Essays on empirical likelihood in economics

Gao, Z.

Publication date
2012
Document Version
Final published version

This thesis intends to exploit the roots of Empirical Likelihood and its related methods in mathematical programming and computation. The roots will be connected and the connections will induce new solutions for the problems of estimation, computation, and generalization of Empirical Likelihood.

Zhengyuan Gao BSc in Computational Mathematics (Xi'an Jiaotong University), Msc in Economics and Econometrics (University of Southampton). Major research interests are econometrics and mathematical economics.

Thela Thesis
Academic Publishing Services

533
Essays on Empirical Likelihood in Economics Zhengyuan Gao

tinbergen *institute*

Essays on Empirical Likelihood in Economics

Zhengyuan Gao

Research Series
Universiteit van Amsterdam

ESSAYS ON EMPIRICAL LIKELIHOOD IN ECONOMICS

ISBN

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **533** of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Essays on Empirical Likelihood in Economics

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 29 mei 2012, te 10:00 uur

door

Zhengyuan Gao

geboren Guangzhou, China

Promotiecommissie:

Promotor:	Prof. dr. H. P. Boswijk
Co-Promotor:	Dr. K.J. van Garderen
Overige leden:	Prof. dr. Y. Kitamura
	Prof. dr. R.J. Smith
	Prof. dr. J. F. Kiviet
	Prof. dr. J. G. de Gooijer

Faculteit Economie en Bedrijfskunde

CONTENTS

1

INTRODUCTION

1.1 EL, MATHEMATICAL PROGRAMMING, AND COMPUTATION

This thesis intends to exploit the roots of Empirical likelihood (EL) and EL's related methods in mathematical programming and computation. The roots will be connected and the connections will induce new solutions for the problems of estimation, computation, and generalization of EL.

In economics, the study of resource allocation under scarcity often refers to optimization or mathematical programming. Identifying optima was firstly proposed by Fermat and Lagrange in their calculus-based formulas, while the way of solving this optimal problem was initiated by Newton and Gauss using their iterative computational methods. The modern optimization which utilizes the dual theory for attaining optima was developed by Kantorovich and was introduced to economics by von Neumann to solve the primal production maximization problem or the dual, cost minimization problem. The existence of the dual of a primal problem requires some regularity conditions which would construct a feasible solution set. In most cases, simplifications of these conditions appear as constraints in optimization problems. Such problems are known as constrained optimization problems. Constrained optimization appears to be *a crucial link* of connecting EL and its related methods to economic and econometric problems. The difference between EL (and its related methods) and many other extremum estimation methods is that EL uses constraints in its estimation.

Prior to EL, in econometrics and statistics, the essential likelihood based topic that relates to mathematical programming and computation is Maximum Likelihood Estimation (MLE). The connection between mathematical programming, MLE and computational algorithms can be traced back to Wald (1943) and Kiefer and Wolfowitz (1952). Recently Owen (1988, 2001) introduced a nonparametric likelihood-based method, Empirical Likelihood, which pushes the implementations of estimation

methodology from the classical optimization to the constrained optimization. While EL itself is a statistical problem, applications of EL to economics require the integration of the advanced techniques of mathematical programming and related computational methods.

EL has attracted a lot of attention in econometrics since Qin and Lawless (1994) incorporated estimating equations into EL. The estimating equations in this modified EL play the same role as moment conditions in the Generalized Methods of Moments (GMM) which is currently one of the most popular estimation approaches in econometrics. Moreover, the estimators in both EL (Qin and Lawless, 1994) and GMM share many similar statistical features. Therefore, EL estimation in the framework of Qin and Lawless (1994) has been recognized as a moment-based estimation method in econometrics. For a sample of n observations, the moment-based EL is as follows:

$$\max_{\theta,p_1,\ldots,p_n} \left\{ \prod_{i=1}^{n} np_i \,\middle|\, \sum_{i=1}^{n} p_i m(X_i,\theta) = 0,\ p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1 \right\}.$$

The likelihood ratio function $\prod_{i=1}^{n} np_i$ is called, variously, an objective function, cost function, energy function, or energy functional in different applications of mathematical programming. The estimating function $m(X_i,\theta)$ is of main interest in all moment-based estimation methods.

A remarkable connection between the moment-based estimation method and the mathematical programming problem is established in Kitamura and Stutzer (1997) via the parameters in the dual problem. The use of dual parameters also appears in Owen (1990) and Qin and Lawless (1994) for Lagrangian type testing. Kitamura and Stutzer (1997) use the dual parameter to derive an alternative representation (or objective function) of moment-based estimation methods. Their method which they refer to as Exponential Tilting is based on optimizing the Kullback-Leibler divergence, instead of the log-likelihood ratio, subjected to moment constraints. The dual problem shows an alternative way of incorporating moment constraints. The moment constraints no longer appear directly in the objective functions as they do in GMM or other minimum distance methods. The moment constraints are controlled dually by the Lagrangian multiplier instead and then appear indirectly in the modified objective functions.

General speaking, "duality theory" in optimization means the simultaneous study of a *pair* of optimization problems, the

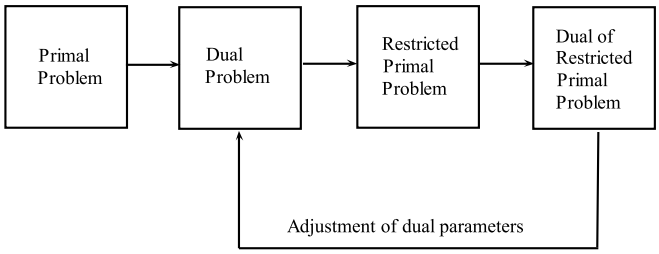


Figure 1: An outline of the primal-dual method.

initial problem which is also called the "primal problem" and the dual problem. The aim of dual problem is to obtain more information about the primal problem. For EL and its related methods, the information of constraints and the information of optimal "weights" w of these constraints are presented in a single objective function by the duality theory. Using the auxiliary dual parameters, Smith (1997) and Newey and Smith (2004) show that a class of estimators including Exponential Tilting, continuous updating GMM and EL, will have better statistical properties than the original GMM whose weights are not necessarily optimal.

Although they are important improvements from the statistical point of view, EL and its related moment-based methods still face several practical difficulties when they are applied in economic models. One of these issues addressed in Chapter 2 and 3 is about non-linear constraints. Non-linearity represents a multitude of complex economic phenomena, and in turn, applications in structural econometrics ubiquitously give rise to problems formulated as nonlinear constrained optimization problems. The duals of these non-linear constrained optimization problems are generally large, complex and infinite dimensional. As a result, this class of optimization problems present significant computational challenges.

As we know, nonlinear equations in general have no closed form solutions but only numerical solutions. Optimization can be thought of as a way of finding approximating solutions to some equations. If the equations are nonlinear, optimization problems also become nonlinear. The computational mechanism, however, is developed from a linear or quasi-linear environment. Thus when one attempts to solve a nonlinear optimization problem, one should first think about transferring the problem to a linear or quasi-linear problem. Then firstly one needs to make a choice for linearizing the problem:

(a): should I linearize the nonlinear problem first and solve a linearized optimization problem?

(b): should I optimize the nonlinear problem first and obtain a set of solution equations to be linearized?

The approach (a) often relates to approximation and the second approach often relates to the optimization given the first order condition. In general, these two steps do not commute.

Chapter 2 suggests a localization method to solve the nonlinear optimization problem of EL, which belongs to category (a). This idea originally appeared in Newton's algorithmic scheme. Le Cam (1974) used Newton's scheme in statistics to solve the irregular likelihood (objective) function caused by the highly degree of nonlinearity. Unlike classic MLE, the moment-based EL estimator depends on its dual optimal value. Le Cam's localization technique might not be proper in the dual problem. As we will see, if the likelihood function in the primal problem of EL has (highly) nonlinear constraints, the dual problem will have a single objective function of a minimax problem with an infinite dimensional Lagrangian multiplier. I apply the primal-dual scheme together with Netwon-Le Cam's localization to solve this problem. From Figure 1 above, one can see the principle role of the dual problem: it helps to restrict the solution set of the primal problem. It implies that a tractable dual representation would be helpful for finding the solution. By this property and the primal likelihood problem, we will have a Kitamura-Stutzer type duality. We will apply this dual result to a restricted problem - a locally linearized (linear-quadratic) representation of the primal problem. Finally the dual of this restricted problem will give a signal how to adjust the multiplier and thereafter the value of the primal likelihood function. Like Newton's algorithm, this is an iterative scheme. Statistical properties of this scheme will depend only on the construction of the estimator rather than the number of iterations in the algorithm.

Chapter 3 considers a rather different problem: how to make an optimal decision when agents face various uncertainties in a dynamic economic system. We will see that this problem is again a mathematical programming problem. The solution of this problem seeks to optimize an objective defined over many points in time taking into account how the objective function changes if there is a small change in the choice path. Usually, economists apply dynamic programming to study this case and the optimization strategy is based on splitting the problem into smaller subproblems. The equation that describes one

of these subproblems is called a Bellman equation. We will see that this Bellman type problem can be dually formulated as a mathematical programming with equilibrium constraints where the constraints are Bellman type equations. Then we can easily introduce robustness concerns of uncertainty to the new dual problem even when it is hard to introduce it to the original problem. The solution of the dual problem will be approximately represented by a linear function.

The connection between the contents of these two chapters is the estimation method. The estimation in Chapter 3 relies on the approximating solution of the robust decision problem and it belongs to the category (b). The robust decision problem in Chapter 3 induces non-differentiable components when we construct the constraints for estimation. For non-differentiable components, there is a risk of obtaining inconsistent gradients of the objective functionals. One may approximate these gradients, but in our specific setting (with robustness concerns) the approximate gradient obtained is not a true gradient of anything: neither of the continuous functional nor the discrete functional. To avoid the effects of non-differentiability, we transfer this problem to the one in the category (a). In other words, while the objective functions in the category (b) are continuously differentiable before linearization, we might obtain non-differentiable components after linearizing such problems.

In Chapter 3, one can see that using a usual localization procedure we show that the estimation problem can be trivially discretized first and then the discretized problem will be represented by a continuous local objective function. In a number of subfields in control theory, localization techniques are designed specifically for optimization in dynamic contexts.

The last chapter considers a generalization of EL dealing with the primal-dual concern. Duality, roughly speaking, is a fundamental concept that underlies many aspects of extremum principles in natural systems. In EL and its related methods, the optimal probabilistic weights of the moment constraints are convexifying these constraint functions. This is quite natural in the light of the results of constrained optimization, since best approximation by convex sets is a particular case of convex optimization and for convex systems, the mathematical theory of duality is well established due to the existence of a common symmetric framework (Hahn-Banach Theorem). Therefore, a naive guess is that a primal-dual relation must exist for a general class of EL and its related methods. Chapter 4 is to verify this

conjecture and to give a specific class for dually representing GEL.

2

AN EMPIRICAL LIKELIHOOD BASED LOCAL ESTIMATION

2.1 INTRODUCTION

To estimate an economic model, the model is often represented in terms of a family of probability measures $\mathcal{E}_\theta = \{P_\theta; \theta \in \Theta\}$ depending on a parameter θ in $\Theta \in \mathbb{R}^d$. By adjusting the value of θ one can choose which P_θ best fits the data. There is a literature within econometrics that considers how to attain a suitable measure P_θ by comparing a specified *moment condition function*

$$\int m(x,\theta) dP_\theta(x) = \mathbb{E}_\theta[m(X,\theta)],$$

a $k \times 1$ vector with $k \geq d$, with its *sample counterpart*

$$\int m(x,\theta) dP_n(x) = \frac{1}{n}\sum_{i=1}^{n} m(X_i,\theta),$$

where P_n is the empirical distribution. Note that although P_θ is indexed by θ, the distribution of $m(X,\theta)$ does not necessarily fully depend on θ. The notation P_θ should be interpreted as a pseudo measure of $m(X,\theta)$ and the specification of this measure depends on the value of θ. In this chapter, we assume the random variable X_i to be i.i.d. A particular correspondence between $\mathcal{E}_\theta$ and $m(X,\theta)$ is established by Empirical Likelihood (EL) (Qin and Lawless, 1994; Kitamura and Stutzer, 1997). EL has been embedded into some more general problems, e.g. Csiszar (1984); Smith (1997); Baggerly (1998); Newey and Smith (2004). The aims of these methods are similar: to optimize a criterion function of θ, for example a likelihood ratio, subject to constraints based on $m(X,\theta)$.

The choice of criterion functions matters for the efficiency and robustness of an estimator. To balance the tradeoff between these two objectives, Schennach (2007) suggests a two-step inference method by switching the empirical discrepancy between

Kullback-Leibler and likelihood ratio. Kitamura et al. (2009) suggest using Hellinger's distance as the criterion. In this chapter, we will focus on a representation of the classical likelihood ratio. Classical likelihood ratio does have problems of maintaining robustness, but several ways of re-constructing the likelihood have been proposed that remedy this problem. This chapter will consider a "localization" technique of representing the likelihood ratio function of $m(X,\theta)$ when it has poor behavior over some critical points. The method was first suggested in statistics for parametric Maximum Likelihood Estimation by Le Cam even earlier than his 1974 published notes (Le Cam, 1974).

The "local" here is the analog of "differential". If one fixes a particular θ_0 in Θ and investigates what happens to the likelihood ratio function with parameter sequences of the form $\theta = \theta_0 + \delta_n\tau$, with $\delta_n \to 0$ as n goes to infinity, then δ_n yields a sort of differentiation rate just as the differentiation rate in basic calculus, and then the whole localization problem can be analyzed as a kind of differentiability problem. The term τ is called local parameter since it is an index for local features. This technique often appears in the evaluation of local power of test statistics and statistical experiments, see van der Vaart (1998) and Le Cam and Yang (2000).

The advantage of studying the EL problem under the localized representation is significant. For instance, the likelihood of EL includes a vector of implied probabilities $(p(X_1,\theta),\ldots,p(X_n,\theta))$ where $\theta \in \Theta \subset \mathbb{R}^d$. Localization considers the probability vector $(p(X_1,\theta),\ldots,p(X_n,\theta))$ on a neighborhood of some θ^* and returns numbers instead of functions. In addition, a well-behaved local representation ensures the existence of the derivative of this representation. By definition, when the derivative exists, small changes will not blow up the approximation of the original likelihood ratio function and this representation is therefore robust to these changes. Apart from theoretical advantages, the method is also computationally attractive. Unlike global approaches where complexity grows exponentially with the dimension of the parameter, the local method can handle the growing number of $(p(X_1,\theta),\ldots,p(X_n,\theta))$ by cutting a large problem into many small, but easily computable local problems. The local approach can avoid some peculiar points that break down the computational routines.

2.2 EMPIRICAL LIKELIHOOD

EL considers a finite dimensional parameter θ and an increasing number of $(p(X_1,\theta),\ldots,p(X_n,\theta))$. EL simultaneously finds the optimal θ and the optimal $(p(X_1,\theta),\ldots,p(X_n,\theta))$ that satisfy the required moment constraints $\sum_{i=1}^{n} m(X_i,\theta)p(X_i,\theta) = 0$. Its criterion is:

$$\sup_{p_i,\theta}\left\{\sum_{i=1}^{n}\log np_i \mid p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i m_i(\theta) = 0\right\},$$

where p_i is shorthand for $p(X_i,\theta)$ given the value θ. An explicit expression for the optimal p_i's can be derived using the Lagrangian method and gives the solution:

$$\tilde{p}_i(\theta) := \frac{1}{n}\frac{1}{1+\lambda_n^T m_i(\theta)},$$

where $\tilde{p}_i(\theta)$ is called the *implied probability*. The candidate solutions belong to the family

$$\mathcal{E}_\theta := \{\tilde{P}_\theta : \theta \in \Theta, \int m(X,\theta)d\tilde{P}_\theta = 0\},$$

where $d\tilde{P}_\theta(x_i) = \tilde{p}_i(\theta)d\mu$ for a counting measure μ[1]. The λ_n is the solution of:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{m_i(\theta)}{1+\lambda_n^T m_i(\theta)}\right] = 0. \tag{2.1}$$

Let the log-likelihood ratio of the implied probability between any two parameter values θ_1 and θ_2 be:

$$\Lambda_n(\theta_1,\theta_2) := \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{\tilde{p}_i(\theta_1)}{\tilde{p}_i(\theta_2)}\right]$$

and define the average log-likelihood ratio of the implied probability given θ and counting numbers $1/n$ as

$$\Lambda_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\log n\tilde{p}_i(\theta).$$

The constraint $0 \leq \tilde{p}_i \leq 1$ requires that the inequality $1+\lambda_n^T m_i(\theta) \geq 1/n$ always holds. The *population* $\lambda(\theta) := \lim_{n\to\infty}\lambda_n$ must lie in

1 The family $\mathcal{E}_\theta$ obtains both continuous measures and discrete measures. The definition will become clear once we introduce the infinite divisiblility concept.

a convex and closed set $\Gamma_\theta = \lim_{n\to\infty} \cup_{i=1}^{n} \Gamma_{\theta,i}$. For fixed n, the set (σ-algebra) $\Gamma_{\theta,n}$ is defined as a collection of subsets of

$$\{\lambda_n : 1 + \lambda_n^T m(X_i,\theta) \geq 1/n,\ i = 1,\ldots,n,\ \theta \in \Theta\}.$$

In the following, we will consider the case where the derivatives of $m(X,\theta)$ do not exist for some X. A weaker consistency result for the EL estimator is required. Here are the conditions:

Condition 2.1. (i) $M(\theta) := \mathbb{E}[m(X,\theta)]$ exists for all $\theta \in \Theta$ and has a unique zero at $\theta = \theta_0$.

(ii) θ_0 is a well-separated point in $M(\theta)$ such that

$$\inf_{\theta : d(\theta,\theta_0) \geq \epsilon} |M(\theta)| > |M(\theta_0)| = 0,$$

where ϵ is an arbitrary value larger than zero and $d(\cdot,\cdot)$ is any distance function on Θ.

(iii) $m(X,\theta)$ is continuous in θ,

$$\lim_{\theta' \to \theta} \left\| m(X,\theta) - m(X,\theta') \right\| = 0.$$

(iv) Let ∞ be the one-point compactification of Θ, then there exists a continuous function $b(\theta)$ bounded away from zero, such that

(1) $\sup_{\theta\in\Theta} \|m(X,\theta)\| / b(\theta)$ is integrable,
(2) $\liminf_{\theta\to\infty} \|M(\theta)\| / b(\theta)$ is larger than 1, and
(3) $\limsup_{\theta\to\infty} \|m(X,\theta) - M(\theta)\| / b(\theta) < 1$.

(v) $\sum_{i=1}^{n} \left[m(x_i,\theta_0) m(x_i,\theta_0)^T\right] / n$ exists and has full rank.

Condition 2.1 (i) ensures the model is identified for a small neighborhood of θ_0. (ii) is a local separability condition. (iii) is used to obtain the continuity of the Lagrangian multiplier. (iv) is an envelope assumption; it is used to obtain some dominated convergence results. The one-point (Alexandroff) compactification allows us to let θ approach any boundary place of Θ, even if Θ is not compact and may extend indefinitely. The usual proof of EL consistency (Qin and Lawless, 1994) requires the existence of the continuous derivative of $m(X,\theta)$ and that the derivative is of full rank. Condition 2.1 is less restrictive because it allows for irregular cases where the usual "delta method" does not work, e.g. when $m(x,\theta)$ is non-differentiable. Condition 2.1 (i)-(iv) are the standard M-estimator conditions in Huber (1981) and are very weak in the context of parametric models.

Theorem 2.2. *If Condition 2.1 holds, then every sequence T_n satisfying*

$$T_n := \arg\sup_{\theta\in\Theta} \sum_{i=1}^{n} \log n\tilde{p}_i(\theta) = \arg\sup_{\theta\in\Theta} n\Lambda_n(\theta)$$

will converge to θ_0 almost surely.

Remark 2.3. Kitamura and Stutzer (1997) relax the assumptions in Qin and Lawless (1994) and Kitamura et al. (2004) and obtain consistency of the estimator based on Wald's approach (Wald, 1949). Newey and Smith (2004) assume the differentiability of Lagrangian multiplier rather than that of $m(x,\theta)$. Schennach (2007) gives another consistency proof for a non-differentiable objective function and avoids applications of a Taylor expansion. The differentiability of the moment restriction, however, is assumed in order to obtain a valid approximation for the Lagrangian $\lambda(\theta)$. In this chapter, the assumptions are similar to the standard M-estimator conditions in Huber (1981), thus the differentiability assumption is not required.

2.3 GAUSSIAN PROPERTIES, METRIZATION AND LOCALIZATION OF EL

To get a standard result for the estimator, the criterion function has to satisfy some regularity conditions. The attempt in this chapter is to consider the situations where regularity conditions may be violated, so a weaker counterpart of the conditions is called for. Here the problem appears. The Lagrangian multiplier λ_n gives a dual of the constraint function. But the functional form of λ_n has no closed-form representation, since it is the solution of Equation (2.1) depending on sample size and parameter values. Because of this feature, the general techniques such as empirical processes of studying irregular behavior of the functions are not directly applicable. While the usual asymptotic theorems that rely on differentiability and smoothness of the criterion functions in moment-based estimations are too restrictive in the situations on which we focus, we need alternative conditions and specifications.

We come back to the probability family $\mathcal{E}_\theta$. If $\mathcal{E}_\theta$ belongs to an ideal space, e.g. a complete separable metric (Polish) space, the ordinary topology for assessing weak convergence of $\tilde{P} \in \mathcal{E}_\theta$

is *relative compactness*[2]. The relative compactness will induce a well-defined likelihood ratio function.

Condition 2.4. For any sequence

$$\Lambda_n((X_1,\ldots,X_n),\theta) = n^{-1}\sum_{i=1}^{n}\log n\tilde{p}(\theta,X_i)$$

evaluated at any fixed θ, there are constants a_n such that

$$\Lambda_n((X_1,\ldots,X_n),\theta) - a_n$$

forms a relatively compact sequence.

Lemma 2.5. *Given Condition 2.4, for every fixed n and θ, the random variable $\log n\tilde{p}(\theta,X_i)$ has bounded variance.*

Note that Condition 2.4 still allows a single (or a certain proportional) value of $\log n\tilde{p}_i(\theta)$ to go to infinity at some x_i if only the speed is slower than exponential rate of n. Condition 2.4 and Lemma 2.5 imply that $\Lambda_n(\theta)$ can be thought of as a well-defined random variable or a realization of a likelihood ratio process $\Lambda((X_1,\ldots,X_n),\theta)$ with indices n and θ. In the following subsection, we will see how to connect this process with a local representation.

2.3.1 *Approximation for an Infinitely Divisible Family*

A non-closed form λ_n induces a non-closed form vector

$$(\tilde{p}_1(\theta),\ldots,\tilde{p}_n(\theta)).$$

It is better to transfer the attention of the implied probability vectors to a family of probability measures

$$\mathcal{E}_\theta := \{\tilde{P}_\theta : \theta\in\Theta, \int m(X,\theta)d\tilde{P}_\theta = 0\},$$

where the discrete vector $(\tilde{p}_1(\theta),\ldots,\tilde{p}_n(\theta))$ satisfies

$$\sum_{i}^{n} m(x_i,\theta)\tilde{p}_i(\theta) = 0.$$

2 A sequence of statistics S_n associated with $\tilde{P}$ is *relatively compact* if for every $\epsilon>0$ there is a number $b(\epsilon)$ and an integer $N(\epsilon)$ such that $n\geq N(\epsilon)$ implies $\tilde{P}\{|S_n|>b(\epsilon)\}<\epsilon$.

If a random variable ξ, for every natural number n, can be represented as the sum

$$\xi = \xi_{1,n} + \xi_{2,n} + \cdots + \xi_{n,n}$$

of n i.i.d random variables $\xi_{1,n}, \ldots, \xi_{n,n}$, then ξ is called *infinitely divisible* (Gnedenko and Kolmogorov, 1968, p. 78). A probability distribution is said to be infinitely divisible if and only if it can be represented as the distribution of the sum of an arbitrary number of i.i.d random variables. A family of such distributions is often referred to as an *infinitely divisible family*. In our case, for arbitrary sample size n and fixed θ, the log-likelihood ratio process is

$$\Lambda((X_1, \ldots, X_n), \theta) = \log n\tilde{p}(X_1, \theta) + \cdots + \log n\tilde{p}(X_n, \theta).$$

Every additional term $\log n\tilde{p}(X_i, \theta)$ is an i.i.d increment of this log-likelihood ratio process.

Condition 2.6. Let $\varpi_{n,\theta}(\cdot) = \sum_{i=1}^{n} \left[\frac{m(X_i,\theta)}{1+(\cdot)m(X_i,\theta)} \right]$. Given θ,

$$\lim_{n\to\infty} \varpi_{n,\theta}(\cdot)$$

is independent of $m(X_i, \theta)$ for $0 < i \leq n$.

In a localization approach, when θ is given, λ_n as a solution of the nonlinear equation (2.1) depends on the aggregated element $\varpi_{n,\theta}(\cdot)$. For sufficient large n, an aggregated element may be independent of its individual element. For example, the average of a summation of i.i.d variables will converge to a Gaussian random variable, but the i.i.d variable itself does not necessary to be Gaussian. When this is the case for λ_n, λ_n, which is independent of $m(X_i, \theta)$, is simply a stochastic factor in all $\log n\tilde{p}(X_i, \theta)$, $0 < i \leq n$.

For given a sufficient large n and θ, then

$$\log n\tilde{p}(X, \theta) := -\log(1 + \lambda_n^T m(X, \theta))$$

is random and $(\log n\tilde{p}(X_1, \theta), \ldots, \log n\tilde{p}(X_n, \theta))$ is a random vector where λ_n is a stochastic factor for n-sample size problems. For sufficient large n, Condition 2.6 implies that $\log n\tilde{p}(X_i, \theta)$ are identically distributed and independent. Then one can think that the integral of the log-likelihood ratio process $\int \log \frac{dP_\theta}{dP_0}(x) dP_0(x)$ for given sample size n, $\sum_i \log n\tilde{p}(X_i, \theta)$, as representing an infinite divisible process ξ in n additive terms $\xi_{1,n} + \xi_{2,n} + \cdots +$

$\xi_{n,n}$[3]. Thus $\mathcal{E}_\theta$ does not merely include the family of distributions that satisfy the constraint $\int m(x,\theta)d\tilde{P}_\theta(x)$, it also requires the sample average of the log-likelihood ratio process of $\tilde{P}_\theta$ to be infinitely divisible. It is clear now that EL *inherits* the moment constraint from moment-based methods and *inherits* the infinitely divisibility from likelihood ratio based methods.

An infinitely divisible family $\mathcal{E}$ admits a representation $\mathcal{E} = \mathcal{E}_1 \times \cdots \times \mathcal{E}_n = \otimes_{i=1,\ldots,n}\mathcal{E}_i$ based on n copies of the so called divisor $\mathcal{E}_i$, where n could be arbitrarily large and $\times$ denotes the direct product. The family $\mathcal{E}$ is called *divisible* with divisor $\mathcal{E}_i$. There are several well known infinitely divisible families, e.g. Poisson and Gaussian families.

It has been proved by Gnedenko and Kolmogorov (1968, Theorem 17.5) that any infinitely divisible family can be approximated by a finite number of Poisson type measures. This is an extremely useful result. It basically means that the infinitely divisible family constructed by $\{\log n\tilde{p}(X,\theta)\}$ can be approximated by a finite number of Poisson measures[4]. We know that the Poisson family can be related to the Gaussian family, for example via Hellinger's affinity. We will use this property to deduce a representation of the likelihood ratio process.

Theorem 2.7. *If $\tilde{P}_\theta$ is infinitely divisible then when $n \to \infty$, the log-likelihood $\log d\tilde{P}_{\theta+\delta_n\tau_n}/d\tilde{P}_\theta$ can be approximated by a linear quadratic expression such that the difference*

$$\sum_{i=1}^{n} \log \frac{d\tilde{P}_{\theta+\delta_n\tau_n}}{d\tilde{P}_\theta}(x_i) - \left[\tau_n^T S_{\theta,n} - \frac{1}{2}\tau_n^T K_{\theta,n}\tau_n\right] \tag{2.2}$$

tends to zero in probability for any bounded sequence $\{\tau_n\}$ with a random vector $S_{\theta,n}$ and a deterministic matrix $K_{\theta,n}$.

The infinite divisible feature gives us a useful representation for the likelihood ratio process, a linear quadratic expression with a local parameter τ_n. With this expression, we can construct our estimator without bothering with non-linear optimization, since the parameter in (2.2) is re-parametrized by τ_n which appears linearly and quadratically in the equation. Furthermore, neither the computational algorithm nor the weakly convergent statistics involve any differentiation requirements.

3 More details about such a construction are discussed in Le Cam and Yang (2000, Chapter 5), although in most cases, they use $\log(1+(p_\theta/p_\vartheta)^{-1/2}-1)$ instead of $\log(p_\theta/p_\vartheta)$ directly.

4 We give a short description about Poissonization in the appendix.

Remark 2.8. In the proof, we will show a relation for univariate Gaussian families. For any pair of Gaussian measures G_θ and G_ϑ, there will be a linear-quadratic expression to relate them. Therefore, the integral of $(dG_\theta/dG_\vartheta)^{1/2}$ w.r.t. G_ϑ will have a linear quadratic representation. Then we show that if $\tilde{P}_\theta$ is infinitely divisible, $(d\tilde{P}_\theta/d\tilde{P}_\vartheta)^{1/2}$ will be approximately equal to $(dG_\theta/dG_\vartheta)^{1/2}$, so $(d\tilde{P}_\theta/d\tilde{P}_\vartheta)^{1/2}$ will also have a linear quadratic representation.

Remark 2.9. The linear-quadratic approximations to the log-likelihood ratios can possibly be used with other minimum contrast estimators, but such constructions only lead to asymptotically sufficient estimates, in the sense of Le Cam, when the contrast function mimics the properties of log-likelihood function, at least locally.

Remark 2.10. From a computational aspect, when confronted with the nonlinear optimization, the Hessian matrix of the problem in some cases is difficult to evaluate especially in regions that are either extremely flat or very erratic. It is then computationally more efficient to consider the local optimization and avoid a singular or non-invertible Hessian matrix rather than calculate the global second order derivative of the objective function.

Remark 2.11. Theorem 2.7 shows that with a proper choice of δ_n, the log-likelihood ratio can be approximated by a linear-quadratic representation. One of the main focus of this representation is the quadratic term. As we known, for a pair of Gaussian measures (G_θ, G_ϑ) with dominating measure μ we will have

$$\begin{aligned}
&\int \left(\frac{dG_\theta}{dG_\vartheta}\right)^{\frac{1}{2}} dG_\vartheta = \int dG_\theta^{\frac{1}{2}} dG_\vartheta^{\frac{1}{2}} d\mu \\
=&\mathbb{E}\exp\left\{\sum_{i=\theta,\vartheta}\frac{1}{2}\left[L(i) + \mathbb{E}\log(\frac{dG_i}{d\mu})\right]\right\} \\
=&\left[\exp -\frac{1}{4}\left(K(\theta,\theta)+K(\vartheta,\vartheta)\right)\right]\cdot\mathbb{E}\exp\left(\sum_{i=\theta,\vartheta}\frac{1}{2}L(i)\right) \qquad (2.3)\\
=&\exp\left\{\frac{1}{4}\left[2K(\theta,\vartheta)-K(\theta,\theta)-K(\vartheta,\vartheta)\right]\right\}, \qquad (2.4)
\end{aligned}$$

[5] where $L(i) := \{\log(dG_i/d\mu) - \mathbb{E}\log(dG_i/d\mu)\}$. The property of $L(i)$ includes that it is Gaussian with expectation $\mathbb{E}L(i) = 0$ and covariance kernel $K(\theta,\vartheta) = \mathbb{E}L(\theta)L(\vartheta)$ and we have

$$\mathbb{E}L(i)^2 = K(i,i).$$

Let $q(\theta,\vartheta) = -8\log\int dG_\theta^{\frac{1}{2}} dG_\vartheta^{\frac{1}{2}} d\mu$. Since the quadratic term is deterministic in the neighborhood of θ_0, we can use interpolation to find $K(\cdot,\cdot)$. With an arbitrary mid-point u, three-point interpolation gives us:

$$K(\theta,\vartheta) = -\left(q(\theta,\vartheta) - q(\theta,u) - q(u,\vartheta)\right).$$

For small $|\theta - \vartheta|$, to speed up the computation, one could use an approximated value $\Lambda_n(\theta,\vartheta)$ instead of $q(\theta,\vartheta)$[6].

