UDC 330.115

T. ERMOLIEVA
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *ermol@iiasa.ac.at*.

Y. ERMOLIEV
International Institute for Applied Systems Analysis, Laxenburg, Austria; V.M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, e-mail: *ermoliev@iiasa.ac.at*.

P. HAVLIK
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *havlik.petr@gmail.com*.

A. LESSA-DERCI-AUGUSTYNCZIK
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *augustynczik@iiasa.ac.at*.

N. KOMENDANTOVA
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *komendan@iiasa.ac.at.*

T. KAHIL
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *kahil@iiasa.ac.at*.

J. BALKOVIC
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *balkovic@iiasa.ac.at*.

R. SKALSKY
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *skalsky@iiasa.ac.at*.

C. FOLBERTH
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *folberth@iiasa.ac.at*.

P.S. KNOPOV
V.M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, e-mail: *knopov1@yahoo.com*; *knopov1@gmail.com*.

G. WANG
China Agricultural University (CAU), Beijing, China, e-mail: *gangwang@cau.edu.cn*.

CONNECTIONS BETWEEN ROBUST STATISTICAL ESTIMATION, ROBUST DECISION=MAKING WITH TWO-STAGE STOCHASTIC OPTIMIZATION, AND ROBUST MACHINE LEARNING PROBLEMS[1]

Abstract. The paper discusses connections between the problems of two-stage stochastic programming, robust decision-making, robust statistical estimation, and machine learning. In the conditions of uncertainty, possible extreme events and outliers, these problems require quantile-based criteria, constraints, and "goodness-of-fit" indicators. The two-stage STO problems with quantile-based criteria can be effectively solved with the iterative stochastic quasigradient (SQG) solution algorithms. The SQG methods provide a new type of machine learning algorithms that can be effectively used for general-type nonsmooth, possibly discontinuous, and nonconvex problems, including quantile regression and neural network training. In general problems of decision-making, feasible solutions and concepts of optimality and robustness are characterized from the context of decision-making situations. Robust ML

[1]The development of robust decision-making, statistical estimation, machine learning and Big Data analysis problems, respective solution procedures and case studies, is supported by the joint project between the International Institute for Applied Systems Analysis (IIASA) and National Academy of Sciences of Ukraine (NASU) on "Integrated robust modeling and management of food-energy-water-land use nexus for sustainable development". The work has received partial support from the Ukrainian National Fund for Strategic Research, grant No. 2020.02/0121, and project CPEA-LT-2016/10003 jointly with Norwegian University for Science and technology. The paper contributes to EU PARATUS (CL3-2021-DRS-01-03, SEP-210784020) project on "Promoting disaster preparedness and resilience by co-developing stakeholder support tools for managing systemic risk of compounding disasters".

approaches can be integrated with disciplinary or interdisciplinary decision-making models, e.g., land use, agricultural, energy, etc., for robust decision-making in the conditions of uncertainty, increasing systemic interdependencies, and “unknown risks.”

Keywords: two-stage STO, robust decision-making and statistical estimation, robust quantile regression, machine learning, general problems of robust decision making, systemic risks, uncertainties.

INTRODUCTION

Various problems of decision-making under uncertainty, statistics, big data analysis, artificial intelligence (AI) can be formulated or can be reduced to the two-stage stochastic optimization (STO) problems. For example, these are problems inherent to engineering, economics, finance, operations research, etc., that involve minimization or maximization of an objective or a goal function when randomness is present in model’s data and parameters, e.g., observations, costs, prices, returns, crop yields, temperature, precipitation, soil characteristics, water availability, emissions, return periods of natural disasters, etc. Uncertain parameters can be interpreted as environment-determining variables [1], that conditions the performance of the system under investigation. Randomness can enter the problems in several ways:

1) through stochastic (exogenous or/and endogenous) parameters, e.g., costs, prices, returns, crop yields;

2) stochastic resources, e.g., water, land, biomass, investments;

3) random occurrence of exogenous natural disasters depleting resources and assets;

4) stochastic endogenous events (systemic risks) induced by decisions of various agents.

Non-normal probability distributions of stochastic parameters and percentile-based criteria functions. Stochastic variables can be characterized by means of a probability distribution (parametric or nonparametric) function or can be represented by probabilistic scenarios. Probability distributions of stochastic parameters are often non-normal, heavy tailed and even multimodal. For example, Fig. 1 depicts probability distribution of wheat yields for several countries — major grain producers. Horizontal axis denotes yield (in kilograms per hectare of harvested land) and vertical axis shows the number of years (frequency) the corresponding yield occurred in the 1960–2012 period. Cumulative distribution refers to the percentage of total of the yield occurrences at or below the value on the horizontal axis. Low wheat yields on the left-hand side visualized in all four panels can cause imbalances in grain supply-demand chains and thus lead to prices increase, market disturbances, trade bans, etc. These low values correspond to about 20 % percentiles of the crop yield observations and can correspond to production years characterized by bad weather conditions, e.g., low precipitation or/and high temperature in important grain growth periods/months (e.g., grain filling period).

If probability distribution functions of stochastic parameters are non-normal, non-symmetrical, or even heavy tailed, such decision-making or/and parameter estimation criteria as Mean-Variance, Ordinary Least Square (OLS) or Root Mean Squared Error (RMSE) are not appropriate. They rely only on the first two moments, i.e., mean and variance, which characterize normal distributions. Thus, the information about the tails of the distributions is not accounted for. Extreme values (outliers) can distort the results, i.e., the criteria are not robust. In statistics and machine learning, the OLS and RMSE estimates can be misleading, and the effects can be different for different subsets of data sample.

