© 2019 г. А.В. НАЗИН, д-р физ.-мат. наук (nazine@ipu.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва),
А.С. НЕМИРОВСКИЙ, д-р физ.-мат. наук (nemirovs@isye.gatech.edu)
(ISyE, Технологический институт Джорджии, Атланта, США),
А.Б. ЦЫБАКОВ, д-р физ.-мат. наук (alexandre.tsybakov@ensae.fr)
(CREST, ENSAE, Франция),
А.Б. ЮДИЦКИЙ, канд. техн. наук (anatoli.juditsky@univ-grenoble-alpes.fr)
(LJK, Университет Гренобль Альпы, Гренобль, Франция)

АЛГОРИТМЫ РОБАСТНОЙ СТОХАСТИЧЕСКОЙ ОПТИМИЗАЦИИ НА ОСНОВЕ МЕТОДА ЗЕРКАЛЬНОГО СПУСКА[1]

Предлагается подход к построению робастных неевклидовых итеративных алгоритмов выпуклой композитной стохастической оптимизации, основанный на усечении стохастических градиентов. Для таких алгоритмов устанавливаются субгауссовские доверительные границы точности при слабых предположениях о хвостах распределения шума в выпуклой и сильно выпуклой постановках. Также предлагаются робастные оценки точности стохастических алгоритмов общего вида.

Ключевые слова: робастные итеративные алгоритмы, алгоритмы стохастической оптимизации, выпуклая композитная стохастическая оптимизация, метод зеркального спуска, робастные доверительные множества.

DOI: 10.1134/S000523101909006X

1. Введение

В данной статье рассматривается задача *выпуклой композитной стохастической оптимизации*

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

где X — компактное выпуклое подмножество конечномерного векторного пространства E с нормой $\|\cdot\|$, ω — случайная переменная на вероятностном пространстве Ω с распределением P, функция ψ выпуклая и непрерывная, а функция $\Phi:\ X\times\Omega\to\mathbf{R}$. Предполагается, что математическое ожидание

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

[1] Работа А.Б. Юдицкого поддержана грантом PGMO 2016-2032H и совместно А.Б. Юдицкого с А.С. Немировским — грантом NSF CCF-1523768. Работа А.В. Назина поддержана Российским научным фондом (грант №16-11-10015). Работа А.Б. Цыбакова поддержана институтом GENES и грантом Labex Ecodec (ANR-11-LABEX-0047).

конечно при всех $x \in X$ и является выпуклой и дифференцируемой функцией от x. При этих предположениях задача (1) имеет решение с оптимальным значением $F_* = \min_{x \in X} F(x)$.

Предполагается, что доступен механизм (оракул), который для заданной на входе точки $(x, \omega) \in X \times \Omega$ возвращает случайный градиент — вектор $G(x, \omega)$, удовлетворяющий условиям

$$\mathbf{E}\{G(x,\omega)\} = \nabla\phi(x) \quad \text{и} \quad \mathbf{E}\left\{\|G(x,\omega) - \nabla\phi(x)\|_*^2\right\} \leqslant \sigma^2, \quad \forall\, x \in X, \tag{2}$$

где $\|\cdot\|_*$ — сопряженная норма к $\|\cdot\|$, а $\sigma > 0$ — некоторая постоянная. Цель данной работы состоит в построении $(1-\alpha)$-надежных приближенных решений задачи (1), т.е. решений $\widehat{x}_N$, основанных на N вызовах стохастического оракула и удовлетворяющих условию

$$\mathrm{Prob}\{F(\widehat{x}_N) - F^* \leqslant \delta_N(\alpha)\} \geqslant 1 - \alpha, \quad \forall\, \alpha \in (0,1), \tag{3}$$

с насколько возможно малым $\delta_N(\alpha) > 0$.

Заметим, что задачи стохастической оптимизации типа (1) возникают в контексте минимизации пенализованного риска, где доверительные границы (3) напрямую преобразуются в доверительные границы для точности получаемых оценок. В этой статье отыскиваются границы (3) с $\delta_N(\alpha)$ порядка $\sqrt{\ln(1/\alpha)/N}$. Такие границы часто называют субгауссовскими доверительными границами. Стандартные результаты о субгауссовских доверительных границах для алгоритмов стохастической оптимизации предполагают конечность экспоненциальных или субэкспоненциальных моментов стохастического шума оракула $G(x,\omega) - \nabla\phi(x)$ (ср. [1–3]). В настоящей работе строятся робастные стохастические алгоритмы, которые удовлетворяют субгауссовским границам типа (3) при существенно менее ограничительном условии (2).

Напомним, что понятие робастности процедур принятия статистических решений было введено Дж. Тьюки [4] и П. Хубером [5–7] в 1960-е гг., что послужило основой для последующей разработки робастных алгоритмов стохастической аппроксимации. В частности, в 1970–1980-х гг. алгоритмы, устойчивые к широким классам распределений шумов, были предложены для задач стохастической оптимизации и параметрической идентификации. Их асимптотические (с ростом размера выборки) свойства были хорошо изучены, см., например, [8–16] и ссылки в них. Важный вклад в развитие робастного подхода внес Я.З. Цыпкин. Так, исследованию итеративных алгоритмов робастной идентификации уделено значительное место в монографиях [17, 18].

Интерес к проблематике робастного оценивания возобновился в 2010-х гг. в связи с необходимостью разработки статистических процедур, устойчивых к шумам с тяжелыми хвостами в задачах большой размерности. Некоторые недавние работы [19–29] связаны с развитием метода медианы средних (median-of-means) [30] для построения оценок, удовлетворющих субгауссовским доверительным границам при шумах с тяжелыми хвостами. Так, в [27] была использована процедура median-of-means для построения $(1-\alpha)$-надежной версии стохастической аппроксимации с усреднением ("пакетного" алгоритма) в постановке стохастической оптимизации, подобной (1). Были

разработаны и другие оригинальные подходы [31–35], в частности, использующие геометрическую медиану для робастной оценки сигналов и ковариационных матриц с субгауссовыми гарантиями [34, 35]. Также возобновился интерес к робастным итеративным алгоритмам. Так, было показано, что робастность алгоритмов стохастической аппроксимации может быть повышена за счет использования геометрической медианы стохастических градиентов [36, 37]. Другой вариант процедуры стохастической аппроксимации для вычисления геометрической медианы был изучен в [38, 39], где особая структура задачи — ограниченность стохастических градиентов — позволила построить $(1-\alpha)$-надежные оценки при чрезвычайно слабом предположении о хвостах распределения шума.

В настоящей статье рассматривается подход к построению робастных стохастических алгоритмов методом усечения стохастических градиентов. Показывается, что этот метод удовлетворяет субгауссовским доверительным границам. В разделах 2 и 3 определяются основные компоненты рассматриваемой задачи оптимизации. В разделе 4 дается определение алгоритма робастного стохастического зеркального спуска и для него устанавливаются доверительные границы. В разделе 5 строятся робастные оценки точности для стохастических алгоритмов общего вида. Наконец, в разделе 6 устанавливаются робастные доверительные границы для задач, в которых F имеет квадратичный рост. Приложение содержит доказательства утверждений статьи.

2. Обозначения и определения

Пусть E — конечномерное вещественное векторное пространство с нормой $\|\cdot\|$, а E^* — сопряженное к E пространство. Обозначим через $\langle s, x\rangle$ значение линейной функции $s \in E^*$ в точке $x \in E$ и через $\|\cdot\|_*$ сопряженную к $\|\cdot\|$ норму на E^*, т.е.

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

Рассмотрим непрерывную выпуклую функцию $\theta : B \to \mathbf{R}$ на единичном шаре

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

обладающую следующим свойством:

$$\langle \theta'(x) - \theta'(x'), x - x'\rangle \geqslant \|x - x'\|^2, \quad \forall x, x' \in B, \tag{4}$$

где $\theta'(\cdot)$ — непрерывная в $B^o = \{x \in B : \partial\theta(x) \neq \emptyset\}$ версия субградиента $\theta(\cdot)$, а $\partial\theta(x)$ — субдифференциал функции $\theta(\cdot)$ в точке x, т.е. множество всех субградиентов в данной точке. Иными словами, функция $\theta(\cdot)$ сильно выпукла на B с коэффициентом 1 относительно нормы $\|\cdot\|$. Будем называть $\theta(\cdot)$ нормализованной прокс-функцией. Примерами таких функций являются:

- $\theta(x) = \dfrac{1}{2}\|x\|_2^2$ для $(E, \|\cdot\|) = (\mathbf{R}^n, \|\cdot\|_2)$;
- $\theta(x) = 2\mathrm{e}(\ln n)\|x\|_p^p$ при $p = p(n) := 1 + \dfrac{1}{2\ln n}$ для $(E, \|\cdot\|) = (\mathbf{R}^n, \|\cdot\|_1)$;

- $\theta(x) = 4\mathrm{e}(\ln n)\sum\limits_{i=1}^{n}|\lambda_i(x)|^p$ при $p = p(n)$, для $E = S_n$, где S_n — пространство симметричных $n \times n$ матриц, оснащенное ядерной нормой $\|x\| = \sum\limits_{i=1}^{n}|\lambda_i(x)|$, а $\lambda_i(x)$ — собственные значения матрицы x.

Здесь и далее через $\|\cdot\|_p$ обозначается ℓ_p-норма в $\mathbf{R}^n$, $p \geqslant 1$. Без ограничения общности, будем предполагать в дальнейшем, что

$$0 = \arg\min_{x\in B}\theta(x).$$

Введем также обозначение

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

Пусть теперь X — выпуклое компактное подмножество в E, и пусть $x_0 \in X$ и $R > 0$ таковы, что $\max_{x\in X}\|x - x_0\| \leqslant R$. Снабдим X прокс-функцией

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

Обратим внимание, что $\vartheta(\cdot)$ сильно выпукла с коэффициентом 1 и

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

Пусть $D := \max\limits_{x,x'\in X}\|x - x'\|$ — диаметр множества X. Тогда $D \leqslant 2R$.

Далее также понадобится расхождение Брегмана

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

В дальнейшем через C и C' обозначаются положительные числовые константы, не обязательно одинаковые в различных случаях.