2.3.2 *Comparison with Other Conditions*

The standard EL ratio can be put into the form of the linear quadratic representation in (2.2) but this requires some additional assumptions, e.g. differentiability of $m(X,\theta)$. The following proposition establishes this relation.

Proposition 2.12. *Suppose that in addition to Condition 2.1, the following holds*

(i) the model is just-identified, $\partial m(X,\theta)/\partial\theta < \infty$ for any X, the rank of $\mathbb{E}[\partial m(X,\theta)/\partial\theta]|_{\theta_0}$ equals $\dim(\theta)$,

(ii) $\frac{1}{n}\sum_{i=1}^n [m_i(\theta)m_i(\theta)^T]$ and $\frac{1}{n}\sum_{i=1}^n [\lambda_n^T m_i(\theta)]^2$ are both finite for any positive n, even as $n \to \infty$,

---

5 The derivation of (2.3) and (2.4) is as follows. Since $\mathbb{E}[\exp(\log(dG_i/d\mu))] = 1$, then we have

$$\mathbb{E}\exp\left[L(i) + \mathbb{E}\log(\frac{dG_i}{d\mu})\right] = \left[\mathbb{E}e^{L(i)}\right]\cdot e^{\mathbb{E}\log(\frac{dG_i}{d\mu})} = 1$$

By the log-normal property, $\mathbb{E}\exp L(i) = e^{\frac{1}{2}K(i,i)}$, we have

$$e^{\frac{1}{2}K(i,i)}\cdot e^{\mathbb{E}\log(\frac{dG_i}{d\mu})} = 1 \iff \mathbb{E}\left[\log(\frac{dG_i}{d\mu})\right] = -\frac{1}{2}K(i,i)$$

thus we have (2.3). For $\mathbb{E}\exp[L(\theta) + L(\vartheta)]$, we have $2K(\theta,\vartheta)$. Combining $2K(\theta,\vartheta)$ and $K(i,i)$ gives us (2.4).

6 The concern is that the square root density computing may induce rounding error. In fact $\frac{1}{2}\log\int(dG_\theta/dG_\vartheta)^{1/2}dG_\vartheta$ approximately equal to $\frac{1}{2}\sum_i \log(dG_\theta/dG_\vartheta)(x_i)$ when x_i is generated by G_ϑ.

then the log-likelihood ratio between $\tilde{p}_{\theta_0}$ *and* $\tilde{p}_{\theta_0+\delta_n\tau}$ *can be approximated by:*

$$2\sum_{i=1}^{n}\log\frac{\tilde{p}_{\theta_0+\delta_n\tau_n}}{\tilde{p}_{\theta_0}}(x_i)=\delta_n\tau_n^T A_1+\frac{1}{2}\delta_n^2\tau_n^T A_2\tau_n^T+o_p(1) \qquad (2.5)$$

where A_1 *is* $\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0)$ *and* A_2 is $\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta^T}$.

The expansion (2.5) is obtained simply by Taylor expansion and the result therefore does not apply to the nonstandard applications we are interested in. However, the result is intuitive as it mimics the standard Local Asymptotic Normal (LAN) property for parametric models, see e.g. van der Vaart (1998, pp 104). The relation between (2.5) and (2.2) is also quite clear: the first term is τ_n times a random vector, and the second term is its variance.

Remark 2.13. With the additional normality assumption on the average of $m_i(\theta_0)$ and assuming $\delta_n=n^{-1/2}$ we will of course have:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}m_i(\theta_0)$$
$$\rightsquigarrow\mathcal{N}\left(0,\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta^T}\right).$$

Asymptotic normality of the EL estimator is established by equation (2.5) with additional conditions on the continuity or the boundedness of second derivative of the moment restriction functions, e.g. Qin and Lawless (1994), Newey and Smith (2004) or Kitamura et al. (2004).

Remark 2.14. An alternative way of deducing this asymptotic normality is via Differentiability in Quadratic Mean (DQM). This entails the existence of a vector of measurable functions $S_{\theta_0,n}$ such that

$$\int\left[\tilde{p}_{\theta_0+\delta_n\tau}^{1/2}-\tilde{p}_{\theta_0}^{1/2}-\frac{1}{2}\delta_n\tau^T S_{\theta_0,n}\tilde{p}_{\theta_0}^{1/2}\right]^2 d\mu=o(\|\delta_n\|^2), \qquad (2.6)$$

where $\delta_n\to 0$. Note the relation between the derivatives of the square root density and the score function (when it exists):

$$2\frac{1}{\sqrt{\tilde{p}_\theta}}\frac{\partial}{\partial\theta}\sqrt{\tilde{p}_\theta}=\frac{\partial}{\partial\theta}\log\tilde{p}_\theta.$$

If along a path, the square root of the implied probability $\theta \mapsto \sqrt{\tilde{p}_\theta}$ is differentiable, then DQM basically means that a expansion of the square root of $\tilde{p}_\theta$ is valid and the remainder term is negligible in $L^2(\mu)$ norm. The term $S_{\theta,n}$ can be considered as the score function of the implied probability $\tilde{p}_\theta$ at θ_0. DQM implies that the condition does not require the pointwise definition of the derivative of $m(\theta, X)$ therefore it is less restrictive.

The implied probability includes the term $m(\theta, X)$ which is not always differentiable in nonstandard cases that we want to consider. It therefore deserves more effort to relax the restrictive condition on differentiability. In fact, Theorem 2.7 implies that the log-likelihood ratio belongs to the LAN family. The result is already good enough for constructing an efficient (or asymptotic sufficient) estimator. The expression in (2.2) is much weaker than the regular conditions and DQM. It only states that log-likelihood ratios of implied probabilities can be approximated by a linear-quadratic expression.

2.4 ESTIMATION

2.4.1 *Local Estimation*

By the result (2.2) in Theorem 2.7, we can study the behavior of a pair $(\tilde{P}_{\theta+\delta_n\tau_n}, \tilde{P}_\theta)$ by looking at the log-likelihood ratio process $\Lambda_n(\theta + \delta_n\tau_n, \theta)(X)$ with index τ_n. The log-likelihood ratio process admits linear quadratic approximations as $n \to \infty$, with the term $\tau_n S_n$ linear in τ_n and the term $\tau_n^T K_n \tau_n$ quadratic in τ_n. The numerical values of the approximation depend on the concentrated point θ and its local neighborhoods. With these ideas in mind, we will show the following steps of constructing a local type estimator. The explanation of each step is given after the definition.

Definition. Given Condition 2.1, we define the following Le Cam type local EL estimator in 5 steps:

Step 1. Find an auxiliary estimate θ_n^* using a δ_n-*consistent* estimator and restricted such that it lies in Θ_n (a δ_n-sparse discretization of Θ).

Step 2. Construct a matrix K_n with $K_{n,i,j} = u_i^T K_n u_j$, $i, j = 1, 2, \ldots, d$, given by

$$\begin{aligned} K_{n,i,j} = -\big\{ &\Lambda_n[\theta_n^* + \delta_n(u_i + u_j), \theta_n^*] \\ &- \Lambda_n[\theta_n^* + \delta_n u_i, \theta_n^*] - \Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*] \big\} \end{aligned}$$

and $\{u_1, \ldots, u_d\}$ is a linearly independent set of directional vectors in $\mathbb{R}^d$ selected in advance.

Step 3. Construct the linear term:

$$u_j^T S_n = \Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*] + \frac{1}{2} K_{n,j,j}.$$

Since all the right hand side values are known, S_n can be computed and is a proper statistic.

Step 4. Construct the adjusted estimator:

$$T_n = \theta_n^* + \delta_n K_n^{-1} S_n.$$

Step 5. Return the value of $\sum_i \log n\tilde{p}_{T_n}(x_i)$ and if it is larger than $\sum_i \log n\tilde{p}_{\theta_n^*}(x_i)$ then the estimator is T_n, otherwise it is θ_n^*. When $T_n \neq \theta_n^*$, if the difference between $\lambda_n(\theta_n^*)$ and $\lambda_n(T_n)$ is larger than a certain criterion value then go back to Step 2 and replace θ_n^* with T_n.

STEP 1 The *δ_n-sparse* (discretization of the) parameter space in Step 1 is suggested by Le Cam (see Le Cam and Yang (2000, p 125)). It requires a sequence of subsets $\Theta_n \subset \Theta$ satisfying the following conditions (i) that for any $\theta \in \Theta$ and any constant $b \in \mathbb{R}^+$, the ball $B(\theta, b\delta_n)$ contains a finite number of elements of Θ_n, independent of n, and (ii) that there exist a $c \in \mathbb{R}^+$ such that any $\theta \in \Theta$ is within a distance $c\delta_n$ of a point of Θ_n. If we think of Θ_n as nodes of a grid with a mesh that gets finer as n increases, then (i) says that the grid does not get too fine too fast and (ii) says that the mesh refines fast enough to have nodes close to any point in the original space Θ. In other words, asymptotically θ_n^* should be close enough to θ_0. Another interpretation of δ_n-sparsity is from a Bayesian perspective. That is for arbitrary priors, the corresponding posteriors essentially concentrate on the small vicinities shrinking at the rate δ_n.

STEP 2 As in the remark in previous section, the covariance matrix in Step 2 is an analog to the covariance kernel in Gaussian processes. For a stationary Gaussian process, the covariance kernel is smooth and differentiable in quadratic mean, the covariance kernel can be written as

$$\begin{aligned}
&\operatorname{Cov}\left(\frac{1}{\delta_n}(G_{\theta+u\delta_n} - G_\theta), \frac{1}{\delta_n}(G_{\vartheta+u\delta_n} - G_\vartheta)\right) \\
&= \frac{1}{\delta_n^2}\left(2C(\theta-\vartheta) - C(\theta-\vartheta+u\delta_n) - C(\theta-\vartheta-u\delta_n)\right) \\
&\to -\left.\frac{\partial^2 C(h)}{\partial h^2}\right|_{h=\theta-\vartheta},
\end{aligned}$$

where $C(\theta, \vartheta) := \mathrm{Cov}\{G_\theta, G_\vartheta\}$. Since K_n is an analog to the covariance kernel, the construction of K_n is nothing else but a finite difference of $\Lambda_n(\cdot,\cdot)$ which is analogous to the second derivative of the covariance kernel.

STEP 3 AND 4 With a control term K_n which is asymptotically determined, all the randomness of the log-likelihood ratio is contained in the first term, S_n. Step 3 is to extract the randomness from $\Lambda_n(\cdot,\cdot)$ and construct the linear term. Step 4 is to construct the estimator. To verify these two steps, we need to ensure that the covariance kernel in (2.2) is invertible.

Proposition 2.15. *The matrices $K_{\theta,n}$ in (2.2) are almost surely positive definite. Any cluster point K_θ of $K_{\theta,n}$ in $P_{\theta,n}$-law is invertible.*

If $K_n - K_{\theta,n}$ converges to zero, then K_n is also invertible. This result will be given in the following Theorem. If K_n is positive definite, by substituting $S_n = K_n\delta_n^{-1}(T_n - \theta_n^*)$ into the linear quadratic expression:

$$\tau_n^T K_n \delta_n^{-1}(T_n - \theta_n^*) - \frac{1}{2}\tau_n^T K_n \tau_n = -\frac{1}{2}\delta_n^{-2}\left[T_n - (\theta_n^* + \delta_n\tau_n)\right]^T K_n \times \left[T_n - (\theta_n^* + \delta_n\tau_n)\right] + \frac{1}{2}\delta_n^{-2}\left[T_n - \theta_n^*\right]^T K_n \left[T_n - \theta_n^*\right],$$

we have a quadratic expression of T_n and $(\theta_n^* + \delta_n\tau_n)$. The maximal value of this approximating representation of the log-likelihood ratio is achieved when $\theta_n^* + \delta_n\tau_n = T_n$. In other words, $\delta_n^{-1}(T_n - \theta_n^*)$ is the estimator for the local parameter τ_n.

Remark 2.16. The construction was originally proposed by Le Cam (1974). He supposed that there is a special interest in the likelihood function at particular points where Taylor's expansion fails, e.g. for the Laplace distribution. The advantage of the construction is that the quadratic term does not depend very much on the particular auxiliary estimation method that is used to obtain the value of θ_n^* and the construction is only determined in a local neighborhood of the particular point.

Remark 2.17. One may be concerned with the δ_n-consistency requirement for the auxiliary estimator. For a simple i.i.d. case, the δ_n is set to $n^{-1/2}$, the requirement is the same as asking for an $\sqrt{n}$-consistent auxiliary estimator. Any $\sqrt{n}$-consistent estimator should be, in principle, good enough from the estimation perspective, because the auxiliary estimator θ_n^* is at least in a neighborhood of θ_0. However, in practice, it may be hard to find a well behaved moment restriction function around θ_0. The use

of local EL estimator is to overcomes the problem and improve the auxiliary estimator. We suppose that θ_n^* is located within a range $n^{-1/2}$ of the true value, then a local method would give a refinement. When consistency and asymptotic normality are treated separately, one could take good care of consistency first and then use localization method to improve the final result or one could take care of the concentration of distribution first and then correct the bias by localization.

Theorem 2.18. *Given conditions 2.1 and 2.4, T_n, S_n and K_n have following properties:*

(i) $K_n^{-1}S_n - K_{\theta,n}^{-1}S_{\theta,n}$ and $K_n - K_{\theta,n}$ converge to zero in $\tilde{P}_{\theta,n}$-law where $(K_{\theta,n}, S_{\theta,n})$ is in (2.2).

(ii) $\delta_n^{-1}(T_n - \theta)$ is bounded in $\tilde{P}_{\theta,n}$ -law.

(iii) if Equation (2.6) holds and the moment restrictions are just-identifying, the sequence of models $\{\tilde{P}_{\theta,n} : \theta \in \Theta\}$ is LAN and

$$\delta_n^{-1}(T_n - \theta_0) \rightsquigarrow \mathcal{N}(0, \Omega)$$

where $\Omega = \mathbb{E}\frac{\partial m(x,\theta_0)}{\partial \theta}^T \left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1} \mathbb{E}\frac{\partial m(x,\theta_0)}{\partial \theta^T}$.

The LAN theory is useful in showing that many statistical models can be approximated by Gaussian models. In the parametric likelihood framework, when the original model P_θ is smooth in the parameters, i.e. DQM, the local parameter $\tau_n = \delta_n^{-1}(\theta_0 - \theta_n^*)$ can be used to construct a log likelihood ratio based on $P_{\theta_0+\tau_n\delta_n}$ that is asymptotically $\mathcal{N}(\tau_n, I_{\theta_0}^{-1})$. Here we use LAN in a moment based setting without further parametric assumptions. Once LAN is established, asymptotic optimality of estimators and of tests can be expressed in terms of LAN properties.

Remark 2.19. Some other articles also utilize local information based on an EL framework. Donald et al. (2003) propose re-sampling data from a local EL estimated distribution. Kitamura et al. (2004) consider another localized EL based on conditional moment restrictions and use them to re-construct a smooth global profile likelihood function. Smith (2005) extends moment smoothing to GEL. These methods construct smooth objective functions, implicitly or explicitly. Our solution is to discretize the parameter space and then construct local log-likelihood ratios as local objective functions.

Theorem 2.18 gives an asymptotic result on the weak convergence of the estimator. In the theorem, the limit distribution is based on a kind of Cramér-Rao type lower bound and is

essentially a pointwise result. In order to obtain a result in a neighborhood rather than at a single point, we will now state and prove a minimax type theorem on the risk of any estimator.

Before giving the theorem, we need to introduce a technical concept of δ_n-*regularity*. This concept expresses the desirable requirement that a small change in the parameter should not change the distribution of estimator too much. For the estimator sequence T_n, if the difference between the distributions of $\delta_n^{-1}(T_n - \theta_0 - \delta_n\tau)$ and $\delta_n^{-1}(T_n - \theta_0)$ tends to zero under $P_{\theta_0+\delta_n\tau,n}$-law and $P_{\theta_0,n}$-law respectively, then T_n is called δ_n-regular at the point θ_0.

Theorem 2.20. *Given Condition 2.1 and letting W be a non-negative bowl shaped loss function, if T_n is δ_n-regular on all Θ, then for any estimator sequence Z_n of τ, one has*

$$\lim_{b\to\infty}\lim_{c\to\infty}\lim_{n\to\infty}\inf_{n}\sup_{|\tau|\le c}\mathbb{E}_{\theta_0+\delta_n\tau}\left[\min(b, W(Z_n-\tau))\right] \ge \mathbb{E}[W(\xi)]$$

where ξ has a Gaussian distribution $\mathcal{N}(0, K^{-1})$. The lower bound is achieved by $Z_n = \delta_n^{-1}(T_n - \theta_0)$.

A loss function is "bowl-shaped" if the sublevel sets $\{u : W(u) \le a\}$ are convex and symmetric around the origin. The value b is used to construct a bounded function $\min(b, W(Z_n - \tau_n))$. We let c go to infinity in order to cover a general case. The expectation $\mathbb{E}_{\theta_0+\delta_n\tau}[\cdot]$ is taken w.r.t. a measure $\mathcal{M}$ of the set $\{\theta : |\theta - \theta_0| \le \delta_n\tau_n\}$ while $\mathbb{E}[\cdot]$ is taken w.r.t. a distribution of $K^{-1/2} \times \mathcal{N}(0, I)$ on ξ.

The theorem can be interpreted as follows. When using the auxiliary estimator θ_n^* in the likelihood ratio, this induces randomness to the local parameter τ_n. By using the LAN result in Theorem 2.18, we can attach the local parameter τ_n with a Gaussian measure. By the Gaussian prior assumption of τ_n, one can express the convergent procedure as a procedure of updating a Gaussian prior, while for a centered Gaussian prior, this procedure is to update the prior covariance matrix Γ^{-1}. The δ_n-regularity condition implies that K_n will converge uniformly in a neighborhood of θ_0 for arbitrary measure $\mathcal{M}$. Thus the covariance will converge to a the posterior covariance matrix $(K+\Gamma)^{-1}$. The Gaussian randomness introduces a new random variable ξ that has the posterior covariance matrix $(K+\Gamma)^{-1}$. The lower bound of the Bayes risk of this Gaussian variable is obtained by letting Γ go to zero, corresponding to initial values of τ widely spread. This is the local asymptotic minimax theorem. It is based on the minimax criterion and gives a lower

bound for the maximum risk over a small neighborhood of the parameter θ. Because the local EL can achieve this lower bound, it is an asymptotically optimal estimator.

2.4.2 *Global Estimation*

From our previous discussion, we can see that the construction of T_n relies on the assumption that, locally, the logarithms of likelihood ratios admit approximations of a quadratic nature:

$$\tau_n^T K_n \delta_n^{-1}(T_n - \theta_n^*) - \frac{1}{2}\tau_n^T K_n \tau_n = -\frac{1}{2}\delta_n^{-2}\left[T_n - (\theta_n^* + \delta_n\tau_n)\right]^T K_n \times \left[T_n - (\theta_n^* + \delta_n\tau_n)\right] + \frac{1}{2}\delta_n^{-2}\left[T_n - \theta_n^*\right]^T K_n \left[T_n - \theta_n^*\right].$$

The method is to use the auxiliary estimate from local expansions to construct the estimate T_n and covariance kernel K_n. As we can see, the kernel K_n controls the quality of representation. In the local case, because of δ-sparsity and the auxiliary θ_n^*, the kernel matrix $K = \lim K_n$ is deterministic. If the quadratic nature can be extended to a global region, namely linear quadratic representation valid on the whole or at least a large part of the parameter space as the linear-quadratic representation (2.2), then K is not necessarily fixed; in fact, it should vary with θ.

If we are worried about irregular situations, e.g. flatness, non-smoothness, non-differentiability, on a large part of Θ for Λ_θ, then we should turn to a global method that can mimic the good properties of our local estimator. The difficulty of extending the linear quadratic representation to all of the parameter space Θ comes from the randomness of K_θ. In this section, we will encounter this problem, where the form of the population K_θ could be random. In order to emphasize this property, we will use the notation:

$$\mathbb{K}_\theta(T_n - \theta) := (T_n - \theta)^T K_{\theta,n}(T_n - \theta) \tag{2.7}$$

for the global case, namely $\mathbb{K}_\theta(\cdot)$ is a quadratic random function on $\theta \in \Theta$. Conditional on θ, the linear quadratic approximation still holds, like (2.2). When K is random, the approximation belongs to the Locally Asymptotically Mixed Normal (LAMN) class. The reason for using quadratic form in (2.7) instead of the linear-quadratic one is that both "linear" and "quadratic" terms will be influenced by the choice of parametrization, it is better to use one term rather than two to express this variational effect of θ. Because the constructed estimator will depend on the value of θ, from now on, we will denote T_n as $T_{\theta,n}$.

For practical application, it is more convenient to work in a well-defined space rather than an arbitrary functional space on Θ. We want to attach to the parameter space Θ a vector space $\mathcal{F}(\Theta)$ induced by K_θ which is the "kernel" of this space $\mathcal{F}(\Theta)$. To see the necessity of doing this, one should compare to the situation in LAN. The LAN result can be stated as a pointwise result in θ. Thus the convergence and optimality properties are also valid pointwise like Theorem 2.18, or just a small region around θ like Theorem 2.20. If one wants to extend the result to the whole Θ space, namely, a uniform result on Θ, then one needs an appropriate uniformity condition. If the subspace $\mathcal{F}(\Theta)$ is constructed by K_θ, then an estimate using K_θ will be uniformly valid. The following condition is to define such a subspace.

Condition 2.21. (i) The vector space $\mathcal{F}(\Theta)$ is a linear subspace with finite signed measures and with finite support on Θ. Let

$$\mathcal{F}(\Theta) := \{\nu : \nu(\Theta) = 0\}.$$

We require that for infinite number of observations, any measure $\nu \in \mathcal{F}(\Theta)$ for centering $T_\theta - \theta$ is constructed by K_θ.

(ii) The inner product for $\mathcal{F}(\Theta)$ is defined by bilinear function

$$\langle \theta, \vartheta \rangle_K = \frac{1}{2}\{K_{\theta+\vartheta} - K_\theta - K_\vartheta\} = -\frac{1}{2}\{K_{\theta-\vartheta} - K_\theta - K_\vartheta\}$$

and $\theta \mapsto \sqrt{K_\theta}$ is the semi-norm on $\mathcal{F}(\Theta)$.

(iii) Let $\bar{\Lambda}_\theta(\vartheta, \theta^*) = -\{\mathbb{K}_\theta(T_\theta - \vartheta) - \mathbb{K}_\theta(T_\theta - \theta^*)\}/2$. The function $\bar{\Lambda}_\theta(\cdot,\cdot)$ is additive on its domain such that

$$\bar{\Lambda}_\theta(\vartheta, u) = \bar{\Lambda}_\theta(\vartheta, \theta^*) + \bar{\Lambda}_\theta(\theta^*, u).$$

Condition (i) is to give a tractable structure for the problem. The *null space* $\mathcal{F}(\Theta)$ is of particular interest. For any subset in Θ, K_θ is a quadratic form. The restriction of K_θ is to symmetrize $K(\cdot,\cdot)$. The restriction of $T_\theta - \theta$ is to make T as a centering estimator. (ii) is also called *the polarization condition* which ensures that the parallelogram law holds. This condition attaches a kind of Hilbert space characteristics to the linear space $\mathcal{F}(\Theta)$. One can compare (ii) with Theorem 2.7 where the canonical Gaussian family is attached to the Hilbert space. The condition basically says that we want the constructed quadratic K_θ to inherit such a property automatically. (iii) imposes a feature of presumed midpoints as Step 2 in local estimator's construction. This relation is to connect the quadratics with logarithms.

The motivation of these conditions is to ensure that the entire family $\mathcal{E}_\theta$ is globally approximable by a heteroskedastic Gaussian family where the log-likelihood contains constructed K_θ only and K_θ varies slowly enough as θ varies. Then conditional on any local value θ, one could also approximate $\mathcal{E}_\theta$ by a Gaussian family. The heteroskedastic approximation is uniformly feasible over θ. These conditions are extracted from the properties of *Quadric*, the solution rings of algebraic linear quadratic equations. The simplification allows us to work on a more concrete topological structure rather than on rings.

The aim of a global construction is to find centering values of K_θ and an estimator $T_{\theta,n}$ based on K_θ. Consider the EL log-likelihood ratio

$$\Lambda_n(s) = \frac{1}{n}\sum_{i=1}^{n} \log n\tilde{p}_i(s) = \Lambda_n(s,\theta_0),$$

where θ_0 induces the counting measure. For any initial parameter value t, minimizing $\Lambda_n(s)$ is equivalent to finding some ideal centering values which make the difference

$$\Lambda_n(s) + \frac{1}{2}[\hat{\mathbb{K}}(T_{\theta,n} - s) - \hat{\mathbb{K}}(T_{\theta,n} - t)]$$

tends to zero in probability by updating t, where $\hat{\mathbb{K}} = \mathbb{K}_{T_{\theta,n}}$ and $T_{\theta,n}$ are our estimators.

Definition. Given conditions 2.1 and 2.21, the global estimators are constructed as follow:

Step 1. Give two initial values t and s, compute the EL value $\Lambda_n(t)$.

Step 2. Find a $T_{\theta,n}$ to solve the linear equation system:

$$\langle T_{\theta,n} - s, t - s\rangle_{K_t} = \bar{\Lambda}_t(s,t) + \frac{1}{2}\mathbb{K}_t(s-t) \tag{2.8}$$

where the inner product and $\bar{\Lambda}_t(\cdot,\cdot)$ are defined in condition 2.21 and $\mathbb{K}_t(\cdot)$ is constructed by Step-2 in the local method .

Step 3. Adjust the parameter t by solving the quadratic problem:

$$\frac{1}{2}[\hat{\mathbb{K}}(T_{\theta,n} - s) - \hat{\mathbb{K}}(T_{\theta,n} - t)] = \Lambda_n(s).$$

If $|t^* - t|$ is larger than a certain critical value, e.g. 10^{-5}, then go back to step 1 and use t^* as the starting value.

It is a recursive algorithm type construction. Step 2 and 3 involve heavier computational tasks compared to the local method. But the whole setting avoids computing derivative and is feasible for arbitrary values s and t.

2.5 ROBUSTNESS

The robustness we consider here is essentially Huber's idea that an estimator is insensitive to perturbations of the model. This includes insensitivity to outliers if we think of outliers as being generated by a different model with a small probability. When the assumed structure of the model is incorrect or the DGP is wrongly specified, one can detect, in principle, the misspecification by various testing procedures. This kind of testing and the deletion of potential outliers, however, directly affects the inference procedures. It should in principle condition on the outcome of the test and take explicitly into account the statistical properties of deleting outliers. In this section we do not consider testing for misspecification and cleaning the data in advance, but analyze an inference procedure that explicitly takes into account that the model can, to a certain extent, be misspecified.

The sensitivity of EL estimation results from the unboundedness of the moment constraints. We borrow Huber's setting to illustrate this problem. Consider our estimator T_n as a statistical functional of an empirical measure P_n such that

$$\frac{1}{n}\sum_{i=1}^{n} m(T_n, X_i) = \frac{1}{n}\sum_{i=1}^{n} m(T(P_n), X_i) = 0. \tag{2.9}$$

The functional $T(P_n)$ it is defined as:

$$T(P_n) := \arg\sup_{\theta\in\Theta} \sum_{i=1}^{n} \log n\tilde{p}_i(\theta),$$

A natural robustness requirement on a statistical functional is the boundedness of its influence function. The influence function of a given statistical functional $T(\cdot)$ is:

$$IF(x, T, P_n) := \lim_{\epsilon\to 0} \frac{T((1-\epsilon)P_n + \epsilon\Delta_x) - T(P_n)}{\epsilon}$$

for all x's that make the limit exist, where Δ_x is the probability measure giving mass 1 to x. An alternative way is to think of T as a linear functional which is continuous w.r.t. a weak(-star) topology, namely the map

$$P \mapsto \int \psi dP = T(P)$$

from the space of all probability measures on the sample space to $\mathbb{R}$ is continuous whenever ψ is bounded and continuous. If

ψ is not bounded, a single error can completely upset $T(P_n)$; if ψ is not continuous, a mass on these discontinuity points may cause a significant change for $T(P_n)$ with even a small change in subsample of $\mathcal{X}$. Note that the influence function asks for stronger condition. Because $IF(x, T, P_n)$ is defined by the functional derivative in Δ_x direction while in the linear functional it implies bounded and continuity.

For example, Ronchetti and Trojani (2001) show that the influence function of exactly identified GMM is

$$[\mathbb{E}\partial m_i(T(P_\theta))/\partial T]^{-1} m_i(T(P_\theta))$$

at a single observation x_i. The influence function is unbounded if $m(\theta)$ is unbounded or if the derivative is not defined. An unbounded influence function implies an unbounded asymptotic bias of a statistic at a single-point contamination of the model.

White (1982) shows that Maximum Likelihood (ML) defined in terms of the Kullback-Leibler divergence is robust. EL inherits a lot of properties from ML, so one may expect it is also robust. However, Schennach (2007) gives a counterexample. Suppose the outliers of sample space $\mathcal{X}$ give $\sup_{x\in\mathcal{X}} m_i(\theta) = \infty$ so that $\inf_\theta \sum m_i(\theta)/n \neq 0$ for any $\theta \in \Theta$ but $\mathbb{E}[\|m(\theta, X)\|^2] < \infty$. The λ associated with these outliers' moment restriction functions will give strong penalties so that the values of λ will stay close to zero independently of the value of θ. The implied density $\tilde{p}_i$ of each outlier's moment restriction function equals $1/n$. When the sample size is very small, this means that the effects of relative weights on the outliers are strong, the criterion function $\sum_i^n \log n\tilde{p}_i(\theta)$ will be quite flat on θ. In large samples, the intrinsically misspecified EL estimator will have a slower convergence rate, although it may be consistent.

As we see, the statistical functional of EL estimation is a solution of (2.9). The moment restriction function $m(T(P_n), x)$ as in the previous discussion is discontinuous on those peculiar points causing unbounded "influence function" of $T(P_n)$ and (2.9) is non-differentiable on these points as well, thus EL is not a robust estimation procedure because of the non-robust moment restriction functions, not the EL procedure itself. If one can eliminate the outlier's influences, one will keep the robustness of EL. Localization can prevent such misbehavior.

If a peculiar point x_k drives $T(P_n)$ unbounded in a direction $u_i \in \Theta$, then the moment restrictions function $m(T(P_n), x_i)$ becomes unbounded and λ_n may not be zero because

$$\sum_i m(T(P_n), x_i) \neq 0.$$

However, we have a strong belief that the model is correctly specified. Thus we force λ_n to zero for the moment restriction and exclude the effect from the peculiar point. We achieve this goal by selecting a direction of the local estimation which guides λ_n to zero.