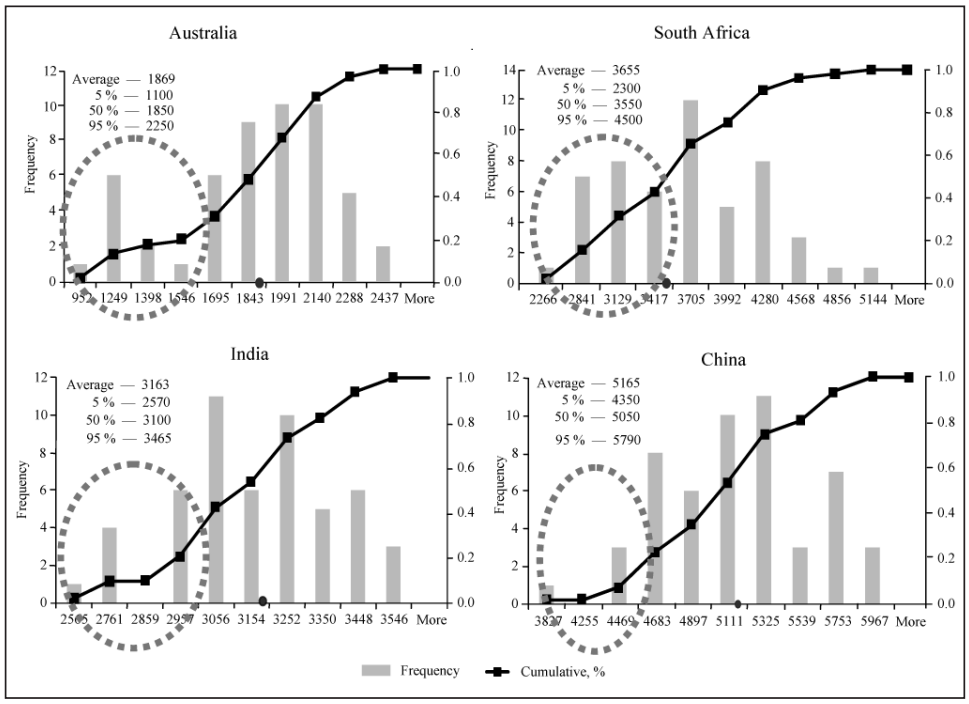

Fig. 1. Empirical wheat yield distributions by major grain producers, 1960– 2012, FAO data analysis

For decision-making, statistical estimation and machine learning problems in the presence of non-normal, heavy tailed and possibly multimodal probability distributions it is more natural to use the median or other quantile-based criteria instead of the mathematical expectation. The problems can be formulated in the form of a two-stage STO with discontinuous chance (probabilistic or quantile-based) constraints used for robust decision-making under uncertainty and risks [2]. For nonsmooth, possibly discontinuous and nonconvex problems, including quantile-based regression and neural networks training, the discussion about the application and convergence of the stochastic quasigradient (SQG) methods can be found, e.g., in ([3–7]). For the case of a general endogenous reinforced learning, the convergence of the SQG procedure was proven in [8] based on the results of nondifferentiable optimization providing a new type of machine learning algorithms solving the problem of distributed decentralized models' linkage under asymmetric information and uncertainty. In this case, the SQG procedure enables to organize an iterative computerized negotiation between individual systems (models) representing Intelligent Agents (IA). The models act as "agents" that communicate with a "central hub" (a regulator) and take decisions in order to maximize the "cumulative reward."

Two-stage decisions. Often, decisions (actions) or parameter estimation have to be performed ex-ante before the true values (realizations) of the uncertain parameters become known or observed [9–15]. Sometimes, the observations can be only partial or incomplete, i.e., incomplete "learning". These situations happen, for example, in the process of agricultural production planning under weather variability and market risks [10–13]; water reservoir management [14]; investments in irrigation and crop storage facilities [10]; energy technologies investment planning [13], and in many other application areas.

The ex-ante decisions may require revisions and corrections after the information about the uncertain parameter realization becomes known (i.e., after "learning" or partial "learning" of uncertain parameters values). Which means, that the ex-ante decisions/solutions can incur the costs for their correction, revision, or reversion. Thus,

there are two types (two-stage) of decisions. The term "two-stage" does not necessary mean two consequent time intervals. Rather, they correspond to steps in the decision process. The ex-ante decision x may be "here-and-now" decision whereas ex-post decisions y correspond to all future actions to be taken in different time periods in response to the environment created by the chosen x and the observed value of the uncertain parameter ω in that specific time period [10–13]. In another instance, the x, y solutions may represent sequences of control actions over a given time horizon. In the case of dynamical systems, there may also be an additional group of variables characterizing the state of the system [14, 15]. These problems often emerge in operations research models, economics and system analysis, in the theory of optimal control and its applications in engineering, inventory control, etc. Specific applications include: deriving parameters of a statistical model (parametric or nonparametric) or training a machine learning model that maps an input to an output based on examples of input-output pair through minimizing/maximizing a percentile-based "goodness-of-fit" criteria; deciding on optimal dynamic investments allocation into new technologies (irrigation, energy, agricultural, water management) to minimize costs or/and maximize profits accounting for various norms and constraints; deciding when to release water from a multipurpose reservoir for hydroelectric power generation, agricultural and industrial production, household requirements, environmental constraints, food protection; defining insurance coverage and premiums to minimize risk of bankruptcy of insurers and risk of overpayments of individuals.

Basic model of a two-stage decision-making and parameter estimation. Let us illustrate the concept of the two-stage decision making and robust statistical estimation problem with an example of a simplest two-stage STO model.

Assume, there are observations of an uncertain variable ω, which can be associated with a stochastic parameter, e.g., demand for a certain product or resource (water level), a crop yield, a weather parameter (temperature, precipitation, pressure). The choice of the decision $\bar{x} \geq x \geq 0$, to fulfill stochastic demand ω (or the estimate of a crop yield or weather parameter based on observations of ω) can be associated with a function $f(x, \omega)$ reflecting costs of overestimation and underestimation of ω. In the simplest case, $f(x, \omega)$ is a random piecewise linear function $f(x, \omega) = \max\{\alpha(x - \omega), \beta(\omega - x)\}$, where α is the unit overestimation/surplus cost (associated e.g., with storage costs) and β is the unit underestimation/shortage cost (associated e.g., with import costs, which also can be interpreted as borrowing). The problem is to find the level x that is "optimal", in a sense, for all foreseeable random scenarios/observations ω.

The expected cost criterion leads to the minimization of the following function:

$$F(x) = Ef(x, \omega) = E \max\{\alpha(x - \omega), \beta(\omega - x)\}$$

subject to $\bar{x} \geq x \geq 0$ for a given upper bound x. This stochastic minimax problem is also reformulated as a two-stage stochastic programming.

The optimal solution minimizing $F(x)$ and more general stochastic minimax problems defines quantile type characteristics of solutions [2, 10–15], e.g., CVaR risk measures. For example, if the distribution of ω has a density, $\alpha, \beta > 0$, then the optimal solution x minimizing $F(x)$ is the quantile defined as $Pr\{\omega \leq x\} = \beta / (\alpha + \beta)$.