3. Предположения

Рассмотрим выпуклую стохастическую композитную задачу оптимизации (1) на выпуклом компактном множестве $X \subset E$. В дальнейшем предполагается, что функция

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

выпукла на X, дифференцируема в каждой точке множества X и ее градиент удовлетворяет условию Липшица

$$\|\nabla\phi(x') - \nabla\phi(x)\|_* \leqslant L\|x - x'\|, \quad \forall\, x, x' \in X. \tag{5}$$

Предположим также, что функция ψ является выпуклой и непрерывной. Далее предположим, что доступен стохастический оракул, который, имея на

входе точку $(x,\omega)\in X\times\Omega$, возвращает случайный вектор $G(x,\omega)$, удовлетворяющий условиям (2). Кроме того, предполагается, что при любых $a\in E^*$ и $\beta>0$ доступно точное решение задачи на минимум

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

Это предположение выполняется для типичных штрафных функций ψ, таких как выпуклые степенные функции от ℓ_p-нормы (если X — выпуклый компакт в $\mathbf{R}^n$) или отрицательная энтропия $\psi(x)=\kappa\sum\limits_{j=1}^{n}x_j\ln x_j$, где $k>0$ и $0\cdot\ln 0=0$ (если X — стандартный симплекс в $\mathbf{R}^n$). Наконец, предполагается, что доступен вектор $g(\bar x)$, где $\bar x\in X$ — некоторая точка в X, такой что

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

где $\upsilon\geqslant 0$ — некоторая постоянная. Это предположение мотивируется следующим образом.

Во-первых, если *априори* известно, что глобальный минимум функции ϕ достигается во внутренней точке x_ϕ множества X (что часто имеет место в статистических приложениях стохастической оптимизации), имеем $\nabla\phi(x_\phi)=0$. Поэтому, выбирая $\bar x=x_\phi$, можно положить $g(\bar x)=0$, и предположение (6) автоматически выполнено с $\upsilon=0$.

Во-вторых, в общей ситуации можно взять в качестве $\bar x$ произвольную точку множества X и в качестве $g(\bar x)$ геометрическую медиану стохастических градиентов $G(\bar x,\omega_i)$, $i=1,\dots,m$, по m вызовам оракула. Из [34] следует, что если m порядка $\ln\left(\varepsilon^{-1}\right)$ при некотором достаточно малом $\varepsilon>0$, то

$$\mathrm{Prob}\{\|g(\bar x)-\nabla\phi(\bar x)\|_*>\upsilon\sigma\}\leqslant\varepsilon. \tag{7}$$

Таким образом, доверительные границы, полученные ниже, останутся в силе с точностью до поправки ε в вероятности уклонений.

4. Границы точности алгоритма РСЗС

Далее везде считается, что выполнены предположения, сформулированные в разделе 3. Введем композитное проксимальное преобразование

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

где $\beta>0$ — параметр настройки.

Для $i=1,2,\dots$ определим алгоритм *Робастного Стохастического Зеркального Спуска* (РСЗС) следующими рекуррентными соотношениями:

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

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

Здесь $\beta_i > 0$, $i = 0, 1, \ldots$, и $\lambda > 0$ — параметры настройки, которые будут определены ниже, а $\omega_1, \omega_2, \ldots$ — независимые одинаково распределенные (н.о.р.) реализации случайной величины ω, соответствующие вызовам оракула на каждом шаге алгоритма.

Приближенное решение задачи (1) после N итераций определим как взвешенное среднее

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

В случае, когда глобальный минимум ϕ достигается во внутренней точке X и $\upsilon = 0$, определение (10) упрощается. В этом случае, заменяя $\|\bar{x} - x_{i-1}\|$ на верхнюю границу D и полагая $\upsilon = 0$ и $g(\bar{x}) = 0$ в (10), усеченный стохастический градиент вычисляем по формуле

$$y_i = \begin{cases} G(x_{i-1}, \omega_i), & \text{если } \|G(x_{i-1}, \omega_i)\|_* \leqslant LD + \lambda, \\ 0 & \text{в противном случае.} \end{cases}$$

Следующее утверждение описывает некоторые полезные свойства рекурсии зеркального спуска (9). Обозначим:

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

и

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

где $x^N = (x_0, \ldots, x_N)$.

Предложение 1. Пусть $\beta_i \geqslant 2L$ при всех $i = 0, 1, \ldots$ и пусть $\widehat{x}_N$ определено в (11)*, где x_i – итерации* (9) *при любом y_i, не обязательно заданным соотношением* (10)*. Тогда для любого $z \in X$ имеем*

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

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

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

где z_i – случайный вектор со значениями в X, зависящий только от $x_0, \xi_1, \ldots, \xi_i$.

Используя предложение 1, получим следующие границы для математического ожидания ошибки $F(\widehat{x}_N) - F_*$ приближенного решения задачи (1), основанного на алгоритме РСЗС. В дальнейшем будем обозначать через $\mathbf{E}\{\cdot\}$ математическое ожидание по распределению $\omega^N = (\omega_1, \ldots, \omega_N) \in \Omega^{\otimes N}$.

Следствие 1. Обозначим: $M = LR$. *Пусть* $\lambda \geqslant \max\{M, \sigma\sqrt{N}\} + \upsilon\sigma$ *и* $\beta_i \geqslant 2L$ *при всех* $i = 0, 1, \ldots$ *Пусть* $\widehat{x}_N$ *– приближенное решение* (11), *где* x_i *– итерации РСЗС, определяемые соотношениями* (9) *и* (10). *Тогда*

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

В частности, если $\beta_i = \bar{\beta}$ *при всех* $i = 0, 1, \ldots$, *где*

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

то выполняются неравенства

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

Более того, в этом случае неравенство с явными константами имеет вид

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

Из этого результата видно, что если порог усечения λ достаточно велик, то математическое ожидание ошибки предложенного алгоритма оценивается аналогично математическому ожиданию ошибки стандартного алгоритма зеркального спуска с усреднением, т.е. алгоритма, в котором стохастические градиенты берутся без усечения: $y_i = G(x_{i-1}, \omega_i)$.

Следующая теорема дает доверительные границы для предложенного алгоритма.

Теорема 1. Пусть $\beta_i = \bar{\beta} \geqslant 2L$ *при всех* $i = 0, 1, \ldots$ *и пусть* $1 \leqslant \tau \leqslant N/\upsilon^2$,

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

Пусть $\widehat{x}_N$ *– приближенное решение* (11), *где* x_i *– итерации РСЗС, определяемые соотношениями* (9) *и* (10). *Тогда существует случайное событие* $\mathcal{A}_N \subset \Omega^{\otimes N}$ *вероятности не менее* $1 - 2e^{-\tau}$ *такое, что при всех* $\omega^N \in \mathcal{A}_N$ *выполняется неравенство*

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

В частности, выбирая $\bar\beta$ по формуле (16)*, при всех $\omega^N \in \mathcal{A}_N$ имеем*

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

где $C_1 > 0$ и $C_2 > 0$ — числовые постоянные.

Значения постоянных C_1 и C_2 в (19) можно получить из доказательства теоремы, ср. границу в (П.12).

Доверительная граница (19) в теореме 1 содержит два члена, соответствующие детерминированной и случайной ошибкам. В отличие от случая шума с "легким хвостом" (см., например, [40]) и от границы в среднем (17) детерминированная ошибка $LR^2[\tau \vee \Theta]/N$ зависит от τ. Заметим также, что теорема 1 дает субгауссовскую доверительную границу (ср. порядок стохастической ошибки $\sigma R\sqrt{[\tau \vee \Theta]/N}$). Однако при этом порог усечения λ зависит от доверительного уровня τ. Это может быть неудобно при реализации алгоритмов. Некоторые простые, но более грубые, доверительные границы могут быть получены, если использовать универсальный порог, не зависящий от τ, а именно, $\lambda = \max\{\sigma\sqrt{N}, M\} + \upsilon\sigma$. В частности, верно следующее утверждение.

Теорема 2. Пусть $\beta_i = \bar\beta \geqslant 2L$ при всех $i = 0, 1, \ldots$ и пусть $N \geqslant \upsilon^2$. Положим

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

Пусть

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

где x_i – итерации РСЗС, определяемые соотношениями (9) *и* (10)*. Тогда существует случайное событие $\mathcal{A}_N \subset \Omega^{\otimes N}$ вероятности не менее $1 - 2e^{-\tau}$ такое, что при всех $\omega^N \in \mathcal{A}_N$ выполняется неравенство*

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

$$\leqslant \frac{C}{N}\left(\bar\beta R^2\Theta + \tau R \max\left\{\sigma\sqrt{N}, M\right\} + \tau\bar\beta^{-1}\max\{N\sigma^2, M^2\}\right).$$

В частности, выбирая $\bar\beta$ по формуле (16)*, при всех $\omega^N \in \mathcal{A}_N$ имеем*

$$F(\widehat{x}_N) - F_* \leqslant \frac{\bar\beta}{N}\sup_{z\in X}\varepsilon(x^N, z) \leqslant C\max\left\{\frac{LR^2[\tau \vee \Theta]}{N},\ \tau\sigma R\sqrt{\frac{\Theta}{N}}\right\}. \tag{20}$$

Значения числовых постоянных C в теореме 2 можно получить из ее доказательства, ср. границу в (П.12).

5. Робастные доверительные границы для методов стохастической оптимизации

Рассматриваются произвольные алгоритмы решения задачи (1) по N вызовам стохастического оракула. Пусть имеется последовательность $\big(x_i, G(x_i, \omega_{i+1})\big)$, $i = 0, \ldots, N$, где $x_i \in X$ — точки поиска некоторого стохастического алгоритма, а $G(x_i, \omega_{i+1})$ — соответствующие наблюдения стохастического градиента. Предполагается, что x_i зависит только от $\{(x_{j-1}, \omega_j), j = 1, \ldots, i\}$. Приближенное решение задачи (1) определим в виде:

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

Целью является построение доверительного интервала субгауссовской точности для величины $F(\widehat{x}_N) - F_*$. Для этого воспользуемся следующим фактом. Заметим, что для любого $t \geqslant L$ значение

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

является верхней границей точности приближенного решения $\widehat{x}_N$:

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

(см. лемму 1 в Приложении). Этот факт верен для любой последовательности точек $x_0, \ldots, x_N$ в X вне зависимости от того, как они получены. Однако поскольку функция $\nabla\phi(\cdot)$ не известна, оценка (22) на практике не реализуема. Заменяя в (21) градиенты $\nabla\phi(x_{i-1})$ на их усеченные оценки y_i, определенные в (10), получим реализуемый аналог величины $\epsilon_N(t)$:

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

Заметим, что вычисление $\widehat{\epsilon}_N(t)$ сводится к решению задачи вида (8) с $\beta = 0$. Таким образом, оно не сложнее, чем, например, один шаг алгоритма РСЗС. Замена $\nabla\phi(x_{i-1})$ на y_i вносит случайную ошибку, для компенсации которой необходимо несколько увеличить $\widehat{\epsilon}_N(t)$, чтобы получить надежную верхнюю границу для $\epsilon_N(t)$. А именно, добавим к $\widehat{\epsilon}_N(t)$ величину

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

где $\tau > 0$.