Note that the auxiliary estimator in local EL, θ_n^* in the range of order δ_n, does admit a good quadratic approximation of the log-likelihood ratios. This regularizes the matrix in the quadratic term to be positive semi-definite. In step 2 of the local EL's construction, we let

$$K_{n,i,j} = -\left\{\Lambda_n[\theta_n^* + \delta_n(u_i + u_j), \theta_n^*] - \Lambda_n[\theta_n^* + \delta_n u_i, \theta_n^*] \right. \\ \left. -\Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*]\right\}.$$

The direction u in $K_{n,i,j}$ is constructed by a bisection type method. The correctly specified model must satisfy an auxiliary condition $f(\theta_0) = \lambda_n^T(\sum_i m(\theta_0, x_i)) = 0$. $\lambda_n(\theta_0)$ is dual parameter of $\sum_i m(\theta_0, x_i)$, when $\sum_i m(\theta_0, x_i) = 0$ is zero then $\lambda_n(\theta_0)$ is zero. Suppose $f(\theta_n^*)$ is positive. We select a direction u such that $f(\theta_n^* + u) = -f(\theta_n^*)$. By the mean value theorem, $f(\theta)$ must have at least one root $f(\theta) = 0$ between $\tilde{\theta}$ and θ_n^*. The constructed $K_{n,i,j}$ and S_n using this direction will make the local estimator T_n leave from θ_n^* to θ_0.

Remark 2.22. One may ask why we should use robust estimation rather than give a mis-specification test or clean the data first. Essentially, there is no "mis-specification" in the model, the moment constraint functions are correctly specified, although certain sample points may lead to discontinuity and the unboundedness problems previously described. Applying mis-specification test may reject this essentially correct model. An insightful argument is given by Huber (1981).

> Even if the original batch of observations consists of normal observations interspersed with some gross errors, the cleaned data will not be normal, and the situation is even worse when the original batch derives from a genuine non-normal distribution, instead of from a gross-error framework. Therefore the classical normal theory is not applicable to cleaned samples, and the actual performance of such a two-step (clean and test) procedure may be more difficult to work out than that of a straight robust procedure.
> -(Huber, 1981, Chapter 1)

Remark 2.23. Schennach (2007) shows that an Exponential Tilting Empirical Likelihood (ETEL) estimation is robust and almost as efficient as EL. However, it is still less efficient in the higher order than EL. Moreover, the procedure of ETEL changes the divergence criterion in the intermediate step[7]. This action will discard not only the weights of outliers but also some informative weights used to capture the fat tail feature of sample distributions. Ronchetti and Trojani (2001) construct a Huber-type GMM estimators based on a bounded self-standardized norm of the given orthogonality function. They also show that imposing this robustness correction has an impact on the power of the mis-specification test.

2.6 CONCLUSION

In this chapter, we propose a new local EL method. We discuss its construction and have derived theoretical properties. The construction is based on the infinite divisibility property which is one of the crucial features in stochastic processes; to the best of our knowledge, this feature has not yet been applied to EL. When the implied probability of EL is embedded in the infinitely divisible class, the log-likelihood ratio admits a local representation. Our local estimator is built on the basis of this representation. The consistency, local asymptotic normality, and asymptotic optimality of this estimator have been established. These results depend on conditions that are weaker than usual and allow for applications when the standard regularity conditions are violated.

7 In first step, the criterion function is to minimize the log-likelihood ratio over empirical entropy while in the second step the criterion function is to minimize the log-likelihood ratio over sample average.

APPENDIX TO CHAPTER 2

PROOF OF THEOREMS

Proof of Theorem 2.2

The Lagrangian of EL is

$$L = \sum_{i=1}^{n} \log(np_i) - n\lambda^T \sum_{i=1}^{n} p_i m_i(\theta) - \gamma \left(\sum_{i=1}^{n} p_i - 1 \right),$$

where λ and γ are Lagrange multipliers. Setting the partial derivative of L w.r.t. p_i equal to zero will give $\gamma = n$ and the implied probability $\tilde{p}_i = 1/(\gamma + n\lambda_n^T m_i(\theta))$. By the implicit function theorem, the partial derivative of $\sum_{i=1}^{n} \log \tilde{p}_i$ w.r.t. λ gives a function $Im(\cdot,\cdot)$ of λ_n and θ such that

$$\begin{aligned} \frac{\partial \sum \log \tilde{p}_i}{\partial \lambda} &:= Im(\lambda_n, \theta) = 0, \qquad (2.10) \\ \Longrightarrow \frac{1}{n} \sum_{i=1}^{n} \frac{m_i(\theta)}{1 + \lambda_n^T m_i(\theta)} &= \sum_{i=1}^{n} \tilde{p}_i(\theta) m_i(\theta) \end{aligned}$$

where λ_n is unique for fixed n and θ. Note that $Im(\lambda_n, \theta) = 0$ for $\forall \theta \in \Theta$ and θ is continuous hence $Im(\cdot)$ is continuous in θ. By the continuity of $m(X, \theta)$ and the representation of $Im(\cdot)$, we know that λ_n is also continuous on θ. The proof of the uniqueness of $\lambda(\theta)$ is as follows: because the set $\Gamma(\theta) = \lim_{n\to\infty} \cap_{i=1,\ldots,n} \{\lambda | 1 + \lambda^T m(X_i, \theta) > 1/n\}$ is convex if it does not vanish, the function of $\log p$ is strictly concave on λ, so $\lambda(\theta)$ exists and is unique.

With these, the properties of likelihood ratio are shown in as follows. Equation (2.10) can be re-written as

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \left[1 - \frac{\lambda_n^T m_i(\theta)}{1 + \lambda_n^T m_i(\theta)} \right] m_i(\theta) = 0 \\ \Longrightarrow \frac{1}{n} \sum_{i=1}^{n} m_i(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \frac{m_i(\theta) \lambda_n^T m_i(\theta)}{1 + \lambda_n^T m_i(\theta)} \\ &= \underbrace{\left[\sum_{i=1}^{n} \tilde{p}_i(\theta) m_i(\theta) m_i(\theta)^T \right]}_{(*)} \lambda_n. \end{aligned}$$

Condition 2.1 (v) states that $n^{-1}\sum_i^n m_i(\theta)m_i(\theta)^T$ is positive definite, let $\mathbf{c}$ be larger than any eigenvalue of $n^{-1}\sum_i^n m_i(\theta)m_i(\theta)^T$ and let v be the corresponding eigenvector. The convex combination of $m_i(\theta)m_i(\theta)^T$ over $\{\tilde{p}_i(\theta)\}$ in $(*)$ is bounded by $v^T\mathbf{c}v$. Let $E_v = v^T\mathbf{c}v$. According to condition 2.1 (iv), $m_i(\theta)$ has an envelop function $b(\theta)$ such that $\liminf_\theta |m(\theta,X)|/b(\theta) \geq 1$, then

$$\lim_{n\to\infty} |\lambda_n|/b^{'}(\theta) \geq 1$$

for any θ where $b^{'}(\theta) = b(\theta)/E_v$.

Let's first prove the existence of $\Lambda(\theta)$:

$$\lim_{n\to\infty}\int \log\frac{1}{n}\frac{n}{1+\lambda(n,\theta)^T m(x,\theta)}dP(x) = \tag{2.11}$$

$$\mathbb{E}\lim_{n\to\infty}\log\frac{1}{1+\lambda(n,\theta)^T m(X,\theta)} = \mathbb{E}\log\frac{1}{1+\lambda(\theta)^T m(X,\theta)} = \Lambda(\theta).$$

The first convergence is by the LLN and the second equation is obtained by the dominated convergence Theorem, since $[1+\lambda(\theta)^T m(x,\theta)]^{-1}$ is bounded and $\lambda(\theta)$ exists.

Next we prove the continuity of $\Lambda(\theta)$. The envelop functions $b^{'}(\theta)$ and $b(\theta)$ are integrable and continuous (Condition 2.1), $\lambda(\theta)^T m(X,\theta)$ is bounded by a continuous function. Thus $\Lambda(\theta)$ is continuous and is bounded by an envelop function $b^{''}(\theta) = \max(b^{'}(\theta), b(\theta))$ such that

$$\sup_\theta \|\Lambda_n(\theta) - \Lambda(\theta)\|/b^{''}(\theta) < 1 \tag{2.12}$$

Now prove the identifiability of EL estimation. Choose a compact set $\Theta_c \subset \Theta$ such that for given ϵ

$$\sup_{\theta\in\Theta_C} |\Lambda(\theta)|/b^{''}(\theta) \geq 1-\epsilon.$$

By (2.12), LLN applied to $\Lambda_n(\theta)$ implies

$$\begin{aligned}\sup_\theta \frac{\|\Lambda_n(\theta)-\Lambda(\theta)\|}{b''(\theta)} &< \frac{\|\Lambda_n(\theta)\| - \|\Lambda(\theta)\| + 2\|\Lambda(\theta)\|}{b''(\theta)} - \epsilon \\ &< \frac{\|\Lambda_n(\theta)\| - \|\Lambda(\theta)\| + 2\sup_{\theta\in\Theta_C}|\Lambda(\theta)|}{b''(\theta)} - \epsilon \\ &< \frac{\|\Lambda_n(\theta)-\Lambda(\theta)\| + 2\sup_{\theta\in\Theta_C}|\Lambda(\theta)|}{b''(\theta)} < 1-3\epsilon\end{aligned}$$

The first inequality uses triangle inequality, the second one uses supremum property, and the third one uses triangle inequality again. Therefore

$$|\Lambda_n(\theta) - \Lambda(\theta)| \leq (1-3\epsilon)b''(\theta)$$
$$\leq \frac{1-3\epsilon}{1-\epsilon} \sup_{\theta\in\Theta_C} |\Lambda(\theta)| \leq (1-\delta) \sup_{\theta\in\Theta_C} |\Lambda(\theta)|$$

for $\forall\theta \in \Theta_c$. This inequality implies

$$\sup_{\theta\in\Theta_c} |\Lambda_n(\theta)| \leq \sup_{\theta\in\Theta_c} |\Lambda(\theta)| + \epsilon$$

asymptotically for any $\theta \in \Theta_c$. Thus if $\theta_0 \in \Theta_c$, then

$$\{T_n \subset \Theta_c\} \subset \left\{ \sup_{\theta\in\Theta_c} \Lambda_n(\theta) \leq \Lambda(\theta_0) + o_p(1) \right\},$$

where the probability of the event on the right side converges to one as $n \to \infty$. Because the compact set Θ could be shrinking to an arbitrary neighborhood of θ_0, the EL estimator T_n is consistent.

Proof of Theorem 2.7

Before proving the Theorem, we need to introduce a relation for univariate Gaussian families. For any pair of Gaussian measures in $\mathcal{G}_\Theta = \{G_\theta, \theta \in \Theta\}$, $G_\theta \subset \mathcal{E}_\theta$, there will be an expression to relate both of them as follows:

$$dG_\theta = \exp\left[\langle Y_\vartheta, \theta\rangle - \frac{1}{2}\|\theta\|^2\right] dG_\vartheta, \tag{2.13}$$

$$\exp\left[\frac{1}{2}\langle Y_\vartheta, \vartheta + \theta\rangle\right]$$

where $\vartheta, \theta \in \Theta$. The bilinear product in this expression is $\langle Y_\vartheta, \theta\rangle = \int_0^1 Y_\vartheta(t) G_\theta(dt)$ where Y_ϑ is a univariate Gaussian process. This is a random variable (functional integral or Wiener integral) with mean zero and variance $\|\theta\|^2 \leq \infty$ [8]. If dG_θ and dG_ϑ are defined as (2.13), the integral of $(dG_\theta/dG_\vartheta)^{1/2}$ w.r.t. G_ϑ will has a linear quadratic representation.

8 This expression is called weak form expression and is often used for generalizing Gaussian processes.

Proof. Le Cam and Yang (2000, Proposition 4.1) show that the affinity between two Poissonized $d\tilde{P}_\theta, d\tilde{P}_\vartheta$ is

$$\int \sqrt{d\tilde{P}_\theta d\tilde{P}_\vartheta} = \exp\left\{-\frac{1}{2}\|\theta-\vartheta\|^2\right\}.$$

Since Gnedenko and Kolmogorov (1968, Theorem 17.5) show that finite many number of Poisson type measures can approximate any infinitely divisible family and EL is embedded in an infinitely divisible family, we know the above expression is applicable over here. The Hellinger affinity for Gaussian family is

$$\int \sqrt{dG_\theta dG_\vartheta} = \int \exp\left[\frac{1}{2}\langle Y_\vartheta, \vartheta+\theta\rangle - \frac{1}{4}(\|\theta\|^2+\|\vartheta\|^2)\right] dG_\vartheta.$$

The Gaussian property of $\langle Y_\vartheta, \vartheta+\theta\rangle$implies that is log-normal distributed, then by log-normal property there is:

$$\int \exp\left[\frac{1}{2}\langle Y_\vartheta, \vartheta+\theta\rangle\right] dG_t = \exp\left(\frac{1}{8}\|\theta+\vartheta\|^2\right).$$

Because only metric distance is going to be studied in $\int\sqrt{dG_\theta dG_\vartheta}$, we attach a Hilbert space to $\mathcal{G}$. The parallelogram identity for Hilbert space induces

$$\|\theta+\vartheta\|^2+\|\theta-\vartheta\|^2 = 2\left(\|\theta\|^2+\|\vartheta\|^2\right),$$

so

$$2\left(\|\theta\|^2+\|\vartheta\|^2\right)+\|\theta+\vartheta\|^2 = -\|\vartheta-\theta\|^2$$

Therefore, $\sqrt{dG_\theta dG_\vartheta} = \exp(-\|\theta-\vartheta\|^2/8)$ which is isometric to $\int\sqrt{d\tilde{P}_\vartheta d\tilde{P}_\theta} = \exp(-\|\theta-\vartheta\|^2/2)$. If Fubini's theorem holds, the expression

$$2\log\int\left(\frac{d\tilde{P}_\theta}{d\tilde{P}_\vartheta}\right)^{\frac{1}{2}} d\tilde{P}_\vartheta \approx 8\log\int\left(\frac{dG_\theta}{dG_\vartheta}\right)^{\frac{1}{2}} dG_\vartheta$$

implies

$$\int\left[\log\frac{d\tilde{P}_\theta}{d\tilde{P}_\vartheta}\right] d\tilde{P}_\vartheta = 4\int\left[\log\frac{dG_\theta}{dG_\vartheta}\right] dG_\vartheta$$

so that we can use the Gaussian expression (2.13) for the log-likelihood ratio process.

By Karhunen–Loève Theorem (Kallenberg, 2002), the Gaussian process Y_θ can be expressed as

$$Y_\theta = \sum_{j=1}^{\infty}\xi_j \mathbf{u}_j(\theta)$$

where $\{\mathbf{u}_j\}$ constitutes an orthonormal basis for the Hilbert space $\mathcal{G}$ and ξ_j are Gaussian random variables and stochastically independent. Now let $\mathbf{u}_j(\cdot) = \sum_i^m \tau_i \mathbf{e}_i(\cdot)$ where $\mathbf{e}$ is a unit basis for the local parameter space and τ_i are linear coefficients for $\mathbf{e}_i(\cdot)$. Let j indicate the index of a basis on the Hilbert space and i indicate the index of a basis on the local parameter space. Then the inner product in the Hilbert space can be expressed using local parameter coordinates such that $\langle Y_\vartheta, \theta\rangle = \sum_i^m \tau_i\theta_i\langle \mathbf{e}(\vartheta), \xi\rangle = \tau^T(\theta\tilde{\xi})$ where $\tilde{\xi}$ is also Gaussian because of the linear property. Let $\theta\tilde{\xi} = S_\theta^{'}$ and $\mathbb{E}(\theta\tilde{\xi})^2 = K_\theta^{'}$, then

$$\|\theta\|^2 = \mathbb{E}[\tau^T(\theta\xi)]^2 = \tau^T K_\theta^{'}\tau.$$

From (2.13), we have

$$\int \left[\log \frac{d\tilde{P}_\vartheta}{d\tilde{P}_\theta}\right] d\tilde{P}_\theta = \tau^T S_\theta - \frac{1}{2}\tau^T K_\theta \tau.$$

where $S_\theta = \int S_\theta^{'} d\tilde{P}_\theta$ and $K_\theta = \int K_\theta^{'} d\tilde{P}_\theta$. For a finite dimensional Gaussian vector based on n realizations Gaussian process, we have the sample counterparts τ_n, $S_{\theta,n}$ and $K_{\theta,n}$. We conclude that the EL ratio is approximately equal to the log-likelihood ratio of $\mathcal{G}$, which for the sample of size n is $\tau_n^T S_{\theta,n} - \tau_n^T K_{\theta,n}\tau_n/2$. □

Proof of Theorem 2.18

Proof. (i) When θ is given, by equation (2.2)

$$\begin{aligned}\Lambda_n(\theta + \delta_n\tau_n, \theta) =& \tau_n^T S_{\theta,n} - \frac{1}{2}\tau_n^T K_{\theta,n}\tau_n + o_{\tilde{p}_\theta}(1) \qquad (2.14)\\ =& -\frac{1}{2}\Big[(K_{\theta,n}^{-1}S_{\theta,n} - \tau_n^T)^T K_{\theta,n}(K_{\theta,n}^{-1}S_{\theta,n} - \tau_n^T))\\ & -(S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n})\Big] + o_{\tilde{p}_\theta}(1).\end{aligned}$$

Similarly,

$$\begin{aligned}\Lambda_n(\theta + \delta_n\tau_n, \theta) &= \tau_n^T K_n \delta_n^{-1}(T_n - \theta_n^*) - \frac{1}{2}\tau_n^T K_n \tau_n \qquad (2.15)\\ &= -\frac{1}{2}\Big[(\delta_n(T_n - \theta) - \tau_n^T)^T K_n(\delta_n(T_n - \theta) - \tau_n^T)) \qquad (2.16)\\ &\quad -(\delta_n(T_n - \theta))^T K_n(\delta_n(T_n - \theta))\Big].\end{aligned}$$

The difference between (2.14) and (2.15) tends to zero as $n \to \infty$. Non-negativity of K_n and $K_{\theta,n}$ shows that each of the four

quadratic terms in (2.16) and (2.14) must be non-negative. If $S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$ converges to $(\delta_n(T_n-\theta))^T K_n(\delta_n(T_n-\theta))$, then

$$(\delta_n(T_n-\theta)-\tau_n^T)^T K_n(\delta_n(T_n-\theta)-\tau_n^T)) \to (K_{\theta,n}^{-1}S_{\theta,n}-\tau_n^T)^T K_{\theta,n}(K_{\theta,n}^{-1}S_{\theta,n}-\tau_n^T)).$$

So one can conclude that $K_n - K_{\theta,n} \to 0$ and $S_n - S_{\theta,n} \to 0$.

Now consider the opposite case $(\delta_n(T_n-\theta))^T K_n(\delta_n(T_n-\theta)) \nrightarrow S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$. By a standard property of quadratic functions, we can have for some positive-definite matrix C

$$(\delta_n(T_n-\theta))^T K_n(\delta_n(T_n-\theta)) + C \to S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$$

Then for some vector Δ such that $\delta_n \Delta^T K_n \Delta \delta_n = C$, there is

$$(\delta_n(T_n-\theta+\Delta))^T K_n(\delta_n(T_n-\theta+\Delta)) \to S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$$

So $T_n + \Delta$ is optimal estimator for τ_n, because

$$(\delta_n(T_n-\theta+\Delta)-\tau_n^T)^T K_n(\delta_n(T_n-\theta+\Delta)-\tau_n^T)) \to (K_{\theta,n}^{-1}S_{\theta,n}-\tau_n^T)^T K_{\theta,n}(K_{\theta,n}^{-1}S_{\theta,n}-\tau_n^T)).$$

But this contradicts with our definition of T_n.

Thus $(\delta_n(T_n-\theta))^T K_n(\delta_n(T_n-\theta))$ converges to $S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$. It implies K_n converges to $K_{\theta,n}$ in probability and $\delta_n(T_n-\theta)$ converges to $K_{\theta,n}^{-1}S_{\theta,n}$.

(ii) By Proposition 2.15, we know that clustering points K_θ of $K_{\theta,n}$ are invertible. Since $\delta_n(T_n-\theta)$ converges to $K_{\theta,n}^{-1}S_{\theta,n}$, the limit of $\delta_n(T_n-\theta)$ is $K_\theta^{-1}S_{\theta,n}$. The Gaussian variable $S_{\theta,n}$ is second moment bounded. So the term $\delta_n(T_n-\theta)$ is bounded in probability.

(iii) We know the DQM condition implies (2.2), thus the linear-quadratic equation (2.2) may coincide with S_n and K_n by (i). The log-likelihood process can be rewritten as a centered log-likelihood process $\Xi_n(\cdot)$ plus a shift item $b_n(\cdot)$:

$$\delta_n \Lambda_n(\theta,\vartheta)(x) = \overbrace{\frac{1}{n}\delta_n \sum_{i=1}^n \log\frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i) - \int \log\frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x)dP_0}^{\Xi_n(\theta)} + \underbrace{\int \log\frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x)dP_0}_{b_n(\theta)} + o_p(1).$$

Let $\delta_n = n^{-1/2}$. Given fixed $\lambda(\cdot)$ values in the constraint of equation (2.1), Theorem 2.7 says that $\log\frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i)$ in $\Xi_n(\eta)$ can

be replaced by a linear quadratic formulae w.r.t. τ_n, namely $\log \frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i)$ belongs to a smooth functional class $\mathcal{C}^2$. Therefore the process $\theta \mapsto \Xi_n(\theta)$ is an empirical process and $\Xi_n(\theta) \rightsquigarrow \Xi(\theta)$ by Donsker's Theorem, see van der Vaart (1998, Example 19.9) where $\Xi(\theta)$ is a Gaussian process. Note that $\Xi(\theta)$ has mean $\int \Xi(\theta) dP_0 = 0$ and covariance kernel $\mathbb{E}\Xi^2(\theta)$ under P_0. The log-normal property implies that $\mathbb{E}\exp[\Xi(\theta) + b(\theta)] = 1$ with the expectation taken under P_0 and $b(\theta) = \lim_{n\to\infty} b_n(\theta)$. Log normal property of $\exp\Xi(\cdot)$ gives $b(\theta) = -(1/2)\mathbb{E}\Xi^2(\theta)$. By Proposition 2.12 and equation (2.2), we can show that

$$\begin{aligned}\Xi_n(\theta) =& S_{\theta,n}\\ b(\theta) =& -\frac{1}{2}K_\theta,\end{aligned}$$

and when $\theta = \theta_0$

$$\begin{aligned}\Xi_n(\theta_0) =& \mathbb{E}\left[\frac{\partial m(X,\theta_0)}{\partial\theta}^T \left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\right]\delta_n\sum_i^n m_i(\theta_0)\\ b(\theta_0) =& -\frac{1}{2}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T \left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta^T}.\end{aligned}$$

□

Proof of Theorem 2.20

Discussion: The proof follows the strategies of van der Vaart (Proposition 8.6 1998) and Le Cam and Yang (Theorem 6.1 1990). The difficulty comes from the expectation conditional on the local parameter τ. Note that the measure $\mathcal{M}$ has not yet been specified. If one can in Bayesian fashion give a prior distribution on $\mathcal{M}$, then what we need to study is the posterior distributions given this "local prior measures". In fact, the δ_n-sparse condition already implies that for arbitrary priors, the corresponding posteriors concentrate on the small shrinking neighborhood of θ_0.

Proof. First look at the population log-likelihood ratio

$$\begin{aligned}\Lambda(\theta+\tau,\theta) = -\frac{1}{2}\Big[&(K_\theta^{-1}S_\theta - \tau)^T K_\theta (K_\theta^{-1}S_\theta - \tau)\\ &-(S_\theta^T K_\theta^{-1} S_\theta)\Big] + o_{\tilde{p}_\theta}(1)\end{aligned}$$

which implies that the term $(K_\theta^{-1}S_\theta - \tau)^T K_\theta (K_\theta^{-1}S_\theta - \tau))$ is χ^2 distributed. The quadratic form of a Gaussian variable ξ, $\xi^T\xi$,

can generate exactly the same distribution. As Theorem 2.7 shows that the approximation of Gaussian family is feasible. For any value of θ, there will be such a ξ_θ whose distribution is equivalent to $K_\theta^{-1}S_\theta - \tau$ and has the variance $K_\theta^{-1/2}$. Then we have the expression

$$\tau = K_\theta^{-1}S_\theta - \xi_\theta,$$

which shows that τ consists of two Gaussian variables $K_\theta^{-1}S_\theta$ and ξ_θ. Thus we are able to impose a Gaussian structure on the measure $\mathcal{M}$.

Now we can look at the expectation $\min(b, \mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n\tau])$ which is bounded by b. Since both "prior" and "posterior" concentrate around θ_0 and are Gaussian, the updating information only occurs for covariance matrix. Let τ be a Gaussian random variable centered at 0 with inverse covariance Γ. The conjugate property indicates the posterior of τ can be written as:

$$Z_n = \delta_n^{-1}(\tilde{T}_n - \theta_0) = (K_n + \Gamma)^{-1/2}K_n\delta_n^{-1}(T_n - \theta_0),$$

especially when $\Gamma = 0$, $Z_n = \delta_n^{-1}(T_n - \theta_0)$. By Anderson's Lemma[9], for bounded W, there is

$$\mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n\tau] \geq \mathbb{E}[W(Z_n)|\theta_0 + \delta_n\tau].$$

Since $K_n\delta_n^{-1}(T_n - \theta_0) \sim \mathcal{N}(0, I)$, the lower bound of $\mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n\tau]$ is

$$\mathbb{E}\left\{W\left[(K_n + \Gamma)^{-1/2} \times \mathcal{N}(0, I)\right]|K_n + \Gamma\right\}.$$

The measure of $\theta_0 + \delta_n\tau$ is replaced by $K_n + \Gamma$ because of the Gaussian property, namely the update of covariance matrix. Note that K_n and Γ are independent with $\mathcal{N}(0, I)$. With the condition $K_n \rightsquigarrow K_\theta$ in $\tilde{P}_\theta$ law, the limit becomes

$$\mathbb{E}\left\{W\left[(K_\theta + \Gamma)^{-1/2} \times \mathcal{N}(0, I)\right]\right\}.$$

When c is very large, the probability of normal prior $|\tau| > c$ is small enough thus

$$\liminf_n \sup_{|\tau| \leq c} \mathbb{E}\left\{W\left[(K_n + \Gamma)^{-1/2} \times \mathcal{N}(0, I)\right]\right\} \geq \mathbb{E}\left\{W\left[(K_\theta + \Gamma)^{-1/2} \times \mathcal{N}(0, I)\right]\right\} - \Delta$$

9 For a symmetric distribution, shifting an integral function of it to a new position will product higher expected value, see van der Vaart (1998, Lemma 8.5).

for Δ with a small enough Euclidean norm $|\Delta|$. Especially, when Γ go to zero or say the measure $\mathcal{M}$ degenerates to a point eventually, $Z_n = \delta_n^{-1}(T_n - \theta_0)$ obtains the lower bound $\mathbb{E}[W(K_\theta^{-1/2}) \times \mathcal{N}(0, I)]$. If $W = 1$ and $K_\theta = K$, by Theorem 2.18(iii) we achieve the efficient bound of semi-parametric estimators. □

OTHER TECHNICAL DETAILS

Proof of Lemma 2.5

Proof. Note that for every fixed n and θ, $\log n\tilde{p}(\theta, X_i)$ is an independent random variable. Condition 2.4 is often used to deduce the upper bound of likelihood ratio function. The purpose of imposing this assumption is to get a bounded variance of $\log n\tilde{p}(\theta, X_i)$. Compactness implies, for a given ϵ, the existence of a number $a_n < \infty$ such that $\Pr\{|\Lambda_n(\theta)| > a_n/2\} < \epsilon/2$. Let $\Lambda_{n,k}(\theta) = n^{-1}\sum_i^k \log n\tilde{p}(\theta, X_i)$. Levy's inequality says that for any $k \leq n$ and $a_n \geq 0$, it holds that

$$\Pr\left[\sup_k |\Lambda_{n,k}(\theta)| \geq \frac{a_n}{2}\right] \leq 2\Pr\left[|\Lambda_n(\theta)| \geq \frac{a_n}{2}\right] < \epsilon.$$

Then by taking differences, $n^{-1}\log n\tilde{p}_k(\theta) = \Lambda_{n,k}(\theta) - \Lambda_{n,k-1}(\theta)$, we have

$$\Pr\left[\sup_k \frac{|\log n\tilde{p}_k|}{n} \geq a_n\right] \leq \Pr\left[\sup_k \frac{|\Lambda_{n,k}(\theta) - \Lambda_{n,k-1}(\theta)|}{n} \geq a_n\right]$$
$$\leq \Pr\left[\sup_k \frac{2|\Lambda_{n,k}(\theta)|}{n} \geq a_n\right] < 2\Pr\left[|\Lambda_n(\theta)| \geq \frac{a_n}{2}\right] < \epsilon.$$

It indicates that for every $a_n > 0$ the quantity

$$\sup_k \Pr\left[n^{-1}\log n\tilde{p}(\theta, X_k)\right] \geq a_n$$

tends to zero as n goes to infinity. Thus the random variable $\log n\tilde{p}(\theta, X_i)$ has bounded variance. □

Poisson Approximation for Arbitrary Infinitely Divisible Families

Let $\phi(t)$ and $\phi_n(t)$ be the characteristic functions of distributions in $\mathcal{E}$ and $\mathcal{E}_n$. By the infinitely divisible property, $\phi(t) = [\phi_n(t)]^n$

or $\phi_n(t) = [\phi(t)]^{1/n}$. Two characteristic functions have the following relation:

$$n(\phi_n(t) - 1) = n(\sqrt[n]{\phi(t)} - 1) = n\left(e^{\frac{1}{n}\log\phi(t)} - 1\right)$$
$$= n\left(1 + \frac{1}{n}\log\phi(t) + o(\frac{1}{n}) - 1\right) \to \log\phi(t),$$

or say $\exp(n(\phi_n(t) - 1)) \to \phi(t)$. The concrete construction of characteristic function in $\mathcal{E}_{\theta,n}$ depends on the discrete Fourier transform of $\Lambda(X,\theta)$ on j segments e.g. $\inf\Lambda(X) < c_1 < c_2 < \cdots < c_j < \sup\Lambda(X)$ which implies that

$$\lim_{j\to\infty}\sum_{k=1}^{j} a_k(i)e^{itc_k} = \int e^{it\Lambda(X)}dF_n = \phi_n(t),$$

where $a_n(k) = n(F_n(c_k) - F_n(c_{k-1}))$ is the Fourier coefficient[10] and F_n is the measure for $\Lambda_n(\theta)$. Combined with the expression above, one can see that a characteristic function of finite many number of Poisson measures (compound Poisson measures) approximates $\phi(t)$:

$$\exp\sum_{i=1}^{j}(na_i)\left(e^{it\Lambda(x_i,\theta)} - 1\right) \to \phi(t) \tag{2.17}$$

where $j \to \infty$ and $\{na_i\}_{i=1,\ldots,j}$ converges to a measure. To see the argument of (2.17), let $V(\cdot)$ be a Poisson process (a random measure) with Poisson parameter γ such that $\mathbb{E}V(\mathcal{A}) = \gamma(\mathcal{A})$ for a set $\mathcal{A}$. For any function v in infinite divisible family, the characteristic function of v is $\phi(t) = \exp\{\int(e^{itv} - 1)d\gamma\}$.

The approximation can be viewed as constructing a new family which approximately equals the infinite divisible $\mathcal{E}_\theta$. Firstly select a Poisson variable ν (again a random measure) such that $\mathbb{E}\nu(\Lambda(X)) = 1$ for any log-likelihood ratio $\Lambda(X)$ and then carry out n-draws from the direct product $\otimes_{i=1,\ldots,\nu}\mathcal{E}_{\theta,i}$, ν copies $\mathcal{E}_{\theta,i}$. The result is called a poissonized family.

