Monotone Approximation of Decision Problems

Naveed Chehrazi, Thomas A. Weber
Department of Management Science and Engineering, Stanford University, Stanford, California 94305
{chehrazi@stanford.edu, webert@stanford.edu}

Many decision problems exhibit structural properties in the sense that the objective function is a composition of different component functions that can be identified using empirical data. We consider the approximation of such objective functions, subject to general monotonicity constraints on the component functions. Using a constrained B-spline approximation, we provide a data-driven robust optimization method for environments that can be sample-sparse. The method, which simultaneously identifies and solves the decision problem, is illustrated for the problem of optimal debt settlement in the credit-card industry.

Subject classifications: B-splines; monotone approximation; nonparametric/semiparametric methods; robust optimization; sample-sparse environments.
Area of review: Special Issue on Computational Economics.
History: Received September 2008; revision received November 2009; accepted December 2009. Published online in *Articles in Advance* June 3, 2010.

1. Introduction

In this paper we study the approximation of decision problems, subject to structural monotonicity constraints, in view of robustly determining optimal actions. A key motivation for imposing monotonicity constraints on the approximation of a decision problem is to combine information from available data with structural insights obtained from a model-based analysis of the problem. Structural monotonicity constraints naturally arise when viewing the objective function as a composition of several "component functions," each of which directly relates to an observable phenomenon that may naturally exhibit a monotone behavior, such as a supply curve that is increasing in price, a demand curve that is decreasing in price, a response curve that is increasing in a certain stimulus, or a production function with a decreasing slope due to decreasing returns to scale.

Our approach has five key features. First, it is semiparametric in the sense that each element of the class of approximating functions (splines of a certain degree and for a certain set of breakpoints or "knots") can be represented by a unique, finite-dimensional parameter vector.[1] With this it is possible to arbitrarily closely approximate *any* continuous component function (and not just functions within a narrow parameterized class). Standard parametric approximation methods do not have this feature. For example, in a classical linear regression one varies parameters so as to select a linear model by minimizing a quadratic cost criterion, which may lead to a significant residual error whenever the actual underlying system is nonlinear. In addition, the method provides a natural way of incorporating monotonicity constraints with respect to first and second derivatives of the component function by translating them into simple restrictions in the parameter space.[2]

Second, the approach combines the problems of model estimation and optimization by evaluating the approximation error using any user-defined robustness criterion, such as average performance, worst-case performance, expected gain, or competitive ratio. The approach is therefore in the spirit of recent advances in operational statistics (Shanthikumar and Liyanage 2005, Lim et al. 2006, and Besbes et al. 2010), which emphasize the importance of using a metric for the evaluation of an approximation error that can be directly related to the cost of mistakes in the underlying decision. An optimal robust decision is obtained together with an optimal approximation error by solving a pair of nested optimization problems.

Third, the method works equally well in data-rich and data-poor (or sample-sparse) environments, because it treats interpolation and approximation in the same unifying framework. In a sample-sparse environment the intersample behavior of a component function we wish to approximate is not well understood. This may be the case even though the information at a given sample point is extremely accurate. We therefore distinguish between intersample and intrasample uncertainty. Our method deals with both sources of parameter uncertainty jointly by fitting a stochastic process instead of a deterministic function (cf. Figure 1). The resulting stochastic model can be used to design optimal experiments for the acquisition of additional data points, so as to maximize the value of the additional information.

Fourth, in fitting the stochastic process, we use information derived from the data as well as the structural insights about the behavior of the component functions. The latter

Figure 1.   (a) Approximation of response curve in sample-sparse environment. (b) Detail with confidence interval.
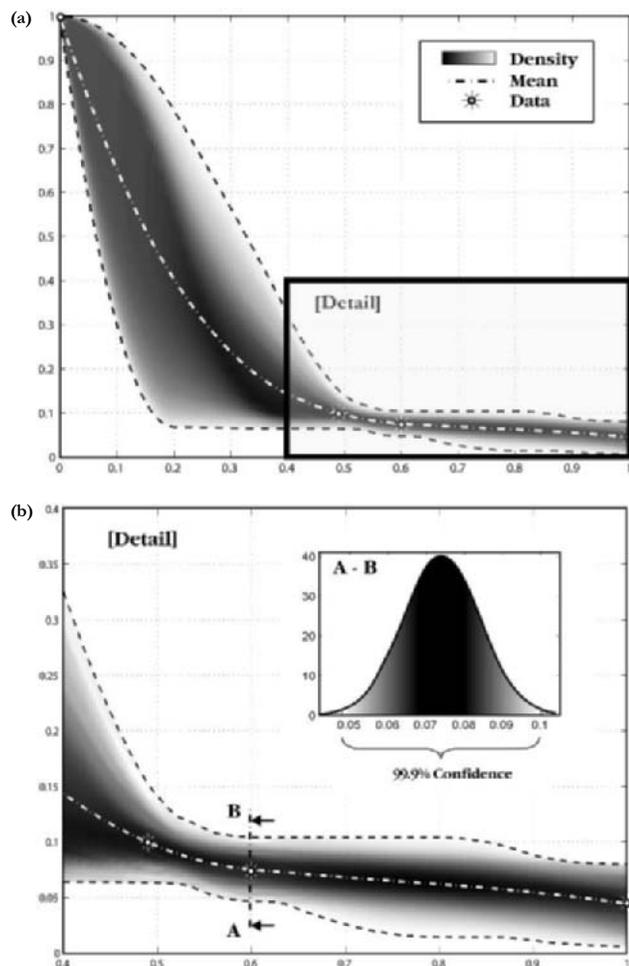
implies a set of "consistent parameters" (a bounded polyhedron as we show), whereas the former yields a conditional probability distribution over these consistent parameters.

Fifth, the method produces not only an optimal decision relative to a given robustness criterion (and a "best approximation error"), but also a probability measure that indicates the likelihood of optimality for any feasible action. The latter constitutes a valuable input for sensitivity analysis.

We illustrate the method for the practical problem of designing optimal settlement offers to collect delinquent credit-card debt (cf. §5).

1.1. Related Literature

a. Robust Optimization and Identification.   Work on the problem of joint system identification and optimization dates back to Haimes et al. (1971), who formulate a bicriterion vector-optimization problem that includes a quadratic error measure for the identification of a finite number of model parameters based on sample observations and a cost function to evaluate the quality of decisions. As detailed further by Haimes and Wismer (1972), several approaches to this multicriteria optimization problem are possible. First, one can consider the minimization of a convex combination of the two criteria (parametric approach). Second, it is possible to alternate between minimizing each criterion individually and using the solution to one problem for the other problem. This recursive two-step approach does not generally converge. Third, by incorporating one of the two criteria as a constraint when optimizing the other criterion, the trade-offs between optimization and identification can be made apparent (ε-criterion approach). Fourth, one can follow a "robust" procedure by optimizing a worst-case version of the model, in which the parameters have been chosen so as to most adversely affect the optimization result (maximin approach). The last two approaches resurface in more recent work on robust optimization (see, e.g., Ben-Tal and Nemirovski 1998, 2002; or Bertsimas and Sim 2004), which studies the dependence of a mathematical program, consisting of an objective function and a constraint set, on the set of unknown parameters either with or without a probabilistic structure incorporating a priori information about likelihood of the different parameter realizations. It aims at solving *robust* versions of the original problems, approximating uncertainty sets, e.g., by suitable ellipsoids, to obtain computationally tractable programs. In that same spirit, our method generates a linearly constrained parameter domain (a bounded polyhedron) that represents the decision maker's knowledge about the model structure and which defines the support of the robustness criterion.

b. Shape-Preserving Approximation.   The problem of "monotone approximation" (i.e., the approximation of a monotone function f with monotone functions of a certain class) was introduced by Shisha (1965).[3] Early efforts in this area concentrated on bounding the approximation error by the modulus of continuity of derivatives of f. Such "Jackson-type" inequalities were established, for example, by De Vore (1977a, b) for polynomials and splines, respectively, and by Anastassiou and Yu (1992) for wavelets. Much of the subsequent developments have focused on the special case where the available data points satisfy the monotonicity constraints themselves, so that monotone interpolation becomes feasible. McAllister et al. (1977), based in part on Passow and Roulier (1977), provide an algorithm that determines an interpolating shape-preserving (i.e., increasing, convex) spline. The monotonicity domain for the parameters of spline polynomials is generally not simple. For example, in the case of cubic $\mathscr{C}^1$-splines, the monotonicity domain for the data-normalized coefficients consists in the nontrivial union of a triangle and an ellipsoid (Fritsch and Carlson 1980). This stands in contrast to the computationally much simpler polyhedron obtained here using so-called B-splines (see, e.g., De Boor 1978/2001).[4]

B-splines have been used by Dierckx (1980) for approximating convex functions, and by Ramsay (1988) for approximating increasing functions. Kocić and Milovanović (1997) and Kvasov (2000) survey recent advances in shape-preserving approximation of data by polynomials and splines, which include more general shape constraints such as symmetry, unimodality, or discontinuities. The extant work on shape-constrained spline interpolation has produced a multitude of algorithms that yield various candidates for interpolation functions, depending on the precise constraints, smoothness requirements, and degrees of the splines considered. However, for the purpose of jointly identifying and optimizing a decision model it is necessary to characterize the *class* of suitable models, i.e., by constructing finite-dimensional and computationally simple parameter domains, rather than by pinpointing particular candidate functions. In that sense, our work is related to recent papers by Beliakov (2000, 2002), who examines conditions for the approximation of functions using least-squares splines subject to shape constraints.[5]

c. Applications. Shape-preserving approximation techniques have seen only very few applications in economics and management science.[6] Wang and Judd (2000) approximate a two-dimensional value function in a savings allocation problem using two-dimensional shape-constrained splines. In the statistics literature, there has been work on the parametric estimation of functions subject to economically motivated shape constraints, such as isotone regression (Robertson et al. 1988) or concave regression (Hanson and Pledger 1976). This work is principally concerned with model identification and relies on the assumption that the chosen class of parametric regressors is appropriate for the model under consideration. The criteria used to measure the approximation error (such as mean square error or maximum absolute error) are typically disconnected from the underlying decision problem. Besbes et al. (2010), motivated in part by this deficiency, develop a hypothesis test for the significance of a parameterization based on sample data of objective-function values when comparing the parameter-induced optimum to a nonparametric baseline estimate of the objective function. In sample-sparse environments, however, parametric methods are of limited use only, because the data does not induce any reasonable assumptions on the model class. We therefore concentrate on nonparametric (or semiparametric) methods.

A much-needed assumption for many of the available nonparametric estimation methods in the literature is that a "well-diversified" and "sufficiently large" data sample is available. However, in many applications, such as the problem of designing optimal credit-card settlement offers discussed in §5, adding a new sample point may be fairly expensive. As a result, a method is needed that performs well in such sample-sparse environments by systematically using all available structural information to identify primitives of the problem in view of making good decisions

(and not necessarily to minimize an abstract error criterion). We assume that the structural information is available in the form of monotonicity constraints and require an entire class of approximating functions (consistent with these constraints) to remain "in play" until a decision has been selected, tightly coupling the optimization and identification.

1.2. Performance Comparison

To illustrate the performance gain that can be achieved using our method in a sample-sparse environment, we compare it to three standard nonparametric estimation procedures: (i) kernel-based, (ii) K-nearest-neighbor, and (iii) orthogonal-series.[7] Each of these methods provides an estimate y of a dependent variable $\tilde{y}$, conditional on the realization x of an independent variable $\tilde{x}$. The estimated functional relationship, $f(x) = E[\tilde{y} \mid \tilde{x} = x]$, is then typically used in place of the unknown relationship in the decision problem. Another common feature of these methods is that they all seek to replace the true functional relationship $y = f(x)$ by an approximate relationship,

$$\hat{f}(x) = \sum_{j=1}^{n} \alpha_j W_{n,j}(x), \tag{1}$$

which is obtained as a superposition of n "weight functions," $W_{n,j}(\cdot)$, $j \in \{1, \ldots, n\}$. Specifically, in methods (i) and (ii) n is the cardinality of the data sample $z = \{(x_j, y_j)\}_{j=1}^{n}$, it is $\alpha_j = y_j$, and the weight function is given by $W_{n,j}(x) = \Lambda((x - x_j)/\beta) / \sum_{l=1}^{n} \Lambda((x - x_l)/\beta)$, where $\Lambda: \mathbb{R} \to \mathbb{R}$ is a suitable kernel function. Note that in method (ii) the (variable) scaling factor $\beta = \beta(x)$ is chosen so as to include the K nearest data points closest to x in (1). Common shapes of $\Lambda(\cdot)$ include rectangular, triangular, and quadratic (Epanechnikov). In method (iii) n is the order of the (truncated) orthogononal series, the $W_{n,j}$s are basis functions for the approximation class (e.g., $\mathscr{C}^0[a, b]$), and the α_js are the corresponding Fourier coefficients.

In sample-sparse environments (see, e.g., the application in §5), methods (ii) and (iii) are unsuitable. In method (ii), any nontrivial support diameter may lead to excessive averaging, which makes it impossible to obtain function estimates that include higher-frequency components (which allow for fast variations and thus flexible intersample behavior in sample-sparse regions). Method (iii) does allow for higher frequencies, yet in its standard form provides only a single trend estimate, without necessarily interpolating the data. At least in data-sparse settings, the latter is not useful when the accuracy at the few sample points is fairly large. We also note that the quality of the approximation is limited by the well-known sampling theorem (Shannon 1949) that relates the maximum frequency in (1) ("bandwidth" B) to the required distance between two sample points ("Nyquist rate" $1/(2B)$). Lastly, standard orthogonal-series methods are at present unsuited for use in conjunction with monotonicity constraints.[8]

Consider now method (i), kernel-based estimation.[9] Similar to method (ii), the $W_{n,j}$s are scaled by a (constant) scaling factor $\beta > 0$ to avoid trivial estimation of the relation $y = f(x)$ between sample points. As before, this introduces—via averaging—an approximation error in the function estimate $\hat{f}(x)$, especially when the (few) available data points are fairly accurate. Furthermore, the function estimate inherits the properties of the kernel functions. Because the latter are typically not monotone, this complicates the consideration of shape constraints. The following example illustrates the performance of our procedure based on B-splines in comparison to method (i).

