Algorithms of Robust Stochastic Optimization Based on Mirror Descent Method

Alexander NAZIN

V.A. Trapeznikov Institute of Control Sciences RAS,

Moscow, Russia, nazine@ipu.ru

jointly with A. Nemirovsky, A. Tsybakov, and A. Juditsky

Workshop on Optimization and Applications
School of Applied Mathematics, MPTI

September 27, 2019

Outline

1. Brief Introduction (see [1][a] for the details)
2. Notation and Definitions
3. Assumptions
4. Accuracy bounds for Algorithm RSMD
5. Robust Confidence Bounds for Stoch. Optim. Methods
6. Conclusion
7. References

[a]Published in the special issue devoted to 100th anniversary of Ya.Z.Tsypkin; see editorial paper: B.T.Polyak, A.V.Nazin, Avtomat. i Telemekh., 2019, 9, 6–8.

1 Introduction

Convex stochastic composite optimization problem:

$$\min_{x \in X} F(x), \qquad F(x) = \mathbf{E}\{\Phi(x, \omega)\} + \psi(x), \tag{1}$$

where X is a compact convex subset of a (primal) fin.-dim. real vector space E with a norm $\|\cdot\|$,

$\omega \in \Omega$ is a random variable with distribution P,

$\Phi : X \times \Omega \to \mathbf{R}$, and

ψ is convex and continuous.

Suppose a well defined and finite expectation

$$\phi(x) := \mathbf{E}\{\Phi(x, \omega)\} = \int_{\Omega} \Phi(x, \omega) dP(\omega), \quad \forall x \in X,$$

being convex and differentiable function of x. Therefore, problem (1) has a solution with optimal value

$$F_* = \min_{x \in X} F(x).$$

It is assumed that a mechanism (oracle) is available, which for the input point $(x, \omega) \in X \times \Omega$ returns a stochastic gradient $G(x, \omega)$ satisfying, $\forall\, x \in X$,

$$\mathbf{E}\{G(x,\omega)\} = \nabla\phi(x), \quad \mathbf{E}\{\|G(x,\omega) - \nabla\phi(x)\|_*^2\} \le \sigma^2, \ (2)$$

where $\|\cdot\|_*$ is the conjugate norm, and $\sigma > 0$ is a constant.

The goal is to construct $(1-\alpha)$-reliable approximate solutions $\widehat{x}_N$ of problem (1), based on N calls of stochastic oracle and satisfying condition

$$\mathsf{P}\{F(\widehat{x}_N) - F^* \le \delta_N(\alpha)\} \ge 1 - \alpha, \quad \forall\, \alpha \in (0,1), \qquad (3)$$

with as small as possible $\delta_N(\alpha) > 0$.

Remark: Here in [1], the bounds (3) are founded with $\delta_N(\alpha)$ of order $\sqrt{\ln(1/\alpha)/N}$. Such bounds are often called sub-Gaussian confidential bounds, and standard results for stochastic optimization algorithms imply the finiteness of the exponential or subexponential moments of the stochastic "noise" of oracle $G(x,\omega) - \nabla\phi(x)$ (cf. [2, 3, 4]). Here, the robust stochastic algorithms are constructed that satisfy sub-Gaussian bounds of type (3) with a significantly less restrictive condition (2).

Recall that the notion of robustness of statistical decision making procedures was introduced by J. Tukey [5] and P. Huber [6, 7, 8] at years of 1960, which led to the study of robust stochastic approximation algorithms. In particular, at the 1970s–1980s, the algorithms that are resistant to wide classes of noise distributions were proposed for problems of stochastic optimization and parametric identification. Their asymptotic (with increasing sample size) properties have been well studied, see, for example, [9, 10, 11, 12, 13, 14, 15, 16, 17] and references therein.

An important contribution to the development of the robust approach was made by Ya.Z. Tsypkin. Thus, a significant place in the monographs [18, 19] is devoted to the study of iterative robust identification algorithms.

The interest in the issues of robust estimation resumed in the 2010s due to the need to develop statistical procedures that are resistant to noise with heavy tails in large-scale problems. Some recent works
[20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30] are related to the development of the method of average median (*median-of-means*) [31] for constructing estimates that satisfy the sub-Gaussian confidence bounds for noise with heavy tails. So, in [28], a median-of-means procedure was used to construct $(1-\alpha)$-reliable version of stochastic approximation with averaging ("batch" algorithm) in the formulation of stochastic optimization similar to (1).

Other original approaches were developed [32, 33, 34, 35, 36], in particular, using the *geometric median* for robust estimation of signals and covariance matrices with sub-Gaussian guarantees [35, 36]. Also renewed interest in robust iterative algorithms. Thus, it was shown that the robustness of stochastic approximation algorithms can be enhanced by using the geometric median of stochastic gradients [37, 38]. Another variant of the stochastic approximation procedure for calculating the geometric median was studied in [39, 40], where the special structure of the problem — the limitation of stochastic gradients — allowed us to construct $(1-\alpha)$-reliable estimates with an extremely weak assumption about tails of noise distribution.

Here, we discuss an approach to the construction of robust stochastic algorithms by truncating stochastic gradients. It is shown that this method satisfies the sub-Gaussian confidence bounds. Sections 2 and 3 define the main components of the optimization problem under consideration. Section 4 defines the robust stochastic mirror descent algorithm and establishes confidence bounds for it. Section 5 builds robust accuracy estimates for general stochastic algorithms. Finally, Section 5.1 establishes robust confidence bounds for problems in which F has a quadratic growth. The appendix in [1] contains all the proofs.

