ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 8, pp. 699–717.

SPECIAL ISSUE

A Search Method for Stochastic Non-Stationary Optimization of Functions with Hölder Gradient

I. A. Akinfiev[*,a], **O. N. Granichin**[*,**,b], **and E. Yu. Tarasova**[*,c]

[*] *Saint Petersburg State University, St. Petersburg, Russia*
[**] *Institute for Problems in Mechanical Engineering, Russian Academy of Sciences, St. Petersburg, Russia*
e-mail: [a] *i@iakinfiev.ru,* [b] *o.granichin@spbu.ru,* [c] *elizaveta.tarasova@spbu.ru*

Received June 23, 2025
Revised June 30, 2025
Accepted July 4, 2025

Abstract—We propose a gradient-free method of stochastic optimization with perturbation at the input which is designed to track changes in the minimum point of a function with Hölder gradient, with observations subject to almost arbitrary (unknown-but-bounded) noise. Similar methods are widely used in adaptive control problems (energy, logistics, robotics, goal tracking), optimization of noisy systems (biomodeling, physical experiments), and online learning with drift of the data parameters (finance, streaming analytics). The efficiency of the algorithm is tested under conditions that mimic tracking the evolution of human expectations in reinforcement learning problems based on human feedback when tracking the center of a cluster of problems in queueing systems. Search methods with input perturbations have been actively developed in the works by B.T. Polyak since 1990.

Keywords: tracking, input perturbations, randomization, stochastic optimization, gradient-free methods, reinforcement learning via human feedback, queueing systems, unknown-but-bounded disturbances

DOI: 10.31857/S0005117925080023

1. INTRODUCTION

The problem of minimizing a function (functional) $f(x)$ is at the heart of solving many practical problems, from control of engineering systems to machine learning. Closed-form solutions are often not available due to high dimensionality, nonlinearities, or the lack of an explicit form. Even when the function is defined explicitly, the practical applicability of the existing approaches is limited by computational resources, measurement inaccuracies, or rounding errors. Traditional iterative gradient methods are efficient when finding the minimum of smooth or differentiable functions. However, in real-world problems, situations often arise where computing the gradient is difficult or impossible. Typically, the objective function is subject to stochastic disturbances, or its explicit form is unknown. In practice, the optimized function is often defined by some oracle, and by making requests (function arguments) to this oracle, it is possible to obtain certain realizations. The availability of measurements of the gradient itself is feasible with the implementation of special measuring devices for specific tasks or through finite difference approximations, which are inefficient in the presence of a high-level noise in the obtained measurements. In such cases, alternative approaches are required that do not rely on the information about gradients.

A significant contribution to the development of the theory and methods of stochastic optimization was made by B.T. Polyak and his research group. Their research covers a wide range of issues, including gradient methods [1], pseudo-gradient adaptation and learning algorithms [2–4], and methods for accelerating convergence [5–7]. Even nowadays, the two papers [8, 9] provide

comprehensive answers when analyzing the convergence of general-type iterative stochastic algorithms in terms of mean-square deviations, as well as in the linear case in terms of error covariance matrices.

A new search method of stochastic approximation proposed in the 1990 paper [10] not only develops the overall direction of random search algorithms [11], but also significantly advances the entire general theory of iterative optimization algorithms. This paper shows that, if the observed values of the optimized function are corrupted by noise, the proposed algorithm has the asymptotically optimal rate of convergence in the sense that it is impossible to find a faster algorithm among all possible iterative optimization algorithms for a sufficiently broad class of functions. A similar algorithm was previously proposed in [12], and consistency of estimates generated by it was justified in the presence of almost arbitrary noise in the observations. In the English-language literature, similar methods have been called SPSA (Simultaneous Perturbation Stochastic Approximation), see [13, 14]. A salient feature of these gradient-free methods is that, regardless of the dimensionality of the problem, the oracle needs to be called only once or twice per iteration, with arguments being chosen over a randomly generated line through the current point (it is what is referred to as randomization of the algorithm). A detailed analysis of the history of development of search algorithms of stochastic approximation with perturbation at the input, as well as the properties of the estimates generated by these methods are provided in [15–17].

A limitation of classical iterative zero-order stochastic optimization methods (those which do not use the values of the gradient), such as the Kiefer-Wolfowitz procedure [18] in the multivariate case, is the need to repeatedly compute the function at each iteration. This becomes especially impractical in dynamical environments where the target function $f_n(x)$ changes over time. A similar situation arises, for example, in optimization problems related to real-time systems. It turned out that methods like the previously proposed search algorithms of stochastic optimization with input perturbation remain to be to be efficient in this situation when replacing the decreasing step-sizes over time with constant ones, [19, 20]. Later, it was possible to formulate and justify the properties of a distributed algorithm of this type, combined with a consensus algorithm [20].

In practice, [21, 22], statistical uncertainties are often encountered which do not have second statistical moment. For example, stable distributions, such as Levi–Pareto, are better at describing the prices of stocks and commodities than Gaussian distributions. In [24], the properties of the estimates provided by the SPSA algorithm under such conditions were studied. In the present paper, these studies are extended to the case of optimization of the non-stationary mean-risk functional.

2. STATEMENT OF THE PROBLEM

We consider discrete time $n = 0, 1, \ldots$, defined by the label of step (iteration), and we denote by $\{F_n(\cdot,\cdot)\colon \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}\}$ the set of functions in two vector variables, which are all differentiable with respect to the first argument. At every step n, observations

$$y_n = F_n(x_n, w_n) + v_n \tag{1}$$

are performed at known (chosen) points x_n (experimental design), where the w_ns are uncontrollable disturbances defined over a probabilistic space Ω and having identical unknown distribution $P_w(\cdot)$, and v_n is the (perhaps non-random) observation noise.

Let $\mathcal{F}_{n-1}$ denote the σ-algebra of all random events that have been realized up to the time instant n; $\mathbb{E}$ be the symbol of mathematical expectation; $\mathbb{E}_{\mathcal{F}_{n-1}}$ denote the conditional mathematical expectation relative to the σ-algebra $\mathcal{F}_{n-1}$.

We are interested in the minimization of the following nonstationary mean risk functional:

$$f_n(x) = \mathbb{E}_{\mathcal{F}_{n-1}} F_n(x, w) = \int\limits_{\mathbb{R}^q} F_n(x, w) P_w(dw) \to \min_x . \tag{2}$$

The goal is to evaluate the minimum point θ_n of the function $f_n(x)$; i.e., to find

$$\theta_n = \arg\min_x f_n(x).$$

Accuracy of the estimate x of the points θ_n is addressed through use of the scalar Lyapunov functions

$$V_n(x) = \|x - \theta_n\|^{\rho+1} = \sum_{i=1}^{n} |x^{(i)} - \theta_n^{(i)}|^{\rho+1},$$

where θ_n are the vectors to be found, and $\rho \in (0, 1]$ is the Hölder exponent for the gradients of the functions $V_n(x)$. In the sequel, we write $\|\cdot\|_{\rho+1}$ to denote the $l_{\rho+1}$-norm and $\langle\cdot,\cdot\rangle$ for the inner product in $\mathbb{R}^d$.

To characterize the behavior of the estimates of the minimum points of the non-stationary functional (2), we present two definitions.

Definition 1. The sequence $\hat{\theta}_n$ of the estimates of the minimum points θ_n is said to be $l_{\rho+1}$-*stabilized*, if there exists $C > 0$ such that

$$\mathbb{E}V_n(\hat{\theta}_n) \le C \quad \forall\, n.$$

Definition 2. The number L is referred to as the *asymptotic upper bound* for the estimation errors in the $l_{\rho+1}$-norm, if the sequence of estimates $\{\hat{\theta}_n\}$ of the minimum points θ_n satisfy

$$\overline{\lim_{n\to\infty}} \mathbb{E}V_n(\hat{\theta}_n) \le L < \infty.$$

In what follows, we construct the sequence of stabilizing estimates $\{\hat{\theta}_n\}$ in the spirit of Definition 2 under the following conditions satisfied for all $n > 0$:

(A) *The functions $f_n(\cdot)$ are strongly convex in the first argument:*

$$\langle \nabla V_n(x), \nabla f_n(x)\rangle \ge \mu V_n(x).$$

(B) *For all admissible w, the gradients $\nabla F_n(\cdot, w)$ satisfy the condition*

$$\|\nabla F_n(x, w) - \nabla F_n(y, w)\|_1 \le M\|x - y\|_\rho^\rho$$

for a certain constant M.

(C) *The local Lebesgue property:* For every point $x \in \mathbb{R}^d$ there exists a neighborhood U_x and a function $\Phi_x(w)$ such that $\mathbb{E}\Phi_x(w) < \infty$ and $\|\nabla F_n(x', w)\|_2 \le \Phi_x(w)\ \forall x' \in U_x$.

