International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Estimation of Econometric Models by Risk Minimization: A Stochastic Quasigradient Approach

Ermoliev, Y.M., Keyzer, M.A. and Norkin, V.I.

IIASA Interim Report
March 2002

International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: publications@iiasa.ac.at
Web: www.iiasa.ac.at

Interim Report **IR-02-021**

Estimation of econometric models by risk minimization: a stochastic quasigradient approach

Yuri Ermoliev, (ermoliev@iiasa.ac.at),
Michiel Keyzer, (M.A.Keyzer@SOW.ECON.VU.NL), and
Vladimir I. Norkin, (norkin@d130.icyb.kiev.ua)

Approved by

Günther Fischer
Leader, Land Use Change Project

March, 2002

Stichting Onderzoek Wereldvoedselvoorziening van de Vrije Universiteit (Centre for World Food Studies of the Free University, Amsterdam)

International Institute for Applied Systems Analysis, Laxenburg, Austria

Contents

Abstract

The paper presents a risk minimization approach to estimate a flexible form that meets a priori restrictions on slope and curvature by means of constraints on both the estimated parameters and the function values. The resulting constrained risk minimization combines parametric and nonparametric estimation and contains integrals and implicit constraints. Within econometrics, simulation has become a common tool to solve problems of this kind. However, it appears that in our case, the simulation approach only applies when the model is linear in parameters, has simple constraints on parameters and a quadratic risk function. To deal with other cases, we use a stochastic optimization technique known as the stochastic quasi-gradient method for stationary and nonstationary problems with Cesàro averaging. This method is also applicable to an expanding series of random observations, and produces asymptotically (weakly) convergent estimates.

About the Authors

Dr. Yuri Ermoliev has been Head of the Department of Mathematical Methods of Operations Research at the Institute of Cybernetics of the Ukrainian Academy of Sciences, Kiev, since 1969. He was first employed with IIASA in the System and Decision Sciences Program from 1979 to 1984, undertaking research in non-differentiable and stochastic optimization problems. In 1991 he was a visiting professor with the University of California at Davis. Dr. Ermoliev's scientific interests are modeling of decision-making processes in the presence of risks and uncertainties, stochastic and dynamic systems optimization, optimization on networks, and nonlinear dynamics. His major publications include Stochastic Programming Methods (1976), Stochastic Models in Economics (1979), and Numerical Techniques for Stochastic Optimization (1988). Other publications concern the study of path-dependent adaptation processes, pollution control problems, energy and agriculture modeling, reliability theory, and optimization of discontinuous systems, in particular, discrete event systems optimization.

Dr. Michiel A. Keyzer is professor of mathematical economics and Director of the Centre for World Food Studies, Free University (SOW-VU), Amsterdam, The Netherlands. Professor Keyzer's main activities are in research and research co-ordination in the areas of mathematical economics and economic model building. He has led studies on development planning in Bangladesh, Indonesia, Nigeria, West Africa and Lebanon, on reform of the Common Agricultural Policy (1995), and on farm restructuring and land tenure in reforming socialist economies for IFAD and the World Bank. His major publications are in the field of general equilibrium, modelling (with Ginsburgh, 1997). Recent publications are with Ermoliev and Norkin (2000), and Gerlagh (2001).

Dr. Vladimir I. Norkin received his MS degree from Moscow Institute of Physics and Technology (1974). Since 1974 he has been affiliated with the Glushkov Institute of Cybernetics, Kiev, Ukraine, receiving his Ph.D. (1983) and Doctor (1998) degrees in Operations Research from this Institute. He has been collaborating with IIASA since 1993, firstly with Optimization under Uncertainty project and later with the project on Risk, Modeling and Society. He published a monograph Methods of Non-convex Optimization (Nauka, Moscow, 1987) and a number of papers on non-differentiable, stochastic, integer and global optimization, computation of economic equilibrium, discrete event systems optimization, catastrophic risk management.

Estimation of econometric models by risk minimization: A stochastic quasigradient approach

Yu. Ermoliev, M.A. Keyzer and V.I. Norkin

1. Introduction

Econometric regression techniques have developed in two directions. One focuses on flexible adaptation of the regression curve to the data. In this area, nonparametric techniques such as kernel density regression have gained popularity because they are easy to use and permit to estimate full conditional distributions without any normality assumptions (see e.g., Haerdle, 1993; Yatchew, 1998). The other emphasizes imposition of a priori constraints, so as to improve the identification of parameters and to narrow the gap between economic theory and empirical applications. Estimation techniques based on simulation, such as simulated maximum likelihood, permit to impose a priori restrictions of integral or implicit form, both on function values and on derivatives, while random sampling from continuous distributions facilitates the identification of a large number of parameters (Gouriéroux and Monfort, 1996). Simulation-based techniques were designed to deal with situations when the criterion function of the optimization is not tractable, either because it contains integrals without a closed form solution, or because it is only implicitly available, say, as the outcome of a process model. They circumvent this problem by proceeding in two steps. The first step conducts a random sampling of error terms, latent variables or variables that are only measured as continuous distribution to replace the integrals of the risk function, or other expressions that are not available in closed form, by a discretized form whose summations run over empirical distributions with a finite number of observations. The second step applies common estimation approaches such as maximum likelihood, pseudo maximum likelihood, or Generalized Method of Moments (GMM) to the discretized problem with sampled data.

The simulation approach is workable if the results of these summations can be stored in a tractable number of terms that are independent of the vector of parameters to be estimated. This is for example the case if the discretized form reduces to linear least squares, where the sums enter calculations only once since the optimization has a closed form solution, and the matrices to be inverted have the dimension of the parameter vector, rather than of the number of observations. However, in the cases mentioned above, that is whenever the model is nonlinear in parameters, or has integral constraints that depend on parameters, or when the loss function itself is not quadratic, it becomes extremely cumbersome, if not impossible, to evaluate these large expressions with sufficient accuracy at every iteration. By contrast, the SQG-algorithm directly addresses the approximation by updating parameter values so as to find an optimum during simulations, rather than conducting sampling prior to parameter estimation.

Flexibility is related to this imposition of constraints, because the less observations are available, the more inflexible the specification will have to be so as to maintain identification of the parameters (see e.g., Davidson and MacKinnon, 1993; Wets, 1998). Barnett et al. (1991) have, among others, proposed to overcome this limitation through Bayesian methods that could provide additional identifying restrictions. Various other semiparametric techniques were developed. For example, the parametric step may be used to compute an error term whose distribution is estimated non-parametrically, and from which a larger data set is created through random sampling for the next step of parametric estimation (Robinson, 1988). However, using the empirical distribution may be restrictive in this respect, because it cannot venture beyond the limited number of observations.

Overview

The paper presents an alternative to the simulation approach to constrained risk minimization by replacing the two-step procedure by an SQG-algorithm (Ermoliev, 1976, 1988) with Cesàro averaging (Nemirowski and Yudin, 1983), applicable also in case of growing number of observations. Besides the empirical distribution we also consider a kernel estimate of the density to sample the data from. This eases the identification of parameters since it enables us to sample from an infinite number of points on continuous densities rather than from the finite number of data points of the empirical distribution. Under relatively mild identifiability restrictions, we prove convergence of this algorithm, with probability one, to the unique (global) minimum of the risk function and we describe the consistency properties of the procedure.

The paper proceeds as follows. Section 2 introduces the computational method using the relatively simple problem of computing the mathematical expectation by sampling from the given data set and by considering the typical problem of supply function estimation as an example. Section 3 applies the computational approach to the problem of prediction and parameter estimation for a general form that is linear in parameters, with data sampled from the empirical distribution. In section 4 data sampling is from a smooth density estimated by kernel density regression. Section 5 establishes the consistency of the estimators for both cases. Section 6 concludes.

2. Introducing the approach

To introduce our approach and the main concepts, we review in this section seven ways of computing the mathematical expectation for a given sample of size *N*. Next we describe the issues arising for the typical candidate problem of estimating a supply function.

Computation of the mathematical expectation

We start from the simple problem of computing the mathematical expectation, since this enables us to focus on the method of calculation rather than on the problem itself. The sample consists of *N* observations $y^1,...,y^N$ of the vector of observable dependent random variable $y \in R^m$ distributed according density $g(y)$. There are several ways for computing the sample mean, as follows:

(i) *Direct calculation* obtains the mean as

$$\beta^N = \frac{1}{N}\sum_{k=1}^{N} y^k \tag{2.1}$$

(ii) *Unconstrained least squares* calculation solves the convex problem

$$\min_{\beta} F^N(\beta), \tag{2.2}$$

for $F^N(\beta) = \frac{1}{2}\sum_{k=1}^{N}\left\|y^k - \beta\right\|^2$. For this minimization, (2.1) represents first-order optimality conditions.

(iii) *Stepwise calculation* proceeds in $N-1$ steps, starting from $\beta_1 = y^1$, according to the recursion:

$$\beta^{k+1} = \beta^k - \rho_k(\beta^k - y^k), \qquad k = 1,...,N-1 \tag{2.3}$$

for $\rho_k = 1/k$, i.e. exact solution of least square problem (2.2) can be obtained by $N-1$ steps of iterative procedure (2.3).

(iv) *Constrained least squares* can account for the a priori information that the true sample mean is known to belong to the compact convex set *B*, say, $B = \{\beta : \underline{\beta} \le \beta \le \overline{\beta}\}$:

$$min_{\beta \in B} F^N(\beta), \tag{2.4}$$

The problem can be solved iteratively, using a standard (deterministic) constrained optimization algorithm.