2 Notation and Definitions

Let E be finite-dimensional real vector space with norm $\|\cdot\|$, and E^* be conjugate to space E. Denote by $\langle s, x\rangle$ the value of linear function $s \in E^*$ at point $x \in E$ and by $\|\cdot\|_*$ the conjugate to norm $\|\cdot\|$ on E^*, i.e.

$$\|s\|_* = \max_x\{\langle s, x\rangle : \ \|x\| \leq 1\}, \quad s \in E^*.$$

Consider continuous convex function $\theta : B \to \mathbf{R}$ on a unit ball

$$B = \{x \in E : \ \|x\| \leq 1\},$$

possessing the following property:

$$\langle \theta'(x) - \theta'(x'), x - x'\rangle \geq \|x - x'\|^2, \quad \forall x, x' \in B, \qquad (4)$$

where $\theta'(\cdot)$ — continuous in $B^o = \{x \in B : \partial\theta(x) \neq \emptyset\}$ subgradient version $\theta(\cdot)$, and $\partial\theta(x)$ — subdifferential of function $\theta(\cdot)$ at point x, i.e. the set of all subgradients at a given point. In other words, function $\theta(\cdot)$ is strongly convex on B with coefficient 1 relative to norm $\|\cdot\|$. We will call $\theta(\cdot)$ the normalized proxy function. Examples of such functions are:

- $\theta(x) = \frac{1}{2}\|x\|_2^2$ for $(E, \|\cdot\|) = (\mathbf{R}^n, \|\cdot\|_2)$;
- $\theta(x) = 2\mathrm{e}(\ln n)\|x\|_p^p$ at $p = p(n) := 1 + \frac{1}{2\ln n}$ for $(E, \|\cdot\|) = (\mathbf{R}^n, \|\cdot\|_1)$;
- $\theta(x) = 4\mathrm{e}(\ln n)\sum_{i=1}^n |\lambda_i(x)|^p$ at $p = p(n)$, for $E = S_n$, where S_n — space of symmetric $n \times n$ matrices, equipped with nuclear norm $\|x\| = \sum_{i=1}^n |\lambda_i(x)|$, and

$\lambda_i(x)$ — eigenvalues of matrix x.

Here and further, $\|\cdot\|_p$ is denoted by ℓ_p-norm in $\mathbf{R}^n$, $p \geq 1$.
Without loss of generality, we will assume further that

$$0 = \operatorname*{argmin}_{x \in B} \theta(x).$$

Also, we introduce the notation

$$\Theta = \max_{x \in B} \theta(x) - \min_{x \in B} \theta(x) \geq \tfrac{1}{2}.$$

Now let X be a convex compact subset in E, and let
$x_0 \in X$ and $R > 0$ be such that $\max_{x \in X} \|x - x_0\| \leq R$.
We supply X with proxy function

$$\vartheta(x) = R^2 \theta\left(\frac{x - x_0}{R}\right).$$

Note that $\vartheta(\cdot)$ is strongly convex with coefficient 1 and

$$\max_{x\in X}\vartheta(x)-\min_{x\in X}\vartheta(x)\le R^2\Theta.$$

Let $D:=\max_{x,x'\in X}\|x-x'\|$ be diameter of set X. Then $D\le 2R$.

You will also need the Bregman diversion here

$$V_x(z)=\vartheta(z)-\vartheta(x)-\langle\vartheta'(x),z-x\rangle,\quad\forall\, z,x\in X.$$

In the following, C and C' denote positive numeric constants, not necessarily the same in different cases.

3 Assumptions

Consider a convex stochastic composite optimization problem (1) on a convex compact set $X \subset E$. Assume in the following that function

$$\phi(x) = \mathbf{E}\{\Phi(x, \omega)\}$$

is convex on X, differentiable at each set point X, and its gradient satisfies Lipschitz condition

$$\|\nabla\phi(x') - \nabla\phi(x)\|_* \leq L\|x - x'\|, \qquad \forall\, x, x' \in X. \qquad (5)$$

Assume also, that function ψ is convex and continuous. Further assume that stochastic oracle is available that, having an input point $(x, \omega) \in X \times \Omega$, returns random

vector $G(x, \omega)$, satisfies the conditions (2). In addition, it is assumed that for any $a \in E^*$ and $\beta > 0$ an exact solution is available to the minimum problem

$$\min_{z \in X} \{\langle a, z \rangle + \psi(z) + \beta \vartheta(z)\}.$$

The assumption is fulfilled for typical ψ penalty functions, such as convex power functions of ℓ_p-norm (if X is a convex compact in $\mathbf{R}^n$) or negative entropy $\psi(x) = \kappa \sum_{j=1}^n x_j \ln x_j$, $\kappa > 0$ (if X – standard simplex in $\mathbf{R}^n$). Finally, it is assumed that vector $g(\bar{x})$ is available, where $\bar{x} \in X$ is a point in set X, such that

$$\|g(\bar{x}) - \nabla \phi(\bar{x})\|_* \leq \upsilon \sigma \tag{6}$$

with a constant $\upsilon \geq 0$. This assumption is motivated as

follows.