NUMERICAL EXAMPLE. For the Epanechnikov kernel, with

$$W_{n,j}(x) = \left[\sum_{l=1}^{n} \left(1 - \left(\frac{x - x_l}{\beta} \right)^2 \right) \mathbf{1}_{[x_l - \beta, \, x_l + \beta]} \right]^{-1}$$
$$\cdot \left(1 - \left(\frac{x - x_j}{\beta} \right)^2 \right) \mathbf{1}_{[x_j - \beta, \, x_j + \beta]},$$

given the support diameter $2\beta > 0$, Figure 2(a) shows the kernel-based estimation procedure applied to a sample-sparse debt-collection problem (cf. §5, Segment 2) to obtain an estimate $\hat{f}(x)$ of a "response rate," where $x \in [0, 1]$ corresponds to a settlement rate, and where the data sample $z = \{(0, 1), (0.49, 0.0996), (0.6, 0.0743), (1, 0.0457)\}$ is given. The decision maker is interested in maximizing the expected revenue $g(\hat{f}(x), x) = x\hat{f}(x)$, depicted in Figure 2(b). As can be seen, kernel-based estimation incorrectly suggests an optimal settlement rate of 100%, given various reasonable bandwidths, although the data itself suggest that a settlement rate of 49% is superior. Figures 2(c) and 2(d) show the corresponding results for the monotone approximation method used in this paper. As detailed in §5, the method predicts that the optimal average performance is obtained at a settlement rate of 20%. The expected return ("spin") at the optimal settlement rate suggested by our method is about 80% larger than what is achieved by the kernel-based estimator.[10] The performance difference

Figure 2. (a) Approximation of the response-rate function using Epanechnikov kernel. (b) Expected income (spin) using Epanechnikov kernel. (c) Approximation of the response-rate function by a stochastic process using the proposed monotone approximation method. (d) Stochastic process of expected income (spin) using the proposed monotone approximation method.
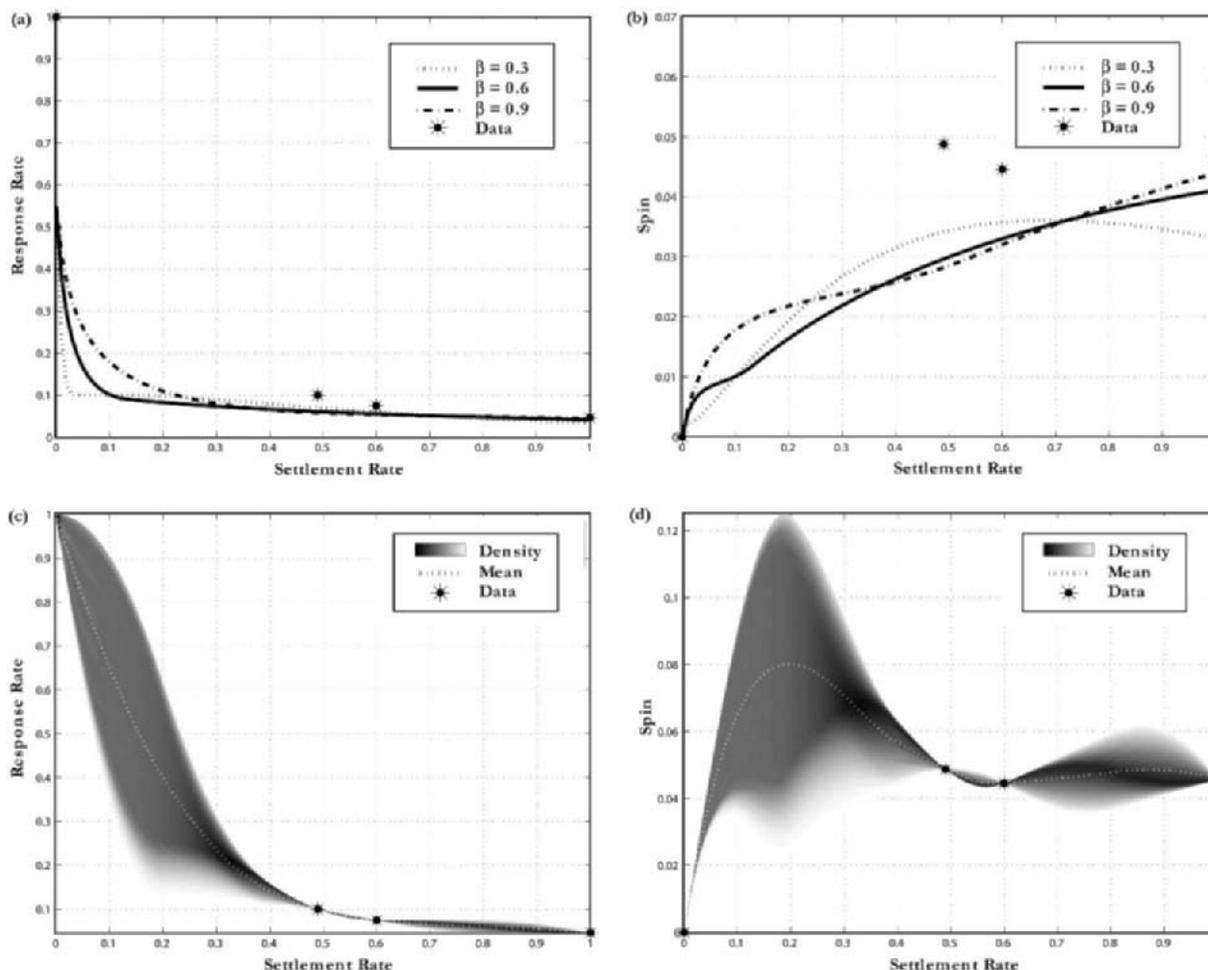
for a different robustness measure (competitive ratio) is similarly large. The main driver for the difference is that the new method considers all functional relationships consistent with the existing knowledge about the structure of the decision problem, which provides significant additional information, especially in sample-sparse settings.

1.3. Outline

The paper is organized as follows. Section 2 introduces the decision problem, its decomposition into component functions, and the basic problem of approximating the component functions subject to structural monotonicity constraints. Section 3 provides characterizations of the "monotonicity domain" and the "robust approximation set" conditional on observed sample data. Spline parameters in the latter set guarantee that approximations are within a prespecified approximation-error bound and satisfy the monotonicity constraints. Section 4 examines informational issues related to the use of sample data, in terms of both consistent belief updates and information acquisition by selecting an optimal experiment. In §5 we analyze a debt-settlement problem as an application of the method in a sample-sparse setting. Section 6 concludes.

2. The Model

2.1. Decision Problem

We consider the decision problem of finding

$$x^* \in \arg\max_{x \in X} g(f(x), x), \tag{2}$$

where $X = [a, b] \subset \mathbb{R}$, with $-\infty < a < b < \infty$, is a given choice set, $f \colon X \to \mathbb{R}$ is a *component function*, and $g \colon \mathbb{R} \times X \to \mathbb{R}$ is a *structural objective function*. We make the following assumptions on the primitives of the decision problem (2).

A1. $g \in \mathcal{C}_L^0$, *i.e., the structural objective function g is Lipschitz-continuous.*

A2. *For some $r \geqslant 1$: $f \in \mathcal{C}^r[a, b]$, i.e., the component function f is r-times continuously differentiable.*

We assume that although the structural objective function g is known, an analytic expression for the component function is generally not available. In other words, the decision maker knows how to compute his *objective function*

$$\pi(x) = g(f(x), x) \tag{3}$$

only after f has been identified. For example, if x is the price of a product and $f(x)$ denotes a (downward-sloping) market demand curve, then a revenue-maximizing firm's objective function $\pi(x)$ is given by $g(f(x), x) = xf(x)$.

REMARK 1. The situation in which the objective function itself is unknown corresponds to the special case where $g(f(x), x) = f(x)$.

REMARK 2. The reason for writing the objective function in (2) as a composition of a structural objective function and a component function is that this enables the decision maker to use both his knowledge about properties of the component function and his knowledge about how the component function affects the objective function before committing to a specific feasible action $x \in X$. Using this decomposition, the decision maker can solve the identification problem and the optimization problem jointly. The standard procedure would be to first solve the identification problem to obtain a component-function approximation $\hat{f} \colon X \to \mathbb{R}$. The decision maker's *approximate objective function* is then

$$\hat{\pi}(x) = g(\hat{f}(x), x),$$

leading to the approximate solution

$$\hat{x} \in \arg\max_{x \in X} g(\hat{f}(x), x).$$

This procedure, however, ignores the model uncertainty resulting from the approximation of f by $\hat{f}$ at the optimization step.

Lastly, we assume that the decision maker knows that the component function satisfies a set of monotonicity constraints.

M1. The component function f satisfies a first-order monotonicity constraint of the form

$$\gamma_1 f'(x) \geqslant 0, \quad \forall x \in X, \tag{4}$$

given some real number γ_1.

M2. If $r \geqslant 2$, then the component function f satisfies a second-order monotonicity constraint of the form

$$\gamma_2 f''(x) \geqslant 0, \quad \forall x \in X, \tag{5}$$

given some real number γ_2.
Note that if $\gamma_1 = 0$ or $\gamma_2 = 0$, then the corresponding constraints (M1 and/or M2) are not active. Without loss of generality, γ_1 and γ_2 can be chosen in $\{-1, 0, 1\}$.

2.2. Component-Function Approximation

Data relative to the component function f is available in the form of the *sample $z = \{(x_j, y_j)\}_{j=1}^n$* containing n noisy observations $z_j = (x_j, y_j)$ with $a = x_1 < x_2 < \cdots < x_n = b$, which are such that

$$y_j = f(x_j) + \epsilon_j \tag{6}$$

for all $j \in \{1, \ldots, n\}$, where ϵ_j are realizations of independent zero-mean random variables $\tilde{\epsilon}_j$.

REMARK 3. In applications, multiple observations of the component function evaluated at the same set of x_j-values can sometimes be obtained. If for a given x_j the i.i.d.

sample $y_{j,1}, \ldots, y_{j,m}$ is available, then by the strong law of large numbers $\bar{y}_{j,m} = (y_{j,1}, \ldots, y_{j,m})/m \xrightarrow{a.s.} f(x_j)$ as $m \to \infty$. Furthermore, by the central limit theorem the sample mean $\bar{y}_{j,m}$ is asymptotically normal.[11] Thus, in practice we collapse the vector $(y_{j,1}, \ldots, y_{j,m})$ to its sample mean $y_j = \bar{y}_{j,m}$ and assume that the corresponding sample error is normal, i.e., $\tilde{\epsilon}_j \sim N(0, \hat{\sigma}_j^2)$, with zero mean and variance $\hat{\sigma}_j^2 = \frac{1}{m(m-1)} \sum_{l=1}^{m} (y_{j,l} - \bar{y}_{j,m})^2$.

Given a family of functions $s(\cdot; \theta): \mathbb{R} \to \mathbb{R}$, indexed by a parameter vector θ in a finite-dimensional parameter space Θ, we henceforth use the *approximation error*

$$
\begin{aligned}
e(\theta \mid z) &= \left\| (w_1(s(x_1; \theta) - y_1), \ldots, w_n(s(x_n; \theta) - y_n)) \right\|_\infty \\
&= \max_{j \in \{1, \ldots, n\}} |w_j(s(x_j; \theta) - y_j)|,
\end{aligned}
\tag{7}
$$

which represents the (weighted) maximum absolute deviation of $s(\cdot; \theta)$ from the sample z, where $w = (w_1, \ldots, w_n)$ is a given vector of positive weights.[12] The inclusion of weights in the error norm is isomorphic to having different standard deviations for different y_js (cf. Remark 3). The weight is typically chosen larger at a point with smaller standard deviation, e.g., $w_j = 1/\hat{\sigma}_j$.

2.3. Robust Decision Making

The set $\Theta_\gamma \subset \Theta$ containing parameters for functions $s(\cdot; \theta)$ that satisfy the monotonicity constraints M1 and M2 for a given $\gamma = (\gamma_1, \gamma_2)$ is referred to as *monotonicity domain*. Relative to the family of functions that is considered for the approximation of a component function, the set Θ_γ contains all component-function approximations that satisfy the monotonicity requirements, i.e., all those that could possibly be relevant for the decision problem (2), even in the complete absence of empirical data.

Given a sample z, which contains data on the component function, we can, for any approximation error $\varepsilon \geq 0$, determine the set $\widehat{\Theta}_{\varepsilon, \gamma}(z) \subset \Theta_\gamma$ that contains all those parameter vectors θ for which the maximum absolute deviation $e(\theta \mid z) \leq \varepsilon$. The set $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ is referred to as the *robust approximation set* (cf. Figure 3). It is used as a starting point for what we term a *robust* solution to the decision problem (2). More specifically, considering component-function approximations in $\{s(\cdot; \theta)\}_{\theta \in \Theta_\gamma}$, we say that $\hat{x}^*(z)$ is a *robust solution to the decision problem* (2) (*relative to the sample z*) if

$$
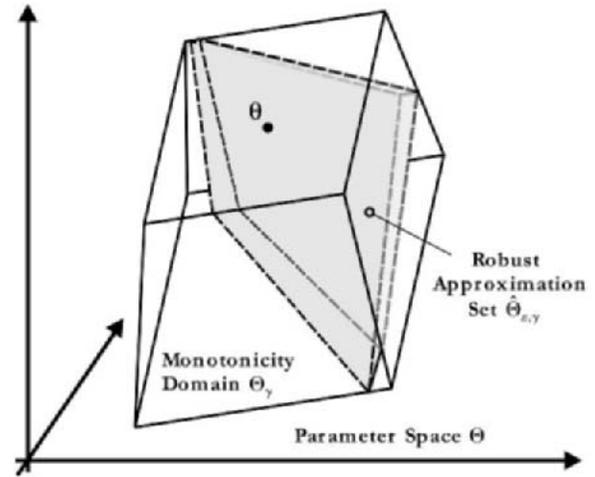\hat{x}^*(z) \in \arg\max_{x \in \bar{X}(z)} \mathcal{R}[g(s(x; \theta), x) \mid \theta \in \Theta_\gamma],
\tag{8}
$$

where

$$
\bar{X}(z) = \bigcup_{\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]} \widehat{X}(\varepsilon, z),
\tag{9}
$$

and

$$
\widehat{X}(\varepsilon, z) = \arg\max_{x \in X} \mathcal{R}[g(s(x; \theta), x) \mid \theta \in \widehat{\Theta}_{\varepsilon, \gamma}(z), z]
\tag{10}
$$

Figure 3. Model primitives in the parameter space.

for all ε in the set of *admissible* approximation errors, $[\underline{\varepsilon}, \bar{\varepsilon}]$. The lower bound $\underline{\varepsilon}$ cannot be smaller than $\underline{\varepsilon}_\gamma(z)$, which is the minimum achievable error such that $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ is nonempty (cf. Equation (25) in §3.3) and $\bar{\varepsilon}$ is a maximum acceptable error, specified by the decision maker.[13] Given a nonempty measurable subset $\widehat{\Theta}$ of the parameter space Θ, the *robustness criterion* $\mathcal{R}[g(s(x; \theta), x) \mid \theta \in \widehat{\Theta}]$ is a real-valued functional that evaluates its argument $g(s(x; \theta), x)$ at all $\theta \in \widehat{\Theta}$. Following are a number of well-known robustness criteria. Let $\widehat{F}(\theta \mid z)$ be a cumulative distribution function (cdf) that represents the decision maker's beliefs over realizations of θ in $\widehat{\Theta}$, conditional on the observed data sample z.[14] The following are standard robustness criteria for $x \in X$:

$$
\begin{aligned}
&\mathcal{R}_{\text{Avg}}[g(s(x; \theta), x) \mid \theta \in \widehat{\Theta}, z] \\
&\quad = \int_{\widehat{\Theta}} g(s(x; \theta), x) \, d\widehat{F}(\theta \mid z): \textit{Average Performance;}
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{R}_{\text{WC}}[g(s(x; \theta), x) \mid \theta \in \widehat{\Theta}, z] \\
&\quad = \inf\{g(s(x; \theta), x): d\widehat{F}(\theta \mid z) > 0\}: \\
&\qquad \textit{Worst-Case Performance;}
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{R}_{\text{EG}}[g(s(x; \theta), x) \mid \theta \in \widehat{\Theta}, z] \\
&\quad = \int_{\widehat{\Theta}} \left[g(s(x; \theta), x) - \max_{\hat{x} \in X}\{g(s(\hat{x}; \theta), \hat{x})\} \right] d\widehat{F}(\theta \mid z): \\
&\qquad \textit{Expected Gain;}
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{R}_{\text{CR}}[g(s(x; \theta), x) \mid \theta \in \widehat{\Theta}, z] \\
&\quad = \int_{\widehat{\Theta}} \left[\frac{g(s(x; \theta), x)}{\max_{\hat{x} \in X}\{g(s(\hat{x}; \theta), \hat{x})\}} \right] d\widehat{F}(\theta \mid z): \\
&\qquad \textit{Competitive Ratio.}
\end{aligned}
$$

Other robustness criteria can be used. All robustness criteria are defined analogously when they are not conditioned on z. Our definition of a robust solution $\hat{x}^*(z)$ to the decision problem (2) in (8)–(10) guarantees that the

decision maker does not pay excessive attention to the data when *jointly* considering the problems of approximating the objective function and determining a best action. In relation (10) the decision maker determines the set of optimal actions $\hat{X}(\varepsilon, z)$, considering all component-function candidates $s(\cdot; \theta)$ that achieve a specified error bound ε relative to the available data sample z. In (9) the decision maker determines the superset of actions that may be optimal, given *any* allowable error $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]$.[15] Finally, in (8) an optimal robust action is determined by testing any potential solution against all parameter realizations $\theta \in \Theta_\gamma$, without conditioning on z. The decision-relevant information contained in the data sample z is transmitted from problem (10) to problem (8) via the set $\bar{X}(z)$. For example, when the amount of data is very large, then because of the fact that the conditional distribution $\hat{F}(\theta \mid z)$ in (10) is sharply concentrated around the true parameter value, the set $\hat{X}(\varepsilon, z)$ becomes insensitive to the error ε, so that the cardinality of $\bar{X}(z)$ is "small."

