Київський національний університет імені Тараса Шевченка
Факультет комп'ютерних наук та кібернетики

Рандомізована паралельна декомпозиція для розв'язання операторних включень

Виконав:
Білінський Павло
3 курс
Кафедра Обчислювальної математики

Науковий керівник:
Семенов Володимир Вікторович

13 травня 2021 р.

Зміст

1 Постановка задачі

Нехай в гільбертовому просторі $\mathcal{H}$ задані багатозначні оператори $A_1, A_2, \ldots, A_m$. Потрібно знайти такі $x \in \mathcal{H}$, що

$$0 \in A_1(x) + A_2(x) + \cdots + A_m(x) \tag{1}$$

Зазвичай ми припускаємо, що оператори A_i - монотонні, більше того, максимально монотонні. В роботі [?] був запропонований алгоритм:

Алгоритм

1) Задаємо $\{\lambda_n\} \subseteq (0, +\infty),\ x_1 \in \mathcal{H};\ n := 1.$
2) Знаходимо елементи

$$x_{i,n} = J_{\lambda_n A_i} x_n,\ i = \overline{1, m}.$$

3) Покладаємо

$$x_{n+1} = \frac{1}{m} x_{1,n} + \frac{1}{m} x_{2,n} + \ldots + \frac{1}{m} x_{m,n}$$

$n := n + 1$, *переходимо на крок 2.*

Також в роботі [1] було встановлено, що послідовність середніх по Чезаро:

$$\frac{\sum_{i=1}^{n} \lambda_i x_i}{\sum_{i=1}^{n} \lambda_i} \tag{2}$$

за певних умов слабко збігається до розв'язку включення 1.

На практиці часто обчислення значень $J_{\lambda_n A_i} x_n$ для всіх $i = \overline{1, m}$ потребує багато обчислювальних ресурсів. Щоб пришвидшити інтераційний процес, на кожному кроці можемо обчислювати лише певну кількість цих значень. Отримаємо

Рандомізований алгоритм

1) Задаємо $\{\lambda_n\} \subseteq (0, +\infty),\ x_1 \in \mathcal{H};\ n := 1.$
2) Випадковим чином вибираємо k *індексів* $\{i_1, i_2, ..., i_k\}$ *з множини* $\{1, 2, \dots, m\}$*,* $k < m$
3) Для кожного з обраних індексів обчислюємо елементи

$$x_{i,n} = J_{\lambda_n A_i} x_n,\ i \in \{i_1, i_2, ..., i_k\}.$$

3) Покладаємо

$$x_{n+1} = \frac{1}{k} x_{1,n} + \frac{1}{k} x_{2,n} + \ldots + \frac{1}{k} x_{i_k,n}$$

$n := n + 1$*, переходимо на крок 2.*
Постає питання про властивості цього алгоритму, а саме:

1. Збіжність ітераційного процесу. Оскільки алгоритм - рандомізований, то доцільніше говорити про збіжність послiовностi випадкових величин (в середньому, майже напевно).

2. Оцінка швидкості збіжності.

Потрібно:

- Встановити істотні припущення про оператори A_i та спосіб отримання вибірки.

- Сформулювати і довести теореми про збіжність, а також інші проміжні результати.

- Розглянути важливі частинні випадки

2 Актуальність проблеми

План:

1. Застосування в оптимізації: опуклі недиференційовні функції, субградієнт, , необхідна умова екстремуму.

2. Приклади задач:

   - мінімізація математичного сподівання
   - Мінімізація суми функцій: $f_1 + f_2 \to \min$, тоді необхідна умова екстремуму має вигляд $0 \in (\partial f_1 + \partial f_2)(x)$. Якщо другий доданок має спеціальний вигляд, це дозволяє застосовувати кращі алгоритми.

3. Де виникає $0 \in (A + B)(x)$

   (а) Умовний естремум: $f \to \min_C$. Необхідна умова: $0 \in \partial f(x) + N_C x$

   (б) Пошук сідлової точки функції Лагранжа: $0 \in \begin{pmatrix} \nabla_1 L(x, y) \\ -\nabla_2 L(x, y) \end{pmatrix} + N_{C \times D}(x, y)$

4. Застосування в машинному навчанні `https://arxiv.org/pdf/1403.5074.pdf`. Розподілені обчислення, федеративне навчання, stochastic gradient descend.

3 Стаття Bianchi: ергодична збіжність

3.1 Постановка задачі

Розглядається сім'я операторів $A(s,x) : E \times \mathcal{H} \to 2^{\mathcal{H}}$. Наявність першої змінною дає змогу представити випадковий вибір одного оператора із сукупності на кожному кроці.

