Essays on empirical likelihood in economics

Gao, Z.

Publication date
2012

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.