Function $F(x)$ is convex, therefore the SQG algorithm can be defined as the following discontinuous adaptive machine learning process:

$$x^{k+1} = \min\{\max\{0, x^k - \rho_k \xi^k\}, \bar{x}\}, \ k = 0, 1, \ldots,$$

where $\xi^k = \alpha$, if the current level of production x^k exceeds the observed demand ω and $\xi^k = -\beta$ otherwise, ρ_k is a positive step size. The SQG method can be viewed as

an adaptive machine learning process which is able to learn the optimal level x through sequential adjustments of its current levels $x^0, x^1, x^2, \ldots, x^k, \ldots$ to observable (or simulated) scenarios/observations $\omega^0, \omega^1, \omega^2, \ldots, \omega^k, \ldots$. The SQG process is a convergent with probability 1 sequential estimation procedure. Problem of minimizing $F(x)$ illustrates the essential difference between the so-called scenario analysis aiming at the straightforward calculation of $x(\omega)$ for various scenarios of ω and the STO optimization approach, the STO model produces one solution that is optimal ("robust") against all possible stochastic scenarios ω [2, 10–16]. The model of this section illustrates typical difficulties in dealing with optimization of continuously differentiable expectation functions $F(x)$ when sample functions $f(x, \omega)$ are nonsmooth.

1. CONNECTIONS AND NEW PROBLEMS OF STATISTICS AND STOCHASTIC OPTIMIZATION

Let us formulate a general stochastic optimization (STO) model, which can be appropriately revised to capture various problems of decision-making under uncertainty, statistics, big data analysis, artificial intelligence (AI): find x minimizing (maximizing) the expectation functional:

$$F(x) = Ef(x, \theta) = \int f(x, \theta) P(x, d\theta), \tag{1}$$

$$x \in X, \tag{2}$$

where the set of feasible solutions X can be a set of scalar quantities, a set of vectors or a set of abstract elements, e.g., a set of probability density functions. In general problems X is described by using similar to (1) expectation functionals. The random "loss" functions $f(x, \theta)$ are often non-smooth and discontinuous functions.

The main complexity of the STO model (1), (2) is that exact evaluation of $F(x)$ is practically impossible. This may be due to various reasons: probability measure $P(x, d\theta)$ is unknown or only partially known, random function $f(x, \theta)$ is analytically intractable, or the evaluation of $F(x)$ is analytically intractable despite well-defined $f(x, \theta)$ and P. All these makes impossible to use standard optimization methods. When exact evaluation of the objective function $F(x)$ (or derivatives) is not possible, the SQG methods enable effective solution of the problems. SQG methods perform a sequential revision of approximate solutions towards the optimal using newly acquired information on the system, obtained via either direct on-line observations or(and) simulations. This feature is especially practical in decision-making problems with decision-dependent exogenous uncertainties and risks, in statistical estimation, ML and AI problems.

Standard statistical problems are formulated as the minimization of the type (1) functionals in the case when the probability measure P is unknown but the sample $\theta^1, \ldots, \theta^N$ of observations drawn randomly according to P is available. It is assumed that P does not depend on X.

Statistics (statistical decision theory) deals with situations in which the model of uncertainty and the optimal solution are defined by unknown sampling model P. The main issue is to recover P by using available samples. In other words, the desirable optimal solutions $x = x^*$ is associated with P (or its parameters), the performance of x^* can be observed from available random data on its performance.

STO models were introduced for decision making problems under uncertainty arising in operation research and systems analysis which are typically described by a large number of decision variables and uncertainties. These models deal with fundamentally different situations. The uncertainty, feasible solutions, and performance of the optimal solution are not given by the sampling model. All of these have to be

characterized from the context of the decision-making situation. As a consequence, multiple performance indicators, constraints, and dependencies among decisions and uncertainties play a key role. Thus, in STO, which in fact arose as an extension of linear and non-linear programming with their sophisticated computation techniques, the accent is on solving problems (1), (2) with large number of decisions variables, random parameters and constraints.

The classical statistics has been developed on the basis of asymptotic analysis requiring large samples of historical data. In this case, additional apriori information on "true" to be estimated parameters of the sampling model can be ignored, and the multidimensional optimization problem (1), (2) can be separated into independent small optimization problems. Consequently, a large place in statistics is occupied by the search for closed form solutions and simple computational procedure. In particular, this is often possible due to simple structure of the loss function $f(x,\theta)$ which is often defined as a quadratic function characterizing a distance between true parameters $x^* = \theta$ and its estimate x.

New important problems in statistics and STO have to confront situations with small data samples, cases of missing observations or absence of direct observations. For these problems, experiments may be dangerous or even simply impossible. These new problems require explicit joint treatment of all relevant interdependent observable, partially observable and non-observable variables by using various prior information in the form of additional constraints describing these interdependencies. These leads to high dimensions. The key issue is the representation of interdependencies enabling to organize pseudo-sampling based on proper characterization of probability measure P by using all available information.

Consequently, these new problems are formulated as general constrained STO problems where estimation of unknown probability measure P is directly associated with goals of overall decision-making problem. Since only specific data are essential for desirable decisions, the combined consideration of statistical estimation within overall decision-making problem can considerably reduce the quality and quantity of estimated information, e.g., the accuracy of the true parameters of P including even requirements on the uniqueness of P. New type of estimation problems arise which can be called as downscaling problems.

Consider some important statistical estimation problems which can be formulated as STO model (1), (2). Instead of asymptotic analysis, this provides the natural criterion of efficiency which can be used to evaluate the convergence to optimal solutions with respect to increasing number of real observations, resampling schemes, and pseudo sampling procedures. This section characterizes also loss functions which are typical for statistics.

1.1. Regression estimation. Assume that a random function $u(v)$ for each element v from a set V corresponds a random element $u(v)$ of the set U. Assume that $V \subset R^l$, $U \subset R^1$. Let P is a joint probability measure defined on pairs $\theta = (u, v)$. The regression function is defined as the conditional mathematical expectation

$$r(v) = E(u \mid v) = \int uP(U \mid v). \tag{3}$$

It is easy to see that $r(v)$ minimizes the functional (providing it is well defined)

$$F(x(v)) = E(u(v) - x(v))^2, \tag{4}$$

where $Eu^2(v) < \infty$, $Ex^2(v) < \infty$.

It follows from the fact that

$$\min_{x(v)} F(x(v)) = E \min_{x} E[(u(v) - x)^2 \mid v],$$

$$\frac{d}{dx} = E[(u(v) - x)^2 \mid v] = -2(E(u \mid v) - x) = 0.$$

The estimation of $r(v)$ is usually considered in the set of functions given in a parametric form $\{r(x,v), x \in X\}$. In this case, the criterion (4) can be rewritten as

$$F(x) = E(u(v) - r(x,v))^2 = E(r(v) - r(x,v))^2 + E(u(v) - r(x,v))^2,$$

i.e., the minimum of $F(x)$ is attained at the function $r(x,v)$ which is close to $r(v)$ in the metric $L_2(P)$ defined as $\sqrt{E(r(v) - r(x,v))^2}$.

