, the estimator $\bar{f}^n(x)$ converges to $\mu(x)$ almost surely for $\forall x \in \boldsymbol{X}$. We conclude that $\bar{f}^n(x)$ is a (strongly) consistent estimator of the true function $\mu(x), x \in \boldsymbol{X}$. $\square$

Chapter 3: Almost Sure Convergence of the Bias-Adjusted Kalman Filter

3.1 Introduction

In this chapter we consider the topic of predicting and estimating a background signal in a sequentially recursive manner. This statistical issue arises in economics (Van and Dana, 2003), forecasting (Gardner, 1985; Hyndman et al., 2008), simulation optimization (Fu, 2002), bioinformatics (Giegerich, 2000), signal processing (Bouzeghoub et al., 2000) and dynamic programming (Powell, 2007, Ch.6). The underlying signal may represent a stationary or nonstationary series of stock prices, inventory levels, expensive experimental outcomes, genetic sequence alignments or sensor readings of a mechanical system. In most of these applications, the mean value of the signal is unknown to the researcher and can only be estimated from noisy observations with some prespecified variance structure. The impact of the noise can be smoothed with new incoming observations. In its most typical form, the procedure recursively updates the estimates according to

$$\bar{\theta}_{n+1} = (1 - \alpha_n)\bar{\theta}_n + \alpha_n \hat{X}_{n+1} \tag{3.1.1}$$

where $\bar{\theta}_n$ is the current estimate at time n, $\hat{X}_{n+1}$ is the observation obtained at time $(n+1)$, and α_n is a weight between 0 and 1 and is commonly referred to (e.g., in dynamic programming and stochastic approximation) as a *stepsize*. It is also known by other terminologies such as *learning rate* (machine learning), *smoothing constant* (forecasting) or *gain* (signal processing). The magnitude of α_n determines the rate at which new information is combined with the existing knowledge.

Define $\{\theta_n\}_{n=1}^{\infty}$ as the underlying sequence of signals. The optimal choice of stepsize for (3.1.1) depends on the stationarity of $\{\theta_n\}_{n=1}^{\infty}$. For observations that come from a nonstationary series, it is very often the case (e.g. in dynamic programming) that the learning process undergoes an initial unstable phase where the estimates fluctuate erratically, and then converge to an eventual value, assuming the underlying signal converges to a limit point. Nonstationarity arises when either the starting estimate is way off target or the underlying signal itself is of a nonstationary nature due to the physical system. In such scenarios, typically we desire the α_n to be relatively high in the early learning stage so as to drive the estimating process towards the potential true parameter, and then we reduce it later.

Equation (3.1.1) can be viewed as a form of stochastic approximation (SA), introduced by Robbins and Monro (1951) and Kiefer and Wolfowitz (1952). Pflug (1988) provided an overview of some deterministic and adaptive stepsize rules, and points out that the major drawback of applying a deterministic stepsize rule is the strong dependence between its performance and the initial estimate. It is suggested that during the progress of the estimation procedure, it could be advantageous to adopt certain adaptive rules that enable the stepsizes to vary based on the collected

observations.

The study of optimal stepsizes are also carried out in many other practical realms. Gardner (1985) compared various stepsize rules developed in the forecasting community. In the field of reinforcement learning. Darken and Moody (1991) tackled the issue of allowing the stepsize to evolve at multiple rates, and uses time-dependent deterministic models to calculate the stepsize. The usage of time-varying stepsize sequences has been suggested by researchers in signal processing and adaptive control to boost the speed of convergence of filter coefficients; see for example Mikhael et al. (1986) and Bouzeghoub et al. (2000) for a general review. A number of techniques have been proposed where the stepsize is computed adaptively as a function of prediction or estimation errors (Kesten, 1958; Saridis, 1970). Another approach is to adjust the stepsize by a correction term that is a function of the gradient of the error measure that needs to be minimized (Kushner and Yang, 1995; Douglas and Mathews, 1995; Douglas and Cichocki, 1998; Schraudolph, 1999). However, most gradient adaptive stepsize methods suffer the limitation of using a smoothing parameter for updating the stepsize values. The ideal parameter value depends on the specific problem and changes with each coefficient being estimated, causing intractable computational issues when the number of parameters is large.

Among the rich family of stepsize structures, the Kalman Filter stepsize is broadly used in adaptive control applications. The Kalman Filter technique estimates system states sequentially by incorporating two parts: previous estimates and newly recorded information. The relative weights placed on these two parts are controlled by the Kalman Filter stepsize α_n, which is calculated adaptively by

minimizing the expected value of the squared prediction error

$$\mathbb{E}\left[(\hat{\varepsilon}_n)^2\right] = \mathbb{E}\left[\left(\hat{X}_{n+1} - \bar{\theta}_n\right)^2\right];$$

see for example Stengel (1994) and Choi and Van Roy (2006). Suppose $\theta_n = \theta_{n-1} + u_n$ is the sequence of system states with $\mathbb{E}(u_n) = 0$ and $Var(u_n) = \rho^2$, and $\hat{X}_n = \theta_n + \varepsilon_n$ is the sequence of measurements of system states with $\varepsilon_n \sim N(0, \sigma^2)$. Then the computation of Kalman Filter stepsizes depends on both of the measurement noise and the process noise, according to the equations:

$$\alpha_n = \frac{p_{n-1}}{p_{n-1} + \sigma^2},$$

$$p_n = (1 - \alpha_n)p_{n-1} + \rho^2.$$

George and Powell (2006) first introduced the bias-adjusted Kalman Filter procedure (BAKF) specifically to handle non-stationary system signals. The BAKF procedure enables the stepsize to be dependent on the estimation bias of $\bar{\theta}_n$, which allows α_n to be responsive to the level of measurement noise compared to the rate of change of the underlying system signals. For more details and in-depth motivations of the BAKF method, see Powell (2007, Ch.6). One appealing advantage of the BAKF procedure is that it adaptively learns non-stationary signals, which leads to a much faster approximation process in approaching an accurate estimate. This feature is crucial in many applications where each observation can be highly expensive in financial costs or time (see Simão et al. (2009; 2010)).

It has been well established in the stochastic approximation literature that the conditions on α_n which ensure the almost sure convergence of the estimating sequence $\bar{\theta}_n$ to the right limit point are the following two expressions:

$$\sum_{n=0}^{\infty} \alpha_n = \infty, \tag{3.1.2}$$

$$\sum_{n=0}^{\infty} \alpha_n^2 < \infty. \tag{3.1.3}$$

The interpretation is that essentially we require the stepsize to diminish to zero towards the infinite time horizon so that (3.1.1) converges. However, α_n should not vanish too rapidly since otherwise the process will be trapped at a suboptimal value. Although there are plenty of stepsize algorithms satisfying the conditions, many of them suffer from low convergence rates in practice (Ryzhov et al., 2015). On the contrary, the BAKF procedure exhibits remarkable performance in many practical situations, but a thorough consistency analysis is yet to be explored. Ryzhov (2018) has shown convergence of the BAKF stepsize in the sense of L_2, but we are interested in extending this result to almost sure convergence. Therefore, based on Ryzhov (2018) and George and Powell (2006), our plan for this chapter is to show the almost sure convergence of the estimation process and thus the strong consistency of $\bar{\theta}_n$ to its limiting value.

We need to emphasize that the empirical performance of BAKF is not our major concern, since the practical superiority of BAKF has been demonstrated by

George and Powell (2006), together with a comprehensive comparison of BAKF against a broad choice of other stepsize rules. The sole objective of our work is to establish the theoretical properties of BAKF in a stronger sense of convergence. The rest of this chapter is organized as follows. Section 3.2 introduces the model that induces BAKF and some theoretical results derived by Ryzhov (2018) and George and Powell (2006). Section 3.3 extends the L_2 convergence in Ryzhov (2018) to the almost sure case. Section 3.4 concludes and propose some future work in the theoretical part of BAKF.

3.2 Review of Model, Notations and Definitions

We begin with the model settings of the BAKF stepsize algorithm. Consider a deterministic sequence $\{\theta_n\}_{n=1}^{\infty}$ which represents the underlying non-stationary system states that varies over time. We are able to obtain noisy measurements

$$\hat{X}_n = \theta_n + \varepsilon_n$$

where $\{\varepsilon_n\}_{n=1}^{\infty}$ is a sequence of independent random noise with $\mathbb{E}(\varepsilon_n) = 0$ and $Var(\varepsilon_n) = \sigma^2$, assumed known to the researchers. Presumably, the observation sequence $\hat{X}_n$ is much more volatile than the underlying signal, that is, the rate of change of θ_n is much smaller than the variance σ^2. Thus, a smoothing technique needs to be applied to dampen the impact of the noisy measurements on our estimates $\bar{\theta}_n$, which is defined by (3.1.1). We denote by $\bar{\theta}_0$ the initial estimate and assume that $\theta_n \to \theta^*$.

The BAKF chooses the stepsize α_n in equation (3.1.1) by minimizing the expected squared error of the smoothed estimate $\bar{\theta}_n$ with respect to θ_n on the interval $[0,1]$. That is:

$$\alpha_n = \arg\min_{\alpha\in[0,1]} \mathbb{E}\left[\bar{\theta}_n(\alpha) - \theta_n\right]^2. \tag{3.2.1}$$

George and Powell (2006) derive the closed form solution of (3.2.1) as

$$\alpha_n = \frac{\lambda_n\sigma^2 + (\beta_{n+1})^2}{(1+\lambda_n)\sigma^2 + (\beta_{n+1})^2} \tag{3.2.2}$$

where

$$\begin{aligned}\beta_{n+1} &= \mathbb{E}[\theta_{n+1} - \bar{\theta}_n] \\ &= \theta_{n+1} - \mathbb{E}[\bar{\theta}_n]\end{aligned} \tag{3.2.3}$$

represents the bias in the estimate from the previous iteration, and λ_n is the coefficient of the variance of $\bar{\theta}_n$ defined by

$$Var(\bar{\theta}_n) = \lambda_n\sigma^2$$

and can be calculated due to George and Powell (2006) by

$$\lambda_n = \begin{cases} \alpha_0^2, & n = 1 \\ \alpha_{n-1}^2 + (1-\alpha_{n-1})^2\lambda_{n-1}, & n \geq 2, \end{cases}$$

where α_0 is an deterministic initial stepsize value in $[0, 1]$. In a real application, β_{n+1} would also need to be estimated by a statistical procedure, but for our consistency analysis, we focus on its explicit form in (3.2.3) and the stepsize α_n in (3.2.2). Ryzhov (2018) has already established some propositions and theorems about BAKF which we summarize below.

The first two propositions show the boundedness of both the expectation and variance of our estimates.

Proposition 3.2.1. *Define $\zeta_n = \mathbb{E}\bar{\theta}_n$. The sequence $\{\zeta_n\}$ is uniformly bounded in n, regardless of the stepsize used for updating.*

Proposition 3.2.2. *For all n, $0 \leq \lambda_n \leq 1$.*

The following Theorem 3.2.1 establishes the L_1 convergence of our estimates and the bias term. The next Theorem 3.2.2 proves that both of the variance of $\bar{\theta}_n$ and the stepsize sequence vanish to zero.

Theorem 3.2.1. *$\zeta_n \to \theta^*$ and thus $\beta_n \xrightarrow{L_1} 0$.*

Theorem 3.2.2. *$\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty} \alpha_n = 0$.*

With these propositions and theorems, Ryzhov (2018) further shows that $\bar{\theta}_n$ converges to θ^* in L_2. To see this, write

$$\begin{aligned}
\mathbb{E}\left[\bar{\theta}_n - \theta^*\right]^2 &= \mathbb{E}\left[\bar{\theta}_n - \zeta_n + \zeta_n - \theta^*\right]^2 \\
&= \mathbb{E}\left[\bar{\theta}_n - \zeta_n\right]^2 + (\zeta_n - \theta^*)^2 + 2(\zeta_n - \theta^*)\mathbb{E}\left(\bar{\theta}_n - \zeta_n\right) \\
&= Var(\bar{\theta}_n) + (\zeta_n - \theta^*)^2 + 0
\end{aligned}$$

$$= \lambda_n \sigma^2 + (\zeta_n - \theta^*)^2$$

$$\to 0.$$

The third equation is due to the definition of ζ_n and the last convergence is by Theorems 3.2.1 and 3.2.2.

To extend this result to almost sure convergence, typically we need the two conditions in (3.1.2) and (3.1.3). Recall that George and Powell (2006) have proved that $\alpha_n \geq \frac{1}{n+1}$ for all n which implies that $\sum_{n=0}^{\infty} \alpha_n = \infty$. They also provide strong empirical evidence that indicates $\alpha_n \leq \frac{c}{n+1}$ for all n with some constant c, but they have not been able to prove it as of their writing. In the next section we introduce the rigorous proof of this inequality, which would lead to the fulfillment of the second condition $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$.

3.3 Main Results

We begin with the following additional assumption:

Assumption 1. *There exists some fixed positive constant h such that for all n, the sequence $\{\theta_n\}$ satisfies*

$$|\theta_n - \theta^*| \leq \frac{h}{2(n+1)^{3/2}};$$

and $\alpha_n \geq \frac{1}{n+1}$ for all n.

In words, we make a stronger assumption on the rate of convergence of the underlying system states in order to obtain stronger convergence of the estimator.

However, note that the measurement variance σ^2 can still be arbitrarily large.

For simplicity of the proof, we now suppose that $\theta^* = 0$ without loss of generality. To show the almost sure convergence of $\{\bar{\theta}_n\}$, we first present the following result on the convergence rate of the prediction bias term β_n.

Lemma 3.3.1. *Under Assumption 1, the sequence $\{\beta_n\}$ satisfies*

$$|\beta^{n+1}| \leq \frac{2h}{\sqrt{n+1}}$$

for all n.

Proof. Recall $\theta^* = 0$ and by Assumption 1 for all n we have

$$\begin{aligned}
|\theta_n| &\leq \frac{h}{2(n+1)^{3/2}} \\
&= \frac{1}{2}\frac{\sqrt{n}}{\sqrt{n+1}}\frac{1}{n}\frac{h\sqrt{n}}{(n+1)} \\
&= \frac{1}{2}\frac{1}{\sqrt{1+1/n}}\left(1+\frac{1}{n}-1\right)\frac{h\sqrt{n}}{(n+1)} \\
&\leq \left(\sqrt{1+\frac{1}{n}}-1\right)\frac{h\sqrt{n}}{(n+1)} \qquad (3.3.1) \\
&= h\frac{\sqrt{n+1}-\sqrt{n}}{n+1} \qquad (3.3.2)
\end{aligned}$$

To see why the last inequality holds, consider the concave function $f(x) = \sqrt{x}$. For any two nonnegative points x and y, the concavity of $f(x)$ implies

$$f(y) - f(x) \leq f'(x)(y-x) \qquad (3.3.3)$$

where the derivative $f'(x) = (1/2)x^{-1/2}$. Set $x = 1 + 1/n$ and $y = 1$. The inequality (3.3.3) becomes

$$\begin{aligned} & f(1) - f(1+\frac{1}{n}) \leq f'(1+\frac{1}{n})(-\frac{1}{n}) \\ \Longrightarrow\ & 1 - \sqrt{1+\frac{1}{n}} \leq \frac{1}{2}\frac{1}{\sqrt{1+1/n}}(-\frac{1}{n}) \\ \Longrightarrow\ & \sqrt{1+\frac{1}{n}} - 1 \geq \frac{1}{2}\frac{1}{\sqrt{1+1/n}}\frac{1}{n} \\ & \qquad = \frac{1}{2}\frac{1}{\sqrt{1+1/n}}(1+\frac{1}{n}-1) \end{aligned}$$

which is the inequality in (3.3.1).

