Исследования вариаций субградиентного метода на классе слабо выпуклых задач

Есения Андреевна Лушко

Физтех-школа прикладной математики и информатики, МФТИ
научный руководитель: Федор Сергеевич Стонякин

23 мая, 2023

Актуальность и смежные задачи

1. Хорошо изучено, что субградиентные методы в сравнении с ускоренными субградиентными методами на классе выпуклых гладких функций являются более устойчивыми к погрешностям.[1]
2. Современные исследования показывают, что для невыпуклых задач, есть нижняя оценка на область сходимости.[2]

Цель исследования – сделать подобные результаты на классе слабо выпуклых задач с острым минимумом, т.к. они возникают в различных приложениях. [3]

[1] Devolder, O., Glineur, F. Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. // Math. Program. 146, 37–75 (2014).

[2] Fedyor Stonyakin, Ilya Kuruzov, Boris Polyak: Stopping rules for gradient methods for non-convex problems with additive noise in gradient

[3] Damek Davis and Dmitriy Drusvyatskiy and Courtney Paquette: The nonsmooth landscape of phase retrieval // IMA Journal of Numerical Analysis – 2020, Vol. 40, P. 2652–2695

Известные исследования рассматриваемого класса задач

Теорема 1. Оценка сходимости субградиентного метода для слабо выпуклых функций с острым минимумом [4]

Пусть f μ-слабо выпукла и имеет α-острый минимум $(\alpha, \mu > 0)$, а точка $x_0 : \min_{x_* \in X_*} \|x_0 - x_*\|_2 \leqslant \frac{\alpha\gamma}{\mu}$ для некоторого фиксированного $\gamma \in (0; 1)$. Тогда для метода

$$x_{k+1} := \mathrm{Proj}_Q \left(x_k - h_k \nabla f\left(x_k\right)\right), \quad k = 0, 1, 2, \ldots,$$

$$\text{где } h_k := \frac{f\left(x_k\right) - f_*}{\|\nabla f\left(x_k\right)\|_2^2} \left(\text{ при } \nabla f\left(x_k\right) \neq 0\right),$$

верно неравенство

$\min_{x_* \in X_*} \|x_{k+1} - x_*\|_2^2 \leqslant \prod_{i=0}^{k} \left(1 - \frac{\alpha^2(1-\gamma)}{\|\nabla f(x_i)\|_2^2}\right) \min_{x_* \in X_*} \|x_0 - x_*\|_2^2,$

[4] Davis, D., Drusvyatskiy, D., MacPhee, K.J. et al. Subgradient Methods for Sharp Weakly Convex Functions. J Optim Theory Appl 179, 962–982 (2018).

Постановка задачи

Выделяются две вариации метода:

1. На каждом шаге метода доступен субградиент целевой функции с аддитивной неточностью:

$$\left\|\widetilde{\nabla} f\left(x_k\right)-\nabla f\left(x_k\right)\right\|_2 \leq \Delta,$$

где Δ положительно, но достаточно мало.

———————————————————————————————————

2. На каждом шаге погрешность присутствует как для значения целевой функции, так и для значении субградиента:

$$\left\|\widetilde{\nabla} f\left(x_k\right)-\nabla f\left(x_k\right)\right\|_2 \leq \Delta, \quad\left|\widetilde{f(x_k)}-f\left(x_k\right)\right| \leq \delta,$$

где $\delta, \Delta>0$, но достаточно малые.

Анализ метода с неточной информацией о субградиенте

Пусть функция f μ-слабо выпукла, M-липшицева и имеет α-острый минимум ($\alpha, \mu > 0$). Пусть начальная точка удовлетворяет условию:

$$\|x_0 - x_*\|_2 \leq \frac{\alpha}{2\mu\sqrt{2}},$$

Тогда субградиентный метод с шагом Б.Т.Поляка:

$$x_{k+1} := \mathrm{Proj}_Q\left(x_k - h_k \widetilde{\nabla} f\left(x_k\right)\right), \quad k = 0, 1, 2, \ldots,$$

при наличии неточной информации значения субградента на каждом шаге $\left(\left\|\widetilde{\nabla} f\left(x_k\right) - \nabla f\left(x_k\right)\right\|_2 \leq \Delta\right)$ сходится и в худшем случае можно гарантировать

$$\|x_k - x_*\|_2^2 = O(\Delta^{2-\frac{2}{\lambda}})$$

для достаточно большого k.

Клиппинг для случая с неточным оракулом в субградиенте [5]

(a) Если $\left\|\widetilde{\nabla} f\left(x_k\right)\right\|_2 \leq M \sqrt[\lambda]{\Delta}$, то $h_k=\frac{f(x_k)-f_*}{M^2}$ и:

$$x_{k+1}:=\operatorname{Proj}_Q\left(x_k-\frac{f(x_k)-f_*}{M^2} \cdot \widetilde{\nabla} f\left(x_k\right)\right)$$

(b) Если $\left\|\widetilde{\nabla} f\left(x_k\right)\right\|_2 > M \sqrt[\lambda]{\Delta}$, то $h_k=\frac{f(x_k)-f_*}{\left\|\widetilde{\nabla} f(x_k)\right\|_2^2}$ и:

$$x_{k+1}:=\operatorname{Proj}_Q\left(x_k-\frac{f(x_k)-f_*}{\left\|\widetilde{\nabla} f\left(x_k\right)\right\|_2^2} \widetilde{\nabla} f\left(x_k\right)\right)$$

[5] Why gradient clipping accelerates training: a theoretical justification for adaptivity