Предложение 2. Пусть $\big(x_i, G(x_i,\omega_{i+1})\big)_{i=0}^{N}$ *— траектория стохастического алгоритма, для которого* x_i *зависит только от* $\{(x_{j-1},\omega_j),\ j=1,\dots,i\}$. *Пусть* $0<\tau\leqslant N/\upsilon^2$ *и пусть* $y_i=y_i(\tau)$ *— усеченные стохастические градиенты, определенные в* (10), *где порог* $\lambda=\lambda(\tau)$ *выбран в виде* (18). *Тогда при любом* $t\geqslant L$ *величина*

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

является верхней границей для $\epsilon_N(t)$ *с вероятностью* $1-2\mathrm{e}^{-\tau}$, *так что*

$$\mathrm{Prob}\big\{F(\widehat{x}_N)-F_*\leqslant\Delta_N(\tau,t)\big\}\geqslant 1-2\mathrm{e}^{-\tau}.$$

Так как $\Delta_N(\tau,t)$ монотонно возрастает по t, поэтому L известно, то достаточно использовать эту границу при $t=L$. Заметим, что хотя $\Delta_N(\tau,t)$ и дает верхнюю границу для $\epsilon_N(t)$, предложение 2 не гарантирует, что $\Delta_N(\tau,t)$ достаточно близко к $\epsilon_N(t)$. Тем не менее для алгоритма РСЗС с постоянным шагом это свойство имеет место, как явствует из следующего утверждения.

Следствие 2. Пусть в условиях предложения 2 *векторы* $x_0,\dots,x_N$ *задаются соотношениями РСЗС рекурсии* (9)–(10), *где* $\beta_i=\bar{\beta}\geqslant 2L$, $i=0,\dots,N-1$. *Тогда*

$$\bar{\rho}_N(\tau)\leqslant N\epsilon_N(\bar{\beta})+4R\sqrt{5\Theta\max\{N\sigma^2,M^2\tau\}}+$$
$$+16R\max\Big\{\sigma\sqrt{N\tau},M\tau\Big\}+20\bar{\beta}^{-1}\max\big\{N\sigma^2,M^2\tau\big\}. \tag{24}$$

Если, более того,

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

то

$$\bar{\rho}_N(\tau)\leqslant N\epsilon_N(\bar{\beta})+C_3LR^2[\Theta\vee\tau]+C_4\sigma R\sqrt{N[\Theta\vee\tau]},$$

и с вероятностью не менее $1-4\mathrm{e}^{-\tau}$ *величина* $\Delta_N(\tau,\bar{\beta})$ *удовлетворяет неравенствам*

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

где $C_3>0$ *и* $C_4>0$ *— числовые постоянные.*
Значения постоянных C_3 и C_4 можно вывести из доказательства данного следствия.

6. Робастные доверительные границы для задач с квадратичным ростом

В этом разделе предполагается, что F — функция квадратичного роста на X в следующем смысле (ср. [41]). Пусть функция F непрерывна на X и пусть $X_* \subset X$ — множество ее точек минимума на X. Тогда F называется *функцией квадратичного роста на* X, если существует постоянная $\kappa > 0$ такая, что для любого $x \in X$ найдется $\bar{x}(x) \in X_*$, для которого выполняется неравенство

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

Заметим, что всякая сильно выпуклая на X функция F с коэффициентом сильной выпуклости κ является функцией квадратичного роста на X. Однако предположение о сильной выпуклости, если одновременно с ним накладывать еще и условие Липшица с константой L на градиент F, имеет тот недостаток, что за исключением случая, когда $\|\cdot\|$ является евклидовой нормой, отношение L/κ зависит от размерности пространства E. Например, в важных случаях, когда $\|\cdot\|$ является ℓ_1-нормой, ядерной нормой, нормой полной вариации, и ряде других, можно легко проверить (ср. [2]), что не существует функции с липшицевым градиентом, для которой отношение L/κ было бы меньше, чем размерность пространства. Замена сильной выпуклости на условие роста (26) устраняет эту проблему, см. примеры в [41]. С другой стороны, предполагать (26) в задаче композитной оптимизации вполне естественно по той причине, что во многих интересных примерах составляющая ϕ гладкая, а негладкая часть ψ целевой функции сильно выпуклая. В частности, если $E = \mathbf{R}^n$ и норма $\|\cdot\|$ является ℓ_1-нормой, это позволяет рассматривать такие сильно выпуклые компоненты, как отрицательная энтропия $\psi(x) = \kappa \sum\limits_{j=1}^{n} x_j \ln x_j$ (если X — стандартный симплекс в $\mathbf{R}^n$), $\psi(x) = \gamma(\kappa)\|x\|_p^p$ с $1 \leqslant p \leqslant 2$ и соответствующим выбором $\gamma(\kappa) > 0$ (если X — выпуклый компакт в $\mathbf{R}^n$) и др. Во всех этих случаях условие (26) выполнено с известной постоянной κ, что позволяет применять подход [2, 42] для улучшения доверительных границ стохастического зеркального спуска.

Алгоритм РСЗС для квадратично растущих функций определим поэтапно. На каждом этапе для специально подобранных $r > 0$ и $y \in X$ решается вспомогательная задача

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

с использованием РСЗС. Здесь

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

Инициализируем алгоритм, выбирая произвольные $y_0 = x_0 \in X$ и $r_0 \geqslant \max\limits_{z \in X} \|z - x_0\|$. Положим $r_k^2 = 2^{-k} r_0^2$, $k = 1, 2, \ldots$ Пусть C_1 и C_2 — числовые постоянные в границе (19) теоремы 1. Для заданного параметра $\tau > 0$ и

$k = 1, 2, \ldots$ определим значения

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

Здесь $\lfloor t \rfloor$ — наименьшее целое число, большее или равное $t > 0$. Обозначим:

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

Пусть теперь $k \in \{1, 2, \ldots, m(N)\}$. На k-м этапе алгоритма решаем задачу минимизации F на множестве $X_{r_{k-1}}(y_{k-1})$, вычисляем ее приближенное решение $\widehat{x}_{N_k}$ в соответствии с (9)–(11), где заменяем x_0 на y_{k-1}, X на $X_{r_{k-1}}(y_{k-1})$, R на r_{k-1}, N на N_k, и полагаем

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

и

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

Предполагается, что на каждом этапе k алгоритма доступно точное решение задачи минимизации

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

при любых $a \in E$ и $\beta > 0$. На выходе k-го этапа алгоритма получаем $y_k := \widehat{x}_{N_k}$.

Теорема 3. Предположим, что $m(N) \geqslant 1$, т.е. хотя бы один этап описанного выше алгоритма завершен. Тогда существует случайное событие $\mathcal{B}_{m(N)} \subset \Omega^{\otimes N}$ вероятности не менее $1 - 2m(N)e^{-\tau}$ такое, что для $\omega^N \in \mathcal{B}_{m(N)}$ приближенное решение $y_{m(N)}$ после $m(N)$ этапов алгоритма удовлетворяет неравенству

$$F(y_{m(N)}) - F_* \leqslant C\max\left\{\kappa r_0^2 2^{-N/4},\ \kappa r_0^2 \exp\left(-\frac{C'\kappa N}{L[\tau \vee \Theta]}\right),\ \frac{\sigma^2[\tau \vee \Theta]}{\kappa N}\right\}. \tag{28}$$

Теорема 3 показывает, что для функций квадратичного роста можно существенно уменьшить детерминированную составляющую ошибки, сделав ее экспоненциально убывающей по N. Стохастическая составляющая ошибки также существенно уменьшается. Заметим, что множитель $m(N)$ логарифмического порядка слабо влияет на вероятность уклонений. Действительно, из (27) следует, что $m(N) \leqslant C\ln\left(\frac{C'\kappa^2 r_0^2 N}{\sigma^2(\tau \vee \Theta)}\right)$. Пренебрегая этим множителем в вероятности уклонений и рассматривая стохастическую составляющую ошибки, видим, что доверительная граница теоремы 3 является приближенно субэкспоненциальной, а не субгауссовской.

7. Заключение

Рассмотрены алгоритмы гладкой стохастической оптимизации в ситуации, когда распределения шумов наблюдений имеют тяжелые хвосты. Показано, что при усечении наблюдений градиента с соответствующим порогом можно построить доверительные множества для приближенных решений, аналогичные имеющимся в случае "легких хвостов". Стоит отметить, что порядок детерминированной ошибки в полученных границах является субоптимальным — он значительно больше оптимальных оценок ($O(LR^2N^{-2})$ в случае выпуклой целевой функции и $O(\exp(-N\sqrt{\kappa/L}))$ в сильно выпуклом случае), достигаемых ускоренными алгоритмами [3, 40]. С другой стороны, предлагаемый подход не может быть использован для робастизации ускоренных алгоритмов, так как при его применении для таких алгоритмов накапливается смещение, вносимое усечением градиентов. Задача построения ускоренных робастных стохастических алгоритмов с оптимальными гарантиями остается открытой.

ПРИЛОЖЕНИЕ

П.1. Предварительные замечания. Начнем со следующего известного результата.