First, if a *priori* is known, that global minimum of function ϕ attains at an interior point x_ϕ for set X (that holds often in statistical applications of stochastic approximation), we have $\nabla\phi(x_\phi) = 0$. Therefore, chosing $\bar{x} = x_\phi$, one can put $g(\bar{x}) = 0$, and assumption (6) holds automatically with $\upsilon = 0$.

Second, in general, one can chose $\bar{x}$ in the capacity of a point for set X and $g(\bar{x})$ in the capacity of geometric median for stochastic gradients $G(\bar{x}, \omega_i)$, $i = 1, \ldots, m$, over m oracle calls. From [35] follows, that if m is of order $\ln(\varepsilon^{-1})$ with some sufficiently small $\varepsilon > 0$, then

$$\mathsf{P}\{\|g(\bar{x}) - \nabla\phi(\bar{x})\|_* > \upsilon\sigma\} \leq \varepsilon. \tag{7}$$

Thus, the confidence bounds obtained below will remain valid up to the ε correction in the probability of deviations.

4 Accuracy bounds for Algorithm RSMD

Further everywhere it is considered that the assumptions formulated in the section 3 are fulfilled. We introduce a composite proximal transform

$$\begin{aligned} \mathrm{Prox}_{\beta,x}(\xi) &:= \operatorname*{argmin}_{z\in X}\left\{\langle \xi, z\rangle + \psi(z) + \beta V_x(z)\right\} = \qquad (8)\\ &= \operatorname*{argmin}_{z\in X}\left\{\langle \xi - \beta\vartheta'(x), z\rangle + \psi(z) + \beta\vartheta(z)\right\}, \end{aligned}$$

where $\beta > 0$ is a setting parameter.

For $i = 1, 2, \ldots,$ by the following recurrent relations, define

the algorithm of *Robust Stochastic Mirror Descent* (RSMD):

$$x_i = \mathrm{Prox}_{\beta_{i-1}, x_{i-1}}(y_i), \quad x_0 \in X, \tag{9}$$

$$y_i = \begin{cases} G(x_{i-1}, \omega_i), & \text{if } \|G(x_{i-1}, \omega_i) - g(\bar{x})\|_* \\ & \leq L\|\bar{x} - x_{i-1}\| + \lambda + \upsilon\sigma, \\ g(\bar{x}), & \text{otherwise.} \end{cases} \tag{10}$$

Here $\beta_i > 0$, $i = 0, 1, \ldots$ and $\lambda > 0$ are setting parameters, which will be defined below, and $\omega_1, \omega_2, \ldots$ are independent identically distributed (i.i.d.) realizations of a random variable ω, corresponding to the oracle calls at each step of the algorithm.

The approximate solution of the problem (1) after N iterations is defined as the weighted average

$$\widehat{x}_N = \left[\sum_{i=1}^{N} \beta_{i-1}^{-1}\right]^{-1} \sum_{i=1}^{N} \beta_{i-1}^{-1} x_i. \qquad (11)$$

In the case when the global minimum of function ϕ is reached at the interrior point of set X and $\upsilon = 0$, the definition of (10) is simplified. In this case, replacing $\|\bar{x} - x_{i-1}\|$ on the upper bound D and putting $\upsilon = 0$ and $g(\upsilon) = 0$ in (10), the truncated stochastic gradient is calculated by formula

$$y_i = \begin{cases} G(x_{i-1}, \omega_i), & \text{if } \|G(x_{i-1}, \omega_i)\|_* \leq LD + \lambda, \\ 0, & \text{otherwise.} \end{cases}$$

The following statement describes some useful properties of mirror descent recursion (9). Denote:

$$\xi_i = y_i - \nabla\phi(x_{i-1})$$

and

$$\varepsilon(x^N, z) = \sum_{i=1}^{N} \beta_{i-1}^{-1}[\langle \nabla\phi(x_{i-1}), x_i - z\rangle + \psi(x_i) - \psi(z)] \\ + \tfrac{1}{2} V_{x_{i-1}}(x_i), \tag{12}$$

where $x^N = (x_0, \ldots, x_N)$.

Proposition 1 *Let be $\beta_i \geq 2L$ for all $i = 0, 1, ...,$ and let $\widehat{x}_N$ be defined in (11), where x_i are iterations (9) for any y_i, not necessarily given by (10). Then for any $z \in X$ we have*

$$\left[\sum_{i=1}^{N} \beta_{i-1}^{-1}\right] [F(\widehat{x}_N) - F(z)] \quad \leq \quad \sum_{i=1}^{N} \beta_{i-1}^{-1} [F(x_i) - F(z)] \leq \varepsilon(x^N, z)$$

$$\leq V_{x_0}(z) \quad + \quad \sum_{i=1}^{N} \left[\frac{\langle \xi_i, z - x_{i-1} \rangle}{\beta_{i-1}} + \frac{\|\xi_i\|_*^2}{\beta_{i-1}^2}\right] \quad (13)$$

$$\leq 2V_{x_0}(z) + \quad \sum_{i=1}^{N} \quad \left[\frac{\langle \xi_i, z_{i-1} - x_{i-1} \rangle}{\beta_{i-1}} + \frac{3}{2} \frac{\|\xi_i\|_*^2}{\beta_{i-1}^2}\right], \quad (14)$$

where z_i is random vector with values in X, dependent only on $x_0, \xi_1, \ldots, \xi_i$.