REMARK 4. The robust solution $\hat{x}^*(z)$ in (8)–(10) implies the *optimal approximation error*

$$\hat{\varepsilon}^*(z) = \inf\{\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]: \hat{x}^*(z) \in \hat{X}(\varepsilon, z)\}, \qquad (11)$$

which is the smallest allowable error ε such that $\hat{X}(\varepsilon, z)$ contains $\hat{x}^*(z)$.

REMARK 5. To understand the hierarchical nature of a robust solution to the decision problem (2) that satisfies (8)–(10), it is useful to consider a simple example without monotonicity constraints (i.e., $\gamma = 0$). Let $s(x; \theta) = 2\theta - x$, where the parameter $\theta \geqslant 0$ is unknown, and let $g(s(x; \theta), x) = s(x; \theta)x = (2\theta - x)x$. Given a sample $z = \{(x_j, y_j)\}_{j=1}^n$ and $\varepsilon \geqslant \underline{\varepsilon}$, it is $\theta \in \hat{\Theta}_{\varepsilon, \gamma}(z) = \{\vartheta \in \Theta_\gamma: \max_{j \in \{1, \ldots, n\}} |y_j - (2\vartheta - x_j)| \leqslant \varepsilon\}$ if and only if $\theta_1(\varepsilon, z) \leqslant \theta \leqslant \theta_2(\varepsilon, z)$, for some appropriate functions θ_1, θ_2. Taking average performance as a robustness criterion, solving the optimization problem (10) yields

$$\hat{X}(\varepsilon, z) = \{E[\tilde{\theta} \mid \theta \in \hat{\Theta}_{\varepsilon, \gamma}(z), z]\}$$

$$= \arg\max_{x \in X} \int_{\theta_1(\varepsilon, z)}^{\theta_2(\varepsilon, z)} (2\theta - x)x \, d\hat{F}(\theta \mid z),$$

where $X = [0, \bar{x}]$ for some large enough $\bar{x} > 0$. The solution to (9) becomes

$$\bar{X}(z) = \{E[\tilde{\theta} \mid \theta_1(\varepsilon, z) \leqslant \theta \leqslant \theta_2(\varepsilon, z), z]: \varepsilon \geqslant \underline{\varepsilon}\},$$

so that by Equation (8) we obtain as the optimal robust decision $\hat{x}^*(z)$ the point in $\bar{X}(z)$ that is closest to $\bar{\theta} = E[\tilde{\theta}]$.

REMARK 6. If the hierarchy between problems (8) and (10) is eliminated by replacing the monotonicity domain Θ_γ in (8) by the robust approximation set $\hat{\Theta}_{\bar{\varepsilon}, \gamma}(z)$, then the decision maker can choose $\bar{\varepsilon}$ so as to ignore certain unfavorable parameter values. Using such an "ostrich-algorithm," the decision maker may be led to overly optimistic actions when the initially observed sample suggests "good news."[16]

3. Monotone Approximation

We now consider the approximation of a twice continuously differentiable real-valued component function f. By M1 and M2, f satisfies the monotonicity constraint

$$\min\{\gamma_1 f'(x), \gamma_2 f''(x)\} \geqslant 0 \qquad (12)$$

for a given $\gamma = (\gamma_1, \gamma_2)$. The key idea is to select a family of approximating functions $s(\cdot; \theta)$, so that the monotonicity constraint (12) can be transformed into a parameter constraint such that

$$\min\{\gamma_1 s^{(1)}(x; \theta), \gamma_2 s^{(2)}(x; \theta)\} \geqslant 0 \quad \Leftrightarrow \quad \theta \in \Theta_\gamma \qquad (13)$$

for an appropriate *monotonicity domain* $\Theta_\gamma \subseteq \Theta$. The latter is determined explicitly in Corollary 1 below.

3.1. Approximating Functions

The set of approximating functions is the linear space $\mathscr{S}_k(\nu)$ of splines of degree $k \geqslant 1$, defined on the interval $[a, b]$ with an ordered vector of knots $\nu \in \mathscr{V}_{a, b}$, where

$$\mathscr{V}_{a, b} = \big\{(\nu_0, \ldots, \nu_{\kappa+1}) \in \mathbb{R}^{\kappa+2}:$$
$$a = \nu_0 < \nu_1 < \cdots < \nu_{\kappa+1} = b \text{ and } \kappa \in \mathbb{N}\big\}.$$

Let $\nu \in \mathscr{V}_{a, b}$ and $\kappa = \dim(\nu) - 2$. A *spline* $s \in \mathscr{S}_k(\nu)$ is a function $s: [a, b] \to \mathbb{R}$ that satisfies the following two conditions.

S1. For all $i \in \{0, \ldots, \kappa\}$: $s|_{[\nu_i, \nu_{i+1}]} \in \mathscr{P}_k$, i.e., the restriction of the function s to the interval $[\nu_i, \nu_{i+1}]$ is a polynomial of degree k;

S2. $s \in \mathscr{C}^{k-1}[a, b]$, i.e., the function s and its derivatives up to order $k - 1$ are continuous on $[a, b]$.

REMARK 7. (i) The components of the knot vector ν do not have to be strictly increasing if condition S2 is appended as follows: if $\nu_i = \cdots = \nu_{i+l}$ for some i, l with $l \leqslant k$, then s is required only to have continuous derivatives up to order $k - l - 1$ at the point ν_i. (ii) It is possible to use splines of degree $k = 0$, i.e., piecewise-constant functions, by dropping condition S2 altogether and replacing the closed intervals $[\nu_i, \nu_{i+1}]$ in condition S1 by the right-open intervals $[\nu_i, \nu_{i+1})$. The right endpoint $s(b)$ is then either undefined or specified separately (see, e.g., Dierckx 1993).

For any $i \in \{-k, \ldots, \kappa\}$, let $N_{i, k+1}: [a, b] \to \mathbb{R}$ denote a (normalized) B-spline of degree k with knots $\nu_i, \ldots, \nu_{i+k+1}$, where we assume that[17]

$$\nu_{-k} = \nu_{-k+1} = \cdots = \nu_{-1} = \nu_0 = a, \quad \text{and}$$

$$b = \nu_{\kappa+1} = \nu_{\kappa+2} = \cdots = \nu_{\kappa+k} = \nu_{\kappa+k+1}.$$

It can be shown that every spline $s \in \mathscr{S}_k(\nu)$ has a unique *B-spline representation* as a linear combination of B-splines, in the form

$$s(x; \theta) = N_{k+1}(x)\theta, \quad \text{for all } x \in [a, b], \qquad (14)$$

where

$$\theta = (\theta_{-k}, \ldots, \theta_{\kappa})^T \in \mathbb{R}^{k+\kappa+1}$$

is the *B-spline-coefficient vector*, and

$$N_{k+1} = (N_{-k, k+1}, \cdots, N_{\kappa, k+1})$$

is a row vector of $k + \kappa + 1$ B-splines, forming a basis for the spline space $\mathscr{S}_k(\nu)$.

REMARK 8. For our developments below it proves useful to introduce an equivalent representation of s in parameters $\hat{\theta} = L_{k+1}^{-1}\theta$,

$$s(x; \hat{\theta}) = \hat{N}_{k+1}(x)\hat{\theta} = N_{k+1}(x)L_{k+1}L_{k+1}^{-1}\theta = s(x; \theta), \quad (15)$$

where $\hat{N}_{k+1} = N_{k+1}L_{k+1}$, and L_{k+1} is a $(k + \kappa + 1) \times (k + \kappa + 1)$ lower-triangular matrix with ones as the only nonzero entries, so that

$$L_{k+1}^{-1} = \begin{bmatrix} 1 & & & & 0 \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \cdots & \cdots & \\ 0 & & & -1 & 1 \end{bmatrix}. \quad (16)$$

We recall that splines of any positive order can be used to approximate continuous functions on a compact interval in the sup-norm.

LEMMA 1 (DENSENESS OF SPLINES IN $\mathscr{C}^0[a, b]$).[18] *Let $k \geqslant 0$. For any given $f \in \mathscr{C}^0[a, b]$ and $\varepsilon > 0$, there exist $\nu \in \mathscr{V}_{a,b}$ and $s \in \mathscr{S}_k(\nu)$ such that $\sup_{x \in [a,b]} |f(x) - s(x)| < \varepsilon$.*

3.2. Monotonicity Domain

Given an element of the spline space $\mathscr{S}_k(\nu)$ in the representation (14) or (15), we seek to convert the monotonicity constraint (12) (i.e., M1 and M2) into an equivalent parameter constraint of the form $\theta \in \Theta_\gamma$, where Θ_γ is an appropriate monotonicity domain.

The first derivative of the ordered basis N_{k+1} can be written, using standard properties of B-spline functions (cf. Appendix B),[19] in the form

$$N_{k+1}^{(1)} = (0, N_k)\Delta_{k+1}, \quad (17)$$

where $N_k = (N_{-k+1, k}, \ldots, N_{\kappa, k})$ forms an ordered basis for $\mathscr{S}_{k-1}(\nu)$, defined over the same set of knots as N_{k+1},

$$\Delta_{k+1} = (L_{k+1}D_{k+1})^{-1} = D_{k+1}^{-1}L_{k+1}^{-1},$$

and

$$D_{k+1} = \frac{1}{k} \begin{bmatrix} \nu_0 - \nu_{-k} & & & 0 \\ & \nu_1 - \nu_{-k+1} & & \\ & & \ddots & \\ 0 & & & \nu_{k+\kappa} - \nu_\kappa \end{bmatrix}.$$

Similarly, the second derivative of the ordered basis N_{k+1} can be expressed in the form

$$N_{k+1}^{(2)} = (0, (0, N_{k-1})) \begin{bmatrix} 0 & 0 \\ 0 & \Delta_k \end{bmatrix} \Delta_{k+1}. \quad (18)$$

In terms of the representation (15), the first and second derivatives of $s(\cdot; \hat{\theta})$ are therefore

$$s^{(1)}(x; \hat{\theta}) = \hat{N}_{k+1}^{(1)}(x)\hat{\theta}, \quad (19)$$

$$s^{(2)}(x; \hat{\theta}) = \hat{N}_{k+1}^{(2)}(x)\hat{\theta}, \quad (20)$$

for all $x \in [a, b]$, where

$$\hat{N}_{k+1}^{(1)} = (0, N_k)D_{k+1}^{-1} \quad \text{and}$$

$$\hat{N}_{k+1}^{(2)} = (0, (0, N_{k-1})) \begin{bmatrix} 0 & 0 \\ 0 & D_k^{-1}L_k^{-1} \end{bmatrix} D_{k+1}^{-1}.$$

PROPOSITION 1 (FIRST-ORDER MONOTONICITY). *Let $k \geqslant 0$, $\nu \in \mathscr{V}_{a,b}$, and let $s \in \mathscr{S}_k(\nu)$ be given in the B-spline representation (15), $s = \hat{N}_{k+1}\hat{\theta}$, for some $\hat{\theta} = (\hat{\theta}_{-k}, \hat{\theta}_{-k+1}, \ldots, \hat{\theta}_\kappa)^T \in \mathbb{R}^{k+\kappa+1}$. Then*
  (i) *s is nondecreasing on $[a, b]$ if*

$$\hat{\theta}' = (\hat{\theta}_{-k+1}, \ldots, \hat{\theta}_\kappa)^T \geqslant 0; \quad (21)$$

  (ii) *if $k \in \{0, 1, 2\}$ and s is nondecreasing on $[a, b]$, then condition (21) must hold.*

Part (i) states that the first-order monotonicity constraint M1 is implied by the linear constraint (21). Part (ii) guarantees that for splines of at most second degree this condition is also necessary.

To ensure that the second derivative of a spline $s \in \mathscr{S}_k(\nu)$ is nonnegative it suffices to have $I_{k,0}\Delta_k'\hat{\theta}' \geqslant 0$, where

$$\Delta_k' = (\text{diag}(\nu_1 - \nu_{-k+1}, \ldots, \nu_{k+\kappa} - \nu_\kappa)L_k)^{-1} \quad (22)$$

and $I_{k,0}$ is the $(k + \kappa) \times (k + \kappa)$ identity matrix whose first element is zero.

PROPOSITION 2 (SECOND-ORDER MONOTONICITY). *Let $k \geqslant 1$, $\nu \in \mathscr{V}_{a,b}$, and let $s \in \mathscr{S}_k(\nu)$ be given in the B-spline representation (15), $s = \hat{N}_{k+1}\hat{\theta}$, for some $\hat{\theta} = (\hat{\theta}_{-k}, \hat{\theta}_{-k+1}, \ldots, \hat{\theta}_\kappa)^T \in \mathbb{R}^{k+\kappa+1}$. Then*
  (i) *$s^{(1)}$ is nondecreasing on $[a, b]$ if*

$$I_{k,0}\Delta_k'\hat{\theta}' \geqslant 0; \quad (23)$$

  (ii) *if $k \in \{1, 2, 3\}$ and $s^{(1)}$ is nondecreasing on $[a, b]$, then condition (23) must hold.*

Condition (23) is equivalent to

$$\frac{\hat{\theta}_{-k+1}}{\nu_1 - \nu_{-k+1}} \leqslant \frac{\hat{\theta}_{-k+2}}{\nu_2 - \nu_{-k+2}} \leqslant \cdots \leqslant \frac{\hat{\theta}_\kappa}{\nu_{k+\kappa} - \nu_\kappa}.$$

REMARK 9. Linear inequality conditions (similar to the ones in Propositions 1 and 2) on the B-spline coefficient $\hat{\theta}$ of a spline $s \in \mathcal{S}_k(\nu)$ can be obtained to guarantee that a higher-order derivative $s^{(r)}$, for some $r \in \{1, \ldots, k\}$, is nonnegative-valued a.e. on $[a, b]$.

