



Math-Net.Ru

All Russian mathematical portal

D. M. Dvinskikh, S. S. Omelchenko, A. V. Gasnikov, A. I. Turin, Accelerated gradient sliding for minimizing a sum of functions, *Dokl. RAN. Math. Inf. Proc. Upr.*, 2020, Volume 492, 85–88

DOI: 10.31857/S268695432003008X

Use of the all-Russian mathematical portal Math-Net.Ru implies that you have read and agreed to these terms of use

<http://www.mathnet.ru/eng/agreement>

Download details:

IP: 176.57.72.126

August 4, 2023, 18:21:08



УДК 519.853.62

## УСКОРЕННЫЙ ГРАДИЕНТНЫЙ СЛАЙДИНГ-МЕТОД В ЗАДАЧАХ МИНИМИЗАЦИИ СУММЫ ФУНКЦИЙ

© 2020 г. Д. М. Двинских<sup>1</sup>, С. С. Омельченко<sup>2,\*</sup>, А. В. Гасников<sup>1,\*\*</sup>, А. И. Тюрин<sup>3</sup>

Представлено академиком РАН Ю.Г. Евтушенко 10.03.2020 г.

Поступило 20.03.2020 г.

После доработки 26.03.2020 г.

Принято к публикации 03.04.2020 г.

Предложен новый способ обоснования ускоренного градиентного слайдинга Дж. Лана, позволяющий распространить технику слайдинга на сочетание ускоренных градиентных методов с ускоренными методами редукции дисперсии. Получены новые оптимальные оценки для решения задач минимизации суммы гладких сильно выпуклых функций с гладким регуляризатором.

*Ключевые слова:* ускоренный градиентный слайдинг Дж. Лана, ускоренные методы редукции дисперсии, гладкие сильно выпуклые функции

**DOI:** 10.31857/S268695432003008X

Многие задачи анализа данных (машинного обучения) приводят к необходимости решения задач минимизации функции вида суммы (эмпирический риск) с большим числом слагаемых, отвечающих объему выборки [1, 4, 14, 15]. В последнее десятилетие активно развиваются численные методы оптимизации функции вида суммы [1, 4, 6, 9]. В частности, были получены оптимальные методы (ускоренные методы редукции дисперсии) для такого класса задач, когда слагаемые в сумме – гладкие (сильно) выпуклые функции, см., например, [9]. Были исследованы задачи, в которых дополнительно в минимизируемую функцию вносится аддитивно, возможно, негладкий, но выпуклый/сильно выпуклый композитный член (по терминологии анализа данных вносится слагаемое, отвечающее “регуляризации”), являющийся проксимально дружественным [1, 10], т.е. задача минимизации такого члена с квадратичной добавкой – простая задача. В настоящей работе предлагается способ получения оптимальных оценок для случая, когда композит-

ный член будет выпуклым (сильно выпуклым) гладким, но уже не будет проксимально дружественным. Не предполагается проксимальная дружелюбность и у слагаемых в сумме.

В разделе 1 техника ускоренного градиентного слайдинга Дж. Лана [9, section 8.2] будет объяснена с помощью популярной в последнее время конструкции катализ [1, 11]. Обнаруженный способ позволил распространить область приложений техники слайдинга на интересующий нас класс задач. В разделе 2 результаты обобщаются на различные негладкие постановки задач, в частности на обобщенные линейные модели [15] и другие модели, допускающие эффективное сглаживание [2, 13].

### 1. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Рассмотрим следующую задачу:

$$F(x) = f(x) + g(x) = f(x) + \frac{1}{m} \sum_{k=1}^m g_k(x) \rightarrow \min_x \quad (1)$$

где  $f$  и  $g_k$  имеют  $L_f$  и  $L_g$  – липшицевы градиенты в 2-норме, а функция  $F$  –  $\mu$ -сильно выпуклая в 2-норме, причем  $\mu \ll L_f$ . В задаче (1) введем дополнительное условие  $m \leq L_g/\mu$ . Результат Дж. Лана [9, section 8.2] заключается в том, что для решения рассмотренной задачи с заданной точностью достаточно  $\tilde{O}(\sqrt{L_f/\mu})$  вычислений  $\nabla f$  и  $\tilde{O}(\sqrt{L_g/\mu})$  вычислений  $\nabla g$ , т.е.  $\tilde{O}(m\sqrt{L_g/\mu})$  вычислений  $\nabla g_k$ . При этом не важно с какой именно точностью  $\varepsilon$ . Эта точность будет входить под логарифмами в