1.2. Quantile based regression. The conditional expectation $r(v)$ provides a satisfactory representation of stochastic dependencies $u(v)$ when they are well approximated by two first moments, e.g., for normal distributions. For general (possibly, multimodal) distributions it is more natural to use the median or other quantiles instead of the expectation. Let us define the quantile regression function $r_\rho(v)$ as the maximal value y satisfying equation

$$P(u(v) \geq y \mid v) = \rho(v), \tag{5}$$

where $0 < \rho(v) < 1$. It can be shown that function $r_\rho(v)$ minimizes the functional

$$F(x(v)) = E(\rho(v)x(v) + \max\{0, u(v) - x(v)\}). \tag{6}$$

This is due to the following. First of all, we have

$$\min_{x(v)} F(x(v)) = E \min_{x} E[\rho(v)x + \max\{0, u(v) - x(v) \mid v\}].$$

Assume that probability $P(d\theta)$ has continuous density function $p(\theta)$, $P(d\theta) = p(\theta)d\theta$. Then from the optimality condition for internal stochastic minimax problem follows that optimal solution x satisfies the equation

$$\rho(v) - \int_x^\infty P(d\theta \mid v) = 0, \tag{7}$$

i.e., indeed, it satisfies (5).

Let us note, that the minimization of more general at the first glance functional

$$F(x(v)) = E(a(v)x(v) + \max\{\alpha(v)(u(v) - x(u)), \beta(v)(x(v) - u(v))\})$$

is reduced to the minimization of (6) with $\rho(v) = (a + \beta)(\alpha + \beta)^{-1}$. The median corresponds to the case when $a \equiv 0$, $\alpha = \beta$. The existence of optimal solution requires $a < \alpha$. The literature on the support vector (as Huber [7]) uses

$$F(x(v)) = E \max\{u(v) - x(v) - \varepsilon, x(v) - u(v) + \varepsilon\}.$$

There is interesting effect of ε on the optimality condition (7) and, thus, corresponding modification of the basic idea (5).

The estimation of $r_\rho(v)$ can be considered by a parametric set of functions $\{r_\rho(x,v), x \in X\}$. In this case, the criterion (6) is rewritten as

$$F(x) = E[\rho(v)r(x,v) + \max\{0, u(x) - r(x,v)\}]$$

or equivalently

$$F(x) = E \max[\alpha(v)(u(v) - r(x,v)), \beta(v)(r(x,v) - u(v))],$$

for

$$\alpha(v) > 0,\ \beta(v) > 0,\ \rho(v) = \alpha(v) / (\alpha(v) + \beta(v)).$$

Assuming $r(\cdot, v)$ in (6) is a convex function for all $v \in V$, $F(x)$ is also a convex function. If $r(\cdot, v)$ is a linear function for all $v \in V$, then $F(x)$ can be minimized by linear programming methods.

If $|u(v)| < \text{const}$, $v \in V$, then $|F(x) - F(r_\rho)| < \text{const}\, E|r_\rho(v) - r(x,v)|$, i.e., the minimum of $F(x)$ in (6) is attained at the function $r(x,v)$ which is closed to $r_\rho(v)$ in the metric $L_1(P)$ defined as $E|r_\rho(v) - r(x,v)|$.

1.3. Density estimation. Assume $P(d\theta)$ has a density function $p(\theta)$. Consider an arbitrary density function $x(\theta) \ge 0$, $\int x(\theta)d\theta = 1$ and the expectation functional

$$F(x(\theta)) = E \ln x(\theta) = \int \ln x(\theta)\, p(\theta) d\theta. \tag{7}$$

The maximum of $F(x(\theta))$ (if it exists) is attained at the function $p(\theta)$. The proof follows from the Jensen's inequality, which states that for a concave function Ψ and integrable $\Phi(\theta)$, $E|\Phi(\theta)| < \infty$, $E\Psi(\Phi(\theta)) \le \Psi(E\Phi(\theta))$.

Let $\Psi(\Phi) = \ln \Phi$, $\Phi(\theta) = x(\theta) / p(\theta)$, then

$$F(x(\theta)) - F(p(v)) = \int \ln \frac{x(\theta)}{p(\theta)} p(\theta)\, d\theta \le \ln \int \frac{x(\theta)}{p(\theta)} p(\theta)\, d\theta = \ln 1 = 0.$$

Example 1. Let $\theta^1, \ldots, \theta^N$ be available (not necessarily distinct) observations of θ. The sample mean value of $F(x(v))$ is calculated as

$$F^N(x(\theta)) = \frac{1}{N} \sum_{i=1}^{N} \ln x(\theta^i) = \int P_\delta^N(\theta) \ln x(\theta)\, d\theta, \tag{8}$$

where $p^N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta(\theta - \theta^i)$ is the empirical density defined by the Dirac function $\delta(\cdot)$.

The maximization of $F^N(x(\theta))$ with respect to feasible density functions $x(\theta) \ge 0$, $\int x(\theta)d\theta = 1$ yields the empirical density function

$$x^N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta(\theta - \theta^i). \tag{9}$$

2. PROBLEMS WITH SMALL SAMPLES AND HIGH DIMENSIONS OF θ

Standard statistical estimation and STO approaches for solving (1), (2) are based on the ability to obtain observations according to probability measure P. In fact, the justification of these methods, e.g., their consistency (convergence) and efficiency, rely on asymptotic analysis requiring infinite number of observations. For new problems, in particular, arising in the study of global change processes, we often have large number of unknown interdependent variables θ, x and only very restricted samples of real observations.

Experiments to generate new real observations may be extremely expensive, dangerous or simply impossible. The natural approach for dealing with new problems can be based on using all additional information on P. The main issue is not to obtain a good estimation of P, but to achieve a robust solution of (1), (2). A key problem is the design of pseudosampling procedures enabling to generate samples of closed to reality observations ensuring the robustness (in a sense) of the solution. High dimensions and general cases of random functions $f(x,\theta)$ in (1) will require often rather sophisticated STO computational methods. Given observations $\theta^1, \ldots, \theta^N$ of vector θ, function $F(x(\cdot))$ in (1) is estimated often by using the empirical density

$$p^N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta(\theta - \theta^i), \tag{10}$$

i.e., as the sample mean functional

$$F^N(x) = \int f(x,\theta)\, p_N(\theta)\, d\theta = \frac{1}{N} \sum_{i=1}^{N} f(x,\theta^i). \tag{11}$$

Here we use notation x instead of $x(\theta)$ for the sake of simplicity, i.e., instead of vector-function $x(\theta)$ we view $x(\theta)$ as a vector $x \in R^n$.