COROLLARY 1 (MONOTONICITY DOMAIN). *Let $k \in \{1, 2\}$, $\nu \in \mathcal{V}_{a,b}$, and let $s \in \mathcal{S}_k(\nu)$ be given in the B-spline representation* (14), $s = N_{k+1}\theta$, *for some $\theta \in \mathbb{R}^{k+\kappa+1}$. For any $\gamma = (\gamma_1, \gamma_2)$ it is*

$$\min\{\gamma_1 s^{(1)}(x; \theta), \gamma_2 s^{(2)}(x; \theta)\} \geqslant 0$$

$$\Leftrightarrow \quad \theta \in \Theta_\gamma = \{\vartheta \in \mathbb{R}^{k+\kappa+1}: M_\gamma \vartheta \geqslant 0\},$$

where

$$M_\gamma = \begin{bmatrix} 0 & \gamma_1 I_k \\ 0 & \gamma_2 I_{k,0} \Delta'_k \end{bmatrix} L_{k+1}^{-1} \in \mathbb{R}^{2(k+\kappa) \times (k+\kappa+1)},$$

and I_k denotes the $(k+\kappa) \times (k+\kappa)$ identity matrix.

The last result provides a characterization of the monotonicity domain Θ_γ, combining the insights from the first two propositions. It gives necessary and sufficient parameter restrictions for the corresponding approximation functions to satisfy the monotonicity constraints M1 and M2.[20]

3.3. Robust Approximation Sets

Given a data sample $z = \{(x_j, y_j)\}_{j=1}^n$ (of size $n \geqslant 3$) and an allowable approximation error ε, we consider the subset

$$\widehat{\Theta}_{\varepsilon, \gamma}(z) = \{\theta \in \Theta_\gamma: e(\theta \mid z) \leqslant \varepsilon, s(\cdot; \theta) \in \mathcal{S}_k(\nu)\} \quad (24)$$

of the monotonicity domain Θ_γ, which we term a *robust approximation set*. This set contains all parameter vectors θ that are compatible with z and ε. Relative to the available data and the allowable approximation error, candidates for robust decisions (i.e., decisions that maximize a robustness criterion as introduced in §2.3) are judged against all parameters in Θ_γ (cf. Equations (8)–(10)).

PROPOSITION 3 (BOUNDEDNESS). *Let $k \in \{1, 2\}$. For any $\gamma \neq 0$ and nonnegative ε, the robust approximation set $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ is a bounded polyhedron.*

The last result ensures that no matter what the approximation error ε or the monotonicity constraints on the component function (in terms of $\gamma \neq 0$), the parameter set of interest is a compact subset of a finite-dimensional Euclidean space, allowing the use of standard numerical algorithms when solving the robust decision problem. In the case where $\gamma = 0$, e.g., when no structural knowledge about the monotonicity of the component function is available, the robust approximation set $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ may become unbounded.

Let $\underline{\varepsilon}_\gamma(z)$ denote the smallest error such that $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ is nonempty, i.e.,

$$\underline{\varepsilon}_\gamma(z) = \inf\{\varepsilon \geqslant 0: \widehat{\Theta}_{\varepsilon, \gamma}(z) \neq \varnothing\}. \quad (25)$$

If $\underline{\varepsilon}_\gamma(z) = 0$, then the approximation error is zero, and interpolation of the sample data is possible.[21]

To determine if for a given ε the robust approximation set is empty or not, we formulate the following "sample monotonicity conditions" in analogy to the original monotonicity conditions M1 and M2.

$\widehat{\text{M1}}$. For a given real constant γ_1 the sample z satisfies a "first-order divided-difference condition" of the form

$$\gamma_1 \delta_j(z) \equiv \gamma_1 \frac{y_{j+1} - y_j}{x_{j+1} - x_j} \geqslant 0 \quad (26)$$

for all $j \in \{1, \ldots, n-1\}$.

$\widehat{\text{M2}}$. For a given real constant γ_2, the sample z satisfies a "second-order divided-difference condition" of the form

$$\gamma_2 \delta_j^2(z) \equiv \gamma_2 \left(\frac{y_{j+1} - y_j}{x_{j+1} - x_j} - \frac{y_j - y_{j-1}}{x_j - x_{j-1}}\right) \geqslant 0 \quad (27)$$

for all $j \in \{2, \ldots, n-1\}$.

In general, there is no reason to believe that the (noisy) data points in z always satisfy the sample monotonicity conditions in Equations (26)–(27). As long as the approximation error is allowed to be positive, there is no need to satisfy them strictly. The following ε-relaxations, $\widehat{\text{M1}}_\varepsilon$ and $\widehat{\text{M2}}_\varepsilon$, determine if the sample z can be "adjusted" to a sample $\hat{z} = \{(x_i, \hat{y}_i)\}_{i=1}^n$ such that the adjusted sample $\hat{z}$ satisfies $\widehat{\text{M1}}$ and $\widehat{\text{M2}}$ and at the same time the distance between z and $\hat{z}$ (in the sup-norm) does not exceed ε. Each one of these can be conveniently formulated in terms of a linear program (LP).

$\widehat{\text{M1}}_\varepsilon$. The value of the LP

$$\begin{cases} \min_{t, \hat{y}} \sum_{\ell=1}^n t_\ell & \text{s.t.} & -\gamma_1(\hat{y}_{j+1} - \hat{y}_j) \leqslant t_j, \\ (1 \leqslant i \leqslant n) & & 0 \leqslant t_i, \\ (1 \leqslant j \leqslant n-1) & & -\varepsilon \leqslant w_i(\hat{y}_i - y_i) \leqslant \varepsilon, \end{cases}$$

is zero, given the constants $\varepsilon \geqslant 0$ and $\gamma_1 \in \{-1, 0, 1\}$.

$\widehat{\text{M2}}_\varepsilon$. The value of the LP

$$\begin{cases} \min_{t, \hat{y}} \sum_{\ell=1}^n t_\ell & \text{s.t.} & -\gamma_2 \left(\frac{\hat{y}_{j+1} - \hat{y}_j}{x_{j+1} - x_j} - \frac{\hat{y}_j - \hat{y}_{j-1}}{x_j - x_{j-1}}\right) \leqslant t_j, \\ (1 \leqslant i \leqslant n) & & 0 \leqslant t_i, \\ (2 \leqslant j \leqslant n-1) & & -\varepsilon \leqslant w_i(\hat{y}_i - y_i) \leqslant \varepsilon, \end{cases}$$

is zero, given the constants $\varepsilon \geqslant 0$ and $\gamma_2 \in \{-1, 0, 1\}$.

We say that the sample z satisfies the ε-relaxation of the sample monotonicity condition $\widehat{\text{M1}}$ or $\widehat{\text{M2}}$ if the optimal value of the corresponding LP vanishes. The sample z satisfies both ε-relaxations *jointly* if the LPs in $\widehat{\text{M1}}_\varepsilon$ and $\widehat{\text{M2}}_\varepsilon$ admit a common solution $(t; \hat{y}) = (t_1, \ldots, t_n; \hat{y}_1, \ldots, \hat{y}_n)$, and their respective optimal values are both zero. It is now possible to use these conditions to guarantee that the monotone approximation set $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ is nonempty.

PROPOSITION 4. *Let* $\gamma_1 \in \{-1, 0, 1\}$ *and* $\varepsilon \geqslant 0$ *be given. If the sample z satisfies the relaxed sample monotonicity condition* $\widehat{M1}_\varepsilon$*, then there exists* $\nu \in \mathcal{V}_{a,b}$ *such that* $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ *is nonempty for* $\gamma = (\gamma_1, 0)$.

REMARK 10. (i) Proposition 4 generalizes the main finding in Pruess (1993) for monotone spline interpolation from $k = 3$ to $k \geqslant 1$. (ii) The proof of our result is constructive in the sense that it provides a simple condition N1 for the knot vector ν, which can be used directly for a numerical implementation.

PROPOSITION 5. *Let* $\gamma \in \{-1, 0, 1\}^2$ *and* $\varepsilon \geqslant 0$ *be given. If the sample z satisfies the relaxed sample monotonicity conditions* $\widehat{M1}_\varepsilon$ *and* $\widehat{M2}_\varepsilon$ *jointly, then there exists* $\nu \in \mathcal{V}_{a,b}$ *such that* $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ *is nonempty.*

REMARK 11. As for Proposition 4, the proof of Proposition 5 is constructive, and provides an algorithm for finding an appropriate knot vector, in terms of conditions $\overline{N1}$ and $\overline{N2}$, which together are less restrictive than condition N1 mentioned in Remark 10.

4. Information for Robust Decisions

As pointed out in §2.3, conditional on observing the data sample $z = \{x_j, y_j\}_{j=1}^n$, an optimal robust decision $\hat{x}^*(z)$ together with an optimal approximation error $\hat{\varepsilon}^*(z)$ (cf. Remark 4) can be obtained by solving the nested pair of optimization problems (8)–(10). The robustness criterion generally depends on the decision maker's beliefs about the realization of the parameter vector θ. We now investigate two important issues, the belief update conditional on sample data, and the acquisition of new information by designing an optimal experiment.

4.1. Belief Update

The decision maker's prior beliefs about the distribution of the parameter vector θ on the monotonicity domain Θ_γ, given by the cdf $F(\theta)$, can be updated using the data sample z. In Remark 12 below we discuss how such beliefs can be obtained in practice. Let the updated beliefs be represented by the cdf $F(\theta \mid z)$. Furthermore, for any $\theta \in \Theta_\gamma$, let $h(z \mid \theta)$ be the probability density for the realization of the random data sample z conditional on the parameter vector θ. Because the y_js are independent given θ, by Remark 3 we have

$$h(z \mid \theta) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi}\hat{\sigma}_j} \exp\left[-\frac{(y_j - s(x_j; \theta))^2}{2\hat{\sigma}_j^2}\right].$$

Given $F(\theta)$ and $h(z \mid \theta)$, the updated beliefs are obtained by Bayes' rule,

$$dF(\theta \mid z) = \frac{h(z \mid \theta) dF(\theta)}{\int_{\Theta_\gamma} h(z \mid \vartheta) dF(\vartheta)}. \tag{28}$$

REMARK 12. In practice, the question might arise as to which prior beliefs $F(\theta)$ might plausibly be used for

determining the posterior distribution in Equation (28). We outline three (in part implicit) approaches.

(i) *Uniform prior on* Θ_γ. Without any available information, a decision maker might simply assume the (entropy-maximizing) uniform distribution over the monotonicity domain Θ_γ. The updated beliefs $\hat{F}(\theta \mid z)$ over the robust approximation set $\widehat{\Theta}_{\varepsilon, \gamma}(z) \subset \Theta_\gamma$ are given by

$$\hat{F}(\theta \mid z) = \mathbf{1}_{\{\theta \in \widehat{\Theta}_{\varepsilon, \gamma}(z)\}} \frac{F(\theta \mid z)}{P(\tilde{\theta} \in \widehat{\Theta}_{\varepsilon, \gamma}(z))}. \tag{29}$$

(ii) *Uniform posterior on iso-error sets.* If no other information about the structural properties of the component function is available, the decision maker may want to consider all parameter realizations θ that produce a given approximation error, i.e., are such that all

$$\theta \in \Xi_{\varepsilon, \gamma}(z) \equiv \{\vartheta \in \Theta_\gamma: e(\vartheta \mid z) = \varepsilon\},$$

are equally likely. Therefore, the updated beliefs can be written as

$$dF(\theta \mid z) = \int_{\mathbb{R}_+} \lambda(\theta \mid \varepsilon, z) \, d\varphi(\varepsilon \mid z), \tag{30}$$

where $\lambda(\cdot \mid \varepsilon, z)$ is the standard Lebesgue measure over $\Xi_{\varepsilon, \gamma}(z),$[22] and for $\varepsilon \geqslant 0$:

$$\varphi(\varepsilon \mid z) = P\left(\max_{j \in \{1, \ldots, n\}} w_j |y_j - s(x_j; \tilde{\theta})| \leqslant \varepsilon \mid z\right).$$

Assuming that $f \in \mathcal{S}_k(\nu)$ (motivated by Lemma 1), the error probability φ can be derived from the probability distribution function of $\tilde{\epsilon}_j = y_j - s(x_j; \tilde{\theta})$. By Remark 3 the independent error terms $\tilde{\epsilon}_j$ are normally distributed with mean zero and variance $\hat{\sigma}_j^2$. Thus, $\varphi(\varepsilon \mid z)$ can be written in the form

$$\varphi(\varepsilon \mid z) = \prod_{j=1}^n P(w_j^2(y_j - s(x_j; \tilde{\theta}))^2 \leqslant \varepsilon^2).$$

Note that each $P(w_j^2(y_j - s(x_j; \tilde{\theta}))^2 \leqslant \varepsilon^2)$ is in fact a χ_1^2 distribution with mean $w_j^2 \sigma_j^2$ and variance $2w_j^2 \sigma_j^2$. The beliefs $\hat{F}(\theta \mid z)$ over $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ can be obtained from Equation (29) as in (i).

(iii) *Uniform posterior on iso-difference sets.* Similar to the last method, the decision maker may consider all parameter realizations θ, satisfying

$$\theta \in \hat{\Xi}_{\eta, \gamma} = \{\vartheta \in \Theta_\gamma: s(x_j; \vartheta) - y_j = \eta_j, \, 1 \leqslant j \leqslant n\},$$

i.e., located at the same position relative to the sample z, as equally likely. The updated beliefs are then

$$dF(\theta \mid z) = \int_{\mathbb{R}^n} \lambda(\theta \mid \eta, z) \, d\varphi(\eta \mid z),$$

where the measure λ is as in (ii), and

$$\varphi(\eta \mid z) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi}\hat{\sigma}_j} \exp\left[-\frac{(\eta_j)^2}{2\hat{\sigma}_j^2}\right].$$

The beliefs $\hat{F}(\theta \mid z)$ over $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ are obtained from Equation (29) as in (i).