Случай малой нормы субградиента:

$$\|x_{k+1} - x_*\|_2^2 \leqslant \|x_0 - x_*\|_2^2 \left(1 - \frac{\alpha^2}{M^2} \cdot \frac{\sqrt{2}-1}{\sqrt{2}}\right)^{k+1} + \frac{\Delta^2}{2\mu\alpha\left(1-\sqrt{2}\right)} \times$$
$$\times \left(\left(1 - \frac{\alpha^2}{M^2} \cdot \frac{\sqrt{2}-1}{\sqrt{2}}\right)^{k+1} - 1\right)$$

Итого, после k итераций в данном сценарии гарантируется, что $\|x_{k+1} - x_*\|_2^2 = O(\Delta^2)$

Случай большой нормы субградиента:

$$\|x_{k+1} - x_*\|_2^2 \leqslant \|x_0 - x_*\|_2^2 \prod_{i=0}^{k+1} \left(1 - \frac{\alpha^2}{\left\|\widetilde{\nabla} f(x_k)\right\|_2^2} \cdot \frac{\sqrt{2}-1}{\sqrt{2}}\right) + \frac{\Delta^{2-\frac{2}{\lambda}} M}{2\mu\alpha\left(1-\sqrt{2}\right)} \times$$
$$\times \left(\left(1 - \frac{\alpha^2}{M^2} \cdot \frac{\sqrt{2}-1}{\sqrt{2}}\right)^{k+1} - 1\right)$$

Итого, после k итераций в данном сценарии гарантируется, что $\|x_{k+1} - x_*\|_2^2 = O(\Delta^{2-\frac{2}{\lambda}})$

Постановка задачи для второго сценария

Пусть теперь у нас тот же класс функций, тот же субградиентный метод с шагом Б.Т. Поляка. При наличии в такой постановке погрешности в целевой функции и субградиенте $\left\|\widetilde{\nabla} f\left(x_k\right)-\nabla f\left(x_k\right)\right\|_2 \leq \Delta, \quad\left|\widetilde{f\left(x_k\right)}-f\left(x_k\right)\right| \leq \delta$, где $\delta, \Delta>0$, но достаточно малые, что выполнено $\alpha^2-2 \Delta \alpha \geqslant \frac{\alpha^2}{\sqrt{2}}$, то верно, что в худшем случае можно гарантировать

$$\left\|x_k-x_*\right\|_2^2=O\left(\delta^{1-\frac{1}{\lambda}}\right)$$

для достаточно большого k, если на начальную точку наложить условие $\left\|x_0-x_*\right\|_2 \leq \frac{\alpha}{2 \mu \sqrt{2}}$.

Анализ значения погрешности целевой функции

Выделяются две вариации значения погрешности:

1. Значение целевой функции меньше реального:

$$f(x_k) - f_* \geqslant \widetilde{f(x_k)} - f_*$$

---

2. Значение целевой функции больше реального:

$$f(x_k) - f_* < \widetilde{f(x_k)} - f_*$$

Формально, это просто значит, что

$$\left| \widetilde{f(x_k)} - f(x_k) \right| \leq \delta,$$

раскрывается с разными знаками. Здесь $\delta > 0$, но достаточно мало.

Следствие при $\delta = 0$.

$$\min_{x_* \in X_*} \|x_{k+1} - x_*\|_2^2 \leqslant \prod_{i=0}^{k} \left(1 - \frac{\frac{\alpha^2}{2} - 2\Delta\alpha}{\left\|\widetilde{\nabla} f\left(x_k\right)\right\|_2^2}\right) \|x_0 - x_*\|_2^2$$

Теперь гарантируется сходимость к точному минимуму со скоростью геометрической прогрессии. Этот результат улучшает уже имеющийся результат для данного сценария, хоть и делает дополнительные предположения.

Клиппинг для сценария с погрешностью в субградиенте и целевой функции

Пусть $\left\|\widetilde{\nabla} f\left(x_k\right)\right\|_2 \leqslant M \sqrt[\lambda]{\delta}$. Тогда $h_k = \frac{f(x_k) - f_*}{M^2}$ и невязку можно переписать как:

$$\|x_{k+1} - x_*\|_2^2 \leqslant \|x_k - x_*\|_2^2 \left(1 - \frac{\alpha^2 - 2\Delta\alpha - \frac{\alpha^2}{\sqrt{2}}}{M^2}\right) + \frac{\delta\alpha^2}{M^2\mu\sqrt{2}}$$

Пусть $\left\|\widetilde{\nabla} f\left(x_k\right)\right\|_2 > M \sqrt[\lambda]{\delta}$. Тогда $h_k = \frac{f(x_k) - f_*}{\left\|\widetilde{\nabla} f(x_k)\right\|_2^2}$ и невязку можно переписать как:

$$\|x_{k+1} - x_*\|_2^2 \leqslant \|x_k - x_*\|_2^2 \left(1 - \frac{\alpha^2 - 2\Delta\alpha - \frac{\alpha^2}{\sqrt{2}}}{\left\|\widetilde{\nabla} f\left(x_k\right)\right\|_2^2}\right) + \frac{\delta^{1-\frac{1}{\lambda}}\alpha^2}{M^2\mu\sqrt{2}}$$

Итоги для сценария с погрешностью в субградиенте и целевой функции

Итого, можно гарантировать окрестность сходимости:

$$\|x_k - x_*\|_2^2 = O(\delta^{1-\frac{1}{\lambda}})$$

Спасибо за внимание!