Besides the sample $\theta^1, \ldots, \theta^N$ of size N, there often exist additional information on unknown P. Let us denote by Ξ the set of admissible distributions consistent with available information. The robust solution can be defined as $x \in X$ which minimizes

$$F(x) = \max_{P \in \Xi} \int f(x, \theta) P(d\theta) = \int f(x, \theta) P(x, d\theta) \ , \tag{12}$$

where $P(x, d\theta)$ denotes the worst-case distribution. Thus, we have to deal with a general case of STO (1), (2) where probability measure is, in general, analytically intractable and it is affected by decision x.

The empirical density function $p_N(\theta)$ may not belong to Ξ, therefore, $p_N(\theta)$ and estimator (11) of function $F(x)$ have to be modified according to (12). In particular, model (12) can be applied for the estimation of density function $x(\theta)$ by maximizing function (8) for $x(\theta) \in \Xi$. If $p_N(\theta) \notin \Xi$, this would not lead to the empirical density $p_N(\theta)$. In other words, the empirical distribution function $p_N(\theta)$ which is constructed only on the basis of empirical observations may not belong to Ξ.

The representation of available information on $P(d\theta)$ plays the key role in cases of small sample size N. Besides the sample, class Ξ may include information on some quantiles of $P(d\theta)$, its shape, marginal distributions, and moments. The representation of Ξ by given marginal distributions plays an important role in multidimensional samples of θ. In this case, it is possible to include any components but maintain fixed the one dimensional marginal. Although there is no general systematic procedure to elicit a class Ξ, and the procedure will depend on the particular application, commonly used ways of eliciting such a class are the following. We can derive from the sample straightforward nonparametric empirical density function $p_N(\theta)$ defined by (10). It says that $P(d\theta)$ is concentrated at observed points with equal probability $1/N$. Density $p_N(\theta)$ can be used to derive new observations. Although any other observation according to $p_N(\theta)$ is a repetition of the already observed points $\theta^1, \ldots, \theta^N$, yet, this is exactly the main idea of the bootstrapping. Clearly, it would be preferable to use pseudo-sampling schemes based on better nonparametric density estimators. The sample $\theta^1, \ldots, \theta^N$ may have certain tendencies, e.g., clustering around some regions in admissible set Θ, $\theta \in \Theta \subset R^l$. In the case of low dimensions (small l) this tendency can be utilized by a histogram. This nonparametric density estimation function provides the possibility to sample new plausible observations within "windows" of the histogram.

The use of histograms with fixed mesh in the case of high dimensions leads to "exponential explosion" of computations (say, only 10 intervals per dimension lead to 10^l equispaces Cartesian mesh points). Since the available sample $\theta^1, \ldots, \theta^N$ tends to come from regions where the density is relatively high, the main idea of better nonparametric estimators is to use observations $\theta^1, \ldots, \theta^N$ as anchor points from which the fine structure of density can be examined further. In fact, for solving problem (1), (2) the philosophy of nonparametric density estimation can be used only as guidance without actually precise estimation. However, nonparametric density estimation uses only information which are contained in the sample. The key task is to combine this information with other available information.

Assume that in accordance with sample $\theta^1, \ldots, \theta^N$, and our beliefs we can subdivide the set Θ into a collection of sets $\{C_k, k = 1, \ldots, K\}$. Some of them may correspond to clusters of available observations whereas others may reflect experts opinions on the degree of uncertainty and its heterogeneity across the admissible set Θ;

for instance, we can distinguish some critical zones ("catalogues of earthquakes") which may cause significant losses $f(x,\theta)$ although with high degree of uncertainty. Accordingly, the additional beliefs can be given in terms of a "quantile" class

$$\Xi = \left\{ P : \int_{C_s} P(d\theta) = \alpha_s, s = 1, \dots, S \right\}, \tag{13}$$

where $\sum_{s=1}^{S} \alpha_s = 1$; more generally — in terms of ranges of probabilities

$$\Xi = \left\{ P : \alpha_s \le \int_{C_s} P(d\theta) \le \beta_s, s = 1, \dots, S \right\}, \tag{14}$$

where α_s, β_s are given numbers such that $\sum_{s=1}^{S} \alpha_s \le 1 \le \sum_{s=1}^{S} \beta_s$. This class is considered as "the most natural elicitation mechanism. It is important that both classes (13), (14) for problem (12) lead to the type (11) mean value functions

$$F(x) = \sum_{s=1}^{S} \gamma_s f(x, \hat{\theta}^s), \ \sum_{s=1}^{S} \gamma_s = 1, \ \gamma_s > 0,$$

where $\hat{\theta}^s$ belongs to C_s. Namely:

Proposition 1. For any function $f(x,\theta)$ assumed to be integrable w.r.t. all P in Ξ defined by (13):

$$\max_{P \in \Xi} \int f(x,\theta) P(d\theta) = \sum_{s=1}^{S} \alpha_s f(x, \hat{\theta}^s), \tag{15}$$

where

$$f(x, \hat{\theta}^s) = \max_{\theta \in C_s} f(x,\theta), \ s = 1, \dots, S. \tag{16}$$

In the case of Ξ defined by (14):

$$\max_{P \in \Xi} \int f(x,\theta) P(d\theta) = \sum_{s=1}^{S} \hat{\gamma}_s f(x, \hat{\theta}^s), \tag{17}$$

where $\hat{\theta}^s$ is defined by (16) and $\hat{\gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_s)$ is such that

$$\hat{\gamma} = \arg\max \left\{ \sum_{s=1}^{S} \gamma_s f(x, \hat{\theta}^s | \alpha_s \le \gamma_s \le \beta_s, \sum_{s=1}^{S} \gamma_s = 1 \right\}. \tag{18}$$

Corollary 1. Formula (15) generalizes the standard mean value estimator (12). Let $\{\theta^1, \dots, \theta^S\}$ be distinct observations of Θ and N_s the number of times θ^s has been observed in C_s Let $\alpha_s = N_s / N$, where N is the sample size. Then from (15) follows that

$$\max_{P \in \Xi} \int f(x,\theta) P(d\theta) = \sum_{s=1}^{S} \frac{N_s}{N} f(x, \theta^s), \tag{19}$$

assuming that function $f(x,\cdot)$ is constant in C_s.