4.2. Information Acquisition

Let us now consider the question of how to best improve the quality of the optimal decision by selecting from all available experiments one of maximum value. For this we consider a family of available experiments $\mathscr{F} = \{\tilde{r}_\alpha\}_{\alpha \in A}$ indexed by α in the compact set $A \subset [0, 1]$. Each experiment $\tilde{r}_\alpha$ is a random variable with realizations in the (nonempty, measurable) sample space R. Conditional on a true parameter vector $\theta \in \Theta_\gamma$, it is distributed with the conditional cdf $\rho_\alpha(r \mid \theta)$. Let $\hat{X}_0(\varepsilon, z)$ be the set of "default" decisions maximizing a robustness criterion $\mathscr{R}$ (cf. §2.3), so that by Equation (10)

$$x_0 \in \hat{X}_0(\varepsilon, z)$$
$$\Rightarrow \max_{x \in X} \mathscr{R}[g(s(x; \theta), x) \mid \theta \in \hat{\Theta}_{\varepsilon, \gamma}(z), z]$$
$$= \mathscr{R}[g(s(x_0; \theta), x_0) \mid \theta \in \hat{\Theta}_{\varepsilon, \gamma}(z), z] \equiv \mathscr{R}_0^*(\varepsilon, z);$$

and let $\hat{X}_{\alpha, r}(\varepsilon, z)$ be the set of "contingent" decisions maximizing the robustness criterion conditional on having observed the realization r of the signal $\tilde{r}_\alpha$, so that

$$x_{\alpha, r} \in \hat{X}_{\alpha, r}(\varepsilon, z) \Rightarrow \mathscr{R}_{\alpha, r}^*(\varepsilon, z)$$
$$= \mathscr{R}[g(s(x_{\alpha, r}; \theta), x_{\alpha, r}) \mid \theta \in \hat{\Theta}_{\varepsilon, \gamma}(z), z, \tilde{r}_\alpha = r],$$

where

$$\mathscr{R}_{\alpha, r}^*(\varepsilon, z) \equiv \max_{x \in X} \mathscr{R}[g(s(x; \theta), x) \mid \theta \in \hat{\Theta}_{\varepsilon, \gamma}(z), z, \tilde{r}_\alpha = r],$$

for all $r \in R$. From Equations (8) and (9) we obtain the corresponding optimal values of the robustness criterion,

$$\bar{\mathscr{R}}_0^*(z) = \max_{x \in \bar{X}_0(z)} \mathscr{R}[g(s(x; \theta), x) \mid \theta \in \Theta_\gamma],$$

$$\bar{\mathscr{R}}_{\alpha, r}^*(z) = \max_{x \in \bar{X}_{\alpha, r}(z)} \mathscr{R}[g(s(x; \theta), x) \mid \theta \in \Theta_\gamma],$$

where

$$\bar{X}_0(z) = \bigcup_{\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]} \hat{X}_0(\varepsilon, z) \quad \text{and} \quad \bar{X}_{\alpha, r}(z) = \bigcup_{\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]} \hat{X}_{\alpha, r}(\varepsilon, z).$$

Remark 4 yields the optimal approximation errors $\hat{\varepsilon}_{\alpha, r}^*(z)$ and $\hat{\varepsilon}_0^*(z)$ corresponding to the contingent and default solutions $\hat{X}_{\alpha, r}(\varepsilon, z)$ and $\hat{X}_0(\varepsilon, z)$. The expected value of the experiment $\tilde{r}_\alpha$ is then

$$V(\alpha \mid z) = \int_R (\mathscr{R}_{\alpha, r}^*(\hat{\varepsilon}_r^*(z), z) - \mathscr{R}_0^*(\hat{\varepsilon}_0^*(z), z)) \, d\bar{\rho}_\alpha(r \mid z),$$

where $\bar{\rho}_\alpha(r \mid z)$ is the cdf for the realizations of the signal $\tilde{r}_\alpha$ obtained by

$$\bar{\rho}_\alpha(r \mid z) = \int_{\Theta_\gamma} \rho_\alpha(r \mid \theta, z) \, dF(\theta \mid z) = \int_{\Theta_\gamma} \rho_\alpha(r \mid \theta) \, dF(\theta \mid z),$$

with conditional cdf $F(\theta \mid z)$ as in §4.1. An optimal experiment maximizes the value of information over all available experiments. Correspondingly, we refer to finding

$$\alpha^*(z) \in \arg\max_{\alpha \in A} V(\alpha \mid z)$$

as the decision maker's *(robust) information acquisition problem*. The solution to this problem (which always exists, as long as $V(\alpha \mid z)$ is continuous in α, for A is compact) defines an optimal experiment in terms of its expected improvement to the robustness criterion.

5. Application: Optimal Debt-Settlement Offers

In October 2009, the Federal Reserve Statistical Release (G.19) reported the 2008 U.S. total consumer credit outstanding in excess of 2.5 trillion dollars, of which about 40% is revolving debt such as unpaid credit-card balances. The debt-collection sector of the credit-card industry, which is concerned with the retrieval of outstanding loans from "delinquent" accounts (with an outstanding repayment significantly past the agreed due dates), can be viewed as a multibillion dollar industry in and of itself. A credit-card issuer's debt-collection process typically consists of different collection phases that vary among the various account segments (e.g., according to the holder's FICO score, the outstanding balance, the account age, or business versus personal use). A common feature across all segments and collection phases is that a large fraction (typically in the order of 40%–50%) of all delinquent accounts fail to repay *any* of their debt. Moreover, a delinquent account holder's assets tend to "decay" due to financial distress, bankruptcy costs, and claims by other lenders, which implies declining prospects for collecting the outstanding amount as time passes. Hence, credit-card issuers increasingly look favorably upon the option of extending early settlement offers to delinquent account holders. The design of optimal settlement offers can be interpreted as a revenue management problem, in which a bank sets the settlement rate so as to maximize the product of collected amount (corresponding to rate times outstanding balance) and a response rate (describing the likelihood of payback).

Optimization Model. We consider two segments of accounts. Each segment is modeled in the aggregate with the total outstanding balance and hidden characteristics such as willingness and ability to pay off the balance. These segment-specific characteristics are summarized by a "type" vector θ, which from the bank's point of view is distributed on an appropriate type space Θ. The type vector describes how a segment's average response rate, $f(x) \approx s(x; \theta) \in [0, 1]$, decreases in the settlement rate $x \in [0, 1]$, where $s(0; \theta) = 1$. For a given data sample $z = \{(x_j, y_j)\}_{j=1}^n$ and an allowable approximation error ε the relevant set of types is $\hat{\Theta}_{\varepsilon, \gamma}(z) \subset \Theta_\gamma$, where $\gamma = (-1, 0)$. The bank's goal is to maximize its aggregate payoff $g(s(x; \theta), x) = xs(x; \theta)$ using an appropriate robustness criterion (cf. §2.3).

The Data Set. Data obtained from a major credit-card company, covering the time between 2005 and 2007, includes full financial information for different types of credit-card accounts, segmented by attributes such as FICO score, line of credit, outstanding balance, or mortgage status. We focus on two segments of a charge-card product consisting of about 12,000 and 33,000 similar accounts,[23] respectively, each subdivided into three smaller groups. In the settlement-offer experiment two groups were offered a reduced settlement rate, and the remaining group served

Table 1. Summary statistics as a function of settlement rate: (a) sample size; (b) response rate.

(a)

	Settlement rate		
	0.49	0.60	1
Segment 1	1,964	8,244	2,320
Segment 2	5,032	22,710	5,561

(b)

Segment 1 (%)	4.74	4.36	2.87
Segment 2 (%)	9.96	7.43	4.57

as control. The response of an account holder in a given group is considered positive if the demanded portion of the outstanding balance is fully paid in a timely manner. Averaging the individual responses in each group (cf. Remark 3) yields the response rates in Table 1(b). Due to the costliness of settlement-offer trials, and a potentially negative reputation effect, the available experiments are sparse in X, with only $n = 4$ data points available.

Numerical Analysis. The response-rate samples in Table 1 satisfy $\widehat{\text{M1}}$ and $\widehat{\text{M2}}$ for $\gamma = (-1, 0)$, so that by Proposition 5 there exists a knot vector ν such that the robust approximation set $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ is nonempty for any $\varepsilon \geqslant 0$. Indeed, for $k = 2$, the vector $\nu = (0, 0.2, 0.4, 0.6, 0.8, 1)$ guarantees that $\underline{\varepsilon}_\gamma(z) = 0$ in Equation (25), i.e., interpolation is possible. Moreover, by Proposition 3 the set $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ is a bounded polyhedron, described by the following set of linear inequalities,

$$\begin{cases} \|W(\bar{N}_{k+1}\theta - Y)\|_\infty \leqslant \varepsilon, \\ M_\gamma \theta \geqslant 0, \end{cases}$$

where $\bar{N}_{k+1}$ and M_γ are specified in the proof of Proposition 3 and Remark 9, respectively, $Y = (y_1, \ldots, y_n)^T$ is a data vector, and $W = \text{diag}(w_1, \ldots, w_n)$ is a diagonal matrix of positive weights.[24] For any $\varepsilon \geqslant 0$, the elements of $\widehat{\Theta}_{\varepsilon, \gamma}(z)$ can be obtained as a convex hull of its vertices (or "vertex solutions"). Figure 4 depicts the vertex solutions as well as their associated response and spin curves for the two account segments, when $\varepsilon = 0$. Figure 5 shows both the response-rate and spin distributions, based on a uniform posterior on iso-difference sets (cf. Remark 12(iii)), inside a 99.9% confidence region (corresponding to $\bar{\varepsilon}$).

Optimal Robust Decisions. We choose $\bar{\varepsilon} \in \{3\%, 4.2\%\}$ so as to provide 99.9% confidence (with the smaller of the two values for Segment 2; cf. Figure 1(b)), and then solve Equations (8)–(10) to determine an optimal robust decision with respect to average performance and competitive ratio as criteria. Figure 6 shows the robustness criteria used in Equation (10) for $\varepsilon = \bar{\varepsilon}$. As illustrated, Segment 1's expected spin (respectively, competitive ratio) is maximized

at a settlement rate of 18% (respectively, 17%). For Segment 2 both robustness criteria are maximized at a settlement rate of 20%. Interpreting the competitive ratio, the expected income at the optimal settlement rate is on average about 90% of the highest achievable repayment, in either segment.

The solution to the optimization problem (10) is depicted in Figure 8(a) for Segment 2 when $\varepsilon \in [0, \bar{\varepsilon}]$ and $\bar{\varepsilon} = 3\%$. The optimal decision $\widehat{x}^*(z)$ is unique and does not vary for $\varepsilon \in [0, \bar{\varepsilon}]$ due to the quality of the available data. This implies that the decision maker can be quite confident in the quality of $\widehat{x}^*(z) = 20\%$ as an optimal robust settlement offer.[25] Indeed, this offer increases the bank's expected payoff by at least 60% over the best available settlement-rate experiment. Figure 8(b) shows the average-performance robustness criteria in problems (8) and (10) together for Segment 2. It can be seen that without any data, a settlement rate of about 65% becomes optimal. Interestingly, this rate is consistent with one of the bank's initial experiments of offering to settle at a rate of 60%.

Distribution of Optimal Robust Decisions. The probability measure $d\widehat{F}_x$ for the distribution of the maximizer $x(\theta) \in \arg\max_{x \in X} g(s(x; \theta), x)$, for $\theta \in \widehat{\Theta}_{\varepsilon, \gamma}(z)$, can be obtained from the measure $d\widehat{F}(\theta \mid z)$ in §4.1. Because $g(s(x; \theta), x)$ is continuous in θ, the support of $d\widehat{F}_x$ is compact. For $\varepsilon = \bar{\varepsilon}$, and given a uniform posterior on iso-difference sets (cf. Remark 12(iii)), Figure 7 illustrates the distribution of settlement offers that are optimal for some parameter realizations.

Information Acquisition. Figure 9 compares the expected spin if another data point at a different settlement rate were available for either segment. As an example of how to interpret the curves, Figure 9(b) shows that if we knew the response rate for a 15% settlement offer for Segment 2, then we could expect a 0.3% spin increase. The outstanding aggregate balance of the accounts under consideration amounts to more than $100 million dollars; a 0.3% spin increase therefore has a significant absolute impact on the bank's collection revenues. Using these curves, the solution to the information acquisition problem in §4.2 (for $A = [0, 1]$ and $\tilde{r}_\alpha$ as the response to offers at the rate of α) for Segment 1 (Segment 2) is to conduct an additional settlement experiment at the rate of 20% (15%).

6. Conclusion

The proposed robust approximation method has been shown to yield good results in sample-sparse environments. Our approach has two key features that aim at making maximum use of the available data points, as few as they may be. First, the method allows the direct use of structural knowledge about the problem, in the form of first- and second-order monotonicity constraints. It therefore supplements the data sample with information about the shape

Figure 4.   (a) Vertices of $\widehat{\Theta}_{0,(-1,0)}(z)$ with response rates and associated expected income (spin) for Segment 1. (b) Vertices of $\widehat{\Theta}_{0,(-1,0)}(z)$ with response rates and associated expected income (spin) for Segment 2.
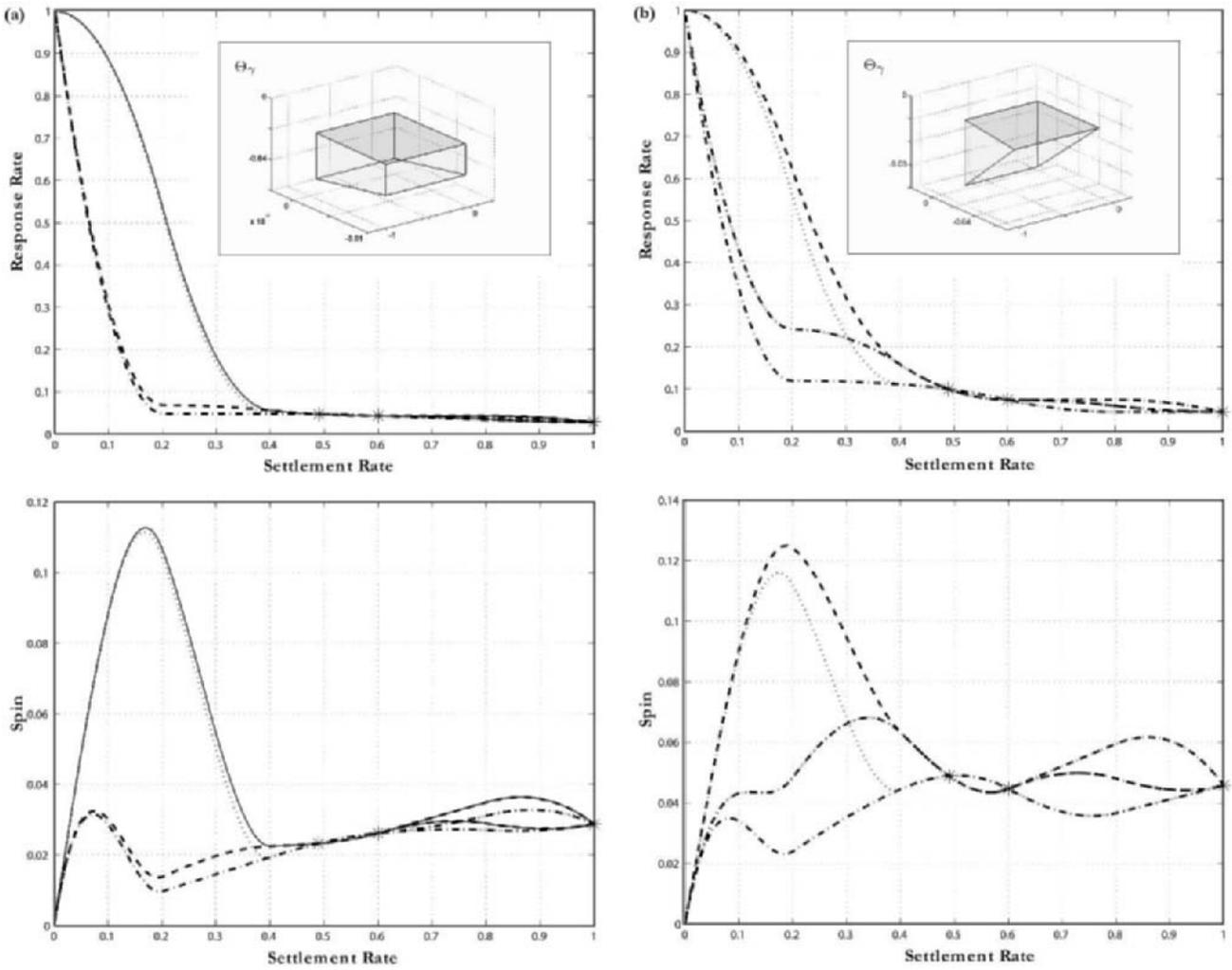