Now we will show that $|\mathbb{E}\bar{\theta}_n| \leq h^*/\sqrt{n+1}$ for all n and some constant h^* by induction. For $n = 1$, since $\bar{\theta}_0$ is deterministic that may not be zero and α_0 is an arbitrary starting value in $(0, 1)$, we can find such a constant $h^* \geq h$ such that

$$\begin{aligned} |\mathbb{E}\bar{\theta}_1| &= |\mathbb{E}[(1-\alpha_0)\bar{\theta}_0 + \alpha_0\hat{X}_1]| \\ &\leq (1-\alpha_0)|\bar{\theta}_0| + \alpha_0|\mathbb{E}\hat{X}_1| \\ &= (1-\alpha_0)\bar{\theta}_0 + \alpha_0\theta_1 \\ &= \frac{(1-\alpha_0)\bar{\theta}_0\sqrt{2} + \alpha_0 h}{\sqrt{2}} \\ &\leq \frac{h^*}{\sqrt{2}}. \end{aligned}$$

Suppose that $|\mathbb{E}\bar{\theta}_n| \leq h^*/\sqrt{n+1}$ holds for some n. Then we have

$$|\mathbb{E}\bar{\theta}_{n+1}| = |\mathbb{E}[(1-\alpha_n)\bar{\theta}_n + \alpha_n\hat{X}_{n+1}]|$$

$$
\begin{aligned}
&= |(1-\alpha_n)\mathbb{E}\bar{\theta}_n + \alpha_n \mathbb{E}\hat{X}_{n+1}| \\
&\leq (1-\alpha_n)|\mathbb{E}\bar{\theta}_n| + \alpha_n |\mathbb{E}\hat{X}_{n+1}| \\
&\leq \left(1 - \frac{1}{n+1}\right)|\mathbb{E}\bar{\theta}_n| + |\theta_{n+1}| \\
&\leq \left(1 - \frac{1}{n+2}\right)\frac{h^*}{\sqrt{n+1}} + h\frac{\sqrt{(n+1)+1} - \sqrt{n+1}}{(n+1)+1} \\
&= h^*\frac{\sqrt{n+1}}{n+2} + h^*\frac{\sqrt{n+2} - \sqrt{n+1}}{n+2} \\
&= \frac{h^*}{\sqrt{(n+1)+1}}
\end{aligned}
$$

where the second inequality holds because $1/(n+1) \leq \alpha_n \leq 1$ and the third inequality holds because of equation (3.3.2). Therefore, for all n, we have $|\mathbb{E}\bar{\theta}_n| \leq h^*/\sqrt{n+1}$. Then for all n, we have

$$
\begin{aligned}
|\beta_{n+1}| &= |\theta_{n+1} - \mathbb{E}\bar{\theta}_n| \\
&\leq |\theta_{n+1}| + |\mathbb{E}\bar{\theta}_n| \\
&\leq \frac{h}{2(n+1)^{\frac{3}{2}}} + \frac{h^*}{\sqrt{n+1}} \\
&\leq \frac{h^*}{\sqrt{n+1}} + \frac{h^*}{\sqrt{n+1}} \\
&= \frac{2h^*}{\sqrt{n+1}}
\end{aligned}
$$

which completes the proof. □

For simplicity of derivations, we define the following additional notations

$$
b = \left\lceil \frac{4h^2}{\sigma^2} \right\rceil,
$$

$$l = \max\{b+1, 4\},$$

$$c = l + b,$$

where $\lceil \cdot \rceil$ is the ceiling function. Obviously, all b, l, c are positive integers satisfying $1 < l < c < 2l$ and $c + l - 1 > 1$. By Lemma 3.3.1, we have

$$\begin{aligned}(\beta_{n+1})^2 &\leq \frac{4h^2}{n+1} \\ &\leq \frac{b\sigma^2}{n+1}\end{aligned}$$

for all n. We now proceed to our main theorem, which shows that $\alpha_n = \mathcal{O}\left(1/(n+1)\right)$, and whence $\sum_n \alpha_n^2 < \infty$.

Theorem 3.3.1. *Under Assumption 1, we have*

$$\lambda_{n-1} \leq \frac{l}{n-c} \quad \textit{and} \quad \alpha_{n-1} \leq \frac{c}{n}$$

for all $n \geq c + l$.

Proof. We will establish the proof by induction. First for $n = c + l$, by Proposition 3.2.2 we have

$$\lambda_{c+l-1} \leq 1 = \frac{l}{c+l-c}.$$

Now suppose that $\lambda_{n-1} \leq l/(n-c)$ for $n > c + l$. We will show that

$$\lambda_{n-1} \leq \frac{l}{n-c} \implies \alpha_{n-1} \leq \frac{c}{n} \implies \lambda_n \leq \frac{l}{n+1-c}.$$

By Lemma 3.3.1 and equation 3.2.2, we have

$$\begin{aligned}
\alpha_{n-1} &= \frac{\lambda_{n-1}\sigma^2 + \beta_n^2}{(1+\lambda_{n-1})\sigma^2 + \beta_n^2} \\
&\leq \frac{\sigma^2 l/(n-c) + \sigma^2 b/n}{\sigma^2(1 + l/(n-c)) + \sigma^2 b/n} \\
&\leq \frac{\sigma^2 l/(n-c) + \sigma^2 b/(n-c)}{\sigma^2(1 + l/(n-c)) + \sigma^2 b/(n-c)} \\
&= \frac{l+b}{(n-c)+l+b} \\
&= \frac{c}{(n-c)+c} \\
&= \frac{c}{n}.
\end{aligned} \tag{3.3.4}$$

Thus we can represent $\alpha_{n-1} = a_n/n$, where $1 \leq a_n \leq c$. Then

$$\frac{a_n}{n} = \alpha_{n-1} = 1 - \frac{\sigma^2}{(1+\lambda_{n-1})\sigma^2 + \beta_n^2}$$

which implies

$$\begin{aligned}
\frac{\sigma^2}{(1+\lambda_{n-1})\sigma^2 + \beta_n^2} &= 1 - \frac{a_n}{n} \\
\implies 1 + \lambda_{n-1} + \frac{\beta_n^2}{\sigma^2} &= \frac{n}{n-a_n} \\
\implies \lambda_{n-1} + \frac{\beta_n^2}{\sigma^2} &= \frac{a_n}{n-a_n},
\end{aligned}$$

whence

$$\lambda_{n-1} \leq \frac{a_n}{n-a_n}. \tag{3.3.5}$$

We will consider two cases $a_n \leq l$ and $a_n > l$ respectively.

First, if $a_n \leq l$, we derive

$$\begin{aligned}
\lambda_n &= \alpha_{n-1}^2 + (1-\alpha_{n-1})^2 \lambda_{n-1} \\
&= \left(\frac{a_n}{n}\right)^2 + \left(1 - \frac{a_n}{n}\right)^2 \lambda_{n-1} \\
&\leq \left(\frac{a_n}{n}\right)^2 + \left(1 - \frac{a_n}{n}\right)^2 \frac{a_n}{n-a_n} \\
&= \frac{a_n^2 + a_n(n-a_n)}{n^2} \\
&= \frac{a_n}{n} \\
&\leq \frac{l}{n+1-c},
\end{aligned}$$

where the first inequality holds because of (3.3.5), and the last inequality holds because $a_n \leq l$ and $c > 1$ (recall $1 < l < c < 2l$).

On the other hand, in the case when $a_n > l$, we write

$$\begin{aligned}
\lambda_n &= \alpha_{n-1}^2 + (1-\alpha_{n-1})^2 \lambda_{n-1} \\
&= \left(\frac{a_n}{n}\right)^2 + \left(1 - \frac{a_n}{n}\right)^2 \lambda_{n-1} \\
&\leq \left(\frac{a_n}{n}\right)^2 + \left(1 - \frac{a_n}{n}\right)^2 \frac{l}{n-c}
\end{aligned}$$

where the last inequality is due to the induction assumption mentioned at the beginning of our proof. Thus, to show $\lambda_n \leq l/(n+1-c)$, it suffices to show

$$\left(\frac{a_n}{n}\right)^2 + \left(1 - \frac{a_n}{n}\right)^2 \frac{l}{n-c} \quad \leq \quad \frac{l}{n+1-c},$$

which implies

$$\begin{aligned}\frac{a_n^2}{n^2} &\leq \frac{l}{n+1-c} - \left(1-\frac{a_n}{n}\right)^2 \frac{l}{n-c} \\ &\leq l\frac{n^2(n-c)-(n-a_n)^2(n+1+c)}{n^2(n+1+c)(n-c)} \\ &\leq l\frac{(2na_n-a_n^2)(n-c)-(n-a_n)^2}{n^2(n+1+c)(n-c)}\end{aligned}$$

which is equivalent to the inequality

$$n^2 l\left[(2na_n-a_n^2)(n-c)-(n-a_n)^2\right]-a_n^2n^2(n+1+c)(n-c) \quad \geq \quad 0.$$

That is

$$l\left[(2na_n-a_n^2)(n-c)-(n-a_n)^2\right]-a_n^2(n+1+c)(n-c)\geq 0.$$

For simplicity, we define the notation

$$f(n;a_n) \coloneqq l\left[(2na_n-a_n^2)(n-c)-(n-a_n)^2\right]-a_n^2(n+1+c)(n-c).$$

Then we expand $f(n;a_n)$ and complete the square in terms of the parabola in n:

$$\begin{aligned}f(n;a_n) &= (l(2a_n-1)-a_n^2)n^2-\left[l(a_n^2+2ca_n-2a_n)-a_n^2(2c-1)\right]n \\ &\quad + a_n^2(c-1)(l-c) \\ &= (l(2a_n-1)-a_n^2)\left[n-\frac{l(a_n^2+2ca_n-2a_n)-a_n^2(2c-1)}{2(l(2a_n-1)-a_n^2)}\right]^2\end{aligned}$$

$$+\frac{4(l(2a_n-1)-a_n^2)a_n^2(c-1)(l-c)-[l(a_n^2+2ca_n-2a_n)-a_n^2(2c-1)]^2}{4(l(2a_n-1)-a_n^2)}.$$

Thus, to show that $f(n;a_n)\geq 0$, it is sufficient to show the following three inequalities:

$$\begin{aligned} l(2a_n-1)-a_n^2 &> 0, && (3.3.6)\\ c+l &\geq \frac{l(a_n^2+2ca_n-2a_n)-a_n^2(2c-1)}{2(l(2a_n-1)-a_n^2)}, && (3.3.7)\\ f(c+l;a_n) &\geq 0, && (3.3.8) \end{aligned}$$

because when (3.3.6) and (3.3.7) hold, we have

$$f(n;a_n)\geq f(c+l;a_n)$$

for all $n\geq c+l$. At that point, (3.3.8) leads to $f(n;a_n)\geq 0$.

Recall $1<l<a_n\leq c$. To show inequality (3.3.6), it is sufficient to show

$$l>\frac{a_n^2}{2a_n-1},$$

where

$$g(a_n):=\frac{a_n^2}{2a_n-1}$$

is an increasing function of a_n with $1<l<a_n\leq c$. To see this, take the derivative

of $g(a_n)$ with respect to a_n

$$\begin{aligned} g'(a_n) &= \left(\frac{a_n^2}{2a_n - 1}\right)' \\ &= \frac{2a_n(2a_n - 1) - 2a_n^2}{(2a_n - 1)^2} \\ &= \frac{2a_n(a_n - 1)}{(2a_n - 1)^2} \\ &> 0. \end{aligned}$$

Thus $g(a_n) \leq g(c)$ and to show $l > g(a_n)$, it is sufficient to show $l > g(c) = c^2/(2c-1)$, which is equivalently to show $l(2c-1) - c^2 > 0$. This holds since

$$\begin{aligned} l(2c-1) - c^2 &= l(2(l+b) - 1) - (l+b)^2 \\ &= 2l^2 + 2lb - l - l^2 - 2lb - b^2 \\ &= l(l-1) - b^2 \\ &> 0, \end{aligned}$$

where the last inequality hold because by definition $l \geq b + 1$. Hence, (3.3.6) is satisfied.

To show inequality (3.3.7), we can see that, since (3.3.6) is valid, it is sufficient to show

$$2(c+l)(l(2a_n - 1) - a_n^2) - \left[l(a_n^2 + 2ca_n - 2a_n) - a_n^2(2c-1)\right] \geq 0$$

where the left hand side of the above inequality can be manipulated as

$$
\begin{aligned}
&2(c+l)(l(2a_n-1)-a_n^2)-\left[l(a_n^2+2ca_n-2a_n)-a_n^2(2c-1)\right]\\
&=4a_ncl-2cl+4a_nl^2-2l^2-2a_n^2c-2a_n^2l-a_n^2l-2a_ncl+2a_nl+2ca_n^2-a_n^2\\
&=2a_ncl+4a_nl^2+2a_nl-3a_n^2l-a_n^2-2cl-2l^2\\
&=(2a_nl-2l^2)+(2a_ncl+2a_nl^2-3a_n^2l)+(2a_nl^2-a_n^2-2cl)\\
&\geq(2l\cdot l-2l^2)+(2a_n\cdot a_nl+a_nl^2-3a_n^2l)+(2l\cdot l^2-(2l)^2-2(2l)l)\\
&=0+(2a_n^2l+a_nl^2-3a_n^2l)+(2l^3-8l^2)\\
&\geq(2l^2\cdot l+l\cdot l^2-3l^2\cdot l)+(2l^3-8l^2)\\
&=0+2l^2(l-4)\\
&\geq 0,
\end{aligned}
$$

where the first and the second inequality holds because $l<a_n\leq c<2l$ and the last inequality hold since in fact $l\geq 4$ by construction. Hence, (3.3.7) is satisfied.

To show inequality (3.3.8), we can see that

$$
\begin{aligned}
&f(c+l;a_n)\\
&=2a_ncl^2+2a_ncl+2a_nl^3+2a_nl^2-2a_n^2l^2-2a_n^2l-c^2l-2cl^2-l^3\\
&=(2a_ncl^2-2a_n^2l^2)+(2a_ncl-2a_n^2l)+(2a_nl^2-l^3)+(2a_nl^3-c^2l-2cl^2)\\
&\geq(2a_na_nl^2-2a_n^2l^2)+(2a_na_nl-2a_n^2l)\\
&+(2l\cdot l^2-l^3)+(2l\cdot l^3-(2l)^2l-2\cdot 2l\cdot l^2)\\
&=0+0+l^3+2l^3(l-4)
\end{aligned}
$$

$$\geq 0,$$

where the first inequality holds since $l < a_n \leq c < 2l$ and the final inequality holds again since $l \geq 4$. Hence, (3.3.8) is satisfied. Now all of the conditions (3.3.6)-(3.3.8) are satisfied which means we have $f(n; a_n) \geq 0$, whence

$$\lambda_n \leq \frac{l}{n+1-c}$$

when $a_n > l$.

Up until now, we have proved that

$$\lambda_{n-1} \leq \frac{l}{n-c} \implies \alpha_{n-1} \leq \frac{c}{n} \implies \lambda_n \leq \frac{l}{n+1-c}.$$

Therefore, by mathematical induction, for all $n \geq c + l$, we have

$$\lambda_{n-1} \leq \frac{l}{n-c}$$

and

$$\alpha_{n-1} \leq \frac{c}{n},$$

which completes the proof. □

Combining the result in Theorem 3.3.1 and in George and Powell (2006), we have shown that

$$\frac{1}{n+1} \leq \alpha_n \leq \frac{c}{n+1}$$

which means both of the conditions $\sum_{n=0}^{\infty} \alpha_n = \infty$ and $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$ are satisfied. This implies that $\bar{\theta}_n$ converges to θ^* almost surely due to the well-known theory of stochastic approximation (Kushner and Yin, 2003).

3.4 Conclusion and Future Work

In this chapter, we reviewed the bias-adjusted Kalman filter procedure which is an algorithm for finding an optimal stepsize in a stochastic approximation process with non-stationary observations. We generalized the convergence of the estimator to the system states, from L_2 convergence, proved in previous literature, to almost sure convergence, by verifying that the Kalman filter stepsize satisfies the well-known conditions for almost sure convergence of stochastic approximation. This result has not been previously available in the literature and is particularly noteworthy because BAKF is the first algorithm to employ an optimal stepsize that can automatically adapt to the level of noise in a measured signal. Optimal stepsizes have been proven to dramatically enhance the performance of stochastic approximation algorithm in both dynamic programming and signal processing, and it is important to see that the legitimacy of BAKF stepsizes is substantiated by asymptotic consistency. Therefore we can be confident that the procedure will not fundamentally misdirect us in real implementation.