Лемма 1. Предположим, что ϕ и ψ удовлетворяют предположениям раздела 3, и пусть $x_0, \dots, x_N$ – некоторые точки множества X. Обозначим:

$$\varepsilon_{i+1}(z) := \langle \nabla\phi(x_i), x_{i+1} - z\rangle + \langle \psi'(x_{i+1}), x_{i+1} - z\rangle + LV_{x_i}(x_{i+1}).$$

Тогда для любого $z \in X$ выполнено неравенство

$$F(x_{i+1}) - F(z) \leqslant \varepsilon_{i+1}(z).$$

Кроме того, для $\widehat{x}_N = \frac{1}{N}\sum\limits_{i=1}^{N} x_i$ имеем

$$F(\widehat{x}_N) - F(z) \leqslant N^{-1}\sum_{i=1}^{N}[F(x_i) - F(z)] \leqslant N^{-1}\sum_{i=0}^{N-1}\varepsilon_{i+1}(z).$$

Доказательство. Используя свойство $V_x(z) \geqslant \frac{1}{2}\|x - z\|^2$, выпуклость функций ϕ и ψ и условие Липшица на $\nabla\phi$, для любого $z \in X$ будем иметь:

$$\begin{aligned}
F(x_{i+1}) - F(z) &= [\phi(x_{i+1}) - \phi(z)] + [\psi(x_{i+1}) - \psi(z)] = \\
&= [\phi(x_{i+1}) - \phi(x_i)] + [\phi(x_i) - \phi(z)] + [\psi(x_{i+1}) - \psi(z)] \leqslant \\
&\leqslant [\langle \nabla\phi(x_i), x_{i+1} - x_i\rangle + LV_{x_i}(x_{i+1})] + \\
&\quad + \langle \nabla\phi(x_i), x_i - z\rangle + \psi(x_{i+1}) - \psi(z) \leqslant \\
&\leqslant \langle \nabla\phi(x_i), x_{i+1} - z\rangle + \langle \psi'(x_{i+1}), x_{i+1} - z\rangle + LV_{x_i}(x_{i+1}) = \\
&= \varepsilon_{i+1}(z).
\end{aligned}$$

Просуммировав эту границу по i от 0 до $N-1$ и воспользовавшись выпуклостью F, получим второе утверждение леммы. □

В дальнейшем будем обозначать через $\mathbf{E}_{x_i}\{\cdot\}$ условное математическое ожидание при фиксированном x_i.

Лемма 2. Пусть выполнены предположения раздела 3 и пусть x_i и y_i удовлетворяют рекурсии РСЗС, ср. (9) *и* (10)*. Тогда*

$$
\text{(П.1)}\qquad
\begin{aligned}
&\text{(a)}\quad \|\xi_i\|_* \leqslant 2(M+\upsilon\sigma)+\lambda,\\
&\text{(b)}\quad \|\mathbf{E}_{x_{i-1}}\{\xi_i\}\|_* \leqslant (M+\upsilon\sigma)\left(\frac{\sigma}{\lambda}\right)^2+\frac{\sigma^2}{\lambda},\\
&\text{(c)}\quad (\mathbf{E}_{x_{i-1}}\{\|\xi_i\|_*^2\})^{1/2} \leqslant \sigma+(M+\upsilon\sigma)\frac{\sigma}{\lambda}.
\end{aligned}
$$

Доказательство. Обозначим $\chi_i = 1_{\{\|G(x_{i-1},\omega_i)-g(\bar{x})\|_*>L\|x_{i-1}-\bar{x}\|+\lambda+\upsilon\sigma\}}$. Заметим, что по построению $\chi_i \leqslant \eta_i := 1_{\{\|G(x_{i-1},\omega_i)-\nabla\phi(x_{i-1})\|_*>\lambda\}}$. Имеем

$$
\begin{aligned}
\xi_i = y_i-\nabla\phi(x_{i-1}) &= [G(x_{i-1},\omega_i)-\nabla\phi(x_{i-1})](1-\chi_i)+[g(\bar{x})-\nabla\phi(x_{i-1})]\chi_i =\\
&= [G(x_{i-1},\omega_i)-g(\bar{x})](1-\chi_i)+[g(\bar{x})-\nabla\phi(x_{i-1})] =\\
&= [G(x_{i-1},\omega_i)-\nabla\phi(x_{i-1})]+[g(\bar{x})-G(x_{i-1},\omega_i)]\chi_i.
\end{aligned}
$$

Следовательно,

$$
\|\xi_i\|_* \leqslant \|G(x_{i-1},\omega_i)-g(\bar{x})\|_*(1-\chi_i)+\|g(\bar{x})-\nabla\phi(x_{i-1})\|_* \leqslant 2(M+\upsilon\sigma)+\lambda.
$$

Кроме того, поскольку $\mathbf{E}_{x_{i-1}}\{G(x_{i-1},\omega_i)\} = \nabla\phi(x_{i-1})$, имеем

$$
\begin{aligned}
\|\mathbf{E}_{x_{i-1}}\{\xi_i\}\|_* &= \left\|\mathbf{E}_{x_{i-1}}\{[(G(x_{i-1},\omega_i)-\nabla\phi(x_{i-1}))-(g(\bar{x})-\nabla\phi(x_{i-1}))]\chi_i\}\right\|_* \leqslant\\
&\leqslant \mathbf{E}_{x_{i-1}}\{[\|G(x_{i-1},\omega_i)-\nabla\phi(x_{i-1})\|_*+\|g(\bar{x})-\nabla\phi(x_{i-1})\|_*]\chi_i\} \leqslant\\
&\leqslant \mathbf{E}_{x_{i-1}}\{\|G(x_{i-1},\omega_i)-\nabla\phi(x_{i-1})\|_*\varsigma_i\}+(M+\upsilon\sigma)\mathbf{E}_{x_{i-1}}\{\varsigma_i\} \leqslant\\
&\leqslant \frac{\sigma^2}{\lambda}+(M+\upsilon\sigma)\left(\frac{\sigma}{\lambda}\right)^2.
\end{aligned}
$$

Далее,

$$
\|\xi_i\|_* \leqslant \|G(x_{i-1},\omega_i)-\nabla\phi(x_{i-1})\|_*(1-\chi_i)+\|g(\bar{x})-\nabla\phi(x_{i-1})\|_*\chi_i
$$

и

$$
\begin{aligned}
\mathbf{E}_{x_{i-1}}\{\|\xi_i\|_*^2\}^{1/2} &\leqslant \mathbf{E}_{x_{i-1}}\left\{\|G(x_{i-1},\omega_i)-\nabla\phi(x_{i-1})\|_*^2\right\}^{1/2}+\\
&\quad+\mathbf{E}_{x_{i-1}}\left\{\|g(\bar{x})-\nabla\phi(x_{i-1})\|_*^2\chi_i\right\}^{1/2} \leqslant\\
&\leqslant \sigma+(M+\upsilon\sigma)\mathbf{E}\{\chi_i\}^{1/2} \leqslant \sigma+(M+\upsilon\sigma)\mathbf{E}_{x_{i-1}}\{\eta_i\}^{1/2} \leqslant\\
&\leqslant \sigma+(M+\upsilon\sigma)\frac{\sigma}{\lambda}.
\end{aligned}
$$

□

В следующей лемме даются границы для отклонений сумм $\sum\limits_i\langle\xi_i, x_{i-1}-z\rangle$ и $\sum\limits_i\|\xi_i\|_*^2$.

Лемма 3. Пусть выполнены предположения раздела 3 и пусть x_i и y_i удовлетворяют рекурсии РСЗС, ср. (9) *и* (10).

(i) *Если* $\tau \leqslant N/\upsilon^2$ *и* $\lambda = \max\left\{\sigma\sqrt{\frac{N}{\tau}}, M\right\} + \upsilon\sigma$, *то для произвольного* $z \in X$

$$\text{(П.2)} \qquad \operatorname{Prob}\left\{\sum_{i=1}^{N}\langle\xi_i, z - x_{i-1}\rangle \geqslant 16R\max\{\sigma\sqrt{N\tau}, M\tau\}\right\} \leqslant e^{-\tau}$$

и

$$\text{(П.3)} \qquad \operatorname{Prob}\left\{\sum_{i=1}^{N}\|\xi_i\|_*^2 \geqslant 40\max\{N\sigma^2, M^2\tau\}\right\} \leqslant e^{-\tau}.$$

(ii) *Если* $N \geqslant \upsilon^2$ *и* $\lambda = \max\left\{\sigma\sqrt{N}, M\right\} + \upsilon\sigma$, *то для произвольного* $z \in X$

$$\text{(П.4)} \qquad \operatorname{Prob}\left\{\sum_{i=1}^{N}\langle\xi_i, z - x_{i-1}\rangle \geqslant 8(1+\tau)R\max\{\sigma\sqrt{N}, M\}\right\} \leqslant e^{-\tau}$$

и

$$\text{(П.5)} \qquad \operatorname{Prob}\left\{\sum_{i=1}^{N}\|\xi_i\|_*^2 \geqslant 8(2+3\tau)\max\{N\sigma^2, M^2\}\right\} \leqslant e^{-\tau}.$$

Доказательство. Положим $\zeta_i = \langle\xi_i, z - x_{i-1}\rangle$ и $\varsigma_i = \|\xi_i\|_*^2$, $i = 1, 2, \ldots$ Используя лемму 2, нетрудно проверить, что выполняются неравенства

$$\text{(П.6)} \qquad \begin{aligned} &\text{(a)} && |\mathbf{E}_{x_{i-1}}\{\zeta_i\}| \leqslant D\left[(M+\upsilon\sigma)(\sigma/\lambda)^2 + \sigma^2/\lambda\right], \\ &\text{(b)} && |\zeta_i| \leqslant D[2(M+\upsilon\sigma) + \lambda], \\ &\text{(c)} && (\mathbf{E}_{x_{i-1}}\{\zeta_i^2\})^{1/2} \leqslant D[\sigma + (M+\upsilon\sigma)\sigma/\lambda] \end{aligned}$$

и

$$\text{(П.7)} \qquad \begin{aligned} &\text{(a)} && \mathbf{E}_{x_{i-1}}\{\varsigma_i\} \leqslant [\sigma + (M+\upsilon\sigma)\sigma/\lambda]^2, \\ &\text{(b)} && \varsigma_i \leqslant [2(M+\upsilon\sigma) + \lambda]^2, \\ &\text{(c)} && (\mathbf{E}_{x_{i-1}}\{\varsigma_i^2\})^{1/2} \leqslant [\sigma + (M+\upsilon\sigma)\sigma/\lambda]\,[2(M+\upsilon\sigma) + \lambda]. \end{aligned}$$

Далее несколько раз применяется неравенство Бернштейна и каждый раз используются одинаковые обозначения r, A, s для величин, которые задают равномерные верхние оценки для соответственно математического ожидания, максимального модуля и среднеквадратического уклонения случайных величин.

