IIASA

International Institute for Applied Systems Analysis • A-2361 Laxenburg • Austria
Tel: +43 2236 807 • Fax: +43 2236 71313 • E-mail: info@iiasa.ac.at • Web: www.iiasa.ac.at

INTERIM REPORT IR-98-009 / March

Monte Carlo Optimization and Path Dependent Nonstationary Laws of Large Numbers

Yuri M. Ermoliev (ermoliev@iiasa.ac.at)
Vladimir I. Norkin (norkin@umc.kiev.ua)

Approved by
GordonMacDonald (macdon@iiasa.ac.at)
Director, *IIASA*

Abstract

New types of laws of large numbers are derived by using connections between estimation and stochastic optimization problems. They enable one to "track" time-and-path dependent functionals by using, in general, nonlinear estimators. Proofs are based on the new stochastic version of the second Lyapunov's method. Applications to adaptive Monte-Carlo optimization, stochastic branch and bounds method and minimization of risk functions are discussed.

Contents

Monte Carlo Optimization and Path Dependent Nonstationary Laws of Large Numbers

Yuri M. Ermoliev (ermoliev@iiasa.ac.at)
Vladimir I. Norkin (norkin@umc.kiev.ua)

1 Introduction

A rather general stochastic optimization (STO) problem may be regarded as estimating the minimal value F^* of the integral

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

and a corresponding optimal parameters (solution) x^* from a subset X of Euclidean space R^n. The Monte Carlo optimization (MCO) problem arises when the probability measure $\mathbf{P}$ and(or) the sample function f are not known explicitly but only in terms of other explicitly known measures and functions. The estimation of F^*, x^* can be viewed then as a Monte Carlo simulation procedure with adaptive adjustments of parameters $x \in X$. A typical example of MCO problem arises in catastrophic risk management [4, 5], where the sample performance function $f(x,\theta)$ and the probability distribution of θ are defined implicitly through complex dynamic interactions between spatial patterns of catastrophes, decisions and damages. The estimation of F^*, x^* is a significant generalization of the standard Monte Carlo estimation problem which corresponds to the case when x^* is known. A stochastic optimization procedure produces estimates $x^k,\ k = 1,2,\ldots,$ of optimal solution x^* by using samples θ^k from distribution $\mathbf{P}(x^k,\cdot)$. An important question is whether

$$F^k = k^{-1}\sum_{s=1}^{k} f(x^s,\theta^s) \longrightarrow F^*, \quad k \longrightarrow \infty, \tag{2}$$

with probability 1, i.e. whether F^* can be estimated from available path dependent observations $\zeta^s = f(x^s,\theta^s),\ s = 1,2,\ldots$. Another important question is whether values F^k and $F_x^k = k^{-1}\sum_{s=1}^{k} f_x(x^s,\theta^s)$ can be used as estimates of the current value $F(x^k)$ and its gradient $F_x(x^k)$ (or a subgradient in the case of nonsmooth function F), i.e. with probability 1

$$\lim_{k\to\infty}(F^k - F(x^k)) = 0, \quad \lim_{k\to\infty}(F_x^k - F_x(x^k)) = 0, \tag{3}$$

assuming that values $f_x(x^s,\theta^s)$ are known. The convergence in (2), (3) can be derived easily (see Theorem 1.2 and Example 3.1) if one knows that x^k converges with probability 1 to the set of optimal solutions X^*.

Unfortunately, the convergence $x^k \longrightarrow X^*$ itself is often derived only from convergence (3), i.e. when estimates $F^k,\ F_x^k$ track path dependent values $F(x^k),\ F_x(x^k), k \longrightarrow \infty$ without assumption on the convergence of $\{x^k\}$. An important example of such a situation occurs in so-called adaptive Monte Carlo optimization (AMCO) as it was discussed in

[4] for catastrophic risk management problems. In this case the direct sampling of "low probability – high consequences" events θ^k from distributions $\mathbf{P}(x^k,\cdot)$ may be time consuming and the prposed AMCO procedure makes use of the information in the sample as it is collected to sequentially improve the efficiency of the sampling procedure itself jointly with the adjustment of x^k. For this purpose the probability measure $\mathbf{P}$ at each step k is modified by chosing it from a family of distributions $\mathbf{P}(x^k,y,\cdot)$ indexed by a vector y. A value $y=y^k$ specifies the distribution $\mathbf{P}(x^k,y^k,\cdot)$ from which θ^k is drown at step k. At each step k y^k is adjusted towards increasing of an efficiency criteria $\psi(x^k,y^k)$ of the estimate x^k. The feasibility of such approach essentially depends on the ability to estimate the value $\psi(x^k,y^k)$ and the gradient $\psi_y(x^k,y^k)$ as in (3) despite changes x^k, $k=0,1,...$ in parameter x.

The assertion (2) sometimes can be derived from the following known results.

Theorem 1.1 *(see, for instance, [7]). Let $\{\zeta^k,\ k=1,2,\ldots\}$ be a sequence of random variables (possibly vector valued) with finite first moments $z^k=\mathbf{E}\zeta^k$ and*

$$\sum_{k=1}^{\infty}\frac{1}{k}\mathbf{E}\|\zeta^k-z^k\|<+\infty. \tag{4}$$

Then with probability one

$$\lim_{k\to\infty}\frac{1}{k}\sum_{s=1}^{k}(\zeta^s-z^s)=0.$$

Let us notice that (4) implies $\liminf_k \mathbf{E}\|\zeta^k-z^k\|=0$, i.e. the accuracy of observations ζ^k must increase with increase of k.