Nevertheless, one major concern of the BAKF stepsize is the bias term β_n whose calculation involves the unknown underlying sequence of states θ_n. Obviously, in many applications, the bias term is also unknown and must be approximated. The

study by George and Powell (2006) proposes an approximate version of the BAKF stepsize given by the plug-in estimator

$$\bar{\alpha}_n = \frac{\bar{\lambda}_n \bar{\sigma}_n^2 + (\bar{\beta}_{n+1})^2}{(1+\bar{\lambda}_n)\bar{\sigma}_n^2 + (\bar{\beta}_{n+1})^2} \tag{3.4.1}$$

where

$$\bar{\beta}_{n+1} = (1-\nu_n)\bar{\beta}_n + \nu_n(\hat{X}_{n+1} - \bar{\theta}_n)$$

$$\bar{\sigma}_n^2 = \frac{\bar{\delta}_n - \bar{\beta}_n}{1+\bar{\lambda}_{n-1}}$$

where $\nu_n = 1/(n+1)$ is a deterministic stepsize, δ_n is the expected value of the squared prediction errors with

$$\begin{aligned}\delta_n &\coloneqq \mathbb{E}\left[\left(\hat{X}_n - \bar{\theta}_{n-1}\right)^2\right] \\ &= (1+\lambda_{n-1})\sigma^2 + \beta_n^2,\end{aligned}$$

which is further approximated by

$$\bar{\delta}_n = (1-\nu_{n-1})\bar{\delta}_{n-1} + \nu_{n-1}(\hat{X}_n - \bar{\theta}_{n-1})^2,$$

and $\bar{\lambda}_n$ is the approximation of λ_n that is calculated recursively by

$$\bar{\lambda}_n \coloneqq \bar{\alpha}_{n-1}^2 + (1-\bar{\alpha}_{n-1})^2\,\bar{\lambda}_{n-1}.$$

Notice that under these settings, $\bar{\alpha}_n$ no longer coincides with the solution to the

optimization problem in equation 3.2.1. Compared to the deterministic α_n in equation 3.2.2, $\bar{\alpha}_n$ is the plug-in estimator by replacing the corresponding deterministic quantities with its stochastic approximation counterpart. Our future efforts are to extend the consistency results of the estimator $\bar{\theta}_n$ to the case with the stochastic stepsize $\bar{\alpha}_n$.

Appendix A: Convergence of Knowledge Gradient Policy

In Section 1.3, the measurement decision x^n is due to a exogenous decision-making policy. One such policy is the *knowledge gradient* (KG) policy, which was first proposed and studied in Frazier et al.(2008; 2009) and extended by Ryzhov et al. (2012). In recent years, KG was applied in combination with Bayesian methods, one of which is the *knowledge gradient with discrete priors* (KGDP). This policy KGDP was first introduced by Chen et al. (2015), and later was extended to *KGDP with resampling* (KGDP-R) by He and Powell (2016) and further extended to *KGDP with resampling and regeneration* (KGDP-R&R) by Huang et al. (2018). We refer readers to Huang et al. (2018, Table 1) for a summary of papers on knowledge gradient policies.

In Chapter 2, we have shown the consistency of the estimator of the true (cost) function under the assumption that the information collection policy (i.e., KGDP) would potentially guarantee us enough observations for each alternative in the set $\boldsymbol{X} = \{1, 2, ..., M\}$. In this appendix, let $\boldsymbol{N}$ be the number of experiments we are able to perform under the budget. We would like to address two more topics with regard to KGDP as mentioned in Section 1.4. The first topic is the formulation of KGDP when there is more than one delivery region (as in Huang et al. (2018) and our

motivating example), in which case there are R different true cost functions for the R regions. Another topic, as a follow up to the first one, is whether or not the marginal information gain of measuring one more alternative can be shown to have limit zero so that we could maximize the potency of the $(\boldsymbol{N})$ experiments under the budget. In this appendix, we will consider R regions and we begin with some modifications to our previous notations and definitions to accommodate this change. Notice that our previous consistency proof still holds for each region considered individually.

A.1 Modified Notations and Definitions

The following notations and definitions are designed for proofs related to the KGDP and are only used for expressions and derivations in this appendix.

Definition A.1.1. We rewrite the candidate set for the region r at the time n as

$$\mathcal{C}_r^n = \{f_{r1}^n, f_{r2}^n, ..., f_{rL}^n\}.$$

We still use a fixed number L of candidate functions over the entire sequence of experiments.

Definition A.1.2. In Definition 1.3.2 we defined the true curve for only one region. Similarly, we define an unknown function $\mu_r(x), x \in \boldsymbol{X}$ that is assumed to be isotonic with respect to the simple order $1 < 2 < ... < M$, on the set $\boldsymbol{X} = \{1, 2, ..., M\}$. We call $\mu_r(x)$ the *true function*, or the *true curve* of the region r. If we allocate x_r delivery resources to region r, then the $\mu_r(x_r)$ is interpreted as the true cost of such

an allocation.

Definition A.1.3. Consider the following optimization problem as described in Huang et al. (2018):

$$\min_{x_1,...,x_R} \sum_{r=1}^{R} \mu_r(x_r) \tag{A.1.1}$$
$$\text{s.t.} \quad \sum_{r=1}^{R} x_r = M, \; x_r \in \boldsymbol{X}$$

Denote by $\mathscr{X}$ the feasible region of $\mathbf{x}$ in the above resource allocation minimization model (A.1.1), where $\mathbf{x} = (x_r)_{\forall r}$ is the R-dimensional vector of integer variables representing a resource allocation solution. Therefore, $\mathbf{x} = (x_r)_{\forall r} \in \mathscr{X}$ and $x_r \in \boldsymbol{X} = \{1, 2, ..., M\}$. Italic $\boldsymbol{X}$ denotes the set of alternatives and scripted $\mathscr{X}$ denotes the set of all possible recourse allocation solutions.

Definition A.1.4. Denote by S^n the *state variable* that contains our belief about the candidates and the R true curves at time n. We write

$$S^n = \{p_{11}^n, ..., p_{rl}^n, ..., p_{RL}^n\} = \{p_{rl}^n\}_{\forall r,l}$$

where $r = 1, 2, ..., R$, $l = 1, 2, ..., L$, and p_{rl}^n is the probability that the l-th curve in the candidate set $\mathcal{C}_r^n$, that is, f_{rl}^n, is the true curve for region r. The state S^n is updated in the Bayesian fashion described in Section 1.3.3 and 1.3.4.

Definition A.1.5. Denote by $\bar{f}_r^n$ the estimator of the true curve μ_r on $\boldsymbol{X}$ up to the nth information collection experiment. Keep $\mathbb{L} = \{1, 2, ..., L\}$ as the index set for $\mathcal{C}$.

Then as in Definition 1.3.8, $\bar{f}_r^n$ is formulated as a weighted sum of all the candidate functions in $\mathcal{C}_r^n$:

$$\bar{f}_r^n(x) = \sum_{l \in \mathbb{L}} p_{rl}^n f_{rl}^n(x),\ x \in \boldsymbol{X}$$

and the overall cost for all the R regions given a resource allocation $\mathbf{x} = (x_r)_{\forall r}$ is calculated as

$$\sum_{r=1}^{R} \bar{f}_r^n(x_r) = \sum_{r=1}^{R} \sum_{l \in \mathbb{L}} p_{rl}^n f_{rl}^n(x_r),\ \forall \mathbf{x} \in \mathscr{X}.$$

In the main part of this thesis we have proved the asymptotically strong consistency of $\bar{f}_r^n$ as an estimator of μ_r.

Since we now have more than one region under consideration, in each iteration we need to decide one region to be measured in addition to one choice of alternative. Moreover, the algorithm evaluates only one region-resource combination (r, m) in each measurement. Therefore we have:

Definition A.1.6. The sequence of decisions, denoted by $\boldsymbol{x}_n, n = 0, 1, ..., \boldsymbol{N}-1$, is a series of bi-vectors each of which is a region-resource combination that we choose to measure in the next experiment. We use $\boldsymbol{x}$ when referring to a decision generically. The corresponding measurements are $\hat{y}^{n+1}(\boldsymbol{x}_n), n = 0, 1, ..., \boldsymbol{N} - 1$.

Finally we still keep $\mathscr{F}$ as the set of all finite and non-increasing monotonic functions defined on the discrete set of alternatives $\boldsymbol{X}$.

A.2 Knowledge Gradient

As mentioned in both He and Powell (2016) and Huang et al. (2018), the knowledge gradient at a specific measurement decision $\boldsymbol{x}$ reflects the expected marginal gain on the value of information acquired from evaluating $\boldsymbol{x}$; and in the nth experiment the KG policy selects the decision $\boldsymbol{x}$ that maximizes the KG value. Precisely, the knowledge gradient with discrete prior (KGDP), denoted by $\nu^{KGDP,n}(\boldsymbol{x})$, is the expected improvement between the best estimated value in the nth iteration and the best expected estimated value in the $(n+1)$st iteration if the previous decision is $\boldsymbol{x} = (r, m)$.

A.2.1 Decomposition of KGDP

Based on the Huang et al. (2018, Eqn.(7)), we decompose the KGDP as follows:

$$
\begin{aligned}
&\nu^{KGDP,n}(\boldsymbol{x}) \\
&= \mathbb{E}^n \left[\min_{\mathbf{x}\in\mathscr{X}} \sum_{r'=1}^{R} \bar{f}^n_{r'}(x_{r'}) - \min_{\mathbf{x}\in\mathscr{X}} \sum_{r'=1}^{R} \bar{f}^{n+1}_{r'}(x_{r'}) \middle| S^n, \boldsymbol{x} = (r,m) \right] \\
&= \min_{\mathbf{x}\in\mathscr{X}} \sum_{r'=1}^{R} \bar{f}^n_{r'}(x_{r'}) - \mathbb{E}^n \left[\min_{\mathbf{x}\in\mathscr{X}} \sum_{r'=1}^{R} \bar{f}^{n+1}_{r'}(x_{r'}) \middle| S^n, \boldsymbol{x} = (r,m) \right] \\
&= \min_{\mathbf{x}\in\mathscr{X}} \sum_{r'=1}^{R} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) \\
&\quad - \mathbb{E}_{f^n_{rl'}} \mathbb{E}_{\hat{y}^{n+1}|f^n_{rl'}} \left[\min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) + \sum_{l=1}^{L} p^{n+1}_{rl} f^n_{rl}(x_r) \right\} \middle| S^n, \boldsymbol{x}, \mu_r = f^n_{rl'} \right] \\
&:= A_n - B_n(\boldsymbol{x}),
\end{aligned}
$$

where $\mu_r = f^n_{rl'}$ indicates that $f^n_{rl'}$ is the true curve among the L candidates in $\mathcal{C}^n_r$. Our ultimate goal is to show that the limit of $\nu^{KGDP,n}(\boldsymbol{x})$ is zero as n goes to infinity. The intuition is that as we measure the decision $\boldsymbol{x} = (r, m)$ infinitely many times and gradually learn the true curve, the marginal value of information for measuring $\boldsymbol{x}$ one more time would be diminishing to zero. Note that we drop the $\boldsymbol{x}$ in the notation A_n because this part is irrelevant to the region-resource we measure.

Given that $f^n_{rl'}$ is the true curve, the $(n+1)$th observation, which is a measurement taken on the region r and alternative m, is given by $\hat{y}^{n+1} \sim \mathcal{N}(f^n_{rl'}(m), \sigma^2)$ with density

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(\hat{y} - f^n_{rl'}(m)\right)^2}{2\sigma^2}\right].$$

For simplicity, we drop the time index and use $\hat{y}$ to denote observations generically when the integration (in the later proof) is with respect to the observation. Before calculating $B_n(\boldsymbol{x})$, we denote by $p^{n+1}_{rl|l'}$ the posterior probability of the curve f^n_{rl} (being the true curve) given that $f^n_{rl'}$ is the real true curve. Writing $\hat{y} = f^n_{rl'}(m) + W$ where $W \sim \mathcal{N}(0, \sigma^2)$, we have by Bayes rule:

$$\begin{aligned} p^{n+1}_{rl|l'} &= \frac{\exp\left[-\left(f^n_{rl'}(m) + W - f^n_{rl}(m)\right)^2/(2\sigma^2)\right] p^n_{rl}}{\sum_{l''=1}^{L} \exp\left[-\left(f^n_{rl'}(m) + W - f^n_{rl''}(m)\right)^2/(2\sigma^2)\right] p^n_{rl''}} \\ &= \frac{\exp\left[-\left(\hat{y} - f^n_{rl}(m)\right)^2/(2\sigma^2)\right] p^n_{rl}}{\sum_{l''=1}^{L} \exp\left[-\left(\hat{y} - f^n_{rl''}(m)\right)^2/(2\sigma^2)\right] p^n_{rl''}}. \end{aligned}$$

Note that $p^{n+1}_{rl|l'}$ depends on m which is given by the decision $\boldsymbol{x} = (r, m)$.

We now rewrite $B_n(\boldsymbol{x})$ as:

$$
\begin{aligned}
&B_n(\boldsymbol{x}) \\
&= \mathbb{E}_{f^n_{rl'}} \mathbb{E}_{\hat{y}|f^n_{rl'}} \left[\min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) + \sum_{l=1}^{L} p^{n+1}_{rl} f^n_{rl}(x_r) \right\} \middle| S^n, \boldsymbol{x}, \mu_r = f^n_{rl'} \right] \\
&= \sum_{l'=1}^{L} \left(\mathbb{E}_{\hat{y}|f^n_{rl'}} \left[\min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) + \sum_{l=1}^{L} p^{n+1}_{rl|l'} f^n_{rl}(x_r) \right\} \middle| S^n, \boldsymbol{x} = (r,m) \right] \right) \cdot p^n_{rl'} \\
&= \sum_{l'=1}^{L} \left[\int_{\hat{y}} \min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) + \sum_{l=1}^{L} f^n_{rl}(x_r) \frac{\exp\left[-\left(\hat{y} - f^n_{rl}(m)\right)^2 / (2\sigma^2) \right] p^n_{rl}}{\sum_{l''=1}^{L} \exp\left[-\left(\hat{y} - f^n_{rl''}(m)\right)^2 / (2\sigma^2) \right] p^n_{rl''}} \right\} \right. \\
&\qquad \left. \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(\hat{y} - f^n_{rl'}(m)\right)^2}{2\sigma^2} \right] d\hat{y} \right] \cdot p^n_{rl'} \\
&= \sum_{l'=1}^{L} \left[\int_{\hat{y}} \min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) \sum_{l''=1}^{L} \exp\left[-\frac{\left(\hat{y} - f^n_{rl''}(m)\right)^2}{2\sigma^2} \right] p^n_{rl''} \right. \right. \\
&\qquad \left. + \sum_{l=1}^{L} f^n_{rl}(x_r) \exp\left[-\frac{\left(\hat{y} - f^n_{rl}(m)\right)^2}{2\sigma^2} \right] p^n_{rl} \right\} \\
&\qquad \times \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\left(\hat{y} - f^n_{rl'}(m)\right)^2 / (2\sigma^2) \right]}{\sum_{l''=1}^{L} \exp\left[-\left(\hat{y} - f^n_{rl''}(m)\right)^2 / (2\sigma^2) \right] p^n_{rl''}} d\hat{y} \Bigg] \cdot p^n_{rl'} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{\hat{y}} \left[\min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) \sum_{l''=1}^{L} \exp\left[-\frac{\left(\hat{y} - f^n_{rl''}(m)\right)^2}{2\sigma^2} \right] p^n_{rl''} \right. \right. \\
&\qquad \left. + \sum_{l=1}^{L} f^n_{rl}(x_r) \exp\left[-\frac{\left(\hat{y} - f^n_{rl}(m)\right)^2}{2\sigma^2} \right] p^n_{rl} \right\} \\
&\qquad \times \frac{\sum_{l'=1}^{L} \exp\left[-\left(\hat{y} - f^n_{rl'}(m)\right)^2 / (2\sigma^2) \right] p^n_{rl'}}{\sum_{l''=1}^{L} \exp\left[-\left(\hat{y} - f^n_{rl''}(m)\right)^2 / (2\sigma^2) \right] p^n_{rl''}} \Bigg] d\hat{y} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \left[\min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) \sum_{l''=1}^{L} \exp\left[-\frac{\left(\hat{y} - f^n_{rl''}(m)\right)^2}{2\sigma^2} \right] p^n_{rl''} \right. \right. \\
&\qquad \left. \left. + \sum_{l=1}^{L} f^n_{rl}(x_r) \exp\left[-\frac{\left(\hat{y} - f^n_{rl}(m)\right)^2}{2\sigma^2} \right] p^n_{rl} \right\} \right] d\hat{y}
\end{aligned}
$$

Thus for $r' \in \{1, 2, ..., R\} \backslash \{r\}$,

$$B_n(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \min_{\mathbf{x} \in \mathscr{X}} \left\{ \sum_{r' \neq r} \sum_{l=1}^{L} p_{r'l}^n f_{r'l}^n(x_{r'}) \sum_{l''=1}^{L} \exp\left[-\frac{(\hat{y} - f_{rl''}^n(m))^2}{2\sigma^2} \right] p_{rl''}^n \right.$$
$$\left. + \sum_{l=1}^{L} f_{rl}^n(x_r) \exp\left[-\frac{(\hat{y} - f_{rl}^n(m))^2}{2\sigma^2} \right] p_{rl}^n \right\} d\hat{y} \qquad \text{(A.2.1)}$$

A.2.2 The Limit of KGDP

In this section we establish the main theorem that shows the KGDP asymptotically vanishes to zero.