1^o. Докажем сначала утверждение (i). Начнем со случая $M \leqslant \sigma\sqrt{\frac{N}{\tau}}$. Из (П.6) следует, что в этом случае

$$\text{(П.8)} \qquad \begin{aligned} |\mathbf{E}_{x_{i-1}}\{\zeta_i\}| &\leqslant 2D\sigma^2/\lambda \leqslant 4R\sigma\sqrt{\frac{\tau}{N}} =: r, \\ |\zeta_i| &\leqslant A := 3\lambda D \leqslant 6R\lambda, \\ (\mathbf{E}_{x_{i-1}}\{\zeta_i^2\})^{1/2} &\leqslant s := 2D\sigma \leqslant 4R\sigma. \end{aligned}$$

Используя (П.8) и неравенство Бернштейна для мартингалов (см., например, [43]), получим

$$\operatorname{Prob}\left\{\sum_{i=1}^{N}\zeta_i \geqslant 16R\sigma\sqrt{N\tau}\right\} \leqslant \operatorname{Prob}\left\{\sum_{i=1}^{N}\zeta_i \geqslant Nr + 3s\sqrt{N\tau}\right\} \leqslant$$

$$\leqslant \exp\left\{-\frac{9\tau}{2+\frac{2}{3}\frac{3\sqrt{\tau}A}{s\sqrt{N}}}\right\} \leqslant$$

$$\leqslant \exp\left\{-\frac{9\tau}{2+3\bigl(1+\upsilon\sqrt{\tau/N}\bigr)}\right\} \leqslant e^{-\tau}$$

для всех $\tau > 0$, удовлетворяющих условию $\tau \leqslant 16N/(9\upsilon^2)$. С другой стороны, в рассматриваемом случае выполняются неравенства (ср. (П.7) и (П.8))

$$\mathbf{E}_{x_{i-1}}\{\varsigma_i\} \leqslant \underbrace{4\sigma^2}_{=:r}, \quad \varsigma_i \leqslant \underbrace{9\lambda^2}_{=:A}, \quad (\mathbf{E}_{x_{i-1}}\{\varsigma_i^2\})^{1/2} \leqslant \underbrace{6\lambda\sigma}_{=:s}.$$

Таким образом,

$$Nr + 3s\sqrt{\tau N} = 4N\sigma^2 + 18\lambda\sigma\sqrt{\tau N} = 22N\sigma^2 + 18\upsilon\sigma^2\sqrt{N\tau} \leqslant 40N\sigma^2$$

для $0 < \tau \leqslant N/\upsilon^2$. Применяя снова неравенство Бернштейна, получим

$$\operatorname{Prob}\left\{\sum_{i=1}^{N}\varsigma_i \geqslant 40N\sigma^2\right\} \leqslant \exp\left\{-\frac{9\tau}{2+\bigl(3+3\upsilon\sqrt{\tau/N}\bigr)}\right\} \leqslant e^{-\tau}$$

для всех $\tau > 0$, удовлетворяющих условию $\tau \leqslant N/\upsilon^2$.

2^o. Предположим теперь, что $M > \sigma\sqrt{\frac{N}{\tau}}$, так что $\lambda = M + \upsilon\sigma$ и $\sigma^2 \leqslant$ $\leqslant M^2\tau/N$. Тогда

$$|\mathbf{E}_{x_{i-1}}\zeta_i| \leqslant 4R\sigma^2/\lambda \leqslant \underbrace{4RM\tau/N}_{=:r}, \quad |\zeta_i| \leqslant R(2(M+\upsilon\sigma)+\lambda) = \underbrace{6R(M+\upsilon\sigma)}_{=:A},$$

$$(\mathbf{E}_{x_{i-1}}\{\zeta_i^2\})^{1/2} \leqslant 4R\sigma \leqslant \underbrace{4MR\sqrt{\tau/N}}_{=:s}.$$

Далее,

$$Nr + 3s\sqrt{\tau N} = 4RM\tau + 12RM\tau = 16RM\tau,$$

и, снова применяя неравенство Бернштейна, получим

$$\operatorname{Prob}\left\{\sum_{i=1}^{N}\zeta_i \geqslant 16RM\tau\right\} \leqslant \exp\left\{-\frac{9\tau}{2+\frac{2}{3}\frac{3\sqrt{\tau}A}{s\sqrt{N}}}\right\} \leqslant$$
$$\leqslant \exp\left\{-\frac{9\tau}{2+\left(3+3\upsilon\sigma/M\right)}\right\} \leqslant$$
$$\leqslant \exp\left\{-\frac{9\tau}{5+3\upsilon\sqrt{\tau/N}}\right\} \leqslant e^{-\tau}$$

для всех $\tau > 0$, удовлетворяющих условию $\tau \leqslant 16N/(9\upsilon^2)$. Кроме того, в рассматриваемом случае

$$\mathbf{E}_{x_{i-1}}\{\varsigma_i\} \leqslant 4\sigma^2 \leqslant \underbrace{4\tau M^2/N}_{=:r},\ \ \varsigma_i \leqslant \underbrace{9\lambda^2}_{=:A},\ \ (\mathbf{E}_{x_{i-1}}\{\varsigma_i^2\})^{1/2} \leqslant \underbrace{6\lambda\sigma}_{=:s} \leqslant 6\lambda M\sqrt{\tau/N}.$$

Теперь

$$Nr + 3s\sqrt{\tau N} = 4\tau M^2 + 18\lambda\sigma\sqrt{\tau N} \leqslant 22M^2\tau + 18\upsilon\sigma^2\sqrt{N\tau} \leqslant 40M^2\tau$$

при $0 < \tau \leqslant N/\upsilon^2$. Применяя еще раз неравенство Бернштейна, получим

$$\operatorname{Prob}\left\{\sum_{i=1}^{N}\varsigma_i \geqslant 40\tau M^2\right\} \leqslant \exp\left\{-\frac{9\tau}{2+\left(3+3\upsilon\sqrt{\tau/N}\right)}\right\} \leqslant e^{-\tau}$$

для всех $\tau > 0$, удовлетворяющих условию $\tau \leqslant N/\upsilon^2$.

3^o. Теперь рассмотрим случай $\lambda = \max\{M, \sigma\sqrt{N}\} + \sigma\upsilon$, и пусть $M \leqslant \sigma\sqrt{N}$, так что $\lambda = \sigma(\sqrt{N} + \upsilon)$. Рассуждаем точно так же, как в доказательстве (i). В силу (П.6) имеем

$$|\mathbf{E}_{x_{i-1}}\{\zeta_i\}| \leqslant 4R\frac{\sigma}{\sqrt{N}} =: r,$$
$$(\mathbf{E}_{x_{i-1}}\{\zeta_i^2\})^{1/2} \leqslant 4R\sigma =: s,$$
$$|\zeta_i| \leqslant 6R\lambda \leqslant 12R\sigma\sqrt{N} = 3s\sqrt{N},$$

так что, используя неравенство Бернштейна, получим

$$\operatorname{Prob}\left\{\sum_{i=1}^{N}\zeta_i \geqslant 8R\sigma\sqrt{N}(\tau+1)\right\} \leqslant \operatorname{Prob}\left\{\sum_{i=1}^{N}\zeta_i \geqslant Nr + (2\tau+1)s\sqrt{N}\right\} \leqslant$$
$$\leqslant \exp\left\{-\frac{(2\tau+1)^2 s^2 N}{2s^2N + \frac{2}{3}3s^2N(2\tau+1)}\right\} \leqslant \exp\left\{-\frac{(2\tau+1)^2}{2+2(2\tau+1)}\right\} \leqslant e^{-\tau}.$$

Из (П.7) также имеем

$$\mathbf{E}_{x_{i-1}}\{\varsigma_i\} \leqslant \underbrace{4\sigma^2}_{=:r},$$

$$(\mathbf{E}_{x_{i-1}}\{\varsigma_i^2\})^{1/2} \leqslant 6\lambda\sigma \leqslant 12\sigma^2\sqrt{N} =: s,$$

$$\varsigma_i \leqslant 9\lambda^2 \leqslant 36\sigma^2 N = 4s\sqrt{N}.$$

Теперь, снова применив неравенство Бернштейна, получим

$$\operatorname{Prob}\left\{\sum_{i=1}^N \varsigma_i \geqslant 16N\sigma^2 + 24N\sigma^2\tau\right\} = \operatorname{Prob}\left\{\sum_{i=1}^N \varsigma_i \geqslant Nr + (2\tau+1)s\sqrt{N}\right\} \leqslant$$

$$\leqslant \exp\left\{-\frac{(2\tau+1)s^2N}{[2+2(2\tau+1)]s^2N}\right\} \leqslant e^{-\tau}.$$

Доказательство границ (П.4) и (П.5) в случае $M > \sigma\sqrt{N}$ и $\lambda = M + \sigma\upsilon$ вытекает из таких же рассмотрений. □

П.2. Доказательство предложения 1. Докажем сначала неравенство (13). Ввиду (8) условие оптимальности в (9) имеет вид

$$\langle y_{i+1} + \psi'(x_{i+1}) + \beta_i[\vartheta'(x_{i+1}) - \vartheta'(x_i)], z - x_{i+1}\rangle \geqslant 0, \ \ \forall\, z \in X,$$

или в эквивалентной форме

$$\langle y_{i+1} + \psi'(x_{i+1}), x_{i+1} - z\rangle \leqslant \beta_i\langle[\vartheta'(x_{i+1}) - \vartheta'(x_i)], z - x_{i+1}\rangle =$$

$$= \langle\beta_i V'_{x_i}(x_{i+1}), z - x_{i+1}\rangle =$$

$$= \beta_i[V_{x_i}(z) - V_{x_{i+1}}(z) - V_{x_i}(x_{i+1})], \ \ \forall\, z \in X,$$

где последнее равенство вытекает из следующего замечательного тождества (см., например, [44]): для любых u, u' и $w \in X$

$$\langle V'_u(u'), w - u'\rangle = V_u(w) - V_{u'}(w) - V_u(u').$$

Принимая во внимание определение $\xi_i = y_i - \nabla\phi(x_{i-1})$, получим

$$\langle\nabla\phi(x_i), x_{i+1} - z\rangle + \langle\psi'(x_{i+1}), x_{i+1} - z\rangle \leqslant \beta_i[V_{x_i}(z) - V_{x_{i+1}}(z) - V_{x_i}(x_{i+1})] -$$

$$-\langle\xi_{i+1}, x_{i+1} - z\rangle. \tag{П.9}$$

Из леммы 1 и условия $\beta_i \geqslant 2L$ следует, что

$$F(x_{i+1}) - F(z) \leqslant \varepsilon_{i+1}(z) \leqslant$$

$$\leqslant \langle\nabla\phi(x_i), x_{i+1} - z\rangle + \langle\psi'(x_{i+1}), x_{i+1} - z\rangle + \frac{\beta_i}{2}V_{x_i}(x_{i+1}).$$