Theorem 1.2 *(see Neveu [13] and [9], [18]). Let $\mathcal{F}_k$ be a flow of nondecreasing σ-algebras, random variables ζ^k are $\mathcal{F}_k$-measurable, nonnegative numbers $n_k\uparrow\infty$, $k=1,2,\ldots$. Denote $z^k=\mathbf{E}\{\zeta^k|\mathcal{F}_{k-1}\}$ conditional mean value of ζ^k with respect to $\mathcal{F}_{k-1}$. Suppose that $\mathbf{E}\|\zeta^k-z^k\|<\infty$ and*

$$\sum_{k=1}^{\infty}\frac{1}{n_k^2}\mathbf{E}\{(\zeta^k-z^k)^2|\mathcal{F}_{k-1}\}<\infty\quad a.s.$$

Then with probability 1

$$\lim_{k}\frac{1}{n_k}\sum_{s=1}^{k}(\zeta^s-z^s)=0.$$

Theorems 1.1 and 1.2 (with $n_k=k$), in particular, state that if z^k a.s. converge to a random limit z^*, then with probability 1

$$\lim_{k}\frac{1}{k}\sum_{s=1}^{k}\zeta^s=z^*.$$

The aim of this paper is to develop a framework enabling one to derive assertions of type (3) and (2) for more general estimators than arithmetic mean, what is critically important for the design of Monte Carlo optimization procedures (see Examples 3.3-3.5). These results have their origin in the theory of stochastic optimization [3], [6] (see p.177). We use the fact that the estimation of expectation $z^*=\mathbf{E}\zeta$ can be viewed as a Monte Carlo optimization of the integral $H(z)=\int\|z-\zeta\|^2\mathbf{P}(d\zeta)$ with unknown probability measure $\mathbf{P}$ and easily calculated stochastic estimate $2(z-\zeta)$ of the gradient H_z. The

estimation of a nonstationary expectation $z^k = \mathbf{E}\zeta^k$, $k = 1, 2, \ldots$, can similarly be viewed as the minimization of the nonstationary function $H_k(z) = \int \|z - \zeta^k\|^2 \mathbf{P}(d\zeta^k)$. If $\{\zeta^k\}$ are uniformly bounded then convergence properties (2), (3) can be derived from general results on nonstationary optimization (see references in [6], p.152). The main idea of this article resembles this approach, but we derive assertions (2), (3) and more general ones from a special (nonstationary) form of the law of large number (LLN). To prove this form of LLN we use a new rather general stochastic version of the second Lyapunov's method (Theorem 2.3). Theorem 2.3 is important itself for the convergence analysis of various adaptive Monte Carlo procedures. In the paper we assume that involved random variables are integrable in power $(1 + \epsilon)$ $(0 < \epsilon \leq 1)$.

We give two versions (Theorems 2.1, 2.2) of a nonstationary law of large numbers for dependent random variables, in particular, analogs of Theorems 1.1, 1.2, and with more general (in contrast to arithmetic mean) rules for averaging of random variables. Theorems 2.1, 2.2 also utilize additional information on z^k that z^k belongs to a convex set $Z \subset R^n$. Theorem 2.2 shows how to track moving means $z^k = \mathbf{E}\{\zeta^k | \mathcal{F}_{k-1}\}$ by using only observations ζ^k in the case when z^k does not converge to any limit.

It opens up a way to use different estimates of F^*, $F(x^N)$, $F_x(x^N)$, not only arithmetic means as in (2), (3). In Section 3 we discuss various applications of Theorems 2.1, 2.2 to Monte Carlo optimization problems, estimation problems, adaptive Monte Carlo method, stochastic branch and bound procedures, minimization of risk functions. The proofs of all results are given in the last Section 4.

2 Nonstationary Laws of Large Numbers

Let $(\Omega, \Sigma, \mathbf{P})$ be a probability space with a flow of nondecreasing σ-algebras $\mathcal{F}_k \subseteq \mathcal{F}_{k+1} \subseteq \Sigma$, $k = 1, 2, \ldots$. Let random variables $\zeta^k(\omega) : \Omega \longrightarrow R^n$ are measurable with respect to $\mathcal{F}_k$, $k = 1, 2, \ldots$. Denote (changing with k, i.e. nonstationary) conditional mathematical expectations

$$z^k(\omega) = \mathbf{E}\{\zeta^k(\omega) | \, \mathcal{F}_{k-1}\}. \tag{5}$$

In particular, one can take $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_k = \sigma\{\zeta^1, \ldots, \zeta^k\}$ and $z^k(\omega) = \mathbf{E}\{\zeta^k(\omega) | \, \zeta^1, \ldots, \zeta^{k-1}\}$, $k > 1$. Consider the following estimators ($\overline{\zeta}^1 = 0$):

$$\overline{\zeta}^{k+1}(\omega) = \Pi_Z \left(\overline{\zeta}^k(\omega) - \sigma_k(\overline{\zeta}^k(\omega) - \zeta^k(\omega), \right), \quad k = 1, 2, \ldots, \tag{6}$$

and an auxiliary sequence ($\overline{z}^1 = 0$):

$$\overline{z}^{k+1}(\omega) = \Pi_Z \left(\overline{z}^k(\omega) - \sigma_k(\overline{z}^k(\omega) - z^k(\omega) \right), \quad k = 1, 2, \ldots, \tag{7}$$

where Z is a convex set from R^n, Π_Z is the (orthogonal) projection operator on the set Z, random variables σ_k are $\mathcal{F}_{k-1}$-measurable and satisfy conditions:

$$0 \leq \sigma_k \leq 1, \quad \lim_k \sigma_k = 0, \quad \sum_{k=1}^{+\infty} \sigma_k = +\infty \quad \text{a.s.}; \tag{8}$$

$$\sum_{k=1}^{\infty} \mathbf{E}\{\sigma_k^{1+\epsilon} \|\zeta^k(\omega) - z^k(\omega)\|^{1+\epsilon}\} \leq C < +\infty \tag{9}$$

for some ϵ, $0 < \epsilon \leq 1$.

