Градиентный слайдинг для Седловых Задач с композитами

Антышев Тихон Глебович, студент Б05-904, МФТИ

Научный руководитель - Гасников Александр Владимирович
Научный консультант - Бородич Екатерина Дмитриевна

Долгопрудный,
2023

References

Dmitry Kovalev and Alexander Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. 2022. URL `arxiv.org/abs/2205.05653`.

Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to distributed optimization under similarity. 2022a. URL `arxiv.org/abs/2205.15136`.

TaeHo Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm, 2021.

Dmitry Kovalev, Alexander Gasnikov, and Peter Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling, 2022b.

Постановка задачи

Актуальность

С развитием методов машинного обучения появляется всё больше узкоспециализированных оптимизационных задач.

Мы исследуем седловые задачи с определёнными свойствами.

Свойства

В задаче присутствует композит $f(x) + F(x, y)$

$\nabla f(x)$ затратно вычислять, как например в Personalized Federated Learning[1]:

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} \frac{\lambda}{2} \left\| \sqrt{W} X \right\|^2 + \sum_{m=1}^{M} f_m(x_m, y_m) - \frac{\lambda}{2} \left\| \sqrt{W} Y \right\|^2,$$

Необходимо предложить

Итеративный алгоритм, подходящий под данную задачу.

[1] *Smith, Virginia and Chiang, Chao-Kai and Sanjabi, Maziar and Talwalkar, Ameet* Federated multi-task learning // 2017

Предварительные утверждения

- Мы рассматриваем следующую седловую задачу:

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x) + F(x, y) \tag{1}$$

Предположения

- **Предположение 1** $f(x) : \mathbb{R}^{d_x} \to \mathbb{R}$ - L_f-гладкая и выпукла на $\mathbb{R}^{d_x}$.
- **Предположение 2** $F(x, y) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ - L_F-гладкая на $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, μ-сильно выпукла на $\mathbb{R}^{d_x}$ для фиксированного y and μ-сильно выпукла на $\mathbb{R}^{d_y}$ для фиксированного x.

Определение 1 (L-глакость). $f(x) : \mathbb{R}^n \to \mathbb{R}$ L-гладкая с $L > 0$, если её градиент L-Липшицев:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \tag{2}$$

Определение 2 (μ-сильная выпуклость). $f(x) : \mathbb{R}^n \to \mathbb{R}$ μ-сильно выпуклая с $\mu > 0$, если $\forall\, x,\, y \in \mathbb{R}^n$ выполняется

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \tag{3}$$

Градиентный Слайдинг

- Сначала используем алгоритм из Kovalev et al. [2022a], который используется для решения

$$\min_x p(x) + q(x)$$

- Заменяем ∇p на оператор $\begin{pmatrix} \nabla f(x) \\ 0 \end{pmatrix}$ и ∇q на $\begin{pmatrix} \nabla_x F(x, y) \\ -\nabla_y F(x, y) \end{pmatrix}$

Проблема: необходимо решать подзадачу численно на каждой итерации:

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} A_\eta^k(x, y)$$

$$A_\theta^k(x, y) := \langle \nabla f(x_g^k), x \rangle + F(x, y) + \frac{1}{2\theta_x}\|x - x^k\|^2 - \frac{1}{2\theta_y}\|y - y^k\|^2$$

Итоговый алгоритм

Algorithm Седловой Слайдинг

Входные данные: Начальные точки $x^0 \in \mathbb{R}^{d_x}$, $y^0 \in \mathbb{R}^{d_y}$
Параметры: $\alpha,\ \theta,\ \eta > 0,\ N \in \{1, 2, ...\}$
for $k = 0, 1, 2, \ldots, N-1$ **do**
    $(x_f^k, y_f^k) \approx \arg\min\limits_{x \in \mathbb{R}^{d_x}} \max\limits_{y \in \mathbb{R}^{d_y}} A_\theta^k(x, y)$

$$A_\theta^k(x, y) := \langle \nabla f(x^k), x \rangle + F(x, y) + \frac{1}{2\theta}\|x - x^k\|^2 - \frac{1}{2\theta}\|y - y^k\|^2 \quad (4)$$