(v) *Calculation by iterative sampling* is a stochastic method. It performs a sequence of random drawings y^t (with replacement) from the sample $y^1,...,y^N$, and updates the estimated mean on the basis of the newly sampled value. It uses the a priori information $\beta \in B$ and computes the mean as the limit point of the sequence

$$\beta^{t+1} = \Pi_B(\beta^t - \rho_t(\beta^t - y^t)), \qquad t = 1,2,... \tag{2.5}$$

where Π_B denotes the projection on the set *B* (an interval if $m = 1$ or a hypercube in the present case) and ρ_t is a suitable step-size, for example, $\rho_t = 1/t$. That this in fact is a stochastic optimization procedure can be seen as follows. Let us notice that the random vector $\xi^t = \beta^t - y^t$ is an estimate of the gradient (called stochastic gradient) $F^N_\beta(\beta^t)$ for

$$F^N(\beta) = \frac{1}{2} E\|y - \beta\|^2 = \frac{1}{2}\int \|y - \beta\|^2 dG^N(y) \tag{2.6}$$

where $G^N(y)$ is a distribution from which the empirical distributions are obtained and we explicitly write β^N_* rather than β^* to emphasize that this is not the population parameter. We note that Procedure (2.5) drives β^t towards the minimum of the function (2.6), without requiring the distribution $G^N(y)$ to be known explicitly. It only assumes that $F^N(\beta)$ exists, i.e. that the distribution has a finite second moment. The sequence generated by (2.5) converges to the minimum of $F^N(\beta)$ with probability one (Ermoliev, 1976, 1988). It is easy to see that β^N_* indeed solves $min_{\beta \in B} F^N(\beta)$, since $\beta = Ey$ satisfies the optimality conditions:

$$F^N_\beta(\beta) = (\beta - Ey) = 0, \tag{2.7}$$

where Ey is the expectation according to the density $g(y)$. Therefore, procedure (2.5) attempts to reach the stationary point $\beta = Ey$ of $F^N(\beta)$ by moving from any current point estimate β^t to β^{t+1} in the direction opposite to the stochastic gradient ξ^t, for $\xi^t = (\beta^t - y^t)$, and such that

$$E[\xi^t \mid \beta^t] = F_\beta^N(\beta^t). \tag{2.8}$$

(vi) *Solving the stochastic optimization problem for a previously estimated probability distribution*. Notice that the given sample of *N* observations can be thought of as having been generated through the generalized density

$$g^N(y) = \frac{1}{N}\sum_{k=1}^N \delta(y - y^k) \tag{2.9}$$

where $\delta(y - y^k)$ is the delta-measure concentrated at point y^k. The mean can also be computed by solving the explicitly defined stochastic optimization problem $min_{\beta \in B} F^N(\beta)$, for

$$F^N(\beta) = \frac{1}{2}\int \|y - \beta\|^2 g^N(y)dy = \frac{1}{2}\int \|y\|^2 g^N(y)dy - \frac{1}{2}\left\|y^N\right\|^2 + \frac{1}{2}\left\|y^N - \beta\right\|^2, \tag{2.10}$$

where $y^N = \int y g^N(y)dy = \frac{1}{N}\sum_{k=1}^N y^k$.

Procedure (2.5) will solve this problem. The alternative is to integrate (2.10) analytically. This leads back to the empirical data set as in (2.4). Yet the empirical density (2.9) is highly nonsmooth. It would seem a natural idea to avoid nonsmoothness by incorporating more information about the actual distribution $G(y)$, such as smoothness or a shape restriction. This amounts to replacing the empirical density (2.6) by a smooth density $g_\theta^N(y)$ that approaches true density $g(y)$ for $\theta \to 0$ and $N \to +\infty$. Let $K_\theta(y, y^k)$ be a symmetric probability density of the form $K_\theta(y, y^k) = \frac{1}{\theta^m} K\left(\frac{y - y^k}{\theta}\right)$, $\int K(y)dy = 1$, e.g. the normal density. Replacing the empirical measure by the Kernel estimate leads to:

$$g_\theta^N(y) = \frac{1}{N\theta^m}\sum_{k=1}^N K\left(\frac{y - y^k}{\theta}\right), \theta > 0 \tag{2.11}$$

It must be noted, however, that this approach could lead to a biased estimate, since the first moment $y_\theta^N(y)$ might not be the sample mean. This approach can be characterized as semiparametric because the density is obtained by a non-parametric (kernel density) method.

(vii) *Minimizing squared deviation between non-parametric and parametric estimate.*

The same estimate can be obtained if we rewrite the objective of (2.10) as

$$F_\theta^N(\beta) = \frac{1}{2}\int \|y\|^2 g_\theta^N(y)dy - \frac{1}{2}\left\|y_\theta^N\right\|^2 + \frac{1}{2}\left\|y_\theta^N - \beta\right\|^2, \tag{2.12}$$

where the first moment $y_\theta^N = \int y\, g_\theta^N(y)dy$ serves as non-parametric estimate. For given θ, only the third term matters, which minimizes the squared deviation between the parametric and the non-parametric estimate. Once this non-parametric estimate has been computed, the optimization problem becomes trivial. The problem will not be as simple once explanatory variables are taken into account, and it is at that stage that it becomes meaningful to compare the relative merits of these seven approaches. Here we only notice that formulations (i)-(iv) invoke deterministic techniques of computation, which are perfectly straightforward for small to medium size samples but less practical if N is very large viz. infinite. In those cases, stochastic methods (v)-(vii) become relevant alternatives.

The seven approaches mentioned compute the same arithmetic sample mean over N observations. The Law of Large Numbers says that this mean is a consistent estimator of β^*. This estimation shifts the concern from computing the mean of the finite sample to estimating the mean of a population of a much larger size. If in (2.6) the distribution function $G^N(y)$ were known, and equal to the distribution function of the population $G(y)$, with well-defined moments, estimation of β^* would only be a problem of computation. However, in practice $G(y)$ is unknown and only the fixed sample $\{y^k\}$ of size N is available. The common strategy is to show that the sequence $\{\beta_*^N\}$, $N = 1,2,...$ converges to β^* for N going to infinity. We return to this issue in section 5.

Constrained risk minimization

Next, the scope of application of the proposed approach may be illustrated by means of a typical regression problem. Let y denote the endogenous variable "net supply" and x the exogenous variable "price". Consider the supply function $y(x)$ obtained in the microeconomic theory of the firm as the derivative of the profit function: $y(x) = \partial\Pi(x)/\partial x$ (Varian, 1992), where the profit function $\Pi(x)$ is taken to be continuously differentiable, convex, and homogeneous of degree one in prices p, and increasing in the prices of the goods for which the firm is a net seller and decreasing in those where it is a net buyer. The task in estimation is to identify the m net supply functions on the basis of a finite sample of size N, with elements indexed k and observations (y^k, x^k) and on $\Pi^k \equiv x^{k\prime} y^k$:

$$y_i^k = \frac{\partial\Pi(x^k)}{\partial x_i} + \varepsilon_i^k \qquad i = 1,...,m \tag{2.13}$$

with error term ε_i^k where the profit function has a parametric specification

$$\Pi(x) = H(x, \beta^*) \tag{2.14}$$

and $\beta^* \in B \subset R^n$ is, for known parameter set B, the unknown true parameter value, and we have to identify the vector function $\partial H(x, \beta^*)/\partial x$.

Theoretical restrictions

In general, the parametric form $H(x,\beta^*)$ itself is unknown, because theory only imposes general restrictions such as convexity, monotonicity and homogeneity of degree zero. The common procedure is to postulate a form that offers sufficiently flexible adjustment to the observed data, while meeting theoretical restrictions. It is relatively easy to find functional expansions (e.g. basis in a functional space, Gallant (1981)) such as $H(x,\beta) = h(x)'\beta = \sum_{j=1}^{n} h_j(x)\beta_j$, where $h_j(x)$ is j-th column of matrix $h(x)$, that combine flexibility with linearity in parameters. Convexity, monotonicity and homogeneity are readily imposed by requiring the "building blocks" $h_j(x)$ to have the same properties and by restricting parameters β_j to nonnegative values.

However, nonnegativity of β_j again reduces the flexibility of the model, especially with respect to cross effects. Another way to ensure monotonicity and convexity properties is to place restrictions on the derivatives within a given domain X. For example, monotonicity of $H(x,\beta)$ with respect to the variable x_i means that at any $x \in X$, $\partial H(x,\beta)/\partial x_i = \sum_{j=1}^{n} \partial h_j(x)/\partial x_i \beta_j \geq 0$. This can be expressed by the scalar inequality:

$$\varphi_i(\beta) = \min_{x\in X} \sum_{j=1}^{n} \partial h_j(x)/\partial x_i \, \beta_j \geq 0 \tag{2.14a}$$

and, similarly, monotonicity in all nonnegative directions is given by

$$\varphi(\beta) = \min_{x\in X} \min_{\ell\in L_+^r} \sum_{j=1}^{n} \left(\sum_i \ell_i \partial h_j(x)/\partial x_i\right)\beta_j \geq 0, \tag{2.14b}$$

for $L_+^r = \{\ell \in R_+^r \mid \|\ell\| \leq 1\}$, and the concavity property can be guaranteed by the inequality

$$\psi(\beta) = \max_{x\in X} \max_{\ell_1,\ell_2\in L^r} \sum_{j=1}^{n} \left(\sum_{i,h} \ell_i \ell_h \frac{\partial^2 h_j(x)}{\partial x_i \partial x_h} \right)\beta_j \leq 0, \tag{2.14c}$$

$L^r = \{\ell \in R^r \mid \|\ell\| \leq 1\}$. We will assume that β belongs to some convex compact set $B \in R^n$, given by a system of linear and nonlinear convex inequalities. In addition, one may impose lower and upper bounds on the model, $\underline{y}(x) \leq h(x)'\beta \leq \bar{y}(x)$, $x \in X$, that also amount to constraints on β, namely:

$$\psi_1(\beta) = \max_{x\in X} (\underline{y}(x) - h(x)'\beta) \leq 0, \tag{2.14d}$$

$$\psi_2(\beta) = \max_{x\in X} (h(x)'\beta - \bar{y}(x)) \leq 0. \tag{2.14e}$$

We remark that the inequalities generated by these constraints are convex in β, and therefore fit within a convex programming framework, but the evaluation of the functions $\varphi, \varphi_i, \psi, \psi_1, \psi_2$ will generally be a difficult task.

Finally, integral restrictions may have to be imposed on the model. For example, suppose that the variable x expresses geographical coordinates and $H(x,\beta)$ the supply distribution, as before. If empirical data are available on aggregate supply

$$I(\beta) = \int_X H(x,\beta)dx = \sum_{j=1}^{n} \beta_j \int_X h_j(x)dx \tag{2.14f}$$

in region X, the parameter β may have to satisfy a linear equality. Alternatively, if aggregate supply is known to lie within given bounds, the parameter β is constrained by linear inequalities $a \le I(\beta) \le b$.