Next theorem presents a strong law of large number for dependent random variables, centered by conditional means. It is used further in the proof of Theorem 2.2.

Theorem 2.1 *Assume (5)-(9). Then* $\lim_k(\overline{\zeta}^k(\omega) - \overline{z}^k(\omega)) = 0$ *a.s. In particular, if* $\zeta^k : \Omega \longrightarrow Z$ *and* $\sigma_k = 1/k$ *then*

$$\lim_k \frac{1}{k}\sum_{s=1}^{k}(\zeta^s(\omega) - z^s(\omega)) = 0 \quad a.s. \tag{10}$$

Corollary 2.1 *Suppose additionally that sequence of conditional means* $\{z^k(\omega)\}$ *converges to a convex deterministic set* $Z^* \subseteq Z$*:*

$$\lim_k dist(z^k(\omega), Z^*) = 0 \quad a.s.$$

Then estimators $\{\overline{\zeta}^k(\omega)\}$ *converge to the same set:*

$$\lim_k dist(\overline{\zeta}^k(\omega), Z^*) = 0 \quad a.s.,$$

where $dist(z, Z^*) = \inf_{x\in Z^*}\|z - x\|$.

Corollary 2.2 *Suppose that sequence* $\{z^k(\omega)\}$ *a.s. converges to some random limit:*

$$\lim_k z^k(\omega) = z^*(\omega) \in Z \quad a.s.$$

Then sequence $\{\overline{\zeta}^k(\omega)\}$ *is a consistent estimate of this limit:*

$$\lim_k \overline{\zeta}^k(\omega) = z^*(\omega) \quad a.s..$$

In particular, if $\zeta^k(\omega) : \Omega \longrightarrow Z$ *and* $\sigma_k = 1/k$*, then*

$$\lim_k \frac{1}{k}\sum_{s=1}^{k}\zeta^s(\omega) = z^*(\omega) \quad a.s.$$

Corollary 2.3 *Assume additionally to (5)-(8), that random variables* $\zeta^k(\omega) : \Omega \longrightarrow Z$ *are independent,* $\sigma_k = 1/k$*,* $\mathcal{F}_0 = \{\emptyset, \Omega\}$*,* $\mathcal{F}_k = \sigma\{\zeta^1, \ldots, \zeta^k\}$*. Then* $z^k = \mathbf{E}\{\zeta^k(\omega)|\mathcal{F}_{k-1}\} = \mathbf{E}\zeta^k(\omega)$*, and (10) takes on a standard form*

$$\lim_k \frac{1}{k}\sum_{s=1}^{k}(\zeta^s(\omega) - z^s) = 0 \quad a.s. \tag{11}$$

Remark 2.1 *Denote*

$$\lambda_{kn} = \sigma_k \prod_{i=k+1}^{n}(1 - \sigma_i), \;\; 1 \geq k < n, \;\; \lambda_{nn} = \sigma_n. \tag{12}$$

Obviously,

$$y^{n+1} = \sum_{k=1}^{n}\lambda_{kn}(\zeta^k - z^k), \tag{13}$$

where $\{\lambda_{kn}\}$ *satisfy conditions*

$$\max_{1\leq k\leq n}\lambda_{kn} \longrightarrow 0, \quad \sum_{k=1}^{n}\lambda_{kn} \longrightarrow 1, \quad as \;\; n \longrightarrow \infty. \tag{14}$$

For a general averaging procedure (13), (14) a weak law of large numbers $y^n \xrightarrow{P} 0$ *(in probability) is known (see, for example, [1]). Theorem 2.1 presents a strong law of large numbers for dependent random variables with specific averaging coefficients (12) such that averaging (13) can be made iteratively.*

Remark 2.2 *The idea of estimators (6) comes from the theory of stochastic quasi-gradient methods [3]. It was observed [3], pp.130, 161 (see also [6], p.177), that the law of large numbers can be interpreted as a stochastic gradient procedure for solving some quadratic STO problem. We can interpret the expression* $(\overline{\zeta}^k(\omega) - \zeta^k(\omega))$ *in (6) as a gradient (in* z*) of the function*

$$f(z, \zeta^k(\omega)) = \frac{1}{2}\|z - \zeta^k(\omega)\|^2 - \frac{1}{2}\mathbf{E}(\zeta^k(\omega))^2,$$

at $z = \overline{\zeta}^k(\omega)$*. Let us consider*

$$F_k(z) = \mathbf{E}f(z, \zeta^k(\omega)) = \frac{1}{2}\|z - \mathbf{E}\zeta^k\|^2 - \frac{1}{2}(\mathbf{E}\zeta^k)^2, \tag{15}$$

which achieves its minimum at $z = \mathbf{E}\zeta^k$*. So procedure (6) can be viewed as an attempt to minimize function* $F_k(z)$ *(15) by STO procedure (6) with the projection on a convex set* Z*.*

Next theorem shows how to track mean values $z^k(\omega) = \mathbf{E}\{\zeta^k(\omega)|\mathcal{F}_{k-1}\}$, if they do not converge to any limit. Results of this kind are required for constraint Monte Carlo optimization (see [3, 6, 10] and examples 3.4, 3.5).