(D) *The rate of drift of the minimum point satisfies the following conditions:*

a: $\|\theta_n - \theta_{n-1}\|_1 \le A$;

alternatively, if $\{\theta_n\}$ is a sequence of random variables, then

$$\mathbb{E}_{\mathcal{F}_{n-1}}\|\theta_n - \theta_{n-1}\|_{\rho+1}^{\rho+1} \le A^{\rho+1},$$

b: $\mathbb{E}_{\mathcal{F}_{n-1}} \|\nabla_x F_n(x, w) - \nabla_x F_{n-1}(x, w)\|_1 \leqslant B\|x - \theta_{n-1}\|_1^\rho$,

c: $\mathbb{E}_{\mathcal{F}_{n-1}} \|\nabla_x F_n(\theta_n, w_n)\|_{\rho+1}^{\rho+1} \leqslant C$,

d: $\mathbb{E}_{\mathcal{F}_{2n-2}} |F_{2n}(x, w_{2n}) - F_{2n-1}(x, w_{2n-1})|^{\rho+1} \le DV_{2n-2}(x) + E$.

(E) *The observation noise v_n satisfies the condition*

$$|v_{2n} - v_{2n-1}| \le \sigma_v,$$

or

$$\mathbb{E}_{\mathcal{F}_{2n-2}}\{|v_{2n} - v_{2n-1}|^{\rho+1}\} \le \sigma_v^{\rho+1}$$

if it has random nature.

Note that the last condition is valid for arbitrary deterministic bounded sequences $\{v_n\}$. Condition (C) allows for interchanging the integration and differentiation operations when justifying the stabilizability of the estimates. Conditions of the form (D) cover both the random walk drift and directed drift in a certain direction. For instance, the following condition based on (D) is presented in [1]:

$$\theta_n = \theta_{n-1} + a + \xi_n,$$

where ξ_n is a zero-mean random variable, and a is trend. Stabilizability of the estimates generated by the algorithm under conditions (D) shows its applicability to a wide range of problems.

3. A SEARCH RANDOMIZED ESTIMATION ALGORITHM

Assume that the sequence $\{\Delta_n\}$ of trial simultaneous perturbations fed to the input of the algorithm is a realization of a sequence of independent Bernoulli vectors in $\mathbb{R}^d$ with components being independent random variables taking values $\pm\frac{1}{\sqrt{d}}$ with probability 0.5. Let us pick an initial vector $\theta_0 \in \mathbb{R}^d$. We will estimate the sequence $\{\theta_n\}$ of the minimum points by the sequence $\{\hat{\theta}_n\}$ defined by the following stochastic optimization algorithm with trial simultaneous input perturbations:

$$\begin{cases} \hat{\theta}_{2n-1} = \hat{\theta}_{2n-2} \\ x_{2n} = \hat{\theta}_{2n-2} + \beta\Delta_n,\ x_{2n-1} = \hat{\theta}_{2n-2} - \beta\Delta_n \\ \hat{\theta}_{2n} = \hat{\theta}_{2n-2} - \frac{\alpha}{2\beta}\Delta_n(y_{2n} - y_{2n-1}), \end{cases} \tag{3}$$

where α and β are the step-size parameters. To substantiate the stabilizability property of the estimates generated by algorithm (3), we adopt yet another assumption:

(F) *The random vectors Δ_n and w_{2n}, w_{2n-1} are independent of each other as well as of $\mathcal{F}_{n-1}$. If $\{v_n\}$ are assumed to have random nature, then Δ_n do not depend on v_{2n}, v_{2n-1}.*

4. STABILIZATION OF ESTIMATES

Denote $H = A + \alpha\beta M$, where A and M are constant bounds on the rate of drift and change of gradients, respectively.

Theorem 1. *Let conditions (A)–(F) be satisfied and let the parameters α, β be chosen in such a way as to guarantee the constant $K > 0$ defined later in the proof to be less than unity.*

Then, for any initial choice $\hat{\theta}_0$ with $E\|\hat{\theta}_0 - \theta_0\|^{\rho+1} < \infty$, the estimates generated by algorithm (3) *are being stabilized in the following sense:*

$$\overline{\lim_{n\to\infty}} \mathbb{E}\|\hat{\theta}_n - \theta_n\|_{\rho+1} \leqslant \left(\frac{L}{K}\right)^{\frac{1}{\rho+1}},$$

where L is also defined at the end of the proof.

Conditions (A)–(C) and (E)–(F) are standard when proving the consistency of estimates generated by stochastic optimization algorithms with input perturbations; see [18]. Mean-square stabilizability of the estimates provided by algorithm (3) has been earlier proved in [19] under more stringent assumptions.

Proof of Theorem 1 and the precise definition of the constants K and L are presented in the Appendix.

5. SIMULATION IN THE RLHF-SCENARIO

In reinforcement learning based on feedback from humans (Reinforcement Learning from Human Feedback, RLHF), a key challenge is working with noisy and unstable data, [26, 27]. Human evaluations often contain random errors and may change over time, hence complicating the optimization process. In particular, in tasks related to fine tuning of language models (Large Language Models, LLM), RLHF is used to improve the quality of text generation, align with user preferences, and minimize undesirable model behavior. However, the subjectivity and variability of human evaluations create significant difficulties for traditional optimization methods.

5.1. The Model

In the simulations, we examine the efficiency of the search algorithm under conditions close to reality; i.e., in the presence of heavy-tailed noise (Pareto distribution) and preference drift ([28, 30]), which mimics the evolution of human expectations. Three scenarios are considered: Moderate drift, near-stationary preferences, and stationary preferences with asymmetric noise. This allows for the assessment of stability and adaptability of the algorithm under RLHF conditions and checking its applicability to tasks related to LLM training and other systems where human feedback plays a key role.

The goal of simulations is to test the ability of RLHF agents to adapt to a reward model shaped from noisy and changing human evaluations. We then

— model heavy-tailed noise (Pareto distribution) describing uncertainty and rare but significant deviations in estimates;

— introduce a preference drift model that simulates the gradual change in human expectations;

— note that all functions and parameters are formulated in conditions (A)–(F) in Section 2.

Each agent has to minimize the discrepancy between its own estimate of the parameter and the true value set by the reward model, despite noise and dynamics of target preferences; a search algorithm is used for the minimization.

The RLHF-based reward model is specified as follows:

$$F_n(\mathbf{x}) = -\sum_{i=1}^{m}(x_i - x_n^*)^{1.35}, \tag{4}$$

where the target parameter x_n^* drifts in time n as

$$x_n^* = x_{n-1}^* + \delta, \quad x_0^* = 5,$$

thus, reflecting a change in preferences.

Choosing x_n from the feedback, we obtain

$$y_n = F_n(\mathbf{x}) + v_n,$$

where v_n is noise that models uncertainty in the feedback channel. Two types of noise were used in the simulations:

— symmetric noise $v_i = Z_i \cdot \mathrm{sgn}_i$, where $Z_i \sim \mathrm{Pareto}(\beta, \sigma)$, $\mathrm{sgn}_i \sim \mathrm{Uniform}(\{-1, 1\})$;

— asymmetric noise $v_i = Z_i$, where $Z_i \sim \mathrm{Pareto}(\beta, \sigma)$, which potentially reflects a tendency to overestimate.

Table 1 presents the basic parameters of the numerical simulation. They cover the structure of the experiment, settings of the algorithm (so-called hyper-parameters), as well as the characteristics of noise and drift scenarios, which model feedback instabilities.

Table 1. Parameters of simulations

Parameter	Description	Value
Agent's initial estimate	Initial point for learning	$\hat{\theta}_0 = 0$
Number of iterations	Number of adaptation steps	$N = 1000$
Number of runs	Amount of independent experiments	$m = 1000$
Hyper-parameters		
Adaptation step	Conservative step (for stability)	$\gamma = 0.05$
Level of perturbation	Amplitude to estimate the gradient	$c = 0.1$
Characteristics of noise		
Shape parameter	Defines weights of tails	$\beta = 1.6$
Scale	Intensity of deviations	$\sigma = 2.0$
Rate of drift	Moderate drift	$\delta = 0.01$
	Near-stationary mode	$\delta = 0.0001$
Type of noise	Random deviations	Symmetric
	Systematic bias	Asymmetric

5.2. Simulation Scenarios

To analyze the adaptability of the algorithm, we consider three scenarios:

1. Moderate drift of preferences ($\delta = 0.01$) and symmetric noise (referred to as noise with symmetric distribution). This scenario mimics gradual changes in target parameters in the presence of random errors in the estimates.
2. Near-stationary preferences ($\delta = 0.0001$) and symmetric noise. Within this scenario we test accuracy of tuning under conditions close to stable ones.
3. Stationary preferences ($\delta = 0.0001$) and asymmetric noise (referred to as noise with asymmetric distribution). This scenario corresponds to a systematic distortion of feedback; i.e., a permanent overestimation.

5.3. Agent Adaptation Process

The agent updates its estimate $\hat{\theta}$ of the parameter based on the observed values of y (rewards) obtained from the model. The algorithm follows the iterative scheme described in (3).

