Безградиентные методы решения негладких выпуклых задач стохастической оптимизации с тяжелыми хвостами на выпуклом компакте

Корнилов Никита
Научный руководитель: д.ф.-м.н. А.В. Гасников

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра "Интеллектуальные системы"

28 июня, 2023

Постановка задачи и цели

Цель

Построить эффективные алгоритмы для решения негладких задач с тяжелыми хвостами и получить оценки на скорость их сходимости.

Рассмотрим негладкую выпуклую задачу стохастической оптимизации на выпуклом компакте $\mathcal{X} \subset \mathbb{R}^d$ с функцией $f(x) : \mathbb{R}^d \to \mathbb{R}$

$$\min_{x \in \mathcal{X}} f(x) \triangleq \mathbb{E}_\xi[f(x, \xi)],$$

где значения функции доступны только через зашумленный оракул нулевого порядка, т.е.

$$\phi(x, \xi) = f(x, \xi) + \delta(x).$$

Мы рассматриваем двухточечный оракул нулевого порядка. Это значит, что на запрос из двух точек $x, y \in \mathcal{S}$ мы получаем $\phi(x, \xi)$ и $\phi(y, \xi)$ с одним и тем же ξ.

Предположения

Проблема

Оракул содержит сильный двухкомпонентный шум. Случайный шум с тяжелыми хвостами, а также детерминированный враждебный шум.

1. **Выпуклость**
   Функция $f(x,\xi)$ выпукла по x на $\mathcal{X}$ для любого ξ.
2. **Ограниченность $1+\kappa$ момента**
   Функция $f(x,\xi)$ является $M_2(\xi)$-Липшиц непрерывной по x в l_2-норме, т.е. для любых $x_1, x_2 \in \mathcal{X}$
   $$|f(x_1,\xi) - f(x_2,\xi)| \leq M_2(\xi)\|x_1 - x_2\|_2.$$
   К тому же, пусть существуют $\kappa \in (0,1]$ и M_2 такие, что $\mathbb{E}_\xi[M_2^{1+\kappa}(\xi)] \leq M_2^{1+\kappa}$.
   Например, в гладком случае $||\nabla f(x,\xi)||^{1+\kappa} \leq M_2^{1+\kappa}(\xi)$.
3. **Ограниченность враждебного шума**
   Для всех $x \in \mathcal{X} : |\delta(x)| \leq \Delta < \infty$

Решение

Решение

Предлагается использовать разностную аппроксимацию градиента по двум точкам с евклидовой сферы и клиппировать его. Этот вектор используется как вектор обновления в алгоритмах первого порядка.

Литература

Предлагается безградиентный метод для решения задач, где у случайного шума ограничен лишь второй момент

- Dvinskikh D. et al. Gradient-Free Optimization for Non-Smooth Minimax Problems with Maximum Value of Adversarial Noise – 2022

Расписана идея клиппирования и доказательства верхней оценки

- Zhang J., Cutkosky A. Parameter-free Regret in High Probability with Heavy Tails – 2022

Рассмотрен другой подход к борьбе с тяжелыми хвостами путём модификации зеркального градиентного спуска

- Vural N. M. et al. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance - 2022

Гладкая аппроксимация

Для того чтобы сделать гладкую аппроксимацию негладкой оптимизируемой функции мы семплируем вектор **e** из равномерного распределения на евклидовой сфере $S^2 := \{\mathbf{e} : ||\mathbf{e}||_2 = 1\}$.
Определим гладкую аппроксимацию как

$$f^\tau(x) = \mathbb{E}_{\mathbf{e}\sim S^2}[f(x + \tau\mathbf{e})].$$

Её градиент вычисляется по формуле

$$\nabla f^\tau(x) = \mathbb{E}_{\mathbf{e}\sim S^2}\left[\frac{d}{\tau} f(x + \tau\mathbf{e})\mathbf{e}\right].$$

Конечная разность для приближения градиента по двум точкам

$$g(x, \xi, \mathbf{e}) = \frac{d}{2\tau}(\phi(x + \tau\mathbf{e}, \xi) - \phi(x - \tau\mathbf{e}, \xi))\mathbf{e}$$

для константы сглаживания $\tau > 0$.