Using Proposition 1, we obtain the following bounds for the mathematical expectation of error $F(\widehat{x}_N) - F_*$ for an approximate solution of the problem (1), based on the RSMD algorithm. In what follows, we denote by $\mathbf{E}\{\cdot\}$ the expectation of the distribution $\omega^N = (\omega_1, ..., \omega_N) \in \Omega^{\otimes N}$.

Corollary. *Denote $M = LR$. Let be $\lambda \geq \max\{M, \sigma\sqrt{N}\} + \upsilon\sigma$ and $\beta_i \geq 2L$ for all $i = 0, 1, ...$ Let $\widehat{x}_N$ be aproximate solution (11), where x_i are iterations of RSMD algorithm, defined by relations (9) and (10). Then*

$$\mathbf{E}\{F(\widehat{x}_N)\} - F_* \leq \left[\sum_{i=1}^{N} \beta_{i-1}^{-1}\right]^{-1} \left[R^2\Theta + \sum_{i=1}^{N} \left(\frac{2D\sigma}{\beta_{i-1}\sqrt{N}} + \frac{4\sigma^2}{\beta_{i-1}^2}\right)\right]. \tag{15}$$

In particular, if $\beta_i = \bar{\beta}$ *for all* $i = 0, 1, ...,$ *where*

$$\bar{\beta} = \max\left\{2L, \frac{\sigma\sqrt{N}}{R\sqrt{\Theta}}\right\}, \tag{16}$$

then inequalities are fulfilled:

$$\mathbf{E}\{F(\widehat{x}_N)\} - F_* \leq \frac{\bar{\beta}}{N}\mathbf{E}\left\{\sup_{z\in X}\varepsilon(x^N, z)\right\} \leq C\max\left\{\frac{LR^2\Theta}{N}, \frac{\sigma R\sqrt{\Theta}}{\sqrt{N}}\right\}. \tag{17}$$

Moreover, in this case the inequality with explicit constants has the form

$$\mathbf{E}\{F(\widehat{x}_N)\} - F_* \leq \max\left\{\frac{2LR^2\Theta}{N} + \frac{4R\sigma(1+\sqrt{\Theta})}{\sqrt{N}}, \frac{2R\sigma(1+4\sqrt{\Theta})}{\sqrt{N}}\right\}.$$

This result shows that if the truncation threshold λ is large enough, then the expected error of the proposed algorithm is estimated similarly to the expected error of the standard mirror descent algorithm with averaging, i.e. algorithm in which stochastic gradients are taken without truncation: $y_i = G(x_{i-1}, \omega_i)$.

The following theorem gives confidence bounds for the proposed algorithm.

Theorem. *Let be* $\beta_i = \bar{\beta} \geq 2L$ *for all* $i = 0, 1, ...,$ *and let be* $1 \leq \tau \leq N/\upsilon^2$,

$$\lambda = \max\left\{\sigma\sqrt{\frac{N}{\tau}}, M\right\} + \upsilon\sigma. \tag{18}$$

Let $\widehat{x}_N$ be aproximate solution (11), where x_i are iterations RSMD, defined by relations (9) and (10). Then there is a random event $\mathcal{A}_N \subset \Omega^{\otimes N}$ of probability at least $1 - 2e^{-\tau}$ such that for all $\omega^N \in \mathcal{A}_N$ the inequalities hold:

$$\begin{aligned} F(\widehat{x}_N) - F_* &\leq \bar{\beta} N^{-1} \sup_{z \in X} \varepsilon(x^N, z) \leq \\ &\leq C N^{-1} \left(\bar{\beta} R^2 \Theta + R \max \left\{ \sigma \sqrt{\tau N}, M\tau \right\} \right. \\ &\quad \left. + \bar{\beta}^{-1} \max\{N\sigma^2, M^2\tau\} \right). \end{aligned}$$

In paticular, chosing $\bar{\beta}$ by (16), we have for all $\omega^N \in \mathcal{A}_N$

$$F(\widehat{x}_N) - F_* \leq \max \left\{ C_1 \frac{LR^2[\tau \vee \Theta]}{N}, \; C_2 \sigma R \sqrt{\frac{\tau \vee \Theta}{N}} \right\} \quad (19)$$

where $C_1 > 0$ and $C_2 > 0$ are numerical constants.

The values of the numerical constants C_1 and C_2 in (19) can be obtained in the proof of the theorem, see [1].

Confidence bound (19) in Theorem contains two terms corresponding to deterministic and random errors. Unlike the case of noise with a "light tail" (see, for example, [41]) and the average (17) of the deterministic error $LR^2[\tau \vee \Theta]/N$ depends on the boundary depending on τ. Note also that Theorem gives the sub-Gaussian confidence bound (cf. the order of the stochastic error $\sigma R\sqrt{[\tau \vee \Theta]/N}$). However, the truncation threshold λ depends on the confidence level τ. This can be inconvenient when implementing algorithms. Some simple, but coarser confidence bounds can be obtained by using a universal threshold independent of τ, namely,

$\lambda = \max\{\sigma\sqrt{N}, M\} + \upsilon\sigma$. In particular, the following statement is true.