Instead of (9) assume that for some ϵ, $0 < \epsilon \leq 1$,

$$\mathbf{E}\|\zeta^k - z^k\|^{1+\epsilon} \leq C < \infty, \tag{16}$$

and

$$\sum_{k=1}^{\infty} \sigma_k^{1+\epsilon} < \infty. \tag{17}$$

Theorem 2.2 *Assume (5) – (8), (16), (17). Then*

$$\lim_k (\overline{\zeta}^k(\omega) - \overline{z}^k(\omega)) = 0 \quad a.s. \tag{18}$$

Suppose additionally that

$$\lim_k \frac{\|z^{k+1} - z^k\|}{\sigma_k} = 0 \quad a.s., \tag{19}$$

then

$$\lim_k (\overline{z}^k - \Pi_Z(z^k) = 0 \quad a.s., \tag{20}$$

and hence

$$\lim_k (\overline{\zeta}^k(\omega) - \Pi_Z(z^k(\omega)) = 0 \quad a.s. \tag{21}$$

Thus if $z^k(\omega) \in Z$ then estimator $\overline{\zeta}^k$ a.s. tracks a moving mean z^k as $k \longrightarrow \infty$.

The proof of Theorem 2.2 rests on Theorem 2.1 and the following general statement.

Let $v_k \geq 0, \sigma_k \geq 0, \gamma_k, w_k, k \geq 1$, be random variables. Suppose that each of the following conditions is fulfilled with probability 1:

$$v_{k+1} \leq v_k - \sigma_k w_k + \gamma_k, \quad k \geq 1; \tag{22}$$

$$\lim_k \sigma_k = 0, \quad \sum_{k=1}^{\infty} \sigma_k = +\infty; \tag{23}$$

$$v_1 + \sum_{k=1}^{\infty} \gamma_k < +\infty. \tag{24}$$

$$\text{If } \liminf_s v_{k_s} > 0, \text{ then } \liminf_s w_{k_s} > 0; \tag{25}$$

$$\text{If } \limsup_s v_{k_s} < \infty, \text{ then } \limsup_s |w_{k_s}| < \infty. \tag{26}$$

Theorem 2.3 *If conditions (22)-(26) are fulfilled a.s., then* $\lim_k v_k = 0$ *with probability 1.*

Conditions (22)-(26) can be viewed as a stochastic version of the second Lyapunov's method generalizing related results from [2, 3, 8, 11, 12, 16, 18]). The essential new feature is that the estimates w_k of "derivatives" of the corresponding Lyapunov function are not necessary nonnegative. It is important for various applications. The proof of Theorem 2.3 requires essentially new approaches. The requirement (23) is standard for iterative stochastic procedures; (24) usually follows from convergence of some martingale or quasimartingale. For example, let $\gamma_k = a_k + b_k$, where $\{a_k\}$ generates an a.s. convergent martingale with respect to some flow of nondecreasing σ-algebras $\mathcal{F}_k$; $b_k \geq 0$, and

$$\mathbf{E}v_1 + \sum_{k=1}^{\infty} \mathbf{E}b_k < +\infty.$$

Then (24) is fulfilled. Conditions (25), (26) relate values v_k and w_k and are easily checked for concrete situations.

3 Applications

Example 3.1 Concurrent estimation in Monte Carlo optimization. *Consider the minimization of function (1), where* $f(\cdot, \theta)$ *is a convex function,* X *is a convex compact set in* R^n. *There is a number of iterative stochastic optimization procedures generating a sequence* $x^k \longrightarrow X^*$ *a.s.,* $F(x^k) \longrightarrow F^*$ *a.s. on the basis of independent samples* $\{\theta^s\}$ *of* θ. *But these methods, as a rule, do not comprise construction of estimates for* F^*. *A natural way to produce such estimates is to construct a sequence*

$$y^k = \frac{1}{k} \sum_{s=1}^{k} f(x^s, \theta^s), \quad k = 1, 2, \ldots,$$

concurrently with the minimization sequence x^k. *Denote* $\zeta^k = f(x^k, \theta^k)$ *and introduce* σ*-algebra* $\mathcal{F}_k = \sigma\{x^1, \theta^1, x^2, \ldots, x^k, \theta^k\} \subset \Sigma$, *generated by random variables* $\{x^1, \theta^1, x^1, \ldots, x^k, \theta^k\}$. *Obviously,* ζ^k *is measurable with respect to* $\mathcal{F}_k$ *and* $\mathbf{E}\{\zeta^k | \mathcal{F}_{k-1}\} = F(x^k) \longrightarrow F^*$ *a.s. If* $|f(x, \theta)|$ *is majorized for any* $x \in X$ *by an integrable function* $C(\theta)$, $\mathbf{E}C^{1+\epsilon}(\omega) < \infty$, $0 < \epsilon \leq 1$, *then* $y^k \longrightarrow F^*$ *a.s. according to Corollary 2.1.*

Another important problem is the estimation of gradients $\nabla F(x^k) \longrightarrow 0$ *on the basis of observations* $\nabla f(x^k, \theta^k)$ *to evaluate closeness of* x^k *to* $X^* = \{x | \nabla F(x) = 0\}$. *If* $\nabla f(\cdot, \theta)$ *is a lipschitzian mapping with integrable in* $(1 + \epsilon)$*-power Lipschitz constant and* $x^k \longrightarrow X^*$, *then likewise*

$$g^k = \frac{1}{k} \sum_{s=1}^{k} \nabla f(x^s, \theta^s) \longrightarrow 0 \ \ a.s.$$

Example 3.2 Adaptive Monte Carlo method *(see [12, 17, 19, 20]). Suppose that a real random variable* $\xi(x)$ *depending on a parameter* $x \in X \subset R^n$ *has a common mean* $\mathbf{E}\xi(x) = m^*$, *which is to be estimated. Let us denote the variance function* $F(x) = \mathbf{E}(\xi(x) - m^*)^2$ *with gradient* $\nabla F(x) = 2\mathbf{E}\xi(x)\nabla\xi(x)$ *and estimate* m^* *adaptively changing parameter* x *towards minimal values of* $F(x)$*:*

$$m^k = m^{k-1} + \frac{1}{k}(\xi^k - m^{k-1}), \quad m^0 = 0, \quad k = 1, 2, \ldots,$$

where $\{\xi^k\}$ *are independent observations of* $\{\xi(x^k)\}$, $\{x^k\}$ *is a sequence minimizing* $F(x)$ *over* X. *Then by Corollary 2.1 estimates* $\{m^k\}$ *a.s. converge to* m^*.