<sup>1</sup> Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, Россия

<sup>2</sup> Московский физико-технический институт (национальный исследовательский университет), Долгопрудный, Московская обл., Россия

<sup>3</sup> Национальный исследовательский университет “Высшая школа экономики”, Москва, Россия

\*E-mail: sergery.omelchenko@phystech.edu

\*\*E-mail: gasnikov@yandex.ru

приведенные далее оценки, а для наглядности логарифмические сомножители было решено опустить. Далее оговорки о точности решения возникающих подзадач также опускаются, поскольку все это влияет только на логарифмические сомножители в итоговых оценках, которые опущены. Здесь и далее  $\tilde{O}(\cdot) = O(\cdot)$  с точностью до логарифмического множителя.

Наложим еще одно дополнительное условие  $mL_f \leq L_g$ . Применим к рассмотренной задаче технику каталист [1, 11]. Отметим также, что если использовать технику каталист в варианте [1, 8], то применение данной техники не привносит дополнительного логарифмического множителя. Тогда вместо исходной задачи (1) потребуется  $\tilde{O}(\sqrt{L/\mu})$  раз решать задачу вида

$$f(x) + g(x) + \frac{L}{2} \|x - x^k\|_2^2 \rightarrow \min_x, \quad (2)$$

где  $L$  по построению должно удовлетворять неравенству  $\mu \leq L \leq L_f$ . Задачу (2) можно решать неускоренным композитным градиентным методом [1, 12], считая  $g(x) + \frac{L}{2} \|x - x^k\|_2^2$  композитом. Число итераций такого метода будет совпадать с числом вычислений  $\nabla f$  и равно  $\tilde{O}(L_f/(L + \mu))$ . Но в условиях задачи не предполагалась проксимальная дружелюбность функции  $g$ , поэтому возникающую на каждой итерации неускоренного композитного градиентного метода задачу вида (детали см. в препринте [7])

$$\begin{aligned} \langle \nabla f(\tilde{x}^l), x - \tilde{x}^l \rangle + \frac{L_f}{2} \|x - \tilde{x}^l\|_2^2 + \\ + g(x) + \frac{L}{2} \|x - x^k\|_2^2 \rightarrow \min_x, \end{aligned} \quad (3)$$

в свою очередь, необходимо будет решать. Для решения задачи (3) можно использовать ускоренный композитный метод редукции дисперсии [1, 9, 10], считая  $\frac{L_f}{2} \|x - \tilde{x}^l\|_2^2 + \frac{L}{2} \|x - x^k\|_2^2$  композитом. Число вычислений  $\nabla g_k$  для такого метода будет  $\tilde{O}(\sqrt{mL_g(L_f + L)})$ . Точнее говоря, оценка имеет вид:  $\tilde{O}(m + \sqrt{mL_g/(L_f + L)})$ . Однако в виду предположений  $mL_f \leq L_g$ ,  $L \leq L_f$ :

$$\tilde{O}(m + \sqrt{mL_g/(L_f + L)}) = \tilde{O}(\sqrt{mL_g/(L_f + L)}).$$

Таким образом, общее число вычислений  $\nabla g_k$  будет

$$\begin{aligned} \tilde{O}(m\sqrt{L/\mu}) + \tilde{O}(\sqrt{L/\mu}) \cdot \tilde{O}(L_f/(L + \mu)) \times \\ \times \tilde{O}(\sqrt{mL_g/(L_f + L)}). \end{aligned} \quad (4)$$

Первое слагаемое появилось из-за того, что в каталисте требуется считать  $\nabla F$  на каждой итерации.

Выбирая параметр  $L$  ( $\mu \leq L \leq L_f$ ) так, чтобы выражение (4) было минимальным, получим (с учетом сделанных предположений  $mL_f \leq L_g$  и  $\mu \ll L_f$ ), что  $L \approx L_f$ . Следовательно, имеет место

**Теорема 1.** При  $mL_f \leq L_g$  задачу (1) можно решить с помощью описанной выше техники за  $\tilde{O}(\sqrt{L_f/\mu})$  вычислений  $\nabla f$  и  $\tilde{O}(\sqrt{mL_g/\mu})$  вычислений  $\nabla g_k$ .

Последняя оценка в  $\tilde{O}(\sqrt{m})$  раз лучше оценки, которую можно получить, используя исходный ускоренный градиентный слайдинг Дж. Лана [9]. Несложно заметить [9], что приведенные в теореме 1 оценки оптимальны с точностью до логарифмических множителей.