Proof of Proposition 2.12

The proof is based on Taylor expansions. Note that

$$m(x,\theta_0 + \delta_n\tau) = m(x,\theta_0) + \delta_n\frac{\partial m(x,\theta_0)}{\partial\theta^T}\tau + o_p(\delta_n^2). \tag{2.18}$$

10 The Stieltjes sum, a discrete version of stochastic integral.

Let $\theta \in \{\theta || \theta - \theta_0| \leq |\tau|\delta_n\}$, $|\tau|$ is a vector with elements equal to their absolute values. The result

$$\lambda_n(\theta) = \left(\sum_{i=1}^{n}[m_i(\theta)m_i(\theta)^T]/n\right)^{-1}\sum_{i=1}^{n}m_i(\theta)/n + o_p(n^{-1/2})$$

holds uniformly for θ in a neighborhood of θ_0, see the proofs in Qin and Lawless (1994, Lemma 1) or Owen (2001, Theorem 2.2). For the empirical log-likelihood at θ, by noting that $\lambda_n^T m_i$ is close to zero and using a second order approximation for $\log(1+\lambda_n^T m_i)$,we obtain:

$$\sum_{i=1}^{n}\log\tilde{p}_\theta = \sum_{i=1}^{n}\left[\lambda_n(\theta)^T m_i(\theta) - \frac{1}{2}\left(\lambda_n(\theta)^T m_i(\theta)m_i(\theta)^T\lambda_n(\theta)\right)\right] - n\log n + o_p(1).$$

The remainder term is based on bounding $\sum_{i=1}^{n}\left(\lambda_n^T m_i\right)^3$ for which Owen (1990) showed in Lemma 3 that it is of order $o_p(1)$. Note that his γ_i is our $\lambda_n^T m_i(\theta)$. Note that

$$\lambda_n(\theta)^T m_i(\theta) = \left(\sum_{i=1}^{n}\frac{m_i(\theta)}{n}\right)^T\left[\sum_{i=1}^{n}\frac{1}{n}\left(m_i(\theta)m_i(\theta)^T\right)\right]^{-1}m_i(\theta)$$

and after summation equals the squared term:

$$\sum_{i=1}^{n}\lambda(\theta)^T m_i(\theta)m_i(\theta)^T\lambda_n(\theta) = \left(\sum_{i=1}^{n}\frac{m_i(\theta)}{n}\right)^T\left[\sum_{i=1}^{n}\frac{1}{n}\left(m_i(\theta)m_i(\theta)^T\right)\right]^{-1}\left(\sum_{i=1}^{n}\frac{m_i(\theta)}{n}\right).$$

So adding these two terms we obtain:

$$\sum_{i=1}^{n}\log\tilde{p}_\theta = \frac{1}{2}\left(\sum_{i=1}^{n}\frac{m_i(\theta)}{n}\right)^T\left[\sum_{i=1}^{n}\frac{1}{n}\left(m_i(\theta)m_i(\theta)^T\right)\right]^{-1} \times\left(\sum_{i=1}^{n}\frac{m_i(\theta)}{n}\right) - n\log n + o_p(1).$$

It implies:

$$2\sum_{i=1}^{n}\log\frac{\tilde{p}_{\theta_0+\delta_n\tau}}{\tilde{p}_{\theta_0}}(x_i)=\left(\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0+\delta_n\tau)\right)^T\times$$
$$\left(\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0+\delta_n\tau)m_i(\theta_0+\delta_n\tau)^T]\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0+\delta_n\tau)-$$
$$\left(\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)\right)^T\left(\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0)m_i(\theta_0)^T]\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0)+o_p(1).$$

It follows from the approximation of λ above. Using equation (2.18) we can further simplify the terms involving $\theta+\delta_n\tau$. We obtain for the middle term:

$$\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0+\delta_n\tau)m_i(\theta_0+\delta_n\tau)^T]=\frac{1}{n}\sum_{i=1}^{n}\Big[[m_i(\theta_0)m_i(\theta_0)^T]+$$
$$\delta_n\tau\left(\frac{\partial m_i(\theta_0)}{\partial\theta^T}\right)^T m_i(\theta_0)+\frac{(\delta_n\tau)^2}{4}\left(\frac{\partial m_i(\theta_0)}{\partial\theta^T}\right)^T\frac{\partial m_i(\theta_0)}{\partial\theta^T}+o_p(\delta_n^3)\Big]$$
$$=\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0)m_i(\theta_0)^T]+\frac{1}{n}\delta_n O_p(n^{1/2})+o_p(\delta_n^2)+o_p(\delta_n^3).$$

With the big bracket becoming

$$n\left[\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\delta_n\frac{\partial m_i(\theta_0)}{\partial\theta^T}\tau\right]^T\left(\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0)m_i(\theta_0)^T]\right)^{-1}$$
$$\times\left[\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\delta_n\frac{\partial m_i(\theta_0)}{\partial\theta^T}\tau\right]$$
$$=n\left[\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\delta_n\mathbb{E}\frac{\partial m_i(\theta_0)}{\partial\theta^T}\tau+\delta_n O(n^{-1/2}(\log\log n)^{1/2})\right]^T$$
$$\times\left(\mathbb{E}\left(m(x,\theta_0)m(x,\theta_0)^T\right)\right)^{-1}$$
$$\times\left[\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\delta_n\mathbb{E}\frac{\partial m_i(\theta_0)}{\partial\theta^T}\tau+\delta_n O(n^{-1/2}(\log\log n)^{1/2})\right]$$
$$=2\delta_n\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta^T}\tau\left(\mathbb{E}\left(m(x,\theta_0)m(x,\theta_0)^T\right)\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+$$
$$\delta_n^2\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta^T}\tau\left(\mathbb{E}\left(m(x,\theta_0)m(x,\theta_0)^T\right)\right)^{-1}\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}\tau+$$
$$\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)\left(\mathbb{E}\left(m(x,\theta_0)m(x,\theta_0)^T\right)\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+o_p(\delta_n^3)$$

where $O(n^{-1/2}(\log\log n)^{1/2})$ is used to bound the difference of the sample average and the expectation of a random vector. Thus the local EL is

$$2\sum_{i=1}^{n}\log\frac{\tilde{p}_{\theta_0+\delta_n\tau_n}}{\tilde{p}_{\theta_0}}(x_i)=\delta_n\tau_n^T A_1++\frac{1}{2}\delta_n^2\tau_n^T A_2\tau_n^T+o_p(1)$$

where A_1 is $\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0)$ and A_2 is $\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta^T}$.

Note that $O(n^{-1/2}(\log\log n)^{1/2})\times\delta_n\sum_{i=1}^{n}m_i(\theta_0)=o_p(1)$ and

$$\lim_{n\to\infty}A_n\cdot\sum_{i=1}^{n}[m_i(\theta_0+\delta_n\tau)-m_i(\theta_0)]/n=o_p(1)$$

with $A_n=\sum_{i=1}^{n}m_i(\theta_0)\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}$ by the continuity of $m_i(\theta)$.

Proof of Proposition 2.15

To prove K_θ is invertible, we will prove K_θ is almost surely positive definite. Le Cam's first lemma implies that

$$\mathbb{E}\exp\left[\tau^T S_\theta-\frac{1}{2}\tau^T K_\theta\tau\right]=1. \tag{2.19}$$

Because (2.19) holds for all τ, we can use a symmetrized method to simplify (2.19). For a given value τ and $-\tau$, we have

$$\mathbb{E}\left\{\exp\left[\tau^T S_\theta-\frac{1}{2}\tau^T K_\theta\tau\right]+\exp\left[-\tau^T S_\theta-\frac{1}{2}\tau^T K_\theta\tau\right]\right\}=2.$$

By $\cosh\tau^T S_\theta=(\exp\tau^T S_\theta+\exp(-\tau^T S_\theta))/2$, we have

$$\mathbb{E}[(\cosh\tau^T S_\theta)\exp(-\tau^T K_\theta\tau/2)]=1. \tag{2.20}$$

Assume there is some τ such that $\tau^T K_\theta\tau$ is negative, then

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_{\{\tau^T K_\theta\tau>0\}}(\cosh\tau^T S_\theta)\exp(-\tau^T K_\theta\tau/2)\right] \\ \leq\mathbb{E}\left[(\cosh\tau^T S_\theta)\exp(-\tau^T K_\theta\tau/2)\right]=1\end{aligned} \tag{2.21}$$

where $\mathbb{I}_{\{\cdot\}}$ is an indicator function. However, since

$$\exp(-\tau^T K_\theta\tau/2)>1$$

when $\tau^T K_\theta \tau$ is negative and $(\cosh \tau^T S_\theta) > 1$,

$$\underbrace{\mathbb{E}\left[\mathbb{I}_{\{\tau^T K_\theta \tau > 0\}} (\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)\right]}_{>0} +$$
$$\underbrace{\mathbb{E}\left[\mathbb{I}_{\{\tau^T K_\theta \tau \leq 0\}} (\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)\right]}_{\geq 1}$$
$$= \mathbb{E}\left[(\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)\right] > 1$$

we have a contradiction with equation (2.20) unless the set $\{\tau^T K_\theta \tau \leq 0\}$ is empty. Therefore, K_θ is positive definite and hence invertible.

3

ROBUST DECISIONS IN DYNAMIC CHOICE MODELS

3.1 INTRODUCTION

3.1.1 *Relevant Literature*

Dynamic discrete choice models build on the assumption of rational behavior. The rationality appears in the expectation form and utility maximization when an agent makes his/her decisions dynamically. This model exploits the intrinsic evolution structure and captures the endogenous effect of agents' actions. Self-expectations, however, are difficult to be characterized completely in practice because of the uncertainty of expectations. If an agent has an incomplete information set, his expectation based on such a set will be incomplete as well. The incompleteness will induce stochastic characteristics of his/her expectation. On the other hand, if this agent is aware of the incompleteness and will make a decision accounting for this effect rather than insisting on the maximal principle of utility, he/she may follow a strategy called robust decision. Such an agent is often referred to as a boundedly rational agents. Incomplete information has been addressed in many dynamic discrete choice papers (Aguirregabiria and Mira, 2002, 2007; Bajari et al., 2007) while boundedly rational agent and robust decision are often addressed in the dynamic control literature in especially for macroeconomic models (Hansen and Sargent, 2007). This chapter will consider the robust decision problem in the framework of dynamic discrete choice models[1]. We will identify robustness issues in dynamic discrete choice models and then develop a tractable prediction approach under a more flexible assumption on consumer's optimal behavior and a more flexible model specification.

1 This setup is often referred to as a (mathematical) dynamic game in control related subjects.

Estimation for single agent dynamic discrete choice models was initiated by Wolpin (1984), Pakes (1986) and Rust (1987). Estimating the parameters of these structural models requires solving an optimization problem with a nested dynamic programming problem. The computation is non-trivial for both the likelihood function evaluation (outer-loop) and the fixed point iteration (inner-loop). The difficulties mainly come from two sources, the exponential growth of the state space and the existence of multiple equilibria (Aguirregabiria and Mira, 2010).

A common approach for breaking down the complexity growth is to specify the expectation. When the expectation is evaluated with respect to a specified distribution, like the private information, the expectation will be computed in terms of analytic algebra function instead of numerical integration. In this way, randomization can break down the curse of dimensionality caused by the multivariate integration in the dynamic programming. Keane and Wolpin (1994) suggest solving the dynamic programming problem by Monte Carlo integration and interpolation. Later, Rust (1997) proves that a certain random Bellman operator with particular parametric structure can break the curse of dimensionality. This method works well for unique equilibrium models. But if there are multiple equilibria, the expectation will depend on the equilibrium that gives locally optimal utility. Estimation for multiple equilibrium may need to compute the full set of equilibria in order to select an optimal expression. Usually, this is a two-step procedure. In the first step, with some estimated or approximated components, one will establish an objective function of the expected utility and then use this function in the second step. In the second step, one will solve the optimization problem of this objective function with respect to the structural parameters. For this kind of two-step approach (Aguirregabiria and Mira, 2007; Pesendorfer and Schimidt-Dengler, 2003; Bajari et al., 2007; Pakes et al., 2008), the final estimation result depends heavily on the constructed objective function in the first step. Two issues may complicate the inferential result. 1. If the expectation is accompanied by parametric assumptions, then the true model should not be far from the parametric model, otherwise the bias in the first step will be accumulated afterward. 2. If the expected function involves non/semi-parametric components, the associated curse of dimensionality problem will occur.

For inferential or prediction purposes, the most important criterion is minimum mean square error (MSE). But two essential

reasons make it difficult to incorporate the MSE criterion in the two-step procedure. The *first* reason is the concern of the misspecification problems at the second step of the estimate. The second step objective functions is either a pseudo likelihood or an approximating moment constraint. They are not exact parametric functions or constraints but approximating ones. Fernandez-Villaverde et al. (2006) show that the higher order bias from the first step dynamic programming solution will give a first-order biased-effect on the likelihood function in the second step. Therefore, for the concern of consistency, achieving an exact fit in the first step is more important than minimizing MSE. But this may lead to a potential issue of over-fitting.

The *second* reason is the small sample problem. In agent-based modeling, data resources are often very limited and it may be unwise to solely rely on pure statistical inferential decisions to judge the correctness of the models. "The statistical mentality that 'all structural models are rejected, therefore none of them are any good' does a disservice and contributes to a radicalization of some members of the profession"-Rust (2008). Therefore, it seems that statistical prediction or inference should not be ranked in the first place for structural estimation and modeling.

In sum, given the complicated nesting dynamic programming structure, people often prefer to "fit" a given model (non)- parametrically rather than to obtain a (non)-parametric function with prediction ability. However, in practice, there is no universally acceptable true model. It means that in principle even an unbiased estimator for a hypothetical true model does not necessarily give us a good estimate for the underlying model. In this chapter, we give up maintaining the aim of consistent approximation. We look for an alternative method which allows for little bias but will yield more flexibility in return. In dynamic models, this purpose can be explained in terms of bounded rationality, to be more specific, *the worst case* of Bellman's optimality principle, where the numerically approximated solution is a distance ϵ away from the fixed point value function.

3.1.2 *Contributions*

Our approach is motivated by the constrained maximum likelihood (Rust (1987, Nested Fixed Points - NXFP), Aguirregabiria and Mira (NPML 2007, -Nested Pseudo Maximum Likelihood)) and the constrained optimization (Su and Judd (2011, Mathematical Programming with Equilibrium Constraints-MPEC)). Our

approach does not rely on strong parametric assumptions on the specification of the incomplete information and addresses the robustness concern in the model. We also provide some techniques to solve the problems that accompany with the appearance of boundedly rational agents.

The first contribution of this paper is proposing an alternative approach to estimate a dynamic discrete choice model with boundedly rational agents. This is a game-theoretic approach. The departure from the rationality induces the so called *ϵ-equilibrium* or *robust equilibrium* (Radner, 1981, 1986). In ϵ-equilibrium, each player is satisfied to get within a small distance of his best response function. In other words, given the ϵ-equilibrium strategy profile, the player will not gain more than ϵ expected payoff by altering his strategy. In our context, ϵ-equilibrium will be converted to an ϵ-variation of a fixed point condition. Suppose one fixed point condition is equivalent to one equilibrium model, then ϵ-variation will induce a set of multiple equilibria where dynamic programming is difficult to implement. This chapter will consider how to select a single representative model from the set of equilibria. The selected model is able to explain those alternative models that satisfy the ϵ-fixed point condition.

The second contribution is to exploit the asymptotic properties of the solution of the dynamic programming problem with boundedly rational agents. The asymptotic behavior is not standard since the solution is inconsistent within a small neighborhood but the solution will not get out of this neighborhood in probability. This non-trivial consistency result forms the quantitative intuition of boundedly rational agent in the ϵ-equilibrium.

The third contribution is to apply a localization method to handle the moment constraints with non-smooth population functions. Due to the "bias" or "inconsistency" from the first step, one should expect that the population function is non-smooth even at the true parameter value. Thus the usual sample smoothing methods may not qualify for the handling of a non-differentiable population moment function. The localization method transfers this global infeasible problem about structural parameters into a feasible problem about local parameters.

3.2 STOCHASTIC DYNAMICS

The agent's action is taken in a given period and may affect both current and future profits. The evolution of an agent is modeled via the dynamic programming problem with an infinite discrete time horizon $t = 1,2,\ldots,\infty$. Dynamic programming is able to exploit the intrinsic evolution process and construct a controllable scheme. In period t the state variable is denoted as s_t. The vector of all states for all $N-$agents is commonly observed as $\mathbf{s}_t = (s_{1t},\ldots,s_{Nt}) \in \mathcal{S}$, where $\mathcal{S} \subset \mathbb{R}^N$ is the entire state space. Given the current common information $\mathbf{s}_t$, an agent will choose his/her action(s) $\mathbf{a}_t = (a_{1t},\ldots,a_{Nt}) \in \mathcal{A}$ in each period to adjust his/her expected utility. The action space $\mathcal{A}$ includes only a finite number of actions. The expectation is driven by private shocks. These shocks, collected in $\Xi_t = (\xi_{1t},\ldots,\xi_{Nt})$, are unobservable information for the analysts and are observed only by the agents. They are treated as an N-dimensional independent *random vector*. The distribution of ξ_{it} is denoted $G(\cdot)$.

For simplicity, the current action and state are denoted by $(\mathbf{a},\mathbf{s})$ and $(\mathbf{a}',\mathbf{s}')$ denote the next period action and state. If the evolution pattern is first order Markovian, then given an action $\mathbf{a}$, the state $\mathbf{s}$ will transit to $\mathbf{s}'$ with a transition density $\mathbf{p}(\mathbf{s}'|\mathbf{s},\mathbf{a})$. With this transition density, we can construct the transition density of the full Markov chain from $\mathbf{s}_1$ to $\mathbf{s}_T$ as:

$$\mathbf{p} = (p(\mathbf{s}_2|\mathbf{s}_1,\mathbf{a}_1),\ldots,p(\mathbf{s}_{T+1}|\mathbf{s}_T,\mathbf{a}_T)).$$

We assume the Markov transition density is *homogeneous*, meaning that given action $\mathbf{a}$, states $\mathbf{s}$ and $\mathbf{s}'$, the transition density $p(\mathbf{s}'|\mathbf{s},\mathbf{a})$ is homogeneous in time.

Let $\Pi(\mathbf{a}_t,\mathbf{s}_t,\Xi_t)$ be the N-vector profit for all agents in period t. Because we consider discrete choice models, we assume that $\Pi(\mathbf{a}_t,\mathbf{s}_t,\Xi_t)$ contains independent elements. Then by implication we can consider operations of individual elements independent of each other. Let $\pi(\mathbf{a}_t,\mathbf{s}_t,\xi_t)$ be a single representative of an element in $\Pi(\mathbf{a}_t,\mathbf{s}_t,\Xi_t)$. A *rational* agent who makes his decisions by maximizing the expected current and discounted future profits,

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty}\beta^{\tau-t}\pi(\mathbf{a}_\tau,\mathbf{s}_\tau,\xi_\tau)|\mathbf{s}_t\right], \tag{3.1}$$

where $\beta \in (0,1)$ is the discount factor. The expectation is taken w.r.t. the distribution constructed by Markov chain $\mathbf{p}$ and $G(\cdot)$. The shock ξ_t depends on the state $\mathbf{s}_t$ therefore $G(\xi)$ is a condi-

tional distribution given $\mathbf{s}$. The primitives of the model include the discount factor β and the transition densities $p(\mathbf{s}'|\mathbf{s},\mathbf{a})$.[2]

If the distribution $G(\cdot)$ can be associated with some parameter θ, then it is also a primitive factor. The parametric expected profit function is:

$$\mathbb{E}_\theta \left[\sum_{\tau=t}^{\infty} \beta^{\tau-t} \pi(\mathbf{a}_\tau, \mathbf{s}_\tau, \xi_\tau) | \mathbf{s}_t \right], \tag{3.2}$$

where θ is the unknown parameter. The difference between (3.1) and (3.2) is whether or not the value function can be parametrized by θ.

Markov Perfect Equilibria (MPE) i.e. equilibria where the optimal actions of agents will depend only on the state at time t, are frequently used in various dynamic discrete choice models. By MPE condition, each action depends on the current state and its current private shock. Let $\sigma(\mathbf{s},\xi)$ be a Markov policy, or strategy function or decision rule that maps the current state and shock onto an action $\sigma(\cdot,\cdot) : \mathcal{S} \times \mathbb{R}^N \to \mathcal{A}$. The Bellman equation for the dynamic model (3.1) is

$$V(\mathbf{s},\xi) = \max_{\mathbf{a}\in\mathcal{A}} \left\{ \pi(\mathbf{a},\mathbf{s},\xi) + \beta \int \int V(\mathbf{s}',\xi') dP(\mathbf{s}'|\mathbf{s},\mathbf{a}) dG(\xi') \right\}, \tag{3.3}$$

where $V(\mathbf{s},\xi)$ is the value function. Since the expression may involve discrete- or real-valued ξ, the notation of the underlying value function is set up in terms of integrals rather than sums. The associated policy function or decision function is:

$$\sigma(\mathbf{s},\xi) := \arg\max_{\mathbf{a}\in\mathcal{A}} \left\{ \pi(\mathbf{a},\mathbf{s},\xi) + \beta \int \int V(\mathbf{s}',\xi') dP(\mathbf{s}'|\mathbf{s};\mathbf{a}) dG(\xi') \right\}.$$

The above dynamic programming problem suffers from a serious curse of dimensionality if the dimension of ξ is high. In this case, the numerical integral w.r.t. $G(\xi)$ is difficult to evaluate and thus the expectation is poorly approximated. By Bellman's principle of optimality, Rust (1987) suggests to substitute the

2 Publicly known transition densities are not standard in the literature, but it is a compromise by relaxing the specification of probability $G(\cdot)$. In the literature, $G(\cdot)$ is often assumed being multinomial logit which leads to a conditional choice probability $F(\mathbf{a}|\mathbf{s})$ which is also a multinomial logit see Hotz and Miller (1993, Lemma 3.1). This $F(\mathbf{a}|\mathbf{s})$ in turn leads to $p(\cdot|\cdot,\mathbf{a})$. Both are extremely sensitive to the choice of $G(\cdot)$. In the boundedly rational case, G is not consistently estimated and leads to a very unrealistic $p(\cdot|\cdot,\mathbf{a})$.

decision rule into equation (3.3) and to integrate out the private shocks, then:

$$V(\mathbf{s},\sigma(\mathbf{s})) = \mathbb{E}_{\xi}\left[\pi(\sigma(\mathbf{s}),\mathbf{s},\xi) + \beta \int V(\mathbf{s}',\sigma(\mathbf{s}'))dP(\mathbf{s}'|\mathbf{s},\sigma(\mathbf{s}))\right]. \tag{3.4}$$

The function $V(\mathbf{s},\sigma(\mathbf{s}))$ illustrates that the expected profit at the beginning of a period satisfies Bellman's optimality principle. Equation (3.4) is called the *integrated Bellman equation* and $V(\mathbf{s},\sigma(\mathbf{s}))$ is called the *ex ante value function* (Bajari et al., 2007). Because the ex ante value function $V(\mathbf{s},\sigma(\mathbf{s}))$, or $V(\mathbf{s})$ in short, is not a function of ξ but of $\mathbf{s}$ only, the solution of the problem (3.4) will be a function of $\mathbf{s}$ only. This is a way of dealing with the unknown distribution of ξ.

We assume the following standard conditions:

Condition 3.1. Conditional Independence (CI): Given $\mathbf{s}_t$, the random variables ξ_{it} and ξ_{js} are independent for any $s > t$ and $i \neq j$.

Condition 3.2. Additive Separability (AS): A private shock appears additively in the profit function. $\pi(\mathbf{a},\mathbf{s},\xi) = \bar{\pi}(\mathbf{a},\mathbf{s}) + \xi$.

3.3 ROBUST DECISION

In this section, we will use an ϵ-approximating value function $\hat{V}$ to solve dynamic programming problem dealing with the robustness concern. This problem is developed from equation (3.4). Also we will give a policy function iteration method for this problem. All the results in this section focus on the first step of a 2-step estimation procedure. The second step will be discussed in the next section.

3.3.1 *The Kernel-based Constrained Optimization*

The motivation of the constrained optimization procedure is to put the thorny part of the problem into the constraints and set up an objective function of controllable coefficients of the constraints. We will consider the ϵ-variation fixed point condition defined below as a single inequality constraint. The set of MPEs will satisfy this constraint. Then we will use a local basis function to approximate the unknown ex ante value function.

Given the MPE assumption, we use a vector of *Radial Basis Functions (RBF)* $\Phi(\cdot)$ to approximate the ex ante value function

in (3.4) since we assume a permutation structure of the elements of $\mathbf{s}$ so that the $V(\cdot)$ depends on the length of $\mathbf{s}$, not the ordering of its elements:

$$\hat{V}(\mathbf{s}) = \rho^T \Phi(\mathbf{s}). \tag{3.5}$$

The standard inner product for a functional space of Φ defines a kernel function as follows. Select T points $(\mathbf{s}_1, \ldots, \mathbf{s}_T)$, and calculate the kernel:

$$K_{ij} := k(\mathbf{s}_i, \mathbf{s}_j) := \langle \Phi(\mathbf{s}_i), \Phi(\mathbf{s}_j) \rangle.$$

This kernel is a positive definite kernel (Gram) matrix. By letting the dimension of ρ and Φ go to infinity, we can approximate the function $V(\cdot)$ arbitrarily well. But in practice, a finite number of basis functions has to be chosen. Similar approximations can also be found in Bajari et al. (2007) who assume a profit function and a value function that are linear in the unknown parameters in order to simplify the computation. However, equation (3.5) has a different meaning. The coefficient ρ in (3.5) does not necessarily correspond to a structural parameter as in Bajari et al. (2007) at all, but the coefficient ρ itself can be used to describe how complex[3] the approximation is. The norm of ρ, $\|\rho\|$, is a regularizer for the complexity of the solution $\hat{V}$. Thus we are trying to *learn* the unknown function $\hat{V}(\cdot)$ and we can adjust the *learning* ability of function $\rho^T \Phi(\mathbf{s})$ by tuning ρ. We will include $\|\rho\|^2/2$ in the objective function in order to penalize the complexity of potential approximations.

Condition 3.3. (ϵ-equilibrium) The ϵ-equilibrium set includes Markov Perfect Equilibria and can be represented in terms of $\hat{V}(\mathbf{s})$ in equation (3.5).

Condition 3.3 is to convert a dynamic programming problem with robust decisions to an expression in ϵ-equilibria. Here ϵ is a *deterministic* value determined by prior deliberation. The ϵ-equilibria imply that agents are indifferent or cannot distinguish within a set that covers their optimal action functions, for example any V:

$$V(\mathbf{s}, \sigma(\mathbf{s})) - \frac{1}{2}\epsilon \leq V \leq V(\mathbf{s}, \sigma(\mathbf{s})) + \frac{1}{2}\epsilon,$$

would be a feasible solution for equation (3.4). Therefore, by Condition 3.3 the profit $\mathbb{E}_\xi \pi(\sigma(\mathbf{s}, \xi), \mathbf{s}, \xi)$ based on the decision

3 The complexity refers the number of implementing basis functions i.e. see Vapnik (1998).

$\sigma(\mathbf{s})$ for fully rational agents is in fact acceptable for boundedly rational agents. Mathematically, we relax the rational expectation condition based on $\mathbb{E}_{\xi}[\cdot]$ and require only that the solutions lie within an ϵ-distance of the exact rational expectation solution. We will refer to the resulting solution set as the ϵ-tube defined as follows:

$$\left| V(\mathbf{s};\sigma(\mathbf{s})) - \pi(\sigma(\mathbf{s},\xi),\mathbf{s},\xi) - \beta \int V(\mathbf{s}',\sigma(\mathbf{s}))dP(\mathbf{s}'|\mathbf{s},\sigma(\mathbf{s})) \right| \leq \epsilon. \tag{3.6}$$

Obviously, when ϵ goes to zero, equation (3.6) corresponds to the exact rational expectation solution in (3.4). The $\mathbb{E}_{\xi}[\cdot]$ in (3.4) corresponds to the exact fixed point and the principle of optimal decision in Bellman's equations, however, due to the uncertainty of the functional form of $G(\xi)$, it is quite common for an agent to realize that his/her profits will differ from the expected values and make decisions allowing for uncertainty. Mathematically, the difference between (3.6) and the integrated Bellman equation is that (3.6) accommodates the stochastic effects using a deterministic inequality. Relations between deterministic deviation ϵ and uncertainty can be traced back to the work of Debreu (1972) and Simon (1957).

Therefore, instead of using the equality function (3.4), we establish an inequality where the solution of the integrated Bellman equation (3.4) is included. Equation (3.6) can be interpreted as a so-called ϵ-insensitive loss function or a pseudo-metric by Vapnik (1998) such that:

$$|\hat{V} - \Gamma\hat{V}|_{\epsilon} = \max\left\{0, |\hat{V} - \Gamma\hat{V}| - \epsilon\right\}, \tag{3.7}$$

where Γ is an operator on $V(\mathbf{s}) \in \mathcal{V}$ such that $\Gamma : \mathcal{V} \mapsto \mathcal{V}$ and defined as

$$\Gamma V(\mathbf{s}) := \pi(\sigma(\mathbf{s}),\mathbf{s},\xi) + \beta \int V(\mathbf{s}',\sigma(\mathbf{s}))dP(\mathbf{s}'|\mathbf{s};\sigma(\mathbf{s})).$$

The stochastic feature of $\pi(\sigma(\mathbf{s}),\mathbf{s},\xi)$ implies that Γ is a *random* Bellman's operator (Rust, 1997) so Γ still depends on ξ. For the space $\mathcal{V}$ which in general is a sub-space of a Banach space, we will only consider a special discretized version such that Γ shares similar properties with the *deterministic* Bellman's operator. This approach has been used previously e.g. in (Rust et al., 2002) where $\pi(\sigma(\mathbf{s},\xi),\mathbf{s},\xi)$ is discretized into finite states as non-deterministic lattices.

In this chapter, we let equation (3.6) suggest a different approach. The random operator is limited within a deterministic

ϵ tube. All the randomness of ΓV will be treated equivalent to the single deterministic counterpart inside the ϵ-tube[4]. We also need a condition for this random operator:

Condition 3.4. (E) Ergodic distribution $G(\xi)$: Γ is a measure-preserving mapping on the measurable space $\mathcal{V}$.

By Condition 3.3, the solution set of (3.6) will be represented in terms of $\hat{V}$. We call the equation (3.6) ϵ-tube which is a ϵ-*fixed point condition* in the optimization procedure. The primal problem of minimizing the complexity of the approximating value function subject to the ϵ-fixed point condition of (3.6) is:

$$\begin{aligned}
&\min_{\rho,\zeta,\zeta^*} \quad \frac{1}{2}\|\rho\|^2 + \sum_{t=1}^{T}\left[\frac{C}{|\mathcal{S}|}(\zeta_t + \zeta_t^*)\right], \\
&\quad s.t. \quad \rho^T\Phi(\mathbf{s}_t) - \Gamma\hat{V}(\mathbf{s}_t) \le \epsilon + \zeta_t,\ -\rho^T\Phi(\mathbf{s}_t) + \Gamma\hat{V}(\mathbf{s}_t) \le \epsilon + \zeta_t^*, \\
&\qquad \zeta_t \ge 0,\ \zeta_t^* \ge 0 \text{ for all } 0 \le t \le T, \\
&\text{with } \Gamma\hat{V}(\mathbf{s}_t) := \pi(\mathbf{a}_t, \mathbf{s}_t, \xi_t) + \beta \sum_{\mathbf{s}_{t+1}\in\mathcal{S}} (\rho^T\Phi(\mathbf{s}_{t+1}))\mathbf{p}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}),
\end{aligned} \tag{3.8}$$

ζ and ζ^* are slack variables, $|\mathcal{S}|$ is the number of states and C is some constant. The ϵ-fixed point condition sets up a 2ϵ-wide "tube" for the fitting curve. The slack variables ζ and ζ^* play roles as soft margins for that tube, like dual variables in the L^1-penalty[5].