Theorem A.2.1. *The knowledge gradient with discrete priors at* $\boldsymbol{x} = (r, m)$ *goes to zero as n goes to infinity. That is* $\lim_{n \to \infty} \nu^{KGDP,n}(\boldsymbol{x}) = 0$.

Proof. We will proceed by calculating $\lim_{n \to \infty} A_n$ and $\lim_{n \to \infty} B_n(\boldsymbol{x})$ respectively and then show that $\lim_{n \to \infty} A_n = \lim_{n \to \infty} B_n(\mathbf{x})$. From our previous proof of Theorem 2.4.1, for each region r, the function $f_{rl}^n(x_r)$ converges to $\mu_r(x_r)$ almost surely on $\boldsymbol{X}$ and $\mathcal{C}_r^n$. That is

$$\lim_{n \to \infty} f_{rl}^n(x_r) = \mu_r(x_r) \qquad \text{(A.2.2)}$$

for all $x_r \in \boldsymbol{X}$ and $f_{rl}^n \in \mathcal{C}_r^n$. In addition for each region r, the estimator converges to the true curve almost surely, that is,

$$\lim_{n \to \infty} \bar{f}_r^n(x_r) = \mu_r(x_r) \qquad \text{(A.2.3)}$$

for all $x_r \in \boldsymbol{X}$. We then have for any $\mathbf{x} = (x_r)_{\forall r} \in \mathscr{X}$:

$$\lim_{n\to\infty} A_n = \lim_{n\to\infty}\left[\min_{\mathbf{x}\in\mathscr{X}} \sum_{r'=1}^{R}\sum_{l=1}^{L} p_{r'l}^n f_{r'l}^n(x_{r'})\right]$$

$$= \min_{\mathbf{x}\in\mathscr{X}}\left\{\lim_{n\to\infty} \sum_{r'=1}^{R}\sum_{l=1}^{L} p_{r'l}^n f_{r'l}^n(x_{r'})\right\} \tag{A.2.4}$$

$$= \min_{\mathbf{x}\in\mathscr{X}}\left\{\lim_{n\to\infty} \sum_{r'=1}^{R} \bar{f}_{r'}^n(x_{r'})\right\}$$

$$= \min_{\mathbf{x}\in\mathscr{X}}\left\{\sum_{r'=1}^{R} \lim_{n\to\infty}\bar{f}_{r'}^n(x_{r'})\right\}$$

$$= \min_{\mathbf{x}\in\mathscr{X}}\left\{\sum_{r'=1}^{R} \mu_{r'}(x_{r'})\right\}. \tag{A.2.5}$$

Notice that here (A.2.4) is valid since all f_{rl}^n and μ_r are in $\mathscr{F}$ and the minimum operator is taken over a finite set.

When finding the limit of $B_n(\boldsymbol{x})$, one key step is to interchange the limit and the integral operator. Recall that $f_{rl}^n \in \mathcal{C}_r^n$ and $f_{r'l}^n \in \mathcal{C}_{r'}^n$, we define

$$\psi^n(\hat{y}) = \min_{\mathbf{x}\in\mathscr{X}}\left\{\sum_{r'\neq r}\sum_{l=1}^{L} p_{r'l}^n f_{r'l}^n(x_{r'})\cdot \sum_{l''=1}^{L}\exp\left[-\frac{(\hat{y}-f_{rl''}^n(m))^2}{2\sigma^2}\right] p_{rl''}^n \right.$$
$$\left. + \sum_{l=1}^{L} f_{rl}^n(x_r)\exp\left[-\frac{(\hat{y}-f_{rl}^n(m))^2}{2\sigma^2}\right] p_{rl}^n\right\}$$

which is everything inside the integral sign in (A.2.1), and

$$\phi_{\mathbf{x}_0}^n(\hat{y}) = \sum_{r'\neq r}\sum_{l=1}^{L} p_{r'l}^n f_{r'l}^n(x_{r'}^0)\cdot \sum_{l''=1}^{L}\exp\left[-\frac{(\hat{y}-f_{rl''}^n(m))^2}{2\sigma^2}\right] p_{rl''}^n$$
$$+ \sum_{l=1}^{L} f_{rl}^n(x_r^0)\exp\left[-\frac{(\hat{y}-f_{rl}^n(m))^2}{2\sigma^2}\right] p_{rl}^n$$

where the vector $\mathbf{x}_0 \in \mathscr{X}$ is a particular realization of the resource allocation to the R regions and $\mathbf{x}_0 = (x_r^0)_{r=1}^R$. Here , we choose the $\mathbf{x}_0 \in \mathscr{X}$ such that

$$\mathbf{x}_0 \quad = \quad \arg\max_{\mathbf{x}\in\mathscr{X}} \left|\phi_{\mathbf{x}}^n\left(\hat{y}\right)\right|$$

which implies that

$$\left|\psi^n\left(\hat{y}\right)\right| \leq \left|\phi_{\mathbf{x}_0}^n\left(\hat{y}\right)\right|. \tag{A.2.6}$$

For each $r \in \{1, 2, ..., R\}$ and $l \in \{1, 2, ..., L\}$, p_{rl}^n is bounded between 0 and 1. Thus there exists $p_{rl}^* \in [0, 1]$ and a subsequence $\left\{n_k^{rl}\right\}_{k=1}^{\infty}$ along which

$$\lim_{k\to\infty} p_{rl}^{n_k^{rl}} = p_{rl}^*, \; \forall r \in \{1, 2, ..., R\}, \; \forall l \in \{1, 2, ..., L\}.$$

Furthermore there exists a general subsequence

$$\{n_k\}_{k=1}^{\infty} \subset \cup_{r=1}^{R} \cup_{l=1}^{L} \left\{n_k^{rl}\right\}_{k=1}^{\infty}$$

along which for all $r \in \{1, 2, ..., R\}$ and all $l \in \{1, 2, ..., L\}$

$$\lim_{k\to\infty} p_{rl}^{n_k} = p_{rl}^*.$$

Along this general subsequence $\{n_k\}_{k=1}^{\infty}$, since each function $f_{rl}^n(x_r)$ converges to $\mu_r(x_r)$, we also have $\lim_{k\to\infty} f_{rl}^{n_k}(x_r) = \mu_r(x_r)$. Now we define

$$
\begin{aligned}
&\psi\left(\hat{y}\right) \\
&= \min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \mu_{r'}(x_{r'}) \cdot \exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2} \right] + \mu_r(x_r) \exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2} \right] \right\} \\
&= \min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'=1}^{R} \mu_{r'}(x_{r'}) \cdot \exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2} \right] \right\}
\end{aligned}
$$

and

$$
\begin{aligned}
&\phi_{\mathbf{x}_0}\left(\hat{y}\right) \\
&= \sum_{r'\neq r} \mu_{r'}(x^0_{r'}) \cdot \exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2} \right] + \mu_r(x^0_r) \exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2} \right] \\
&= \sum_{r'=1}^{R} \mu_{r'}(x^0_{r'}) \cdot \exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2} \right]
\end{aligned}
$$

We first apply the dominated convergence theorem in Koralov and Sinai (2007, Thm.3.26) (or Çınlar (2011, Thm.4.16)) to show that

$$
\lim_{n\to\infty} \int_{-\infty}^{+\infty} \phi^n_{\mathbf{x}_0}\left(\hat{y}\right) d\hat{y} = \int_{-\infty}^{+\infty} \lim_{n\to\infty} \phi^n_{\mathbf{x}_0}\left(\hat{y}\right) d\hat{y} = \int_{-\infty}^{+\infty} \phi_{\mathbf{x}_0}\left(\hat{y}\right) d\hat{y}. \tag{A.2.7}
$$

Second, we apply the dominated convergence theorem in Kallenberg (2006, Thm.1.21) together with $\psi^n\left(\hat{y}\right) \to \psi\left(\hat{y}\right)$, $\phi^n_{\mathbf{x}_0}\left(\hat{y}\right) \to \phi_{\mathbf{x}_0}\left(\hat{y}\right)$, $\left|\psi^n\left(\hat{y}\right)\right| \leq \phi^n_{\mathbf{x}_0}\left(\hat{y}\right)$ and (A.2.7), showing that

$$
\lim_{n\to\infty} \int_{-\infty}^{+\infty} \psi^n\left(\hat{y}\right) d\hat{y} = \int_{-\infty}^{+\infty} \lim_{n\to\infty} \psi^n\left(\hat{y}\right) d\hat{y} = \int_{-\infty}^{+\infty} \psi\left(\hat{y}\right) d\hat{y}. \tag{A.2.8}
$$

Finally we use (A.2.8) to calculate the limit of $B_n(\mathbf{x})$. We begin with the proof of (A.2.7).

We have shown that $\lim_{n\to\infty} f_{rl}^n(x_r) = \mu_r(x_r)$ and $\lim_{n\to\infty} \bar{f}_r^n(x_r) = \mu_r(x_r)$ for all $x_r \in \boldsymbol{X}$ and $f_{rl}^n \in \mathcal{C}_r^n$. If we further define

$$\bar{f}_r^{e_n} = \sum_{l''=1}^{L} \exp\left[-\frac{(\hat{y} - f_{rl''}^n(m))^2}{2\sigma^2}\right] p_{rl''}^n$$

and

$$\bar{f}_r^{n,e_n}(x_r) = \sum_{l=1}^{L} f_{rl}^n(x_r) \exp\left[-\frac{(\hat{y} - f_{rl}^n(m))^2}{2\sigma^2}\right] p_{rl}^n,$$

then $\exp\left[-(\hat{y} - f_{rl''}^n(m))^2/(2\sigma^2)\right]$ is a continuous function of the value of $f_{rl''}^n(m)$, and the product $f_{rl}^n(x_r)\exp\left[-(\hat{y} - f_{rl}^n(m))^2/(2\sigma^2)\right]$ is a continuous function of the value of $f_{rl}^n(x_r)$ and $f_{rl}^n(m)$. Thus we have

$$\lim_{n\to\infty} \exp\left[-\frac{(\hat{y} - f_{rl''}^n(m))^2}{2\sigma^2}\right] = \exp\left[-\frac{(\hat{y} - \mu_r(m))^2}{2\sigma^2}\right]$$

and

$$\lim_{n\to\infty} f_{rl}^n(x_r) \exp\left[-\frac{(\hat{y} - f_{rl}^n(m))^2}{2\sigma^2}\right] = \mu_r(x_r) \exp\left[-\frac{(\hat{y} - \mu_r(m))^2}{2\sigma^2}\right].$$

Because $\bar{f}_r^{e_n}$ and $\bar{f}_r^{n,e_n}(x_r)$ can be viewed as the weighted sum of

$$\exp\left[-\frac{(\hat{y} - f_{rl''}^n(m))^2}{2\sigma^2}\right]$$

and

$$f_{rl}^{n}(x_r)\exp\left[-\frac{\left(\hat{y}-f_{rl}^{n}(m)\right)^2}{2\sigma^2}\right]$$

respectively with weights sum to one (those p_{rl}^{n}'s), applying the same technique in the consistency proof of Theorem 2.4.1 we would have

$$\lim_{n\to\infty}\bar{f}_r^{e_n}=\exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2}\right] \tag{A.2.9}$$

and

$$\lim_{n\to\infty}\bar{f}_r^{n,e_n}(x_r)=\mu_r(x_r)\exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2}\right]. \tag{A.2.10}$$

With the above two equations (A.2.9) and (A.2.3), it is easy to see that

$$\begin{aligned}
\lim_{n\to\infty}\phi_{\mathbf{x}_0}^{n}\left(\hat{y}\right) &= \lim_{n\to\infty}\left\{\sum_{r'\neq r}\sum_{l=1}^{L}p_{r'l}^{n}f_{r'l}^{n}(x_{r'}^{0})\cdot\sum_{l''=1}^{L}\exp\left[-\frac{\left(\hat{y}-f_{rl''}^{n}(m)\right)^2}{2\sigma^2}\right]p_{rl''}^{n}\right. \\
&\qquad \left.+\sum_{l=1}^{L}f_{rl}^{n}(x_r^{0})\exp\left[-\frac{\left(\hat{y}-f_{rl}^{n}(m)\right)^2}{2\sigma^2}\right]p_{rl}^{n}\right\} \qquad \text{(A.2.11)} \\
&= \lim_{n\to\infty}\left\{\left[\sum_{r'\neq r}\bar{f}_{r'}^{n}(x_{r'}^{0})\right]\cdot\bar{f}_r^{e_n}+\bar{f}_r^{n,e_n}(x_r^{0})\right\} \\
&= \lim_{n\to\infty}\sum_{r'\neq r}\bar{f}_{r'}^{n}(x_{r'}^{0})\cdot\lim_{n\to\infty}\bar{f}_r^{e_n}+\lim_{n\to\infty}\bar{f}_r^{n,e_n}(x_r^{0}) \\
&= \sum_{r'\neq r}\mu_{r'}(x_{r'}^{0})\cdot\exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2}\right]+\mu_r(x_r^{0})\exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2}\right] \\
&= \sum_{r'=1}^{R}\mu_{r'}(x_{r'}^{0})\cdot\exp\left[-\frac{\left(\hat{y}-\mu_r(m)\right)^2}{2\sigma^2}\right] \\
&= \phi_{\mathbf{x}_0}\left(\hat{y}\right). \qquad \text{(A.2.12)}
\end{aligned}$$

Note that the above equation (A.2.12) does not depend on the specific $\mathbf{x}_0$ and it

holds for any $\mathbf{x} \in \mathscr{X}$. Therefore we have

$$\lim_{n\to\infty} \phi_{\mathbf{x}}^{n}(\hat{y}) = \phi_{\mathbf{x}}(\hat{y}). \tag{A.2.13}$$

Our next step would be finding an function $\tilde{\phi}_{\mathbf{x}_0}(\hat{y})$ that is integrable with respect to $\hat{y}$ such that $\left|\phi_{\mathbf{x}_0}^{n}(\hat{y})\right| \leq \tilde{\phi}_{\mathbf{x}_0}(\hat{y})$.