Заметим, что в описанном выше подходе с  $g(x)$  общего вида ускоренный метод редукции дисперсии можно заменить на покоординатный спуск или безградиентный метод [5]. Таким образом, можно получить расщепление задачи не только по гладкости или структуре слагаемых, но и по структуре оракула, доступного для каждого из слагаемых. Другой пример такого расщепления см. в [3].

Заметим также, что если в описанном выше подходе ограничиться вариантом каталиста из [1, 11], то все рассуждения можно провести в модельной (для  $f$ ) общности [1].

## 2. ПРИЛОЖЕНИЕ

Заметим, что аналогично случаям задач из [1, 14, 15] описанная выше техника может использоваться и тогда, когда  $g_k$  — не гладкие функции, но допускающие сглаживание [2, 13]. Скажем, двойственное сглаживание по Ю.Е. Нестерову [1, 2, 13]. А именно, предположим, что функции  $g_k$  имеют проксимально-дружелюбные сопряженные функции  $g_k^*$ . В частности, это имеет место для обобщенной линейной модели [15], в которой  $g_k(x) := g_k(\langle a_k, x \rangle)$ . Тогда, регуляризируя сопряженные функции  $g_k^*$  с коэффициентом регуляризации  $\sim \varepsilon$ , где  $\varepsilon$  — желаемая точность (по функции) решения исходной задачи, получим, что  $\varepsilon/2$ -решение сглаженной задачи будет  $\varepsilon$ -решением исходной. При том, что для сглаженной задачи  $L_g \sim \varepsilon^{-1}$ .

Заметим, что с помощью регуляризации исходной задачи [1] описанные выше результаты распространяются с сильно выпуклого случая на просто выпуклый случай. Для этого в постановку выпуклой задачи (1) вносится регуляризация  $+\mu/2 \|x\|_2^2$ , где  $\mu = \varepsilon/R^2$ . Здесь  $\varepsilon$  — желаемая точность реше-

**Таблица 1.** Сравнение алгоритмов

Алгоритм	Сложность	Ссылка
FGM	$O\left(\frac{R}{\varepsilon} \sqrt{s(ms + n^2)}\right)$	[1]
Слайдинг	$\tilde{O}\left(\frac{R}{\varepsilon} \sqrt{ms} \cdot s\right) + \tilde{O}\left(\sqrt{\frac{\lambda_{\max}(C)R^2}{\varepsilon}} \cdot n^2\right)$	данная статья

ния задачи по функции, а  $R = \|x_*\|_2$  – 2-норма решения (на практике можно брать оценку сверху [1]). Из [1] следует, что  $\varepsilon/2$ -решение так регуляризованной задачи будет  $\varepsilon$ -решением исходной задачи (1). Продемонстрируем возможные преимущества предложенного подхода в выпуклом (но не сильно выпуклом случае).

Рассматривается постановка задачи

$$F(x) = \frac{1}{2} \langle x, Cx \rangle + \frac{1}{m} \sum_{k=1}^m g_k(\langle a_k, x \rangle) \rightarrow \min_{x \in \mathbb{R}^n}$$

Предполагаем, что  $|g_k''(y)| = O(1/\varepsilon)$ , матрица  $A = [a_1, \dots, a_m]^T$  имеет  $ms$  ненулевых элементов,  $\max_{k=1, \dots, m} \|a_k\|_2^2 = O(s)$ , где  $1 \ll s \leq n$  и  $C$  – неотрицательно определенная матрица с  $\lambda_{\max}(C) \leq 1/(\varepsilon m)$ . Ускоренный градиентный метод (FGM) [1] будет требовать

$$O\left(\sqrt{\frac{(s/\varepsilon + \lambda_{\max}(C))R^2}{\varepsilon}}\right)$$

итераций для достижения точности  $\varepsilon$  по функции со сложностью одной итерации

$$O(ms + n^2)$$

арифметических операций (а.о.). В настоящей работе предложен подход, который требует

$$\tilde{O}\left(\sqrt{\frac{\lambda_{\max}(C)R^2}{\varepsilon}}\right)$$

итераций ускоренного градиентного метода для квадратичной формы (первого слагаемого). При этом сложность одной такой итерации

$$O(n^2) \text{ а.о.}$$

Также предложенный подход требует

$$\tilde{O}\left(\sqrt{\frac{(ms/\varepsilon)R^2}{\varepsilon}}\right)$$

итераций ускоренного метода редукции дисперсии [1, 9, 10]. При этом сложность одной такой итерации

$$O(s) \text{ а.о.}$$

Для наглядности эти результаты собраны в табл. 1. Из табл. 1 можно сделать вывод, что при  $s \gg 1$ ,  $\lambda_{\max}(C) \leq 1/(\varepsilon m) \ll s/\varepsilon$ , предложенный в данной работе подход имеет лучшую теоретическую сложность, чем ускоренный градиентный метод, который принято было считать наилучшим для данного класса задач.

#### ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Работа поддержана грантами РФФИ 18–31–20005 мол\_а\_вед (раздел 1) и РФФИ 19–31–90062 Аспиранты (раздел 2).

Работа первого автора (Д.М. Двинских) была выполнена при поддержке Министерства науки и высшего образования Российской Федерации (госзадание № 075-00337-20-03).

#### СПИСОК ЛИТЕРАТУРЫ

1. *Гасников А.В.* Современные численные методы оптимизации. Метод универсального градиентного спуска. М.: МФТИ, 2018.
2. *Allen-Zhu Z., Hazan E.* Optimal Black-Box Reductions between Optimization Objectives // *Advances in Neural Information Processing Systems*. 2016. P. 1614–1622.
3. *Beznosikov A., Gorbunov E., Gasnikov A.* Derivative-Free Method for Decentralized Distributed Non-Smooth Optimization // *IFAC 2020. The 21st World Congress of the International Federation of Automatic Control*.
4. *Bottou L., Curtis F.E., Nocedal J.* Optimization Methods for Large-Scale Machine Learning // *Siam Review*. 2018. V. 60. № 2. P. 223–311.
5. *Dvurechensky P., Gasnikov A., Tiurin A.* Randomized Similar Triangles Method: A Unifying Framework for Accelerated Randomized Optimization Methods (Coordinate Descent, Directional Search, Derivative-Free Method) // *arXiv:1707.08486*
6. *Hazan E.* Lecture Notes: Optimization for Machine Learning // *arXiv:1909.03550*
7. *Ivanova A., Gasnikov A., Dvurechensky P., Dvinskikh D., Tyurin A., Vorontsova E., Pasechnyuk D.* Oracle Complexity Separation in Convex Optimization // *arXiv:2002.02706*
8. *Ivanova A., Grishchenko D., Gasnikov A., Shulgin E.* Adaptive Catalyst for Smooth Convex Optimization // *arXiv:1911.11271*

9. *Lan G.* Lectures on Optimization. Methods for Machine Learning // <https://wpw.gatech.edu/guanghui-lan/publications/>
10. *Lan G., Li Z., Zhou Y.* A Unified Variance-Reduced Accelerated Gradient Method for Convex Optimization // *Advances in Neural Information Processing Systems*. 2019. P. 10462–10472.
11. *Lin H., Mairal J., Harchaoui Z.* Catalyst Acceleration for First-Order Convex Optimization: from Theory to Practice // *J. Machine Learning Research*. 2017. V. 18. № 1. P. 7854–7907.
12. *Nesterov Yu.* Gradient Methods for Minimizing Composite Functions // *Math. Prog.* 2013. V. 140. № 1. P. 125–161.
13. *Nesterov Yu.* Smooth Minimization of Non-Smooth Function // *Math. Program.* 2005. V. 103. № 1. P. 127–152.
14. *Shalev-Shwartz S., Ben-David S.* Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press, 2014.
15. *Shalev-Shwartz S., Shamir O., Srebro N., Sridharan K.* Stochastic Convex Optimization // *COLT*. 2009.

## ACCELERATED GRADIENT SLIDING FOR MINIMIZING THE SUM OF FUNCTIONS

**D. M. Dvinskikh<sup>a</sup>, S. S. Omelchenko<sup>b</sup>, A. V. Gasnikov<sup>b</sup>, and A. I. Tyurin<sup>c</sup>**

<sup>a</sup> *Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russian Federation*

<sup>b</sup> *Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russian Federation*

<sup>c</sup> *Higher School of Economics, Moscow, Russian Federation*

Presented by Academician of the RAS Yu.G. Evtushenko

In this article, we propose a new way to justify the accelerated gradient sliding of G. Lan, which allows one to extend the sliding technique to a combination of an accelerated gradient method with an accelerated variance reduced method. We obtain new optimal estimates for solving the problem of minimizing a sum of smooth strongly convex functions with a smooth regularizer.

*Keywords:* convex optimization, fast gradient method, dispersion reduction method, composite optimization, sliding