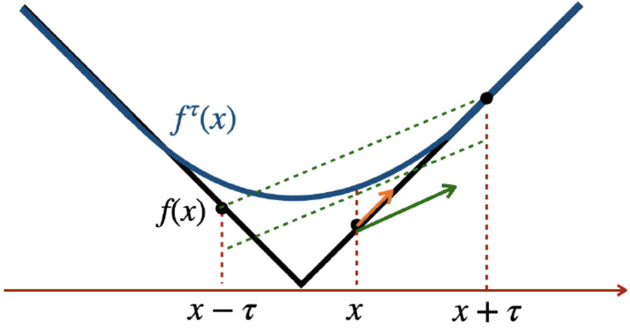

Рис.: Функция модуля и её сглаженный аппроксимация

Качество аппроксимации

В работе мы доказываем ряд свойств гладкого приближения и её градиента, самые важные из них представлены ниже.

Theorem

При предположениях верны следующие неравенства

- *Отличие от целевой функции*

$$\sup_{x \in \mathcal{X}} |f^\tau(x) - f(x)| \leq \tau M_2.$$

- $(1+\kappa)$*-ый момент оценки градиента ограничен*

$$\mathbb{E}_{\xi, \mathbf{e}}[\|g(x, \xi, \mathbf{e})\|_q^{1+\kappa}] \leq 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa} = \sigma_q^{1+\kappa},$$

где $a_q = d^{\frac{1}{q} - \frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q - 1}\}$.

Стохастический Зеркальный Спуск

Пусть прокс-функция $\Psi : \mathbb{R}^d \to \mathbb{R}$ является 1 сильно выпуклой по ℓ_p-норме и непрерывно дифференцируемой. Обозначим её сопряжённую функцию и дивергенцию Брегмана как

$$\Psi^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - \Psi(x)\},$$

$$D_\Psi(y, x) = \Psi(y) - \Psi(x) - \langle \nabla \Psi(x), y - x \rangle.$$

Шаг Алгоритма с размером шагом ν и вектором обновления g_{k+1} определяется по формулам

$$y_{k+1} = \nabla(\Psi^*)(\nabla \Psi(x_k) - \nu g_{k+1}), \quad x_{k+1} = \arg\min_{x \in \mathcal{X}} D_\Psi(x, y_{k+1}).$$

Для краткости будем записывать его через функцию

$$x_{k+1} = SMD\_Step(x_k, \nu g_{k+1}).$$

Клиппирование

Установим константу клиппирования $c > 0$. Оператор клиппирования, применённый к вектору g, задаётся как

$$\hat{g} := \frac{g}{\|g\|} \min(\|g\|, c) := Clip(g, c).$$

Стоит отметить, что $\hat{g}$ даёт смещенную оценку изначального вектора g

$$\|\mathbb{E}[g] - \mathbb{E}[\hat{g}]\|_q \leq \frac{\sigma_q^{1+\kappa}}{c^\kappa}.$$

Чем меньше c, тем больше и смещение. Но при этом второй момент меньше

$$\mathbb{E}[\|\hat{g}\|_q^2] \leq \sigma_q^{1+\kappa} c^{1-\kappa}.$$

Таким образом, выбор константы c позволяет балансировать между быстрой сходимостью и большим отклонением.

Алгоритм с Клиппингом

procedure $\textsc{Zero CLIP}$(Количество итераций T, размер шага ν, константа клиппирования c, прокс-функция Ψ_p, константа сглаживания τ)

  $x_0 \leftarrow \arg\min\limits_{x \in \mathcal{X}} \Psi_p(x)$

  for $k = 0, 1, \ldots, T-1$ **do**

    Независимо сэмплируем $\mathbf{e}_k \sim \mathrm{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$

    Независимо сэмплируем ξ_k

    Оцениваем $g_{k+1} = \frac{d}{2\tau}(\phi(x_k + \tau\mathbf{e}_k, \xi_k) - \phi(x_k - \tau\mathbf{e}_k, \xi_k))\mathbf{e}_k$