Incorporating non-parametric regression in constrained risk minimization

Our approach amounts to choosing parameter values β in a compact convex set *B* generated by constraints as in (2.14), so as to minimize the integral square deviation between a parametric and a non-parametric form, where the non-parametric form offers a representation in a continuum of the discrete and finite set of empirical observations. Formally:

$$\min_{\beta \in B} F_\theta^N(\beta), \tag{2.15}$$

for the risk functions

$$F_\theta^N(\beta) = \int_{R^r} L(y - h(x)'\beta) g_\theta^N(x,y)dxdy, \tag{2.16a}$$

or

$$F_\theta^N(\beta) = \int_{R^r} L(y_\theta^N(x) - h(x)'\beta) g_\theta^N(x)dx, \tag{2.16b}$$

where *x* is the *r*-dimensional vector of exogenous variables, $L(\cdot)$ is some nonnegative, strictly convex loss function, *h(x)* is a vector function compatible with β, $y_\theta^N(x)$ is a kernel density regression function estimated on the basis of N observations and with given parameter value θ (usually a scalar referred to as window size), and, finally, $g_\theta^N(x)$ is the associated, nonparametrically estimated density of *x*. For $\theta = 0$ this density reduces to the empirical one. Thus, if observations (x,y) have density function $g(x,y)$, the underlying true risk function has the form:

$$F(\beta) = \int_{R^r} L(y - h(x)'\beta) g(x,y)dxdy, \tag{2.17a}$$

or

$$F(\beta) = \int_{R^r} L(y(x) - h(x)'\beta) g(x)dx, \tag{2.17b}$$

where $y(x) = \int_{R^m} yg(y \mid x)dy$ is a regression curve and $g(x) = \int_{R^r} g(x,y)dy$ is a marginal density and $g(y \mid x) = g(y,x)/g(x)$ is a conditional density. For $L(\cdot) = \frac{1}{2}\|\cdot\|^2$, we obtain a least squares estimate. Other examples of smooth and nonsmooth convex risk functions can be found in Huber (1981; section 7.3). For the quadratic loss function, the function $F(\beta)$ in (2.17a) or (2.17a) can in principle be written in closed form, and the terms with integrals can be treated as data that can be obtained by means of numerical integration.

This is not possible for general loss functions or when β has to meet convex constraints. Then, we can apply SQG to avoid explicit calculation of the integrals at every iteration. If a convex constraint set B is given by means of complex inequalities (2.14a)-(2.14f), it can be represented by

an infinite system of linear inequalities in β. In this case problem (2.15) becomes a semi-infinite programming problem (see Hettich and Kortanek, 1993).

3. Parameter estimation on the basis of the empirical distribution

This section applies the computational approach to the problem of parameter estimation for a model that is linear in parameters. For this model, we prove strict convexity of the risk function as well as identification and consistency of the estimators. We also consider various constraints on parameters, and estimate these by sequentially outer-approximate the constraint set by a family of randomly chosen linear inequalities on β. Model $Ey = \beta$ of the mean is now replaced by the linear regression model:

$$y = h(x)'\beta^* + \varepsilon \quad \text{for fixed } \beta^* \in B \subset R^n, \tag{3.1}$$

where $h : R^r \to R^n \times R^m$, $h(x)$ is a given continuous *(n×m)*-matrix-function, say, the terms of a multivariate polynom and $h(x)'$ denotes the matrix transpose; $y \in R^m$ is the vector of observable dependent random variables (net supply in (2.13)), for given values of independent random variables $x \in X \subseteq R^r$, β^* is an unknown vector of parameters and ε is an unknown error term (random variable).

The parametric form (3.1) of simultaneous nonlinear equations is quite flexible. If, say, $y(x)$ is a scalar function from a certain functional space Y, then $h(x)'$ can contain the first n elements of a basis in this space and elements of β^* can be thought of as Fourier coefficients for this basis. Hence, a general relationship $y(x)$ can be modeled in (3.1) by allowing for a sufficiently large number n of basis elements. Furthermore, if the functions from Y possess certain properties such as homogeneity, then it is reasonable to choose in Y a basis that possesses the same properties. If y is a vector, the form (3.1) would have to be a vector function. This function could relate specific elements of β^* to different components of y. It might also allow for common coefficients across equations. For example, for vector function (2.13), the elements of β^* are Fourier coefficients of the scalar model $\Pi(x) = H(x)'\beta^*$ as in (2.14), and in this case one has to take $y(x) = \dfrac{\partial H(x)}{\partial x}' \beta^*$.

Regarding estimation, let us assume that there is a finite data set $\{x^1, y^1, ..., x^N, y^N\}$ of inputs-outputs of model (3.1) drawn from some unknown theoretical distribution $G(x,y)$ with density $g(x,y)$. We assume that the data were obtained as through random sampling, that is that the individual observations are independent (i.i.d.). Hence, we disregard any serial correlation. We denote by $G^N(x,y)$ a probability distribution (with density $g^N(x,y)$) reconstructed from observations $\{x^1, y^1, ..., x^N, y^N\}$. This could be an empirical distribution, a nonparametric kernel density estimate or a member of some parametric family. Also denote the marginal distribution by $g^N(x) = \int g^N(x,y)dy$. As a prerequisite for consistency, we formulate an assumption on convergence of $G^N(x,y)$:

Assumption D (convergence of measures). The sequence of measures $G^N(x,y)$ is such that for any continuous function $f(x,y)$ convergence

$$lim_{N\to\infty} E^N f(x,y) = lim_{N\to\infty} \int f(x,y) dG^N(x,y) = Ef(x,y) = \int f(x,y) dG(x,y)$$

holds with probability one. ◊

This assumption certainly holds if $f(x,y)$ is bounded and G^N weakly converge to G as $N \to +\infty$. Given this specification of the model and its data, the underlying risk minimizing problem (2.6) for estimation of parameter β^* can be written:

$$F^N(\beta) = EL(y - h(x)'\beta) = \int L(y - h'(x)\beta) dG^N(x,y) \to min_{\beta\in B} . \qquad (3.2)$$

Thus, this problem is fully specified by the risk function L, the probability measure G^N and the constraint set B. A typical and convenient choice of risk function is $L(\cdot) = \frac{1}{2}\|\cdot\|^2$. In this case, if the optimum β^N of (3.2) lies in the interior of set B, a closed form solution is available:

$$\beta^N = \left(E^N h(x)h(x)'\right)^{-1}\left(E^N h(x)y\right) = \left(\int_X h(x)h(x)' dG^N(x,y)\right)^{-1}\left(\int_X h(x)y dG^N(x,y)\right). \qquad (3.3)$$

Since the inverse on the right-hand side of this expression should exist, to ensure that β^N can be identified, we need further assumptions.

Assumption M1 (Model assumptions).

(a) The model error has zero conditional mean: $E\{\varepsilon \mid x\} = 0$.

(b) True model parameter β^* belongs to constraint set B.

(c) Matrix $h(x)h(x)'$ is nonsingular with positive probability:
$\int_A h(x)h(x)' g^N(x) dx$ is nonsingular for some $A \subseteq R^r$.

(d) Vector-function $h(x)$ is continuous and bounded on X. ◊

Note, since $h(x)h(x)'$ is semipositive definite by definition, condition (c) amounts to requiring $h(x)h(x)'$ to be positive definite on some set of positive measure. It is fulfilled, for example, if $m \geq n$ – the number m of columns of $h(x)$ is greater or equal to the number of rows n – and $h(x)$ contains n linear independent rows on the set of positive measure in X. Condition (c) is a generalization of nonsingularity assumption on matrix $\frac{1}{N}\sum_i h(x_i)h(x_i)'$ for a classical linear regression when $h(x)$ is a column vector-function.

Assumption L (loss function).

The loss function $L(z)$, $z \in R^m$,

(a) is continuous and strictly convex;

(b) $L(0) = 0$ and $L(z) > 0$ for any $z \neq 0$. ◊

Lemma 3.1 (strict convexity of the integral risk function).
Under assumptions M1(c) and L(a) the integral risk function

$$F^N(\beta) = \int_{R^n} L(y - h(x)')dG^N(y,x)$$

is strictly convex.

Proof. By strict convexity of the loss function (Assumption L(a)), for any $z_0 \neq z_1$ and $0 < \lambda < 1$ the following strict inequality holds:

$$L((1-\lambda)z_0 + \lambda z_1) < (1-\lambda)L(z_0) + \lambda L(z_1). \tag{3.4}$$

Let us take $z_0(y,x) = y - h(x)'\beta_0$ and $z_1(y,x) = y - h(x)'\beta_1$ with $\beta_0 \neq \beta_1$. Then, with positive probability $z_0(y,x) \neq z_1(y,x)$, since equality $z_0(y,x) = z_1(y,x)$ implies $h(x)'(\beta_0 - \beta_1) = 0$ and $h(x)h(x)'(\beta_0 - \beta_1) = 0$, which is impossible for x such that $h(x)h(x)'$ is nonsingular. Substituting $z_0(x)$ and $z_1(x)$ into (3.4), we obtain inequality

$$L(y - h(x)'((1-\lambda)\beta_0 + \lambda\beta_1)) \leq (1-\lambda)L(y - h(x)'\beta_0) + \lambda L(y - h(x)'\beta_0), \tag{3.5}$$

where strong inequality happens with positive probability. Integration yields the required property

$$F^N((1-\lambda)\beta_0 + \lambda\beta_1) < (1-\lambda)F^N(\beta_0) + \lambda F^N(\beta_1). \square$$

This lemma only gives sufficient conditions for the integrated risk function to be strictly convex and thus for single-valuedness of solution of (3.2). Assumption L(a) is fulfilled for quadratic but not for least norm estimates.