Figure 5.   (a) Distribution of response rates for Segment 2 (with $\varepsilon = \bar{\varepsilon}$). (b) Distribution of spins for Segment 2 (with $\varepsilon = \bar{\varepsilon}$).
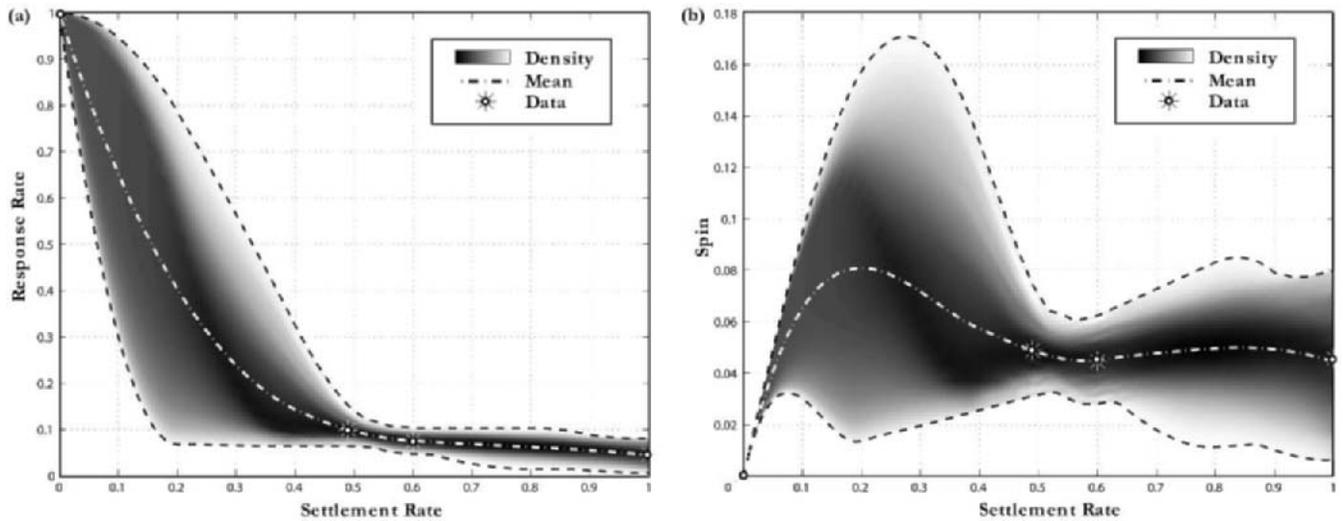
Figure 6. Expected spin and competitive ratio for settlement offers to Segment 1 and Segment 2 ((a)/(b)).
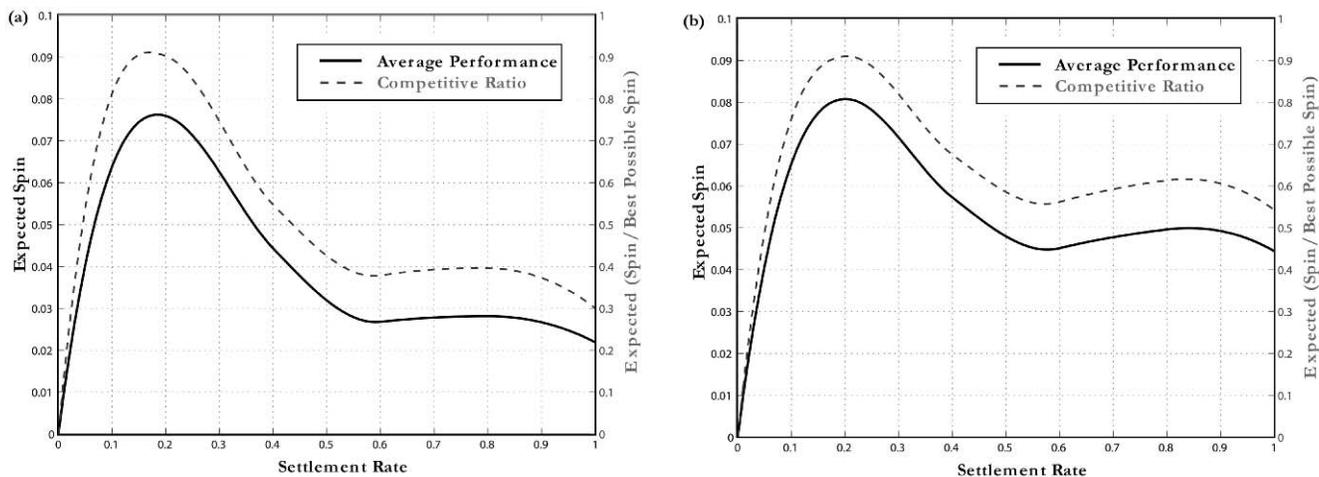
Figure 7. Probability density of optimal settlement offer to Segment 1 (a) and Segment 2 (b).
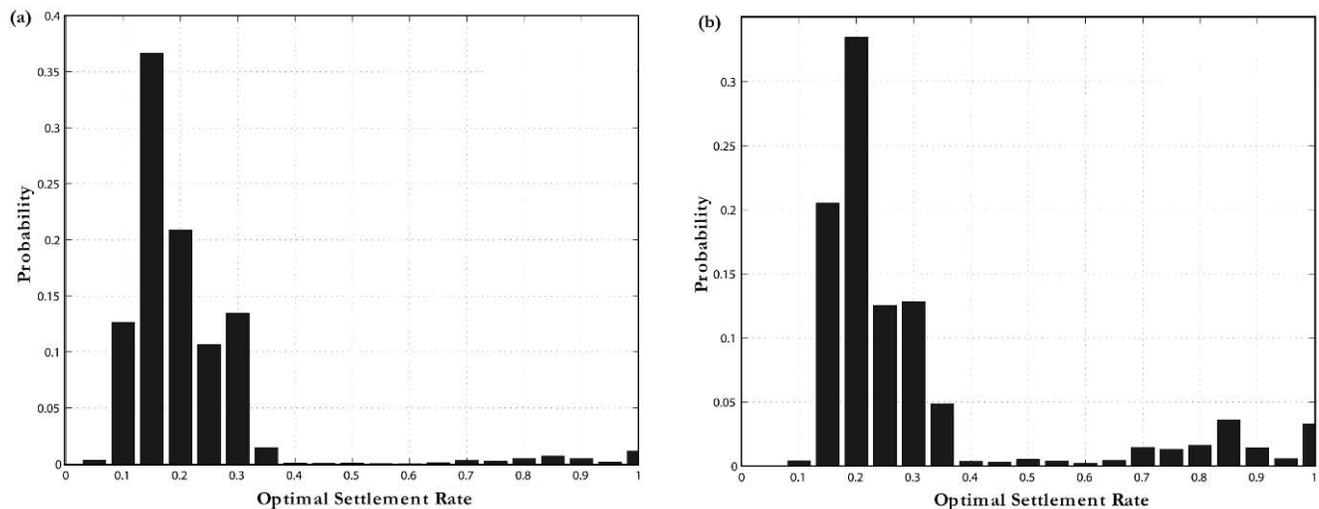
Figure 8. (a) Maximum expected spin and optimal settlement offer as a function of the allowable sample error (for Segment 2). (b) Robustness criteria in Equation (8) (maximized at $\hat{x}^*(\varnothing)$), and in Equation (10) with information obtained from the data sample z (maximized at $\hat{x}^*(z)$).
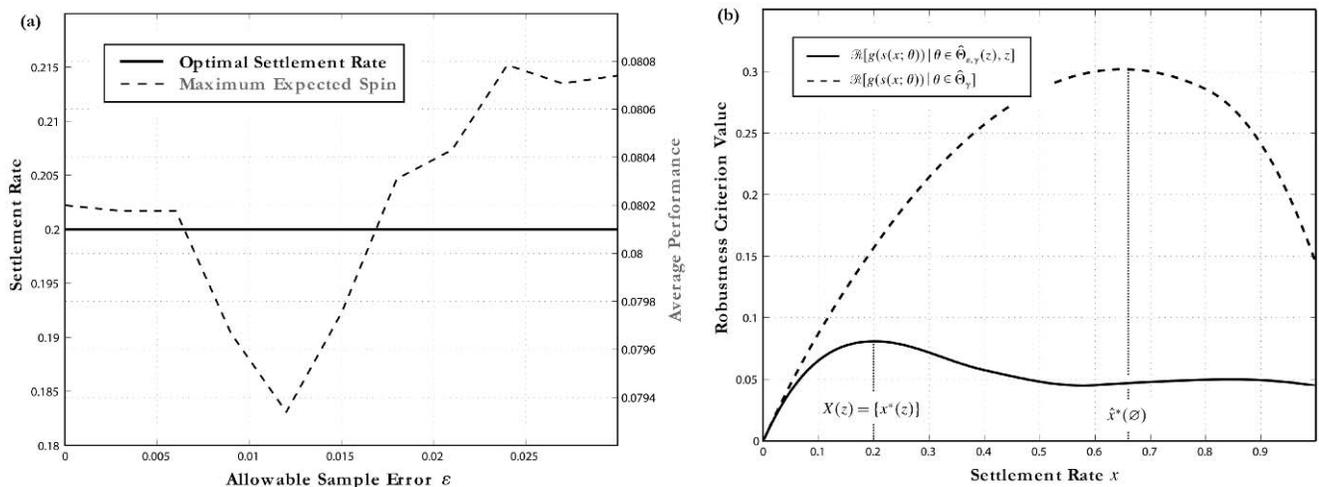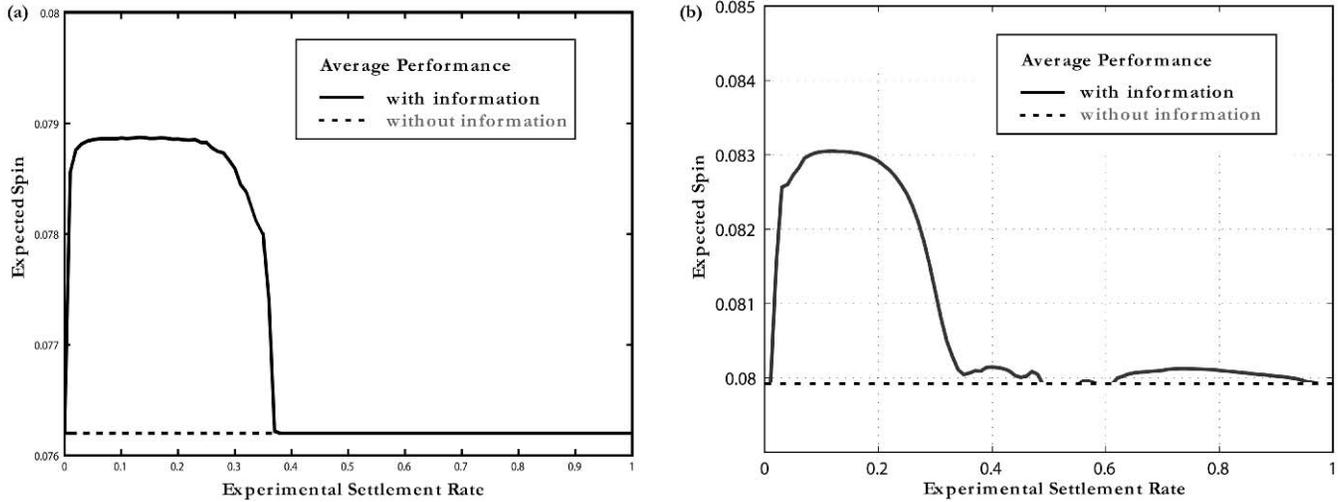
Figure 9.   (a) Expected spin with and without extra information at different settlement rates for Segment 1. (b) Expected spin with and without extra information at different settlement rates for Segment 2.

of the unknown functional relationship. In contrast to most of the extant literature on shape-preserving approximation, we do not constrain the data in order to guarantee the existence of shape-preserving approximations (see, e.g., Fritsch and Carlson 1980), but characterize the parameter domain consistent with the shape constraints as a function of the available data. For any given model error relative to the data, the resulting robust approximation set as a subset of the monotonicity domain is a bounded polyhedron.

The second key feature of the method is that it considers the shape-preserving approximation problem embedded in a decision problem. By linking the optimization and identification problems in Equations (8)–(10), an action and an approximation error (cf. Remark 4) are obtained simultaneously. All models consistent with the structural knowledge are considered up to the very point when the decision maker finally commits to an optimal robust action. An interesting consequence of this jointness is that to strive for the smallest-possible approximation error may not be optimal, given the goal of determining optimal robust decisions which perform well despite possible variations of the model specification. The relevant trade-off is between the increased robustness from considering a larger set of models when relaxing the approximation error, and an improved estimate of the objective function in the decision problem (2) when decreasing the approximation error. The practical application in §5 confirms that it is not always optimal to minimize the approximation error, which underlines the importance of the joint consideration of data and decisions. These findings relate to recent advances in operational statistics that seek to relate the error metric for approximation to the objective function. A contribution of this paper is to combine the use of structural knowledge about the problem (monotonicity constraints) with joint robust identification and decision making.

Appendix A. Proofs

Proof of Proposition 1

Let $k \geqslant 0$, $\nu \in \mathcal{V}_{a,b}$, and let $s \in \mathcal{S}_k(\nu)$ be given in the B-spline representation (15), so that $s = \hat{N}_{k+1}\hat{\theta}$.

(i) By Equation (19), because

$$\hat{N}_{k+1}^{(1)} = (0, N_k)D_{k+1}^{-1}$$
$$= (0, kN_{-k+1,k}/(\nu_1 - \nu_{-k+1}), \ldots, kN_{\kappa,k}/(\nu_{k+\kappa} - \nu_\kappa))$$
$$\geqslant 0,$$

condition (21) is clearly sufficient for the nonnegativity of $s^{(1)}$.