Remark 3.5. If ξ is an additive term in $\pi(\sigma(\mathbf{s},\xi),\mathbf{s},\xi)$ in (3.6), equation (3.6) and equation (3.4) associated with condition 3.2 imply that $|\mu| \le 2\epsilon$ where μ is the expectation of ξ, we have:

$$\left|\rho^T\Phi(\mathbf{s}_t) - \left[\bar{\pi}(\sigma_t, \mathbf{s}_t) + \mu + \beta \sum_{\mathbf{s}_{t+1}\in\mathcal{S}} (\rho^T\Phi(\mathbf{s}_{t+1}))\mathbf{p}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a})\right]\right|$$

less than 2ϵ. This inequality implicitly states that ϵ-deviation implies MPE as in the deterministic Bellman equation when ϵ goes to zero.

Remark 3.6. Minimization in (3.8) captures the main feature of constructing a robust procedure. It states that in order to obtain a small risk (the expected value of a loss function) in the prediction, we need to control both empirical risk, via the

4 The idea of ϵ-tube is similar to the idea of averaging in statistics in sense that it maps a stochastic feature into deterministic parameter values.

5 The dual problem of $\min|f|$ is $\min\zeta$ s.t. $-\zeta \le f \le \zeta$.

ϵ-insensitive loss function, and model complexity, via penalties ρ. The parameter C trades off the complexity and the prediction ability of the approximation.

The key idea of solving problem (3.8) is to construct a Lagrangian from the objective function and setup dual constraints. The Lagrangian function is the following:

$$\begin{aligned} L :=& \frac{1}{2}\|\rho\|^2 + \frac{C}{|\mathcal{S}|}\sum_{t=1}^{T}(\zeta_t + \zeta_t^*) - \sum_{t=1}^{T}(\eta_t\zeta_t + \eta_t^*\zeta_t^*) \\ &- \sum_{t=1}^{T}\alpha_t \cdot \left(\epsilon + \zeta_t + \rho^T\Phi(\mathbf{s}_t) - \Gamma\left(\rho^T\Phi(\mathbf{s}_t)\right)\right) \\ &+ \sum_{t=1}^{T}\alpha_t^* \cdot \left(\epsilon + \zeta_t^* - \rho^T\Phi(\mathbf{s}_t) + \Gamma\left(\rho^T\Phi(\mathbf{s}_t)\right)\right), \end{aligned} \tag{3.9}$$

where α_t^* and α_t are the multipliers for ϵ-inequalities and $\eta_t^*$and η_t are the multipliers for the positive constraints $\zeta_t^* \geq 0$ and $\zeta_t \geq 0$ respectively, and $\alpha_t^*, \alpha_t, \eta_t^*, \eta_t \geq 0$. For the *fitting error* $\rho^T\Phi(\mathbf{s}_t) - \Gamma\left(\rho^T\Phi(\mathbf{s}_t)\right)$, we use a regression-type expression to separate the approximating value function $\rho^T\Phi(\cdot)$:

$$\rho^T\Phi(\mathbf{s}_t) - \Gamma\left(\rho^T\Phi(\mathbf{s}_t)\right) = \rho^T\Psi(\mathbf{s}_t) - \pi(\mathbf{a}_t, \mathbf{s}_t, \xi_t), \tag{3.10}$$

where $\Psi(\mathbf{s}) := \Phi(\mathbf{s}) - \beta\sum_{\mathbf{s}'\in\mathcal{S}}\Phi(\mathbf{s}')\mathbf{p}(\mathbf{s}'|\mathbf{s},\mathbf{a})$. The partial derivatives of L with respect to the primal variables (ρ, ζ, ζ^*) equal to zero for optimality:

$$\begin{aligned} \partial L/\partial\rho &= \rho - \sum_{t=1}^{T}(\alpha_t^* - \alpha_t)\Psi(\mathbf{s}_t) = 0, \\ \partial L/\partial\epsilon &= -\sum_{t=1}^{T}(\alpha_t^* - \alpha_t) = 0, \\ \partial L/\partial\zeta_t^{(*)} &= \frac{C}{|\mathcal{S}|} - \alpha_t^{(*)} - \eta_t^{(*)} = 0, \end{aligned} \tag{3.11}$$

where $(*)$ indicates that analogous results hold for both ζ_t and ζ_t^*. Substituting (3.11) into (3.9) yields the dual optimization problem,

$$\begin{aligned} \min_{\alpha,\alpha^*} \quad & -\frac{1}{2}\sum_{i,j=1}^{T}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle\Psi(\mathbf{s}_i), \Psi(\mathbf{s}_j)\rangle \\ & - \sum_{t=1}^{T}(\alpha_t^* + \alpha_t)\epsilon + \sum_{t=1}^{T}(\alpha_t^* - \alpha_t)\pi(\mathbf{a}_t, \mathbf{s}_t, \xi), \\ s.t. \quad & 0 \leq \alpha^*, \alpha \leq C/|\mathcal{S}| \text{ and } \sum_{t=1}^{T}(\alpha_t^* - \alpha_t) = 0, \end{aligned} \tag{3.12}$$

where the quadratic term is the simplification of

$$\frac{1}{2}\|\rho\|^2 - \sum_{t=1}^{T}(\alpha_t^* - \alpha_t)\left[\rho^T\Phi(\mathbf{s}_t) - \Gamma\left(\rho^T\Phi(\mathbf{s}_t)\right) + \pi(\mathbf{a}_t, \mathbf{s}_t, \xi_t)\right].$$

The dual optimization problem (3.12) is a convex linear quadratic programming. Many available packages can solve this standard problem (3.12). Given the numerical values α and α^*, coefficients ρ and approximating value function $\hat{V}$ are:

$$\begin{aligned}\rho &= \sum_{t=1}^{T}(\alpha_t^* - \alpha_t)\Psi(\mathbf{s}_t),\\ \hat{V}(\mathbf{s}_t) = \rho^T\Phi(\mathbf{s}_t) &= \sum_{t=1}^{T}(\alpha_t^* - \alpha_t)\times \\ &\left[K(\mathbf{s}_t) - \beta\sum_{\mathbf{s}_{t+1}\in\mathcal{S}}\mathbf{p}(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a})K(\mathbf{s}_t)\right]\end{aligned} \tag{3.13}$$

where $K(\mathbf{s}_t) := \langle\Phi(\mathbf{s}_t), \Phi(\mathbf{s}_t)\rangle$. The coefficient vector ρ consists of the dual slack variables α, α^* and basis functions $\Phi(\cdot)$. As a function of α, α^* that dually represent inequality conditions, the coefficient ρ captures the influence from bounded rationality via ϵ. The optimal α, α^* also implies that the estimated ρ should trade off the complexity that preserves the shape of the approximating solution and the feasibility of implementations. When ϵ is changed, the approximation $\hat{V}$ will change as reflected in a change of ρ.

The constraint (3.6) restricts the potential solution set of $\hat{V}$ as a regularized operator. We denote this regularized operator Y such that it maps from a kernel solution $K(\cdot)$ into a restricted inner product $\Omega(K) := \langle \mathrm{Y}\Phi, \mathrm{Y}\Phi\rangle$ where $\Omega(K)$ is the regularization term. The operator Y induces an invariant subspace, because it imposes that interesting "potential candidate" functions should "almost" satisfy the constraint and should be stable under Y transformation.

Before giving the main result of this chapter, we need three technical conditions. Condition 3.7 is the *Lipschitz condition* for the profit function and transition density function. Condition 3.8 is to abstract from unboundedness issues in basis and kernel functions. Both of them will restrict us to a "stable" functional class. Condition 3.9 is the necessary condition for applying empirical process results in Bousquet and Elisseeff (2002). This condition basically implies independent increments.

Condition 3.7. (i) There exists a constant C_π such that for any $\mathbf{a} \in \mathcal{A}$ and $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ such that $|\pi(\mathbf{s}, \mathbf{a}, \xi) - \pi(\mathbf{s}', \mathbf{a}, \xi)| \leq C_\pi |\mathbf{s} - \mathbf{s}'|$.

(ii) For any $\mathbf{a} \in \mathcal{A}$, $\mathbf{s}, \mathbf{s}', \mathbf{t} \in \mathcal{S}$ and $G(\xi)$ with Radon–Nikodym density $g(\xi)$, $|p(\mathbf{s}'|\mathbf{s}, \mathbf{a}) - p(\mathbf{s}'|\mathbf{t}, \mathbf{a})| \cdot |g(\xi)| \leq C_p(\mathbf{s}')\|\mathbf{s} - \mathbf{t}\|$ where $C_p(\mathbf{s}')$ is an envelop function in L^2.

Condition 3.8. There is an upper bound κ such that $\|k(\mathbf{s}, \mathbf{s}')\| < \kappa$ for any $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$. There is an associated upper bound M^* such that $|\hat{V}| = |\rho^T \Phi(\mathbf{s})| < M^*$ for all $\mathbf{s} \in \mathcal{S}$.

Let π_t denote $\pi(\mathbf{a}_t, \mathbf{s}_t, \xi_t)$. We define a fitting penalty function $\tilde{c}(\mathbf{s}_j, \pi_j, \hat{V})$ in terms of a fitting error function in (3.10):

$$\hat{V}(\mathbf{s}_j) - \Gamma \hat{V}(\mathbf{s}_j) = \sum_{j \in \mathcal{S}_i} (\alpha_j^* - a_j) \left\langle \Psi(\mathbf{s}_j), \Psi(\mathbf{s}_j) \right\rangle - \pi(\mathbf{a}, \mathbf{s}_j, \xi)$$

and for exact fixed point condition $\tilde{c}(\mathbf{s}_j, \pi_j, V)$.

Condition 3.9. The form $\tilde{c}(\mathbf{s}_t, \xi_t, \hat{V}) = \hat{V}(\mathbf{s}_t) - \Gamma \hat{V}(\mathbf{s}_t)$ in (3.10) is independent across t.

Theorem 3.10. *Given conditions 3.7 to 3.9, for any action* $\mathbf{a}$ *given state* $\mathbf{s}$ *based on* T *observations, the limit*

$$\lim_{T \to \infty} \Pr \left\{ \frac{1}{T} \sum_{t=1}^{T} \left| \left| \tilde{c}(\mathbf{s}_t, \pi(\mathbf{a}, \mathbf{s}_t, \xi_t), \hat{V}) \right|_\epsilon - \tilde{c}(\mathbf{s}_t, \pi(\mathbf{a}, \mathbf{s}_t, \xi_t), V) \right| > 4\epsilon \right\}$$

equals zero, where $\hat{V}$ *is* $\rho^T \Phi(\mathbf{s})$ *in equation (3.13) and* V *is the solution in the exact fixed point condition. We say that the* $\hat{V}$ *is* non-trivially *consistent to* V.

The concept of *nontrivial* consistency is proposed by Vapnik and Chervonenkis (Vapnik, 1998). This is the concept of consistency over all tractable candidate functions where classic consistency requires selection over all admissible functions for a given sample. While nontrivial consistency requires that the induction principle be consistent even after the "best" functions have been removed. Because these "best" functions are usually the best in some given samples but not the best in other test samples, moreover, these "best" functions are often much more complicated than the second best functions. Theorem (3.10) is another version of law of large numbers which is uniform over all second-best candidate functions. Vapnik and Chervonenkis propose this concept to rule out any function which uniformly does better than all other functions within sample but has little predictive power out of sample.

Remark 3.11. Theorem 3.10 is also a very useful condition for capacity control. The capacity term is a property of the function class which can be measured by entropy numbers. The robustness concern depresses the growth rate of entropy numbers[6] so that it avoids using functional forms that are too complicated. Since the ϵ-approximation makes the fixed point condition converge to a small neighborhood of the exact fixed point condition, the simplification of introducing ϵ does not sacrifice too much consistency. In addition we have a computationally more efficient expression, a linear-quadratic programming problem.

3.3.2 *Iterative Policy Algorithm*

Finding the optimal policy function σ in (3.3) is often considered as a dual problem of (3.3) itself. The common approach of defining policy functions in the dynamic discrete choice literature is to find out a specified conditional choice probability. However, the use of ϵ-fixed point conditions in our framework will make the computation of conditional choice probability *invalid*.

The conditional choice probability is used to set up the transition density of the likelihood function (Rust, 1987; Aguirregabiria and Mira, 2007) or to derive a closed form policy function (Hotz and Miller, 1993; Bajari et al., 2007) or both. In the first step of our approach, we use the ϵ-approximation to nonparametrically solve the dynamic programming problem. If we derive the choice probability from some parametric likelihood, then the bias from ϵ-approximations will be accumulated. Thus it is impossible to derive a closed form policy function from the approximating value function in our framework.

The use of a conditional choice probability is to recover the policy function of (3.3). But it often causes a huge computational burden because of specifying the relation between the policy rule and structural parameters. Su and Judd (2011) give an extensive discussion of this issue and propose a computational alternative. Su and Judd (2011) pointed out that computing a conditional choice probability is not a necessary nor a sufficient way to find out the optimal policy rule in the dynamic discrete choice problem. A good solver in optimization programs will implicitly define the policy function and will implement the augmented likelihood through updating the optimization process.

6 Uniform convergence often requires a exponential bound which has a factor in terms of an entropy number. By Theorem 3.10, Y is a scale operator so that the entropy number changes by scaling.

We use an algorithm oriented method to search the optimal policy rule. With the explicit kernel-based expression (3.13), we can set up an iteration algorithm to obtain the estimated policy function $\hat{\sigma}$. Note that due to the ϵ-loss function, the estimated policy function is not necessarily equivalent to that in the recursive forward iteration method. The ϵ-tube avoids "over-fitting" the model thus $\Gamma\hat{V}$ will not fluctuate within the ϵ-wide tube. Also, an ϵ-optimal policy rule may be different from the one in the exact fixed point problem.

The algorithm is given below:

1. Set the initial policy $\mathbf{a}_0$, select a basis function and its corresponding kernel K.
2. Choose a subset $\mathcal{S}_{\mathbf{a}}$ of states space and ensure that the transition density between any two of them are strictly positive.
3. Given the action $\mathbf{a}_t$ ($\mathbf{a}_0$ for the first evaluation), calculate the kernel matrix $K^a := \{K_{kj} = \langle \Phi(\mathbf{s}_k), \Phi(\mathbf{s}_j) \rangle | \mathbf{s} \in \mathcal{S}_{\mathbf{a}}\}$.
4. Given the profit function π evaluated at policy $\mathbf{a}$ and kernel matrix K, solve the optimization problem (3.12) for α^* and α.
5. Apply (3.13) to obtain the ex ante value function $V(\mathbf{s})$, and then calculate the one-step policy improvement:

   $$\mathbf{a}_{t+1} = \arg\max_{\mathbf{a}} (\alpha^* - \alpha) \left[K^a(\mathbf{a}_t) - \beta \sum \mathbf{p}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) K^a(\mathbf{a}_t) \right].$$

   Set next period action to $\mathbf{a}_{t+1}$, update the policy rule and then go back to Step-3.

The algorithm procedure is similar to the inner iteration of the NXFP (Rust, 1987) except that we implement the kernel-based optimization. Thus one might be temped to compare the results of NXFP with (3.12). However, we have to emphasize that it is inappropriate to use the policy function from the original Bellman problem to judge the correctness of the policy function based on the ϵ-deviated Bellman problem. The two methods focus on different aspects. People who use the approximation to solve the exact Bellman equations have a lot of faith in the fixed point condition and are convinced the equilibria satisfy this condition. A standard method for obtaining the optimal policy function is by solving the following system of linear equations:

$$V^*(\mathbf{s}) = \sum_{\mathbf{a} \in \mathcal{A}} F^*(\mathbf{a}) [\pi^*(\mathbf{a}) + \mu^*(\mathbf{a})] + \beta \sum_{\mathbf{s}'} V^*(\mathbf{s}') \mathbf{p}^*(\mathbf{s}'|\mathbf{s}), \tag{3.14}$$

where μ^* is the expectation of ξ conditional on $\mathbf{a}$ and $F^*(\mathbf{a})$ is the conditional choice probability. P^*, π^*, e^* are vectors that stack the corresponding state-specific elements. Star $*$ represents the elements associated with an equilibrium conditional choice probability. This method has been used for example by Aguirregabiria and Mira (2007).

People using the robust kernel-based optimization are less optimistic and are willing to consider actions that are less than perfect. It is quite intuitive that the solution will converge to (3.14) when ϵ is set to zero in (3.8). The ϵ can therefore be used to express the degree of confidence in Bellman's principle of optimality. We give the following Theorem to formalize this intuition.

Theorem 3.12. *When ϵ goes to zero, the approximating value function obtained by (3.13) is equivalent to the solution $V^*(\mathbf{s})$ in (3.14). The iterative policy algorithms are also equivalent.*

The iterative policy algorithm may not be satisfactory even if the value function approximation is good. A small ϵ might give a highly fluctuating policy function and then the wiggling policy function induces many local equilibria while a large ϵ may give a poor approximation for the value function. Since the iterative policy algorithm depends on the approximating value functions, choosing ϵ too small or too large leads to serious consequences. One may expect the policy iteration not to be a robust approach, in other words, policy iteration is sensitive to the selection of ϵ. To obtain a stable policy function, either one needs to have very precise knowledge on ϵ or impose constraints on the shape of the policy function. Cai and Judd (2010) show that monotonic and concave constraints on policy functions significantly improve the stability of solutions.

3.4 THE SECOND STEP ESTIMATION

In the preceding section, θ is treated as a structural parameter in Equation (3.2) and thus for the integrated Bellman equation (3.4) θ is incoporated in the integrated value function and ρ accommodates the robustness concerns in the ϵ-fixed point condition (3.6). We need to establish the relation between these two systems in order to estimate θ. However, there are two obstacles. First, one has no prior information about the form of the distribution of the random variable ξ, nor of the underlying parameter space. Second, the pseudo parameter ρ is irrelevant

for θ in the first step. In this section, we formulate these obstacles as another constrained optimization problem and then solve it.

The nonparametric smoothing densities are feasible solutions for the first obstacle. Empirical Likelihood (EL Owen, 1988, 1990, 2001) generates a so-called implied density function based on model constraints. EL distributes weights according to the imposed constraints. The input values satisfying the model constraints will be assigned to higher weights while those violating the model constrains will be assigned to lower weights or be penalized to zero if it is an outlier.

We handle the second obstacle by constructing a relation between nonparametric solutions $\hat{V}$ and parametric functions of θ. However, the construction will render the EL or GMM method infeasible, because of a *non-differentiable* moment constraint. In this section, we use a local type EL estimator to solve this problem. For theoretical properties of this method, please refer to Chapter 2.

In Section 3.4.1, we will replace observations $\mathbb{E}_\xi \pi(\cdot, \xi)$ in (3.3) with a parametric function $\mathbb{E}_\theta \pi(\cdot, \xi)$ in order to combine the expression of θ and ρ. The new expression of the fitting error is a combination of nonparametric and parametric forms:

$$\rho^T \Phi(\mathbf{s}) - \mathbb{E}_\theta \pi(\sigma(\mathbf{s}, \xi), \mathbf{s}, \xi) + \beta \int \left[\rho^T \Phi(\mathbf{s}')\right] dP(\mathbf{s}'|\mathbf{s}, \mathbf{a}),$$

where for each observation π_t, $\sigma(\mathbf{s}_t, \theta) = \mathbf{a}_t$ is the sample action. In boundedly rational cases, there will be a set of strategies σ satisfying the ϵ-equilibrium condition, but here we will only consider a local optimal case when the fitting error is zero for each σ in this class. The optimized $\hat{\theta}$ for the zero-fitting error will be interpreted as the structural estimator. The problem in this new construction, as well as in many control oriented problems, is the discontinuity of the policy function $\sigma(\theta, \mathbf{s})$ over θ, e.g. kinks in choice functions. This issue will be enlarged when we plug in a "biased" approximating representation $\rho^T \Phi(\mathbf{s})$. Thus, in subsection 4.2 we need to consider an estimation that is robust towards non-differentiability.

3.4.1 *The Semi-parametric Constraints*

In parametric models, the constraint for the second step estimation usually forms a fixed point condition via the conditional choice probability (Pesendorfer and Schimidt-Dengler, 2003; Aguirregabiria and Mira, 2007) or an equilibrium condition of

the policy function and structural parameters (Su and Judd, 2011; Bajari et al., 2007). A nonparametric model does not specify the functional form of the distribution of ξ nor the dependence on θ, thus we have to seek an alternative constraint.

We recall that the approximating integrated Bellman equation for parametric models is:

$$\hat{V}(\mathbf{s}) = \mathbb{E}_\theta \left[\pi(\mathbf{a},\mathbf{s},\xi) + \beta \int \hat{V}(\mathbf{s}') dP(\mathbf{s}'|\mathbf{s},\mathbf{a}) \right] = \rho^T \Phi(\mathbf{s}). \quad (3.15)$$

If we replace $\pi(\mathbf{a},\mathbf{s},\xi)$ with $\bar{\pi}(\mathbf{s},\mathbf{a}) + \xi$ by Condition 3.2, the expression becomes

$$\begin{aligned} \rho^T \Phi(\mathbf{s}) =& \mathbb{E}_\theta \left[\bar{\pi}(\mathbf{a},\mathbf{s}) + \xi(\mathbf{a}) + \beta \int \rho^T \Phi(\mathbf{s}) dP(\mathbf{s}'|\mathbf{s},\mathbf{a}) \right], \\ =& \bar{\pi}(\mathbf{a},\mathbf{s}) + \mathbb{E}_\theta \xi(\mathbf{a}) + \beta \int \rho^T \Phi(\mathbf{s}) dP(\mathbf{s}'|\mathbf{s},\mathbf{a}). \end{aligned} \quad (3.16)$$

The equality follows from the ϵ-equilibrium condition 3.3. Rewrite equation (3.16) as:

$$\rho^T \Phi(\mathbf{s}) - \left[\bar{\pi}(\mathbf{a},\mathbf{s}) + \beta \int \rho^T \Phi(\mathbf{s}) dP(\mathbf{s}'|\mathbf{s},\mathbf{a}) \right] = \int \xi dG_\theta(\xi|\mathbf{a}) \quad (3.17)$$

With the ergodicity condition on ξ_t and Fubini's Theorem, the sample average of equation (3.17) has the following property:

$$\begin{aligned} & \sum_t^T \frac{\left[\rho^T \Phi(\mathbf{s}_t) - \left(\bar{\pi}(\mathbf{a},\mathbf{s}) + \beta \int \rho^T \Phi(\mathbf{s}_t) dP(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t) \right) \right]}{T} \\ =& \int \left[\sum_t \xi_t / T \right] dG_\theta(\xi|\mathbf{a}) \to \int \xi dG_\theta(\xi|\mathbf{a}). \end{aligned} \quad (3.18)$$

where G_θ is the parametric distribution with unknown parameter.

The ϵ-equilibrium corresponds to a class of policy functions $\sigma(\mathbf{s},\theta)$. We will only consider the robust decision function, where value function of MPE $\sigma(\mathbf{s},\theta)$ will be covered by the ϵ-tube (3.6) such that $\mathbb{E}_\theta \xi(\mathbf{a}) < \epsilon$. Given fixed ϵ, we only need to minimize the following fitting error $m_t(\theta|\mathbf{a})$ w.r.t. θ:

$$\left[\rho^T \Phi(\mathbf{s}_t) - \left(\bar{\pi}(\mathbf{a},\mathbf{s}) + \beta \int \rho^T \Phi(\mathbf{s}_{t+1}) dP(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t) \right) \right] - \mathbb{E}_\theta \xi(\mathbf{a}).$$

As in equation (3.16), the limit of the sample average of $m_t(\theta|\mathbf{a})$ over t is an integration w.r.t. G. When $\theta = \theta_0$, $dG_0(\xi|\mathbf{a})$ implies

that $\int m(\theta_0)dG_0(\xi|\mathbf{a})$ conditional on the action trajectory $\mathbf{a}$. The empirical counterpart of G can be estimated by EL.

In practice, heterogeneous action trajectories of $\mathbf{a}$ brings a great difficulty of obtaining an empirical counterpart of $G_0(\xi|\mathbf{a})$. Beside using $G_0(\xi|\mathbf{a})$, one may prefer to consider $dG_\theta(\xi) = dG_\theta(\xi|\mathbf{a})dP(\mathbf{a})$ where $P(\mathbf{a})$ is the choice probability. Then the conditional moment constraint function $m_t(\theta|\mathbf{a})$ is replaced by $m_t(\theta|\mathbf{a})p(\mathbf{a})$ which we denote $m_t(\theta)$. However, in this case, one will confront an issue of executing the derivative of

$$\int m_t(\theta|\mathbf{a})dG_\theta(\xi|\mathbf{a})dP(\mathbf{a})$$

when $\mathbf{a}$ is discrete. To see this problem, we need to point out that $dG_\theta(\xi) = dG_\theta(\xi|\mathbf{a})dP(\mathbf{a})$ is a mixture distribution. Mixtures of log-concave densities may be log-concave, but in general they are not. For instance, if $\mathbf{a} = \{0,1\}$ and $p(\mathbf{a}) = P(\mathbf{a} = 0)$, the location mixture of standard univariate normal densities $dG_\theta(\xi) = \phi(\xi)$

$$\phi(\xi)p(\mathbf{a}) + \phi(\xi - c\mathbf{a})(1 - p(\mathbf{a}))$$

is log-concave if and only if $c\mathbf{a} \leq 2$. Bi-modal or multi-modal distributions lead to non- continuous derivative and zero-Hessian of the log-likelihood of $G_\theta(\xi)$. The worst scenario is the non-differentiablity [7] of $\int m_t(\theta|\mathbf{a})dG_\theta(\xi)$ at the extrema. In the following sub-section, we propose an approach to approximate the derivative of the likelihood ratio of $G_\theta(\xi)$ without using differentiation techniques.

3.4.2 *Empirical Likelihood and Local Empirical Likelihood*

Let $m_t(\theta)$ denote $\int m_t(\theta|\mathbf{a})dP(\mathbf{a})$. The constraint (3.18) is used to identify the unknown distribution function $G(\xi)$ and speed up the computation convergent rate. We apportion the probabilities $g = (g_1, \ldots g_T)$ for the distribution G. The sum of empirical log-likelihood contribution is $\sum_t \log Tg_t$. In addition, g should satisfy

7 The non-differentiability implies that there is no closed form derivative on $\mathbb{R}$. It is possible that the functional derivative is feasible and tractable.

the common requirements of probabilities such that $g_t \geq 0$ and that $\sum_t g_t = 1$. The EL criterion is:

$$\begin{aligned} &\max_{\theta} \sum_{t=1}^{T} \log T g_t, \\ &\text{s.t.} g_t \geq 0, \quad \sum_{t=1}^{T} g_t = 1, \\ &\sum_{t=1}^{T} g_t m_t(\theta) = 0, \end{aligned} \tag{3.19}$$

where the constraint $\sum_t g_t m_t(\theta) = 0$ expresses the fact that in the second stage we pursue zero fitting error for the ϵ-equilibrium. The Lagrangian is:

$$L' := \sum_{t=1}^{T} \log T g_t - T\lambda \sum_{t=1}^{T} m_t(\theta) + \gamma(\sum_{t=1}^{T} g_t - 1). \tag{3.20}$$

With the help of the Lagrange multiplier λ, we have:

$$\tilde{g}_t(\theta) = \frac{1}{T} \frac{1}{1 + \lambda m_t(\theta)}, \tag{3.21}$$

the so-called implied density g. Substituting (3.21) into (3.19), we have a dual representation of problem (3.19):

$$\begin{aligned} &\min_{\theta} - \sum_{t=1}^{T} \log T \left(1 + \lambda m_t(\theta)\right), \\ &s.t. \frac{1}{T} \sum_{t=1}^{T} \frac{m_t(\theta)}{1 + \lambda m_t(\theta)} = 0. \end{aligned} \tag{3.22}$$

Problem (3.22) is a standard nonlinear optimization problem. The outer-loop of the optimization is to minimize minus the empirical log-likelihood with respect to θ and the inner-loop is to obtain numerical solution of λ.

The outer-loop optimization in (3.22) could be discontinuous due to the discrete feature of $\sigma(\mathbf{s}, \theta)$. Thus the outer-loop search in optimization is unstable. Let H and s denote the Hessian and gradient function of $\sum_t \log T \left(1 + \lambda m_t(\theta)\right)$ with respect to θ. The $k+1$-step Newton iteration used in most computational methods is

$$\theta^{(k+1)} = \theta^{(k)} - H(\theta^{(k)})^{-1} s(\theta^{(k)}). \tag{3.23}$$

The evaluation of Hessian matrix $H(\theta^{(k)})$ requires the second derivative of the log-likelihood function. The numerical derivative is difficult to implement and time consuming because of

nonlinearity and non-differentiability. The Hessian matrix in this region is so flat that singularity problems may occur.

By the localization technique, we have a linear-quadratic representation for the log-likelihood ratios:

$$-\min_{\tau}\sum_{t=1}^{T}\left[\tau^T S_t - \frac{1}{2}\tau^T M_t \tau\right]. \tag{3.24}$$

The construction of S and M of this representation are computed without using derivative argument. The construction of this representation will be described in the next subsection.

The term "local" here is a counterpart of "differential". One fixes a particular θ_0 and investigates what happens to likelihood ratio functions with parameter values of $\theta = \theta_0 + \delta_T \tau$, with $\delta_T \to 0$. Here, δ_T is a kind of "differentiation rate". We will fix the differential rate to $T^{-1/2}$ such that $\theta + T^{-1/2}\tau$ and we will call τ the local parameter.

Remark 3.13. In the parametric case, people usually assume that the unknown private shocks have multinomial distributions. The implied density (3.22) is far more flexible than a parametric density because it can assign the weights arbitrarily. In addition, if the log-likelihood ratio in (3.19) is replaced by entropy $\sum_t T g_t \log T g_t$, Kitamura and Stutzer (1997) show that the implied density will become

$$\tilde{g}_{t,ET}(\theta) = \frac{\exp \lambda m_t(\theta)}{\mathbb{E}_t \exp \lambda m_t(\theta)}, \tag{3.25}$$

which is similar to the multinomial choice probability.

3.4.3 *Construction of the Second-Step Estimator*

Partition the parameter space of θ into several multi-dimensional grids and select one particular value on this grid. Denote the selected value θ^* and let

$$\Lambda_T(\theta_1, \theta_2) = \sum_{t=i}^{T} \log(\tilde{g}_{it}(\theta_1)/\tilde{g}_{it}(\theta_2)).$$

Now run the following scheme at *every grid point* θ^*

1. M_T with $M_{T,q,p} = u_q^T M_T u_p$, $q, p = 1, 2, \ldots, l$, given by

$$\begin{aligned} M_{t,q,p} = -\Big\{ & \Lambda_T[\theta^* + \sqrt{T}(u_q + u_p), \theta^*] \\ & -\Lambda_T[\theta^* + \sqrt{T}u_q, \theta_t^*] - \Lambda_T[\theta^* + \sqrt{T}u_p, \theta^*] \Big\} \end{aligned}$$

where $\{u_1,\ldots,u_l\}$ is a linearly independent set of directional vectors in $\mathbb{R}^l$ selected in advance.

2. Construct a linear term $S_T = \{S_{T,q}\}$, $q = 1,2,\ldots,l$, by the linear-quadratic approximation function:

$$S_{T,q} = \Lambda_T[\theta^* + \sqrt{T}u_q, \theta^*] + \frac{1}{2}M_{T,q,q}.$$

Since all the values on the RHS are known.

3. Construct a one-step improved estimator:

$$\tilde{\theta} = \theta^* + \sqrt{T}M_T^{-1}S_T,$$

4. Obtain the value of $\sum_t \log n\tilde{g}(\tilde{\theta} + T^{-1/2}\tau)$ and compare it with the values from other grid points. If $\tilde{\theta}$ returns a higher likelihood value, it will be selected.

The constructed Hessian type matrix M_T in step 2 is invertible. The advantage of the local method is that the gradient vectors and Hessian matrices are available without any differentiation operation. As we mention earlier, the mixing parametric choice distribution implies an irregular shape of the likelihood function so that gradient and Hessian is less tractable than in the usual case. As one can see, the global irregular shape plays no role in our local quadratic construction and the estimator construction. The calculation of M at every fixed θ^* is independent of the numerical second-order derivatives.