Namely, at every even iteration $k = 2n$, $n = 1, 2, \ldots$.

1. The estimate $\hat{\theta}_{2n-2}$ obtained at the previous even iteration is used (for $n = 1$, $\hat{\theta}_0$ is used).
2. A random vector Δ_n of perturbations is generated, with every component independently taking values $+1$ or -1 with probability 0.5.
3. Two points are considered according to (3):

$$x_{2n} = \hat{\theta}_{2n-2} + \beta\Delta_n, \qquad x_{2n-1} = \hat{\theta}_{2n-2} - \beta\Delta_n.$$

4. The values of the reward are then observed at the perturbed points: y_{2n} (associated with x_{2n}) and y_{2n-1} (associated with x_{2n-1}). These two quantities include both the true value of the function and the noise; i.e., $y_n = F_n(x_n, w_n) + v_n$ in terms of the notation of this paper.
5. The estimate $\hat{\theta}$ updates similarly to the formula given by the third line of system (3); however, with sign "+", since the maximization is performed:

$$\hat{\theta}_{2n} \leftarrow \hat{\theta}_{2n-2} + \frac{\alpha}{2\beta}\Delta_n(y_{2n} - y_{2n-1}).$$

At every odd iteration $k = 2n - 1$, the estimate is being copied: $\hat{\theta}_{2n-1} \leftarrow \hat{\theta}_{2n-2}$.

5.4. Checking Conditions (A)–(F) *for Simulations in the RLHF-Scenario*

(A) *Strong convexity of* $f_n(\mathbf{x})$.

$$\nabla f_n(\mathbf{x}) = -\nabla F_n(\mathbf{x}) = -\left[1.35(x_1 - x_n^*)^{0.35}, \dots, 1.35(x_m - x_n^*)^{0.35}\right]^\top,$$
$$\nabla V_n(\mathbf{x}) = [(\rho+1)\mathrm{sgn}(x_1 - x_n^*)|x_1 - x_n^*|^\rho, \dots, (\rho+1)\mathrm{sgn}(x_m - x_n^*)|x_m - x_n^*|^\rho]^\top,$$
$$\langle \nabla V_n(\mathbf{x}), \nabla f_n(\mathbf{x})\rangle = -1.35(\rho+1)\sum_{i=1}^m |x_i - x_n^*|^{\rho+0.35}.$$

Using the inequality $|x_i - x_n^*|^{\rho+0.35} \geq |x_i - x_n^*|^{\rho+1} a^{-0.65}$ with $a \leq |x_i - x_n^*|$, we obtain

$$\sum_{i=1}^m |x_i - x_n^*|^{\rho+0.35} \geq a^{-0.65}\sum_{i=1}^m |x_i - x_n^*|^{\rho+1} = a^{-0.65} V_n(\mathbf{x}).$$

Therefore,

$$\langle \nabla V_n(\mathbf{x}), \nabla f_n(\mathbf{x})\rangle \leq -1.35(\rho+1)a^{-0.65} V_n(\mathbf{x});$$

i.e., the condition of the form $\langle \nabla V_n(\mathbf{x}), \nabla f_n(\mathbf{x})\rangle \geq \mu V_n(\mathbf{x})$ holds for $\mu = -1.35(\rho+1)a^{-0.65} < 0$. In the minimization of $f_n(\mathbf{x})$, the strong convexity condition in the sense of the scalar inequality above is satisfied with $\mu < 0$.

(B) *The Hölder continuity of the gradient.*

The gradient of the reward function $F_n(x)$ writes

$$\nabla F_n(x) = -1.35\left[(x_1 - x_n^*)^{0.35}, \dots, (x_m - x_n^*)^{0.35}\right]^\top.$$

Then the components of the difference of the gradients have the form

$$\left|(x_i - x_n^*)^{0.35} - (y_i - x_n^*)^{0.35}\right| \leq M'|x_i - y_i|^{0.35},$$

where M' is the Hölder constant, which exist for the function $s \mapsto s^{0.35}$ over bounded intervals.

Substitution to the norm gives

$$\|\nabla F_n(x) - \nabla F_n(y)\|_2^2 = 1.35^2 \sum_{i=1}^m \left|(x_i - x_n^*)^{0.35} - (y_i - x_n^*)^{0.35}\right|^2$$
$$\leq 1.35^2 M'^2 \sum_{i=1}^m |x_i - y_i|^{0.7} \leq M^2 \|x - y\|_2^{0.7},$$

where $M^2 = 1.35^2 M'^2 m^{1-0.7/2}$ is a generalized constant.

Then we have

$$\|\nabla F_n(x) - \nabla F_n(y)\|_2 \leq M\|x - y\|_2^{0.35},$$

which corresponds to condition (B) with $\rho = 0.35$ and $M = 1.35M'm^{0.325}$.

(C) *The local Lebesgue condition.*

Let us fix the point x and consider its neighborhood $U_x = B(x, \varepsilon)$ for some $\varepsilon > 0$. Then, for any $x' \in U_x$ we have

$$\|\nabla F_n(x', w)\|_2^2 = 1.35^2 \sum_{i=1}^m |x_i' - x_n^*|^{0.7} \leq 1.35^2 m R^{0.7},$$

where $R = \sup_{x' \in U_x} \max_i |x_i' - x_n^*| < \infty$, and it is finite by the construction of U_x.

We then can set $\Phi_x(w) = 1.35\sqrt{m}R^{0.35}$, which is independent of w, so that $\mathbb{E}\Phi_x(w) = \Phi_x(w) < \infty$. Condition (C) is satisfied.

(D.a) *Boundedness of the drift of the minimum point.*

Since $\theta_n = x_n^*\mathbf{1}$ and $x_n^* = x_{n-1}^* + \delta$, we have $\|\theta_n - \theta_{n-1}\|_2 = \|\delta\mathbf{1}\|_2 = \delta\sqrt{m}$. Hence, condition (D.a) is satisfied for $A = \delta\sqrt{m}$.

(D.b) *Boundedness of change in the gradient.*

Let $r_i = x_i - x_{n-1}^*$, then

$$|\partial_i F_n(x) - \partial_i F_{n-1}(x)| \le 1.35M'|\delta|^{0.35},$$

where M' is the Hölder constant of the function $s^{0.35}$ over the fesible compact.

Summing up over i we obtain

$$\|\nabla_x F_n(x) - \nabla_x F_{n-1}(x)\|_1 \le 1.35M'm|\delta|^{0.35}.$$

Denote $R = \inf_{x\neq\theta_{n-1}} \|x - \theta_{n-1}\|_1 > 0$; then $\|x - \theta_{n-1}\|_1^\rho \ge R^\rho$, and condition (D.b) is satisfied for

$$B = \frac{1.35M'm\delta^{0.35}}{R^{0.35}}.$$

(D.c) *Boundedness of the gradient at the minimum point.*

Since $\theta_n = x_n^*\mathbf{1}$, we have $\nabla_x F_n(\theta_n) = \mathbf{0}$; therefore, $\|\nabla_x F_n(\theta_n, w_n)\|_{\rho+1}^{\rho+1} = 0$, so that the condition holds for $C = 0$.

(D.d) *Boundedness of change in the function at a step.*

$$F_{n-1}(x,w) = -\sum_{i=1}^m (x_i - x_{n-1}^*)^{1.35},$$

$$F_n(x,w) - F_{n-1}(x,w) = \sum_{i=1}^m \left[(x_i - x_{n-1}^*)^{1.35} - (x_i - x_n^*)^{1.35}\right],$$

$$\left|(x_i - x_n^*)^{1.35} - (x_i - x_{n-1}^*)^{1.35}\right| \le M'|\delta|^{1.35}$$

$$|F_n(x,w) - F_{n-1}(x,w)| \le mM'\delta^{1.35}.$$

Since the noise v is subject to the Pareto distribution with parameter $\beta = 1.6 > \rho + 1 = 1.35$, the moment of order 1.35 does exist, and $\mathbb{E}|v_n - v_{n-1}|^{\rho+1} \le \tilde{E} < \infty$. Therefore, for $D = 0$ and $E = (mM'\delta^{1.35} + \tilde{E})$ condition (D.d) is satisfied:

$$\mathbb{E}_{\mathcal{F}_{2n-2}}|F_{2n}(x, w_{2n}) - F_{2n-1}(x, w_{2n-1})|^{\rho+1} \le DV_{2n-2}(x) + E.$$

(E) *Boundedness of change in the observed noise.*

Consider the observation noise v_n defined via the Pareto noise:

$$v_n = \begin{cases} Z_n \mathrm{sgn}_n, & \text{symmetric noise,} \\ Z_n, & \text{asymmetric noise,} \end{cases}$$

where $Z_n \sim \mathrm{Pareto}(\beta = 1.6, \sigma = 2.0)$, $\mathrm{sgn}_n \sim \mathrm{Uniform}\{-1, 1\}$.