Example 3.3 Estimation of subsets in a stochastic branch and bound method. *In the stochastic branch and bound method [14] a sequence of sets $X^k(\omega) \subset X^{k-1}(\omega)$ is constructed, and one has to estimate the lower bound value $L(\cdot)$ on the limit set $X^* = \lim_k X^k(\omega)$, using independent observations of random variables $\xi(X^k)$ such that $\mathbf{E}\xi(X^k) = L(X^k)$. For this purpose in [14] the following estimate is used:*

$$L_k(X^k) = \frac{1}{k}\sum_{s=1}^{k} \xi(X^k) \longrightarrow L(X^*).$$

Example 3.4 Averaging of gradients. *Let us come back to optimization of (1). Assume for simplicity that random function $f(\cdot,\theta)$ is continuously differentiable, $|f(x,\theta)| \le C(\theta)$ and $\|\nabla f(x,\theta)\| \le C(\theta)$ for $x \in X$, $\mathbf{E}C^{1+\epsilon}(\theta) < \infty$, $0 < \epsilon \le 1$. A method of averaged stochastic gradients (see, for example, [3, 6, 10]) generates a sequence $x^k \in X$ as follows: $x^1 \in X$, $y^1 = 0$,*

$$\begin{aligned} x^{k+1} &= \Pi_X(x^k - \rho_k z^k), \\ z^{k+1} &= z^k - \sigma_k(z^k - \xi^k), \\ \xi^k &= \nabla f(x^k, \theta^k), \quad k = 1, 2, \ldots, \end{aligned}$$

where Π_X is a projection operator on the set X, $\{\theta^k\}$ are independent observations of θ, nonnegative numbers σ_k satisfy (8) with $\epsilon = 1$, and

$$\rho_k \ge 0, \quad \sum_{k=1}^{\infty} \rho_k = +\infty, \quad \sum_{k=1}^{\infty} \rho_k^2 < \infty, \quad \lim_k \frac{\rho_k}{\sigma_k} = 0.$$

Then by Theorem 2.2 estimates z^k of gradients $\nabla F(z^k)$ are asymptotically consistent:

$$\lim_k (z^k - \nabla F(x^k)) = 0 \quad a.s.$$

without an assumption on convergence of x^k. It shows that the method of averaged stochastic gradients for large k practically operates as a deterministic gradient method, what explains its convergence a.s.

Example 3.5 Minimization of risk functions. *In practice a given decision $x \in R^n$ may result in different outcomes $g(x,\omega) \in R^m$ which are affected by "uncertainty" ω ("environment", "state of nature", exogenous factors and etc.). The expected utility is an evaluation (for some utility function $u(\cdot)$)*

$$U(x) = \int u(g(x,\omega))\mathbf{P}(d\omega),$$

which is linear with respect to the probability measure $\mathbf{P}$. This is an essential limitation for the applicability of $U(x)$ to problems where variances of $g(x,\omega)$ and other moments may significantly affect the results of decisions x. The risk functions which are used in applications (for example, the classical Markowitz model, some models based on stochastic dominance, ratios of conditional expectations and others) often have the following form:

$$R(x) = \int r(x, g(x,\omega), \omega)\mathbf{P}(d\omega),$$

which is not linear in $\mathbf{P}$. The calculation of function $r(\cdot,\cdot,\cdot)$ requires the evaluation of the expectation, i.e. in general functions r, R are not known explicitly. Assume that $r(x,z,\omega)$ is calculated exactly for a given (x,z,ω) and consider the sequence

$$\overline{u}(k+1) = \overline{u}(k) + \sigma_k(g(x^k,\omega^k) - \overline{u}(k)), \quad \overline{u}(1) = 0, \quad k = 1, 2, \ldots,$$

where x^k is a current approximate minimizer of $R(x)$, ω^k are independent samples of ω and σ_k satisfy (8). Assume that x^k is generated by a stochastic optimization procedure with step sizes ρ_k, $\|x^{k+1} - x^k\| \leq \rho_k Const$. If $\lim_k \rho_k/\sigma_k = 0$, then from Theorem 2.2 follows that $\|\overline{u}(k) - \mathbf{E}g(x^k, \omega)\| \longrightarrow 0$ for $k \longrightarrow \infty$. Therefore, $r(x^k, \overline{u}(k), \omega^k)$ can be used as an estimate of $r(x^k, \mathbf{E}g(x^k, \omega), \omega^k)$ in the design of stochastic optimization procedures for the risk functions.