(ii) We now show that for $k \in \{0, 1, 2\}$, condition (21) is also necessary for $s = \hat{N}_{k+1}\hat{\theta}$ to be nondecreasing. Let $s \in \mathcal{S}_0(\nu)$. This means that s is a piecewise-constant function. Using Equation (14), we can write s as

$$s = N_1\theta,$$

where $\theta = (\theta_0, \ldots, \theta_\kappa)$ and $N_1 = [\mathbf{1}_{[\nu_0, \nu_1)}, \mathbf{1}_{[\nu_1, \nu_2)}, \ldots, \mathbf{1}_{[\nu_\kappa, \nu_{\kappa+1}]}]$ is an ordered basis for $\mathcal{S}_0(\nu)$, which partitions unity over $[a, b]$. Having $\hat{\theta}' \geqslant 0$ is equivalent to

$$\theta_0 \leqslant \theta_1 \leqslant \cdots \leqslant \theta_\kappa,$$

which is a necessary condition for s to be nondecreasing. Let $s \in \mathcal{S}_1(\nu)$. Using Equation (19), it is $s^{(1)} = (0, N_1)D_2^{-1}\hat{\theta}$, where D_2^{-1} is a diagonal matrix with positive entries. Therefore, $s^{(1)} \geqslant 0$ if and only if $\hat{\theta}' \geqslant 0$. Now let $s \in \mathcal{S}_2(\nu)$. Equation (19) implies that

$$s^{(1)} = (0, N_2)D_3^{-1}\hat{\theta},$$

where the components of N_2 are given by

$$N_{-1,2}(x) = \begin{cases} 1 - (x - \nu_0)/(\nu_1 - \nu_0), & \text{if } x \in [\nu_0, \nu_1], \\ 0, & \text{otherwise,} \end{cases}$$

for $i = -1$;

$$N_{i,2}(x) = \begin{cases} (x - \nu_i)/(\nu_{i+1} - \nu_i), & \text{if } x \in [\nu_i, \nu_{i+1}], \\ 1 - (x - \nu_{i+1})/(\nu_{i+2} - \nu_{i+1}), & \text{if } x \in [\nu_{i+1}, \nu_{i+2}], \\ 0, & \text{otherwise}; \end{cases}$$

for $i \in \{0, \ldots, \kappa - 1\}$, and

$$N_{\kappa,2}(x) = \begin{cases} (x - \nu_\kappa)/(\nu_{\kappa+1} - \nu_\kappa), & \text{if } x \in [\nu_\kappa, \nu_{\kappa+1}], \\ 0, & \text{otherwise}; \end{cases}$$

for $i = \kappa$. Therefore, at $x = \nu_{i+1}$, $i \in \{-1, \ldots, \kappa\}$ we have $s^{(1)}(\nu_{i+1}) = 2\hat{\theta}_i/(\nu_{i+2} - \nu_i)$, which yields the claim, i.e., $s^{(1)}(x) \geqslant 0$ implies that $\hat{\theta}' \geqslant 0$.
This completes our proof. $\square$

Proof of Proposition 2

The proof follows directly from the proof of Proposition 1. $\square$

Proof of Proposition 3

Using Corollary 1, for any given $\varepsilon \geqslant 0$, $\gamma \in \mathbb{R}^2$, $\nu \in \mathcal{V}_{a,b}$, and $z = \{(x_j, y_j)\}_{j=1}^n$, the set $\widehat{\Theta}_{\varepsilon,\gamma}(z)$ is characterized by the inequalities

$$\|W(\bar{N}_{k+1}\theta - Y)\|_\infty \leqslant \varepsilon,$$

$$M_\gamma \theta \geqslant 0,$$

where $W = \text{diag}(w_1, \ldots, w_n)$ is the (diagonal) matrix of weights, $Y^T = (y_1, \ldots, y_n)$ is a data vector, $\theta^T = (\theta_{-k}, \ldots, \theta_\kappa)$ is the parameter vector, and

$$\bar{N}_{k+1} = \begin{bmatrix} N_{-k,k+1}(x_1) & \cdots & N_{\kappa,k+1}(x_1) \\ \vdots & & \vdots \\ N_{-k,k+1}(x_n) & \cdots & N_{\kappa,k+1}(x_n) \end{bmatrix}. \tag{A1}$$

First note that $\widehat{\Theta}_{\varepsilon,\gamma}(z)$ is linearly constrained, and thus a polyhedron. We now show that $\widehat{\Theta}_{\varepsilon,\gamma}(z)$ is bounded. For this we assume that, without any loss in generality, $w_1 = w_2 = \cdots = w_n = 1$. Thus, if $\theta \in \widehat{\Theta}_{\varepsilon,\gamma}(z)$ and $\gamma \neq 0$, then the constant

$$M = \max\left\{ \frac{|y_n - y_1| + 2\varepsilon}{b - a}, \frac{|y_2 - y_1| + 2\varepsilon}{x_2 - a}, \frac{|y_n - y_{n-1}| + 2\varepsilon}{b - x_{n-1}} \right\}$$

$$\cdot (b - a) + \max\{|y_1|, |y_n|\}$$

bounds $|s(\,\cdot\,;\theta)|$ uniformly, i.e., is such that $|s(x;\theta)| \leqslant M$ for all $x \in X$. The fact that M does not depend on θ implies immediately that $\widehat{\Theta}_{\varepsilon,\gamma}(z)$ is bounded. $\square$

Proof of Proposition 4

We can restrict attention to the case where $\gamma_1 \geqslant 0$, because the proof is symmetric for $\gamma_1 < 0$. Without loss of generality, one can assume that the y_j-values of the sample $z = \{(x_j, y_j)\}_{j=1}^n$ are part of a solution to the LP in $\widehat{M1}_\varepsilon$, so that the sample monotonicity condition $\widehat{M1}$ holds. We now show that the following condition on the knot vector $\nu \in \mathcal{V}_{a,b}$ implies that the robust approximation set $\widehat{\Theta}_{\varepsilon,\gamma}(z)$ is nonempty.

N1. The knot vector $\nu = (\nu_0, \nu_1, \ldots, \nu_{\kappa+1}) \in \mathcal{V}_{a,b}$ is such that there is at most one data point in the interior of the support of each $N_{i,k+1}$ for all $i \in \{-k, \ldots, \kappa\}$.

By Proposition 1 it is enough to show that if a knot vector ν satisfies N1, then there exists a $\hat{\theta}$ such that

$$\begin{cases} \bar{N}_{k+1} L_k \hat{\theta} = Y, \\ \hat{\theta}' \geqslant 0, \end{cases} \tag{A2}$$

with $Y = (y_1, \ldots, y_n)^T$ and $\bar{N}_{k+1}$ as in Equation (A1). This solution can be constructed as follows. First, examine the matrix $\bar{N}_{k+1} L_k$: because N_{k+1} forms a basis that partitions unity on $[a, b]$, i.e., $N_{k+1}(1, \ldots, 1)^T = \mathbf{1}_{[a,b]}$, each row of $\bar{N}_{k+1} L_k$ has a structure similar to a cumulative distribution function, except that it starts from 1 and decreases to 0, i.e.,

$$\bar{N}_{k+1} L_k = \begin{bmatrix} 1 & \times & \cdots & \times & 0 & 0 & & \cdots & \\ 1 & & \cdots & 1 & 1 & \times & \cdots & \times & 0 & \cdots \\ & & & \cdots & & & & & \end{bmatrix}$$

where $\times$ denotes entries between 0 and 1, decreasing from left to right. The first relation in Equation (A2) becomes, therefore,

$$\begin{bmatrix} 1 & \times & \cdots & \times & 0 & 0 & & \cdots & \\ 1 & & \cdots & 1 & 1 & \times & \cdots & \times & 0 & \cdots \\ & & & \cdots & & & & & \end{bmatrix} \begin{bmatrix} \hat{\theta}_{-k} \\ \vdots \\ \hat{\theta}_\kappa \end{bmatrix}$$

$$= \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{A3}$$

If we set $\hat{\theta}_{-k} = y_1$, then $\hat{\theta}_i \geqslant 0$ for $i \in \{-k+1, \ldots, \kappa\}$ (obtained from the second relation in Equation (A2)) implies that $\hat{\theta}_i = 0$ whenever $N_{i,k+1}(x_1) \neq 0$. By suppressing the columns corresponding to $\hat{\theta}_{-k}$ and the zero $\hat{\theta}_i$s and then subtracting $(y_1, y_1, \ldots, y_1)^T$ from both sides, we obtain (dropping the first row) an equation of exactly the same structure as Equation (A3). We can therefore repeat this procedure, similar to Gauss elimination, and thus determine all components of $\hat{\theta}$. Note that the right-hand side of Equation (A3) after any step is nonnegative as a consequence of $\widehat{M1}$. This in turn implies that $\hat{\theta}' \geqslant 0$, so that a solution to Equation (A2) has been obtained. $\square$

Proof of Proposition 5

Analogous to the proof of Proposition 4, we can, without any loss in generality, assume that the sample $z = \{(x_j, y_j)\}_{j=1}^n$ satisfies the sample monotonicity conditions $\widehat{M1}$ and $\widehat{M2}$. Furthermore, we assume that $\gamma_2 = 1$. For $\gamma_2 = 0$, the result specializes to Proposition 4. The proof proceeds symmetrically for $\gamma_2 = -1$. Note first that in the following construction condition $\widehat{M1}$ implies condition M1, provided that M2 holds. Therefore, we need only to show that there exists $\nu \in \mathcal{V}_{a,b}$ such that the system

$$\begin{cases} \bar{N}_{k+1} L_k \hat{\theta} = Y, \\ I_{k,0} \Delta'_k \hat{\theta}' \geqslant 0, \end{cases} \tag{A4}$$

where $Y = (y_1, \ldots, y_n)^T$ and $\bar{N}_{k+1}$ as in Equation (A1), possesses a solution. Without loss of generality, we assume that n is odd,[26] $x_1 \in [\nu_0, \nu_1]$, and (x_0, y_0) with $x_0 < \nu_0$ is such that the augmented sample $\hat{z} = \{(x_j, y_j)\}_{j=0}^n$ satisfies $\widehat{M1}$ and $\widehat{M2}$. Furthermore, for all $j \in \{1, \ldots, \lfloor n/2 \rfloor\}$ we denote the intersection point of the straight line through the points $(x_{2(j-1)}, y_{2(j-1)})$ and (x_{2j-1}, y_{2j-1}) with the straight line through the points (x_{2j}, y_{2j}) and (x_{2j+1}, y_{2j+1}), by $(\hat{x}_j, \hat{y}_j)$. First, we show that there exists $\nu \in \mathcal{V}_{a,b}$ such that Equation (A4) has a solution for $k = 1$, i.e., there exists a piecewise-linear function, $s_1(x; \theta)$, with breakpoints in $\{a = \nu_0, \nu_1, \ldots, \nu_\kappa, \nu_{\kappa+1} = b\}$, that passes through (x_j, y_j), $j \in \{1, \ldots, n\}$ and is consistent with M1 and M2. Then we derive a sufficient condition over ν that makes it possible to smooth the piecewise-linear function $s_1(x; \theta)$ to obtain an element of $\mathcal{S}_k(\nu)$. Consider the case where $k = 1$.

$\overline{N1}$. The knot vector $\nu = (\nu_0, \nu_1, \ldots, \nu_{\kappa+1}) \in \mathcal{V}_{a,b}$ satisfies the following two conditions:

(i) $k+1$ (or more) knots are placed between x_{2j-1} and x_{2j} for all $j \in \{1, \ldots, \lfloor n/2 \rfloor\}$; and

(ii) $(\hat{x}_j, \hat{y}_j)$ is between ν_{ℓ_j} and ν_{u_j}, where $\ell_j = \min_{\nu_i \geqslant x_{2j-1}} i$ and $u_j = \max_{\nu_i \leqslant x_{2j}} i$.

Under condition $\overline{N1}$, the parameter vector θ, specified hereafter, defines a piecewise-linear function, $s_1(x; \theta) = N_2(x)\theta$, that satisfies Equation (A4),

$$\theta_i = \frac{y_{2(j-1)} - y_{2j-1}}{x_{2(j-1)} - x_{2j-1}} (\nu_{i+1} - x_{2(j-1)}) + y_{2(j-1)},$$

where $\hat{x}_{j-1} \leqslant \nu_i \leqslant \hat{x}_j$ for $(i, j) \in \{-1, \ldots, \kappa\} \times \{1, \ldots, \lfloor n/2 \rfloor + 1\}$, setting $\hat{x}_0 = a$ and $\hat{x}_{\lfloor n/2 \rfloor + 1} = b$. The intuition behind this construction is shown in Figure A.1 (for $n = 5$).

Now let $k > 1$; we show that there exists a spline $s_k(x; \vartheta) \in \mathcal{S}_k(\nu)$ that is comprised of the line segments connecting $(x_{2(j-1)}, y_{2(j-1)})$ with (x_{2j-1}, y_{2j-1}), for all $j \in \{1, \ldots, \lfloor n/2 \rfloor + 1\}$. Hence, we need only to show that there is a spline in $\mathcal{S}_k(\hat{\nu}_j)$ with $\hat{\nu}_j = (\nu_{\ell_j}, \ldots, \nu_{u_j})$ for $j \in \{1, \ldots, \lfloor n/2 \rfloor\}$ satisfying the following requirements. First, it connects the line segments $s_1(x; \theta)|_{[\nu_{u_{j-1}}, \nu_{\ell_j}]}$ and

Figure A.1. Construction in the proof of Proposition 5.
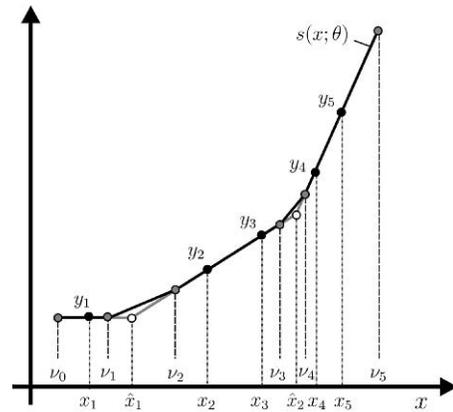
$s_1(x; \theta)|_{[\nu_{u_j}, \nu_{\ell_{j+1}}]}$, where $\nu_{u_0} = \nu_0$, and $\nu_{\ell_{\lfloor n/2 \rfloor + 1}} = \nu_{\kappa+1}$. Second, it has $k - 1$ continuous derivatives at the points ν_{ℓ_j} and ν_{u_j}.

Using the B-spline representation of the second derivative of $s_k(x; \vartheta)$ in $\mathcal{S}_{k-2}(\hat{\nu}_j)$, the above two requirements become

$$\int_{\nu_{\ell_j}}^{\nu_{u_j}} s_k^{(2)}(\xi; \vartheta)\, d\xi = \frac{y_{2j+1} - y_{2j}}{x_{2j+1} - x_{2j}} - \frac{y_{2j-1} - y_{2(j-1)}}{x_{2j-1} - x_{2(j-1)}}, \tag{A5}$$

$$\int_{\nu_{\ell_j}}^{\nu_{u_j}} \left(\int_{\nu_{\ell_j}}^{x} s_k^{(2)}(\xi; \vartheta)\, d\xi \right) dx$$

$$= s_1(\nu_{u_j}; \theta) - s_1(\nu_{\ell_j}; \theta) - \frac{y_{2j-1} - y_{2(j-1)}}{x_{2j-1} - x_{2(j-1)}} (\nu_{u_j} - \nu_{\ell_j}), \tag{A6}$$

$$\frac{d^i}{dx^i} s_k^{(2)}(x; \vartheta)\Big|_{\nu_{\ell_j}} = 0, \quad \text{and} \quad \frac{d^i}{dx^i} s_k^{(2)}(x; \vartheta)\Big|_{\nu_{u_j}} = 0,$$
$$0 \leqslant i < k - 2, \tag{A7}$$

$$s_k^{(2)}(x; \vartheta) \geqslant 0, \quad \forall x \in [\nu_{\ell_j}, \nu_{u_j}], \tag{A8}$$

where $s_k^{(2)}(x; \vartheta) = \sum_{l=-k+2}^{k-1} N_{l,k-1}(x)\vartheta_l$. Equation (A7) forces $\vartheta_{-k+2}, \ldots, \vartheta_{-1}$ and $\vartheta_2, \ldots, \vartheta_{k-1}$ to zero. Therefore, $s_k^{(2)}(x; \vartheta)$ can be written in the form

$$s_k^{(2)}(x; \vartheta) = N_{0,k-1}(x)\vartheta_0 + N_{1,k-1}(x)\vartheta_1.$$

Using the basic properties of B-splines (cf. Appendix B), we can write Equations (A5) and (A6) as

$$\frac{\nu_{u_j-1} - \nu_{\ell_j}}{k-1} \vartheta_0 + \frac{\nu_{u_j} - \nu_{j_{\ell}+1}}{k-1} \vartheta_1 = \delta_{2j}(z) - \delta_{2(j-1)}(z), \tag{A9}$$

and

$$\frac{\nu_{u_j-1} - \nu_{j_\ell}}{k-1} A\vartheta_0 + \frac{\nu_{u_j} - \nu_{j_\ell+1}}{k-1} B\vartheta_1$$

$$= \overbrace{\frac{s_1(\nu_{u_j}; \theta) - s_1(\nu_{\ell_j}; \theta)}{\nu_{u_j} - \nu_{\ell_j}}}^{\equiv \delta_{u_j, \ell_j}(s_1)} - \delta_{2(j-1)}(z), \tag{A10}$$

where

$$A = \sum_{m=0}^{k-2} \left(\frac{1}{k-1-m} \left(\frac{1}{k-m} \sum_{i=0}^{k-1-m} \frac{\nu_{\ell_j+m+i} - \nu_{\ell_j+m}}{\nu_{u_j} - \nu_{\ell_j}} \right) \right)$$
$$+ \frac{\nu_{u_j} - \nu_{u_j-1}}{\nu_{u_j} - \nu_{\ell_j}},$$

and

$$B = \sum_{m=0}^{k-2} \left(\frac{1}{k-1-m} \left(\frac{1}{k-m} \sum_{i=0}^{k-1-m} \frac{\nu_{\ell_j+m+i+1} - \nu_{\ell_j+m+1}}{\nu_{u_j} - \nu_{\ell_j}} \right) \right).$$

Note that $0 \leqslant \delta_{u_j, \ell_j}(s_1) - \delta_{2(j-1)}(z) \leqslant \delta_{2j}(z) - \delta_{2(j-1)}(z)$, because s_1 is convex. Furthermore, there always exists a knot vector ν that satisfies $\overline{\text{N1}}$ and the following condition.