Theorem. *Let be* $\beta_i = \bar{\beta} \geq 2L$ *for all* $i = 0, 1, ...,$ *and let be* $N \geq \upsilon^2$. *Put*

$$\lambda = \max\left\{\sigma\sqrt{N}, M\right\} + \upsilon\sigma.$$

Let be $\widehat{x}_N = N^{-1}\sum_{i=1}^{N} x_i$, *where* x_i *are iterations of RSMD algorithm, defined by relations (9) and (10). Then there is a random event* $\mathcal{A}_N \subset \Omega^{\otimes N}$ *of probability at least* $1 - 2e^{-\tau}$

such that for all $\omega^N \in \mathcal{A}_N$ *the inequalities hold:*

$$\begin{aligned} F(\widehat{x}_N) - F_* &\leq \bar{\beta} N^{-1} \sup_{z \in X} \varepsilon(x^N, z) \leq \\ &\leq C N^{-1} (\bar{\beta} R^2 \Theta + \tau R \max\left\{\sigma\sqrt{N}, M\right\} \\ &\quad + \tau \bar{\beta}^{-1} \max\{N\sigma^2, M^2\}). \end{aligned}$$

In particular, choosing $\bar{\beta}$ *from formula (16), we have*

$$F(\widehat{x}_N) - F_* \leq \tfrac{\bar{\beta}}{N} \sup_{z \in X} \varepsilon(x^N, z)$$

$$\leq C N^{-1} \max\left\{L R^2 [\tau \vee \Theta],\ \tau\sigma R\sqrt{N\Theta}\right\} \qquad (20)$$

for all $\omega^N \in \mathcal{A}_N$.

The values of the numerical constants C in Theorem can be obtained in the proof, see [1].

5 Robust Confidence Bounds for Stochastic Optimization Methods

Arbitrary algorithms for solving problem (1) are considered for N calls of a stochastic oracle. Let there be a sequence $\big(x_i, G(x_i, \omega_{i+1})\big),\ i = 0, ..., N$, where $x_i \in X$ are search points of some stochastic algorithm, and $G(x_i, \omega_{i+1})$ are related observations of the stochastic gradient. It is assumed that x_i depends only on $\{(x_{j-1}, \omega_j), j = 1, \ldots, i\}$. The approximate solution of the problem (1) is defined in the form:

$$\widehat{x}_N = N^{-1} \sum_{i=1}^{N} x_i.$$

The goal is to build a confidence interval of sub-Gaussian accuracy for $F(\widehat{x}_N) - F_*$. To do this, we use the following fact. Note that for any $t \geq L$ the value

$$\epsilon_N(t) = N^{-1} \sup_{z \in X} \left\{ \sum_{i=1}^{N} \left[\langle \nabla \phi(x_{i-1}), x_i - z \rangle + \psi(x_i) - \psi(z) + t V_{x_{i-1}}(x_i) \right] \right\} \quad (21)$$

is upper bound of accuracy of the approximate solution $\widehat{x}_N$:

$$F(\widehat{x}_N) - F_* \leq \epsilon_N(t) \quad (22)$$

(see [1], Lemma 1 in Appendix). This fact is true for any sequence of points $x_0, \ldots, x_N$ in X, regardless of how they are obtained. However, since the function $\nabla \phi(\cdot)$ is not

known, the estimate (22) is not practical in practice. Replacing the gradients $\nabla\phi(x_{i-1})$ in (21) with their truncated estimates y_i defined in (10), we get an implementable analogue of $\epsilon_N(t)$:

$$\widehat{\epsilon}_N(t) = N^{-1} \sup_{z\in X} \left\{ \sum_{i=1}^{N} \left[\langle y_i, x_i - z\rangle + \psi(x_i) - \psi(z) + tV_{x_{i-1}}(x_i) \right] \right\}. \qquad (23)$$

Note that computing $\widehat{\epsilon}_N(t)$ reduces to solving a problem of the form (9) with $\beta = 0$. Thus, it is not more complicated than, for example, one step of the RSMD algorithm. Replacing $\nabla\phi(x_{i-1})$ with y_i introduces a random error, to

compensate for which you need to slightly increase $\widehat{\epsilon}_N(t)$ to get a reliable upper bound for $\epsilon_N(t)$. Namely, we add to $\widehat{\epsilon}_N(t)$ the value

$$\begin{aligned}\bar{\rho}_N(\tau) &= 4R\sqrt{5\Theta\max\{N\sigma^2, M^2\tau\}} + 16R\max\{\sigma\sqrt{N\tau}, M\tau\} + \\ &\quad + \min_{\mu\geq 0}\left\{20\mu\max\{N\sigma^2, M^2\tau\} + \mu^{-1}\sum_{i=1}^{N} V_{x_{i-1}}(x_i)\right\},\end{aligned}$$

where $\tau > 0$.

Proposition 2 *Let be $0 < \tau \leq N/\upsilon^2$, and let $y_i = y_i(\tau)$ be truncated stochastic gradients defined in (10), where the threshold $\lambda = \lambda(\tau)$ is chosen in the form (18). Let $\big(x_i, G(x_i, \omega_{i+1})\big)_{i=0}^{N}$ be the trajectory of a stochastic algorithm for which x_i depends only on*

$\{(x_{j-1}, \omega_j), j = 1, \ldots, i\}$. *Then for any* $t \geq L$

$$\Delta_N(\tau, t) = \widehat{\epsilon}_N(t) + \bar{\rho}_N(\tau)/N$$

is the upper bound for $\epsilon_N(t)$ *with probability* $1 - 2\mathrm{e}^{-\tau}$, *so*

$$\mathsf{P}\left\{F(\widehat{x}_N) - F_* \leq \Delta_N(\tau, t)\right\} \geq 1 - 2\mathrm{e}^{-\tau}.$$

Since $\Delta_N(\tau, t)$ monotonically increases in t, therefore if L is known, it suffices to use this bound for $t = L$. Note that although $\Delta_N(\tau, t)$ gives the upper bound for $\epsilon_N(t)$, the sentence 2 does not guarantee that $\Delta_N(\tau, t)$ is sufficient close to $\epsilon_N(t)$. However, for the RSMD algorithm with a constant step, this property holds, as is clear from the following statement.