Вместе с (П.9) это неравенство влечет, что

$$\varepsilon_{i+1}(z) \leqslant \beta_i\left[V_{x_i}(z) - V_{x_{i+1}}(z) - \frac{1}{2}V_{x_i}(x_{i+1})\right] - \langle\xi_{i+1}, x_{i+1} - z\rangle.$$

С другой стороны, благодаря сильной выпуклости $V_x(\cdot)$ имеем

$$\langle \xi_{i+1}, z - x_{i+1}\rangle - \frac{\beta_i}{2}V_{x_i}(x_{i+1}) = \langle \xi_i, z - x_i\rangle + \langle \xi_{i+1}, x_i - x_{i+1}\rangle - \frac{\beta_i}{2}V_{x_i}(x_{i+1})$$
$$\leqslant \langle \xi_{i+1}, z - x_i\rangle + \frac{\|\xi_{i+1}\|_*^2}{\beta_i}.$$

Соединяя эти неравенства, получим, что

$$\begin{aligned} &F(x_{i+1}) - F(z) \leqslant \varepsilon_{i+1}(z) \leqslant \\ &\leqslant \beta_i[V_{x_i}(z) - V_{x_{i+1}}(z)] - \langle \xi_{i+1}, x_i - z\rangle + \frac{\|\xi_{i+1}\|_*^2}{\beta_i} \end{aligned} \tag{П.10}$$

при всех $z \in X$. Деля (П.10) на β_i и затем суммируя по i от 0 до $N-1$, получим (13).

Докажем теперь границу (14). Применяя лемму 6.1 из [1] с $z_0 = x_0$, будем иметь

$$\forall z \in X, \quad \sum_{i=1}^N \beta_{i-1}^{-1}\langle \xi_i, z - z_{i-1}\rangle \leqslant V_{x_0}(z) + \frac{1}{2}\sum_{i=1}^N \beta_{i-1}^{-2}\|\xi_i\|_*^2, \tag{П.11}$$

где $z_i = \arg\min_{z\in X}\left\{\mu_{i-1}\langle \xi_i, z\rangle + V_{z_{i-1}}(z)\right\}$ зависит только от $z_0, \xi_1, \ldots, \xi_i$. Далее,

$$\sum_{i=1}^N \beta_{i-1}^{-1}\langle \xi_i, z - x_{i-1}\rangle = \sum_{i=1}^N \beta_{i-1}^{-1}[\langle \xi_i, z_{i-1} - x_{i-1}\rangle + \langle \xi_i, z - z_{i-1}\rangle] \leqslant$$
$$\leqslant V_{x_0}(z) + \sum_{i=1}^N \beta_{i-1}^{-1}\langle \xi_i, z_{i-1} - x_{i-1}\rangle + \frac{1}{2}\beta_{i-1}^{-2}\|\xi_i\|_*^2.$$

Соединяя это неравенство с (13), получим (14). □

П.3. Доказательство следствия 1. Заметим, что (15) является непосредственным следствием (13) и оценок для моментов $\|\xi_i\|_*$, приведенных в лемме 2. В самом деле, (П.1)(b) влечет, что в условиях следствия 1

$$|\mathbf{E}_{x_{i-1}}\{\langle \xi_i, z - x_{i-1}\rangle\}| \leqslant D\left[(M + \upsilon\sigma)\left(\frac{\sigma}{\lambda}\right)^2 + \frac{\sigma^2}{\lambda}\right] \leqslant \frac{2D\sigma^2}{\lambda} \leqslant \frac{2D\sigma}{\sqrt{N}}.$$

Далее, ввиду (П.1)(c)

$$\mathbf{E}_{x_{i-1}}\{\|\xi_i\|_*^2\}^{1/2} \leqslant \sigma + (M + \upsilon\sigma)\frac{\sigma}{\lambda} \leqslant 2\sigma.$$

Беря математическое ожидание от обеих частей (13) и используя два последних неравенства, получим (15). Граница (17) доказывается аналогичным образом, с той лишь разницей, что вместо неравенства (13) используется (14). □

П.4. Доказательство теоремы 1. В силу утверждения (i) леммы 3 при условии $\tau \leqslant N/\upsilon^2$ с вероятностью не менее $1 - 2e^{-\tau}$ имеем

$$\sum_{i=1}^{N} \langle \xi_i, z_{i-1} - x_{i-1} \rangle \leqslant 16R \max\{\sigma\sqrt{N\tau}, M\tau\},$$

$$\sum_{i=1}^{N} \|\xi_i\|_*^2 \leqslant 40 \max\{N\sigma^2, M^2\tau\}.$$

Подставляя эти оценки в неравенство (14), получим, что с вероятностью не менее $1 - 2e^{-\tau}$ верно следующее:

$$\bar{\beta} \sup_{z \in X} \varepsilon(x^N, z) \leqslant 2\bar{\beta} V_{x_0}(z) + \sum_{i=1}^{N} \left[\langle \xi_i, z_{i-1} - x_{i-1} \rangle + \frac{3}{2} \bar{\beta}^{-1} \|\xi_i\|_*^2 \right] \leqslant$$

$$\leqslant 2\bar{\beta} R^2 \Theta + 16R \max\{\sigma\sqrt{N\tau}, M\tau\} + 60\bar{\beta}^{-1} \max\{N\sigma^2, M^2\tau\}.$$

Далее, полагая $\bar{\beta} = \max\left\{2L, \frac{\sigma}{R}\sqrt{\frac{N}{\Theta}}\right\}$, будем иметь

$$N[F(\widehat{x}_N) - F(z)] \leqslant \max\{4LR^2\Theta, 2\sigma R\sqrt{N\Theta}\} + 16R \max\{\sigma\sqrt{N\tau}, M\tau\} +$$

$$+ \ 60 \max\{LR^2\tau/2, \sigma R\sqrt{N\Theta}\} \leqslant$$

$$\leqslant \max\{46LR^2\tau, 4LR^2\Theta, 62\sigma R\sqrt{N\Theta}, 16\sigma R\sqrt{N\tau}\} \tag{П.12}$$

при $1 \leqslant \tau \leqslant N/\upsilon^2$. Это влечет (19). □

П.5. Доказательство теоремы 2. Действуем точно так же, как при доказательстве теоремы 1 с той только разницей, что вместо утверждения (i) леммы 3 используем утверждения (ii) той же леммы, которое влечет, что при условии $N \geqslant \upsilon^2$ с вероятностью не менее $1 - 2e^{-\tau}$ выполнены неравенства

$$\sum_{i=1}^{N} \langle \xi_i, z - x_{i-1} \rangle \leqslant 8(1+\tau)R \max\{\sigma\sqrt{N}, M\},$$

$$\sum_{i=1}^{N} \|\xi_i\|_*^2 \leqslant 8(2+3\tau) \max\{N\sigma^2, M^2\}.$$ □

П.6. Доказательство предложения 2. Обозначим:

$$\rho_N(\tau; \mu, \nu) = \nu^{-1} R^2 \Theta + 16R \max\{\sigma\sqrt{N\tau}, M\tau\} +$$

$$+ \ 20(\mu + \nu) \max\{N\sigma^2, M^2\tau\} + \mu^{-1} \sum_{i=1}^{N} V_{x_{i-1}}(x_i).$$

Утверждение предложения является непосредственным следствием следующего результата.

Лемма 4. Положим

$$\bar\rho_N(\tau) = \min_{\mu,\nu>0}\rho_N(\tau;\mu,\nu) =$$

$$= 4R\sqrt{5\Theta\max\{N\sigma^2, M^2\tau\}} + 16R\max\left\{\sigma\sqrt{N\tau}, M\tau\right\} +$$

$$+\min_{\mu>0}\left\{20\mu\max\left\{N\sigma^2, M^2\tau\right\} + \mu^{-1}\sum_{i=1}^{N} V_{x_{i-1}}(x_i)\right\}. \tag{П.13}$$

Тогда при $0 < \tau \leqslant N/\upsilon^2$ *и* $t \geqslant L$ *справедливы неравенства*

$$\begin{aligned} &\text{(a)}\quad \operatorname{Prob}\{\epsilon_N(t) - \widehat{\epsilon}_N(t) \geqslant \bar\rho_N(\tau)/N\} \leqslant 2\mathrm{e}^{-\tau},\\ &\text{(b)}\quad \operatorname{Prob}\{\widehat{\epsilon}_N(t) - \epsilon_N(t) \geqslant \bar\rho_N(\tau)/N\} \leqslant 2\mathrm{e}^{-\tau}. \end{aligned} \tag{П.14}$$

Доказательство леммы. Докажем первое неравенство в (П.14). Напомним, что $\xi_i = y_i - \nabla\phi(x_{i-1})$, $i = 1,\ldots,N$. Ввиду сильной выпуклости $V_x(\cdot)$ имеем для любых $z \in X$ и $\mu \geqslant 0$

$$\langle\xi_i, z - x_i\rangle = \langle\xi_i, z - x_{i-1}\rangle + \langle\xi_i, x_{i-1} - x_i\rangle \leqslant$$

$$\leqslant \langle\xi_i, z - x_{i-1}\rangle + \frac{\mu}{2}\|\xi_i\|_*^2 + \frac{1}{\mu}V_{x_{i-1}}(x_i).$$

Таким образом, для любых $\nu > 0$

$$\sum_{i=1}^{N}\langle\xi_i, z - x_i\rangle \leqslant \sum_{i=1}^{N}\left[\langle\xi_i, z - x_{i-1}\rangle + \frac{\mu}{2}\|\xi_i\|_*^2 + \frac{1}{\mu}V_{x_{i-1}}(x_i)\right] \leqslant$$

$$\leqslant \frac{1}{\nu}V_{x_0}(z) + \sum_{i=1}^{N}\left[\frac{\nu}{2}\|\xi_i\|_*^2 + \langle\xi_i, z_{i-1} - x_{i-1}\rangle + \frac{1}{\mu}V_{x_{i-1}}(x_i) + \frac{\mu}{2}\|\xi_i\|_*^2\right]$$