$\overline{\text{N2}}$. The knot vector $\nu = (\nu_0, \ldots, \nu_{\kappa+1}) \in \mathcal{V}_{a,b}$ is such that

$$B \leqslant \frac{\delta_{u_j, \ell_j}(s_1) - \delta_{2(j-1)}(z)}{\delta_{2j}(z) - \delta_{2(j-1)}(z)} \leqslant A, \quad j \in \{1, \ldots, \lfloor n/2 \rfloor\}.$$

This guarantees the existence of a nonnegative solution $(\vartheta_0, \vartheta_1)$ to Equations (A9)–(A10) and completes our proof. $\square$

Appendix B. Basic Properties of B-Splines

For a more complete list of properties, see, e.g., Dierckx (1993).

1. *Positivity*:

$$N_{i, k+1}(x) \geqslant 0 \quad \text{for all } x.$$

2. *Local support*:

$$N_{i, k+1}(x) = 0 \quad \text{if } x \notin [\nu_i, \nu_{i+k+1}].$$

3. *Boundary values*:

$$N_{i, k+1}^{(l)}(\nu_i) = N_{i, k+1}^{(l)}(\nu_{i+k+1}) = 0, \quad l \in \{0, \ldots, k-1\}.$$

4. *Derivative of a B-spline*:

$$N_{i, k+1}'(x) = k \left\{ \frac{N_{i, k}(x)}{\nu_{i+k} - \nu_i} - \frac{N_{i+1, k}(x)}{\nu_{i+k+1} - \nu_{i+1}} \right\}.$$

Endnotes

1. For a general discussion of semiparametric models and their identification, see, e.g., Powell (1994). The standard econometric approach revolves around estimation, whereas our method is fundamentally aimed at solving decision problems robustly.

2. Constraints on the sign of higher-order derivatives can be imposed in a similar manner.

3. This work is preceded by Wolibner (1951), who establishes the existence of monotonic interpolating polynomials.

4. The definition of B-splines dates back to Curry and Schoenberg (1947).

5. From a technical perspective, we relax the almost ubiquitous assumption in extant results on shape-preserving spline approximation that the knots be located at the data points. The restriction of intersample behavior implied by this assumption artificially constrains the space of feasible models. In view of achieving robustness, being able to place knots at any location in the choice set is key.

6. See, e.g., Judd (1999, Chapters 6 and 12) for a brief overview of shape-preserving numerical methods in economics.

7. See, e.g., Härdle and Linton (1994) and Pagan and Ullah (1999) for surveys of nonparametric estimation techniques.

8. The B-spline approximation method presented in this paper can, in the limit, be interpreted as an orthogonal-series method, for the B-spline polynomials defined on an interval $[a, b]$ form an orthogonal basis of $\mathcal{C}^0[a, b]$ when the maximum distance between two knots tends to zero and/or the degree of the B-splines tends to infinity. However, it allows for shape constraints and does not suffer from most other drawbacks of truncated orthogonal series.

9. Kernel-based estimation was introduced by Rosenblatt (1956) and refined by Nadaraya (1964) and Watson (1964).

10. Due to the distortion caused by each sample on its neighborhood in kernel-based estimation, there is a significant error in the spin at the sample points.

11. This holds, no matter from which distributions the $y_{j, m}$s are obtained, as long as Lindeberg's condition (which guarantees that no one $y_{j, m}$ outweighs the other samples) is satisfied.

12. The results in this paper can be transferred to other error norms. By the well-known norm-equivalence on $\mathbb{R}^n$, for any (weighted) error $\hat{e}(\theta \mid z) = \hat{\|}(w_1(s(x_1; \theta) - y_1), \ldots, w_n(s(x_n; \theta) - y_n))\hat{\|}$, which uses the norm $\hat{\|} \cdot \hat{\|}$ instead of the maximum-norm $\| \cdot \|_{\infty}$, there exist positive constants $\underline{\rho}, \bar{\rho}$ such that $\underline{\rho} e(\theta \mid z) \leqslant \hat{e}(\theta \mid z) \leqslant \bar{\rho} e(\theta \mid z)$ for any admissible θ and z.

13. Because by assumption A1 the structural objective function is Lipschitz, an allowable error on the objective function implies a maximum acceptable error for the component function.

14. Remark 12 in §4.1 provides several ways for obtaining such beliefs.

15. By the maximum theorem (Berge 1959) the set $\bar{X}(z)$ is compact as long as $\bar{\varepsilon}$ is finite, which by the Weierstrass theorem (Bertsekas 1995) guarantees that (8) has a solution.

16. Unsurprisingly, the ostrich bias becomes stronger if we allow the decision maker to first observe the data and then to choose $\bar{\varepsilon}$. For example, if, in the setting of Remark 5, it is impossible that θ is greater than some given upper bound

$\theta_0 > 0$, and the decision maker observes the single data point $(x, 2\theta_0 - x)$, then it is best for him to choose $\bar{\varepsilon} = 0$, effectively ignoring the fact that with probability one the actual θ is smaller than θ_0.

17. On the relevant interval $[a, b]$, the splines generated by $N_{i,k+1}$, $i \in \{-k, \ldots, \kappa\}$ are independent of the choice of the additional knots $\nu_{-k}, \ldots, \nu_{-1}$ and $\nu_{\kappa+2}, \ldots, \nu_{\kappa+k+1}$, as long as they are such that $\nu_{-k} \leqslant \nu_{-k+1} \leqslant \cdots \leqslant \nu_{-1} \leqslant \nu_0 = a$ and $b = \nu_{\kappa+1} \leqslant \nu_{\kappa+2} \leqslant \cdots \leqslant \nu_{\kappa+k} \leqslant \nu_{\kappa+k+1}$.

18. The proof of this result follows from Theorem 6 of De Boor (1978/2001, p. 149).

19. To ensure that all terms in Equation (17) are well defined, we adopt the convention that $0 \cdot \infty = \infty \cdot 0 = 0$. Such expressions may occur if we have coinciding knots, as, for example, at the boundary of $[a, b]$.

20. As De Boor (1978/2001) points out, increasing the number of knots is a desirable substitute for increasing the degree k of an approximating polynomial, especially because the computational complexity increases sharply in k (creating substantial difficulties for $k \geqslant 20$).

21. Given the sample z, the constant $\underline{\varepsilon}_\gamma(z)$ in Equation (25) always exists as a finite nonnegative number. Indeed, it is sufficient to approximate the data by a constant spline, e.g., equal to $\bar{y} = (y_1 + \ldots + y_n)/n$, and set $\varepsilon = \max_{j \in \{1, \ldots, n\}} |\bar{y} - y_j|$.

22. If $\theta \notin \Xi_{\varepsilon, \gamma}(z)$, then $\lambda(\theta \mid \varepsilon, z) = 0$. If $\Xi_{\varepsilon, \gamma}(z) = \varnothing$, then $\lambda(\cdot \mid \varepsilon, z) = 0$.

23. A charge-card balance becomes payable in full at the end of a payment period, contrary to a lending card, which allows for partial repayment at the due date and a debt rollover to the next payment period.

24. Based on the sample variances, for Segment 1 we choose $W = \text{diag}(\infty, 0.4714, 1, 0.5)$, and for Segment 2 it is $W = \text{diag}(\infty, 0.7071, 1, 0.7071)$. Note that the variance for the first (trivial) data point $(0, 1)$ vanishes, because all accounts accept zero repayment; this point is imposed as an interpolation constraint on the response curve.

25. This statement requires that despite the significant intersample belief uncertainty the decision maker is not subject to "ambiguity aversion" (Ellsberg 1961). The plausibility of this assumption is strengthened by the fact that the simple, one-dimensional settlement-rate decision is usually taken in a "noncomparative context" (Fox and Tversky 1995).

26. If n is even, one can extend the sample by an arbitrary data point consistent with $\widehat{M1}$ and $\widehat{M2}$.

Acknowledgments

References

Anastassiou, G. A., X. M. Yu. 1992. Monotone and probabilistic wavelet approximation. *Stochastic Anal. Appl.* **10**(3) 251–264.

Beliakov, G. 2000. Shape preserving approximation using least squares splines. *Approximation Theory Its Appl.* **16**(4) 80–98.

Beliakov, G. 2002. Monotone approximation of aggregation operators using least-squares splines. *Internat. J. Uncertainty, Fuzziness, Knowledge-Based Systems* **10**(6) 659–676.

Ben-Tal, A., A. Nemirovski. 1998. Robust convex optimization. *Math. Oper. Res.* **23**(4) 769–805.

Ben-Tal, A., A. Nemirovski. 2002. Robust optimization—Methodology and applications. *Math. Programming* **92**(3) 453–480.

Berge, C. 1959. *Espaces Topologiques et Fonctions Multivoques*. Dunod, Paris. [English translation: 1963, *Topological Spaces*, Oliver and Boyd, Edinburgh, UK. Reprinted: 1997, by Dover Publications, Mineola, NY.]

Bertsekas, D. P. 1995. *Nonlinear Programming*. Athena Scientific, Belmont, MA.

Bertsimas, D., M. Sim. 2004. The price of robustness. *Oper. Res.* **52**(1) 35–53.

Besbes, O., R. Phillips, A. Zeevi. 2010. Testing the validity of a demand model: An operations perspective. *Manufacturing Service Oper. Management* **12**(1) 162–183.

Curry, H. B., I. J. Schoenberg. 1947. On spline distributions and their limits: The Pólya distribution functions. *Bull. Amer. Math. Soc.* **53**(11) 1114.

De Boor, C. 1978/2001. *A Practical Guide to Splines*. Springer, New York.

De Vore, R. A. 1977a. Monotone approximation by splines. *SIAM J. Math. Anal.* **8**(5) 891–905.

De Vore, R. A. 1977b. Monotone approximation by polynomials. *SIAM J. Math. Anal.* **8**(5) 906–921.

Dierckx, P. 1980. An algorithm for cubic spline fitting with convexity constraints. *Computing* **24**(4) 349–371.

Dierckx, P. 1993. *Curve and Surface Fitting with Splines*. Oxford University Press, Oxford, UK.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quart. J. Econom.* **75**(4) 643–669.

Fox, C. R., A. Tversky. 1995. Ambiguity aversion and comparative ignorance. *Quart. J. Econom.* **110**(3) 585–603.

Fritsch, F. N., R. E. Carlson. 1980. Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.* **17**(2) 238–246.

Haimes, Y. Y., D. A. Wismer. 1972. A computational approach to the combined problem of optimization and parameter identification. *Automatica* **8**(3) 337–347.

Haimes, Y. Y., L. S. Lasdon, D. A. Wismer. 1971. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Trans. Systems, Man, Cybernetics* **1**(3) 296–297.

Hanson, D., G. Pledger. 1976. Consistency in concave regression. *Ann. Statist.* **4**(6) 1038–1050.

Härdle, W., O. Linton. 1994. Applied nonparametric methods. R. F. Engle, D. L. McFadden, eds. *Handbook of Econometrics*, Vol. 4. Elsevier Science, Amsterdam, 2295–2339.

Judd, K. L. 1999. *Numerical Methods in Economics*. MIT Press, Cambridge, MA.

Kocić, L. M., G. V. Milovanović. 1997. Shape preserving approximations by polynomials and splines. *Comput. Math. Appl.* **33**(11) 59–97.

Kvasov, B. I. 2000. *Methods of Shape-Preserving Spline Approximation*. World Scientific, London.

Lim, A. E. B., J. G. Shanthikumar, Z. J. M. Shen. 2006. Model uncertainty, robust optimization, and learning. *Tutorials in Operations Research*. INFORMS, Hanover, MD, 66–94.

McAllister, D. F., E. Passow, J. A. Roulier. 1977. Algorithms for computing shape preserving spline interpolations to data. *Math. Comput.* **31**(139) 717–725.

Nadaraya, E. 1964. On estimating regression. *Theory Probab. Its Appl.* **9** 141–142.

Pagan, A., A. Ullah. 1999. *Nonparametric Econometrics*. Cambridge University Press, Cambridge, UK.

Passow, E., J. A. Roulier. 1977. Monotone and convex spline interpolation. *SIAM J. Numer. Anal.* **14**(5) 904–909.

Powell, J. L. 1994. Estimation of semiparametric models. R. F. Engle, D. L. McFadden, eds. *Handbook of Econometrics*, Vol. 4. Elsevier Science, Amsterdam, 2443–2521.

Pruess, S. 1993. Shape preserving C^2 cubic spline interpolation. *IMA J. Numer. Anal.* **13**(4) 493–507.

Ramsay, J. O. 1988. Monotone regression splines in action. *Statist. Sci.* **3**(4) 425–461.

Robertson, T., F. T. Wright, R. L. Dykstra. 1988. *Order Restricted Statistical Inference*. Wiley, New York.

Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 642–669.

Shannon, C. E. 1949. Communication in the presence of noise. *Proc. Institute Radio Engineers* **37**(1) 10–21. [Reprinted in: 1998, *Proc. IEEE* **86**(2) 447–457.]

Shanthikumar, J. G., L. H. Liyanage. 2005. A practical inventory control policy using operational statistics. *Oper. Res. Lett.* **33**(4) 341–348.

Shisha, O. 1965. Monotone approximation. *Pacific J. Math.* **15**(2) 667–671.

Wang, S.-P., K. L. Judd. 2000. Solving a savings allocation problem by numerical dynamic programming with shape-preserving interpolation. *Comput. Oper. Res.* **27**(5) 399–408.

Watson, G. 1964. Smooth regression analysis. *Sankhya, Ser. A* **26** 359–372.

Wolibner, W. 1951. Sur un polynôme d'interpolation. *Colloquium Mathematicum* **2** 136–137.