Ускоренные методы нулевого порядка в гладкой выпуклой стохастической оптимизации

Хафизов Фанис
khafizov.fa@phystech.edu

Богданов Александр
bogdanov.ai@phystech.edu

Безносиков Александр
beznosikov.an@phystech.edu

Аннотация

Данная работа посвящена задачи оптимизации без доступа к градиенту целевой функции $f(x)$, из-за чего нам надо как-то его оценивать. Рассматривается безградиентный метод JAGUAR, использующий информацию о предыдущих вызовах и требует $\mathcal{O}(1)$ оракульных вызовов. Мы применяем эту аппроксимацию в ускоренном методе градиентного спуска и докажем его сходимость для выпуклой задачи оптимизации. Также сравним метод JAGUAR с другими известными способами на экспериментах.

Ключевые слова методы нулевого порядка, стохастическая оптимизация

1 Введение

1.1 Мотивация

Стохастические градиентные методы являются необходимыми в решении различных оптимизационных задач. Однако в нынешних проблемах машинного обучения возникает потребность оценивать градиент, ввиду, например, дороговизны его подсчета, либо же незнания явного вида минимизируемой функции. Тогда на помощь приходят методы нулевого порядка. Так как имеется доступ только к значениям целевой функции f в различных точках $x \in \mathbb{R}^d$, то необходимо строить методы, аппроксимирующие градиент, используя конечные суммы значений целевой функции.
Одним из возможных усложнений задачи может быть добавлением шума к оракулу: вместо $f(x)$ он будет возвращать $f(x) + \delta(x)$. Более того, шум может быть стохастическим [6] или детерминированным [5].
Другую важную роль в решении оптимизационных задач играют ускоренные методы. Они как правило имеют более быструю сходимость по сравнению со стандартными алгоритмами. Предложенный Нестеровым [7] быстрый градиентный метод является классическим примером.

1.2 Смежные работы

Впервые метод JAGUAR аппроксимации градиента был предложен в [1], доказана сходимость для методов Франка-Вульфа и градиентного спуска для невыпуклой, выпуклой, и сильно-выпуклой задач. Также есть модификация алгоритма Франка-Вульфа для стохастического случая (рассмотрены случаи one-point feedback и two-point feedback).
В работе [2] рассмотрены ускоренные методы первого порядка в невыпуклых и сильно выпуклых задачах оптимизации, содержащих марковский шум. Там при аппроксимации градиента использовался рандомизированный размер батча. Результаты и подходы оттуда адаптированы под метод JAGUAR в данной работе.
State-of-the-art решения в области безградиентной оптимизации собраны в [3]. Представлены рандомизированные методы нулевого порядка, однако лишь для нестохастического случая.
Ускоренный безградиентный метод в гладкой выпуклой стохастической оптимизации был предложен в [4], но лишь в случае two-point feedback, который далеко не всегда реализуется на практике.

1.3 Предложения

Предлагается ускоренный метод, в качестве аппроксимации градиента использующий JAGUAR. Для него получены теоремы о скорости сходимости при разных ограничениях на целевую функцию (выпуклая, сильно-выпуклая, невыпуклая). Теоретическая часть подкреплена численным экспериментом: классификация на датасете небольшого размера и минимизация квадратичной функции. Проведено сравнение с методами l_2-smoothing и полной аппроксимацией градиента.

2 Постановка задачи

2.1 Детерминированный случай

В этом разделе рассматривается детерминированная задача оптимизации:

$$\min_{x \in Q} f(x), \tag{1}$$

где $Q \subset \mathbb{R}^d$ – выпуклое и компактное.
Мы полагаем, что есть доступ только к оракулу нулевого порядка, то есть, мы можем получать только значения $f(x)$, но не градиента $\nabla f(x)$. Следовательно, нужно как-то аппроксимировать градиент $\nabla f(x)$. Также предполагается, что оракул возвращает

$$f_\delta(x) = f(x) + \delta(x). \tag{2}$$

Для аппроксимации градиента используется следующая разностная схема:

$$\widetilde{\nabla}_i f_\delta(x) := \frac{f_\delta(x + \tau e_i) - f_\delta(x - \tau e_i)}{2\tau} e_i. \tag{3}$$

2.2 Стохастический случай

В этом же разделе рассматривается более общая стохастическая задача:

$$\min_{x \in Q} f(x) := \mathsf{E}_{\xi \sim \pi}[f(x, \xi)], \tag{4}$$

где ξ – случайный вектор из неизвестного распределения π. Здесь мы так же считаем, что у нас нет доступа к градиенту $\nabla f(x, \xi)$, а оракул нулевого порядка возвращает зашумленное значение функции $f_\delta(x, \xi) := f(x, \xi) + \delta(x, \xi)$.
Разделяются два вида аппроксимации градиента в стохастическом случае:

- Two-point feedback (TPF)

$$\widetilde{\nabla}_i f_\delta(x, \xi) := \frac{f_\delta(x + \tau e_i, \xi) - f_\delta(x - \tau e_i, \xi)}{2\tau} e_i, \tag{5}$$

- One-point feedback (OPF)

$$\widetilde{\nabla}_i f_\delta(x, \xi^+, \xi^-) := \frac{f_\delta(x + \tau e_i, \xi^+) - f_\delta(x - \tau e_i, \xi^-)}{2\tau} e_i. \tag{6}$$

В случае TPF в обоих точках, где мы вызываем оракула, одно и то же значение ξ, что может быть тяжело реализуемо. Также TPF является частным случаем OPF ($\xi^+ = \xi^- = \xi$).

Список литературы

[1] Alexander Beznosikov Alexander Bogdanov, Andrey Veprikov. New aspects of black box conditional gradient: Variance reduction and one point feedback, 2024.

[2] Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First order methods with markovian noise: from acceleration to variational inequalities, 2023.

[3] Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky, Eduard Gorbunov, Aleksander Beznosikov, and Aleksandr Lobanov. Randomized gradient-free methods in convex optimization, 2024.

[4] Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization, 2020.

[5] Aleksandr Lobanov, Anton Anikin, Alexander Gasnikov, Alexander Gornov, and Sergey Chukanov. Zero-order stochastic conditional gradient sliding method for non-smooth convex optimization, 2023.

[6] Aurelien Lucchi, Frank Proske, Antonio Orvieto, Francis Bach, and Hans Kersting. On the theoretical properties of noise correlation in stochastic optimization, 2022.

[7] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. Proceedings of the USSR Academy of Sciences, 269:543–547, 1983.