(для получения последнего неравенства использовалась лемма 6.1 из [1] с $z_0 = x_0$ так же, как в доказательстве предложения 1). По лемме 3 существует множество $\mathcal{A}_N$ в пространстве реализаций ω^N с вероятностью не менее $1 - 2\mathrm{e}^{-\tau}$, такое что для всех $\omega^N \in \mathcal{A}_N$

$$\sum_{i=1}^{N}\langle\xi_i, z_{i-1} - x_{i-1}\rangle \leqslant 16R\max\{\sigma\sqrt{N\tau}, M\tau\}$$

и

$$\sum_{i=1}^{N}\|\xi_i\|_*^2 \leqslant 40\max\{N\sigma^2, M^2\tau\}.$$

Вспоминая, что $V_{x_0}(z) \leqslant R^2\Theta$, приходим к заключению, что при всех $z \in X$ и всех $\omega^N \in \mathcal{A}_N$ верно неравенство

$$\sum_{i=1}^{N}\langle\xi_i, z - x_i\rangle \leqslant \rho_N(\tau;\mu,\nu).$$

Следовательно, для $\omega^N \in \mathcal{A}_N$ имеем

$$\epsilon_N(t) - \widehat{\epsilon}_N(t) = N^{-1} \sup_{z \in X} \sum_{i=1}^{N} \langle \xi_i, z - x_i \rangle \leqslant N^{-1} \min_{\mu,\nu \geqslant 0} \rho_N(\tau; \mu, \nu) = N^{-1} \bar{\rho}_N(\tau),$$

что доказывает первое неравенство в (П.14). Доказательство второго неравенства в (П.14) аналогично и потому опускается. □

П.7. Доказательство следствия 2. Из определения $\epsilon_N(\cdot)$ заключаем, что

$$\bar{\beta} \sum_{i=1}^{N} V_{x_{i-1}}(x_i) \leqslant \epsilon_N(\bar{\beta}),$$

и получаем (24), подставляя $\mu = 1/\bar{\beta}$. С другой стороны, можно убедиться, что для $\bar{\beta} \geqslant \max\left\{2L, \frac{\sigma\sqrt{N}}{R\sqrt{\Theta}}\right\}$ выполняется неравенство

$$\bar{\rho}_N(\tau) \leqslant N\epsilon_N(\bar{\beta}) + \max\left\{[(20+4\sqrt{5})\sqrt{\Theta} + 16\sqrt{\tau}]R\sigma\sqrt{N}, (4\sqrt{5\Theta\tau} + 26\tau)LR^2\right\} \leqslant$$

$$\leqslant N\epsilon_N(\bar{\beta}) + C_1 R\sigma\sqrt{N[\Theta \vee \tau]} + C_2 LR^2[\Theta \vee \tau].$$

Наконец, так как $\widehat{\epsilon}_N(\bar{\beta}) \leqslant \epsilon_N(\bar{\beta}) + \bar{\rho}_N(\tau)/N$ с вероятностью не менее $1 - 2\mathrm{e}^{-\tau}$ (ср. (П.14)(b)), то

$$\Delta_N(\tau, \bar{\beta}) = \widehat{\epsilon}_N(\bar{\beta}) + \bar{\rho}_N(\tau)/N \leqslant \epsilon_N(\bar{\beta}) + 2\bar{\rho}_N(\tau)/N$$

с той же вероятностью. Это влечет (25). □

П.8. Доказательство теоремы 3.

1^o. Покажем сначала, что для каждого $k = 1, \ldots, m = m(N)$ верно следующее.

Утверждение I_k. Существует случайное событие $\mathcal{B}_k \subseteq \Omega^{\otimes N}$ вероятности не менее $1 - 2ke^{-\tau}$ такое, что для всех $\omega^N \in \mathcal{B}_k$ выполнены неравенства

$$\text{(П.15)} \qquad \begin{aligned} &\text{(a)} \quad \|y_k - \bar{x}(y_k)\|^2 \leqslant r_k^2 = 2^{-k} r_0^2 \ \ \textit{для некоторого } \bar{x}(y_k) \in X_*, \\ &\text{(b)} \quad F(y_k) - F_* \leqslant \frac{\kappa}{2} r_k^2 = 2^{-k-1} \kappa r_0^2. \end{aligned}$$

Доказательство утверждения I_k проведем по индукции. Заметим, что (П.15)(a) выполняется с вероятностью 1 для $k = 0$. Положим $\mathcal{B}_0 = \Omega^{\otimes N}$. Предположим, что (П.15)(a) справедливо для некоторого $k \in \{0, \ldots, m-1\}$ с вероятностью не менее $1 - 2ke^{-\tau}$, и покажем, что тогда верно утверждение I_{k+1}.

Обозначим: $F_*^k = \min\limits_{x \in X_{r_k}(y_k)} F(x)$. Пусть X_*^k — множество точек минимума F на $X_{r_k}(y_k)$. В силу теоремы 1 и определения N_k (ср. (27)) существует событие $\mathcal{A}_k$ вероятности не менее $1 - 2e^{-\tau}$ такое, что для $\omega^N \in \mathcal{A}_k$ после $(k+1)$-го

этапа алгоритма имеем

$$\frac{\kappa}{2}\|y_{k+1}-\bar{x}_k(y_{k+1})\|^2 \leqslant F(y_{k+1})-F_*^k \leqslant$$
$$\leqslant \max\left\{C_1\frac{Lr_k^2[\tau\vee\Theta]}{N_{k+1}}, C_2\sigma r_k\sqrt{\frac{[\tau\vee\Theta]}{N_{k+1}}}\right\}\leqslant$$
$$\leqslant \frac{\kappa}{4}r_k^2=\frac{\kappa}{2}r_{k+1}^2,$$

где $\bar{x}_k(y_{k+1})$ — проекция y_{k+1} на X_*^k. Положим теперь $\mathcal{B}_{k+1}=\mathcal{B}_k\cap\mathcal{A}_k$. Тогда

$$\mathrm{Prob}\{\mathcal{B}_{k+1}\}\geqslant\mathrm{Prob}\{\mathcal{B}_k\}+\mathrm{Prob}\{\mathcal{A}_k\}-1\geqslant 1-2(k+1)e^{-\tau}.$$

Кроме того, по индуктивному предположению на $\mathcal{B}_k$ (и, следовательно, на $\mathcal{B}_{k+1}$) имеем

$$\|y_k-\bar{x}(y_k)\|\leqslant r_k,$$

т.е. расстояние от y_k до множества точек глобального минимума X_* не превосходит r_k, и, значит множество $X_{r_k}(y_k)$ имеет непустое пересечение с X_*. Следовательно, $X_*^k\subseteq X_*$, точка $\bar{x}_k(y_{k+1})$ содержится в X_* и F_*^k совпадает с оптимальным значением F_* исходной задачи. Отсюда заключаем, что

$$\frac{\kappa}{2}\|y_{k+1}-\bar{x}(y_{k+1})\|^2\leqslant F(y_{k+1})-F_*\leqslant\frac{\kappa}{2}r_{k+1}^2=2^{-k}\kappa r_0^2$$

для некоторого $\bar{x}(y_{k+1})\in X_*$.

2^o. Докажем теорему в случае, когда $\overline{N}_1\geqslant 1$. Это условие эквивалентно тому, что $\overline{N}_k\geqslant 1$ при всех $k=1,\ldots,m(N)$, ибо по построению $\overline{N}_1\leqslant\overline{N}_2\leqslant\cdots$ $\cdots\leqslant\overline{N}_{m(N)}$. Предположим, что $\omega^N\in\mathcal{B}_{m(N)}$, так что (П.15) выполняется с $k=m(N)$. Так как $\overline{N}_k\geqslant 1$, то $N_k\leqslant 2\overline{N}_k$. Кроме того, $\overline{N}_{k+1}\leqslant 2\overline{N}_k$. Учитывая эти замечания и определение $m(N)$, получим

$$\text{(П.16)}\qquad N\leqslant\sum_{k=1}^{m(N)+1}N_k\leqslant 2\sum_{k=1}^{m(N)+1}\overline{N}_k\leqslant 2\sum_{k=1}^{m(N)}\overline{N}_k+4\overline{N}_{m(N)}\leqslant 6\sum_{k=1}^{m(N)}\overline{N}_k.$$

Итак, в силу определения $\overline{N}_k$ (ср. (27)) имеем

$$N\leqslant 6\sum_{k=1}^{m(N)}\max\left\{4C_1\frac{L[\tau\vee\Theta]}{\kappa},\,16C_2\frac{\sigma^2[\tau\vee\Theta]}{\kappa^2r_{k-1}^2}\right\}\leqslant$$
$$\leqslant\underbrace{24\sum_{k=1}^{\bar{k}-1}\frac{C_1L[\tau\vee\Theta]}{\kappa}}_{S_1}+\underbrace{96\sum_{k=\bar{k}}^{m(N)}\frac{C_2\sigma^2[\tau\vee\Theta]}{\kappa^2r_{k-1}^2}}_{S_2},$$

где

$$\bar{k} = \min\left\{k:\ \frac{4C_2\sigma^2}{\kappa r_{k-1}^2} \geqslant C_1 L\right\}.$$

Возможны два случая: либо $S_1 \geqslant N/2$, либо $S_2 \geqslant N/2$. Если $S_1 \geqslant N/2$, то

$$\bar{k} \geqslant \frac{C'\kappa N}{L[\tau \vee \Theta]},$$

так что если $\omega^N \in \mathcal{B}_{m(N)}$, то

$$\begin{aligned} F(y_{m(N)}) - F_* &\leqslant \frac{\kappa}{2} r_{m(N)}^2 \leqslant \frac{\kappa}{2} r_{\bar{k}}^2 = \\ &= 2^{-\bar{k}-1}\kappa r_0^2 \leqslant C\kappa r_0^2 \exp\left\{-\frac{C'\kappa N}{L[\tau \vee \Theta]}\right\}. \end{aligned} \tag{П.17}$$

Если же $S_2 \geqslant N/2$, то справедливы неравенства

$$\frac{\kappa^2 r_0^2 N}{\sigma^2[\tau \vee \Theta]} \leqslant \frac{C\kappa^2 r_0^2}{\sigma^2[\tau \vee \Theta]} \sum_{k=\bar{k}}^{m(N)} 2^k \frac{\sigma^2[\tau \vee \Theta]}{\kappa^2 r_0^2} \leqslant C' 2^{m(N)-\bar{k}}.$$