Condition (E) requires the fulfillment of the inequality

$$\mathbb{E}_{\mathcal{F}_{2n-2}} |v_{2n} - v_{2n-1}|^{\rho+1} \le \sigma_v^{\rho+1},$$

where $\rho + 1 = 1.5 < \beta$; i.e., the moment of order 1.5 does exist.

Since v_{2n} and v_{2n-1} are independent, the difference $v_{2n} - v_{2n-1}$ is also a random variable with finite moment of order $\rho + 1$. For the symmetric case (with alternating signs) numerical simulation over 10^6 realizations results in $\mathbb{E}\,|v_{2n} - v_{2n-1}|^{1.5} \approx 53.73$, which allows to admit $\sigma_v^{1.5} = 53.73$. Hence, condition (E) is satisfied with explicitly defined constant $\sigma_v^{\rho+1} = 53.73$.

(F) *Independence of perturbations* Δ_n.

By the construction of the search algorithm and the simulations with the RLHF-model, the vectors Δ_n are generated to be independent of all exogenous factors. The noise v_n is incorporated afterwards and does not depend on the chosen direction of perturbation.

5.5. Metrics for the Estimates and the Results of Simulations

We use a system of empirical metrics to quantify the behavior of the algorithm under the conditions of the optimum drift and the presence of noise with heavy tails. These metrics account for both the accuracy and stability of the estimates and the dynamics of adaptation to changing conditions. The metrics are selected in such a way as to cover both the steady-state characteristics of the algorithm and its behavior throughout optimization. This makes it possible to identify the strengths and weaknesses of the method in various scenarios, from stationary to rapidly changing and noisy ones.

The assessment of the average accuracy of tracking a drifting parameter on the later stages of the algorithm is performed through the average absolute error over the last iterations. The stability of the behavior of the algorithm is determined by the standard deviation of these errors. The range of fluctuations within a single run is characterized by the average minimum and maximum errors across runs, which allows for an evaluation of both the achievable potential and worst-case cases.

The dynamical characteristics of the algorithm are reflected in the metrics of the average time to achieve a given level of accuracy; this provides insight into the rate of adaptation under constraints on the error. The connection to theoretical definitions of stability is ensured through two moments of error: The moment of order ρ, which assesses convergence on average, and the corresponding asymptotic bound that normalizes the error according to the chosen order of the moment. The order used is selected based on the noise parameters to ensure the existence of the corresponding mathematical expectations.

Table 2. Basic metrics of the algorithm

Metrics	Expression
Mean absolute deviation over last 100 iterations	$\mu_{\text{last100}} = \frac{1}{100m} \sum\limits_{n=N-100}^{N-1} \sum\limits_{i=1}^{m} \lvert x_{n,i} - x_n^* \rvert$
Standard deviation of errors over last 100 iterations	$\sigma_{\text{last100}} = \sqrt{\frac{1}{100m-1} \sum\limits_{n=N-100}^{N-1} \sum\limits_{i=1}^{m} \left(\lvert x_{n,i} - x_n^* \rvert - \mu_{\text{last100}} \right)^2}$
Minimum mean deviation over the runs	$\bar{D}_{\min} = \frac{1}{m} \sum\limits_{i=1}^{m} \min_{0 \le n < N} \lvert x_{n,i} - x_n^* \rvert$
Maximum mean deviation over the runs	$\bar{D}_{\max} = \frac{1}{m} \sum\limits_{i=1}^{m} \max_{0 \le n < N} \lvert x_{n,i} - x_n^* \rvert$
Mean convergence time to threshold ϵ	$\bar{T}_\epsilon = \frac{1}{m} \sum\limits_{i=1}^{m} T_{i,\epsilon}$, $T_{i,\epsilon} = \min\{\{n \mid 0 \le n < N, \lvert x_{n,i} - x_n^* \rvert < \epsilon\} \cup \{N\}\}$
$l_{\rho+1}$-metrics of the estimation error	$\mu_{\text{def2,last100}} = \frac{1}{100} \sum\limits_{n=N-100}^{N-1} \frac{1}{m} \sum\limits_{i=1}^{m} \left(\lvert x_{n,i} - x_n^* \rvert^{\rho+1} \right)^{1/2}$

The definitions of the metrics are given in Table 2, a comparison of the results for different metrics is presented in Table 3, and their dynamics are plotted in Fig. 1.

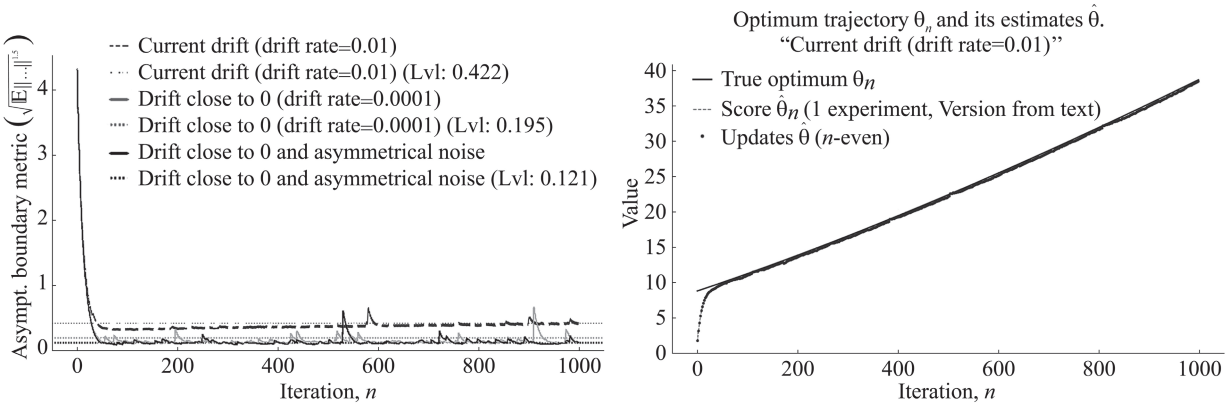

Fig. 1. Quality of the estimates. Left: Plot of the $l_{\rho+1}$ estimate of the error as function of the iteration number n; right: True trajectory of the optimum x_n^* and its estimate x_{n,i_0}.

The results of simulations presented in Table 3 testify to the influence of the environmental parameters on the performance of the algorithm (based on 1000 experiments, $\rho = 0.50$; statistics for the last 100 iterations is presented). With moderate drift ($\delta = 0.01$) and symmetric noise, the agent does track the goal, but with a noticeable average error (0.3012), moderate stability (Std = 0.1682), and rare but significant outliers (maximum 19.7897). High-order metrics take values 0.1788 and 0.4219, with convergence achieved in 20 iterations.

Table 3. Comparison of the results for different types of drift and noise

Metrics	Moderate drift $\delta = 0.01$ (symm)	Near-stationary $\delta = 0.0001$ (symm)	Asymmetric noise $\delta = 0.0001$ (asymm)
Stability metrics $E[\|x - x^*\|^{1.5}]$	0.1788	0.0524	0.0153
$l_{\rho+1}$-metrics of the estimation error	0.4219	0.1954	0.1206
Mean distance $E[\|x - x^*\|]$	0.3012	0.0520	0.0356
Standard deviation of the estimate	0.1682	0.2776	0.0989
Minimum deviation	0.0002	0.0000	0.0000
Maximum deviation	19.7897	57.1922	10.3462
Convergence time (< 1.0)	20	18	18

Decrease of drift down to $\delta = 0.0001$ (near-stationary environment) improves the mean error (0.0520); at the same time it increases instability. Namely, the standard deviation reaches the value 0.2776, and the maximum error attains the level of 57.19. This indicates an increase in sensitivity to noise with heavy tails under weakened drift.

The best results were achieved for asymmetric noise under conditions of weak drift. The error decreases to 0.0356, the variability is bounded (Std = 0.0989), and the maximum deviations are significantly lower (10.3462). Stability metrics (0.0153, 0.1206) and convergence time (18 iterations) also improve.

Hence, decrease of rate of drift increases the accuracy; however, robustness to noise depends on its type. Thus, asymmetric noise implies a better control over extreme errors, perhaps due to the specifics of gradient estimate. This effect requires further analysis.

6. SIMULATION OF THE TASK DISTRIBUTION SYSTEM IN QUEUEING SYSTEM PROBLEMS

Queueing systems, such as modern call centers, are characterized by an incoming flow of tasks having processing times that are often subject to heavy-tailed distributions [31]. This indicates

the presence of a statistically significant share of tasks that require disproportionately large processing times, distinguishing them from systems described by classical exponential or Gaussian distributions. The Pareto distribution can be thought of as a suitable model for describing such phenomena [32], since it accounts for rare but lengthy operations which affect the overall performance of the system [33].