The theoretical properties of this local EL estimator has been derived in Chapter 2. It is shown there that the estimator has a normal limiting distribution within the local neighborhood of the true θ and is asymptotically optimal. The result does not require globally smooth functions and the optimization only depends on local EL values. The estimation concerns local parameters rather than θ directly, thus it avoids evaluating non-differentiable or highly nonlinear functions of θ.

3.5 NUMERICAL ILLUSTRATION

We consider a simple application of Rust's model: optimal replacement of bus engines (Rust, 1987). This is a single agent dynamic discrete choice model, but it is a useful starting example to illustrate how the semiparametric algorithm works. We will use 117 observations of monthly data for the 1975 GMC

model 5308 buses, 37 in total. This is data set a530875 of Rust (1987).

In this setting, the maintenance manager of this fleet of 37 buses has to decide for each bus i how long to operate it before replacing its engine with a new one. The state s_{it} variable is the accumulated miles of the engine in bus i at time t. The decision variable a_{it} is whether to replace the engine $a_{it} = 1$ or maintain the engine $a_{it} = 0$ in bus i at time t. When a bus engine is replaced, it is as good as a new, so the state of the system regenerates to $s_{it} = 0$ when $a_{it} = 1$. The private shock $\xi_{it}(a_{it})$ in this case is the unforeseen cost to bus i in period t given decision a_{it}. The profit function π in our notation for the current example is equal to minus the cost since the maintenance manager takes the number of miles as given. Rust (1987) refers to this as the utility. This π is assumed to be additively separable and is given by:

$$\pi(a_{it}, s_{it}, \theta_1, \theta_2) = \begin{cases} -RC - C(0, \theta_1) - \xi_{it}(1) & \text{if } a_{it} = 1 \\ -C(s_{it}, \theta_1) - \xi_{it}(0) & \text{if } a_{it} = 0, \end{cases}$$

where $C(\cdot)$ is an engine's operating and maintenance cost function. The transition density for s_{it} depends only on miles driven at the beginning of time period $t + 1$:

$$p(s_{it+1}|s_{it}; a_{it}) = \begin{cases} g(s_{it+1} - 0) & \text{if } a_{it} = 1, \\ g(s_{it+1} - s_{it}) & \text{if } a_{it} = 0, \end{cases}$$

where $g(\cdot)$ is a known probability density function that may depend on a parameter θ_2. We will use $g(\cdot) = \theta_2 \exp[-\theta_2(\cdot)]$. In our approximation step (step-1), θ_2 is treated as a fixed value. The specification of cost function is one of the following:

$$\begin{aligned} \text{Quadratic:} \quad & C(s, \theta_1) = \theta_{11} s + \theta_{12} s^2, \\ \text{Power:} \quad & C(s, \theta_1) = \theta_{11} s^{\theta_{12}}, \\ \text{Mixed:} \quad & C(s, \theta_1) = \theta_{11} / (1.1 - s) + \theta_{12} s^{1/2}. \end{aligned} \tag{3.26}$$

Rust (1987) estimates θ_2 and θ_1 separately because only θ_2 affects the transition function. In step-1, we use the estimated result of θ_2 in Rust (1987) as the initial value of θ_2 and then simulate the transition density $p(s', s|\theta_2, a)$. In step-2 we update the values of θ_1 via (3.24). We do this iteratively until the algorithm converges. Thus in step-1, the transition density is taken as known and then new status is given in step-2.

The last equation of (3.26) is slightly different from the original one where the constant is set to 91. The reason is that

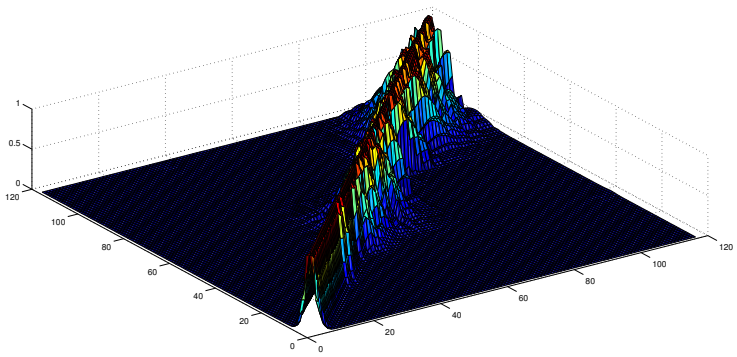


Figure 2: A realization of the kernel function $K(\cdot,\cdot)$ of the cumulated miles $\sum_i s_{it}$ and $\sum_i s_{it'}$ over 117 periods. The x-axis and the y-axis denote time periods t and t'. The z-axis denotes the transition kernel value of $K(\sum_i s_{it}, \sum_i s_{it'})$ for different t and t'.

we re-scale the state variables to the $[0,1]$-interval rather than discretize them into 90 states. The advantage of scaling is that it avoids the case where states in greater numeric ranges dominate those in smaller numeric ranges. Another advantage is that it avoids numerical difficulties in kernel calculation. In kernel calculations, the inner products of basis functions may generate large values which cause problems in the numerical operation.

The kernel matrix used in the approximation of the value function (see Equation (3.13)) is evaluated via $\langle \Psi(\mathbf{s}), \Psi(\mathbf{s}) \rangle$. It can be calculated in each period t as is done in Figure 2. The $\Psi(\mathbf{s})$ equals $\Phi(\mathbf{s}) - \beta \sum \Phi(\mathbf{s}') \mathbf{p}(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ with a fixed discount factor $\beta = 0.98$. The transition density $\mathbf{p}(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ changes over time simply because the action $\mathbf{a}$ changes and this causes the kernel defined by $\langle \Psi(\mathbf{s}), \Psi(\mathbf{s}) \rangle$ to change over time. Figure 2 shows that at the beginning stage the distribution is quite homogeneous but when the odometers accumulate to certain miles the transition distribution among states becomes divergent. Suppose that $\bar{\mathbf{s}}$ is a standard mileage for engine replacement such that it satisfies $V(\bar{\mathbf{s}}, \mathbf{a} = 1) = V(0, \mathbf{a} = 0)$. When the engine is close to this stage, the manager has a greater tendency to replace it. Therefore, the transition among states has significant dissimilarities at the later stage.

The sensitivity loss ϵ controls the goodness of fit of the approximation and furthermore affects the performance of the predictor. To analyze the effects of ϵ, we use different ϵ in the power cost function case. We compare approximation and

ϵ	MSE	r^2	Basis Num
0.005	0.609	0.736	108
0.050	0.614	0.739	63
0.500	0.960	0.690	28

Table 1: Sensitivity analysis of the approximation. By tuning the value of ϵ, we can examine the MSE of the approximation. Changes of the width of the ϵ-tube also affect the complexity of the approximation via the number of used basis functions.

inference ability of this kernel function based on test and estimation samples. With calibrated C, b and θ, Table 1 and Figure 3 give the results of goodness of fit and the predictor ability of ϵ-approximation. The first 50% data is used for estimation while the other is for prediction. The prediction ability is measured in two terms, Mean Squared Error (MSE) and squared correlation coefficient (r^2) which are $T^{-1}\sum_t^T(\rho^T\Phi(s_t)-\pi_t)^2$ and

$$r^2 = \frac{\left(T\sum_t \rho^T\Phi(s_t)\pi_t - \sum_t \rho^T\Phi(s_t)\sum_t \pi_t\right)^2}{\left(T\sum_t(\rho^T\Phi(s_t))^2 - (\sum_t \rho^T\Phi(s_t))^2\right)\left(T\sum_t \pi_t^2 - (\sum_t \pi_t)^2\right)} \tag{3.27}$$

respectively. MSE for test sample can be considered as the information loss due to the inaccurate prediction. The quantity r^2 measures the linear relationship between the approximating value function and the parametric profit function. It is obvious that a small ϵ makes the kernel over-fit and thus ask for more information to construct the fitting. MSE and r^2 are not significantly different for $\epsilon = 0.005$ and $\epsilon = 0.05$, the former case nearly uses the whole 117 sample points while the later case only asks for 63 out of 117. One can find out there is almost no loss by cutting 50% evaluations of samples. The simpler approximation is favored because a complicated basis function leads to intractable kernel computations and optimizations.

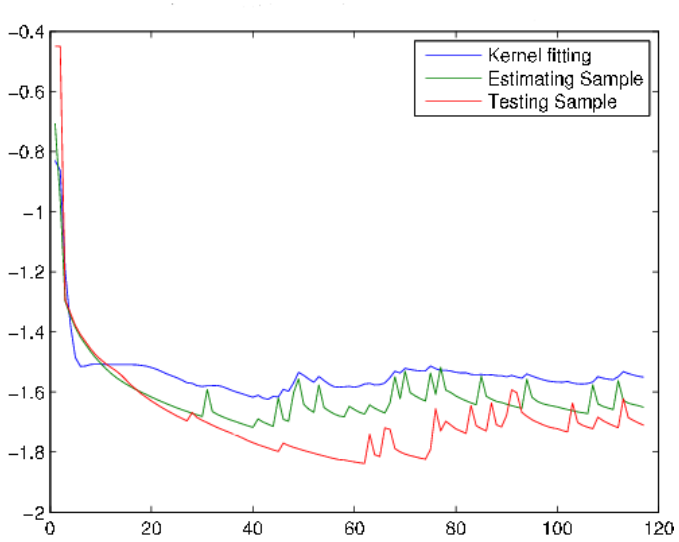
From $\epsilon = 0.05$ to $\epsilon = 0.5$, the prediction accuracy decreases slightly, which, however, is caused by an unpredictable downside shift in the averse direction of the trend. From Figure 5, we can realize that the curve shape of $\epsilon = 0.5$ is more flexible than that of $\epsilon = 0.05$. In addition, the basis number falls to 28, a big gain in simplicity with an insignificant cost of the prediction power. Such a benefit does not appear by increasing ϵ to 1. When $\epsilon = 1$, the fitting has poor MSE and r^2 values. Figure 4 shows the values of $(\alpha - \alpha^*)$ for the sub-state space.

Table 2 describes the performance of different cost functions. The cost function with power functional form shares similar behaviors with that of the quadratic form. The mixed functional form gives unsatisfied outputs for larger structural parameter values. The reason is that the power function and the quadratic function can be easily approximated via linear basis functions but not for the mixed function. To obtain a satisfied fitting of the mixed function, a bigger number of basis functions is required. Therefore the vectors used in the mixed function case is significant larger than the other cases. From Figure 3, we can find out that the value function curves do not seem to converge for large parameters. Because large parameter enlarges the fluctuation and thus causes divergent plots.

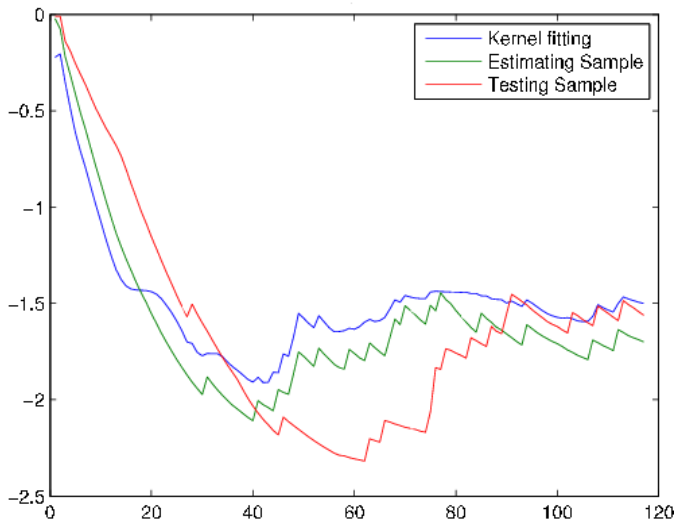
The average accuracy of predicting the validation sets is the cross validation accuracy. We also test the sensitivity of the choice of loss functions for the fixed point condition. It seems that, a small ϵ will lead to an exact classification policy rule. In this case, each decision is strictly isolated with the others and has a narrow inference interval. On the other hand, if we allow a flexible ϵ for the fixed point condition, we can use a simple policy rule for classification and have much larger confidence region for the whole data set.

3.6 CONCLUSION

In this chapter, we propose a new two-step estimation approach for dynamic discrete choice models with boundedly rational agents. In the first step, the method embeds robust decisions caused by bounded rationality into the Bellman's optimal principle. In the second step, the estimated model is accompanied by non-standard moment restrictions. We use the method developed in Chapter 2 to estimate this model. The solution of this problem is based on an ϵ-approximating value function with locally estimated structural parameter. We show that both the value function and the log-likelihood ratio of the local parameter have flexible representations. The result reflects the robustness concern of agents' decision processes.



(a)



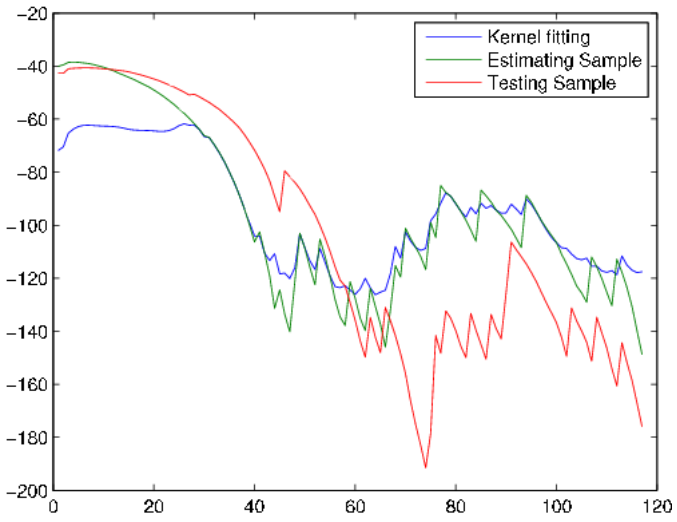
(b)

Figure 3: Prediction ability of the approximation $\rho^T \Phi(\mathbf{s})$ under different functional forms $C(s,\theta)$: (a) Power function $\theta_{11} = 0.1$, $\theta_{12} = 0.1$ (b) Quadratic function with $\theta_{11} = 0.32$, $\theta_{12} = 0.5$ (c) Mixed function $\theta_{11} = 2.5$, $\theta_{12} = 1.1$. One can compare the approximation $\rho^T \Phi(\mathbf{s})$ (in blue) with the value function of the used sample (in green) the testing sample (in red). The x-axis denotes the time period and the y-axis denotes the cost in the value function.

Function type	Param θ_{11}, θ_{12}	MSE	Nr Basis Func
Power	0.1, 0.1	0.031	21
Quadratic	0.3, 0.5	0.719	27
Mixed	2.5, 1.1	0.041	27

Function type	r^2	Likelihood
Power	0.840	−51.908
Quadratic	0.665	−17.957
Mixed	0.645	−20.909

Table 2: Likelihood Estimation: The results of the second step estimation.



(c)

Figure 3 Continued

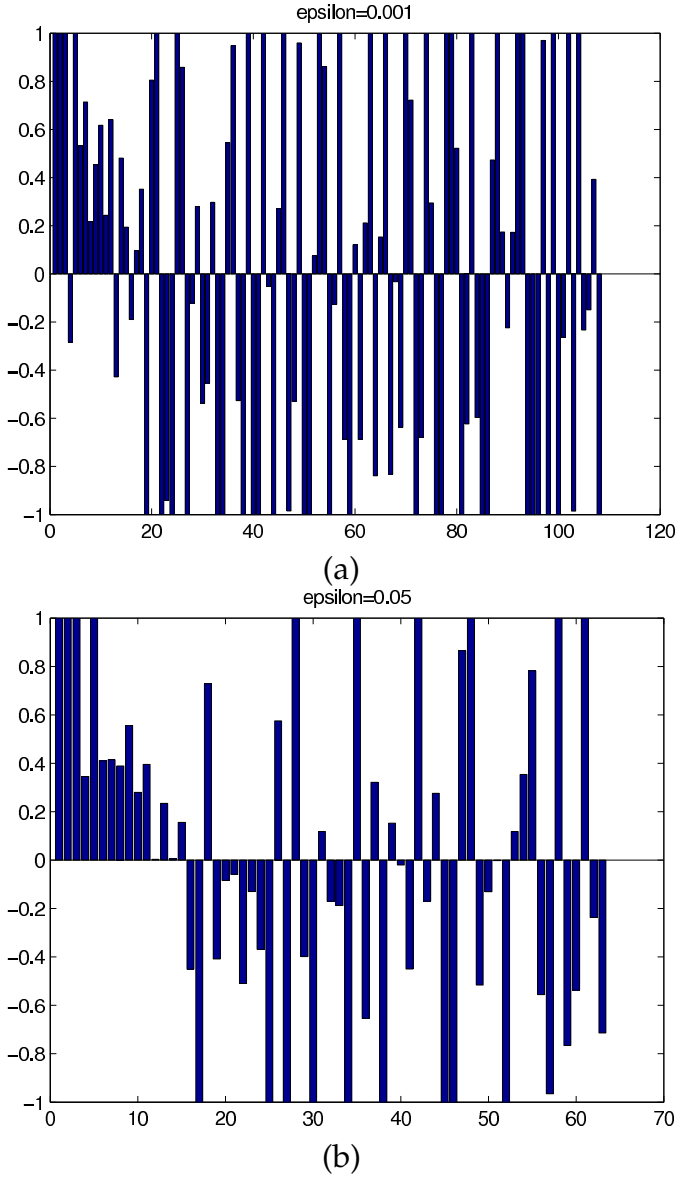
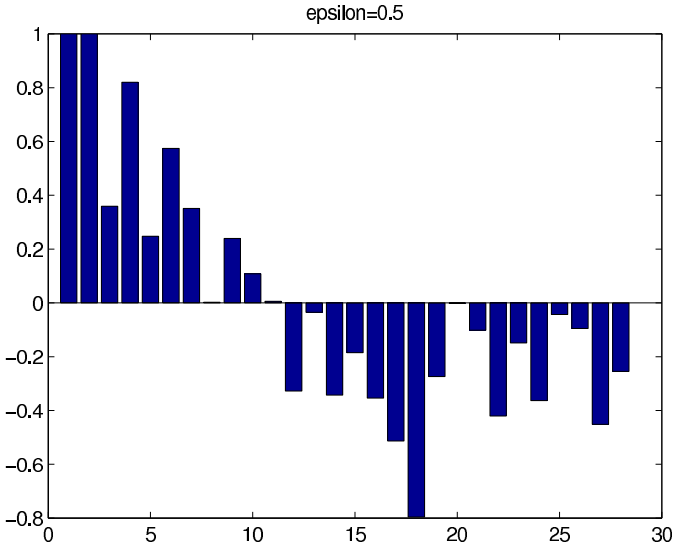
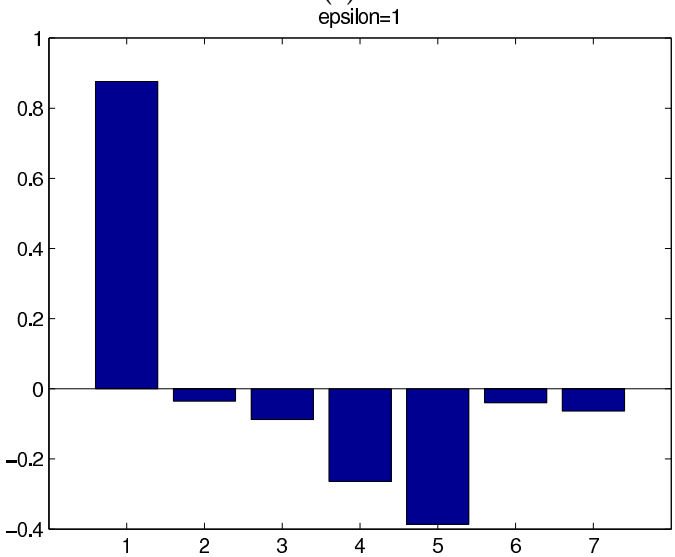


Figure 4: The sensitivity analysis of ϵ reflects on the change of $(\alpha - \alpha^*)$ since $(\alpha - \alpha^*)$ is the dual parameter of ϵ. We have the value of $(\alpha - \alpha^*)$ under (a) $\epsilon = 0.001$, (b) $\epsilon = 0.05$ (c) $\epsilon = 0.5$ and (d) $\epsilon = 1$. The y-axis denotes the value of $(\alpha - \alpha^*)$. The x-axis denotes the number of basis functions used.



(c)



(d)

Figure 4 Continued

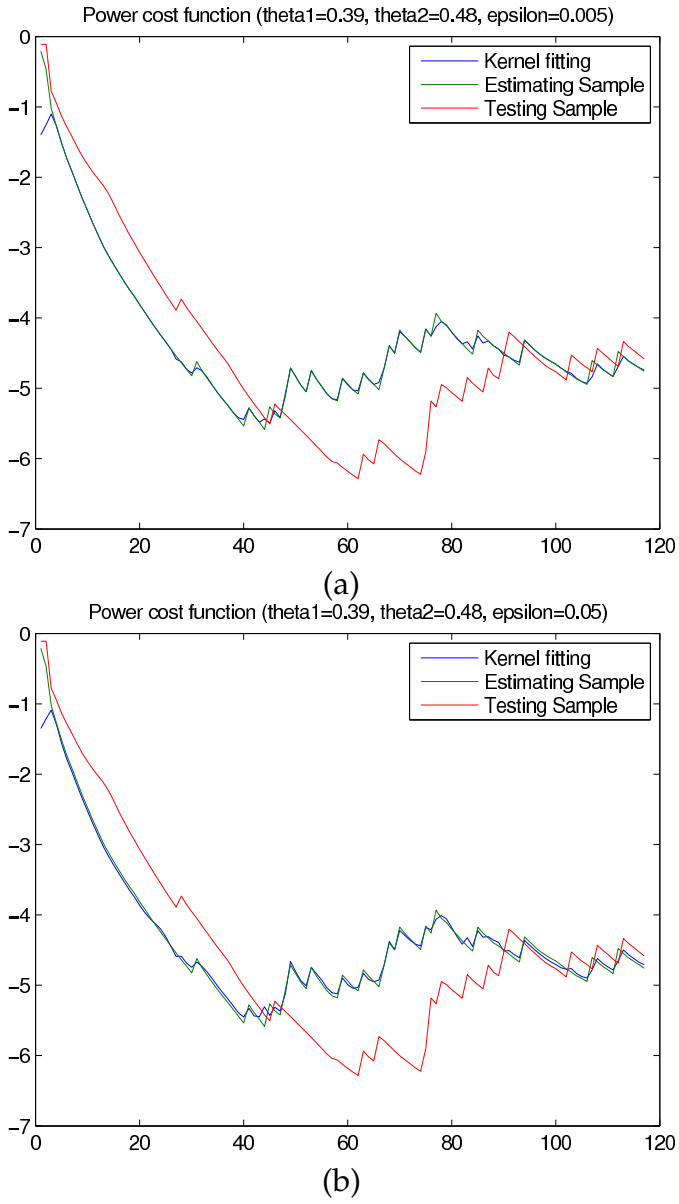


Figure 5: Different ϵ affect the results of the approximation $\rho^T \Phi(\mathbf{s})$. (a) $\epsilon = 0.001$, (b) $\epsilon = 0.05$ (c) $\epsilon = 0.5$ and (d) $\epsilon = 1$. The x-axis denotes the time period and the y-axis denotes the cost in the value function.

(c)

(d)

Figure 5 Continued

APPENDIX TO CHAPTER 3

PROOF OF THEOREMS

Proof of Theorem 3.10

First we need to introduce the definition of approximation stability. The purpose of introducing this definition is to restrict the function class of our potential solution sets.

Approximation Stability:

The approximation function $\hat{V}$ is a function of $\mathcal{S}$, if the $\mathcal{S}$ is changed to $\mathcal{S}^m$, a new approximation follows, $\hat{V}_{\mathcal{S}^m}$ say. We will now define a replacement perturbation by replacing the m-th observation with a new s. So $\mathcal{S}^m := (\mathcal{S} \backslash \{s_m\}) \cup \{s_{new}\}$, where s_{new} is drawn from a relevant distribution. In our case, this is the empirical distribution. The stability means the difference between $\hat{V}_{\mathcal{S}^m}$ and $\hat{V}$ is bounded. In the proof, we will give bounds for $\hat{V}_{\mathcal{S}^m} - \hat{V}$ and show that the approximation is stable.

Sketch of the proof:

The purpose of the proof is to show the fixed point condition with regularized Bellman operator $\Gamma\hat{V}$ converges uniformly to a neighborhood of the fixed point condition with $\hat{\Gamma}V$. A straightforward idea is to check the validity of the following argument:

$$\lim_{T\to\infty} \left| \sum_{t=1}^{T} \left(\hat{V}_t - \Gamma\hat{V}_t\right)/T - \mathbb{E}\left(V - \hat{\Gamma}V\right) \right|_{\epsilon} \to 0.$$

If an empirical process Theorem (e.g. Glivenko-Cantelli) can be applied to the left handside of the argument, then it would be simple. However, there are two difficulties of using an empirical process Theorem to prove the uniform convergence here. The first problem is the ϵ-loss function (ϵ-fixed point constraints) make $\hat{V}_j - \Gamma\hat{V}_j$ not exactly zero and the second is the effect from regularization. Fortunately, Theorem 12 from Bousquet and Elisseeff (2002) (BE Theorem) states that if changing observations

in a given loss function only leads to a bounded variation then there is a Hoeffding type bound[8] for this function.

In our setting, the proof that we will follow can be stated in this way:

If Condition 3.7 (i) holds and

$$\left|\tilde{c}(\mathbf{s},\pi,\hat{V}) - \tilde{c}(\mathbf{s},\pi,\hat{V}_{\mathcal{S}^m})\right| \leq \varsigma \tag{3.28}$$

for all (s,π) where ς depends on T then we have

$$\Pr\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_{\epsilon} - \mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)\right| > \epsilon + \varsigma\right\} \tag{3.29}$$

$$\leq 2\exp\left(-\frac{2T\epsilon^2}{(T\varsigma + M^*)^2}\right). \tag{3.30}$$

So if ς is of order $1/T$, then when $T \to \infty$ the exponential bound (3.30) goes to zero. The standard empirical process of $\tilde{c}(\mathbf{s}_t,\pi_t,V)$ has a Hoeffding bound such that

$$\Pr\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,V) - \mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,V)\right)\right| > \epsilon\right\} \tag{3.31}$$

$$\leq N(\epsilon)\exp\left(-\frac{2T\epsilon^2}{M^{*2}}\right),$$

where $N(\epsilon)$ is the covering number for the function $\hat{V}$ which is smoothed by a RBF kernel. When T increases, $\exp(-2T\epsilon^2/M^{*2})$ will decrease to zero at a faster rate than the bound in (3.30). Therefore, we can incorporate equation (3.31) into (3.29). Since the probability rule implies $2\Pr(A) \geq \Pr(A) + \Pr(B) \geq \Pr(A \cup$

8 The bound slightly differs from original Hoeffding's bound but has similar convergence rate.

B), we can combine the probability $\Pr(A)$ in equation (3.29) and $\Pr(B)$ in equation (3.31) to obtain the following expression[9]:

$$
\begin{aligned}
&2\Pr\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_{\epsilon}-\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)\right|>\epsilon+\varsigma\right\}\\
&\geq\Pr\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_{\epsilon}-\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,V)\right|-\right.\\
&\qquad\left.\left|\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)-\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,V)\right)\right|>2\epsilon+\varsigma\right\} \qquad (3.32)
\end{aligned}
$$

From the right handside of equation (3.32) and using the bound in (3.30), $\left|\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)-\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,V)\right)\right|$ is the population of

$$
\left|\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})-\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,V)\right|.
$$

9 The explicit derivation is as follows: $2\Pr(A)\geq\Pr(A\cup B)$ implies

$$
\begin{aligned}
&2\Pr\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_{\epsilon}-\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)\right|>\epsilon+\varsigma\right\}\\
&\geq\Pr\left\{\left|\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)-\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_{\epsilon}\right|+\right.\\
&\qquad\left.\left|\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,V)-\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,V)\right)\right|>\epsilon+\varsigma\right\}
\end{aligned}
$$

Using triangular inequality, we have:

$$
|x-y|-|z-q|\leq|x-y+z-q|=|x-q+z-y|\leq|x-q|+|z-y|
$$

where $|z-y|$ is $\left|\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)-\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_{\epsilon}\right|$ and $|x-q|$ is $\left|\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,V)-\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,V)\right)\right|$.

We can conclude[10]

$$\Pr\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_\epsilon-\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|>4\epsilon+2\varsigma\right\} \quad (3.33)$$
$$\leq 4\exp\left(-\frac{T\epsilon^2}{2(T\varsigma+M^*)^2}\right).$$

By Condition 3.3,

$$\left|\frac{1}{T}\sum_{t=1}^{T}c(\mathbf{s}_t,\pi_t,\hat{V})-\frac{1}{T}\sum_{t=1}^{T}c(\mathbf{s}_t,\pi_t,V)\right|\leq 2\epsilon$$

for V satisfying the Markov Perfect Equilibrium. If $\varsigma\to 0$ as $T\to\infty$ then the proof ends.

Thus we need to find the bound in (3.28). In Example 1 and Appendix C of Bousquet and Elisseeff (2002), they give an instruction how to use the subgradient of a constructed convex function to derive a stable bound for a support vector regression model. We follow this idea.

Proof. Let's define the specific form of the regularized function:

$$\begin{aligned}
\Upsilon\hat{V}(\mathcal{S}) =& \frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_\epsilon+\frac{1}{2}\left\|\hat{V}\right\|^2\\
=& \frac{1}{T}\sum_{t=1}^{T}\left|\rho^T\Psi(\mathbf{s}_t)-\pi_t\right|_\epsilon+\frac{1}{2}\left\|\rho^T\Phi(\mathbf{s})\right\|^2\\
=& \frac{1}{T}\sum_{t=1}^{T}\left|\sum_{i=1}^{T}(\alpha_i^*-\alpha_i)\left\langle\Psi(\mathbf{s}_i),\Psi(\mathbf{s}_t)\right\rangle-\pi_t\right|_\epsilon\\
&+\frac{1}{2}\left\|\sum_{i=1}^{T}(\alpha_i^*-\alpha_i)K_{it}\right\|^2,
\end{aligned}$$

10 The explicit derivation is as follows: Let $\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_\epsilon-\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,V)\right|>\epsilon+\varsigma$ be an event A and $\left|\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)-\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,V)\right)\right|>\epsilon$ be an event B. The event A'

$$\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_\epsilon-\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,V)\right|-\left|\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,\hat{V})\right)-\mathbb{E}\left(\tilde{c}(\mathbf{s},\pi,V)\right)\right|>2\epsilon+2\varsigma$$

implies the event B'

$$\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t,\pi_t,\hat{V})\right|_\epsilon-\frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t,\pi_t,V)\right|>4\epsilon+2\varsigma$$

so that $\Pr(B')<\Pr(A')$ while $\Pr(A')<\Pr(A\cap B)$.

where $\Psi(\mathbf{s}) = \Phi(\mathbf{s}) - \beta \sum \Phi(\mathbf{s}')\mathbf{p}(\mathbf{s}'|\mathbf{s}, \mathbf{a})$. In order to optimize this function, we need the functional derivative of $\left|\tilde{c}(\mathbf{s}_j, \pi_j, \hat{V})\right|_\epsilon$ w.r.t. $\hat{V}$, see for the definition of the functional derivative and its associated notation:

$$\begin{aligned}[\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}) &:= \partial \left|\rho^T \Psi(\mathbf{s}_t) - \pi_t\right|_\epsilon \\ &= \begin{cases} 0 & \text{if } \left|\rho^T \Psi(\mathbf{s}_t) - \pi_t\right| \leq \epsilon \\ \frac{(\pi_t - \rho^T \Psi(\mathbf{s}_t))}{\left|\rho^T \Psi(\mathbf{s}_t) - \pi_t\right|} & \text{otherwise.} \end{cases}\end{aligned}$$

It is obvious that $|[\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V})| \leq 1$. The derivative of the regularized function for the original sample $\mathcal{S}$ is:

$$\partial \mathrm{Y} \hat{V}(\mathcal{S}) = \frac{1}{T} \sum_{t=1}^{T} [\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}) + \rho^T \Phi(\mathbf{s}) = 0, \tag{3.34}$$

and for the replacement sample $\mathcal{S}^m$ is:

$$\partial \mathrm{Y} \hat{V}(\mathcal{S}^m) = \frac{1}{T} \sum_{t=1}^{T} [\delta]\tilde{c}(\mathbf{s}_j, \pi_j, \hat{V}_{\mathcal{S}^m}) + \rho_{\mathcal{S}^m}^T \Phi(\mathbf{s}_m) = 0. \tag{3.35}$$

Next, we construct an auxiliary convex function:

$$\begin{aligned}\mathcal{A}(f) = &\left\langle \frac{1}{T} \sum_{t=1}^{T} [\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}) - \frac{1}{T} \sum_{t=1}^{T} [\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}_{\mathcal{S}^m}), f - \rho_{\mathcal{S}^m}^T \Phi(\mathbf{s}) \right\rangle \\ &+ \frac{1}{2} \left\| f - \rho_{\mathcal{S}^m}^T \Phi(\mathbf{s}_m) \right\|^2.\end{aligned}$$

It is obvious that when $f = \rho_{\mathcal{S}^m}^T \Phi(\mathbf{s}_m)$, $\mathcal{A}(\rho_{\mathcal{S}^m}^T \Phi(\mathbf{s}_m)) = 0$. Furthermore, the functional derivative of $\mathcal{A}(f)$ w.r.t. f is

$$\begin{aligned}\partial \mathcal{A}(f) = &\frac{1}{T} \sum_{t=1}^{T} [\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}) - \frac{1}{T} \sum_{t=1}^{T} [\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}_{\mathcal{S}^m}) \\ &+ \left(f - \rho_{\mathcal{S}^m}^T \Phi(\mathbf{s}_m) \right) \\ = &\frac{1}{T} \sum_{t=1}^{T} [\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}) + f.\end{aligned}$$

The second equation uses (3.35).