Corollary 3.1 (identifiability and asymptotic consistency). *For a quadratic risk function $L(\cdot)$, under assumptions M1(a)-(c) problem (3.2) has a unique solution β_N^* that converges to a true vector parameter β^* as $N \to +\infty$.*

Proof. Consider a true risk function

$$F(\beta) = \frac{1}{2}E\|y - h(x)'\beta\|^2 = \frac{1}{2}E\|y - E[y \mid x] + E[y \mid x] - h(x)'\beta\|^2 =$$
$$= \frac{1}{2}E\|\varepsilon\|^2 + E\varepsilon' h(x)'(\beta^* - \beta) + \frac{1}{2}(\beta^* - \beta)'Eh(x)h(x)'(\beta^* - \beta).$$

Since by Assumption M1(a), $E[\varepsilon \mid x] = 0$, it follows that

$$F(\beta) = \frac{1}{2}E\|\varepsilon\|^2 + \frac{1}{2}(\beta^* - \beta)'Eh(x)h(x)'(\beta^* - \beta).$$

As $h(x)h(x)'$ is positive semidefinite, by Assumption M1(b), $Eh(x)h(x)'$ can be written as the sum of a positive semidefinite and a positive definite term, and is therefore positive definite and

hence nonsingular. Thus by assumption M1(b), β^* is the unique minimizer of $F(\beta)$ on B. By Assumption D, $lim_{N\to\infty} F^N(\beta) = F(\beta)$, for any β with probability one. By convexity, sequence $F^N(\beta)$ uniformly converges to $F(\beta)$ (in any compact set) with probability one. Thus, minimizers β_N^* converge to the unique minimum β^* of the limit function $F(\beta)$ with probability one.□

Calculation by an SQG-algorithm with Cesàro averaging

Solving risk minimization problem (3.2) for the general case can be a nontrivial task. First, the approximate distribution G^N, for instance, a nonparametric estimate, may not permit to take, in combination with a general risk function, the integral in (3.2) in closed form. Secondly, the risk function L could be nonsmooth as in $L(\cdot) = \|\cdot\|$ or

$$L(\beta) = \sum_i a_N(R_i)(y_i - h(x_i)'\beta), \tag{3.6}$$

where R_i is the rank of the number $r_i = y_i - h(x_i)'\beta$ in the row $(r_1,...,r_N)$, $a_N(\cdot)$ is some monotonic weighting function satisfying $\sum_i a_N(i) = 0$, the so-called R-estimates, see Huber (1981, section 7.3). Thirdly, the convex constraint set B may take the form of a large (possibly infinite) number of linear inequalities.

Nonsmooth optimization methods make it possible to solve (3.2) in the general case. For example, if the distribution G^N derived from the given sample of size *N* does not lead to a closed form for the integral, approximate by a stochastic quasi-gradient (SQG-) method is possible, as in (2.5). For this, one constructs a sequence of length M of estimators:

$$\beta^t = \Pi_B(\beta^t - \rho_t \xi^t), \qquad t = 1,2,...,M \le +\infty, \tag{3.7}$$

where $\beta^1 \in B$, the stochastic quasigradient ξ^t, i.e. the estimator of a (sub)gradient $\nabla F^N(\beta)$ at β^t, of the form:

$$\xi^t = \xi(\tilde{x}^t, \tilde{y}^t) = -h(\tilde{x}^t)\nabla L(\tilde{y}^t - h(\tilde{x}^t)'\beta^t), \tag{3.8}$$

pairs $(\tilde{x}^t, \tilde{y}^t)$ are independently (for different t) drawn from the given G^N. If G^N is the empirical measure, problem (3.2) reduces to:

$$min_{\beta\in B} \frac{1}{N}\sum_{k=1}^N L(y_k - h(x_k)'\beta), \tag{3.9}$$

and one can take ξ^t in (3.7) as a deterministic gradient vector of the form:

$$\xi^t = \frac{1}{N}\sum_{k=1}^N h(x_k)\nabla L(y_k - h(x_k)'\beta^t) \tag{3.10}$$

The convergence of a sequence of estimators for $M \to \infty$ is established in the next theorem.

Theorem 3.1 (convergence of SQG method (3.7)). *Consider the process (3.7) and assume that* ρ_t *is* $\sigma(\beta^1,...,\beta^t)$*-measurable and satisfies*

$$\rho_t \geq 0 \ \ a.s., \ \ \sum_{t=1}^{\infty}\rho_t = +\infty \ \ a.s., \qquad \sum_{t=1}^{\infty} E\rho_t{}^2 < \infty . \tag{3.11}$$

Then, with probability one, β^M *converges as* $M \to \infty$ *to the global optimum of* $F^N(\beta)$ *on a compact set B.*◊

Noting that, by uniform boundedness of *h(x)*, $\left\|\xi^t\right\| < const$, the proof of this theorem follows from Theorem A.1(a) of the Appendix.

Furthermore, to reduce the variance of the estimator β^M, we can consider the Cesàro averaging of the sequence $\{\beta^t\}$:

$$\overline{\beta}^t = (1-\sigma_t)\overline{\beta}^{t-1} + \sigma_t\beta^t = \sum_{\tau=1}^{t}\rho_\tau\beta^\tau \Big/ \sum_{\tau=1}^{t}\rho_\tau , \qquad t = 1,2,...,M \leq +\infty , \tag{3.12}$$

for $\overline{\beta}^1 = \beta^1$, $t > 1$, and where $\sigma_t = \rho_t \Big/ \sum_{\tau=1}^{t}\rho_\tau$. Under condition (3.11), the estimate $\overline{\beta}^t$ converge jointly with β^t to the optimum B_N^* of (3.2). Suppose that the stepsize multipliers ρ_t are deterministic and are sufficiently small, then, by assumption M1(b) and Theorem A.1(b), we can compute a bound on the variance of this estimator around the true value for the case that $L(\cdot) = \frac{1}{2}\|\cdot\|^2$, as follows:

$$E\left\|\overline{\beta}^t - \beta^*\right\|^2 \leq \left(E\left\|\beta^1 - \beta^*\right\|^2 + C\sum_{\tau=1}^{t}\rho_\tau^2\right) \Big/ \left(L_N\sum_{\tau=1}^{t}\rho_\tau\right), \tag{3.13}$$

where $C = sup_t\left\|\xi^t\right\|^2$, and L_N is a lower bound for the minimal eigenvalue of $E^N h(x)h(x)'$. We will see in the next section that the consistency properties are stronger for the Cesàro estimate than for the SQG-estimate.

Finally, the stochastic optimization method is particularly well suited to calculate a large number of derived statistics. Note that so far we have only considered the problem of point estimation of parameters β on the basis of *N* observations. In applications, it is usually necessary to compute several other values which are functions of β that involve integrals, or expected values. An obvious example is the function $F^N(\beta)$. Though process (3.7) is designed to compute an optimum of $F^N(\beta)$, it does not evaluate this function. Yet, the calculation is readily performed. For example, in case of least squares, the following consistent estimate is available (Ermoliev and Norkin, 1998):

$$F^{t+1} = \Pi_F(F^t - \nu_t(F^t - L(y^t - h(x^t)'\beta^t)), \qquad t = 1,2,... \tag{3.14}$$

where

$$\nu_t \geq 0 , \qquad lim_t\, \nu_t/\rho_t = 0 , \qquad \sum_t \nu_t = +\infty , \qquad \sum_t \nu_t^2 < +\infty ,$$

and the projection $\Pi_F(\cdot)$ is on the interval $[0,\overline{F}]$. Sequence (3.14) gives a consistent estimate of the $min_{\beta\in B} F^N(\beta)$, and provides an indication of the quality of fit for model (3.1). Other statistics that involve the calculation of expectations can be evaluated in a similar way.

Approximation of the constraint set

In the previous section, we stressed the possibly complex structure of constraint set B, which may include inequalities like (2.14a)-(2.14e). Here we present a tractable approximation technique to address this difficulty.

Suppose that $\{x^1,...,x^t\}$ is an i.i.d. sample of variable x, obtained during *t* iterations of the stochastic quasi-gradient method. In this case, the function (2.14a) can be approximated by

$$\varphi_i^t(\beta) = min_{x\in\{x^1,...,x^t\}} \sum_{j=1}^n \partial h_j(x)\big/\partial x_i\ \beta_j\ .$$

By construction $\varphi_i^t(\beta)$ is a convex function and bounded from below: $\varphi_i^t(\beta)\geq\varphi_i(\beta)$. If functions $\partial h_j(x)\big/\partial x_i$ are continuous on X and $\{x^1,...,x^t\}$ are sampled from a positive measure on X, then by the law of large numbers, functions $\varphi_i^t(\beta)$ pointwise converge to $\varphi_i(\beta)$ with probability one. However, because of the convexity of $\varphi_i^t(\beta)$ and the separability of R^n, the function $\varphi_i^t(\beta)$ uniformly converges to $\varphi_i(\beta)$ on any compact set in R^n with probability one (see Rockafellar, 1970).

Alternatively, to approximate function (2.14b), we must, besides $\{x^1,...,x^t\}$, also independently sample the directions $(\ell^1,...,\ell^t)$ from the set $L_+^r=\{\ell\in R_+^r\mid\|\ell\|\leq 1\}$ and evaluate the sequence:

$$\varphi^t(\beta) = min_{(x,\ell)\in\{(x^1,\ell^1),...,(x^t,\ell^t)\}} \sum_{j=1}^n \Big(\sum_i \ell_i\ \partial h_j(x)\big/\partial x_i\Big)\beta_j\ . \tag{3.16}$$

By the same argument as for (2.14a), these functions satisfy $\varphi^t(\beta)\geq\varphi(\beta)$, are convex and uniformly convergent to $\varphi(\beta)$ with probability one. Functions from (2.14c)-(2.14e) can be approximated in a similar way, whereas the function from (2.14f) is linear and need not be approximated. Replacing constraint functions by their approximations, we obtain an outer approximation $B^t\subseteq B$ of the original feasible set B. If the replaced inequality constraints satisfy Slater's condition, then B^t converges to B in the sense of "set convergence" as defined in Rockafellar and Wets (1998).

The next step is to formulate the stochastic quasi-gradient estimation procedure for this specification with a nonstationary feasible set:

$$\beta^t = \Pi_{B^t}(\beta^t - \rho_t\xi^t)\,, \tag{3.17}$$

If after iteration T, we stop updating set B^t then method (3.17) becomes (3.7) and the results on convergence (Theorem 3.1), statistics (3.12) and variance estimate (3.13) apply to (3.17) with B

replaced by B^T. The properties of a fully nonstationary estimation procedure are summarized in Theorem A.1 of the Appendix. Remark that sets B^t are defined by a growing system of linear inequalities. Projection on B^t amounts to a quadratic programming problem and every previous projection can be used as a starting point for the next. It is also possible to apply a technique for dropping nonbinding constraints (see, for example, Hiriart-Urruty and Lemaréchal, 1993). Therefore, procedure (3.12) is implementable numerically. We also mention that techniques are available which avoid solving optimization subproblems at each iteration of the stochastic quasi-gradient algorithm, such as the constraint aggregation principle (Ermoliev et al., 1997), Polyak's method (Polyak, 1983), and average gradients methods (Mikhalevich et al., 1987).