By (A.2.2) and (A.2.3), there exists N_ε such that for all l and region r when $n > N_\varepsilon$, we have

$$|f_{rl}^{n}(x_r) - \mu_r(x_r)| < \varepsilon, \qquad \left|\bar{f}_{r}^{n}(x_r) - \mu_r(x_r)\right| < \varepsilon$$

and

$$|f_{rl}^{n}(m) - \mu_r(m)| < \varepsilon.$$

Together with

$$\mu_r(1) \geq \mu(2) \geq ... \geq \mu(M),$$

the previous three inequalities imply that

$$f_{rl}^{n}(x_r) \leq \mu_r(x_r) + \varepsilon \leq \mu_r(1) + \varepsilon, \quad \bar{f}_{r}^{n}(x_r) \leq \mu_r(x_r) + \varepsilon \leq \mu_r(1) + \varepsilon$$

and

$$\mu_r(m) - \varepsilon < f_{rl}^{n}(m) < \mu_r(m) + \varepsilon.$$

The key to the construction of $\tilde{\phi}_{\mathbf{x}_0}(\hat{y})$ lies in the way we bound

$$\exp\left[-\frac{(\hat{y}-f_{rl}^n(m))^2}{2\sigma^2}\right].$$

Up to a scale factor $(\sqrt{2\pi}\sigma)^{-1}$ the $\exp\left[-(\hat{y}-f_{rl}^n(m))^2/(2\sigma^2)\right]$ behaves like the normal density curve of $\hat{y}$ that is centered at $\hat{y}=f_{rl}^n(m)$. Therefore when $n>N_\varepsilon$, geometrically each $\exp\left[-(\hat{y}-f_{rl}^n(m))^2/(2\sigma^2)\right]$ has the same shape and with its axis of symmetry at $\hat{y}=f_{rl}^n(m)$, where $f_{rl}^n(m)\in(\mu_r(m)-\varepsilon,\mu_r(m)+\varepsilon)$. In addition each $\exp\left[-(\hat{y}-f_{rl}^n(m))^2/(2\sigma^2)\right]$ is always positive and reaching its maximum value 1 at $\hat{y}=f_{rl}^n(m)$. Thus we can see that:

- for $\hat{y}>\mu_r(m)+\varepsilon$, we have

$$\exp\left[-(\hat{y}-f_{rl}^n(m))^2/(2\sigma^2)\right]<\exp\left[-(\hat{y}-(\mu_r(m)+\varepsilon))^2/(2\sigma^2)\right];$$

- for $\hat{y}<\mu_r(m)-\varepsilon$, we have

$$\exp\left[-(\hat{y}-f_{rl}^n(m))^2/(2\sigma^2)\right]<\exp\left[-(\hat{y}-(\mu_r(m)-\varepsilon))^2/(2\sigma^2)\right];$$

- for $\hat{y}\in[\mu_r(m)-\varepsilon,\mu_r(m)+\varepsilon]$, we have

$$\exp\left[-(\hat{y}-f_{rl}^n(m))^2/(2\sigma^2)\right]\leq 1.$$

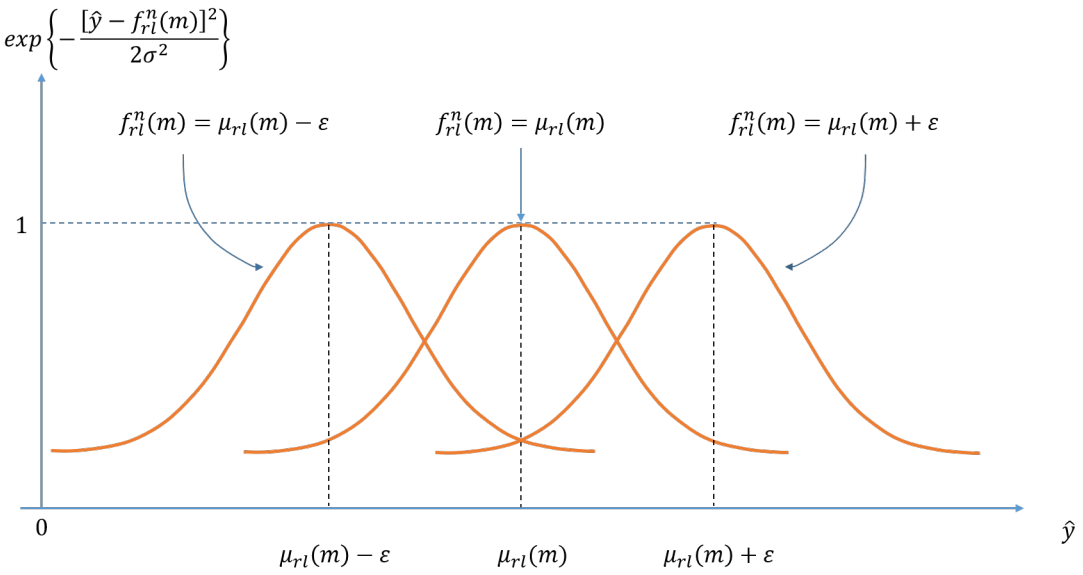

Figure A.2.1: $\exp\left[-\frac{\left(\hat{y}-f_{rl}^n(m)\right)^2}{2\sigma^2}\right]$ centered at $f_{rl}^n(m)\in(\mu_r(m)-\varepsilon,\mu_r(m)+\varepsilon)$

Based on this fact we set

$$\beta(\hat{y})=\begin{cases}\exp\left[-\left(\hat{y}-(\mu_r(m)+\varepsilon)\right)^2/(2\sigma^2)\right], & \hat{y}>\mu_r(m)+\varepsilon,\\ 1, & \hat{y}\in[\mu_r(m)-\varepsilon,\mu_r(m)+\varepsilon],\\ \exp\left[-\left(\hat{y}-(\mu_r(m)-\varepsilon)\right)^2/(2\sigma^2)\right], & \hat{y}<\mu_r(m)-\varepsilon.\end{cases}$$

Graphical illustrations of the construction of $\beta(\hat{y})$ are in Figure A.2.1 and A.2.2 Then $0<\exp\left[-\left(\hat{y}-f_{rl}^n(m)\right)^2/(2\sigma^2)\right]<\beta(\hat{y})$ for $\hat{y}\in(-\infty,+\infty)$. Meanwhile for the previous ε, there exists K_ε such that for $k>K_\varepsilon$ we have $p_{rl}^{n_k}<p_{rl}^*+\varepsilon$. Thus:

$$\begin{aligned}&\left|\phi_{\mathbf{x}_0}^n(\hat{y})\right|\\&=\left|\sum_{r'\neq r}\bar{f}_{r'}^n(x_r^0)\cdot\sum_{l''=1}^{L}\exp\left[-\frac{(\hat{y}-f_{rl''}^n(m))^2}{2\sigma^2}\right]p_{rl''}^n\right.\\&\left.+\sum_{l=1}^{L}f_{rl}^n(x_r^0)\exp\left[-\frac{(\hat{y}-f_{rl}^n(m))^2}{2\sigma^2}\right]p_{rl}^n\right|\end{aligned}$$

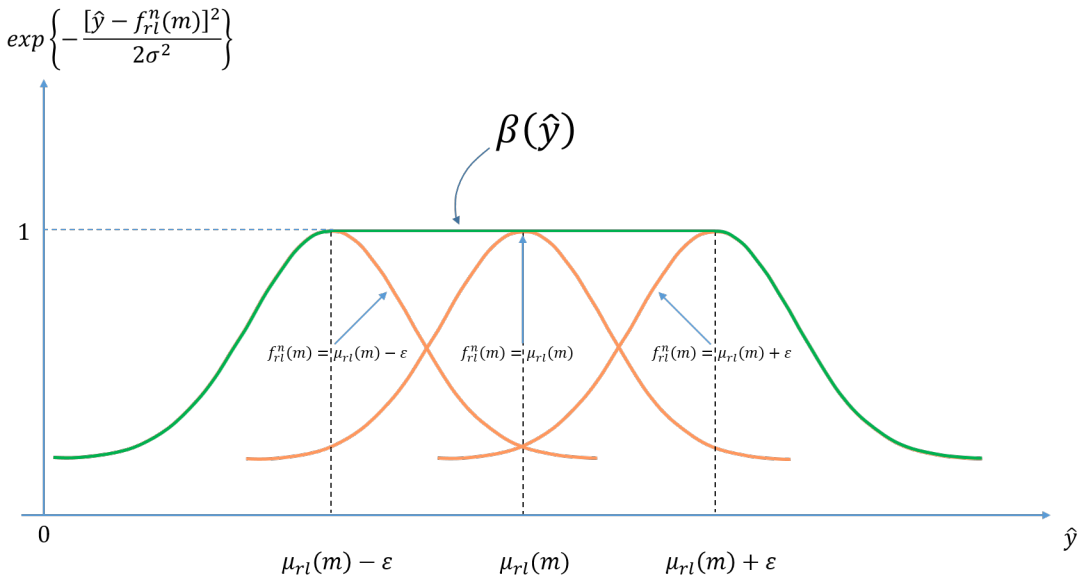

Figure A.2.2: Construction of $\beta(\hat{y})$

$$
\begin{aligned}
&\leq \left| \sum_{r' \neq r} (\mu_r(1)+\varepsilon) \cdot \sum_{l''=1}^{L} \beta(\hat{y})\,(p_{rl}^{*}+\varepsilon) + \sum_{l=1}^{L} (\mu_r(1)+\varepsilon)\beta(\hat{y})\,(p_{rl}^{*}+\varepsilon) \right| \\
&= \left| \sum_{r' \neq r} (\mu_{r'}(1)+\varepsilon) \cdot \beta(\hat{y})(1+L\varepsilon) + (\mu_r(1)+\varepsilon)\beta(\hat{y})(1+L\varepsilon) \right| \\
&= \left| \sum_{r''=1}^{R} (\mu_{r''}(1)+\varepsilon)\beta(\hat{y})(1+L\varepsilon) \right| \\
&:= \tilde{\phi}(\hat{y}).
\end{aligned}
$$

Remembering that $\beta(\hat{y}) > 0$ for all $\hat{y} \in (-\infty, +\infty)$, we can see that $\tilde{\phi}(\hat{y})$ is integrable with respect to $\hat{y}$ because

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} \left|\tilde{\phi}(\hat{y})\right| d\hat{y} \\
&= \int_{-\infty}^{+\infty} \left| \sum_{r''=1}^{R} (\mu_{r''}(1)+\varepsilon)(1+L\varepsilon)\beta(\hat{y}) \right| d\hat{y} \\
&\leq \int_{-\infty}^{+\infty} \sum_{r''=1}^{R} \left| (\mu_{r''}(1)+\varepsilon)(1+L\varepsilon)\beta(\hat{y}) \right| d\hat{y}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{r''=1}^{R} |(\mu_{r''}(1)+\varepsilon)(1+L\varepsilon)| \cdot \int_{-\infty}^{+\infty} \beta(\hat{y}) d\hat{y} \\
&= \sum_{r''=1}^{R} |(\mu_{r''}(1)+\varepsilon)(1+L\varepsilon)| \sqrt{2\pi}\sigma \left\{ \int_{\mu_r(m)+\varepsilon}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\hat{y}-(\mu_r(m)+\varepsilon))^2}{2\sigma^2} \right] d\hat{y} \right. \\
&\quad \left. + \int_{-\infty}^{\mu_r(m)-\varepsilon} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\hat{y}-(\mu_r(m)-\varepsilon))^2}{2\sigma^2} \right] d\hat{y} + \int_{\mu_r(m)-\varepsilon}^{\mu_r(m)+\varepsilon} \frac{1}{\sqrt{2\pi}\sigma} \cdot 1 d\hat{y} \right\} \\
&\leq \sum_{r''=1}^{R} |(\mu_{r''}(1)+\varepsilon)(1+L\varepsilon)| \sqrt{2\pi}\sigma \left\{ \frac{1}{2} + \frac{1}{2} + \frac{2\varepsilon}{\sqrt{2\pi}\sigma} \right\} \\
&= (1+L\varepsilon)\left(\sqrt{2\pi}\sigma + 2\varepsilon \right) \cdot \sum_{r''=1}^{R} |(\mu_{r''}(1)+\varepsilon)| \qquad \text{(A.2.14)}
\end{aligned}
$$

where the quantity in (A.2.14) is finite regardless of the value of ε. Thus, $\left|\phi^n_{\mathbf{x}_0}(\hat{y})\right| \leq \tilde{\phi}(\hat{y})$ and $\tilde{\phi}(\hat{y})$ is integrable. Combined with $\phi^n_{\mathbf{x}_0}(\hat{y}) \to \phi_{\mathbf{x}_0}(\hat{y})$, by the dominated convergence theorem in Çınlar (2011, Thm.4.16) or Koralov and Sinai (2007, Thm.3.26), we showed (A.2.7). That is, $\lambda\left[\phi^n_{\mathbf{x}_0}(\hat{y})\right] \to \lambda\left[\phi_{\mathbf{x}_0}(\hat{y})\right]$ where λ denotes the Lebesgue integral.

Next, since all f^n_{rl} and μ_r are in $\mathscr{F}$, $f^n_{rl}(m)$ and $f^n_{rl''}(m)$ are irrelevant with $\mathbf{x}$ and the minimum operator is over a finite set, we may interchange the "limit" and "min" operator. Then applying equation (A.2.13) implies

$$
\begin{aligned}
&\lim_{n\to\infty} \psi^n(\hat{y}) \\
&= \lim_{n\to\infty} \min_{\mathbf{x}\in\mathscr{X}} \{\phi^n_{\mathbf{x}}(\hat{y})\} \\
&= \min_{\mathbf{x}\in\mathscr{X}} \left\{ \lim_{n\to\infty} \phi^n_{\mathbf{x}}(\hat{y}) \right\} \\
&= \min_{\mathbf{x}\in\mathscr{X}} \{\phi_{\mathbf{x}}(\hat{y})\} \\
&= \psi(\hat{y}), \qquad \text{(A.2.15)}
\end{aligned}
$$

that is, $\lim_{n\to\infty}\psi^n(\hat{y}) = \psi(\hat{y})$. Therefore, we have $|\psi^n(\hat{y})| \leq \left|\phi^n_{\mathbf{x}_0}(\hat{y})\right|$ (by (A.2.6)) for all n and such that $\psi^n(\hat{y}) \to \psi(\hat{y})$ (by (A.2.15)), $\phi^n_{\mathbf{x}_0}(\hat{y}) \to \phi_{\mathbf{x}_0}(\hat{y})$ (by (A.2.12)), and $\lambda\left[\phi^n_{\mathbf{x}_0}(\hat{y})\right] \to \lambda\left[\phi_{\mathbf{x}_0}(\hat{y})\right]$. By the dominated convergence theorem of Kallenberg (2006, Thm.1.21), we have

$$\lambda\left[\psi^n(\hat{y})\right] \to \lambda\left[\psi(\hat{y})\right]$$

which is equivalent to (A.2.8).

We now calculate the limit of $B_n(\boldsymbol{x})$ as follows:

$$\begin{aligned}
&\lim_{n\to\infty} B_n(\boldsymbol{x}) \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'\neq r} \sum_{l=1}^{L} p^n_{r'l} f^n_{r'l}(x_{r'}) \cdot \sum_{l''=1}^{L} \exp\left[-\frac{(\hat{y} - f^n_{rl''}(m))^2}{2\sigma^2}\right] p^n_{rl''} \right. \\
&\qquad\qquad \left. + \sum_{l=1}^{L} f^n_{rl}(x_r) \exp\left[-\frac{(\hat{y} - f^n_{rl}(m))^2}{2\sigma^2}\right] p^n_{rl} \right\} d\hat{y} \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \psi^n(\hat{y})\, d\hat{y} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \lim_{n\to\infty} \psi^n(\hat{y})\, d\hat{y} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \psi(\hat{y})\, d\hat{y} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'=1}^{R} \mu_{r'}(x_{r'}) \cdot \exp\left[-\frac{(\hat{y} - \mu_r(m))^2}{2\sigma^2}\right] \right\} d\hat{y} \\
&= \min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'=1}^{R} \mu_{r'}(x_{r'}) \right\} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\hat{y} - \mu_r(m))^2}{2\sigma^2}\right] d\hat{y} \\
&= \min_{\mathbf{x}\in\mathscr{X}} \left\{ \sum_{r'=1}^{R} \mu_{r'}(x_{r'}) \right\}. \qquad \text{(A.2.16)}
\end{aligned}$$

Comparing (A.2.5) and (A.2.16), we have $\lim_{n\to\infty} A_n = \lim_{n\to\infty} B_n(\boldsymbol{x})$. Hence,

$$\begin{aligned}\lim_{n\to\infty} \nu^{KGDP,n}(\boldsymbol{x}) &= \lim_{n\to\infty} [A_n - B_n(\boldsymbol{x})] \\ &= \lim_{n\to\infty} A_n - \lim_{n\to\infty} B_n(\boldsymbol{x}) \\ &= 0\end{aligned}$$

We conclude the proof that the KGDP evaluated at a particular region-allocation combination $\boldsymbol{x} = (r, m)$ diminishes to zero if we are able to query this $\boldsymbol{x}$ infinitely many times. i.e., $n \to \infty$. □

Appendix B: Isotonic Regression with Finite Measurements

Isotonic Estimator With Finite Measurements

In this section, we discuss the theoretical limiting behavior of isotonic regression estimator when not every $x \in \boldsymbol{X}$ is measured infinitely many times.

B.1 Motivation and Objective Function

In our previous proof of the consistency of the estimator of the cost function, we made a critical assumption that the knowledge gradient (KG) algorithm would measure each alternative $x \in \boldsymbol{X} = \{1, 2, ..., M\}$ infinitely often. In other words, during the information collection process when the number of iterations n goes to infinity, the number of observations on x, namely $w^n(x)$, also goes to infinity. Although such a property is desired for KG and any other learning algorithms beyond KG, it may not always be fulfilled or known to be inherent with the learning algorithms that researchers choose to apply. Learning the properties of isotonic regression when partial alternatives are queried finitely many times can be meaningful for future research in those learning algorithms. Therefore in this section we examine this issue and study the behavior of the isotonic regression estimator under the scenario that

not all of the alternatives are measured infinitely many times. We begin with some key notations and a revisit to the definition of the isotonic regression estimator.

Definition B.1.1. Denote by $\boldsymbol{X}_\infty$ the set of alternatives in $\boldsymbol{X}$ that are measured infinitely often (i.o.) under KG, that is: $\boldsymbol{X}_\infty = \{x \in \boldsymbol{X} : w^n(x) \to \infty \text{ as } n \to \infty\}$. Then denote by $\boldsymbol{X}_\infty^c$ the set of alternatives that are measured finitely often under KG. That is, $\boldsymbol{X}_\infty^c = \{x \in \boldsymbol{X} : w^n(x) \text{ is finite as } n \to \infty\}$. Therefore $\boldsymbol{X}_\infty$ and $\boldsymbol{X}_\infty^c$ form a partition on $\boldsymbol{X}$ with $\boldsymbol{X} = \boldsymbol{X}_\infty \bigcup \boldsymbol{X}_\infty^c$ and $\boldsymbol{X}_\infty \bigcap \boldsymbol{X}_\infty^c = \emptyset$.