To efficiently control such queueing systems, one has to adaptively evaluate the characteristics of the flow and time of service. Below we analyze an application of our stochastic optimization search algorithm (3) to the model of dynamical tuning the estimated expected processing time for different types of tasks; see [34] for a detailed description of the model. We use our method to iteratively optimize the parameters $\hat{\theta}_k, \hat{\theta}_m$, which are adaptive estimates of time of service for each task cluster m and for the system as a whole, k.

A simulation model of a call center was presented in [34]. Task service time in the model is generated from the Pareto distribution, with the parameters being calibrated for each cluster based on the characteristics of lognormal distributions that approximate historical data. The search algorithm (3) is used to refine the estimates $\hat{\theta}_k$, $\hat{\theta}_m$, which in turn are used for the assignment of incoming tasks to agents. The simulation shows the satisfactory performance of the method in the stochastic environment and heavy-tail nature of the task processing time.

6.1. The Model

We consider a system of agents having identical resources and performance. The load of agent i, denoted by q^i, corresponds to the number of tasks in its queue. Each task x_k is characterized by type m and the predicted execution time, calculated via the formula

$$x_{km} = \alpha\hat{\theta}^i_k + (1-\alpha)\hat{\theta}^i_m, \quad \alpha = \frac{\chi|\lambda_m|}{N_m+1},$$
$$\lambda_m(\hat{\theta}_m) = \frac{1}{N_m}\sum_{k\in N_m}\omega_k.\frac{\hat{\theta}_m - t_{km}}{\hat{\theta}_m} \to \min,$$

where $\hat{\theta}^i_k$ is the individual forecast of agent i for task k, $\hat{\theta}^i_m$ is the average predicted time to complete tasks of type m (taking local history into account), α is the weight factor that determines the contribution of the individual forecast and aggregated statistics, and χ is the convergence coefficient. The quantity λ_m characterizes the accuracy of the model prediction for problems of type m type and it is corrected when new observations are received. Here, N_m is the amount of completed tasks of type m, ω_k is the weight of the corresponding error, and t_{km} is the actual time to complete task k of type m.

Such a mechanism for calculating predictions and accuracy let the model adapt to the current quality of forecasts reducing the impact of unreliable data and strengthening the contribution of accumulated statistics with high confidence.

As a new task x_k arrives at step k, it is assigned to the following agent i_k in order to balance the load of the agents:

$$i_k = \arg\min_i \sum_j \left|\frac{q^i_k + x_{km} - q^j_k}{d_{ij}+1}\right|, \tag{5}$$

where q^j is the load of agent j, d_{ij} is the "distance" between agents (for example, based on load or physical location). The agents are connected in a fully connected topology, where each agent interacts with all the others. This ensures global communication with varying influence of agents depending on their relative proximity.

6.2. Description of the Data Set and the Primary Analysis

To demonstrate the efficiency of the developed method, a modeling of the load distribution system was conducted based on real data from an operator call center for September 2023 (over 2.3 million calls). For each inquiry, we recorded the instant of arrival, response time (wait time), the actual duration of call (ACD Time), and customer's segment.

Figure 2 presents two complementary visualizations that reveal the key characteristics of the incoming flow of some clusters and the human resources potential of the call center. The plot on the left shows the hourly intensity of tasks over the ten largest clusters, with the peak load observed for one of the clusters between 10 AM and noon. The plot on the right shows the distribution of active operators (those who received more than 50 calls in a two-hour interval), with maximum values occurring between 8 AM and 4 PM. At the same time, the personnel resources do not always keep up with the sharp fluctuations in incoming traffic. Simulation delays aggregated by time of day generally replicate the dynamics of the actual wait times, including a morning rise around 8–9 AM and an evening peak after 5 PM.

The diagrams presented in Fig. 3 display the distribution of conversation durations for 14 client segments. For the sake of anonymization, all segments have been renamed to numerical identifiers from 1 to 14 (see Table 4). The greatest variability and extended tails of the distribution are observed in segments 11 and 13, whereas segment 2 is characterized by an exclusively short duration range. Segments 3 and 14 also demonstrate a relatively narrow distribution with short medians.

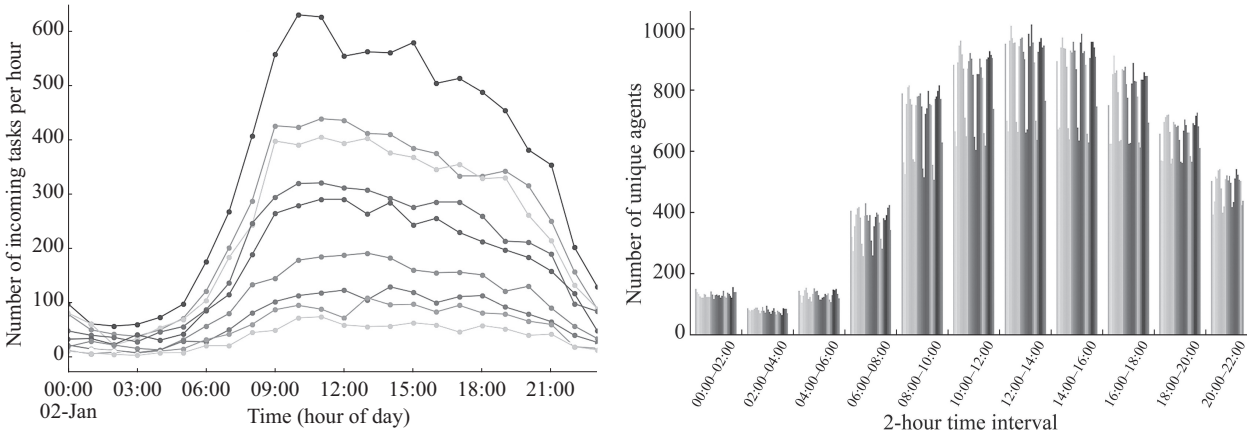

Fig. 2. Dynamics of load and the personnel time commitment. Left: Hourly intensity of tasks (top ten clusters); right: Amount of active agents over two-hour intervals.

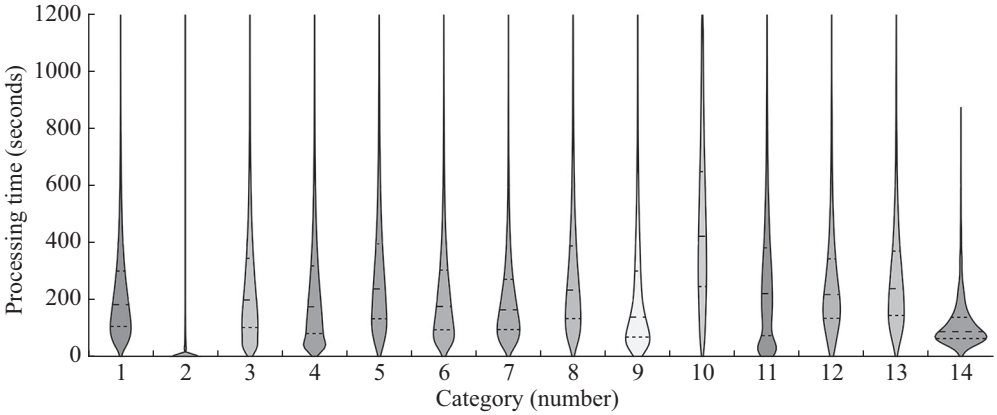

Fig. 3. Duration of calls for the top 14 clusters (max ACD = 1200 sec).

6.3. Results of Simulations

To assess the performance of the proposed method, a simulation of the call center operations was conducted using real data. The results allowed for the evaluation of both the dynamics of task wait times throughout the day and the stability of the load distribution. Figure 4 presents the results of a specific simulation session.

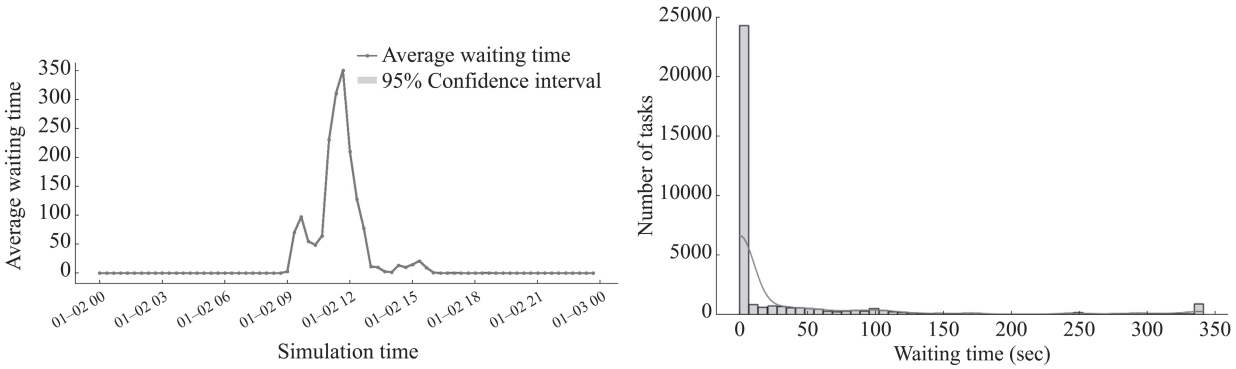

Fig. 4. The results of the simulation model: Analysis of delays in the service. Left: Mean wait time (20-minute intervals), right: The distribution of wait time (98th percentile).