Corollary. *Under the conditions of Proposition 2, let the*

vectors $x_0, \dots, x_N$ *be given by the relations RSMD recursion (9)– (10), where* $\beta_i = \bar{\beta} \geq 2L$, $i = 0, ..., N-1$. *Then*

$$\begin{aligned} \bar{\rho}_N(\tau) \quad \leq \quad & N\epsilon_N(\bar{\beta}) + 4R\sqrt{5\Theta \max\{N\sigma^2, M^2\tau\}} + \\ & + 16R \max\{\sigma\sqrt{N\tau}, M\tau\} \\ & + 20\bar{\beta}^{-1} \max\{N\sigma^2, M^2\tau\}. \end{aligned} \tag{24}$$

If, moreover, $\bar{\beta} \geq \max\left\{2L, \frac{\sigma\sqrt{N}}{R\sqrt{\Theta}}\right\}$, *then*

$$\bar{\rho}_N(\tau) \leq N\epsilon_N(\bar{\beta}) + C_3 L R^2[\Theta \vee \tau] + C_4 \sigma R\sqrt{N[\Theta \vee \tau]},$$

and with probability at least $1 - 4\mathrm{e}^{-\tau}$, *the value* $\Delta_N(\tau, \bar{\beta})$

satisfies the inequalities

$$\epsilon_N(\bar{\beta}) \leq \Delta_N(\tau, \bar{\beta}) \leq 3\epsilon_N(\bar{\beta}) + 2C_3 \frac{LR^2[\Theta \vee \tau]}{N} + 2C_4 \sigma R \sqrt{\frac{[\Theta \vee \tau]}{N}}, \quad (25)$$

where $C_3 > 0$ *and* $C_4 > 0$ *are numerical constants.*

The values of the numerical constants C_3 and C_4 can be derived from the proof of this corollary.

5.1 Robust Confidence Bounds for Quadratic Growth Problems

In this section, it is assumed that F is a quadratic growth function on X in the following sense (cf. [42]). Let a function F be continuous on X and let $X_* \subset X$ be the set of its minimum points on X. Then F is called a *quadratic growth function on* X if there exists a constant $\kappa > 0$ such that for any $\bar{x}(x) \in X_*$ for which the inequality holds

$$F(x) - F_* \geq \frac{\kappa}{2}\|x - \bar{x}(x)\|^2. \tag{26}$$

Note that every strongly convex function F on X with convexity coefficient κ is a quadratic growth function on X. However, the assumption of strong convexity, if

simultaneously imposing a Lipschitz condition with a constant L on the gradient F, has the disadvantage that, except for the case when $\|\cdot\|$ is the Euclidean norm, the ratio L/κ depends on the dimension of the space E. For example, in important cases, when $\|\cdot\|$ is ℓ_1-norm, nuclear norm, norm of full variation, etc., you can easily check (cf. [3]), which is not there are functions with a continuous Lipschitz gradient and with a conditional number — the ratio L/κ — smaller than the dimension of the space. Replacing a strong convexity with a growth condition (26) eliminates this problem, see examples in [42]. On the other hand, the assumption (26) is quite natural in the composite optimization problem, because in many interesting examples the component ϕ is smooth and the non-smooth part ψ of

the objective function is strongly convex. In particular, in the case when $E = \mathbf{R}^n$ and the norm is ℓ_1-norm, this allows us to consider "simple" strongly convex components, such as negative entropy $\psi(x) = \kappa \sum_j x_j \ln x_j$, $\kappa > 0$ (if X is standard simplex in $\mathbf{R}^n$), $\psi(x) = \gamma(\kappa)\|x\|_p^p$ with $1 \le p \le 2$ and corresponding choice $\gamma(\kappa)$ (if X is a convex compact in $\mathbf{R}^n$), and others. In all these cases, condition (26) is fulfilled with a known constant κ, which allows the use of [3, 43] approach to improve the confidence bounds of the stochastic mirror descent.

For simplicity, we consider only the case, when the total number of calls to the oracle N is fixed in advance. The RSMD algorithm for quadratically growing functions will be

determined in stages. At each stage, an auxiliary problem

$$\min_{x \in X_r(y)} F(x)$$

is solved for specially selected $r > 0$ and $y \in X$ by using RSMD. Here

$$X_r(y) = \{x \in X : \ \|x - y\| \leq r\}.$$

We initialize the algorithm by choosing arbitrary $y_0 = x_0 \in X$ and $r_0 \geq \max_{z \in X} \|z - x_0\|$. We set $r_k^2 = 2^{-k} r_0^2$, $k = 1, 2, ...$ Let C_1 and C_2 be numerical constants in the bound (19) of Theorem . For the given parameter $\tau > 0$, and $k = 1, 2, \ldots$ we define the values

$$\overline{N}_k = \max\left\{4C_1 \frac{L[\tau \vee \Theta]}{\kappa}, \, 16C_2 \frac{\sigma^2[\tau \vee \Theta]}{\kappa^2 r_{k-1}^2}\right\}, \quad N_k = \lfloor \overline{N}_k \rfloor. \quad (27)$$