Definition B.1.2. Because isotonic regression is an minimization problem subject to some order constraints, we denote by $\mathscr{F}$ the set of all finite and non-increasing monotonic functions defined on the discrete set of alternatives $\boldsymbol{X}$. From now on to the end of this chapter, our discussion is restricted to all $x \in \boldsymbol{X}$ and $f(x) \in \mathscr{F}$.

Remark. In addition to the monotonicity and discreteness, non-negativity is usually required for some objective functions like an operational cost function in the motivating example described in Section 1.4. However, for the general purpose of learning asymptotic properties of isotonic regression estimator, our discussion and conclusions do not depend on non-negativity hence we do not restrict $f(x)$ to be non-negative.

We still keep $\bar{y}^n = \bar{y}^n(x) = (1/w^n(x)) \sum_{j=1}^{w^n(x)} \hat{y}_j(x), x \in \boldsymbol{X}$ as the sample mean of all the measurements taken on x up to the nth iteration, i.e., $\widehat{\boldsymbol{H}}^n$; and $\mu(x), x \in \boldsymbol{X}$ as the value of the true cost function at x. The sample isotonic regression estimator of $\bar{y}^n$ was previously denoted by $(\bar{y}^n)^*$ when all $x \in \boldsymbol{X}$ are measured infinitely often. We now denoted by $\hat{f}_n^*(x)$ the sample isotonic regression estimator of $\bar{y}^n$ when some

alternatives in $\boldsymbol{X}$ are measured finitely many times. Then $\hat{f}_n^*(x)$ is defined as:

$$\begin{aligned}\hat{f}_n^*(x) &= \arg\min_{f(x)} \sum_{x\in\boldsymbol{X}} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x), \\ &\quad \text{subject to} \quad f(x) \in \mathscr{F}.\end{aligned}$$

We can further decompose the objective function based on the number of measurements of each alternative:

$$\begin{aligned}\hat{f}_n^*(x) &= \arg\min_{f(x)} \sum_{x\in\boldsymbol{X}} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x) \\ &= \arg\min_{f(x)} \left\{ \sum_{x\in\boldsymbol{X}_\infty} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x) + \sum_{x\in\boldsymbol{X}_\infty^c} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x) \right\} \\ &:= \arg\min_{f(x)} \left\{ S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}_\infty^c} \right\} \\ &\text{subject to} \quad f(x) \in \mathscr{F}.\end{aligned}$$

Our goal is to show that for $x \in \boldsymbol{X}_\infty$, $\lim_{n\to\infty} \hat{f}_n^*(x) = \mu(x)$ and for $x \in \boldsymbol{X}_\infty^c$, $\lim_{n\to\infty} \hat{f}_n^*(x)$ exists.

Heuristically, we explain the results by looking at the behavior of $S^n_{\boldsymbol{X}_\infty}$ and $S^n_{\boldsymbol{X}_\infty^c}$ individually, then jointly when finding the optimal minimized value. First, observe that $S^n_{\boldsymbol{X}_\infty^c}$ would converge to some positive random variable. The reason is that in this case we are not measuring x's i.o. so that after a certain time, we would not update $\bar{y}^n(x)$ and $w^n(x)$ any more, and they would converge to some respective constants. Thus $S^n_{\boldsymbol{X}_\infty^c}$ would eventually be equal to a constant number depending on the values of $f(x)$. Second, since $\bar{y}^n(x)$ is a strongly consistent estimator of $\mu(x)$, af-

ter some time large enough, $S^n_{\boldsymbol{X}_\infty}$ would be very close to $\sum_{x\in\boldsymbol{X}_\infty}\left[\mu(x)-f(x)\right]^2 w^n(x)$ but meanwhile, $w^n(x)$ would be very large. Third, note that each term in the summation of $S^n_{\boldsymbol{X}_\infty}$ and $S^n_{\boldsymbol{X}^c_\infty}$ is non-negative; then, to minimize $S^n_{\boldsymbol{X}_\infty}+S^n_{\boldsymbol{X}^c_\infty}$ under isotonic constraints we need to think which of $S^n_{\boldsymbol{X}_\infty}$ or $S^n_{\boldsymbol{X}^c_\infty}$ can we either force to zero, or try to minimize first. As we discussed, as $n\to\infty$, $S^n_{\boldsymbol{X}^c_\infty}$ is very likely to end up with a finite number; and each summand in $S^n_{\boldsymbol{X}_\infty}$ would also be very large as long as $\bar{y}^n(x)-f(x)\neq 0$. Thus the value of $S^n_{\boldsymbol{X}_\infty}$ would become very large and there will be a time large enough such that $S^n_{\boldsymbol{X}_\infty}\gg S^n_{\boldsymbol{X}^c_\infty}$, meaning the value of the objective function would be dominated by $S^n_{\boldsymbol{X}_\infty}$. Therefore, to minimize the objective function we need to set $\hat{f}^*_n(x)=\bar{y}^n(x),\forall x\in\boldsymbol{X}_\infty$, which eventually converges to the true function value $\mu(x)$. We will proceed by first exploring the behavior of $S^n_{\boldsymbol{X}^c_\infty}$ and $S^n_{\boldsymbol{X}_\infty}$respectively, then make comparisons of these two partial sums, which followed by derivations in regard with the minimization under ordered constraints. The final main conclusion will be stated as Theorem (B.4.1).

B.2 Behavior of the Sum with x Measured Finitely Often

We prove the following limiting result for

$$S^n_{\boldsymbol{X}^c_\infty}=\sum_{x\in\boldsymbol{X}^c_\infty}\left[\bar{y}^n(x)-f(x)\right]^2 w^n(x)$$

with $f(x)\in\mathscr{F}$.

Lemma B.2.1.

As $n \to \infty$, the limit of $S^n_{\boldsymbol{X}^c_\infty}$ exists and is a function of $x \in \boldsymbol{X}^c_\infty$.

Proof. By the definition of $\boldsymbol{X}^c_\infty$, there exists a time N_1 such that for all $n > N_1$, $w^n(x) = \tilde{w}(x)$ and $\bar{y}^n(x) = \tilde{y}(x)$, where $\tilde{w}(x)$ and $\tilde{y}(x)$ are some constant numbers which only depend on x. Additionally, $f(x)$ is finite for all $x \in \boldsymbol{X}^c_\infty$ and the cardinality of $\boldsymbol{X}^c_\infty$ is also finite. Therefore, $S^n_{\boldsymbol{X}^c_\infty}$ is a summation of finitely many terms and we have:

$$\begin{aligned}
\lim_{n\to\infty} S^n_{\boldsymbol{X}^c_\infty} &= \lim_{n\to\infty} \sum_{x\in\boldsymbol{X}^c_\infty} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x) \\
&= \sum_{x\in\boldsymbol{X}^c_\infty} \left[\lim_{n\to\infty} \bar{y}^n(x) - f(x)\right]^2 \cdot \lim_{n\to\infty} w^n(x) \\
&= \sum_{x\in\boldsymbol{X}^c_\infty} \left[\tilde{y}(x) - f(x)\right]^2 \tilde{w}(x) \\
&:= \tilde{S}_{\boldsymbol{X}^c_\infty}
\end{aligned}$$

Clearly, the limit $\tilde{S}_{\boldsymbol{X}^c_\infty}$ is finite as a function of all $x \in \boldsymbol{X}^c_\infty$. □

B.3 Behavior of the Sum with x Measured Infinitely Often

In Lemma B.2.1 we showed that after time N_1, the value of $S^n_{\boldsymbol{X}^c_\infty}$ will stop being updated and eventually converge to a finite quantity only depends on $x \in \boldsymbol{X}^c_\infty$ and $f(x) \in \mathscr{F}$. We prove the following lemma:

Lemma B.3.1.

There exists a time τ after which for any $x \in \boldsymbol{X}$ and $f(x) \in \mathscr{F}$,

1) If the true function μ is strictly monotonic, then $\tilde{S}_{\boldsymbol{X}^c_\infty}$ is the lower bound of $\left(S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}^c_\infty}\right)$. The lower bound is achieved at $f(x) = \hat{f}^*_n(x) = \bar{y}^n(x)$, $x \in \boldsymbol{X}_\infty$ for all $n > \tau$.

2) If the true function μ is strictly monotonic, i.e., $\mu(x) \geq \mu(x')$ for $x < x'$, then the lower bound of $\left(S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}^c_\infty}\right)$ is achieved at $f(x) = \hat{f}^*_n(x) = \tilde{f}_n(x)$, $x \in \boldsymbol{X}_\infty$ for all $n > \tau$, where $\tilde{f}_n(x)$ is the isotonic regression estimator of the function $\bar{y}^n(x)$ on the set $\boldsymbol{X}_\infty$.

Proof. (1) Assuming $\mu(x) > \mu(x'), \forall x, x' \in \boldsymbol{X}$ and $x < x'$. Since $\bar{y}^n(x)$ is a consistent estimator of $\mu(x)$ on $x \in \boldsymbol{X}_\infty$ (not $\boldsymbol{X}$), we have $\bar{y}^n(x) \to \mu(x)$ as $n \to \infty$. Therefore there exists a time N_2 such that for $n > N_2$ and all $x < x' \in \boldsymbol{X}_\infty$, we have:

$$|\bar{y}^n(x) - \mu(x)| < \frac{1}{2}|\mu(x) - \mu(x')| = \frac{1}{2}[\mu(x) - \mu(x')] \tag{B.3.1}$$

and

$$|\bar{y}^n(x') - \mu(x')| < \frac{1}{2}|\mu(x) - \mu(x')| = \frac{1}{2}[\mu(x) - \mu(x')]. \tag{B.3.2}$$

Based on (B.3.1) and (B.3.2), we have:

$$\begin{aligned}
\bar{y}^n(x) &> \mu(x) - \frac{1}{2}[\mu(x) - \mu(x')] \\
&= \frac{1}{2}[\mu(x) + \mu(x')] \\
&= \mu(x') + \frac{1}{2}[\mu(x) - \mu(x')] \\
&> \bar{y}^n(x').
\end{aligned} \tag{B.3.3}$$

That is after time N_2 the sample mean will also satisfy strict monotonicity.

By the definition of $\boldsymbol{X}_\infty$, for any $x \in \boldsymbol{X}_\infty \subset \boldsymbol{X}$ we have $w^n(x) \to \infty$ as $n \to \infty$. Thus there exists a time $N_3 > \max(N_1, N_2)$ such that for $n > N_3$ and any $f(x) \in \mathscr{F}$ we have: 1) $S^n_{\boldsymbol{X}^c_\infty} = \tilde{S}_{\boldsymbol{X}^c_\infty}$, and 2) the inequality

$$w^n(x_{i,n}) > \frac{\tilde{S}_{\boldsymbol{X}^c_\infty}}{[\bar{y}^n(x_{i,n}) - f(x_{i,n})]^2} \tag{B.3.4}$$

where we assume that for each time index n there is at least one such $x_{i,n} \in \boldsymbol{X}_\infty$ satisfying $\bar{y}^n(x_{i,n}) \neq f(x_{i,n})$. We are able to make such an assumption because $\bar{y}^n(x_{i,n})$ is normally distributed with mean $\mu(x_{i,n})$ and variance $\sigma^2/w^n(x_{i,n})$. For a continuous random variable $\bar{y}^n(x_{i,n})$ we have $\mathbb{P}(\bar{y}^n(x_{i,n}) = f(x_{i,n})) = 0$. Thus the required $x_{i,n}$ exists almost surely. Additionally, (B.3.4) is valid not only because the denominator is nonzero almost surely, but also because for any n, both $\bar{y}^n(\cdot)$ and $f(\cdot)$ are finite on $\boldsymbol{X}$. Meanwhile,

$$\begin{aligned} S^n_{\boldsymbol{X}_\infty} &= \sum_{x \in \boldsymbol{X}_\infty} [\bar{y}^n(x) - f(x)]^2 \, w^n(x) \\ &\geq [\bar{y}^n(x_{i,n}) - f(x_{i,n})]^2 \, w^n(x_{i,n}) \qquad \text{(B.3.5)} \\ &\geq \tilde{S}_{\boldsymbol{X}^c_\infty}. \qquad \text{(B.3.6)} \end{aligned}$$

Notice that (B.3.6) is true for all $n > N_3$ and any $f(x) \in \mathscr{F}$, with the only assumption being the existence of such $x_{i,n}$. Furthermore, still assuming the existence of such $x_{i,n}$ for each n, we have

$$\lim_{n \to \infty} S^n_{\boldsymbol{X}_\infty} = \infty. \tag{B.3.7}$$

This equation (B.3.7) is valid since in (B.3.5), $[\bar{y}^n(x_{i,n}) - f(x_{i,n})]^2$ is nonzero almost surely and $w^n(x_{i,n}) \to \infty$ as $n \to \infty$.

To minimize the sum $\left(S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}^c_\infty}\right)$ among all $f(x) \in \mathscr{F}$, we need $f(x)$ such that the dominating and unbounded term $S^n_{\boldsymbol{X}_\infty}$ is minimized and hopefully diminishes to zero. As we have shown above, as long as there is at least one $x_{i,n} \in \boldsymbol{X}_\infty$ where $\bar{y}^n(x_{i,n})$ does not agree with $f(x_{i,n})$, $S^n_{\boldsymbol{X}_\infty}$ will dominate the sum. Recall that in (B.3.3) we have shown the strict monotonicity of $\bar{y}^n$, i.e., after time N_3 the function $\bar{y}^n(x)$ is strictly decreasing in x. Therefore, setting $\tau = N_3$ and for $n > \tau$ and $f(x) \in \mathscr{F}$, we have

$$S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}^c_\infty} \geq S^n_{\boldsymbol{X}^c_\infty} = \tilde{S}_{\boldsymbol{X}^c_\infty}. \tag{B.3.8}$$

The equality sign holds when $S^n_{\boldsymbol{X}_\infty} = 0$, which is equivalent to setting $f(x) = \bar{y}^n(x)$ for all $x \in \boldsymbol{X}_\infty$. This is valid again due to the strict monotonicity of $\bar{y}^n(x)$. Therefore, $\tilde{S}_{\boldsymbol{X}^c_\infty}$ is a lower bound on $\left(S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}^c_\infty}\right)$ and is achieved by setting $\hat{f}^*_n(x) = f(x) = \bar{y}^n(x)$ for all $x \in \boldsymbol{X}_\infty$ and $n > \tau$. Finally we need to minimize $\tilde{S}_{\boldsymbol{X}^c_\infty}$ which will be discussed in Theorem B.4.1.

(2) Assuming $\mu(x) \geq \mu(x'), \forall x, x' \in \boldsymbol{X}$ and $x > x'$. In this case we only need to discuss the case when $\mu(x_1) = ... = \mu(x_k)$ and $x_1 \geq ... \geq x_k$. Note that these x_i can be from either $\boldsymbol{X}_\infty$ or $\boldsymbol{X}^c_\infty$. To find the isotonic regression estimator $\hat{f}^*_n$ of $\bar{y}^n$ under such assumption on the true function,we propose a two-step method below:

Step 1: for all $x \in \boldsymbol{X}_\infty$, find

$$\tilde{f}_n(x) \quad = \quad \arg\min_{f(x)} \sum_{x \in \boldsymbol{X}_\infty} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x) \tag{B.3.9}$$

$$\text{subject to} \quad f(x) \in \mathscr{F}$$

That is, $\tilde{f}_n(x)$ is the isotonic regression estimator of the function $\bar{y}^n(x)$ only on the set $\boldsymbol{X}_\infty$.

Step 2: for all $x \in \boldsymbol{X}_\infty^c$, and $n \leq N_1$, find

$$\begin{aligned} \tilde{g}_n(x) &= \arg\min_{g(x)} \sum_{x \in \boldsymbol{X}_\infty^c} \left[\bar{y}^n(x) - g(x)\right]^2 w^n(x) \\ &\quad \text{subject to} \quad g(x) \in \mathscr{F} \text{ and } g(x) = \tilde{f}_n(x), \forall x \in \boldsymbol{X}_\infty \end{aligned}$$

and when $n > N_1$, find

$$\begin{aligned} \tilde{g}_n(x) &= \arg\min_{g(x)} \sum_{x \in \boldsymbol{X}_\infty^c} \left[\tilde{y}(x) - g(x)\right]^2 \tilde{w}(x) \\ &\quad \text{subject to} \quad g(x) \in \mathscr{F} \text{ and } g(x) = \tilde{f}_n(x), \forall x \in \boldsymbol{X}_\infty \end{aligned}$$

where $\tilde{w}(x)$ and $\tilde{y}(x)$, defined in Lemma B.2.1, are some constant numbers only depend on x.