The plot on the left presents the mean wait time for tasks over twenty-minute intervals showing a salient peak during work hours associated with high load. The model efficiently adapts to the changing environment; namely, after a sharp growth of delays at around noon, the mean wait time quickly decreases due to the redistribution of tasks.

The histogram on the right represents the 98-percentile distribution of the wait times. Most of the tasks have been processed in less than 50 sec, which corresponds well to the target SLA-indicators for typical scenarios.

For key clusters, Table 4 presents the values of the predicted processing time z, the amount k of completed requests, the mean actual time t_{avg}, and the maximum duration t_{max}. Clusters 1 to 14 correspond to those presented in Fig. 3, whereas cluster 0 accumulates all other segments outside of the top-14. The quantities z are seen to fit well the empirical means, despite the different statistics for different clusters, which confirms robustness properties of the adaptive prediction based on the developed search algorithm.

Table 4. Results for different key clusters

	0	1	2	3	4	5	6	7
z	143.76	163.14	3.73	174.75	147.62	196.05	159.75	151.75
k	36858	8180	6166	5764	4461	3857	2523	1707
t_{avg}	124.34	158.94	3.71	163.51	135.06	174.91	161.16	157.93
$t_{\max}$	200	200	200	200	200	200	200	200
	8	9	10	11	12	13	14	
z	219.63	151.74	321.53	144.92	147.55	191.31	49.54	
k	1329	943	906	484	265	223	44	
t_{avg}	233.06	153.30	330.74	158.08	156.68	198.90	87.34	
$t_{\max}$	200	200	200	200	200	200	44	

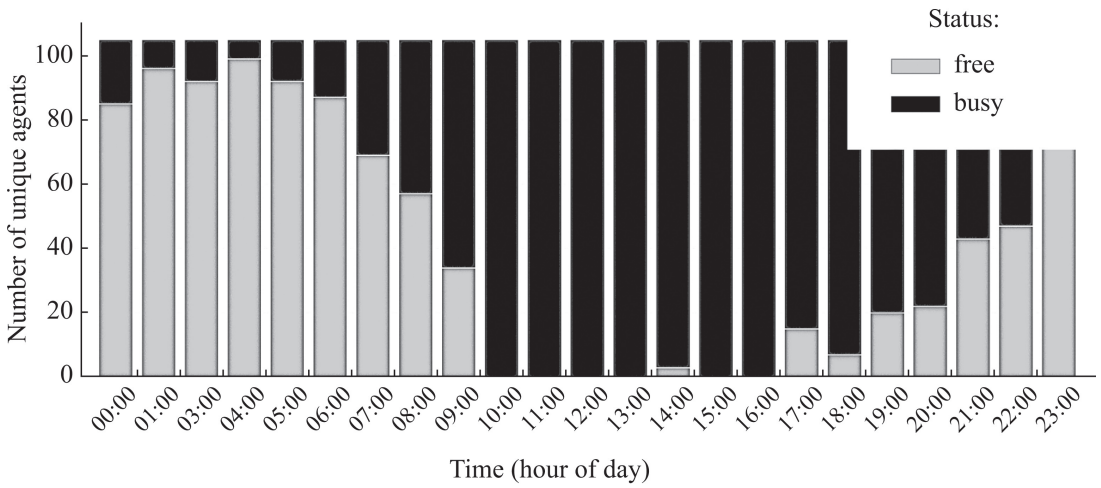

Fig. 5. Hourly load of agents: Comparison of vacant and occupied resources.

The plot presented in Fig. 5 illustrates the hourly workload of operators during the simulation. During the night and morning hours (until 8 AM), a significant portion of agents remains free; however, between 9 AM and 3 PM, there is full utilization of all resources: The number of free agents drops to zero. This coincides with the peak of incoming task flow and stresses the need for an accurate prediction of the duration of processing. In the evening and at night, the load gradually decreases, and the system returns to a balanced state.

Overall, the model demonstrates the ability to correctly adapt to the load, ensuring the mitigation of wait times and an even distribution of tasks throughout the day. The proposed approach allows for efficient resource utilization under conditions of high variability in requests and can be recommended for implementation in distributed support systems with intensive and irregular loads.

7. CONCLUSIONS

In this paper we proposed and thoroughly analyzed a method for estimating the minimum of a functional which varies in time, under conditions where the measurements are subject to noise. This method is based on the pseudogradient approach with randomization and it does not rely on the knowledge of the gradient of the objective function and uses a small number of observations at every iteration. An assumption was made on the boundedness of rate of change (drift) of the extremum of the functional. It is proved that the asymptotic estimation error is bounded from above by $\frac{L}{K}$, where L and K are found from the properties of the objective function, noise characteristics, and the parameters of the algorithm. The validity of the theoretical conclusions was confirmed by the results of numerical simulations which testified to efficient adaptation of RLHF-agents to noisy and dynamical feedback (in particular, heavy-tailed noise and different preference drift rate). The experiments showed that the search algorithm ensures the convergence of the estimates to the target value region.

According to the simulations, the steady-state error and oscillations in the estimates resulting from noise and drift are consistent with theoretical predictions about the boundedness of the asymptotic error. Furthermore, the proposed method was tested through simulations based on real data from an operator call center. Use of empirical characteristics of the flow of requests and processing times demonstrated reliable applicability of the algorithm in dynamical load distribution problems and in predicting service parameters in real service systems.

FUNDING

The theoretical results presented in Sections 1–4 were obtained in the Institute for Problems in Mechanical Engineering, Russian Academy of Science, under the financial support of the Russian Science Foundation (project No. 23-41-00060); the applied part presented in Sections 5 and 6 was implemented under the financial support of the Saint Petersburg State University, project No. 121061000159-6.

APPENDIX

Proof of Theorem 1. Denote the estimation error by $\mathrm{err}_n = \hat{\theta}_n - \theta_n$.

Step 1: Recursive relation for the estimation error. By algorithm (3) we have

$$\hat{\theta}_{2n} = \hat{\theta}_{2n-2} - \frac{\alpha}{2\beta}\Delta_{2n}(y_{2n} - y_{2n-1}),$$

hence,

$$\mathrm{err}_{2n} = \mathrm{err}_{2n-2} - \underbrace{(\theta_{2n} - \theta_{2n-2})}_{\mathrm{drift}_n} - \underbrace{\frac{\alpha}{2\beta}\Delta_{2n}(y_{2n} - y_{2n-1})}_{\mathrm{step}_n}.$$

Step 2: Recursive relation for the estimate of the Lyapunov function $V(x)$. For the vectors $a = \hat{\theta}_{2n-2}$ and $b = \mathrm{drift}_n + \mathrm{step}_n$ we have

$$V_{2n}(\hat{\theta}_{2n}) = V_{2n-2}(\hat{\theta}_{2n} - \mathrm{drift}_n) = V_{2n-2}(a - b) = \|a - b - \theta_{2n-2}\|_{\rho+1}^{\rho+1}$$

by definition. Using the Taylor series expansion of the function $V_{2n-2}(a - b)$ at the point a in the direction $-b$, we obtain

$$V_{2n-2}(a - b) = V_{2n-2}(a) - \langle \nabla V_{2n-2}(a - \delta b), b\rangle, \quad \delta \in [0, 1], \tag{A.1}$$

noting that the gradient $\nabla V_{2n-2}(a - \delta b)$ is computed according to

$$\nabla V_{2n-2}(a - \delta b) = (\rho + 1) \cdot \mathrm{sgn}(\delta) \odot |a - \theta_{2n-2} - \delta b|^{\rho},$$

where $\mathrm{sgn}_n^{(i)}(\delta) = 0$ or ± 1 depending on the sign of the ith component of the vector $a - \theta_{2n-2} - \delta b$; $|a - \theta_{2n-2} - \delta b|^{\rho}$ is the vector of the absolute values of the components of $a - \theta_{2n-2} - \delta b$ to the power ρ, and $\odot$ denotes the componentwise multiplication. The second term in (A.1) can be evaluated as

$$-\langle \nabla V_{2n-2}(a - \delta b), b\rangle \leqslant -\langle (\rho + 1) \cdot \mathrm{sgn}(0) \odot |a - \theta_{2n-2}|^{\rho}, b\rangle + 2^{1-\rho}\delta^{\rho}\|b\|_{\rho+1}^{\rho+1} \leqslant$$

$$-\langle \nabla V_{2n-2}(a), b\rangle + 2^{1-\rho}\|b\|_{\rho+1}^{\rho+1}$$

(see proof of Theorem 1 in [24], p. 93).