4. Parameter estimation on the basis of kernel density distributions

So far, we supposed that the densities $g^N(x,y)$ are known. In this section, we apply nonparametric (kernel density) estimation to obtain this density and we also consider the alternative problem that makes uses of a regression function to estimate the regression function $y^N(x)$ and the marginal density $g^N(x)$. We show that both approaches are equivalent in case of least squares.
Our approach can be looked at as building on Härdle and Mammen (1993) who propose a test to compare the non-parametric and parametric fits. In fact their proposed test function is almost identical to our risk function.[1]

We start from the sample of observations $\{z^k = (y^k, x^k), k = 1,...,N\}$. The empirical distribution defined by this sample can be represented by the empirical density:

$$G^N(z) = \frac{1}{N}\sum_{k=1}^{N}\delta(z - z^k) \tag{4.1}$$

where z denotes the vector (y, x) and $\delta(z - z^k)$ is the delta-measure concentrated at point z^k. Alternatively, for a given smooth kernel density such that $K(z)\geq 0$, $\int_{R^{r+m}} K(z)dz = 1$, we have the smoothed empirical density:

$$g_\theta^N(z) = \frac{1}{N\theta^{r+m}}\sum_{k=1}^{N}K((z - z^k)/\theta). \tag{4.2}$$

The corresponding risk minimization problem similar to (3.2) reads:

$$F_\theta^N(\beta) = \frac{1}{2}\int L(y - h(x)'\beta)g_\theta^N(y,x)dydx \to min_{\beta\in B}. \tag{4.3}$$

There are a number of reasons to use a smooth density $g_\theta^N(y,x)$ rather than the empirical form (4.1). First, at the practical level it is easier to interpret the data pattern on the basis of a smooth density than from the spikes of the empirical density that produce a field of "needles". Second, as N goes to infinity the estimated kernel density $g_\theta^N(y,x)$ uniformly converges to a true density.

[1] Härdle and Mammen (1993) draw many samples of fixed size to calculate both a parametric and a nonparametric regression and in this way derive an empirical distribution of the risk. Here we only draw one sample from the kernel regression but we could repeat the sampling (bootstrapping) and develop similar tests for our estimators.

Unlike the empirical density that only converges weakly, i.e. in distribution. Third, smoothing suppresses outliers, and interpolates to fill data gaps. Fourth, it helps to avoid singularities in data, making it possible to improve the identifiability. Finally, smoothing makes estimators more stable to data updating. In a certain sense, smoothing as defined by the choice of kernel and window width, implements informal knowledge how the data should actually look like, thus supplementing the parametric model that incorporates the a priori information about the relationship between the variables.

Function (4.3) could be a candidate criterion for risk minimization. However, as in (2.12) there is scope for further simplification. Assume that $K(z) = K_x(x)K_y(y)$, where $K_x(\cdot)$, $K_y(\cdot)$ are also densities. Define the conditional and marginal densities

$$g_\theta^N(y \mid x) = \frac{1}{N\theta^{r+m}} \sum_{k=1}^N K_y((y - y^k)/\theta) K_x((x - x^k)/\theta) \Big/ g_\theta^N(x) \tag{4.4}$$

and

$$g_\theta^N(x) = \frac{1}{N\theta^r} \sum_{k=1}^N K_x((x - x^k)/\theta). \tag{4.5}$$

Substituting these in (4.3) yields for $L(\cdot) = 1/2 \|\cdot\|^2$:

$$F_\theta^N(\beta) = \frac{1}{2} \int \|y - h(x)'\beta\|^2 g_\theta^N(y \mid x) dy \, g_\theta^N(x) dx. \tag{4.6}$$

Hence, it becomes possible, as in (2.12), to reduce (4.3) to

$$F_\theta^N(\beta) = 1/2 \int \left[\int \|y\|^2 g_\theta^N(y \mid x) dy - \left\| y_\theta^N(x) \right\|^2 + \left\| y_\theta^N(x) - h(x)'\beta \right\|^2 \right] g_\theta^N(x) dx \tag{4.7}$$

where $y_\theta^N(x) = \int y g_\theta^N(y \mid x) dy$ is a given nonparametric regression curve. For given window size θ, since the first two terms in (4.7) are constants, this amounts to minimizing

$$\overline{F}_\theta^N(\beta) = \frac{1}{2} \int \left\| y_\theta^N(x) - h(x)'\beta \right\|^2 g_\theta^N(x) dx, \tag{4.8}$$

which means that we must compute the value β that minimizes the integral square error of deviation between the nonparametric function $y_\theta^N(x)$ and the parametric function $h(x)'\beta$, as in (2.16). We can deal with this problem either by calculating the terms of the estimation via Monte Carlo before we apply ordinary least squares, or with an SQG-algorithm of section 3. In case of a more general risk function

$$\overline{F}_\theta^N(\beta) = \frac{1}{2} \int L(y_\theta^N(x) - h(x)'\beta) g_\theta^N(x) dx, \tag{4.9}$$

ordinary least squares does not apply. In (4.8) and (4.9) we see data smoothing of two types: *y*-data are smoothed via the nonparametric regression curve, and *x*-data are smoothed via the kernel density estimator. As a stochastic quasi-gradient ξ^t in (3.7), (3.12) and (3.13) one could use an expression

similar to (3.8): $\xi^t = h(x^t)(h(x^t)'\beta^t - y_\theta^N(x^t))$, where x^t is sampled from empirical density $g_\theta^N(x)$, or expression $\xi^t = h(x^t)(h(x^t)'\beta^t - y_\theta^N(x^t))g_\theta^N(x^t)$ with x^t uniformly sampled in X.

5. Consistency of estimators

In this section, we study the consistency properties of a sequence of estimators obtained as exact or approximate solutions of

$$F_\theta^N(\beta) \to min_{\beta \in B^N}, \tag{5.1}$$

with $F_\theta^N(\beta)$, defined in (4.3) or (4.9), an increasing sample size $N \to \infty$ and $\theta(N) \to 0$, $B^N \to B$. Below we consider three kinds of solutions for (5.1), one is the set B_N^* of optimal solutions of (5.1), the other is an approximate solution β^M obtained after M iterations of the SQG-method applied to (5.1), and the third is the corresponding Cesàro average $\overline{\beta}^M$. In case of least squares, we may suppose that the Monte Carlo integration was sufficiently accurate to ensure that all three yield more or less the same solution but in the general case where SQG is required, the distinction is important. We proceed as follows. Referring to a theorem by Bierens (1988), we establish strong point-wise consistency of kernel density and kernel regression estimates. Next, we show uniform convergence of the risk functions. Finally, we turn to the consistency of the parameter estimates, and prove strong consistency for the case with exact solutions, weak consistency for the Cesàro averaging, and weak consistency for the approximate solution in case of a quadratic norm.

Strong point-wise consistency of kernel density and kernel regression estimates

We treat the use of the empirical density in (4.3) as a special case with $\theta = 0$. The kernel regression curve and the kernel density estimate constructed with the sequence of observations $\{y^1, x^1, ..., y^t, x^t, ...\}$ sampled from density $g^N(x,y)$, have the following forms:

$$y_\theta^N(x) = \sum_{k=1}^N y^k K((x - x^k)/\theta) / \sum_{k=1}^N K((x - x^k)/\theta), \tag{5.2a}$$

$$g_\theta^N(x) = \frac{1}{N\theta^r} \sum_{k=1}^N K((x - x^k)/\theta), \tag{5.2b}$$

where $K(x)$ is a density. Functions $g_\theta^N(x)$ and $y_\theta^N(x)$ are random and depend on the path (sequence of observations) $\omega = \{y^1, x^1, ..., y^t, x^t, ...\}$. To unify notations we suppose that $g_0^N(x)$ (i.e. $\theta = 0$) corresponds to the empirical distribution and that $y_0^N(x)$ defines an empirical regression curve, $y_0^N(x) = \sum_{k=1}^N y^k \overline{\delta}(x - x^k)$, where $\overline{\delta}(x - x^k) = 1$ if $x = x^k$ and $\overline{\delta}(x - x^k) = 0$ otherwise.

First, we consider convergence (consistency) of solutions B_N^* of (5.1) to the unique solution β^* of the true estimation problem

$$F(\beta) \to min_{\beta \in B} \ , \tag{5.3}$$

where $F(\beta)$ is given by (2.17), B is a convex compact set in R^n. It can be shown that if $\theta = \theta(N) \to 0$ as $N \to \infty$, then $g_\theta^N(x)$ approximates a marginal density $g(x) = \int_{R^m} g(x,y)dy$ and $y_\theta^N(x)$ approximates the true regression curve $y(x) = \int_{R^m} yg(x,y)dy / g(x)$ (consistency). Many – weak, strong point wise and uniform – consistency results are available (see e.g. Bierens, 1987; Haerdle, 1990). Here we cite a theorem by Bierens (1988) that establishes strong point wise consistency of these estimators. Similar results can be found in Nadaraya (1989) and Noda (1976).

Theorem 5.1 (strong point-wise consistency of kernel density and regression estimates, Bierens, 1988).
Assume that

(i) $\int_{R^r} K(x)dx = 1$, $\int_{R^r} |K(x)|dx < \infty$, $sup_x |K(x)| < \infty$;

(ii) $sup_x g(x) < \infty$, $sup_x |y(x)|g(x) < \infty$, $sup_x \int y^4 g(x,y)dy < \infty$, $\int y^4 g(x,y)dydx < \infty$;

(iii) $\sum_{t=1}^{\infty} N^{-2}\theta(N)^{-3r} < \infty$.

Then, with probability one $lim_{N\to\infty} g_{\theta(N)}^N(x) = g(x)$ *at every continuity point of* $g(x)$, *and* $lim_{N\to\infty} y_{\theta(N)}^N(x) = y(x)$ *at every continuity point of* $y(x)$ *such that* $g(x) > 0$. ◊

Now if we take $|K(x)|$ and $g(x,y)$ to be continuous and to have bounded supports, then conditions (i), (ii) of the above theorem are satisfied and we can state the following lemma:

Uniform convergence of the risk function

Lemma 5.1 (uniform convergence of risk functions).
Assume that

(i) *functions* $K(x)$ *and* $g(x,y)$ *have bounded supports, and* $|K(x)|$ *is bounded*;

(ii) *functions* $y(x)$ *and* $g(x)$ *may be discontinuous only on the set of Lebesgue measure zero;*

(iii) *with probability one at every continuity point of* $g(x)$ $lim_{N\to\infty,\theta(N)\to 0} g_{\theta(N)}^N(x) = g(x)$;

(iv) *with probability one at every continuity point of* $y(x)$ *and* $g(x)$, *for* $g(x) > 0$, $lim_{N\to\infty,\theta(N)\to 0} y_{\theta(N)}^N(x) = y(x)$;

(v) $h(\cdot)$ *and* $L(\cdot)$ *are continuous*.