4 Proofs

Proof of Theorem 2.1. Consider (Lyapunov) function $v(z) = \|z\|^{1+\epsilon}$, $z \in R^n$, $0 < \epsilon \leq 1$. Its gradient $\nabla v(z) = (1+\epsilon)\|z\|^{\epsilon-1}z$ satisfies Hölder condition:

$$\|\nabla v(y) - \nabla v(z)\| \leq (1+\epsilon)\|y - z\|^{\epsilon}.$$

So for any $y,\ z$ the following inequality holds (see [16]):

$$v(y) \leq v(z) - \langle \nabla v(z), z - y\rangle + \|z - y\|^{1+\epsilon}, \tag{27}$$

where $\|\cdot\|$, $\langle\cdot,\cdot\rangle$ denotes Euclidian norm and inner product in R^n. Denote

$$\eta^k(\omega) = \zeta^k(\omega) - z^k(\omega),$$

$$y^k = \overline{\zeta}^k(\omega) - \overline{z}^k(\omega).$$

Obviously, $\mathbf{E}\{\eta^k|\mathcal{F}_{k-1}\} = 0$. By contraction property of the projection operator we have

$$\begin{aligned}
\|y^{k+1}\| &= \|\Pi_Z(\overline{\zeta}^k - \sigma_k(\overline{\zeta}^k - \zeta^k)) - \Pi_Z(\overline{z}^k - \sigma_k(\overline{z}^k - z^k))\| \\
&\leq \|(\overline{\zeta}^k - \sigma_k(\overline{\zeta}^k - \zeta^k)) - (\overline{z}^k - \sigma_k(\overline{z}^k - z^k))\| \\
&= \|y^k - \sigma_k(y^k - \eta^k)\|.
\end{aligned} \tag{28}$$

From (28) and (27) it follows:

$$\begin{aligned}
v(y^{k+1}) &\leq v(y^k) - (1+\epsilon)\sigma_k\|y^k\|^{\epsilon-1}\langle y^k, y^k - \eta^k\rangle \\
&\quad +\sigma_k^{1+\epsilon}\|y^k - \eta^k\|^{1+\epsilon} \\
&\leq v(y^k) - (1+\epsilon)\sigma_k v(y^k) + (1+\epsilon)\sigma_k\|y^k\|^{\epsilon-1}\langle y^k, \eta^k\rangle \\
&\quad +2^{\epsilon}\sigma_k^{1+\epsilon}(\|y^k\|^{1+\epsilon} + \|\eta^k\|^{1+\epsilon}) \\
&= v(y^k) - (1+\epsilon)\sigma_k(1 - \frac{2^{\epsilon}}{1+\epsilon}\sigma_k^{\epsilon})v(y^k) \\
&\quad +(1+\epsilon)\sigma_k\|y^k\|^{\epsilon-1}\langle y^k, \eta^k\rangle + 2^{\epsilon}\sigma_k^{1+\epsilon}\|\eta^k\|^{1+\epsilon} \\
&\leq v(y^k) - (1+\epsilon)\sigma_k(1 - \sigma_k^{\epsilon})v(y^k) \\
&\quad +(1+\epsilon)\sigma_k\|y^k\|^{\epsilon-1}\langle y^k, \eta^k\rangle + 2^{\epsilon}\sigma_k^{1+\epsilon}\|\eta^k\|^{1+\epsilon}
\end{aligned} \tag{29}$$

Denote

$$v_k = v(y^k) \geq 0,$$

$$w_k = (1+\epsilon)(1 - \sigma_k^{\epsilon})v(y^k) \geq 0,$$

$$\chi_k = (1+\epsilon)\sigma_k\|y^k\|^{\epsilon-1}\langle y^k, \eta^k\rangle,$$

$$\gamma_k = 2^{\epsilon}\sigma_k^{1+\epsilon}\|\eta^k\|^{1+\epsilon} \geq 0.$$

Now inequalities (29) can be rewritten in the following form:

$$v_{k+1} \le v_k - \sigma_k w_k + \chi_k + \gamma_k. \tag{30}$$

Obviously, $\mathbf{E}\{\chi_k|\mathcal{F}_{k-1}\} = 0$ and from (9) we have

$$\sum_{k=1}^{\infty} \mathbf{E}\gamma_k = 2^{\epsilon} \sum_{k=1}^{\infty} \mathbf{E}\sigma_k^{1+\epsilon}\|\eta^k\|^{1+\epsilon} < \infty. \tag{31}$$

Taking conditional expectation from both sides of (30) we obtain

$$\mathbf{E}\{v_{k+1}|\mathcal{F}'_k\} \le v_k - \sigma_k w_k + \gamma_k \le v_k + \gamma_k. \tag{32}$$

Convergence of nonnegative random sequences $\{v_k\}$, satisfying (32) with $w_k \ge 0$ and (31) was studied in [2, 18]. From these results it follows that almost sure there exists a limit $(\lim_k v_k)$ and almost sure $\sum_{k=1}^{\infty} \sigma_k w_k < \infty$. From here, nonnegativity of w_k and (8) it follows that $\liminf_k w_k = 0$. But since $w_k = (1+\epsilon)(1-\sigma_k^{\epsilon})v_k$, then almost sure $\lim_k v_k = 0$.

Remark 4.1 *In the proof of the convergence in Theorem 2.1 we basically followed [18], but instead of quadratic Lyapunov function $v(x) = \|x\|^2$ we used $v(x) = \|x\|^{1+\epsilon}$, $0 < \epsilon \le 1$, and inequality (27) from [16].*

Proof of Theorem 2.2. Statement (21) is a consequence of (18) and (20). Condition (9) follows from (16), (17), so the statement (18) follows from Theorem 2.1. Let us deduce (20) from (8) and (19) by using second Lyapunov's method in the form of Theorem 2.3 with function $v(z) = \|z\|^{1+\epsilon}$. By property of the projection operator we have

$$\begin{aligned} \|\overline{z}^{k+1} - \Pi_Z(z^{k+1})\| &\le \|\overline{z}^k - \sigma_k(\overline{z}^k - z^k) - z^{k+1}\| \\ &= \|\overline{z}^k - z^k - \sigma_k(\overline{z}^k - z^k) - (z^{k+1} - z^k)\|. \end{aligned} \tag{33}$$