Keeping the considerations above and using condition (D.a), we have

$$V_{2n}(\hat{\theta}_{2n}) \leqslant V_{2n-2}(\hat{\theta}_{2n-2}) - \langle \nabla V_{2n-2}(\hat{\theta}_{2n-2}), \mathrm{drift}_n + \mathrm{step}_n\rangle + 2(A^{\rho+1} + \|\mathrm{step}_n\|_{\rho+1}^{\rho+1}). \tag{A.2}$$

Step 3: Expansion of the correcting term. According to the model of observations, represent the term step_n as the sum

$$\mathrm{step}_n = \underbrace{\frac{\alpha}{2\beta}\Delta_n\Big(F_{2n}(x_{2n}, w_{2n}) - F_{2n-1}(x_{2n-1}, w_{2n-1})\Big)}_{\text{almost pseudogradient term}} + \underbrace{\frac{\alpha}{2\beta}\Delta_n(v_{2n} - v_{2n-1})}_{\text{noise}}.$$

a. Almost pseudogradient term. Denote $n^{\pm} = 2n - \frac{1}{2} \pm \frac{1}{2}$.

Using the Taylor formula, we first add and subtract the quantity $\sum_{n^{\pm}} \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle$, then the quantity $\langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle$, and finally $\langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle$, to obtain

$$\begin{aligned}
\sum_{n^{\pm}} \pm F_{n^{\pm}}(x_{n^{\pm}}, w_{n^{\pm}}) &= \sum_{n^{\pm}} \pm F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}) + \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2} \pm \delta_{n^{\pm}}\beta\Delta_n, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&= \sum_{n^{\pm}} \pm F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}) + \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2} \pm \delta_{n^{\pm}}\beta\Delta_n, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&= \sum_{n^{\pm}} \pm F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}) + \langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2} \pm \delta_{n^{\pm}}\beta\Delta_n, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle,
\end{aligned}$$

where $\delta_{n^{\pm}} \in [0, 1]$.

Now take the conditional mathematical expectation with respect to the σ-algebra $\mathcal{F}_{2n-2}$. By condition (F), the vectors Δ_n are independent of $w_{n^{\pm}}$ and the σ-algebra $\mathcal{F}_{2n-2}$, hence we have

$$\frac{\alpha}{2\beta}\mathbb{E}_{\mathcal{F}_{2n-2}} \left\{ \Delta_n \sum_{n^{\pm}} \pm F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}) \right\} = 0,$$

since Δ_n are centered, and

$$\frac{\alpha}{2\beta}\mathbb{E}_{\mathcal{F}_{2n-2}} \left\{ \Delta_n \sum_{n^{\pm}} \langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \right\} = 0,$$

since $\mathbb{E}_{\mathcal{F}_{2n-2}}\{\nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}})\} = \nabla_x f_{2n-2}(\theta_{2n-1})$ by condition (C), and the gradient of $f_{2n-2}(\cdot)$ at the minimum point θ_{2n-2} is equal to zero.

As a result, by condition (C) we obtain

$$\mathbb{E}_{\mathcal{F}_{2n-2}} \left\{ \frac{\alpha}{2\beta}\Delta_n \sum_{n^{\pm}} \pm F_{n^{\pm}}(x_{n^{\pm}}, w_{n^{\pm}}) \right\} = \frac{\alpha}{d}\nabla f_{2n}(\hat{\theta}_{2n-2}) + \frac{\alpha}{2\beta}\mathbb{E}_{\mathcal{F}_{2n-2}}\mathrm{corr}_n$$

for the almost pseudogradient term, where

$$\begin{aligned}
\mathrm{corr}_n = \sum_{n^{\pm}} &\langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2} \pm \delta_{n^{\pm}}\beta\Delta_n, w_{n^{\pm}}), \beta\Delta_n \rangle - \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle.
\end{aligned}$$

By conditiond (B) and (D.b), the following estimate holds:

$$\begin{aligned}
\|\mathrm{corr}_n\| &\leqslant M\beta^{\rho}\|\Delta_n\| \left(2\|\Delta_n\|^{\rho} + 2\|\hat{\theta}_{2n-2} - \theta_{2n-2}\|^{\rho}) + 3B\|\hat{\theta}_{2n-2} - \theta_{2n-2}\|^{\rho} \right) \\
&= 2M\beta^{\rho} + (2 + 3B)\|\hat{\theta}_{2n-2} - \theta_{2n-2}\|^{\rho}.
\end{aligned}$$

b. Noise. Take the conditional mathematical expectation with respect to the σ-algebra $\mathcal{F}_{2n-2}$. By the independence of Δ_n on v_{2n}, v_{2n-1} and $\mathcal{F}_{2n-2}$, we obtain

$$\mathbb{E}_{\mathcal{F}_{2n-2}}\left\{\frac{\alpha}{2\beta}\Delta_n(v_{2n}-v_{2n-1})\right\}=0.$$

c. The final estimate of the second term on the right-hand side of Ineq (A.2). By the strong convexity (see condition (A)), we obtain

$$\begin{aligned}
-\mathbb{E}_{\mathcal{F}_{2n-2}}\{\langle\nabla V_{2n-2}(\hat\theta_{2n-2}),\mathrm{drift}_n+\mathrm{step}_n\rangle\} &\leqslant -\frac{\mu\alpha}{d}V_{2n-2}(\hat\theta_{2n-2})\\
-\frac{\alpha}{2\beta}\mathbb{E}_{\mathcal{F}_{2n-2}}\langle\nabla V_{2n-2}(\hat\theta_{2n-2}),\mathrm{drift}_n+\mathrm{corr}_n\rangle &\leqslant -\frac{\mu\alpha}{d}V_{2n-2}(\hat\theta_{2n-2})\\
+2(A+\alpha M\beta^{\rho-1})^2+\left(2+\frac{\alpha}{2\beta}(2+3B)\right)\sum_{i=1}^{d}|\hat\theta^i_{2n-2}-\theta^i_{2n-2}|^{2\rho}&\\
\leqslant -\frac{\mu\alpha}{d}V_{2n-2}(\hat\theta_{2n-2})+\varepsilon V_{2n-2}(\hat\theta_{2n-2})+c_1,&
\end{aligned}$$

where $\varepsilon>0$ and

$$c_1=2(A+\alpha M\beta^{\rho-1})^2+\varepsilon^{\rho-1}\left(2+\frac{\alpha}{2\beta}(2+3B)\right)^{\frac{1-\rho}{\rho+1}}.$$

Step 4: Estimate of the third term on the right-hand side of inequality (A.2). Similarly to the derivations at Step 3 above, the term step_n can be represented as

$$\mathrm{step}_n=\frac{\alpha}{2\beta}\Delta_n\sum_{i=1}^{8}a_i,$$

where

- $a_1=\sum_{n^\pm}\pm F_{n^\pm}(\hat\theta_{2n-2},w_{n^\pm})$;
- $a_2=a_3=\langle\nabla_x F_{n^\pm}(\hat\theta_{2n-2}\pm\delta_{n^\pm}\beta\Delta_n,w_{n^\pm}),\beta\Delta_n\rangle-\langle\nabla_x F_{n^\pm}(\hat\theta_{2n-2},w_{n^\pm}),\beta\Delta_n\rangle$;
- $a_4=a_5=\langle\nabla_x F_{n^\pm}(\hat\theta_{2n-2},w_{n^\pm}),\beta\Delta_n\rangle-\langle\nabla_x F_{n^\pm}(\theta_{n^\pm},w_{n^\pm}),\beta\Delta_n\rangle$;
- $a_6=a_7=\langle\nabla_x F_{n^\pm}(\theta_{n^\pm},w_{n^\pm}),\beta\Delta_n\rangle$;
- $a_8=v_{2n}-v_{2n-1}$.

Respectively, we have

- for a_1: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_1|^{\rho+1}\leqslant DV_{2n-2}(\hat\theta_{2n-2})+E$ by condition (D.d);
- for a_2,a_3: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_i|^{\rho+1}\leqslant M^{\rho+1}\beta^{2\rho+2}$, $i=2,3$, by condition (B);
- for a_4,a_5: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_i|^{\rho+1}\leqslant(M\beta\|\hat\theta_{2n-2}-\theta_{n^\pm}\|_2^\rho)^{\rho+1}\leqslant M^{\rho+1}\beta^{\rho+1}d^{\frac{\rho-1}{2}}V_{n^\pm}(\hat\theta_{2n-2})$, $i=4,5$, by condition (B) and Jensen's inequality;
- for a_6,a_7: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_i|^{\rho+1}\leqslant C$, $i=6,7$, by condition (D.c);
- for a_8: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_8|^{\rho+1}\leqslant\sigma_v^{\rho+1}$ by condition (E).