Then, the approximate risk functions $F_{\theta(N)}^N(\beta)$ *from (5.2) uniformly converge, in every compact* $A \subset R^n$, *to the true risk function (2.17):*

$$F(\beta) = \int_{R^r} L(y(x) - h(x)'\beta)g(x)dx.$$

Proof. Fix any $\beta \in R^n$. Remark that due to the separability of space R^r, convergence with probability one in (iii), (iv) holds not only separately at each point of continuity of $y(x)$ and $g(x)$ but jointly for all such points. So we may consider sample paths $\omega \in \Omega_1$ such that (iii) takes place for all points of continuity of $g(x)$ while (iv) holds at all x, such that $g(\cdot)$ and $y(\cdot)$ are continuous at x and $g(x) > 0$. Thus, the probability measure of the set Ω_1 equals one. Since by assumption (i) and boundedness of $\theta(N)$, all functions $g^N_{\theta(N)}(x)$ have uniformly bounded supports, integral in (5.2) is taken over some bounded set S. Functions $y^N_{\theta(N)}(x)$ are bounded by (i), $h(x)$ is bounded on S by continuity assumption (v), and $g^N_{\theta(N)}(x)$ are bounded due to boundedness of kernel K. Therefore, integrand $\varphi^N(x) = L(y^N_{\theta(N)}(x) - h(x)'\beta) g^N_{\theta(N)}(x)$ in (5.2) is bounded on S and functions $\varphi^N(x)$ converge point-wise with probability one to $\varphi(x) = L(y(x) - h(x)'\beta) g(x)$: (a) for all $x \in S$ such that $g(x) = 0$ and g is continuous at x, by (iii), (b) for all $x \in S$ such that $g(x) > 0$ and y, g are continuous at x, by (iv), (iii), i.e. convergence may fail only for points of discontinuity of $g(x)$ or $y(x)$ that are negligible by (ii). Now for paths $\omega \in \Omega_1$ (i.e. with probability one) convergence of $F^N_{\theta(N)}(\beta)$ to $F(\beta)$ follows from the Lebesgue dominance convergence theorem.□

Strong consistency of exact risk minimizers

Not surprisingly, for a sequence of convex optimization problems (5.1) we have convergence of exact minimizers $B^N_{\theta(N)}$.

Lemma 5.2 (convergence of exact risk minimizers with probability one). If in (5.3) with probability one the convex objective functions $F^N_{\theta(N)}(\beta)$ converge to the strictly convex function $F(\beta)$ (point wise) and feasible sets B^N are monotonically decreasing ($B^N \subseteq B^{N-1}$) and converge to a nonempty convex compact set B (in the sense of set convergence of Rockafellar and Wets (1998)), then any sequence B^*_N of minimizers of (5.1) converges to the minimizer β^* of (5.3) with probability one.◊

Note that by similar arguments, the minimizers of the problem that use the empirical distribution

$$F^N_0(\beta) = \frac{1}{N} \sum_{k=1}^{N} L(y^k - h(x^k)'\beta) \to min_{\beta \in B^N} \tag{5.4}$$

converge to the minimizer β^{**} of

$$F_0(\beta) = \int_{R^{r+m}} L(y - h(x)'\beta) g(x,y) dxdy \to min_{\beta \in B}. \tag{5.5}$$

This minimizer differs from problem (5.3) because here the objective function penalizes deviations between all *y* and $h(x)'\beta$, whereas (5.3) uses the function *y(x)* and thus aggregates

over all y at given x before penalizing, i.e. it does not penalize for the spread of y at given x. The minimizers β^* and β^{**} coincide in case $L(\cdot)=\frac{1}{2}\|\cdot\|^2$, similar to (4.8).

Weak consistency of Cesàro estimates.

Now consider convergence of SQG-approximations β^M, obtained after a finite number $M=M(N)$ of iterations of SQG-method and corresponding Cesàro estimates $\overline{\beta}^M$, to solutions B_N^* of problem (5.1). If to apply SQG-iterations to (5.1) infinitely many times one can approach to B^N and hence to β^* with probability one. The problem is to give a reasonable stopping criterion for the number of iterations. Condition (ii) of Theorem 5.2 below gives such minimal stopping requirements that guarantee (weak) convergence of obtained approximations to the true value β^* as $N\to\infty$.

In a general case to approximate the optimum B_N^*, $F_{\theta(N)}^N(B_N^*)=min_{\beta\in B^N}F_{\theta(N)}^N(\beta)$, one can apply M stochastic quasigradient iterations:

$$\beta^{t+1}=\Pi_{B^N}[\beta^t-\rho_t\xi^t], \qquad \beta^1\in B^1, \qquad t=1,2,...,M, \tag{5.6}$$

where

$$\xi^t\in -h(x^t)\partial L(y_{\theta(N)}^N(x^t)-h(x^t)'\beta^t)$$

is a stochastic gradient of the function $F_{\theta(N)}^N(\beta)$, i.e. the conditional expectation $E\{\xi^t\mid x^1,...,x^t\}\in\partial F_{\theta(N)}^N(\beta^t)$, for $\partial\{\cdot\}$ denoting a subdifferential of the corresponding function. The corresponding Cesàro estimates are of the form

$$\overline{\beta}^{t+1}=\sum_{k=1}^{t+1}\rho_k\beta^k\Big/\sum_{k=1}^{t+1}\rho_k=(1-\sigma_{t+1})\overline{\beta}^t+\sigma_{t+1}\beta^{t+1}, \quad \sigma_{t+1}=\rho_{t+1}\Big/\sum_{k=1}^{t+1}\rho_k. \tag{5.7}$$

We can now prove the following consistency properties.

Theorem 5.2 (weak consistency of Cesàro estimates for a general risk function / stopping criterion for SQG-method with Cesàro averaging).
Assume that
(i) with probability one functions $F_{\theta(N)}^N$ uniformly converge to a strictly convex function F on compact set $A\subset R^n$ and monotonously decreasing feasible sets B^N ($B^N\subseteq B^{N-1}$) converge to a nonempty convex compact set $B\subseteq A$;
(ii) number of iterations of SQG-method $M(N)\to\infty$, average step sizes $\overline{\rho}_N=\sum_{\tau=1}^{M(N)}\rho_\tau^2\Big/\sum_{\tau=1}^{M(N)}\rho_\tau\to 0$ and $\sum_{t=1}^{M(N)}\rho_t\to\infty$ as $N\to\infty$. Then, the Cesàro estimates $\overline{\beta}^{M(N)}$ converge in probability to the true estimate β^* (solution of (5.1)) as $N\to\infty$.

Proof. By Theorem A.1(b) Cesàro estimates $\overline{\beta}^{M(N)}$ satisfy condition

$$EF^N_{\theta(N)}(\overline{\beta}^{M(N)}) - F^*_N \le \left(dist^2(\beta^1, B^*_N) + C\sum_{\tau=1}^{M(N)} \rho_\tau^2 \right) \Big/ \left(2\sum_{\tau=1}^{M(N)} \rho_\tau \right) \le$$

$$\le dist^2(\beta^1, B^*_N) \Big/ \left(2\sum_{\tau=1}^{M(N)} \rho_\tau \right) + C\overline{\rho}_N \to 0 .$$

Then,

$$EF(\overline{\beta}^{M(N)}) - F^* \le E\left| F(\overline{\beta}^{M(N)}) - F^N_{\theta(N)}(\overline{\beta}^{M(N)}) \right| + E\left| F^N_{\theta(N)}(\overline{\beta}^{M(N)}) - F^*_N \right| +$$

$$+ \left| F^*_N - F^* \right| \to 0 ,$$

and hence sequence $F(\overline{\beta}^{M(N)}) - F^*$ converges to zero in probability. By continuity, for any $\varepsilon > 0$ there exists $\gamma(\varepsilon) > 0$ such that $F(\beta) - F^* \ge \gamma(\varepsilon)$ whenever distance $dist(\beta, \beta^*) \ge \varepsilon$. Thus, probability

$$P\{ dist(\overline{\beta}^{M(N)}, \beta^*) \ge \varepsilon \} \le P\{ F(\overline{\beta}^{M(N)}) - F^* \ge \gamma(\varepsilon) \} \to 0 . \square$$

Weak consistency of SQG

We only prove consistency for the quadratic case. Condition (vi) of Theorem 5.3 below gives a minimal (stopping) requirement on the number of iterations $M(N)$ of SQG-method to guarantee (weak) convergence of obtained approximations to true value β^* as $N \to \infty$.

Theorem 5.3 (weak consistency of SQG-estimates in case of quadratic risk function).
Assume that

(i) *with probability one functions* $F^N_{\theta(N)}$ *uniformly converge to a strictly convex function* F *on compact set* $A \subset R^n$ *and monotonously decreasing feasible sets* B^N ($B^N \subseteq B^{N-1}$) *converge to a nonempty convex compact set* $B \subseteq A$;

(ii) $\left\| h(x^t) \right\| \le H$, $\left\| h(x^t) h(x^t)' \right\| \le m$, $\left\| \varepsilon_t \right\|^2 \le N^2$;

(iii) $\beta' h(x^t) h(x^t)' \beta \ge L\|\beta\|^2$ *for all* $\beta \in R^l$, *where* H, m, L, N *are positive constants.*

Suppose that

(iv) *step sizes* ρ_t, $t_0 \le t \le M(N)$, *in procedure (5.6) satisfy conditions*

$$\frac{r}{t^\alpha} \le \rho_t \le \frac{R}{t^\alpha} \le \frac{L}{3m^2}, \quad 0 < \alpha \le \min\{1, 2Lr\},$$

where r, R, b, α *are deterministic positive constants*;

(v) *number of iterations of SQG-method* $M(N) \to \infty$ *as* $N \to \infty$.

Then, SQG-estimates $\beta^{M(N)}$ *converge in probability to the true estimate* β^* *(solution of (5.1)) as* $N \to \infty$.