Here $\rfloor t \lfloor$ denotes the smallest integer greater than or equal to t. Denote:

$$m(N) := \max\left\{k : \sum_{j=1}^{k} N_j \leq N\right\}.$$

Now let be $k \in \{1, 2, \ldots, m(N)\}$. At the k-th stage of the algorithm, solve the minimization problem of F on ball $X_{r_{k-1}}(y_{k-1})$, calculate its approximate solution $\widehat{x}_{N_k}$ according to (9), (11), where we set $x_0 = y_{k-1}$, $X = X_{r_{k-1}}(y_{k-1})$, $R = r_{k-1}$, $N = N_k$, and letting

$$\lambda = \max\left\{\sigma\sqrt{\frac{N}{\tau}}, Lr_{k-1}\right\} + \upsilon\sigma,$$

and

$$\beta_i \equiv \max\left\{2L,\ \frac{\sigma\sqrt{N}}{r_{k-1}\sqrt{\Theta}}\right\}.$$

It is assumed that at each k-th stage of the algorithm an exact solution of the minimization problem

$$\min_{z\in X_{r_{k-1}}(y_{k-1})}\{\langle a, z\rangle + \psi(z) + \beta\vartheta(z)\}$$

is available for any $a \in E$ and $\beta > 0$. At the output of the k-th stage of the algorithm, we obtain $y_k = \widehat{x}_{N_k}$.

Theorem. *Suppose that $m(N) \geq 1$, i.e. at least one stage of the algorithm described above is completed. Then there is a random event $\mathcal{B}_N \subset \Omega^{\otimes N}$ with probability at least $1 - 2m(N)e^{-\tau}$ such that for $\omega^N \in \mathcal{B}_N$ an approximate*

solution of $\widehat{x}_N = y_{m(N)}$ *after the* $m(N)$ *stages of the algorithm satisfies the inequality*

$$F(\widehat{x}_N) - F_* \le C \max \left\{ \kappa r_0^2 \, 2^{-N/4}, \ \kappa r_0^2 \exp\left(-\frac{C' \kappa N}{L[\tau \vee \Theta]} \right), \right. \\ \left. \frac{\sigma^2 [\tau \vee \Theta]}{\kappa N} \right\}. \quad (28)$$

Theorem shows that, for quadratic growth functions, the deterministic error component can be significantly reduced by making it exponentially decreasing in N. The stochastic error component is also significantly reduced. Note that the $m(N)$ factor of logarithmic order has little effect on the probability of deviations. Indeed, from (27) it follows that $m(N) \le C \ln\left(\frac{C\kappa^2 r_0^2 N}{\sigma^2(\tau \vee \Theta)}\right)$. Neglecting this factor in the probability of deviations and considering the stochastic

component of the error, we see that the confidence bound of Theorem is approximately sub-exponential, not sub-Gaussian.

6 Conclusion

Algorithms of smooth stochastic optimization are considered in a situation where the distribution of observations noises has heavy tails. It is shown that by truncating the observations of the gradient with the corresponding threshold, we can construct confidence sets for approximate solutions that are similar to those in the case of "light tails". It should be noted that the order of the deterministic error in the obtained boundaries is suboptimal — it is much more than the optimal estimates ($O(LR^2k^{-2})$ in the case of a convex objective function and $O(\exp(-\sqrt{\kappa/L}))$ in the strongly convex case) achieved accelerated algorithms [4, 41]. On the other hand, the proposed approach cannot be used for robustization of accelerated algorithms, since, when applied to such algorithms, the bias caused by the

truncation of gradients accumulates. The task of building accelerated robust stochastic algorithms with optimal guarantees remains open.

References

[1] *Nazin A., Nemirovski A., Tsybakov A., Juditsky A.* Algorithms of Robust Stochastic Optimization Based on Mirror Descent Method // *Autom. Remote Control.* 2019. V. 80. No 9. P. 1607–1627.

[2] *Nemirovski A., Juditsky A., Lan G., Shapiro A.* Robust stochastic approximation approach to stochastic programming // *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[3] *Juditsky A., Nesterov Y.* Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // *Stochastic Systems*, 4(1):44–80, 2014.

[4] *Ghadimi S., Lan G.* Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework // *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[5] *Tukey J. W.* A survey of sampling from contaminated distributions / *Contributions to probability and statistics*, pages 448–485, 1960.

[6] *Huber P. J.* Robust estimation of a location parameter // *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[7] *Huber P. J. et al.* The 1972 wald lecture robust statistics: A review // *The Annals of Mathematical Statistics*, 43(4):1041–1067, 1972.

[8] P. J. Huber. *Robust Statistics.* John Wiley and Sons, 1981.

[9] *Martin R., Masreliez C.* Robust estimation via stochastic approximation // *IEEE Transactions on Information Theory*, 21(3):263–271, 1975.

[10] Polyak B.T., Tsypkin Y.Z. Adaptive estimation algorithms: convergence, optimality, stability // Autom. Remote Control. 1979. V. 40. No 3. P. 378–389.

[11] *Polyak B.T., Tsypkin Ya.Z.* Robust pseudogradient adaptation algorithms // A&RC. 41:10 (1981). P. 1404–1409.

[12] *Polyak B., Tsypkin J.Z.* Robust identification. *Automatica*, 16(1):53–63, 1980.