By (3.34), the minimum of $\mathcal{A}(f)$ is achieved at $\rho^T\Phi(\mathbf{s})$. because $\partial\mathcal{A}(\rho^T\Phi(\mathbf{s}_m)) = 0$ and $\mathcal{A}(f)$ is a convex function. Thus $\mathcal{A}(f) \leq 0$.

$$
\begin{aligned}
&T \times \left\langle \partial \Upsilon \Gamma \hat{V}(\mathcal{S}) - \partial \Upsilon \Gamma \hat{V}(\mathcal{S}^m), \rho^T\Phi(\mathbf{s}) - \rho_{\mathcal{S}^m}^T\Phi(\mathbf{s}_m) \right\rangle \\
&= \sum_{t\neq m}^{T} \left[[\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}) - [\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}_{\mathcal{S}^m}) \right] \times \left[\rho^T\Psi(\mathbf{s}_t) - \rho_{\mathcal{S}^m}^T\Psi(\mathbf{s}_t) \right] \\
&\quad + [\delta]\tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}) \times \rho^T\Psi(\mathbf{s}_m) - [\delta]\tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}_{\mathcal{S}^m}) \times \rho_{\mathcal{S}^m}^T\Psi(\mathbf{s}_m)
\end{aligned} \tag{3.36}
$$

For a convex function $f(a) + (b-a)f'(a) \leq f(b)$ and $f(b) + (a-b)f'(b) \leq f(a)$, then one can obtain $(f'(a) - f'(b))(a-b) \geq 0$. If $\rho^T\Psi(\mathbf{s}_t) > \rho_{\mathcal{S}^m}^T\Psi(\mathbf{s}_t)$ then $[\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}) > [\delta]\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V}_{\mathcal{S}^m})$ in (3.36) by this convex property of ϵ-loss functions. So the first term of equation (3.36) is positive. Therefore, from the regularized auxiliary function, we have:

$$
\begin{aligned}
T \times \left\{ 0 - \frac{1}{2} \left\| \hat{V} - \hat{V}_{\mathcal{S}^m} \right\|^2 \right\} \geq & T \times \left\{ \mathcal{A}(\hat{V}) - \frac{1}{2} \left\| \hat{V} - \hat{V}_{\mathcal{S}^m} \right\|^2 \right\} \\
\geq [\delta]\tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}) \times \rho^T\Psi(\mathbf{s}_m) & - [\delta]\tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}_{\mathcal{S}^m}) \times \rho_{\mathcal{S}^m}^T\Psi(\mathbf{s}_m).
\end{aligned}
$$

By Condition 3.7 (ii) and $\left| [\delta]\tilde{c}(\mathbf{s}, \pi, \hat{V}) \right| \leq 1$,

$$
\begin{aligned}
\frac{T}{2} \left\| \hat{V} - \hat{V}_{\mathcal{S}^m} \right\|^2 \leq & [\delta]\tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}) \times \rho^T\Psi(\mathbf{s}_m) \\
- [\delta]\tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}_{\mathcal{S}^m}) & \times \rho_{\mathcal{S}^m}^T\Psi(\mathbf{s}_m) \leq 4M^*.
\end{aligned} \tag{3.37}
$$

Note that Condition 3.7 implies that the fitting error is Lipschitz continuous, so there is

$$
|\tilde{c}(\mathbf{s}, \pi, \hat{V}) - \tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}_{\mathcal{S}^m})| \leq C_\pi |\hat{V} - \hat{V}_{\mathcal{S}^m}|.
$$

Now we need to derive a bound for $|\hat{V} - \hat{V}_{\mathcal{S}^m}|$ using equation (3.37). By Condition 3.8 and 3.7 (ii), the Cauchy-Schwarz inequality and the kernel property, we have

$$
|\hat{V} - \hat{V}_{\mathcal{S}^m}| \leq \left\| \hat{V} - \hat{V}_{\mathcal{S}^m} \right\| \times \|k\| \leq \kappa \sqrt{\frac{8M}{T}}. \tag{3.38}
$$

Therefore, for any $\mathbf{s}_m \in \mathcal{S}$, $|\tilde{c}(\mathbf{s}, \pi, \hat{V}) - \tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}_{\mathcal{S}^m})|$ will be bounded by $C_\pi \kappa \sqrt{\frac{8M}{T}}$. Use this bound in (3.37), we will have

$$
\frac{T}{2} \left\| \hat{V} - \hat{V}_{\mathcal{S}^m} \right\|^2 \leq C_\pi \left\| \hat{V} - \hat{V}_{\mathcal{S}^m} \right\|
$$

thus $\left\|\hat{V} - \hat{V}_{\mathcal{S}^m}\right\| \leq 2C_\pi\kappa/T$. Substituting this into (3.38), we have

$$|\tilde{c}(\mathbf{s}, \pi, \hat{V}) - \tilde{c}(\mathbf{s}_m, \pi_m, \hat{V}_{\mathcal{S}^m})| \leq \frac{2C_\pi^2\kappa^2}{T}.$$

The factor decreases to zero when $T \to \infty$. By the BE Theorem, equation (3.33) turns to

$$\begin{aligned}\Pr\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{c}(\mathbf{s}_t, \pi_t, \hat{V})\right|_\epsilon - \frac{1}{T}\sum_{t=1}^{T}\tilde{c}(\mathbf{s}_t, \pi_t, V)\right| > 4\epsilon + 2\varsigma\right\} \\ \leq 4\exp\left(-\frac{T}{2}\left(\frac{\epsilon}{M^*}\right)^2\left(1 + \frac{1}{M^*}(C_\pi\kappa)\right)^{-2}\right),\end{aligned}$$

where $\varsigma = 2C_\pi^2\kappa^2/T$. □

Proof of Theorem 3.12

When ϵ goes to zero, the soft-margins of the kernel-based approximation ζ and ζ^* are slack. We set up the new programming problem and then show that the kernel-based policy iteration algorithm coincides with the exact policy iteration in the limit.

Proof. Let ζ and ζ^* be equal to zero in (3.8), the optimization problem is then reduced to

$$\begin{aligned}&\min_\rho \frac{1}{2}\|\rho\|^2 \\ &s.t. \quad \pi = \rho^T\Psi(\mathbf{s})\end{aligned} \tag{3.39}$$

where $\Psi(\mathbf{s}) = \Phi(\mathbf{s}) - \beta\sum\Phi(\mathbf{s}')\mathbf{p}(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ and the constraint is the fitting error in (3.10). As in (3.11), taking the partial derivative of the Lagrangian with respect to ρ, we have

$$\partial L/\partial\rho = \rho - \sum_{t=1}^{T}\alpha_t^0\Psi(\mathbf{s}_t) = 0,$$

or in the matrix expression $\rho = \boldsymbol{\alpha}^T\Psi$ by stacking α_t and $\Psi(\mathbf{s}_t)$. Similarly, we can set up the dual problem such that

$$\max_\alpha -\frac{1}{2}\boldsymbol{\alpha}^T K^*\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\pi, \tag{3.40}$$

with $\boldsymbol{\alpha} \geq 0$ and kernel matrix $K^* = \langle\Psi, \Psi\rangle$. This is a standard linear quadratic objective function. The derivative of (3.40) with respect to α gives linear system $K^T\boldsymbol{\alpha} = \boldsymbol{\pi}$ and $\boldsymbol{\alpha}$ can be solved by a simple inversion.

When $\boldsymbol{\alpha}$ has a unique solution and does not have a zero element, by the dual space property (Slater condition) we know the constraint $\pi = \rho^T \Psi(s)$ is strictly satisfied.

The uniqueness and non-zero $\boldsymbol{\alpha}$ can be proved as follows. Note that K is invertible and positive definite. In the expression of Ψ, $\beta < 1$ and $|\mathbf{p}(\mathbf{s}'|\mathbf{s},\mathbf{a})| < 1$ imply that $\Psi^T\Psi$ has positive eigenvalues. Hence $\boldsymbol{\alpha}$ is unique. Substitute $K^T\boldsymbol{\alpha} = \boldsymbol{\pi}$ into (3.40), equation (3.40) reduces to $\boldsymbol{\alpha}^T K\boldsymbol{\alpha}/2$. If we suppose $\alpha_1 = 0$, it implies (3.40) has $\pi_1 \leq 0$. Since $\pi_t = [\sum \alpha_t \Psi(\mathbf{s}_t)]\Psi(\mathbf{s}) > 0$ for any t, contradiction. Hence all elements in $\boldsymbol{\alpha}$ should be non-zero.

By the result of Theorem 1 in Aguirregabiria and Mira (2007), we have

$$V^*(\mathbf{s}) = \sum_{\mathbf{a}\in\mathcal{A}} F^*(\mathbf{a})[\pi^*(\mathbf{a}) + e^*(\mathbf{a})] + \beta \sum_{\mathbf{s}'} V^*(\mathbf{s}')\mathbf{p}^*(\mathbf{s}'|\mathbf{s}).$$

The solution of the exact Bellman equations fits our request. □

4

GEOMETRIC INTERPRETATIONS FOR CONSTRAINED GMC AND GEL

Constrained optimization, as a general mathematical tool, can narrow the potential parameter space and refine irregular problems in statistical models. In econometrics, this tool often consists of moment constraints and various criterion functions[1], such as Mahalanobis distance (GMM Hansen, 1982), log-likelihood ratio (Empirical Likelihood, hereafter EL, Owen, 2001), and Kullback-Leibler divergence (Kitamura and Stutzer, 1997). Smith (1997) and Newey and Smith (2004) show that using a general representation, many criterion functions with moment constraints can be covered in the framework of Generalized Empirical Likelihood (GEL). The statistics of members in GEL share similar asymptotic properties. A class competing with GEL is the Generalized Minimum Contrast (GMC) class which has been discussed in Kitamura (2006). This chapter will consider a subclass in GMC whose members have similar properties as those in GEL. By geometry techniques, we connect the representations of GEL and GMC through polyhedral approximations and then we show that a sub-Sobolev class, called $\mathcal{W}$-class, in GMC will have similar properties as GEL and will obtain standard weak convergence result. The result simply induces the Fisher type efficiency.

A discussion about the connection between GEL and GMC has been given by Kitamura (2006). Kitamura (2006) shows that the standard inferential problems in GEL, e.g. EL, Exponential Tiling or continuous updating GMM, can be also presented in the GMC setup. GMC has a primal-dual representation, while GEL is based on the dual parameters of the moment constraints. It implies that GMC has a standard mathematical programming representation as its primal form. This representation is impor-

1 The word "criterion function" has the same meaning as the objective function in constrained optimization problems. Thereafter, these two terminologies are used simultaneously.

tant for both theoretical analyses and practical implementations. However, the Fisher type efficiency of GEL is not derivable if one concentrates on the general form of GMC. The reason is that GMC consists of some criterion functions which are incomparable to those in GEL in the Fisher type information metric. Then what could be the primal representation of GEL? This chapter is intended to give an explicit clue.

4.1 THE MODEL

We consider a criterion function $\rho(\cdot,\cdot)$ that is a contrast function measuring the discrepancy between its inputs. The inputs are the densities or the likelihoods of a general probability family $\mathcal{P}$ such that the Radon–Nikodym derivative exists with respect to (w.r.t.) the counting measure μ:

$$\mathcal{P}_\mu := \left\{ pd\mu | \, p = \frac{dP}{d\mu}, \, P \in \mathcal{P} \right\}$$

So $\rho(\cdot,\cdot) : \mathcal{P}_\mu \times \mathcal{P}_\mu \mapsto \mathbb{R}$. It is not necessary for the contrast function ρ to be a metric. If $\mathcal{P}$ consists of parametric families, we will use the notation $\mathcal{P}_\theta := \{p_\theta d\mu \in \mathcal{P}_\mu, \theta \in \mathbb{R}^d\}$ and let p_0 denote the true density. The minimum dissimilarity between p_θ and p_0 is attainable at $\theta = \theta_0$ w.r.t. $\rho(\cdot,\cdot)$.

The criterion functions in this chapter belong to a general class, $\mathcal{W}$ which is a subspace of the *Sobolev space* $\mathcal{W}^{k,2}$ with $k \geq 2$. To be more specific, any $\rho \in \mathcal{W}^{k,2}$, satisfies:

$$\left(\sum_{j=0}^{k} \left\| \rho^{(j)} \right\|^2 \right)^{1/2} < \infty,$$

where $\rho^{(j)}$ means the j-th (functional) derivative[2] of ρ and $\|\cdot\|$ is L^2-norm. The subspace $\mathcal{W}$-class that we will use only involves those ρs that have the same order as squared Hellinger distance in the sense that for $p_\theta, p_0 \in \mathcal{P}_\theta$:

$$\lim_{\theta \to \theta_0} \frac{\rho(p_\theta, p_0)}{\rho^{(H)}(p_\theta, p_0)} = \frac{\rho(p_{\theta_0}, p_0)}{\rho^{(H)}(p_{\theta_0}, p_0)} = c,$$

where c is a constant and $\rho^{(H)}(p, g)$ is squared Hellinger distance:

$$\rho^{(H)}(p, g) := \frac{1}{2} \int (\sqrt{p} - \sqrt{g})^2 d\mu.$$

2 The notation $\rho^{(j)}(x,y)$ means that $\frac{\partial^j \rho(x,y)}{\partial x^p y^q}$ with any positive p, q and $p+q=j$.

We assume a sample space $\mathcal{X}$. The model specifies d moment restrictions $\mathbb{E}[m(X,\theta)] = 0$ for a unique $\theta_0 \in \mathbb{R}^d$ where $\mathbb{E}[\cdot]$ is taken w.r.t. $p_0 d\mu$ so the model is exactly identified[3]. The estimator is exactly the solution[4] to the moment condition $\sum_{i=1}^n m(x_i,\theta) = 0$. With a sample of n observations we can define the empirical support $\mathcal{X}_n$ as all those points in $\mathcal{X}$ that have been observed in the sample. On $\mathcal{X}_n$ we can define an empirical measure using the Dirac function δ_n such that $\delta_n(x) = 1$ for $x \in \mathcal{X}_n$ and $\delta_n(x) = 0$ otherwise. We consider a mathematical programming problem:

$$(M) := \begin{cases} \min_{\mathbf{p}_n,\theta} & \rho(p_\theta(x), \delta_n(x)n^{-1}) \text{ for any } x \in \mathcal{X}_n \\ s.t. & \mathbf{p}_n = (p_1,\ldots,p_n)^T \in \mathbb{S}_{n-1}, \sum_{i=1}^n p_i m(x_i,\theta) = \mathbf{0}. \end{cases}$$

where $p_\theta(x_i) = p_i$, $\mathbb{S}_{n-1}$ is a regular $(n-1)$-simplex $\{\mathbf{p}_n : \sum_i^n p_i = 1, p_i \geq 0 \text{ for all } i\}$. For every θ, $p_\theta(x_i) = p_i$ is the (weighted empirical) density for x_i under θ. For simplicity, throughout the chapter, we assume that the model (M) satisfies *common regularity assumptions*.

Assumption 4.1.
(i) $N^{-1}\sum_{i=1}^N [m(x_i,\theta)^T m(x_i,\theta)]$ *has finite variance.*
(ii) $\lambda^T m(x_i,\theta)$ *has finite variance where* λ *is introduced in* (4.1).
(iii) $m(\cdot,\theta)$ *belongs to the P-Donsker class.*
(iv) $\Theta \subset \mathbb{R}^d$ *is compact.*
(v) $m(X,\theta)$ *is continuous w.r.t.* θ.

Remark 4.2. The $\mathcal{W}$-class is a sub-class in the GMC. The smoothness requirement of ρ is a crucial feature because the smoothness is a necessary condition for defining a classical information metric in the sense of Fisher. There are many other options for measuring the divergence of distributions. For example, f-divergence is a rather broad class in GMC. However,

3 Note that even if $\theta = \theta_0$, p_{θ_0} is not necessarily equivalent to p_0, since p_{θ_0} is simply a pseudo true density of $m(X,\theta_0)$ and θ_0 itself does not fully capture the parameter information of the distribution of X or $m(X,\theta_0)$. For example, if p_0 is Poisson or log-Gaussian, the implied probability of EL or ET

$$\tilde{p}_{\theta_0}(x_i) = \frac{1}{n}\frac{1}{1+\lambda^T m(x_i,\theta_0)}, \text{ or } \tilde{p}_{\theta_0}(x_i) = \frac{e^{\lambda^T m(x_i,\theta_0)}}{\sum_i e^{\lambda^T m(x_i,\theta_0)}}$$

will not be equivalent to p_0 no matter how many $m(X,\theta_0)$ are used.

4 This chapter is intended to discuss the connection between GEL and GMC. While the exact identification case does diminish the advantages of using over-identified moment constraints in GEL or GMC, the implied density and the analysis based on this implied density do not rely on whether the constraints are over-identified or not.

f-divergence cannot guarantee the smoothness of ρ, e.g., the total variation distance, a special case in the f-divergence. We consider $\mathcal{W}$-class instead of f-divergence because of the important role of the smoothness condition.

Remark 4.3. Many essential contrast functionals belong to $\mathcal{W}$-class. Besides Hellinger distance, Likelihood ratio and χ^2, there are Kullback-Leibler divergence,

$$\rho^{(KL)}(p, g) := \int \left(\log \frac{p}{g}\right) p d\mu$$

Mahalanobis distance,

$$\rho^{(M)}(p, g) := \int (p - g)^2 w^{-1} d\mu$$

Jeffreys divergence, Chernoff information divergence, etc. Eguchi (1985) introduces this class from a differential geometry point of view in statistics.

4.2 DUAL REPRESENTATION ON CONVEX BODIES: CONSTRAINTS

Neither closed form solutions nor explicit expressions of the optimal $p_\theta(x)$ in (M) are available. However, one can get the first-order-condition (FOC) by the Lagrangian method:

$$\mathcal{L}_n = \rho(p_\theta(x), \delta_n(x)n^{-1}) - \lambda^T \left(\sum_{i=1}^n p_i m(x_i, \theta)\right) - \zeta(1 - \sum_i^n p_i). \tag{4.1}$$

Taking derivative of (4.1) w.r.t. $\mathbf{p}_n$, one has the FOC:

$$\nabla_i \rho(p_\theta(x), \delta_n(x)n^{-1}) = \lambda^T m(x_i, \theta) - \zeta, \tag{4.2}$$

where ∇_i is gradient w.r.t. p_i. If we multiply $\{p_1, \ldots, p_n\}$ on both sides and add up these equations, we have

$$\sum_{i=1}^n p_i \nabla_i \rho(p_\theta(x), \delta_n(x)n^{-1}) = -\zeta,$$

where $\lambda^T m(x_i, \theta)$ drops out because $\sum_i p_i m(x_i, \theta) = 0$. Just like EL, if we substitute the expression for ζ into (4.2), we have:

$$\nabla_i \rho(p_\theta(x), \delta_n(x)n^{-1}) - \sum_{i=1}^n p_i \nabla_i \rho(p_\theta(x), \delta_n(x)n^{-1}) = \lambda^T m(x_i, \theta). \tag{4.3}$$

Equation (4.3) shows that the optimal p_θ, the solution of this nonlinear differential equation, is a function of λ. In general, there is no closed form expression for $p_\theta(\lambda)$. A straightforward *conjecture* about the solution of (4.3) is that $p_\theta(\lambda)$ is a function[5] of $\lambda^T m(x_i,\theta)$ only. The following Proposition is to formalize this conjecture.

Proposition 4.4. *The density $p_\theta(\lambda)$ in the problem (M) only depends on $\lambda^T m(x_i,\theta)$ for any $1 \leq i \leq n$.*

One can easily relate this result to GEL (Smith, 1997; Newey and Smith, 2004) whose definition is based on a criterion function of $\lambda^T m(x_i,\theta)$. The setup of GEL is:

$$\hat{\theta} = \arg\min_\theta \max_\lambda \left[\frac{1}{n} \sum_{i=1}^{n} \tilde{\rho}(\lambda^T m(x_i,\theta)) \right]$$

where $\tilde{\rho}(\cdot)$ is strictly concave and $\tilde{\rho}^{(1)}(0) = \tilde{\rho}^{(2)}(0) = -1$. The notation $\tilde{\rho}^{(j)}(0)$ means the j-th derivative of $\tilde{\rho}$ at zero.

As the optimal solution $p_\theta(\lambda)$ in the problem (M) only depends on $\lambda^T m(x_i,\theta)$, we can see that problem (M) is exactly the same as GEL except a different definition of the objective function ρ. However, since ρ is in the $\mathcal{W}$-class, it is smooth, differentiable, continuous and approximately quadratic. Hence, the fact that the objective function is in the $\mathcal{W}$-class plays no additional role other than in GEL in proving consistency. The consistency result for GEL (Newey and Smith, 2004) or for minimum Hellinger distance (Kitamura et al., 2009) can be applied directly to the estimator in the $\mathcal{W}$-class. As a consequence, we have following Proposition which we state without proof.

Proposition 4.5. *For any estimator $\hat{\theta}$ such that*

$$\rho(p_{\hat{\theta}}(x), \delta_n(x) n^{-1}) \leq \rho(p_{\theta_0}(x), \delta_n(x) n^{-1}) + o_p(1),$$

we have for any $\epsilon > 0$, $\Pr(\|\hat{\theta} - \theta_0\| \leq \epsilon) \to 1$.

Remark 4.6. We briefly describe the geometric meaning of Proposition 4.4 but for details, please refer to the proof in the appendix. The geometric role of the moment constraints in (M) is to construct a convex hull of $m(X,\theta)$, $\text{conv}(m(X,\theta))$. The convex hull is approximated by some geometric objects. Mathematically speaking, the border of $\text{conv}(m(X,\theta))$ is empirically approximated by the polyhedron, $\mathcal{P}_m = \{(x_1,\ldots,x_n,\mathbf{b}) | \lambda^T m(x_i,\theta) \leq$

5 Because fundamental solution of the ordinary differential equation $dX/dt = g(X)$ is a function of $g(X)$.

$\mathbf{b}\}$ for the given sample, where $\mathbf{b}$ is chosen optimally. The dual problem of (M) is a linear programming problem of the polyhedron set $\mathcal{P}_m$. Then the optimal solution of this linear programming problem will give a unique representation of $\text{conv}(m(X,\theta))$ and the solution will only depend on $\lambda^T m(x_i,\theta)$. The density $p_\theta(x)$ is the *generator* of $\text{conv}(m(X,\theta))$, so the optimal $p_\theta(x_i)$ will also only depend on $\lambda^T m(x_i,\theta)$.

4.3 WEAK CONVERGENCE ON THE UNIT SPHERE: CRITERION FUNCTIONS

The criterion function ρ plays an important role in weak convergences. The distance between two points in $\mathcal{P}_\theta$ measures the amount of their dissimilarity. The functional form of ρ defines the criterion of such a measurement. We will show that the geometric analysis of weak convergence for the $\mathcal{W}$-class is simple and intuitive.

Given a hypothesis testing problem $H_0 : \theta = \tilde{\theta}$, the Hellinger distance on $\mathcal{P}_\theta$ between p_0 and the empirical distribution $\delta_n(x)n^{-1}$ is:

$$\rho^{(H)}(p_{\tilde{\theta}}(x),\, \delta_n(x)n^{-1}) = -\frac{1}{2}\sum_{i=1}^{n}\left[\sqrt{p_i^*} - \sqrt{\frac{1}{n}}\right]^2,$$

where the notation p_i^* is the solution of problem (M) such that $p_{\tilde{\theta}}(x_i) = p_i^*$ given $\theta = \tilde{\theta}$. Our geometric interpretation of $\rho^{(H)}$ in (M) is the metric distance over a unit sphere $\mathcal{S}^{n-1}$, a n-dimensional Hilbert space $\mathcal{H}$. Note that any square-root likelihood $\xi_i := (p_i^*)^{\frac{1}{2}}$ satisfies the normalizing condition $\sum_i^n \xi_i^2 = 1$ and is nonnegative and is located within $[0,1]$. Thus, ξ can be regarded as a unit vector in the Hilbert space $\mathcal{S}^{n-1}$. The empirical square-root measure $\delta_n(x)n^{-\frac{1}{2}}$, considered as a special case of $(p_i^*)^{\frac{1}{2}}$, also belongs to this space. For any two sequences $\xi^{(1)}$ and $\xi^{(2)}$, the Hellinger distance induces an inner product of this space

$$\cos\beta := \langle \xi_a, \xi_b \rangle = \sum_i^n \xi_i^{(1)}\xi_i^{(2)} = 1 - \frac{1}{2}\sum_{i=1}^{n}\left[\sqrt{\xi_i^{(1)}} - \sqrt{\xi_i^{(2)}}\right]^2$$

which is the so called *Hellinger affinity*. In addition, the double-angle formula $\cos\beta = 1 - 2\sin^2\frac{\beta}{2}$ implies that:

$$4\sin^2\frac{\beta}{2} = \sum_{i=1}^{n}\left[\sqrt{\xi_i^{(1)}} - \sqrt{\xi_i^{(2)}}\right]^2 = -2\rho^{(H)}(p_{\tilde{\theta}}(x),\, \delta_n(x)n^{-1}).$$

We should reiterate that p_θ is a pseudo true density so even $\tilde{\theta} = \theta_0$, $p_{\tilde{\theta}}$ is not necessarily equivalent to p_0.

It is obvious that the angle β can be interpreted as a distance between two probability distributions on $\mathcal{S}^{n-1}$ equipped with the Hellinger metric. The maximal possible distance, corresponding to orthogonal sequences, is given by $\beta = \pi/2$. The double-angle formula clearly shows that the distribution of $\cos\beta$ is equivalent to that of $4\sin^2(\beta/2)$.

Theorem 4.7. *If the hypothesis $H_0 : \theta = \tilde{\theta}$ is true, the statistics $\rho(p_{\tilde{\theta}}(x), \delta_n(x)n^{-1})$ in the $\mathcal{W}$-class converges in distribution to a χ^2_d-distributed random variable when $n \to \infty$.*

Remark 4.8. From basic trigonometrics, we know that when the angle β is very small, $\sin^2\beta \approx \beta^2$. Thus Hellinger distance implies that when $\rho(\cdot,\cdot)$ is small, the distance in the $\mathcal{W}$-class can be described in terms of β^2. Hellinger distance is also an important concept when one defines Fisher's information metric in the spherical geometry. The inner product $\cos\beta$ could be written as $\langle \xi_a, \xi_b \rangle_{g_{ab}} := \int g_{ab}\xi_a\xi_b d\mathcal{S}^{n-1}$ where $d\mathcal{S}^{n-1}$ is volume measure on the sphere associated with the Riemannian metric g_{ab}. If the derivatives of ξ_a and ξ_b are available[6], then $4g_{ab}\partial\xi_a\partial\xi_b$ is the Fisher information metric which can be used to see whether the efficiency bound can be attained on this sphere $\mathcal{S}^{n-1}$.

4.4 CONCLUSION

In this chapter, we show that the GEL problem has a dual representation, a constrained GMC problem for the $\mathcal{W}$ contrast functional class. This connection is established from a geometric perspective. The dual representation transfers the task of estimating implied densities for GEL to a task of solving a standard mathematical programming problem. We also show that the statistics in the $\mathcal{W}$-class share the same first order statistical properties with GEL.

6 The derivative $\partial\xi$ is $d\sqrt{p} = \sqrt{p}d\ell$ where $d\ell$ is the functional derivative of the empirical log-likelihood in semi-parametric models. $(\partial_1\ell, \ldots, \partial_r\ell)^T$ is a co-ordinate basis on the tangent space of $\mathcal{P}_\theta$, $T\mathcal{P}_\theta$, hence one can use it to construct an inner product on $T\mathcal{P}_\theta$.

APPENDIX TO CHAPTER 4

Let $m(\mathbf{X}_n,\theta) = (m(x_1,\theta),\ldots,m(x_n,\theta))$ where $m(x_i,\theta)$ is a $d \times 1$ vector.

Proof of Proposition 4.4

Proof. (*Sufficient condition*) The constraint constructs a set of $m(\mathbf{X}_n,\theta)$, the convex hull of a finite set $\{p_1,\ldots,p_n\}$,:

$$\text{conv}(m(\mathbf{X}_n,\theta)) := \left\{ \sum_{i=1}^{n} p_i m(x_i,\theta) \,|, \sum_{i=1}^{n} p_i = 1, p \succeq 0 \right\},$$

which is called a finitely generated convex set. The symbol $\succeq$ defines a vector inequality in $\mathbb{R}^n$ such that $p_i \geq 0$ for $i = 1,\ldots n$. The set $\text{conv}(m(\mathbf{X}_n,\theta))$ is a *polyhedron* $\mathcal{M}$ (Theorem 19.1 Rockafellar, 1996), namely an intersection of finitely many closed half spaces:

$$\text{conv}(m(\mathbf{X}_n,\theta)) = \mathcal{M} := \left\{ m(x,\theta) \,|\, \lambda_p^T m(x,\theta) \succeq \mathbf{b} \right\},$$

where λ_p^T and $\mathbf{b} = (b_1,\ldots,b_n)^T$ define the half-space for each $m(x_i,\theta)$ such that $\{\lambda_p^T m(x_i,\theta) \geq b_i\}$. Define a function

$$\delta_{\mathcal{M}}(\theta) := \mathbb{I}\{m(\mathbf{X}_n,\theta) \in \mathcal{M}\}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function. The notation

$$\{m(\mathbf{X}_n,\theta) \in \mathcal{M}\} := \{m(x_1,\theta),\ldots,m(x_n,\theta) \in \mathcal{M}\}$$

means that $m(x_i,\theta)$ belongs to $\mathcal{M}$ for $0 < i \leq n$. Laplace's inequality states that $\mathbb{I}\{y\} \leq \exp(y)$ for any $y \geq 0$. The polyhedron set introduces an exponential bound such that:

$$\delta_{\mathcal{M}}(\theta) \preceq \exp\left(\lambda_p^T m(\mathbf{X}_n,\theta) - \mathbf{b}\right).$$

Taking expectation w.r.t. $\mathbf{X}_n$, we have the Chernoff bound

$$\Pr\left(m(\mathbf{X}_n,\theta) \in \mathcal{M}\right) \preceq \mathbb{E}\exp\left(\lambda_p^T m(\mathbf{X}_n,\theta) - \mathbf{b}\right), \tag{4.4}$$

which implies the probability p_θ is controlled by $\lambda_p^T m(\mathbf{X}_n,\theta)$ and $\mathbf{b}$.