Proof. By (i) sets B^N are uniformly bounded for sufficiently large N, hence $\left\| \beta^t - \beta^*_N \right\| \le b$ for $N \ge N_0$ and some constant b. By Theorem A.2

$$E\left\|\beta^{M(N)} - \beta_N^*\right\|^2 \le \frac{Q}{M^{\alpha}(N)}, \qquad Q = \max\{b^2, \frac{3R^2m^2N^2}{Lr-\alpha}\}, \qquad M(N) \ge t_0, \tag{5.8}$$

and $\left\|\beta^{M(N)} - \beta_N^*\right\| \to 0$ in probability as $N \to \infty$. By Lemma 5.2 $\left\|\beta_N^* - \beta^*\right\| \to 0$ with probability one (and hence in probability) as $N \to \infty$.

Since $\left\|\beta^{M(N)} - \beta^*\right\| \le \left\|\beta^{M(N)} - \beta_N^*\right\| + \left\|\beta_N^* - \beta^*\right\|$ then $\left\|\beta^{M(N)} - \beta^*\right\| \to 0$ in probability as $N \to \infty$. □

6. Conclusion

We have described a risk minimization approach to estimate a flexible form that meets a priori restrictions on slope and curvature by means of constraints on both the estimated parameters and the function values. The resulting constrained risk minimization combines parametric and nonparametric estimation and contains integrals and implicit constraints. We found that the simulation approach, which is common in econometrics, only applies when the model is linear in parameters, has simple constraints on parameters and a quadratic risk function. To deal with other cases, we use a stochastic optimization technique known as the stochastic quasi-gradient method with Cesàro averaging. This method is also applicable to an expanding series of random observations, and produces asymptotically (weakly) convergent estimates.

With respect to further research, formulation of tests on parameters and predictions would be a first priority and for this the simulation approach of Härdle and Mammen (1993) discussed in section 4 would seem practicable. Furthermore, building on the nonstationary version presented in section 3, serial correlation between observations could be allowed for.

Appendix: Mathematical background

The following lemma is a stochastic version of the Lyapunov function method for discrete time stochastic processes and is used to prove convergence of SQG-estimates.

Lemma A.1 (Ermoliev and Norkin, 1998).
Let $v_t \geq 0$, $\rho_t \geq 0$, w_t, γ_t, $t \geq 1$, *be a sequence of random variables (scalars). Suppose that each of the following conditions is fulfilled with probability one:*

(i) $v_{t+1} \leq v_t - \rho_t w_t + \gamma_t$, all $t \geq 1$;

(ii) $lim_t\, \rho_t = 0, \quad \sum_{t=1}^{\infty} \rho_t = +\infty$;

(iii) $v_t + \sum_{t=1}^{\infty} \gamma_t < +\infty$;

(iv) for any $\{t_s \to \infty\}$ if $lim\,inf_s\, v_{t_s} > 0$ then $lim\,inf_s\, w_{t_s} > 0$

(v) for any $\{t_s \to \infty\}$ if $lim\,sup_s\, v_{t_s} < +\infty$ then $lim\,sup_s\, | w_{t_s} | < +\infty$.

Then, $lim_t\, v_t = 0$, *with probability one.* ◊

The next lemma establishes the rate of convergence to zero for sequences satisfying some recurrent inequality.

Lemma A.2 (Katkovnik, 1976, p.282)**.**
Let $\{v_t\}_{t\geq t_0}$ *be a sequence of nonnegative numbers such that*

(i) $v_{t+1} \leq \left(1 - \frac{\rho}{t^\alpha}\right) v_t + \frac{C}{t^{\alpha+\gamma}}, \quad v_1 < +\infty;$

(ii) $0 < \alpha \leq 1, \quad 0 < \gamma < \rho, \quad C > 0.$

Then,

$$v_t \leq \frac{Q}{t^\gamma}, \qquad Q = max\{v_{t_0}, \frac{C}{\rho - \alpha}\}, \qquad t \geq t_0 . \Diamond$$

Lemmas A.1 and A.2 are used in the proofs of Theorems A.1 and A.2, respectively.

Theorem A.1 (convergence of stochastic quasi-gradient method and corresponding Cesàro sequence).
Let

(i) *convex functions* $F^t(\beta)$ *uniformly converge to function* $F(\beta)$ *on some compact set* $A \subset R^n$,

(ii) *monotonic (decreasing) sequence of convex compact sets* $\{B^t \subseteq B^{t-1} \subset A\}$ *converges to a compact set* $B \subseteq A$;

(iii) *sequence of approximations* $\{\beta^t\}$ *is constructed by stochastic quasi-gradient method:*
$\beta^{t+1} = \Pi_{B^t}[\beta^t - \rho_t \xi^t(\beta^t)], \qquad \beta^1 \in B^1, \qquad t = 1,2,\ldots,$

with stochastic quasi-gradients $\xi^t(\beta^t)$ *such that* $E\{\xi^t(\beta^t) \mid \beta^t\} \in \partial F^t(\beta^t)$ *and* $\left\|\xi^t(\beta^t)\right\| \leq C$, *adjustment coefficients* $\rho_t \geq 0$ *are measurable with respect to* $\sigma\{\beta^1,\ldots,\beta^t\}$.

Then,

(a) *for adjustment coefficients such that* $\rho_t \geq 0$ *a.s.,* $\sum_{t=1}^{\infty} \rho_t = \infty$ *a.s.,* $\sum_{t=1}^{\infty} E\rho_t^{\,2} < \infty$, *with probability one, the sequence* β^t *converges to* $arg\,min_{\beta \in B} F(\beta)$ *and* $lim_t F(\beta^t) = min_{\beta \in B} F(\beta)$;

(b) *for any nonnegative deterministic adjustment coefficients* ρ_t, *the Cesàro sequence* $\overline{\beta}^t = (1 - \sigma_t)\overline{\beta}^{t-1} + \sigma_t \beta^t = \sum_{\tau=1}^{t} \rho_\tau \beta^\tau \Big/ \sum_{\tau=1}^{t} \rho_\tau$,

satisfies estimates

$$EF(\overline{\beta}^t) - F^* \leq \left(Ed^2(\beta^1, B^*) + C\sum_{\tau=1}^{t} \rho_\tau^2 + 2\sum_{\tau=1}^{t} \rho_\tau E\, max_{\beta \in A}\left|F^\tau(\beta) - F(\beta)\right|\right) \Big/ \left(2\sum_{\tau=1}^{t} \rho_\tau\right)$$

where $d(\beta, B^*) = min_{b \in B^*}\|\beta - b\|$.

Proof. The proof of statement (a) is based on sufficient conditions from Lemma A.1 for convergence to zero of a nonnegative sequence of random variables with probability one. Let B^* be the set of minimizers of F on B, $\beta^* \in B^*$ and $F^* = F(\beta^*)$. Denote $\xi^t = \xi^t(\beta^t)$ and the

conditional expectation by $\bar{\xi}^t = E\{\xi^t(\beta^t)\,|\,\beta^t\}$. As $\|\xi^t\| \le C$ by assumption, it follows that $\|\bar{\xi}^t\| \le C$. Since β^* belongs to all B^t,

$$\|\beta^{t+1} - \beta^*\|^2 = \|\Pi_{B^t}[\beta^t - \rho_t\xi^t] - \beta^*\|^2 \le \|\beta^t - \beta^* - \rho_t\xi^t\|^2 =$$

$$= \|\beta^t - \beta^*\|^2 - 2\rho_t\langle\xi^t, \beta^t - \beta^*\rangle + \rho_t^2\|\xi^t\|^2 =$$

$$= \|\beta^t - \beta^*\|^2 - 2\rho_t\langle\bar{\xi}^t, \beta^t - \beta^*\rangle + 2\rho_t\langle\bar{\xi}^t - \xi^t, \beta^t - \beta^*\rangle + \rho_t^2\|\xi^t\|^2.$$

By convexity

$$F^t(\beta^*) - F^t(\beta^t) \ge \langle\bar{\xi}^t, \beta^* - \beta^t\rangle,$$

and, therefore,

$$\|\beta^{t+1} - \beta^*\|^2 \le \|\beta^t - \beta^*\|^2 + 2\rho_t(F^t(\beta^*) - F^t(\beta^t)) + 2\rho_t\langle\bar{\xi}^t - \xi^t, \beta^t - \beta^*\rangle + \rho_t^2\|\xi^t\|^2$$

$$\le \|\beta^t - \beta^*\|^2 + 2\rho_t(F(\beta^*) - F(\beta^t)) + 2\rho_t\langle\bar{\xi}^t - \xi^t, \beta^t - \beta^*\rangle +$$

$$+ 2\rho_t|F^t(\beta^t) - F(\beta^t)| + 2\rho_t|F^t(\beta^*) - F(\beta^*)| + \rho_t^2\|\xi^t\|^2.$$

Denote $\Delta^t = sup_{\beta\in A}|F^t(\beta) - F(\beta)|$. Now let us introduce function $d(\beta) = \min_{b\in B^*}\|\beta - b\|$ and choose β_t^* such that $\|\beta^t - \beta_t^*\| = d(\beta^t)$. Then, $\|\beta^{t+1} - \beta_t^*\| \le d(\beta^{t+1})$ and we obtain inequality

$$d(\beta^{t+1}) \le d(\beta^t) - 2\rho_t(F(\beta^t) - F^*) + 4\rho_t\Delta^t + C^2\rho_t^2 + \langle\bar{\xi}^t - \xi^t, \beta^t - \beta_t^*\rangle. \qquad \text{(A1)}$$

Denote $\nu_t = d(\beta^t)$, $w_t = 2(F^* - F(\beta^t)) + 4\Delta^t + C\rho_t$, $\gamma_t = 2\rho_t\langle\bar{\xi}^t - \xi^t, \beta^t - \beta_t^*\rangle$. Thus, $\nu_{t+1} \le \nu_t - \rho_t w_t + \gamma_t$, *all* $t \ge 1$. Sequence $\{\mu_t = \sum_{\tau=1}^{t}\gamma_\tau\}$ constitutes a martingale with respect to a sequence of σ-algebras F_t generated by $\{x^1, y^1, ..., x^t, y^t\}$. By assumptions (iii) and (a), martingale $\{\mu_t\}$ a.s. converges, i.e. $\sum_{t=1}^{\infty}\gamma_t < +\infty$ a.s. Quantities $\nu_t, w_t, \rho_t, \gamma_t$ satisfy conditions of Lemma A.1, hence $\{\nu_t\}$ converges to zero with probability one. Since convergence with probability one is preserved under continuous transformations, $lim_{t\to\infty} F(\beta^t) = F^*$ a.s.