Step 3: for all $x \in \boldsymbol{X}$, assign values of $\hat{f}_n^*(x)$ as

$$\hat{f}_n^*(x) = \begin{cases} \tilde{f}_n(x), & x \in \boldsymbol{X}_\infty \\ \tilde{g}_n(x), & x \in \boldsymbol{X}_\infty^c \end{cases} \tag{B.3.10}$$

In the step 2, we would have similar arguments like (B.3.4) and (B.3.6). There exists a time $N_4 > N_1$ such that for $n > N_4$ and any $f(x) \in \mathscr{F}$ we have: 1) $S^n_{\boldsymbol{X}_\infty^c} = \tilde{S}_{\boldsymbol{X}_\infty^c}$,

and 2) the inequality

$$w^n(x_{j,n}) > \frac{\tilde{S}_{\boldsymbol{X}_\infty^c}}{\left[\bar{y}^n(x_{j,n}) - f(x_{j,n})\right]^2} \tag{B.3.11}$$

where we assume that for each n there is at least one such $x_{j,n} \in \boldsymbol{X}_\infty$ with $\bar{y}^n(x_{j,n}) \neq f(x_{j,n})$. Notice again that (B.3.11) is valid since $\bar{y}^n(x_{j,n}) \neq f(x_{j,n})$ almost surely and for any n, both $\bar{y}^n(\cdot)$ and $f(\cdot)$ are finite on $\boldsymbol{X}$. Meanwhile,

$$\begin{aligned} S^n_{\boldsymbol{X}_\infty} &= \sum_{x \in \boldsymbol{X}_\infty} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x) \\ &\geq \left[\bar{y}^n(x_{j,n}) - f(x_{j,n})\right]^2 w^n(x_{j,n}) \\ &\geq \tilde{S}_{\boldsymbol{X}_\infty^c}. \end{aligned} \tag{B.3.12}$$

We still have $\lim_{n\to\infty} S^n_{\boldsymbol{X}_\infty} = \infty$, but $\tilde{S}_{\boldsymbol{X}_\infty^c}$ may not be a lower bound on $\left(S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}_\infty^c}\right)$ since $S^n_{\boldsymbol{X}_\infty}$ may not reach zero. This is because in step 1 we are no longer able to set $f(x) = \bar{y}^n(x)$. Thus it is not guaranteed that $\bar{y}^n(x) - \tilde{f}_n(x) = 0$. Nevertheless, when $n > N_4$, we have $\left(S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}_\infty^c}\right) = \left(S^n_{\boldsymbol{X}_\infty} + \tilde{S}_{\boldsymbol{X}_\infty^c}\right)$ and $S^n_{\boldsymbol{X}_\infty}$ is dominating $\tilde{S}_{\boldsymbol{X}_\infty^c}$. Moreover, setting $f(x) = \tilde{f}_n(x)$ would still give us a lower bound on $\left(S^n_{\boldsymbol{X}_\infty} + \tilde{S}_{\boldsymbol{X}_\infty^c}\right)$. Notice here $\tilde{S}_{\boldsymbol{X}_\infty^c}$ is only the limit of $S^n_{\boldsymbol{X}_\infty^c}$, not the lower bound.

Let $\tau = \max(N_3, N_4)$, we set, for $n > \tau$,

$$\begin{aligned} S^{n,min}_{\boldsymbol{X}_\infty} &= \min_{f(x)\in\mathscr{F}} \sum_{x \in \boldsymbol{X}_\infty} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x) \\ &= \sum_{x \in \boldsymbol{X}_\infty} \left[\bar{y}^n(x) - \tilde{f}_n(x)\right]^2 w^n(x). \end{aligned}$$

Then what we have shown above is that our original objective function $\left(S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}^c_\infty}\right)$ satisfies

$$S^n_{\boldsymbol{X}_\infty} + S^n_{\boldsymbol{X}^c_\infty} = S^n_{\boldsymbol{X}_\infty} + \tilde{S}_{\boldsymbol{X}^c_\infty} \geq S^{n,min}_{\boldsymbol{X}_\infty} + \tilde{S}_{\boldsymbol{X}^c_\infty}$$

for all $n > \tau$. Respectively considering the two functions in (B.3.10), $S^{n,min}_{\boldsymbol{X}_\infty}$ is the minimum value of $S^n_{\boldsymbol{X}_\infty}$ by applying $\tilde{f}_n(x)$ and $\tilde{S}_{\boldsymbol{X}^c_\infty}$ is minimized by applying $\tilde{g}_n(x)$. Furthermore, $\hat{f}^*_n(x)$ in (B.3.10) is a feasible solution to the original minimization problem whose objective value is equal to its global lower bound. Hence this $\hat{f}^*_n(x)$ is the optimal solution to the isotonic regression problem. □

B.4 Main Result

We are now ready to prove the main structural result.

Theorem B.4.1 (Limit of Isotonic Regression Estimator Under Partially Finite Measurements)**.**

For $x \in \boldsymbol{X} = \boldsymbol{X}_\infty \bigcup \boldsymbol{X}^c_\infty$, the isotonic regression estimator $\hat{f}^*_n(x)$ of the sample mean $\bar{y}^n(x)$ satisfies:

$$\lim_{n\to\infty} \hat{f}^*_n(x) = \begin{cases} \mu(x), & \text{if } x \in \boldsymbol{X}_\infty, \\ exists, & \text{if } x \in \boldsymbol{X}^c_\infty. \end{cases}$$

where $\mu(x)$ is the curve of true function on $\boldsymbol{X} = \boldsymbol{X}_\infty \bigcup \boldsymbol{X}^c_\infty$.

Proof. (1) Assuming strict monotonicity of the true function μ. For $x \in \boldsymbol{X}_\infty$, since $\bar{y}^n(x)$ is a strongly consistent estimator of $\mu(x)$ on $x \in \boldsymbol{X}$, we have $\bar{y}^n(x) \xrightarrow{a.s.} \mu(x)$ as

$n \to \infty$. Thus by Lemma B.3.1, as $n \to \infty$, $\lim_{n\to\infty} \hat{f}_n^*(x) = \lim_{n\to\infty} \bar{y}^n(x) = \mu(x), x \in \boldsymbol{X}_\infty$. Meanwhile for each $x \in \boldsymbol{X}_\infty^c$ and the τ in Lemma B.3.1, we have for $n > \tau$:

$$\begin{aligned}\hat{f}_n^*(x) &= \arg \min_{f(x)\in\mathscr{F} \text{ and } f(x)=\bar{y}^n(x), x\in\boldsymbol{X}_\infty} \left\{S^n_{\boldsymbol{X}_\infty^c}\right\} \\ &= \arg \min_{f(x)\in\mathscr{F} \text{ and } f(x)=\bar{y}^n(x), x\in\boldsymbol{X}_\infty} \left\{ \sum_{x\in\boldsymbol{X}_\infty^c} \left[\bar{y}^n(x) - f(x)\right]^2 w^n(x) \right\},\end{aligned}$$

each of which is finite. Hence, $\lim_{n\to\infty} \hat{f}_n^*(x) = \mu(x)$ for $x \in \boldsymbol{X}_\infty$ and $\lim_{n\to\infty} \hat{f}_n^*(x)$ exists for $x \in \boldsymbol{X}_\infty^c$.

(2) Assuming non-strict monotonicity of the true cost function. In this case, for $x \in \boldsymbol{X}_\infty$, since $\tilde{f}_n(x)$ in (B.3.10), as defined in equation (B.3.9), is the isotonic regression estimator of the sample means $\bar{y}^n$ (where itself is strongly consistent), thus by Theorem 1.2.5 and Theorem 1.2.6, $\tilde{f}_n(x)$ is a strongly consistent estimator of $\mu(x)$ on $x \in \boldsymbol{X}_\infty$. That is: $\tilde{f}_n(x) \xrightarrow{a.s.} \mu(x)$ as $n \to \infty$. Again by (B.3.10), as $n \to \infty$,

$$\lim_{n\to\infty} \hat{f}_n^*(x) = \lim_{n\to\infty} \tilde{f}_n(x) = \mu(x), x \in \boldsymbol{X}_\infty.$$

Similarly for each $x \in \boldsymbol{X}_\infty^c$ and the τ specified in Lemma B.3.1, based on (B.3.10) we have for $n > \tau$

$$\hat{f}_n^*(x) = \tilde{g}_n(x)$$

each of which is finite. Hence, $\lim_{n\to\infty} \hat{f}_n^*(x) = \mu(x)$ for $x \in \boldsymbol{X}_\infty$ and $\lim_{n\to\infty} \hat{f}_n^*(x)$ exists for $x \in \boldsymbol{X}_\infty^c$. □

B.5 Simulation

To visualize the main results in Theorem B.4.1, we present the following figures under different simulation setups. Let x (number of trucks) be elements in $\boldsymbol{X}$, $\mathbf{n}$ be number of measurements on each choice of x, μ be the true (cost) function value at each x, and σ be the standard variance; we assume a uniform σ value throughout our simulation study. We use $\mathbf{n} = 10,000$ to simulate the case of an x that is measured infinitely often, and $\mathbf{n} = 20$ for finitely often.

Setup 1: Set $\boldsymbol{X} = \{1, 2, ..., 10\}$ and measure the 1st and 10th x infinitely often, the rest finitely often. The true curve is a decreasing straight line with values at each x specified in Table B.1. The results are in Figure B.5.1.

Setup 2: Set $\boldsymbol{X} = \{1, 2, ..., 11\}$ and measure the 1st, 6th and 11th x infinitely often, the rest finitely often. The true curve is decreasing and convex first then concave, with values at each x specified in Table B.2.

The results are in Figure B.5.2.

Setup 3: Set $\boldsymbol{X} = \{1, 2, ..., 11\}$ and measure the 1st, 6th and 11th x infinitely often, the rest finitely often. The true curve is decreasing and concave first then convex, with values at each x specified in Table B.3.

The results are in Figure B.5.3.

We can see that the isotonic regression estimator is very close to the true curve at points we measure infinitely often (i.e. $\mathbf{n} = 10000$).

x	1	2	3	4	5	6	7	8	9	10
n	10000	20	20	20	20	20	20	20	20	10000
μ	45	40	35	30	25	20	15	10	5	1
σ	Uniformly = 40									

Table B.1: Simulation Setup 1

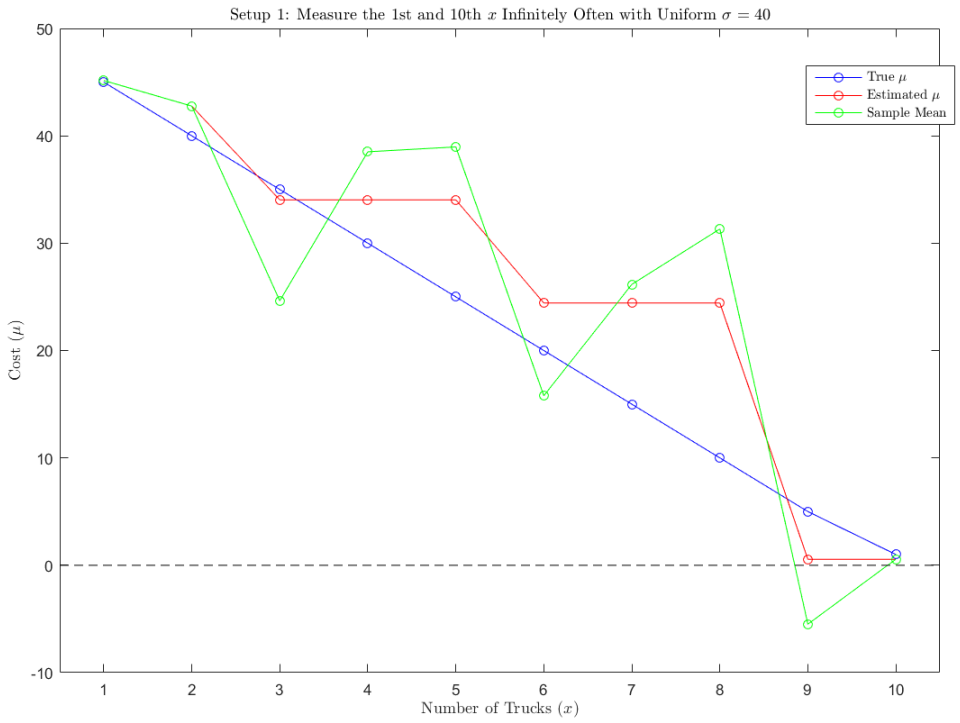

Figure B.5.1: Simulation Setup 1

x	1	2	3	4	5	6	7	8	9	10	11
n	10000	20	20	20	20	10000	20	20	20	20	10000
μ	50	40	33	28	26	25	24	22	17	10	1
σ	Uniformly = 40										

Table B.2: Simulation Setup 2

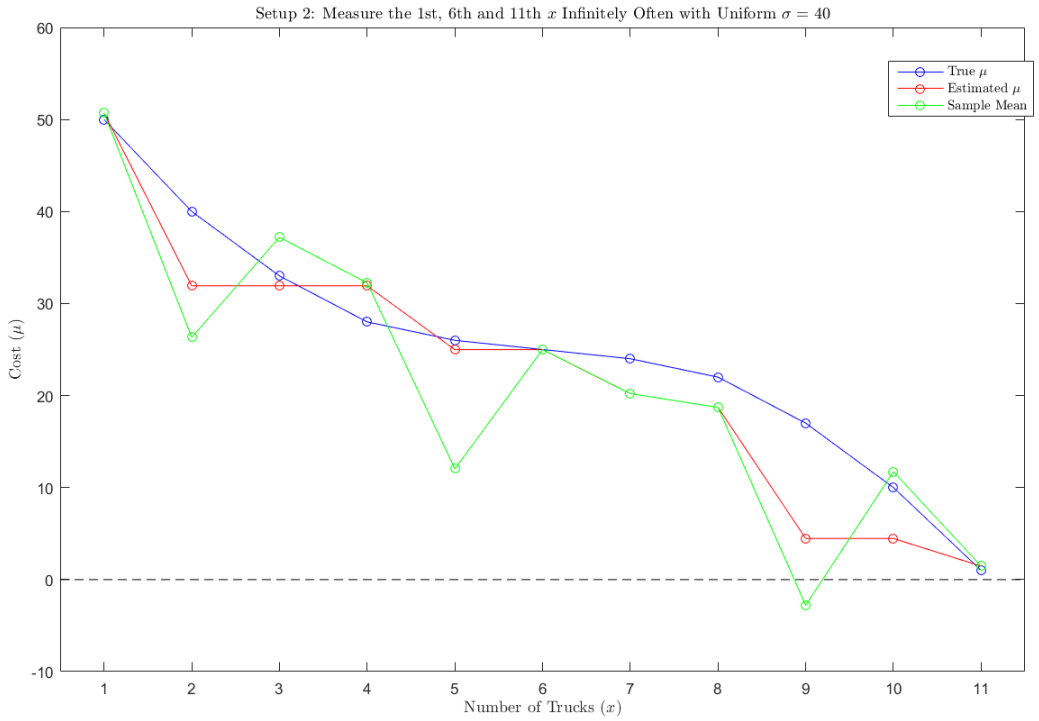

Figure B.5.2: Simulation Setup 2

x	1	2	3	4	5	6	7	8	9	10	11
n	10000	20	20	20	20	10000	20	20	20	20	10000
μ	50	49	47	42	35	25	15	8	3	1	0
σ	Uniformly = 40										

Table B.3: Simulation Setup 3

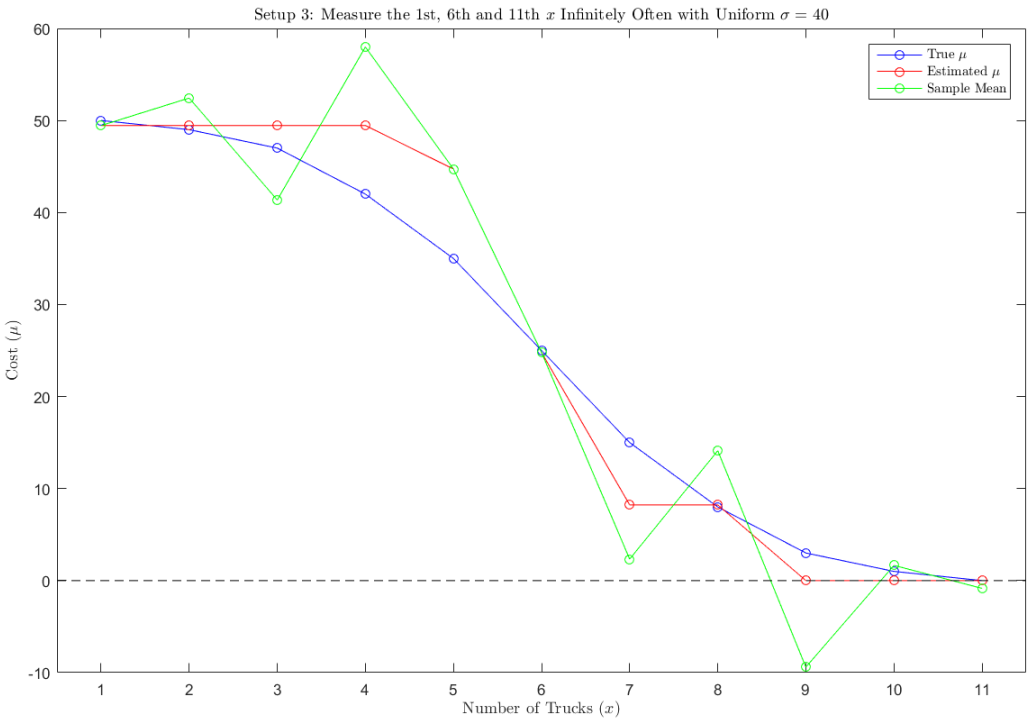

Figure B.5.3: Simulation Setup 3

Bibliography

Acton, S. T. and A. C. Bovik (1998). Nonlinear image estimation using piecewise and local image models. *IEEE Transactions on Image Processing 7*(7), 979–991.