Overall, by Jensen's inequality we obtain

$$\left(\frac{\sum_{i=1}^{8}|a_i|}{8}\right)^{\rho+1}\leqslant\frac{1}{8}\sum_{i=1}^{8}|a_i|^{\rho+1},$$

so that

$$2A^{\rho+1} + 2\mathbb{E}_{\mathcal{F}_{2n-2}}\|\mathrm{step}_n\|_{\rho+1}^{\rho+1} \leqslant 2A^{\rho+1} + 2\cdot 8^{\rho}\left(\frac{\alpha}{2\beta}\right)^{\rho+1}\sum_{i=1}^{7}|a_i|^{\rho+1}$$

$$\leqslant 2A^{\rho+1} + 2^{2\rho}\alpha^{\rho+1}\left(2M^{\rho+1}(\beta^{\rho+1} + d^{\frac{\rho-1}{2}}\sum_{n^{\pm}}V_{n^{\pm}}(\hat{\theta}_{2n-2})) + \frac{2C + DV_{2n-2}(\hat{\theta}_{2n-2}) + E + \sigma_v^{\rho+1}}{\beta^{\rho+1}}\right)$$

$$\leqslant c_2\alpha^{\rho+1}V_{2n-2}(\hat{\theta}_{2n-2}) + c_3,$$

where

$$c_2 = 2^{3\rho+1}M^{\rho+1}\left(d^{\frac{\rho-1}{2}} + \frac{D}{\beta^{\rho+1}}\right)$$

and

$$c_3 = 2A^{\rho+1} + 2^{2\rho}\alpha^{\rho+1}\left(2M^{\rho+1}(\beta^{\rho+1} + 3\cdot 2^{\rho}d^{\frac{\rho-1}{2}}) + \frac{E + 2C + \sigma_v^{\rho+1}}{\beta^{\rho+1}}\right).$$

Step 5: Shaping the recursive inequality for the Lyapunov function. Collecting all estimates obtained above, we arrive at

$$V_{2n} \leqslant V_{2n-2} - (\mu\alpha d^{-1} - \varepsilon - c_2\alpha^{\rho+1})V_{2n-2} + c_1 + c_3.$$

Introducing the notation

$$K = 1 - \mu\alpha d^{-1} + \varepsilon + c_2\alpha^{\rho+1}, \quad L = c_1 + c_3,$$

we obtain

$$V_{2n} \leqslant (1-K)V_{2n-2} + L.$$

By choosing α and ε sufficiently small, the inequality $K < 1$ can be achieved, which implies the assertion of Theorem 1. □

REFERENCES

1. Polyak, B.T., *Introduction to Optimization*, New York, Optimization Software, 1987.
2. Polyak, B.T. and Tsypkin, Ya.Z., Pseudogradient Adaptation and Training Algorithms, *Autom. Remote Control*, 1973, vol. 34, no. 3, pp. 377–397.
3. Polyak, B.T. and Tsypkin, Ya.Z., Adaptive Estimation Algorithms (Convergence, Optimality, Stability), *Autom. Remote Control*, 1979, vol. 40, no. 3, pp. 378–389.
4. Poliak, B.T. and Tsypkin, J.Z., Optimal Pseudogradient Adaptation Algorithms, *Autom. Remote Control*, 1981, vol. 41, no. 8, pp. 1101–1110.
5. Polyak, B.T., Some Methods of Speeding up the Convergence of Iteration Methods, *USSR Comput. Math. Math. Phys.*, 1964, vol. 4, no. 5, pp. 1–17.
6. Polyak, B.T., New Method of Stochastic Approximation Type, *Autom. Remote Control*, 1990, vol. 51, no. 7, pp. 937–946.
7. Polyak, B.T. and Yuditsky, A.B., Acceleration of Stochastic Approximation by Averaging, *SIAM J. Control Optim.*, 1992, vol. 30, no. 4, pp. 838–855.
8. Polyak, B.T., Convergence and Convergence Rate of Iterative Stochastic Algorithms. I. General Case, *Autom. Remote Control*, 1976, vol. 37, no. 12, pp. 1858–1868.
9. Polyak, B.T., Convergence and Convergence Rate of Iterative Stochastic Algorithms. II. The Linear Case, *Autom. Remote Control*, 1977, vol. 38, no. 4, pp. 537–542.
10. Polyak, B.T. and Tsybakov, A.B., Optimal Order of Accuracy for Search Algorithms in Stochastic Optimization, *Problems Inform. Transmiss.*, 1990, vol. 26, no. 2, pp. 126–133.
11. Rastrigin, L.A., *Statisticheskie metody poiska* (Statistical Search Methods), Moscow, Nauka, 1968.
12. Granichin, O.N., Stochastic Approximation with Input Perturbation under Dependent Observation Noises, *Vestn. Leningr. Univ.*, 1989, pp. 27–31.

13. Spall, J.C., Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, *IEEE Trans. Autom. Control*, 1992, vol. 37, no. 3, pp. 332–341.
14. Spall, J.C., A One-measurement Form of Simultaneous Perturbation Stochastic Approximation, *Automatica*, 1997, vol. 33, no. 1, pp. 109–112.
15. Granichin, O.N. and Polyak, B.T., *Randomizirovannye algoritmy otsenivaniya i optimizatsii pri pochti proizvol'nykh pomekhakh* (Randomized Algorithms for Estimation and Optimization under Almost Arbitrary Disturbances), Moscow: Nauka, 2003.
16. Granichin, O., Volkovich, V., and Toledano-Kitai, D., *Randomized Algorithms in Automatic Control and Data Mining*, Springer, 2015.
17. Popkov, A.Yu., Gradient Methods for Nonstationary Unconstrained Optimization Problems, *Autom. Remote Control*, 2005, vol. 66, no. 6, pp. 883–891.
18. Kiefer, J. and Wolfowitz, J., Stochastic Estimation of the Maximum of a Regression Function, *Ann. Math. Stat.*, 1952, vol. 23, no. 3, pp. 462–466.
19. Vakhitov, A.T., Granichin, O.N., and Gurevich, L.S., Algorithm for Stochastic Approximation with Trial Input Perturbation in the Nonstationary Problem of Optimization, *Autom. Remote Control*, 2009, vol. 70, no. 11, pp. 1827–1835.
20. Granichin, O. and Amelina, N., Simultaneous Perturbation Stochastic Approximation for Tracking under Unknown but Bounded Disturbances, *IEEE Trans. Autom. Control*, 2015, vol. 60, no. 6, pp. 1653–1658.
21. Shibaev, I.A., *Bezgradientnye metody optimizatsii dlya funktsii s gel'derovym gradientom* (Gradient-free Optimization Methods for Functions with Hölder Gradient), PhD Dissertation, MIPT, 2024, Dolgoprudny.
22. Shibaev, I., Dvurechensky, P., and Gasnikov, A., Zeroth-order Methods for Noisy Hölder-gradient Functions, *Optim. Lett.*, 2022, vol. 16, pp. 2123–2143.
23. Mandelbrot, B., New Methods in Statistical Economics, *J. Polit. Econ.*, 1963, vol. 71, no. 5, pp. 421–440.
24. Vakhitov, A.T., Granichin, O.N., and Sysoev, S.S., A Randomized Stochastic Optimization Algorithm: Its Estimation Accuracy, *Autom. Remote Control*, 2006, vol. 67, no. 4, pp. 589–597.
25. Granichin, O.N., Stochastic Approximation Search Algorithms with Randomization at the Input, *Autom. Remote Control*, 2015, vol. 76, no. 5, pp. 762–775.
26. Min, T. et al., Understanding Impact of Human Feedback via Influence Functions, *ArXiv preprint arXiv:2501.05790*, 2025.
27. Shen, W. et al., Loose Lips Sink Ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback, *Find. Assoc. Comput. Linguist.: EMNLP*, 2023, pp. 2859–2873.
28. Christiano, P.F. et al., Deep Reinforcement Learning from Human Preferences, *Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–9.
29. Stiennon, N. et al., Learning to Summarize with Human Feedback, *Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 3008–3021.
30. Ouyang, L. et al., Training Language Models to Follow Instructions with Human Feedback, *Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 27730–27744.
31. Gans, N., Koole, G., and Mandelbaum, A., Telephone Call Centers: Tutorial, Review, and Research Prospects, *Manuf. Serv. Oper. Manag.*, 2003, vol. 5, no. 2, pp. 79–141.
32. Anderson, C., *The Long Tail: Why the Future of Business is Selling Less of More*, Hyperion, 2006.
33. Goel, S., Broder, A., Gabrilovich, E., and Pang, B., Anatomy of the Long Tail: Ordinary People with Extraordinary Tastes, *Proc. 3rd ACM Int. Conf. Web Search Data Min. (WSDM)*, New York, Feb. 4–6, 2010, pp. 201–210.
34. Akinfiev, I. and Tarasova, E., Cluster-Aware LVP: Enhancing Task Allocation with Growth Dynamics, *Proc. 15th IFAC Workshop Adapt. Learn. Control Syst. (ALCOS)*, Mexico City, Jul. 2–4, 2025.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board