For $y = \overline{z}^k - z^k - \sigma_k(\overline{z}^k - z^k) - (z^{k+1} - z^k)$ and $z = \overline{z}^k - z^k$ from (27) and (33) it follows:

$$\begin{aligned} \|\overline{z}^{k+1} - \Pi_Z(z^{k+1})\|^{1+\epsilon} &\le \|\overline{z}^k - z^k\|^{1+\epsilon} \\ &\quad -(1+\epsilon)\|\overline{z}^k - z^k\|^{\epsilon-1}\langle \overline{z}^k - z^k, \sigma_k(\overline{z}^k - z^k) + (z^{k+1} - z^k)\rangle \\ &\quad +\|\sigma_k(\overline{z}^k - z^k) + z^{k+1} - z^k\|^{1+\epsilon} \\ &\le \|\overline{z}^k - z^k\|^{1+\epsilon} - (1+\epsilon)\|\overline{z}^k - z^k\|^{\epsilon+1}\sigma_k \\ &\quad +(1+\epsilon)\|\overline{z}^k - z^k\|^{\epsilon}\|z^{k+1} - z^k\| \\ &\quad +2^{\epsilon}(\sigma_k^{1+\epsilon}\|\overline{z}^k - z^k\|^{1+\epsilon} + \|z^{k+1} - z^k\|^{1+\epsilon}) \\ &\le \|\overline{z}^k - z^k\|^{1+\epsilon} \\ &\quad -\sigma_k(1+\epsilon)\|\overline{z}^k - z^k\|^{\epsilon}((1-\sigma_k^{\epsilon})\|\overline{z}^k - z^k\| - \|z^{k+1} - z^k\|/\sigma_k) \\ &\quad +2^{\epsilon}\|z^{k+1} - z^k\|^{1+\epsilon}. \end{aligned} \tag{34}$$

Let us introduce notations

$$v_k = \|\overline{z}^k - z^k\|^{1+\epsilon},$$

$$\begin{aligned} w_k &= (1+\epsilon)\|\overline{z}^k - z^k\|^{\epsilon}((1-\sigma_k^{\epsilon})\|\overline{z}^k - z^k\| - \|z^{k+1} - z^k\|/\sigma_k) \\ &= (1+\epsilon)v_k^{\frac{\epsilon}{1+\epsilon}}((1-\sigma_k^{\epsilon})v_k^{\frac{1}{1+\epsilon}} - \|z^{k+1} - z^k\|/\sigma_k), \end{aligned} \tag{35}$$

$$\gamma_k = 2^{\epsilon}\|z^{k+1} - z^k\|^{1+\epsilon}.$$

Then (34) takes on the form:

$$v_{k+1} \le v_k - \sigma_k w_k + \gamma_k. \tag{36}$$

By conditions (19) and (17)

$$\sum_{k=1}^{\infty} \gamma_k < \infty \quad \text{a.s.} \tag{37}$$

From (36) – (37) it follows that sequences $v_k,\ w_k,\ \gamma_k$ satisfy conditions (22)–(26). By Theorem 2.3

$$\lim_k v_k = 0 \ \text{ a.s.}\square \tag{38}$$

Proof of Theorem 2.3. The proof involves general ideas of arguing from the contradiction which are typical for the convergence analysis of nonmonotone optimization procedures [3, 15]. Let Ω' denotes a set of those $\omega \in \Omega$ that all conditions (22)-(26) are fulfilled simultaneously, $\mathbf{P}(\Omega') = 1$. Fix some $\omega \in \Omega'$. Let us show that $\liminf_k v_k(\omega) = 0$. Suppose the opposite, i.e. that $\liminf_k v_k > 0$. Then from (25) it follows that $\liminf_k w_k > 0$, i.e. for all $k \ge k'$ and some $\epsilon > 0$ we have $w_k \ge \epsilon > 0$. From (22) for $k \ge k'$ it follows that

$$v_{k+1} \le v_k - \sigma_k \epsilon + \gamma_k.$$

Summing up these inequalities from k' to m:

$$0 \le v_{m+1} \le v_{k'} - \epsilon \sum_{k=k'}^{m} \sigma_k + \sum_{k=k'}^{m} \gamma_k \longrightarrow -\infty \quad m \longrightarrow \infty,$$

we obtain a contradiction for m large enough, hence $\liminf_k v_k(\omega) = 0$.

Now let us show that $\limsup_k v_k(\omega) = 0$. Suppose the opposite, i.e. that $\limsup_k v_k(\omega) > 0$. Choose numbers A and B such that

$$0 < A < B < \limsup_k v_k(\omega).$$

Obviously, there exist indices $n_s = n_s(\omega)$ and $m_s = m_s(\omega)$, $s = 1, 2, \ldots$, such that

$$v_{n_s} \le A < v_k \le B < v_{m_s}, \ \ n_s < k < m_s. \tag{39}$$

Since $\limsup_s v_{n_s} \le A < \infty$, then by (26) $\limsup_s |w_{n_s}| < \infty$, i.e. for all sufficiently large $s \ge S$ and some D, $|w_{n_s}| < D$. From (22), (39), (24) it follows that

$$\begin{aligned} 0 \le v_{n_s+1} - v_{n_s} &\le -\sigma_{n_s} w_{n_s} + \gamma_{n_s} \\ &\le \sigma_{n_s} D + \gamma_{n_s} \longrightarrow 0. \quad s \longrightarrow \infty. \end{aligned}$$