Remark 1. Subproblem (16), (17) have explicit simple solutions for loss functions defined in the support vector literature.

Assume $P(d\theta) = p(\theta)d\theta$, $p(\theta) = \int K(y,\theta) Q(dy)$, where $Q(dy)$ satisfies quantile constraints (13) or in terms of ranges of probability measure (14). If $Q(dy)$ satisfies (13), then

$$\max_{P \in \Xi} \int f(x,\theta) P(d\theta) = \int f(x,\theta) \left[\sum_{s=1}^{S} \alpha_s K(\hat{y}^s, \theta) \right] d\theta, \tag{20}$$

where $\hat{y}^s$ is defined similar to (16):

$$\hat{y}^s = \arg\max \{ K(y,\theta), \ y \in C_s \}, \ s = 1, \dots, S. \tag{21}$$

Remark 2. From (21) follows that within approach defined by (12) the kernel-type estimators for the density function arise from assumptions on its shape characterized by $K(y,\theta)$, $Q(dy)$. (This is a natural way to characterize, e.g., seismicity — the Guettenberg–Richter law; distributions of claims in the insurance loss distributions; floods severities — mixed Poison law and so on.) If function $K(\cdot,\theta)$ is constant, we have standard kernel estimates.

There can be various extensions of this approach to cases in which we have additional constraints on moments, monotonicity constraints on $P(d\theta)$, marginal distributions. If

$$\alpha_s = \int\limits_{C_s} K_s(\theta^s,\theta)\,d\theta,$$

where $K(\theta^s,\theta)$, characterizes similarity between θ^s, and θ which can be viewed as a probability density, then

$$\max_{P\in\Xi}\int f(x,\theta)P(d\theta) = \sum_{s=1}^{S} f(x,\hat{\theta}^s)\int\limits_{C_s} K(\theta^s,\theta)\,d\theta. \tag{21}$$

3. SOLUTION PROCEDURES: SQG METHODS FOR ROBUST MACHINE LEARNING

Minimization of (15)–(17) correspond to deterministic minimax problems which can be solved by linear or nonlinear programming methods. These problems are very simple in the case of loss functions $f(x,\theta)$, which are typical for the support vector applications. In more general cases such as (20), the problem is a standard type STO problem (1), (2) with known probability distribution $P(x,d\theta)$. This problem can be solved by STO methods. In particular, we can use pseudo-sampling in addition to the real sample $\theta^1,\dots,\theta^N$. Stochastic SQG methods allow to overcome the complexity arising in cases when $P(x,d\theta)$ is analytically intractable. The combination of pseudo-sampling with SQG can be viewed as the generalization of the bootstrapping to STO procedures. We have to remember that often the main issue is not to predict estimation of P but to find a robust solution of problem (1), (2) what requires only similar to reality pseudo-samples.

3.1. Machine learning. SQG methods can be effectively used for Machine Learning (ML) problems. ML approaches (also utilizing hardly interpretable neural networks) can be combined with disciplinary or interdisciplinary decision-making models, e.g., agricultural, environmental, energy, etc., for decision-making in the conditions of uncertainty, increasing interdependencies and complex analytically intractable systemic (“unknown risks”). ML algorithms are supposed to find natural patterns in data that generate insight and help make better decisions and predictions. In many cases, ML problems utilize Deep Neural Networks characterized by transfer functions, etc. Neural networks are considered to be universal “predictors” (“approximators”, “estimators”). ML grounds on the principles of statistical estimation and learning theory [17, 18]. Robust ML in the presence of outliers is similar to robust statistical learning. Robust ML derives a model which does not deteriorate too much when training and testing with slightly different data (either by adding observations/noise or by taking other dataset). In the presence of potential extreme events, the prediction model has to be able not to ignore, but to predict extreme outcomes. The use of quantile (percentile)-based “loss” or “goodness of fit” functions for ML learning opens up a possibility of using the two-stage STO models [19].

In the traditional statistics theory, one has to know almost everything, including the number of parameters and the types of dependency in order to recover the dependency. Parameter estimation problem is considered to be the dependency

estimation problem. According to the learning theory, it was shown that in order to estimate the dependency from the data, it is sufficient to know some general properties of the set of functions to which the unknown dependencies belong [17]. Machine learning (of a perceptron, a network of neurons, a decision tree, and other models) is being used in patters and image recognition, the type of the model does not affect the general machine learning principles. Major achievement in the machine learning is the implementation in 1986 of the so called back-propagation methods for finding the weights of neurons. The method is based on traditional gradient decent procedures.

3.2. Numerical optimization algorithms for machine learning. Numerical optimization has played an important role in the evolution of ML. For example, the gradient decent procedures, the stochastic gradient decent and their modifications and extensions (stochastic average gradient, stochastic dual coordinate ascent, Nesterov's accelerated gradient, stochastic variance reduced gradient and other) are possible approaches to the solution of large-scale machine learning problems [20, 21]. Many algorithms are based on the results of Robins and Monro [22], Nesterov [20], Polyak and Judinsky [23] for smooth optimization.

However, nonsmooth quantile-based loss, activation, and regularization functions are more typical characteristics of many practical big data and ML problems. In neural networks training, multiple layers and nonsmooth activation functions depend in terms of absolute value, maximum or minimum operations, further complicate the problem. Nonsmooth functions have abrupt bends and even discontinuities requiring methods based on subdifferential and stochastic quasigradient calculus [4–7, 24–27]. In these cases, the SQG methods make use of generalized directional derivatives, subgradients and stochastic quasigradients. The convergence of the SQG methods is proven in rather general cases (see discussion in [1, 3–7, 28], references therein).

3.3. SQG methods for machine learning. The convergence of the SQG method follows from the results of the theory of nonconvex nonsmooth stochastic optimization [1, 3–6, 26, 27]. For highly nonconvex models such as deep neural networks, the SQG methods allow to avoid local solutions. In cases of nonstationary data, the SQGs allow for sequential revisions and adaptation of parameters to the changing environment, possibly, based on offline adaptive simulations.