Уявімо сім'ю операторів, що індексуються множиною E: $\{A_s\}_{s \in E}$. У найпростішому випадку $E = 1, 2, \ldots, n$, тобто, операторів - скінчена кількість. Тепер нехай $(\xi_n(.)) \subset E$ - послідовність випадкових величин, де ξ_n - випадково обраний індекс на кроці n. Тоді $A(\xi_n, \cdot)$ - випадково обраний оператор на кроці n.

Бачимо, що якщо ξ_n реалізує вибір серед $\{1, 2, \ldots, m\}$, де всі наслідки рівноможливі, то це частинний випадок нашої задачі, бо тут обирається лише один оператор із сукупності.

Автор розглядає ітераційну процедуру:

$$x_{n+1} = J_{\lambda_n A(\xi_{n+1}, \cdot)}(x_n) \tag{3}$$

та прагне довести збіжність майже напевно послідовності середніх за Чезаро:

$$\frac{\sum_{i=1}^{n} \lambda_i x_i}{\sum_{i=1}^{n} \lambda_i}$$

Збіжність такої послідовності середніх автор називає *ергодичною збіжністю.*

Автор шукає розв'язки включення

$$0 \in \mathbb{E}[A(\xi, \cdot)] \tag{4}$$

де справа стоїть математичне сподівання Аумана. Для нашого випадку воно має вигляд

$$\mathbb{E}[A(\xi, \cdot)] = \frac{1}{m} \sum_{i=1}^{m} A_i(x)$$

Зрозуміло, що тоді включення 4 еквівалентне:

$$0 \in \sum_{i=1}^{m} A_i(x)$$

що є частинним випадком нашої задачі.

3.2 Основні поняття

1. Випадкові величини
2. Математичне сподівання Аумана

3.3 Основні результати

(in progress)

Теорема 1. *Нехай виконуються припущення 1-5. Нехай також всі області визначення D_s співпадають для всіх s поза μ-нехтовною ("μ-negligible") множиною. Нехай (x_n) - послідовність, що дається формулою 3, а $(\overline{x_n})$ - послідовність середніх Чезаро. ....*

3.4 Висновок

Отже, можна використати результати цієї статті для випадку, коли на кожному кроці з сукупності операторів обирається лише один.

4 Випадок $A = \partial f$

Важливий частинний випадок нашої задачі - коли $A(\cdot) = \partial f(\cdot)$ - субградієнт. Відомо, що якщо f - власна функція (proper function), то $\partial f(\cdot)$ - монотонний оператор [3].

Одним із головних застосувань задачі є пошук екстремумів опуклих недиференційовних функцій. Аналогом теореми Ферма є таке твердження [3]:

Теорема 2 (Теорема Ферма для опуклих функцій). *Нехай $f \in \Gamma_0(\mathcal{H})$, тоді*

$$\arg\min f = zer\, \partial f = \{x \in \mathcal{H} \mid 0 \in \partial f(x)\}$$

Всюди надалі будемо говорити про мінімізацію цільової функції:

$$f \to \min \tag{5}$$

.

Зазвичай ми робимо такі припущеня про $f(x)$:

1. $f : \mathcal{H} \to (-\infty, \infty]$ - власна (proper), замкнена і опукла.
2. Множина розв'язків задачі 5 є непорожньою.
3. МИ маємо доступ до оракула, який для довільної точки x видає випадковий субградієнт $g(x) \in \partial f(x)$.

4.1 Градієнтний спуск

У випадку гладкої функції існує градієнтний метод пошуку екстремуму. Ітераційний процес методу для задачі мінімізації має вигляд:

$$x_{n+1} = x_n - \lambda \nabla f(x_n)$$

, де λ вибирається таким чином, щоби гарантувати збіжність методу. Також є різні оцінки швидкодії градієнтного спуску, які потребують, щоби λ вибиралася за певною формулою.

По аналогії, можемо розглянути субградієнтний метод:

$$x_{n+1} = x_n - \lambda g(x_n) \tag{6}$$

Отримали, замінивши ∇f на $g(x_n) \in \partial f(x_n)$ - довільний субградієнт. Для цього алгоритму справджується така теорема:

Теорема 3 (Збіжність алгоритму субградієнтного спуску). *Нехай окрім згаданих вище умов справджується:*