Hence,

$$\lim_s v_{n_s} = \lim_s v_{n_s+1} = A. \tag{40}$$

Let us sum up both sides of inequalities (22) over k from $n_s + 1$ to $m_s - 1$:

$$v_{m_s} \le v_{n_s+1} - \sum_{k=n_s+1}^{m_s-1} \sigma_k w_k + \sum_{k=n_s+1}^{m_s-1} \gamma_k, \tag{41}$$

and show that

$$\liminf_s \sum_{k=n_s+1}^{m_s-1} \sigma_k w_k \ge 0.$$

Let $w_{k_s} = \min_{n_s<k<m_s} w_k$. Since $\liminf_s v_{k_s} \geq A > 0$, then by (25) $\liminf_s w_{k_s} > 0$, and

$$\liminf_s \sum_{k=n_s+1}^{m_s-1} \sigma_k w_k \geq \liminf_s \left(w_{k_s} \sum_{k=n_s+1}^{m_s-1} \sigma_k \right) \geq 0.$$

Coming in (41) to the limit in s, we obtain by (39), (40), (24) a contradiction

$$B \leq v_{m_s} \leq \limsup_s v_{n_s+1} - \liminf_s \sum_{k=n_s+1}^{m_s-1} \sigma_k w_k + \limsup_s \sum_{k=n_s+1}^{m_s-1} \gamma_k \leq A,$$

i.e. $\limsup_k v_k(\omega) = 0$. □.

References

[1] Borovkov A.A. (1986), *Theory of Probabilities*, Second ed., Nauka, Moscow.

[2] Ermoliev Yu.M. (1969), On the method of generalized stochastic gradients and stochastic quasi-Fejer sequences, *Kibernetika*, 1969, N 2, pp. 73-84 (in Russian, English translation in *Cybernetics*, 1969, Vol. 5, N 2, pp. 208-220).

[3] Ermoliev Yu.M. (1976), *Methods of Stochastic Programming*, Nauka, Moscow. (In Russian).

[4] Ermolieva T. (1997), The Design of Optimal Insurance Decisions in the Presence of Catastrophic Risks, Interim Report IR-97-068, Int. Inst. for Appl. Syst. Analysis, Laxenburg, Austria. (In Internet http://www.iiasa.ac.at/Publications/).

[5] Ermoliev Yu.M., Ermolieva T.Yu. and Norkin V.I. (1997), Spatial Stochastic Model for Optimization Capacity of Insurance Networks Under Dependent Catastrophic Risks: Numerical Experiments, Interim Report IR-97-028, Int. Inst. for Appl. Syst. Analysis, Laxenburg, Austria. (In Internet http://www.iiasa.ac.at/Publications/).

[6] Ermoliev Yu.M. and Wets R.J-B. (Eds.) (1987), *Numerical Techniques for Stochastic Optimization*, Springer, Berlin.

[7] Gihman I.I. and Skorohod A.V. (1971), Theory of random processes, Vol. 1, Nauka, Moscow (In Russian).

[8] Katkovnik V.J. (1976), Linear estimates and stochastic optimization problems, Moscow, Nauka, 1976.

[9] Loève M (1960). Probability Theory. 2nd ed. – D.Van Nostrand Company Inc., Princeton, 1960.

[10] Mikhalevich V.S., Gupal A.M. and Norkin V.I. (1987), Methods of nonconvex optimization, Nauka, Moscow.

[11] Nakonechnyi A.N. (1990), Probabilistic generalization of the second Lyapunov's method, Doklady Akademii Nauk Ukrainian SSR, Ser. A (Proceeding of the Ukrainian Academy of Sciences), No. 2, pp. 18-19.

[12] Nakonechnyi A.N. (1995), Stochastic gradient processes: a review of a convergence theory using second Lyapunov's method, Kibernetika i sistemnyi analiz, 1995, N 1, pp.46-62. (In Russian, English translation in *Cybernetics and System analysis*, Vol. 27, No.1).

[13] Neveu J. Mathematical foundations of the Calculous of Probabilities. – San Francisco: Holden-Day, 1965.

[14] Norkin V. I., Pflug G. Ch. and Ruszczyński A.(1996),
A Branch and Bound Method for Stochastic Global Optimization, Working Paper WP-96-065, Int. Inst. for Appl. Syst. Analysis, Laxenburg, Austria (In Internet http://www.iiasa.ac.at/Publications/).

[15] Nurminski E.A. (1979), Numerical methods for solving deterministic and stochastic minimax problems, Kiev, Naukova dumka, 1979. (In Russian).

[16] Polyak B.T. (1976), Convergence and rate of convergence of iterative stochastic algorithms. I. General Case, *Avtomatika i telemehanika*, 1976, Vol. 37, N 12, pp. 83-94 (In Russian, English translation in *Automation and Remote Control*, Vol. 37, pp.1858-1868).

[17] Pugh E.L. (1966), A gradient technique of adaptive Monte Carlo, SIAM Rev., 1966, Vol.8, N3, pp.346-355.

[18] Robbins H. and Siegmund D. (1971), A convergence theorem for non negative almost supermartingales and some applications, in: *Optimization methods in statistics*, Academic Press, New York, pp. 233-257.

[19] Shpak V.D. (1989), Nonbiased estimates for the solution of an integral equation of the second kind and their application to calculation of reliability indicators for semimarkov systems, Doklady Akademii Nauk Ukrainian SSR, Ser. A (Proceeding of the Ukrainian Academy of Sciences), No. 10, pp. 81-84.

[20] Troubetzkoy E.S. (1991), Optimization of Linear Monte Carlo Calculation, Nucl. Sci. and Eng., 1991, Vol.107, N 4, pp.359-364.