Let us formulate a problem ML of a neural network. It is necessary to define the structure of a net and train it, i.e., estimate the weights/parameters of the network model using the training set for which the solution of the problem is known, i.e., on the available input-output data. The training task is formulated as the problem of minimizing a loss function (empirical risk), which measures the forecast error of the model. The root mean squared error (RMSE) predicts only an average response. For extreme events/scenarios, e.g., crop yield shocks, production and market risks, catastrophe (structural) losses, it is more appropriate to use a quantile-based loss (QL) function [19, 29]. As discussed in the Introduction, the QL allows to identify values higher or lower than the specified quantile (or critical) value of the data in the training set and penalize the positive and negative deviations from the value differently depending on the problem at hand. For example, over- and underestimation can incur different costs associated with over or undersupply of a specific product or resource. As an example, assume, that a hydropower station is undersupplied with water, which has been diverted to be used in other sectors or activities.

The training of a neural net is achieved through the minimization of the loss function $F(x)$ with respect to the net parameters x,

$$F(x) = \sum_{i=1}^{N} F(i, x), \tag{22}$$

where $F(i, x)$ are (nested) neuron-type functions. Each function $F(i, x)$ corresponds to one training object. At each step $k = 0, 1, \ldots$, an object $i = i(k)$ is picked up at random with probability $\mu > 0$ among N alternatives. Starting from the initial parameter x^0, the vector x^k of parameters x is adjusted in the direction opposite to the (sub)gradient

$$\xi(k) = 1 / \mu(i(k)) F_x(i(k), x^k) \tag{23}$$

of the function $F(i, k)$. Let us note that if function f is differentiable at x, then its gradient at x is its subgradient. If function f is nonsmooth at x, there exist a set of subgradients of function f in x, and the function is called subdifferentiable at x. The SQG solution procedures can be applied to the class of generalized differentiable (GD) functions, which include continuously differentiable, convex, concave, weakly convex and concave, semismooth and semiconvex functions. It is easy to see that $[\xi(k) \mid x^k] = F_x(x^k)$. The index $i(k)$ changes from iteration to iteration in order to cover more or less uniformly the set of indices $1, \ldots, m$. Calculation of $\xi(k)$, and in particular, $F_x(i(k), x^k)$ for different cases, including nonsmooth and nonconvex functions F and discrete event systems is discussed, for example, in [1, 3–7, 19] and references therein.

CONCLUSIONS

The paper discusses the connections between the problems of robust decision making, statistical estimation, and machine learning. Robust decision making in the presence of possible extreme events and uncertainties relies on quantile-based goal functions, constraints, and performance indicators. This approach produces solutions, which enable stability and resilience of relevant systems irrespectively of what uncertainty scenario (extreme event) occurs (i.e., the systems become less or even insensitive to the scenarios). Robust statistical estimation also produces solutions, which are not deteriorated by inclusion of additional observations especially from non-normal and possibly heavily skewed distributions. The discussed in this paper quantile-based regression can be considered as a nonparametric estimation method as it does not require specific assumptions about the probability distribution of observation (data) errors. Machine learning grounds on the principles of statistical learning. Using quantile-based “goodness-of-fit” criteria for machine learning derives (a) model(s) which do(es) not deteriorate too much when training and testing with slightly different or additional data. The use of nonsmooth, possibly discontinuous quantile (percentile)-based criteria, “loss” or “goodness of fit” functions for robust decision-making, statistical estimation and machine learning opens-up a possibility of using the two-stage STO models. We illustrated the application the stochastic quasigradient (SQG) methods for the robust decision making and machine learning problems. For general type nonsmooth, possibly discontinuous and nonconvex problems, including for example neural networks training, the discussion about the application and convergence of the stochastic quasigradient (SQG) methods is available in [3–7]. For the case of a general endogenous reinforced learning, the convergence of the SQG procedure was proven in [8] based on the results of nondifferentiable optimization providing a new type of machine learning algorithms solving the problem of distributed decentralized models’ linkage under asymmetric information and uncertainty.

REFERENCES

1. Ermoliev Yu., Wets R.J.-B. (Eds). Numerical Techniques for Stochastic Optimization. Heidelberg: Springer Verlag, 1988. 573 p.
2. Ermoliev Y., Hordijk L. Global changes: Facets of robust decisions. In: Coping with Uncertainty: Modeling and Policy Issue. Marti K., Ermoliev Y., Makowski M., Pflug G. (Eds.). Berlin: Springer Verlag, 2003. P. 4–28.
3. Ermoliev Y. Methods of Stochastic Programming. Moscow: Nauka, 1976. 340 p. (In Russian).
4. Ermoliev Y. Stochastic quasigradient methods. In: Encyclopedia of Optimization. Pardalos P.M. (Ed.). New York: Springer Verlag, 2009. P. 3801–3807.
5. Ermoliev Y. Two-stage stochastic programming: Quasigradient method. In: Encyclopedia of Optimization. Pardalos P.M. (Ed.). New York: Springer Verlag, 2009. P. 3955–3959.
6. Ermoliev Y. Stochastic quasigradient methods in minimax problems. In: Encyclopedia of Optimization. Pardalos P.M. (Ed.). New York: Springer Verlag, 2009. P. 3813–3818.
7. Ermoliev Y., Gaivoronski A. Stochastic quasigradient methods for optimization of discrete event systems. *Annals of Operation Research.* 1992. Vol. 39, Iss. 1. P. 1–39.
8. Ermoliev Y., Zagorodny A.G., Bogdanov V.L., Ermolieva T., Havlik P., Rovenskaya E., Komendantova N., Obersteiner M. Robust food–energy–water–environmental security management: stochastic quasigradient procedure for linkage of distributed optimization models under asymmetric information and uncertainty. *Cybernetics and Systems Analysis*. 2022. Vol. 58, N 1. P. 45–57. https://doi.org/10.1007/s10559-022-00434-5.
9. Ermoliev Y., von Winterfeldt D. Systemic risk and security management. In: Managing Safety of Heterogeneous Systems. Ermoliev Y., Makowski M., Marti K. (Eds.). *Lecture Notes in Economics and Mathematical Systems*. 2012. Vol. 658. P. 19–49.
10. Ermolieva T., Havlik, P., Fran, S., Kahi, T., Balkovi, J., Skalský R., Ermoliev Y., Knopov P.S., Borodina O.M., Gorbachuk V.M. A risk-informed decision-making framework for climate change adaptation through robust land use and irrigation planning. *Sustainability*. 2022. Vol. 14, Iss. 3. P. 1430. https://doi.org/10.3390/su14031430.
11. Ermolieva T., Havlík P., Ermoliev Y., Mosnier A., Obersteiner M., Leclere D., Khabarov N., Valin H., Reuter W. Integrated management of land use systems under systemic risks and security targets: A stochastic global biosphere management model. *Journal of Agricultural Economics*. 2016. Vol. 67, Iss. 3. P. 584–601.
12. Borodina O.M., Kyryziuk S.V., Fraier O.V., Ermoliev Y.M., Ermolieva T.Y., Knopov P.S., Horbachuk V.M. Mathematical modeling of agricultural crop diversification in ukraine: scientific approaches and empirical results. *Cybernetics and Systems Analysis*. 2020. Vol. 56, N 2. P. 213–222. https://doi.org/10.1007/s10559-020-00237-6.
13. Gao J., Xu X., Cao G.-Y., Ermoliev Y., Ermolieva T., Rovenskaya E. Strategic decision-support modeling for robust management of the food–energy–water nexus under uncertainty. *Journal of Cleaner Production*. 2021. Vol. 292. 125995. https://doi.org/10.1016/j.jclepro.2021.125995.
14. Ortiz-Partida J.P., Kahil T., Ermolieva T., Ermoliev Y., Lane B., Sandoval-Solis S., Wada Y. A two-stage stochastic optimization for robust operation of multipurpose reservoirs. *Water Resources Management.* 2019. Vol. 33, Iss. 11. P. 3815–3830. https://doi.org/10.1007/s11269-019-02337-1.
15. Ren M., Xu X., Ermolieva T., Cao G.-Y., Yermoliev Y. The optimal technological development path to reduce pollution and restructure iron and steel industry for sustainable transition. *International Journal of Science and Engineering Investigations*. 2018. Vol. 7, Iss. 73. P. 100–105.
16. Huber P. Robust Statistics. New York; Chichester; Brisbane; Toronto; Singapore: John Wiley & Sons, 1981. 317 p.
17. Vapnik V.N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995. 334 p.
18. Knopov P.S., Kasitskaya E.J. Empirical Estimates in Stochastic Optimization and Identification. Berlin: Springer Verlag, 2002. 250 p.