[13] *Price E., VandeLinde V.* Robust estimation using the robbins-monro stochastic approximation algorithm // *IEEE Transactions on Information Theory*, 25(6):698–704, 1979.

[14] *Stanković S.S., Kovačević B.D.* Analysis of robust stochastic approximation algorithms for process identification // *Automatica*, 22(4):483–488, 1986.

[15] *Chen H.-F., Guo L., Gao A.J.* Convergence and robustness of the robbins-monro algorithm truncated at randomly varying bounds // *Stochastic Processes and their Applications*, 27:217–231, 1987.

[16] *Chen H.-F., Gao A.J.* Robustness analysis for stochastic approximation algorithms // *Stochastics and Stochastic Reports*, 26(1):3–20, 1989.

[17] *Nazin A.V., Polyak B.T., Tsybakov A.B.* Optimal and robust kernel algorithms for passive stochastic approximation // *IEEE Transactions on Information Theory*, 38(5):1577–1583, 1992.

[18] *Tsypkin Ya.Z.* Osnovy informatsionnoy teorii identifikatsii. — M.: Nauka, 1984. (In Russian.)

[19] *Tsypkin Ya.Z.* Informatsionnaya teoriya identifikatsii. — M.: Nauka, 1995. (In Russian.)

[20] *Joon Kwon, Guillaume Lecué and Matthieu Lerasle.* Median of means principle as a divide-and-conquer procedure for robustness, sub-sampling and hyper-parameters tuning. arxiv:2018

[21] *G. Chinot, G. Lecué and M. Lerasle.* Statistical Learning with Lipschitz and convex loss functions. arxiv:2018.

[22] Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice. arXiv preprint arXiv:1711.10306. Annals of Stat., to appear.

[23] Lecué, G., Lerasle, M. and Mathieu, T. (2018). Robust classi?cation via MOM minimization. arXiv preprint arXiv:1808.03106. s

[24] Lerasle, M. and Oliveira, R. I. (2011). Robust empirical mean estimators. arXiv preprint arXiv:1112.3914.

[25] *Lugosi, G. and Mendelson, S.* (2016). Risk minimization by median-of-means tournaments. arXiv preprint arXiv:1608.00757.

[26] Lugosi, G. and Mendelson, S. (2017). Regularization, sparse recovery, and median-of-means tournaments. arXiv preprint arXiv:1701.04112.

[27] Lugosi, G. and Mendelson, S. (2018). Near-optimal mean estimators with respect to general norms. arXiv preprint arXiv:1806.06233.

[28] *Hsu D., Sabato S.* Loss minimization and parameter estimation with heavy tails // *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

[29] *Bubeck S., Cesa-Bianchi N., Lugosi G.* Bandits with heavy tail // *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[30] *Devroye L., Lerasle M., Lugosi G., Oliveira R.I., et al.* Sub-gaussian mean estimators // *The Annals of Statistics*, 44(6):2695–2725, 2016.

[31] *Nemirovski A.S., Yudin D.B. Problem complexity and method efficiency in optimization,* J. Wiley & Sons, New York 1983.

[32] *Lugosi G., Mendelson S., et al.* Sub-gaussian estimators of the mean of a random vector // *The Annals of Statistics*, 47(2):783–794, 2019.

[33] *Catoni O.* Challenging the empirical mean and empirical variance: a deviation study // *Annales de l'IHP Probabilités et Statistiques*, 48(4):1148–1185, 2012.

[34] *Audibert J.-Y., Catoni O., et al.* Robust linear least squares regression // *The Annals of Statistics*, 39(5):2766–2794, 2011.

[35] *Minsker S.* Geometric median and robust estimation in banach spaces // *Bernoulli*, 21(4):2308–2335, 2015.

[36] *Wei X., Minsker S.* Estimation of the covariance structure of heavy-tailed distributions / In *Advances in Neural Information Processing Systems*, pages 2859–2868, 2017.

[37] *Chen Y., Su L., Xu J.* Distributed statistical machine learning in adversarial settings: Byzantine gradient descent / *Proceedings of the*

ACM on Measurement and Analysis of Computing Systems, 1(2):44, 2017.

[38] *Yin D., Chen Y., Ramchandran K., Bartlett P.* Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498*, 2018.

[39] *Cardot H., Cénac P., Chaouch M.* Stochastic approximation for multivariate and functional median / In *Proceedings of COMPSTAT'2010*, pages 421–428. Springer, 2010.

[40] *Cardot H., Cénac P., Godichon-Baggioni A., et al.* Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls // *The Annals of Statistics*, 45(2):591–614, 2017.

[41] *Lan G.* An optimal method for stochastic composite optimization // *Mathematical Programming*, 133(1-2):365–397, 2012.

[42] *Necoara I., Nesterov Y., Glineur F.* Linear convergence of first order methods for non-strongly convex optimization // *Mathematical Programming*, pages 1–39.

[43] *Juditsky A., Nemirovski A.* First order methods for nonsmooth convex large-scale optimization, i: general purpose methods / In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 121–148. MIT Press, 2011.

[44] *Freedman D.A.* On tail probabilities for martingales // *The Annals of Probability*, pages 100–118, 1975.

[45] *Chen G., Teboulle M.* Convergence analysis of a proximal-like minimization algorithm using Bregman functions // *SIAM Journal on Optimization*, 3(3):538–543, 1993.