    $x^{k+1} = x^k + \alpha\eta(x_f^k - x^k) - \eta\left(\nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k)\right)$
    $y^{k+1} = y^k + \alpha\eta(y_f^k - y^k) + \eta\nabla_y F(x_f^k, y_f^k)$
end for
Возвращает: x^N, y^N

Внешняя сходимость

Теорема 1(Антышев 2023): Если для Алгоритма 1 в условиях задачи (1) заданы следующие параметры:

$$\theta = \frac{1}{2L_f}, \quad \eta = \min\left[\frac{1}{4\mu}, \frac{1}{4L_f}\right], \quad \alpha = 2\mu \tag{5}$$

вспомогательная задача (4) решается с такой точностью, что выполнятеся:

$$\left\|B_\theta^k(x_f^k, y_f^k)\right\|^2 \leq \frac{L_f^2}{3}\left\|\begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix}\right\|^2 \tag{6}$$

то для любого числа итераций такого, что

$$N \geq 2\max\left[1, \frac{L_f}{\mu}\log\frac{R_0^2}{\varepsilon}\right] \tag{7}$$

выполняется следующая оценка:

$$\left\|\begin{pmatrix} x^N - x^* \\ y^N - x^* \end{pmatrix}\right\|^2 \leq \varepsilon$$

Лемма 3 (Антышев 2023): Условие остановки внутреннего метода (6) можно записать в терминах ε-точного решения, как

$$\left\| \begin{pmatrix} x_f^k - \overline{x}_f^k \\ y_f^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \leq \varepsilon_k \tag{8}$$

где

$$\varepsilon_k \leq \frac{L_f^2}{3(2L_f + L_F)^2} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \tag{9}$$

Таким образом, нам не нужно выводить сходимость по норме градиента для каждого внутреннего метода.

FOAM как внутренний метод

Градиентный слайдинг требует решения следующей задачи:

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} A_\theta^k(x, y)$$

Теорема 2 (Антышев 2023): Алгоритму FOAM из Kovalev and Gasnikov [2022] для решения внутренней седловой задачи (4) требуется

$$\mathcal{O}\left(\frac{L_F + 2L_f}{\sqrt{\mu(\mu + 2L_f)}} \log \frac{\sqrt{3}(2L_f + L_F)}{L_f R_k}\right) \tag{10}$$

вызовов $\nabla F(x, y)$

Сходимость

Теорема 5 (Антышев 2023): Итоговая оракульная сложность алгоритма, при внутреннем методе FOAM при решении задачи (1)

$$\mathcal{O}\left(\frac{L_f(L_F + 2L_f)}{\mu\sqrt{\mu(\mu + 2L_f)}} \log \frac{\sqrt{3}R_0^2(2L_f + L_F)}{L_f} \log \frac{R_0^2}{\varepsilon}\right) \tag{11}$$

Теорема 7 (Антышев 2023): Итоговая оракульная сложность алгоритма 1, при внутреннем методе EAG-V из Yoon and Ryu [2021]

$$\mathcal{O}\left(\frac{L_f}{\mu} \log \frac{R_0^2}{\varepsilon}\right) \times \mathcal{O}(1) = \mathcal{O}\left(\frac{L_f}{\mu} \log \frac{R_0^2}{\varepsilon}\right) \tag{12}$$

Заключение

1. Предложен алгоритм градиентного слайдинга для седловых задач.
2. Доказанные теоремы:
   - Сходимость слайдинга
   - Эквивалентность критерия остановки ε-точному решению.
3. Получены теоремы сходимости для внутренних методов:
   - First Optimal Algorithm for Minimax Optimization (FOAM)
   - Gradient Descent-Ascent with Extrapolation (GDAE)
   - Extra Anchored Gradient with Varying step-size (EAG-V)