1. *Існує така стала $M > 0$, що для всіх $x \in \mathbb{R}^n$ і для всіх $g(x) \in \partial f(x)$ виконується*

$$\|g(x)\|_2 \leq M$$

Нехай x_* - найближчий розв'язок до x_0, $R_0 = \|x_0 - x_*\|_2$ та $\lambda = \frac{R_0}{M\sqrt{N+1}}$. Тоді через N ітерацій градієнтного спуску маємо

$$f(\overline{x}_N) - f(x_*) \leq \frac{MR_0}{\sqrt{N+1}}$$

де $\overline{x}_N \frac{1}{N+1} \sum_{k=0}^{N} x_k$. Окрім того, для досягнення точності ϵ по функції $(f(\overline{x}_N) - f(x_*))$ метод достатньо запустити на $N = O(\frac{M^2 R_0^2}{\epsilon^2})$

4.2 Проксимальний алгоритм

Нехай цільова функція розкладається є сумою двох частин, і задача має вигляд:

$$F(x) = f(x) + R(x) \to \min_{x \in \mathbb{R}^n}$$

за теоремою Ферма та властивостями субградієнта, ця задача еквівалентна такій:

$$0 \in Ax + Bx, \text{ де } A = \partial f(x), B = \partial R(x)$$

Тоді використовується проксимальний алгоритм, який в класичному випадку має вигляд:

$$x_{n+1} = prox_{\lambda R}(x_n - \lambda \nabla f(x_n)), \text{ де } prox_R(x) = \arg\min_{y \in \mathbb{R}^n} \{R(y) + \frac{1}{2}\|y - x\|_2^2\}$$

Можна отримати субградієнтний проксимальний метод, замінивши градієнт $\nabla(\cdot)$ на довільний субградієнт $g \in \partial(\cdot)$:

$$x_{n+1} = prox_{\lambda R}(x_n - \lambda g(x_n))$$

Зазвичай цей алгоритм застосовують тоді, коли можна легко обчислити $prox_R$.

4.3 Проекційний алгоритм для пошуку умовного екстремуму

Класичний вигляд ітераційного процесу:

$$x_{n+1} = P_C(x_n - \lambda_n \nabla f(x_n))$$

Як і раніше, переходимо від градієнтів $\nabla f(x_n)$ до субградієнтів $g(x_n) \in \partial f(x_n)$, отримаємо проекційний субградієнтний метод:

$$x_{n+1} = P_C(x_n - \lambda_n g(x_n))$$

Можна сформулювати теорему про збіжність, не враховуючи стохастичну природу величини $g(x_n)$. У книзі [4] наведений результат, згідно з яким метод буде збігатися незалежно від того, яким буде $g(x_n)$. Також ця теорема дає оцінку швидкості збіжності.

Інший підхід - розглянути послідовність ітерацій алгоритму (x_n) як послідовність випадкових величин. У книзі [4] доводена збіжність послідовності *найкращих отриманих значень*:

$$\mathbb{E}(f_{best}^{(n)}) \to f_{opt} \text{ при } k \to \infty$$

A Основні визначення та позначення

- $\mathcal{H}$ - гільбертовий простір зі скалярним добутком $\langle \cdot | \cdot \rangle$ та нормою $\| \cdot \|_2$.
- *Субдиференціал* опуклої функції $f(x)$ в точці x - це множина $\partial f(x) := \{x \in \mathcal{H} \mid \forall y \in \mathcal{H} \ \langle y - x | f(y) - f(x) \rangle \geq 0\}$. Субдиференціалом також називається відповідне багатозначне відображення.

  Субградієнтом називається будь-який елемент цієї множини.
- $\Gamma_0(\mathcal{H})$ - множина власних (proper) напівнеперервних знизу функцій з $\mathcal{H}$ в $(-\infty, \infty]$.
- *Проксимальний оператор* $prox_f(x) := \arg\min_{y \in \mathcal{H}} \{f(y) + \frac{1}{2}\|y - x\|_2^2\}$, де $f \in \Gamma_0(\mathcal{H})$.
- *Оператор проектування на множину* $P_C := \arg\min_{y \in C} \|x - y\|^2$. Якщо C - непорожня, замкнена та опукла множина, то для кожного $x \in \mathcal{H}$ існує єдина проекція $P_C(x)$.

Література

[1] В. В. Семенов *Параллельная декомпозиция операторных включений с максимальными монотонными операторами* . Журнал обчислювальної та прикладної математики, 2013, No 2 (112), УДК 517.9 [Journal of Computational & Applied Mathematics]

[2] Pascal Bianchi *Ergodic convergence of a stochastic proximal point algorithm*, 2016, https://arxiv.org/abs/1504.05400

[3] Heinz H. Bauschke, Patrick L. Combettes *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2011 , DOI 10.1007/978-1-4419-9467-7

[4] Amir Beck *First-order methods in optimizations*, 2017, https://doi.org/10.1137/1.9781611974997