19. Ermolieva T., Ermoliev Y., Obersteiner M., Rovenskaya E. Two-stage nonsmooth stochastic optimization and iterative stochastic quasigradient procedure for robust estimation, machine learning and decision making. In: Resilience in the Digital Age. Ch. 4. Roberts F.S., Sheremet I.A. (Eds.). Cham: Springer, 2021. P. 45–74.
20. Nesterov Y. Introductory Lectures on Convex Optimization. New York: Springer New York, 2004. XVIII, 236 p.
21. Reddi S.J., Hefny A., Sra S., Poczos B., Smola A.J. On variance reduction in stochastic gradient descent and its asynchronous variants. In: Advances in Neural Information Processing Systems 28. Cortes C., Lawrence N.D., Lee D.D., Sugiyama M., Garnett R. (Eds.). *Proc. Annual Conference on Neural Information Processing Systems 2015* (7–12 December 2015, Montreal, Canada). Montreal, 2015. P. 2647–2655.
22. Robbins H., Monro S. A stochastic approximation method. *The Annals of Mathematical Statistics*. 1951. Vol. 22, Iss. 3. P. 400–407.
23. Polyak B.T., Juditsk A.B. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization.* 1992. Vol. 30, Iss. 4. P. 838–855.
24. Clarke F.H. Optimization and Nonsmooth Analysis. New York: John Wiley & Sons, 1983. XIII, 308 p.
25. Rockafeller T. The Theory of Subgradient and its Application to Problems of Optimization: Convex and Nonconvex Functions. Berlin: Heldermann Verlag, 1981. 107 p.
26. Ermoliev Y., Shor N. On minimization of nondiferentiable functions. *Kibernetika*. 1967. Vol. 3, N 1. P. 101–102.
27. Ermoliev Y., Norkin V. On nonsmooth and discontinuous problems of stochastic systems optimization. *Europ. J. Oper. Res*. 1997. Vol. 101, Iss. 2. P. 230–244.
28. Gaivoronski A. Convergence properties of backpropagation for neural nets via theory of stochastic quasigradient methods: Part 1. *Optimization Methods and Software*. 1994. Vol. 4, Iss. 2. P. 117–134.
29. Gorbachuk V.M., Ermoliev Y., Ermolieva T., Dunajevskij M.S. Quantile-based regression for the assessment of economic and ecological risks. *Proc. 5th International Scientific Conference on Computational Intelligence* (15–20 April 2019, Uzgorod, Ukraine). Uzgorod, 2019. P. 188–190.

Т. Єрмольєва, Ю. Єрмольєв, П. Гавлик, А. Лесса-Дерсі-Аугустинчик, Н. Комендантова, Т. Кахіл, Дж. Балкович, Р. Скальські, К. Фолберт, П.С. Кнопов, Г. Вонг (Г. Ванг)

ЗВ'ЯЗКИ МІЖ СТІЙКОЮ СТАТИСТИЧНОЮ ОЦІНКОЮ, НАДІЙНИМ ПРИЙНЯТТЯМ РІШЕНЬ ІЗ ДВОЕТАПНОЮ СТОХАСТИЧНОЮ ОПТИМІЗАЦІЄЮ ТА НАДІЙНИМИ ПРОБЛЕМАМИ МАШИННОГО НАВЧАННЯ

Анотація. Розглянуто зв'язки між задачами двоетапного стохастичого програмування, проблемами визначення робастних рішень, робастними методами у статистиці та машинному навчанні. В умовах невизначеності, а також можливого настання екстремальних подій та ситуацій, ці задачі потребують розгляду та оптимізації систем з квантильними критеріями, обмеженнями та індикаторами якості результатів (функціями збитків). Задачі двоетапної стохастичної оптимізації можна ефективно розв'язати ітеративними методами стохастичних квазіградієнтів (SQG). Методи SQG дають змогу розв'язувати негладкі, можливо розривні та неопуклі задачі машинного навчання, наприклад, задачі квантильної регресії та навчання нейронної мережі. Такі поняття, як допустимі розв'язки, оптимальність та робастність у загальних задачах прийняття рішень визначаються конкретною ситуацією прийняття рішень. Задачі робастного статистичного оцінювання та машинного навчання можна інтегрувати у задачі планування дисциплінарних та міждисциплінарних систем, як-от: систем землекористування, сільськогосподарських, енергетичних, тих, що слугують для підтримки прийняття робастних рішень в умовах невизначеностей, зростаючих системних залежностей та невідомих ризиків.

Ключові слова: двоетапна задача стохастичної оптимізації, робастне прийняття рішень та статистичне оцінювання, робастна квантильна регресія, машинне навчання, загальні проблеми прийняття робастних рішень, системні ризики, невизначеності.

Надійшла до редакції 26.12.2022