To prove assertion (b) we follow Nemirovski and Yudin (1983). Taking expectations from both sides of (A1), we obtain

$$Ed(\beta^{\tau+1}) \le Ed(\beta^\tau) - 2\rho_t(EF(\beta^\tau) - F^*) + 4\rho_\tau E\Delta^\tau + C^2\rho_\tau^2.$$

Summing these inequalities from $\tau = 1$ until t, we get

$$0 \le Ed(\beta^{t+1}) \le Ed(\beta^1) - 2(\sum_{\tau=1}^{t}\rho_t EF(\beta^\tau) - F^*\sum_{\tau=1}^{t}\rho_\tau) + 4\sum_{\tau=1}^{t}\rho_\tau E\Delta^\tau + C^2\sum_{\tau=1}^{t}\rho_t^2.$$

By convexity, $EF(\bar{\beta}^t) \le \sum_{\tau=1}^{t}\rho_t EF(\beta^\tau) \Big/ \sum_{\tau=1}^{t}\rho_\tau$. Finally we obtain

$$EF(\bar{\beta}^t) - F^* \le \left(Ed(\beta^1) + 4\sum_{\tau=1}^{t}\rho_\tau E\Delta^\tau + C^2\sum_{\tau=1}^{t}\rho_t^2\right)\Big/\left(2\sum_{\tau=1}^{t}\rho_\tau\right)\square$$

Theorem A2 (rate of convergence of SQG-method in case of quadratic risk function)**.** *Assume that B is a convex set and for all t :*

(i) $\left\|h(x^t)\right\| \le H$, $\left\|h(x^t)h(x^t)'\right\| \le m$, $\left\|\varepsilon_t\right\|^2 \le N^2$;

(ii) $\beta' h(x^t)h(x^t)'\beta \ge L\left\|\beta\right\|^2$ for all $\beta \in R^l$,

where H, m, L, N *are positive constants. Suppose that for all* $t \ge t_0$,

(iii) $\left\|\beta^t - \beta_N^*\right\| \le b$;

(iv) $\frac{r}{t^\alpha} \le \rho_t \le \frac{R}{t^\alpha} \le \frac{L}{3m^2}$, $0 < \alpha \le 1$, $0 < \alpha < 2Lr$;

where r, R, b, α *are deterministic positive constants. Then,*

$$E\left\|\beta^t - \beta_N^*\right\|^2 \le \frac{Q}{t^\alpha}, \quad Q = max\{b^2, \frac{3R^2H^2N^2}{Lr-\alpha}\}, \quad t \ge t_0.$$

Proof. The following estimates hold true

$$\begin{aligned}
\left\|\beta^{t+1} - \beta_N^*\right\|^2 &\le \left\|\beta^t - \beta_N^* - \rho_t[h_t(x^t)(h_t(x^t)'(\beta^t - \beta_N^*) - \varepsilon_t)]\right\|^2 \\
&\le \left\|\beta^t - \beta_N^*\right\|^2 - 2\rho_t(\beta^t - \beta_N^*)'h_t(x^t)h_t(x^t)'(\beta^t - \beta_N^*) + 2\rho_t(\beta^t - \beta_N^*)h_t(x^t)'\varepsilon_t \\
&+ 3\rho_t^2\left\|h_t(x^t)(h_t(x^t)'\right\|^2\left\|\beta^t - \beta_N^*\right\|^2 + 3\rho_t^2\left\|h_t(x^t)\right\|^2\left\|\varepsilon_t\right\|^2 + \\
&\le \left\|\beta^t - \beta_N^*\right\|^2 - 2\rho_t(\beta^t - \beta_N^*)'h_t(x^t)h_t(x^t)'(\beta^t - \beta_N^*) + \\
&+ 2\rho_t(\beta^t - \beta_N^*)h_t(x^t)'\varepsilon_t + 3\mathrm{m}^2\rho_t^2\left\|\beta^t - \beta_t^*\right\|^2 + 3H^2N^2\rho_t^2 \le \\
&\le \left\|\beta^t - \beta_t^*\right\|^2 - \rho_t(2L - 3m^2\rho_t)\left\|\beta^t - \beta_t^*\right\|^2 + \\
&+ 2\rho_t(\beta^t - \beta_t^*)'h_t(x^t)\varepsilon_t + 3H^2N^2\rho_t^2
\end{aligned} \quad . \qquad \text{(A2)}$$

From (A2) by (iii), (iv) for all t we have

$$\left\|\beta^{t+1} - \beta_{t+1}^*\right\|^2 \le \left\|\beta^t - \beta_t^*\right\|^2 - \frac{r}{t^\alpha}L\left\|\beta^t - \beta_t^*\right\|^2 + 2\rho_t(\beta^t - \beta_t^*)'h_t(x^t)\varepsilon_t + \frac{3H^2N^2R^2}{t^{2\alpha}}.$$

Taking expectations from both sides of this inequality and denoting $v_t = E\left\|\beta^t - \beta_t^*\right\|^2$, for all $t \ge t_0$, we get

$$v_{t+1} \le (1 - \frac{Lr}{t^a})v_t + \frac{C}{t^{2\alpha}}, \qquad C = 3H^2N^2R^2 .$$

Then, by Lemma A.2 for $0 < \alpha < \min\{1, Lr\}$ we have

$$v_t \le \frac{Q}{t^\alpha}, \quad Q = \max\{v_{t_0}, \frac{C}{Lr-\alpha}\}. \square$$

This theorem estimates the mean rate of progress of method (4.10) for all *t*. To estimate the rate of convergence we can strengthen (iii) to require boundedness of B. Step sizes ρ_t can be random but by (iv) lie within deterministic bounds (note that $\sum_{t=t_0}^{T=\infty}\rho_t = +\infty$).

References

Barnett, W.A., J. Geweke, and M. Wolfe (1991): 'Semi-nonparametric Bayesian estimation of consumer and factor demand models', in Barnett et al., eds., *Equilibrium Theory and Applications*. Cambridge: Cambridge University Press.

Bierens, H.J. (1987): 'Kernel estimators of regression functions', in: *Advances in Econometrics: fifth world congress* Truman F. Bewley (ed.), Cambridge: Cambridge University Press.

Bierens, H.J. (1988): 'The Nadaraya-Watson kernel regression function estimator', Research Memorandum 1988-58, Free University, Amsterdam.

Davidson R. and J. G. M. MacKinnon (1993): Estimation and Inference in Econometrics, NY: Oxford university press.

Ermoliev, Y.M. (1976): *Methods of stochastic programming*. Moscow: Nauka.

Ermoliev, Yu. M. (1988): 'Stochastic quasigradient methods', in Yu. Ermoliev, R.J.B. Wets, (eds.) *Numerical Techniques for Stochastic Optimization,* Berlin: Springer.

Ermoliev, Y.M., A. Kryazshimski, and A. Rusczchinski (1997): 'Constraint aggregation principle in convex optimization', *Math. Progr.* 76, pp. 353-372.

Ermoliev, Y.M., and V.I. Norkin (1998): 'On the non-stationary law of large numbers for dependent random variables and its application in stochastic optimization', *Kibernetika i sistemnyi analiz (Cybernetics and systems analysis)*, 4: 94-106 (In Russian, English version in: *IIASA Interim Report IR-98-009*, Web: http://www.iiasa.ac.at/Publications).

Ermoliev, Yu., M.A. Keyzer and V. Norkin (2000): 'Global Convergence of the stochastic tatonnement process', *Journal of Mathematical Economics* 34:173-190.

Folmer, C. M.A., M.A. Keyzer, M.D. Merbis, H.J.J. Stolwijk, and P.J.J. Veenendaal (1995) *The common agricultural policy beyond the MacSharry reform*. North Holland, Amsterdam.

Gallant, A.R. (1981): 'On the basis in flexible functional forms and essentially unbiased form: the Fourier form', *Journal of Econometrics*, 15: 211-245.

Gourieroux, Ch., and A. Monfort (1996): *Simulation-based econometric methods.* Oxford: Oxford University Press.

Härdle, W. (1990): *Applied nonparametric regression*, Econometric Society Monograph Series, 19. Cambridge: Cambridge University Press.

Härdle, W. and E. Mammen (1993): 'Comparing Nonparametric versus parametric regression fits', *Annals of Statistics*, 21:1926-1947.

Hiriart-Urruty J-B. and C. Lemaréchal (1993): Convex Analysis and Minimization Algorithms I, II, Berlin: Springer-Verlag.

Huber, P.J. (1981): *Robust Statistics*. New York: Wiley.

Keyzer, M.A. and R. Gerlagh (2001): 'Sustainability and the intergenerational distribution of natural resources entitlements'. In: *Journal of Public Economics* 79: 315-341.

Keyzer, M.A. and V. Ginsburgh (1997): The structure of applied general equilibrium models. Cambridge, MA: MIT Press.

Mikhalevich, V.S., A.M. Gupal, and V.I. Norkin (1987): *Methods of Nonconvex Optimization*. Moscow: Nauka (In Russian).

Nadaraya, E.A. (1989) *Nonparametric estimation of probability densities and regression curves.* Translated from Russian by Samuel Klotz. Amsterdam: Kluwer.

Nemirovski, A.S., and D.B. Yudin (1983): *Complexity of problems and the efficiency of optimization methods*. Moscow: Nauka.

Noda, K. (1976): 'Estimation of the regression function by the Parzen Kernel-type density estimators', *Annals of the Institute of Statistics and Mathematics*, 28: 221-234.

Polyak, B.T. (1983): *Introduction to optimization*. Moscow: Nauka.

Robinson, P. (1988): 'Root-N-Consistent Semi-parametric regression', *Econometrica*, 56: 931-954.

Rockafellar, T. (1970): *Convex Analysis*. Princeton: Princeton University Press.

Rockafellar, T., and R. Wets (1998): *Variational Analysis*. Berlin: Springer.

Wets, R.J.-B (1998): Statistical estimation from an optimisation viewpoint, *Annals of Operations Research*, 84, 79-102.

Varian, H. (1992): *Microeconomic Theory.* New York: Norton University Press.

Yatchew, A. (1998): 'Nonparametric Regression Techniques in Economics', *Journal of Economic Literature*, 36: 669-721.