    Вычисляем клиппированный вектор $\hat{g}_{k+1} = Clip(g_{k+1}, c)$

    Вычисляем шаг $x_{k+1} = SMD\_Step(x_k, \nu\hat{g}_{k+1})$

  end for

  return $\overline{x}_T \leftarrow \frac{1}{T}\sum\limits_{k=0}^{T-1} x_k$

end procedure

Сходимость Алгоритма с Клиппингом

Theorem

Пусть $q \in [2, \infty]$, *количество итераций* T, *константа сглаживания* $\tau > 0$, *прокс-функция* $\Psi_p(x)$ *заранее выбраны. Вычислим константу клиппирования* $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ *и размер шага* $\nu = \frac{\mathcal{D}_\Psi}{c}$, *где* $\mathcal{D}_\Psi^2 = 2 \sup\limits_{x,y \in \mathcal{X}} D_{\Psi_p}(x, y)$ — *диаметр компакта* $\mathcal{X}$.

Пусть точка $\overline{x}_T$ *получена Алгоритмом с Клиппингом с параметрами выше, тогда*

1. $$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \quad (1)$$

2. *С вероятностью больше или равной* $1 - \delta$

$$f(\overline{x}_T) - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi\sigma_q\log\frac{1}{\delta}}{T^{\frac{\kappa}{1+\kappa}}}. \quad (2)$$

Максимально допустимый уровень враждебного шума

Пусть ε желаемая точность по функции, т.е. с вероятностью больше или равной $1-\delta$: $f(\overline{x}_T) - f(x^*) \leq \varepsilon$.

- Если враждебного шума совсем нет $\Delta = 0$:
  Количество итераций $T = \tilde{O}\left(\left(\frac{\mathcal{D}_\Psi \sqrt{d} a_q M_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}\right)$ при $\tau \to 0$.
- Когда $\tau = \frac{\varepsilon}{M_2}$ и $\Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$:
  Количество итераций такое же.
- В противном случае, когда $\Delta > \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$:
  Количество итераций увеличивается в два раза, и Алгоритм не может достигнуть точности меньше, чем $\sqrt{M_2\sqrt{d}\Delta\mathcal{D}_\Psi}$.
- Константы $a_q, \mathcal{D}_\Psi$ во многом зависят от множества $\mathcal{X}$. Рекомендации по выбору q, Ψ также прилагаются.

Зависимость от d

В гладком случае для того, чтобы оценить градиент функции, достаточно использовать $d + 1$ значений функции. Для стохастических методов первого порядка, оптимальное количество вызовов оракула первого порядка пропорционально $\varepsilon^{-\frac{1+\kappa}{\kappa}}$, поэтому для методов нулевого порядка можно ожидать $d\varepsilon^{-\frac{1+\kappa}{\kappa}}$. В этой работе мы получаем оценку $\left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$, которая совпадает только при $\kappa = 1$.

Оптимальна ли эта оценка?

Для гладких задач стохастической выпуклой оптимизации с $(d + 1)$-точечным оракулом нулевого порядка ответ отрицательный, и оптимальная оценка количества значений функции равна $\sim d\varepsilon^{-\frac{1+\kappa}{\kappa}}$.

Результаты

Публикации:

1. Gradient-Free Methods for Non-Smooth Convex Stochastic Optimization with Heavy-Tailed Noise on Convex Compact// under review at Computational Management Science// arXiv preprint arXiv:2304.02442. – 2023

Выступления с докладом:

1. The 13th International Conference on Network Analysis

Выносится на защиту

- Построен основанный на клиппинге безградиентный метод решения стохастических выпуклых задач с тяжелыми хвостами.
- Также построен второй алгоритм при помощи устойчивого Стохастического Зеркального Спуска.
- Для функций с острым минимумом предложен ускоренный с помощью техники рестартов алгоритм.
- Для всех трёх алгоритмов выше произведён анализ скорости сходимости, максимального уровня враждебного шума, зависимостей от параметров.