Aït-Sahalia, Y. and J. Duarte (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics 116*(1-2), 9–47.

Anevski, D. and O. Hössjer (2006). A general asymptotic scheme for inference under order restrictions. *The Annals of Statistics 34*(4), 1874–1930.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics 11*(3), 375–386.

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 641–647.

Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association 84*(405), 289–294.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley: New York.

Barlow, R. E. and E. M. Scheuer (1966). Reliabilitv growth during a development testing progiam. *Technometrics 8*(1), 53–60.

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika 46*(1/2), 36–48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives. ii. *Biometrika 46*(3/4), 328–335.

Bartholomew, D. J. (1961). A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society. Series B (Methodological) 23*(2), 239–281.

Bhattacharyya, G. and J. H. Klotz (1966). The bivariate trend of Lake Mendota. Technical report, University of Wisconsin Madison, Department of Statistics.

Boswell, M. T. (1966). Estimating and testing trend in a stochastic process of poisson type. *The Annals of Mathematical Statistics*, 1564–1573.

Bouzeghoub, M., S. Ellacott, A. Easdown, and M. Brown (2000). On the identification of non-stationary linear processes. *International Journal of Systems Science 31*(3), 273–286.

Brezger, A. and W. J. Steiner (2008). Monotonic regression based on bayesian p–splines: An application to estimating price response functions from store-level scanner data. *Journal of Business & Economic Statistics 26*(1), 90–104.

Broffitt, J. D. (1984). A bayes estimator for ordered parameters and isotonic bayesian graduation. *Scandinavian Actuarial Journal 1984*(4), 231–247.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 607–616.

Brunk, H. D. (1958). On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, 437–454.

Brunk, H. D. (1970). Estimation of isotonic regression. *Nonparametric Techniques in Statistical Inference*.

Brunk, H. D., G. Ewing, and W. Utz (1957). Minimizing integrals in certain classes of monotone functions. *Pacific Journal of Mathematics 7*(1), 833–847.

Cai, B. and D. B. Dunson (2007). Bayesian multivariate isotonic regression splines: Applications to carcinogenicity studies. *Journal of the American Statistical Association 102*(480), 1158–1171.

Carlin, B. P. and T. A. Louis (2008). *Bayesian Methods for Data Analysis*. CRC Press.

Caruana, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 161–168. ACM.

Chakravarti, N. (1993). Sensitivity analysis in isotonic regression. *Discrete Applied Mathematics 45*(3), 183–196.

Chen, S., K.-R. G. Reyes, M. K. Gupta, M. C. McAlpine, and W. B. Powell (2015). Optimal learning in experimental design using the knowledge gradient policy with application to characterizing nanoemulsion stability. *SIAM/ASA Journal on Uncertainty Quantification 3*(1), 320–345.

Cheng, G. (2009). Semiparametric additive isotonic regression. *Journal of Statistical Planning and Inference 139*(6), 1980–1991.

Choi, D. and B. Van Roy (2006). A generalized Kalman filter for fixed point approximation and efficient temporal-difference learning. *Discrete Event Dynamic Systems 16*(2), 207–239.

Çınlar, E. (2011). *Probability and Stochastics*. Springer-Verlag New York.

Darken, C. and J. E. Moody (1991). Note on learning rate schedules for stochastic optimization. In *Advances in Neural Information Processing Systems*, pp. 832–838.

Dawson, D. and L. Magee (2001). The national hockey league entry draft, 1969–1995: An application of a weighted pool-adjacent-violators algorithm. *The American Statistician 55*(3), 194–199.

Douglas, S. C. and A. Cichocki (1998). Adaptive step size techniques for decorrelation and blind source separation. In *Conference Record of the 32nd Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA*, Volume 2, pp. 1191–1195. IEEE.

Douglas, S. C. and V. J. Mathews (1995). Stochastic gradient adaptive step size algorithms for adaptive filtering. In *Proc. International Conference on Digital Signal Processing, Limassol, Cyprus*, Volume 1, pp. 142–147.

Du, S. S. and S. Goel (2018). Improved learning of one-hidden-layer convolutional neural networks with overlaps. *arXiv preprint arXiv:1805.07798*.

Dunson, D. B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association 100*(470), 618–627.

Dunson, D. B. and B. Neelon (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics 59*(2), 286–295.

Durot, C. (2002). Sharp asymptotics for isotonic regression. *Probability Theory and Related Fields 122*(2), 222–240.

Durot, C. and L. Reboul (2010). Goodness-of-fit test for monotone functions. *Scandinavian Journal of Statistics 37*(3), 422–441.

Durot, C. and A.-S. Tocquet (2001). Goodness of fit test for isotonic regression. *ESAIM: Probability and Statistics 5*, 119–140.

Dykstra, R. L. and P. Laud (1981). A bayesian nonparametric approach to reliability. *The Annals of Statistics*, 356–367.

Frazier, P. I., W. B. Powell, and S. Dayanik (2008). A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization 47*(5), 2410–2439.

Frazier, P. I., W. B. Powell, and S. Dayanik (2009). The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing 21*(4), 599–613.

Fu, M. C. (2002). Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing 14*(3), 192–215.

Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting 4*(1), 1–28.

George, A. P. and W. B. Powell (2006). Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming. *Machine Learning 65*(1), 167–198.

Giegerich, R. (2000). A systematic approach to dynamic programming in bioinformatics. *Bioinformatics 16*(8), 665–677.

Gill, J. (2014). *Bayesian Methods: A Social and Behavioral Sciences Approach*. CRC Press.

Gjuvsland, A. B., Y. Wang, E. Plahte, and S. W. Omholt (2013). Monotonicity is a key feature of genotype-phenotype maps. *Frontiers in Genetics 4*, 216.

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Volume 2.

Groeneboom, P. (1989). Brownian motion with a parabolic drift and airy functions. *Probability Theory and Related Fields 81*(1), 79–109.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001a). A canonical process for estimation of convex functions: the "invelope" of integrated Brownian motion $+t^4$. *The Annals of Statistics 29*(6), 1620–1652.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001b). Estimation of a convex function: characterizations and asymptotic theory. *The Annals of Statistics 29*(6), 1653–1698.

Hansohm, J. and X. Hu (2012). A convergent algorithm for a generalized multivariate isotonic regression problem. *Statistical Papers 53*(1), 107–115.

He, X. (2017, August). *Optimal Learning for Nonlinear Parametric Belief Models*. Ph. D. thesis, Princeton University.

He, X. and W. B. Powell (2016). Optimal learning for stochastic optimization with nonlinear parametric belief models. *arXiv preprint arXiv:1611.07161*.

Holmes, C. C. and N. A. Heard (2003). Generalized monotonic regression using random change points. *Statistics in Medicine 22*(4), 623–638.

Huang, Y., L. Zhao, W. B. Powell, Y. Tong, and I. O. Ryzhov (2018). Optimal learning for urban delivery fleet allocation. *Transportation Science*.

Hussian, M., A. Grimvall, O. Burdakov, and O. Sysoev (2005). Monotonic regression for the detection of temporal trends in environmental quality data. *MATCH Commun. Math. Comput. Chem 54*, 535–550.

Hyndman, R., A. B. Koehler, J. K. Ord, and R. D. Snyder (2008). *Forecasting*

with Exponential Smoothing: The State Space Approach. Springer-Verlag Berlin Heidelberg.

Kakade, S. M., V. Kanade, O. Shamir, and A. Kalai (2011). Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pp. 927–935.

Kalish, M. L., J. C. Dunn, O. P. Burdakov, and O. Sysoev (2016). A statistical test of the equality of latent orders. *Journal of Mathematical Psychology 70*, 1–11.

Kallenberg, O. (2006). *Foundations of Modern Probability*. Springer-Verlag New York.

Kesten, H. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics 29*(1), 41–59.

Khodadadi, A. and M. Asgharian (2008). Change-point problem and regression: an annotated bibliography. *COBRA Preprint Series*, 44.

Kiefer, J. and J. Wolfowitz (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics 23*(3), 462–466.

Koralov, L. and Y. G. Sinai (2007). *Theory of Probability and Random Processes*. Springer-Verlag Berlin Heidelberg.

Kotłowski, W., W. M. Koolen, and A. Malek (2016). Online isotonic regression. In *Conference on Learning Theory*, pp. 1165–1189.

Kraft, C. H. and C. van Eeden (1964). Bayesian bio-assay. *The Annals of Mathematical Statistics 35*(2), 886–890.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*(1), 1–27.

Kushner, H. and G. G. Yin (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer New York.

Kushner, H. J. and J. Yang (1995). Analysis of adaptive step-size SA algorithms for parameter tracking. *IEEE Transactions on Automatic Control 40*(8), 1403–1410.

Lavine, M. and A. Mockus (1995). A nonparametric bayes method for isotonic regression. *Journal of Statistical Planning and Inference 46*(2), 235–248.

Lee, C.-I. C. (1996). On estimation for monotone dose¡ªresponse curves. *Journal of the American Statistical Association 91*(435), 1110–1119.

Lee, C.-I. C. et al. (1983). The min-max algorithm and isotonic regression. *The Annals of Statistics 11*(2), 467–477.

Luss, R. and S. Rosset (2014). Generalized isotonic regression. *Journal of Computational and Graphical Statistics 23*(1), 192–210.

Luss, R., S. Rosset, and M. Shahar (2010). Decomposing isotonic regression for efficiently solving large problems. In *Advances in Neural Information Processing Systems*, pp. 1513–1521.

Marshall, A. W. (1970). Discussion on barlow and van zwet's paper. *Nonparametric Techniques in Statistical Inference 1969*, 174–176.

Maxwell, W. L. and J. A. Muckstadt (1985). Establishing consistent and realistic reorder intervals in production-distribution systems. *Operations Research 33*(6), 1316–1341.

Mikhael, W., F. Wu, L. Kazovsky, G. Kang, and L. Fransen (1986). Adaptive filters with individual adaptation of parameters. *IEEE Transactions on Circuits and Systems 33*(7), 677–686.

Mukerjee, H. (1988). Monotone nonparametric regression. *The Annals of Statistics*, 741–750.

Neelon, B. and D. B. Dunson (2004). Bayesian isotonic regression and trend analysis. *Biometrics 60*(2), 398–406.

Pardalos, P. M. and G. Xue (1999). Algorithms for a class of isotonic regression problems. *Algorithmica 23*(3), 211–222.

Pflug, G. C. (1988). Stepsize rules, stopping times and their implementation in stochastic quasigradient algorithms. *Numerical Techniques For Stochastic Optimization*, 353–372.

Powell, W. B. (2007). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons.

Robbins, H. and S. Monro (1951, 09). A stochastic approximation method. *The Annals of Mathematical Statistics 22*(3), 400–407.

Robertson, T. and F. Wright (1975). Consistency in generalized isotonic regression. *The Annals of Statistics*, 350–362.

Robertson, T., F. T. Wright, and R. Dykstra (1988). *Order Restricted Statistical Inference*. Wiley: New York.

Roth, M., T. Buishand, and G. Jongbloed (2015). Trends in moderate rainfall extremes: A regional monotone regression approach. *Journal of Climate 28*(22), 8760–8769.

Ryzhov, I. O. (2018). Convergence analysis of the bias-adjusted Kalman filter. Technical report, University of Maryland at College Park.

Ryzhov, I. O., P. I. Frazier, and W. B. Powell (2015). A new optimal stepsize for approximate dynamic programming. *IEEE Transactions on Automatic Control 60*(3), 743–758.

Ryzhov, I. O., W. B. Powell, and P. I. Frazier (2012). The knowledge gradient algorithm for a general class of online learning problems. *Operations Research 60*(1), 180–195.

Saridis, G. N. (1970). Learning applied to successive approximation algorithms. *IEEE Transactions on Systems Science and Cybernetics 6*(2), 97–103.

Schraudolph, N. N. (1999). Local gain adaptation in stochastic gradient descent.

In *The Ninth International Conference on Artificial Neural Networks*, Volume 2, Edinburgh, London., pp. 569–574.

Sedransk, J., J. Monahan, and H. Chiu (1985). Bayesian estimation of finite population parameters in categorical data models incorporating order restrictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 519–527.

Silvapulle, M. J. and P. K. Sen (2011). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. John Wiley & Sons.

Simão, H. P., J. Day, A. P. George, T. Gifford, J. Nienow, and W. B. Powell (2009). An approximate dynamic programming algorithm for large-scale fleet management: A case application. *Transportation Science 43*(2), 178–197.

Simão, H. P., A. George, W. B. Powell, T. Gifford, J. Nienow, and J. Day (2010). Approximate dynamic programming captures fleet operations for schneider national. *Interfaces 40*(5), 342–352.

Stengel, R. F. (1994). *Optimal Control and Estimation*. Dover Publications, New York.

Stout, Q. F. (2013). Isotonic regression via partitioning. *Algorithmica 66*(1), 93–112.

Stout, Q. F. (2015). Isotonic regression for multiple independent variables. *Algorithmica 71*(2), 450–470.

Sysoev, O. and O. Burdakov (2018). A smoothed monotonic regression via L_2 regularization. *Knowledge and Information Systems*, 1–22.

Thompson, W. (1962). The problem of negative estimates of variance components. *The Annals of Mathematical Statistics*, 273–289.

Tibshirani, R. J., H. Hoefling, and R. Tibshirani (2011). Nearly-isotonic regression. *Technometrics 53*(1), 54–61.

Van, C. and R.-A. Dana (2003). *Dynamic Programming in Economics*. Springer US.

van Eeden, C. (1956). Maximum likelihood estimation of ordered probabilities. In *Indagationes Mathematicae (Proceedings)*, Volume 59 of *A*, pp. 444–455. Stichting Mathematisch Centrum.

van Eeden, C. (1957a). Maximum likelihood estimation of partially or completely ordered parameters. i. In *Indagationes Mathematicae (Proceedings)*, Volume 60 of *A*, pp. 128–136. Elsevier.

van Eeden, C. (1957b). Maximum likelihood estimation of partially or completely ordered parameters. ii. In *Indagationes Mathematicae (Proceedings)*, Volume 60 of *A*, pp. 201–211. Elsevier.

van Eeden, C. (1958). *Testing and estimating ordered parameters of probability distribution*. Ph. D. thesis, University of Amsterdam.

Wang, L. and D. B. Dunson (2011). Bayesian isotonic density regression. *Biometrika 98*(3), 537–551.

Wu, J., M. C. Meyer, and J. D. Opsomer (2015). Penalized isotonic regression. *Journal of Statistical Planning and Inference 161*, 12–24.