Следовательно, в этом случае для $\omega^N \in \mathcal{B}_{m(N)}$ имеем

$$\begin{aligned} F(y_{m(N)}) - F_* &\leqslant \frac{\kappa}{2} r_{m(N)}^2 = \\ &= \frac{\kappa}{2} r_{\bar{k}}^2 2^{-m(N)+\bar{k}} \leqslant \frac{\kappa}{2} r_0^2 2^{-m(N)+\bar{k}} \leqslant C\frac{\sigma^2[\tau \vee \Theta]}{\kappa N}. \end{aligned} \tag{П.18}$$

3^o. Наконец, рассмотрим случай, когда

$$\overline{N}_1 := \max\left\{4C_1\frac{L[\tau \vee \Theta]}{\kappa},\ 16C_2\frac{\sigma^2[\tau \vee \Theta]}{\kappa^2 r_0^2}\right\} < 1. \tag{П.19}$$

Пусть $k_* \geqslant 2$ — наименьшее целое k такое, что $\overline{N}_k \geqslant 1$. Если $k_* > N/4$, то нетрудно видеть, что $m(N) \geqslant N/4$, и, следовательно, для $\omega^N \in \mathcal{B}_{m(N)}$ имеем

$$F(y_{m(N)}) - F_* \leqslant \frac{\kappa}{2} r_{m(N)}^2 \leqslant \frac{\kappa}{2} r_0^2\, 2^{-N/4}. \tag{П.20}$$

Если же $2 \leqslant k_* \leqslant N/4$, то верна цепочка неравенств

$$\begin{aligned} 3\sum_{k=k_*}^{m(N)} \overline{N}_k &\geqslant \overline{N}_{m(N)+1} + \sum_{k=k_*}^{m(N)} \overline{N}_k = \\ = \sum_{k=1}^{m(N)+1} \overline{N}_k - \sum_{k=1}^{k_*-1} \overline{N}_k &\geqslant \sum_{k=1}^{m(N)+1} \overline{N}_k - N/4 \geqslant N/4, \end{aligned}$$

где в первом неравенстве учтено, что $\overline{N}_{m(N)+1} \leqslant 2\overline{N}_{m(N)}$, а последнее следует из определения $m(N)$. Используя эти замечания и то, что $\overline{N}_k/2^k \leqslant \overline{N}_{k_*}/2^{k_*}$ при $k \geqslant k_*$, находим

$$\frac{N}{12} \leqslant \sum_{k=k_*}^{m(N)} \overline{N}_k \leqslant \sum_{k=k_*}^{m(N)} 2^{k-k_*}\overline{N}_{k_*} \leqslant 2^{m(N)-k_*+2},$$

где в последнем неравенстве учтено, что $\overline{N}_{k_*} \leqslant 2\overline{N}_{k_*-1} < 2$. Итак, принимая во внимание (П.15)(b), для $\omega^N \in \mathcal{B}_{m(N)}$ будем иметь

$$F(y_{m(N)}) - F_* \leqslant \kappa r_{k_*}^2 2^{-m(N)+k_*} \leqslant \kappa r_0^2 2^{-m(N)+k_*} \leqslant C\kappa r_0^2/N.$$

Соединяя последнюю оценку с (П.17), (П.18) и (П.20), получим (28). □

СПИСОК ЛИТЕРАТУРЫ

1. *Nemirovski A., Juditsky A., Lan G., Shapiro A.* Robust stochastic approximation approach to stochastic programming // SIAM J. Optim. 2009. V. 19. No. 4. P. 1574–1609.
2. *Juditsky A., Nesterov Y.* Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // Stochast. Syst. 2014. V. 4. No. 1. P. 44–80.
3. *Ghadimi S., Lan G.* Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework // SIAM J. Optim. 2012. V. 22. V. 4. P. 1469–1492.
4. *Tukey J.W.* A survey of sampling from contaminated distributions / I. Olkin, Contribut. Probab. Statist. 1960. P. 448–485.
5. *Huber P.J.* Robust estimation of a location parameter // Ann. Math. Statist. 1964. P. 73–101.
6. *Huber P.J.* The 1972 Wald lecture. Robust statistics: A review // Ann. Math. Statist. 1972. V. 43. No. 4. P. 1041–1067.
7. *Хьюбер П.* Робастность в статистике. М.: Мир, 1984.
8. *Martin R., Masreliez C.* Robust estimation via stochastic approximation // IEEE Transact. Inform. Theory. 1975. V. 21. No. 3. P. 263–271.
9. *Поляк Б.Т., Цыпкин Я.З.* Адаптивные алгоритмы оценивания (сходимость, оптимальность, стабильность) // АиТ. 1979. № 3. С. 71–84.
   Polyak B., Tsypkin Ya.Z. Adaptive Estimation Algorithms: Convergence, Optimality, Stability // Autom. Remote Control. 1979. V. 40. No. 3. P. 378–389.
10. *Поляк Б.Т., Цыпкин Я.З.* Робастные псевдоградиентные алгоритмы адаптации // АиТ. 1980. № 10. С. 91–97.
    Polyak B.T., Tsypkin Ya.Z. Robust Pseudogradient Adaptation Algorithms // Autom. Remote Control. 1981. V. 41. No. 10. P. 1404–1409.
11. *Polyak B., Tsypkin J.Z.* Robust identification. Automatica. 1980. V. 16. No. 1. P. 53–63.
12. *Price E., VandeLinde V.* Robust estimation using the Robbins-Monro stochastic approximation algorithm // IEEE Transact. Inform. Theory. 1979. V. 25. No. 6. P. 698–704.

13. *Stanković S.S., Kovačević B. D.* Analysis of robust stochastic approximation algorithms for process identification // Automatica. 1986. V. 22. No. 4. P. 483–488.
14. *Chen H.-F., Guo L., Gao A.J.* Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds // Stochast. Proc. Appl. 1988. V. 27. No. 2. P. 217–231.
15. *Chen H.-F., Gao A.J.* Robustness analysis for stochastic approximation algorithms // Stochastics. 1988. V. 26. No. 1. P. 3–20.
16. *Nazin A.V., Polyak B.T., Tsybakov A.B.* Optimal and robust kernel algorithms for passive stochastic approximation // IEEE Transact. Inform. Theory. 1992. V. 38. No. 5. P. 1577–1583.
17. *Цыпкин Я.З.* Основы информационной теории идентификации. М.: Наука, 1984.
18. *Цыпкин Я.З.* Информационная теория идентификации. М.: Наука, 1995.
19. *Kwon J., Lecué G., Lerasle M.* Median of means principle as a divide-and-conquer procedure for robustness, sub-sampling and hyper-parameters tuning // arXiv:1812.02435.
20. *Chinot G., Lecué G., Lerasle M.* Statistical learning with Lipschitz and convex loss functions // arXiv:1810.01090.
21. *Lecué G., Lerasle M.* Robust machine learning by median-of-means: theory and practice // arXiv preprint. arXiv:1711.10306. Annals of Stat., 2017, to appear.
22. *Lecué G., Lerasle M., Mathieu T.* Robust classification via MOM minimization // arXiv preprint, 2018. arXiv:1808.03106.
23. *Lerasle M., Oliveira R.I.* Robust empirical mean estimators // arXiv preprint, 2011. arXiv:1112.3914.
24. *Lugosi G., Mendelson S.* Risk minimization by median-of-means tournaments // arXiv preprint, 2016. arXiv:1608.00757.
25. *Lugosi G., Mendelson S.* Regularization, sparse recovery, and median-of-means tournaments // arXiv preprint, 2017. arXiv:1701.04112.
26. *Lugosi G., Mendelson S.* Near-optimal mean estimators with respect to general norms // arXiv preprint, 2018. arXiv:1806.06233.
27. *Hsu D., Sabato S.* Loss minimization and parameter estimation with heavy tails // J. Machin. Learning Res. 2016. V. 17. No. 1. P. 543–582.
28. *Bubeck S., Cesa-Bianchi N., Lugosi G.* Bandits with heavy tail // IEEE Transact. Inform. Theory. 2013. V. 59. No. 11. P. 7711–7717.
29. *Devroye L., Lerasle M., Lugosi G., Oliveira R.I.* Sub-gaussian mean estimators // Ann. Statist. 2016. V. 44. No. 6. P. 2695–2725.
30. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.
31. *Lugosi G., Mendelson S.* Sub-gaussian estimators of the mean of a random vector // Ann. Statist. 2019. V. 47. No. 2. P. 783–794.
32. *Catoni O.* Challenging the empirical mean and empirical variance: a deviation study // Ann. l'IHP Probab. Statist. 2012. V. 48. No. 4. P. 1148–1185.
33. *Audibert J.-Y., Catoni O.* Robust linear least squares regression // Ann. Statist. 2011. V. 39. No. 5. P. 2766–2794.
34. *Minsker S.* Geometric median and robust estimation in Banach spaces // Bernoulli. 2015. V. 21. No. 4. P. 2308–2335.
35. *Wei X., Minsker S.* Estimation of the covariance structure of heavy-tailed distributions / Advances in Neural Information Processing Systems. 2017. P. 2859–2868.

36. *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, ACM, 2017. V. 1. No. 2. P. 44.

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

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

39. *Cardot H., Cénac P., Godichon-Baggioni A.* Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls // Ann. Statist. 2017. V. 45. No. 2. P. 591–614.

40. *Lan G.* An optimal method for stochastic composite optimization // Math. Programm. 2012. V. 133. No. 1–2. P. 365–397.

41. *Necoara I., Nesterov Y., Glineur F.* Linear convergence of first order methods for non-strongly convex optimization // Math. Programm. 2018. P. 1–39.

42. *Juditsky A., Nemirovski A.* First order methods for nonsmooth convex large-scale optimization, I: general purpose methods / Sra, S., Nowozin, S., Wright, S.J. (eds.) Optimization for Machine Learning, MIT Press, Cambridge, MA, 2011. P. 121–148.

43. *Freedman D.A.* On tail probabilities for martingales // Ann. Probab. 1975. V. 3. No. 1. P. 100–118.

44. *Chen G., Teboulle M.* Convergence analysis of a proximal-like minimization algorithm using Bregman functions // SIAM J. Optim. 1993. V. 3. No. 3. P. 538–543.

Статья представлена к публикации членом редколлегии П.В. Пакшиным.

Поступила в редакцию 18.07.2018

После доработки 03.09.2018

Принята к публикации 08.11.2018