Now we need to relate the tangent parameter λ_p in the polyhedron with the multiplier λ in optimization of $\rho(p_\theta(x), \delta_x n^{-1})$.

(*Necessary condition*) By the differentiability of the $\mathcal{W}$-class, the (functional) linearization of $\rho(p_\theta(x), \delta_x n^{-1})$ at the point $p^\#$ is

$$\begin{aligned}\rho(p_\theta(x), \delta_x n^{-1}) =& \rho(p^\#, \delta_x n^{-1}) + (p_\theta - p^\#)^T \nabla \rho(p_\theta, \delta_x n^{-1}) \\ &+ O_p(\|p_\theta - p^\#\|^2).\end{aligned} \tag{4.5}$$

By FOC (4.2), we can replace the gradient vector $\nabla \rho(p_\theta, \delta_x n^{-1})$ in (4.5) by $(\lambda^T m(X, \theta) - \zeta)$. Then within a small region of $p^\#$ where $O_p(\|p_\theta - p^\#\|^2)$ is negligible, minimizing $\rho(p_\theta(x), \delta_x n^{-1})$ in this region is equivalent to minimizing the linear approximation w.r.t. λ, θ and ζ:

$$\mathcal{Q}(p^\#) := \inf_{\lambda,\theta,\zeta} \left\{ \rho(p^\#, \delta_x n^{-1}) + (p_\theta - p^\#) \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] \right\}, \tag{4.6}$$

$$= \inf_{\lambda,\theta,\zeta} \left\{ -p^\# \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \mathcal{Q}^*(\lambda, \theta, \zeta) \right\}, \tag{4.7}$$

where $\mathcal{Q}^*(\lambda, \theta, \zeta) = p_\theta \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \rho(p^\#, \delta_x n^{-1})$ is the conjugate[7] of $\mathcal{Q}(p^\#)$ in the region around $p^\#$. Then by *Legendre–Fenchel transformation* (conjugate functional)[8], the dual programming of problem (4.6) is:

$$\begin{aligned}[\mathcal{Q}(p^\#)]^* =& \mathcal{Q}^*(\lambda, \theta, \zeta) = \inf_{p^\#} \left\{ -p^\# \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \mathcal{Q}(p^\#) \right\}, \\ =& \inf_{p^\#} \Bigg\{ -p^\# \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \\ & \underbrace{\inf_{\lambda,\theta,\zeta} \left\{ -p^\# \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \mathcal{Q}^*(\lambda, \theta, \zeta) \right\}}_{(i)} \Bigg\}.\end{aligned}$$

7 Mathematically, a nonlinear programming problem can be pinned down as many linear programming problems using the linearization technique. In these linear programming problems, finding an optimal linear approximation ($\mathcal{Q}^*(\lambda, \theta, \zeta)$) is a dual or a conjugate problem of finding the optimal parameters for such a linear approximation ($\mathcal{Q}(p^\#)$). The construction of equations (4.6) and (4.7) is a functional *Legendre transformation*.

8 Kitamura (2006) gives a general introduction of Legendre-Fenchel transformation in context of EL and GMC.

We write the expression in a simpler way using a slack variable ω for (i) in the above equation:

$$\inf_{p^{\#},\omega} \quad \left\{ p^{\#} \left[\lambda^T m(\mathbf{X}_n,\theta) - \zeta \right] + \omega \right\} \tag{4.8}$$

$$\text{subject to} \quad p^{\#} \left[\lambda^T m(\mathbf{X}_n,\theta) - \zeta \right] - \mathcal{Q}^*(\lambda,\theta,\zeta) \succeq \omega. \tag{4.9}$$

Constraint (4.9) can be written as $\lambda^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}^*$ where

$$\mathbf{b}^* = (p^{\#})^{-1}(\omega + \mathcal{Q}^*(\lambda,\theta,\zeta)) + \zeta.$$

The polyhedral $\mathcal{M}$ of $m(\mathbf{X}_n,\theta)$ is $\{m(\mathbf{X}_n,\theta) \,|\lambda_p^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}\}$. Constraint (4.9) sets a polyhedral shape using $\lambda^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}^*$. The problem becomes

$$(M') := \begin{cases} \inf_{p^{\#},\omega} & \left\{ p^{\#} \left[\lambda^T m(\mathbf{X}_n,\theta) - \zeta \right] + \omega \right\} \\ \text{subject to} & \lambda^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}^*. \end{cases}$$

The optimal $p^{\#}$ and ω in (M') imply optimal edges of the polyhedron $\lambda^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}^*$. Since the value of $\mathbf{b}$ in (4.4) is arbitrary, we set $\mathbf{b} \equiv \mathbf{b}^*$, then the multiplier λ is equivalent to the tangent parameter λ_p. Except tuning parameter $p^{\#}$ and ω, the variables in (M') are $\lambda^T m(\mathbf{X}_n,\theta)$, $\mathcal{Q}^*(\lambda,\theta,\zeta)$ and ζ. By definition, $\mathcal{Q}^*(\lambda,\theta,\zeta)$ depends on $\lambda^T m(\mathbf{X}_n,\theta)$ and ζ. From the FOC, ζ depends on $\lambda^T m(\mathbf{X}_n,\theta)$. Hence for problem (M'), $\lambda^T m(\mathbf{X}_n,\theta)$ is the only necessary element and $p^{\#}$ will depend on $\lambda^T m(\mathbf{X}_n,\theta)$ only. Note that $p^{\#}$ is the conjugate of p_θ.

By Chernoff bound (4.4) and (M'), the density p_θ will only depend on $\lambda^T m(\mathbf{X}_n,\theta)$. □

Proof of Proposition 4.5

Proof. The consistent result of GEL (Newey and Smith, 2004) can be directly applied to the estimators in $\mathcal{W}$-class. □

Proof of Theorem 4.7

Proof. Completeness of $\mathcal{S}^{n-1}$ (or $\mathcal{H}$) states that every subsequence in $\mathcal{P}_\mu$ equipped with $\rho^{(H)}$ has a limiting point. The consistency result shows that any estimated $\{p_i^*\}_{i\leq n}$ from $\mathcal{W}$-class will converge to $1/n$. All metrics or divergences $\rho(\cdot,\cdot)$ in $\mathcal{W}$-class have the same leading-term as $\rho^{(H)}$. If we consider all the subsequences $\{p_i^*\}_{i\leq n}$ from $\mathcal{W}$-class with the limit point $1/n$

in $\mathcal{S}^{n-1}$ equipped with $\rho^{(H)}$, the subsequences cover all members of p^*s in $\mathcal{W}$-class. In other words, because of completeness and $\mathcal{W}$-class, two sequences of $(p_1^*, \dots p_n^*)$ can be studied in a unified framework even if they are generated by different criterion functions in the problem (M).

We only need to study the convergence result of Hellinger affinity:

$$\begin{aligned}\cos\beta &= \sum_{i=1}^{n} \sqrt{p_i^*} \cdot \sqrt{\frac{1}{n}} = \sum_{i=1}^{n} \frac{1}{n} \left[1 + n\left(p_i^* - \frac{1}{n}\right)\right]^{\frac{1}{2}}, \\ &= \sum_{i=1}^{n} \frac{1}{n} \left[1 + \frac{1}{2} n \left(p_i^* - \frac{1}{n}\right) - \frac{1}{8} n^2 \left(p_i^* - \frac{1}{n}\right)^2 \right. \\ &\quad \left. + O_p\left(|np_i^* - 1|^3\right)\right],\end{aligned}$$

where remainder term will be dropped out asymptotically. The second line is the binomial series expansion because $|n(p_i^* - \frac{1}{n})| \leq 1$ for all i. By consistency, $\lim_{n\to\infty}(p_i^* - \frac{1}{n}) = 0$ for all i. The expression becomes

$$\cos\beta = 1 - \frac{1}{8} \sum_{i=1}^{n} n \left(p_i^* - \frac{1}{n}\right)^2 + o_p(1)$$

the second term is nothing but Neyman's χ^2, also known as Euclidean log-likelihood[9]. Thus, by definition

$$2\sin^2\frac{\beta}{2} = \frac{1}{8} \sum_{i=1}^{n} n \left(p_i^* - \frac{1}{n}\right)^2 \sim \frac{1}{8}\chi_d^2.$$

This weak convergency result will hold for $p_{\tilde{\theta}}$ of any member in the $\mathcal{W}$-class. □

9 Euclidean log-likelihood belongs to the so called Cressie-Read family, $\mathrm{CR}(k) = [2/(k^2+k)] \sum_i^n [n(n/p_i)^k - 1]$, for $k = -2$. Please refer to Owen (Chapter 3.15 and 3.16 2001) for more discussion.

SOME MATHEMATICAL FOUNDATIONS

The goal of this Appendix section is to collect all the basic ingredients necessary for an understanding of the EL developments that follow based on the methodology of mathematical programming and numerical analysis. In order to maintain an accessible level introduction, the material is presented with a minimal amount of mathematical rigor. The mathematical symbols are used specifically for this section.

POLYNOMIALS

The most important ingredients in this section are polynomial chaos, a term coined by Nobert Wiener in 1938 in his work studying the decomposition of Gaussian stochastic processes. We will review the basics of orthogonal polynomials, which play a central role in modern optimization theory. The material is kept to a minimum to satisfy the needs of this thesis. More in-depth discussions of the properties of orthogonal polynomials can be found in many standard books such as Zeidler (1995); Sawyer (2010).

A general polynomial of degree k takes the form

$$Q_k(x) = a_k x^k + a_{k-1} x^{k-1} + \cdots + a_1 x + a_0,\ a_k \neq 0,$$

where a_k is the leading coefficient of the polynomial. Let $\mathbb{Z}^+$ be the set of nonnegative integers. A system of $\{Q_k(x), k \in \mathbb{Z}^+\}$ is an orthogonal system of polynomials with respect to some real positive measure α if the following orthogonality relations hold:

$$\int_{\mathcal{S}} Q_i(x) Q_j(x) d\alpha(x) = \gamma_i \delta_{ij}, \quad i, j \in \mathbb{Z}^+$$

where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$ and $\mathcal{S}$ is the support of the measure α, and γ_i are positive constants often termed normalization constants such that

$$\gamma_i = \int_{\mathcal{S}} Q_i^2(x) d\alpha(x), \quad i \in \mathbb{Z}^+.$$

If $\gamma_i = 1$, the system is orthonormal. Let $\mathbb{P}_k$ be the linear space of polynomials of degree at most k,

$$\mathbb{P}_k = \text{span}\{x^k : k = 0, 1, \ldots, k\}.$$

We begin with a classical Theorem by Weierstrass in approximation theory.

Theorem. *(Weierstrass) Let I be a bounded interval and $\bar{I}$ be the closure of I , let f be continuous on I. Then, for any $\epsilon > 0$, we can find $k \in \mathbb{Z}^+$ and $p \in \mathbb{P}_k$ such that*

$$|f(x) - p(x)| < \epsilon, \qquad \forall x \in \bar{I}.$$

We skip the proof here. Interested readers can find the details in various analysis books, for example, Conway (1990). Note the f is continuous so this Theorem states that any continuous function in a bounded closed interval can be uniformly approximated by polynomials. A natural focus in optimization is to see whether, among all the polynomials of degree less than or equal to a fixed integer k, it is possible to find one that best approximates a given continuous function f uniformly in $\bar{I}$. In other words, we would like to study the existence of $\phi_k(f) \in \mathbb{P}_k$ such that

$$\|f - \phi_k(f)\| = \inf_{\psi \in \mathbb{P}_k} \|f - \psi\|.$$

This problem admits a unique solution. The optimization or mathematical programming problem now is the problem of finding the polynomial class of best uniform approximation of f:

$$\lim_{k \to \infty} \|f - \phi_k(f)\| = 0.$$

For every fixed k, there is an approximation $\phi_k(f)$ of f. When k increases, the approximation becomes better and better. For implementation, it is better to focus on a specific norm $\|\cdot\|$ so that the optimization problem is formulated in terms of a specific normed space. Let's consider the weighted L^2 space:

$$L^2_w(I) := \left\{ v : I \to \mathbb{R} \,\middle|\, \int_I v^2(x) w(x) dx < \infty \right\}$$

with the inner product

$$\langle u, v \rangle_{L^2_w(I)} = \int_I u(x) v(x) w(x) dx, \; \forall u, v \in L^2_w(I),$$

and the norm $\|u\|_{L^2_w(I)} = (\int_I u(x)^2 w(x) dx)^{1/2}$. From now on, we suppose $\{\phi_i(x)\}_{i=0}^k \subset \mathbb{P}_k$ namely ϕ_i form an orthogonal basis, then the inner product is

$$\langle \phi_i(x), \phi_j(x) \rangle_{L^2_w(I)} := \|\phi_i\|^2_{L^2_w(I)} \delta_{i,j}, \; 0 \le i, j \le k.$$

We can introduce a projection operator $P_k : L^2_w(I) \to \mathbb{P}_k$ such that, for any function $f \in L^2_w(I)$,

$$P_k f = \sum_{i=0}^{k} \hat{f}_i \phi_i(x), \quad \hat{f}_i = \frac{1}{\|\phi_i\|^2_{L^2_w(I)}} \langle f, \phi_i(x) \rangle_{L^2_w(I)}.$$

It is called the orthogonal projection of f onto $\mathbb{P}_k$ via the inner product $\langle \cdot, \cdot \rangle_{L^2_w(I)}$ and $\{\hat{f}_i\}$ are the generalized Fourier coefficients. Then we have the following Theorem on $L^2_w(I)$:

Theorem. *For any $f \in L^2_w(I)$ and any $k \in \mathbb{Z}^+$, $P_k f$ is the best approximation in the weighted L^2 norm*

$$\|f - P_k f\|_{L^2_w} = \inf_{\psi \in \mathbb{P}_k} \|f - \psi\|_{L^2_w}.$$

This result relates to the linear-quadratic specification in Chapter 2. Any polynomial $\psi \in \mathbb{P}_k$ can be written in a linearized form $\psi = \sum_{i=0}^{k} c_i \phi_i$ for some real coefficients c_i, $0 \leq i \leq k$. Minimizing $\|f - \psi\|_{L^2_w}$ is equivalent to minimizing $\|f - \psi\|^2_{L^2_w}$, whose derivatives are

$$\begin{aligned} \frac{\partial}{\partial c_i} \|f - \psi\|^2_{L^2_w} =& \frac{\partial}{\partial c_i} \left(\|f\|^2_{L^2_w} - 2 \sum_{i=0}^{k} c_i \langle f, \phi_i \rangle_{L^2_w} + \sum_{i=0}^{k} c_i^2 \|\phi_i\|^2_{L^2_w} \right) \\ =& -2 \sum_{i=0}^{k} c_i \langle f, \phi_i \rangle_{L^2_w} + \sum_{i=0}^{k} c_i^2 \|\phi_i\|^2_{L^2_w}. \end{aligned}$$

Note that by setting the derivatives to zero, the unique minimum is attained when $c_i = \hat{f}_i$, the Fourier coefficients.

An extremely important class of orthogonal polynomials is formed by Hermite polynomials H_k whose weight function is nothing else but the standard normal distribution,

$$w(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad H_k(x) = (-\sqrt{2}x)^k e^{x^2} \frac{d^k}{dx^k} e^{-x^2}.$$

If the normal distribution is the weight for all functions of our interests then the inner product in linear-quadratic form will be represented by $\sum_{i=0}^{k} c_i \langle f, H_i \rangle_{L^2_w}$ and $\sum_{i=0}^{k} c_i^2 \|H_i\|^2_{L^2_w}$ with normal density $w(\cdot)$; for any $f \in \mathbb{P}_k$,

$$-2 \sum_{i=0}^{k} c_i \langle \cdot, H_i \rangle_{L^2_w} + \sum_{i=0}^{k} c_i^2 \|H_i\|^2_{L^2_w}$$

constructs a "field" with Gaussian properties. In Chapter 2, we will see a similar construction of a Gaussian family.

The previous result is intuitive but it is for the one dimensional case only. If one needs to consider a higher dimensional case, the equivalent approximation is the Karhunen-Loève expansion which is a widely used technique for dimension reduction in representing stochastic processes. We will not discuss this content until we face the specific case in the proof of Theorem 2.7 in Chapter 2. But the intuition is similar as that in the one dimensional case: if a problem can be expanded by some polynomials in a certain normed space, we need focus only on the standard representation of this approximation. In Chapter 2, the limit representation is asymptotically locally normal and the localization is a linearization approach. A small perturbation of the likelihood ratio process around the local parameter is equivalent to the derivatives of the squared-L^2 distance[10]. To see this argument, if the approximation is replaced by a parametrized function f_θ, $\partial \| f^{\frac{1}{2}} - f_\theta^{\frac{1}{2}} \|^2 / \partial\theta$ is approximated by the term $\lim_{\epsilon\to 0} \| f_{\theta+\epsilon}^{\frac{1}{2}} - f_\theta^{\frac{1}{2}} \|^2 / \epsilon$. Le Cam and Yang (2000) give the equivalence between Hellinger distance of f_θ and the log-likelihood ratio:

$$\lim_{\epsilon\to 0} \| f_{\theta+\epsilon}^{\frac{1}{2}} - f_\theta^{\frac{1}{2}} \|^2 \doteq \lim_{\epsilon\to 0} \epsilon \log \frac{f_{\theta+\epsilon}}{f_\theta}$$

therefore, one will expect the local log-likelihood ratio process to also have a linear-quadratic representation. This becomes the motivation for deriving the local representation formula of EL.

Feature Spaces over Samples

In the thesis, we make a simple enhancement to the class of nonlinear models by projecting the inputs onto a high-dimensional *feature space*[11], a dot product space embedes all kinds of nonlinear patterns, and applying a linear model there. The idea to overcome the nonlinearity is to first project the inputs into some high dimensional space using a set of basis functions (including the polynomial class) and then apply the linear model in this space instead of directly on the inputs themselves. For example, an input x could be projected into the polyno-

10 In particular, a likelihood ratio process can be approximated by a Hellinger distance (L^2-norm) process for densities in an infinitely divisible family.

11 Polynomial space is a special case of feature space.

mial space $\mathbb{P} := \lim_{k\to\infty} \mathbb{P}_k$ of $x = (x_1, \dots x_D)^T$ by the mapping $\phi(x) = (1, x, x \cdot x, x \cdot x \cdot x, \dots)$ where $x \cdot x = (x_1^2, \dots x_D^2)^T$:

$$\begin{aligned} \text{linear fitting: } & y = f(x) = \mathbf{w}^T x, \\ \text{linearizable fitting: } & y = f(x) = \mathbf{c}^T \phi(x) = \langle \mathbf{c}, \phi(x) \rangle. \end{aligned}$$

For regression problems, x belongs to a sample space $\mathcal{X} \subset \mathbb{R}^D$ rather than a bounded interval. If $\phi \in \mathbb{P}_k$, the function ϕ maps an input vector x into an k-dimensional feature space. An approximation of $f(x)$ in terms of $\langle \mathbf{c}, \phi(x) \rangle$ can be attained as follows:

$$\min_{\mathbf{c}} \frac{1}{2} \|\mathbf{c}\|^2 + \mu \sum_{i=1}^{n} |y_i - \langle \mathbf{c}, \phi(x_i) \rangle| \tag{.10}$$

for n observations on x, $\mathbf{c} \in \mathbb{R}^k$, and μ is the penalty coefficient on the fit which is set in advance. The objective function is similar to the one in Lasso (Hastie et al., 2009). The purpose of using the penalty and optimizating $\mathbf{c}$ is to obtain an optimal representation of $f(x)$ in terms of $\langle \mathbf{c}, \phi(x) \rangle$. The norm of coefficient $\mathbf{c}$ matters the complexity of $\langle \mathbf{c}, \phi(x) \rangle$ while the coefficient $\mathbf{c}$ itself matters the goodness of fit $|y - \langle \mathbf{c}, \phi(x) \rangle|$. Objective function (.10) is to balance these two concerns.

The *feature space* uses the inner product as a similarity measure so that we can represent the "patterns" of regressors as vectors in some inner product space $\mathbb{P}$:

$$\phi : \mathcal{X} \to \mathbb{P}.$$

If ϕ is the polynomial function as before, $\mathbb{P}$ is specified to be a polynomial space. But in general, $\mathbb{P}$ is just an inner product space.

Similarly, on a more abstract level, a reproducing kernel can be defined as a Hilbert space of functions f on $\mathcal{X}$ such that all evaluation functions (the maps $f \to f(x)$) are continuous. The theoretical foundation of this trick is from the following Theorem:

Theorem. *(Riesz representation theorem in a Reproducing Kernel Hilbert Space) If f belongs to a Hilbert space and is continuous, then for each $x \in \mathcal{X} \subset \mathbb{R}^D$ there exists a unique function on $\mathcal{X} \times \mathcal{X}$, called $k(x, x')$, such that $f(x') = \langle f(\cdot), k(\cdot, x') \rangle$ where $k(\cdot, \cdot)$ is symmetric and satisfies the conditions for positive definiteness.*

Theorem. *(Moore-Aronszajn theorem) If $\mathcal{X}$ is a countable set, then for every positive definite function $k(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$ there exists a unique Reproducing Kernel Hilbert Space.*

These two theorems in fact imply that the solution of (.10) depends on $k(\cdot,\cdot)$. If an algorithm is defined solely in terms of inner products in $\mathcal{X}$ then it can be lifted into feature space by replacing occurrences of those inner products by $k(\cdot,\cdot)$; this is sometimes called the kernel trick. This technique is particularly valuable in situations where it is more convenient to compute the kernel than the feature vectors $\phi(x)$ themselves. We will use this trick in Chapter 3.

But the freedom to choose the mapping ϕ will enable us to design a large variety of similarity measures and learning algorithms. Is there a concrete connection between basis functions ϕ and the kernel $k(\cdot,\cdot)$? The connection is based on the following Theorem:

Theorem. *(Mercer's theorem) Suppose $k \in L^\infty(\mathcal{X},\mathcal{X})$ is a symmetric real-valued function such that the integral operator $\mathbb{T} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$*

$$\mathbb{T}f(x) = \int_{\mathcal{X}} k(x,x')f(x')dx'$$

is positive definite with kernel $k(\cdot,\cdot)$

$$\int_{\mathcal{X}\times\mathcal{X}} k(x,x')f(x)f(x')dxdx' \geq 0.$$

Let $\phi_i \in L^2(\mathcal{X})$ be the normalized orthogonal eigenfunctions of $\mathbb{T}$ associated with the eigenvalues $\sigma_i > 0$, then

$$k(x,x') = \sum_{i=1}^{\infty} \sigma_i \phi_i(x)\phi_i(x') = \langle \phi(x), \phi(x') \rangle_\Sigma$$

the series converges absolutely and uniformly for almost all (x,x').

Mercer's Theorem lets us define a similarity measure between $\phi(x)$ and $\phi(x')$ via the kernel $k(\cdot,\cdot)$:

$$\begin{aligned} k(x,x') &:= \langle \phi(x), \phi(x') \rangle_\Sigma \\ &= \phi(x)^T \Sigma \phi(x), \quad \text{if the dimension of } \mathbb{P}_k \text{ is finite.} \end{aligned}$$

$k(\cdot,\cdot)$ is called a reproducing kernel function if $\langle f(\cdot), k(\cdot,x) \rangle = f(x)$ for all $f \in \mathbb{P}$ and Σ positive definite. This technique allows us to carry out computations implicitly in the high dimensional space or even let $n \to \infty$. This leads to computational savings when the dimensionality of the feature space is large compared to the number of data points.

The methods appear to have first been studied in the 1940s by Kolmogorov for countable $\mathcal{X}$ and Nachman (1950) in the

general case. Pioneering work on linear representations of a related class of kernels was done by Schoenberg (1938). Further bibliographical comments about the duality of basis functions and reproducing kernels can be found in van den Berg et al. (1984).

Using Mercer's Theorem, we have shown that one can think of the feature map as a map into a high- or infinite-dimensional Hilbert space. The problem is that a high or infinite-dimensional Hilbert space of $k(\cdot,\cdot)$ corresponds to an infeasible computational problem. This particular issue appears in solving dynamic programming in Chapter 3 where we apply the infinite-dimensional approximation. We will show that to suppress the growing dimension is equivalent to reducing the complexity of computing expectations in dynamic programming.

BIBLIOGRAPHY

Aguirregabiria, V., Mira, P., 2002. Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models. Econometrica 70 (4), 1519–1543.

Aguirregabiria, V., Mira, P., 2007. Sequential estimation of dynamic discrete games. Econometrica 75 (1), 1–53.

Aguirregabiria, V., Mira, P., 2010. Dynamic discrete choice structural models: A survey. Journal of Econometrics 156.

Baggerly, K. A., 1998. Empirical likelihood as a goodness-of-fit measure. Biometrika 85 (3), 535–547.

Bajari, P., Benkard, L., Levin, J., 2007. Estimating dynamic models of imperfect competition. Econometrica 75 (3), 1331–1370.

Bousquet, O., Elisseeff, A., 2002. Stability and generalization. Journal of Machine Learning Research (2), 499–526.

Cai, Y., Judd, K., 2010. Stable and efficient computational methods for dynamic programming. Journal of the European Economic Association 8.

Conway, J., 1990. A Course in Functional Analysis. Springer-Verlag, New York.

Csiszar, I., 1984. Sanov property, generalized i-projection and a conditional limit theorem. The Annals of Probability 12 (3), 768–793.

Debreu, G., 1972. Theory of Value: An Axiomatic Analysis of Economic Equilibrium (Cowles Foundation Monograph). Yale University Press.

Donald, S., Imbens, G. W., Newey, W., 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. Econometrica 117 (1), 55–93.

Eguchi, S., 1985. A differential geometric approach to statistical inference on the basis of contrast functionals. Hiroshima Math Journal 15, 341–391.

Fernandez-Villaverde, J., Rubio-Ramrez, J. F., S. Santos, M., 2006. Convergence properties of the likelihood of computed dynamic models. Econometrica 74 (1), 93–119.

Gnedenko, B., Kolmogorov, A., 1968. Limit Distributions for Sums of Independent Random Variables. Addison-Wesley.

Hansen, L. P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50 (4), 1029–1054.

Hansen, L. P., Sargent, T. J., 2007. Robustness. Princeton University Press, Princeton.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

Hotz, J., Miller, R., 1993. Estimating dynamic models of imperfect competition. The Review of Economic Studies 60 (3), 497–529.

Huber, P., 1981. Robust Statistics. Willy, New York.

Kallenberg, O., 2002. Foundations of Modern Probability. Springer Press.

Keane, M. P., Wolpin, K. I., 1994. The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte carlo evidence. The Review of Economics and Statistics 76 (4), 648–672.

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

Kitamura, Y., 2006. Empirical likelihood methods in econometrics: Theory and practice. Advances in Economics and Econometrics Series: Econometric Society Monographs (No. 43).

Kitamura, Y., Otsu, T., Evdokimov, K., 2009. Robustness, infinitesimal, neighborhoods, and moment restrictions. Working Paper (Cowles Foundation Discussion Paper).

Kitamura, Y., Stutzer, M., 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65 (4), 861–874.

Kitamura, Y., Tripathi, G., Ahn, H., 2004. Empirical likelihood-based inference in conditional moment restriction models. Econometrica 72 (6), 1667–1714.

Le Cam, L., 1974. Notes on Asymptotic Methods in Statistical Decision Theory. Centre de recherches mathematiques, Universite de Montreal.

Le Cam, L., Yang, G., 1990. Asymptotics in Statistics: Some Basic Concepts (Springer Series in Statistics). Springer-Verlag, New York.

Le Cam, L., Yang, G., 2000. Asymptotics in Statistics: Some Basic Concepts Second Edition(Springer Series in Statistics). Springer-Verlag, New York.

Nachman, A., 1950. Theory of reproducing kernels. Transactions of the American Mathematical Society 68 (3), 337–404.

Newey, W., Smith, R. J., 2004. Higher order properties of gmm and generalized empirical likelihood estimators. Econometrica 72 (1), 219–255.

Owen, A., 1988. Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75 (2), 237–249.

Owen, A., 1990. Empirical likelihood ratio confidence regions. The Annals of Statistics 18 (1), 90–120.

Owen, A., 2001. Empirical Likelihood. Chapman & Hall/CRC, Florida.

Pakes, A., 1986. Patents as options: Some estimates of the value of holding european patent stocks. Econometrica 54 (4), 755–784.

Pakes, A., Ostrovsky, M., Berry, S., 2008. Simple estimators for the parameters of discrete dynamic games (with entry/exit examples). The RAND Journal of Economics 38 (2), 373–399.

Pesendorfer, M., Schimidt-Dengler, P., 2003. Identification and estimation of dynamic games. Working Paper.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. The Annals of Statistics 22 (1), 300–325.

Radner, R., 1981. Monitoring cooperative agreements in a repeated principal-agent relationship. Econometrica 49 (5), 1127–1148.

Radner, R., 1986. Can Bounded Rationality Resolve the Prisoners' Dilemma? Vol. Contributions to Mathematical Economics. North-Holland, Amsterdam.

Rockafellar, T., 1996. Convex Analysis. Princeton University Press, Princeton.

Ronchetti, E., Trojani, F., 2001. Robust inference with gmm estimators. Journal of Econometrics 101 (1), 37–69.

Rust, J., 1987. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. Econometrica 55 (5), 999–1033.

Rust, J., 1997. Using randomization to break the curse of dimensionality. Econometrica 65 (3), 487–516.

Rust, J., 2008. Comments on:"structural vs. atheoretic approaches to econometrics, by michael keane". forthcoming in Journal of Econometrics.

Rust, J., Traub, J., Wozniakowski, H., 2002. Is there a curse of dimensionality for contraction fixed points in the worst case? Econometrica 70 (1), 285–329.

Sawyer, W. W., 2010. A First Look at Numerical Functional Analysis (Dover Books on Mathematics). Dover Publications.

Schennach, S. M., 2007. Point estimation with exponentially tilted empirical likelihood. The Annals of Statistics 35 (2), 634–672.

Schoenberg, I. J., 1938. Metric spaces and positive definite functions. Transactions of the American Mathematical Society 44 (1), 522–536.

Simon, H., 1957. A Behavioral Model of Rational Choice. Wiley.

Smith, R., 2005. Local gel methods for conditional moment restrictions. Tech. Rep. CWP15/05.

Smith, R. J., 1997. Alternative semi-parametric likelihood approaches to generalised method of moments estimation. The Economic Journal 107 (441), 503–519.

Su, C.-L., Judd, K., 2011. Constrained optimization approaches to estimation of structural models. forthcoming in Econometrica.

van den Berg, C., Christensen, J. P. R., Ressel, P., 1984. Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions (Graduate Texts in Mathematics). Springer.

van der Vaart, A., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.

Wald, A., 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54 (3), 426–482.

Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Annals of Mathematical Statistics 20 (4), 595–601.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50 (1), 1–25.

Wolpin, K. I., 1984. An estimable dynamic stochastic model of fertility and child mortality. The Journal of Political Economy 92 (5), 852–874.

Zeidler, E., 1995. Applied Functional Analysis: Applications to Mathematical Physics (Applied Mathematical Sciences). Springer.