Stochastic optimization. Coordinate method

Optimization in ML

Aleksandr Beznosikov

Skoltech

12 December 2023

Поворот на 90 градусов

- Простой пример:

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{n} \sum_{i=1}^{n} \|a_i^T x - b_i\|_2^2 \right],$$

где $\{a_i, b_i\}_{i=1}^n$ – обучающая выборка.

Поворот на 90 градусов

- Простой пример:

$$\min_{x\in\mathbb{R}^d}\left[f(x)=\frac{1}{n}\sum_{i=1}^{n}\|a_i^T x-b_i\|_2^2\right],$$

где $\{a_i, b_i\}_{i=1}^n$ – обучающая выборка.

- До этого, мы брали не всю выборку для подсчета градиента, чтобы быть более эффективными. Т.е. использовали только часть строк матрицы A, составленной из a_i **Вопрос:** а как по-другому можно добиться эффективности?

Поворот на 90 градусов

- Простой пример:

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{n} \sum_{i=1}^{n} \|a_i^T x - b_i\|_2^2 \right],$$

  где $\{a_i, b_i\}_{i=1}^n$ – обучающая выборка.
- До этого, мы брали не всю выборку для подсчета градиента, чтобы быть более эффективными. Т.е. использовали только часть строк матрицы A, составленной из a_i **Вопрос:** а как по-другому можно добиться эффективности? если до этого был выбор строк матрицы A, то теперь можно попробовать как-то завязаться на столбцы. **Вопрос:** а что означает «выбор столбцов»?

Поворот на 90 градусов

- Простой пример:

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{n} \sum_{i=1}^{n} \|a_i^T x - b_i\|_2^2 \right],$$

  где $\{a_i, b_i\}_{i=1}^n$ – обучающая выборка.
- До этого, мы брали не всю выборку для подсчета градиента, чтобы быть более эффективными. Т.е. использовали только часть строк матрицы A, составленной из a_i **Вопрос:** а как по-другому можно добиться эффективности? если до этого был выбор строк матрицы A, то теперь можно попробовать как-то завязаться на столбцы. **Вопрос:** а что означает «выбор столбцов»? Выбор координат вектора x.

Производная по направлению

- Часто для более сложных задач к подсчету производных по координатам/направлениям прибегают не из-за удешевления процесса, а из-за доступности только оракула нулевого порядка (значения функции). Потому что производную по направлению $e \in \{e \in \mathbb{R}^d \mid \|e\|_2 \leq 1\}$ можно аппроксимировать через конечную разность:
$$[\nabla f(x)]_i \approx \frac{f(x + \tau e) - f(x - \tau e)}{2\tau}$$
(таким образом можно «собрать» и весь «градиент»).

Координатный метод

Algorithm 1 Координатный метод

Input: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, значения памяти $y_i^0 = 0$ для всех $i \in [n]$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Сгенерировать независимо i_k из $[d]$
3: Вычислить $[\nabla f(x^k)]_{i_k}$
4: $x^{k+1} = x^k - \gamma \cdot d[\nabla f(x^k)]_{i_k} e_{i_k}$
5: **end for**

Output: x^K

Здесь e_{i_k} – iый базисный вектор

Координатный метод

Algorithm 2 Координатный метод

Input: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, значения памяти $y_i^0 = 0$ для всех $i \in [n]$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Сгенерировать независимо i_k из $[d]$
3: Вычислить $[\nabla f(x^k)]_{i_k}$
4: $x^{k+1} = x^k - \gamma \cdot d[\nabla f(x^k)]_{i_k} e_{i_k}$
5: **end for**

Output: x^K

Здесь e_{i_k} – iый базисный вектор

- Зачем в шаге метода есть домножение на d?

Координатный метод

Algorithm 3 Координатный метод

Input: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, значения памяти $y_i^0 = 0$ для всех $i \in [n]$, количество итераций K

1: **for** $k = 0, 1, \dots, K-1$ **do**

2: Сгенерировать независимо i_k из $[d]$

3: Вычислить $[\nabla f(x^k)]_{i_k}$

4: $x^{k+1} = x^k - \gamma \cdot d[\nabla f(x^k)]_{i_k} e_{i_k}$

5: **end for**

Output: x^K

Здесь e_{i_k} – iый базисный вектор

- Зачем в шаге метода есть домножение на d? Для несмещенности того, что мы используем вместо градиента.

Координатный метод: доказательство

- f является L-гладкой и μ - сильно выпуклой.

Координатный метод: доказательство

- f является L-гладкой и μ - сильно выпуклой.
- Уже привычно:

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\gamma\langle d[\nabla f(x^k)]_{i_k} e_{i_k}, x^k - x^*\rangle + \gamma^2\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2.$$

Координатный метод: доказательство

- f является L-гладкой и μ - сильно выпуклой.
- Уже привычно:

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\gamma \langle d[\nabla f(x^k)]_{i_k} e_{i_k}, x^k - x^* \rangle + \gamma^2 \|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2.$$

- Берем условное мат.ожидание по случайности только на итерации k:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] = \|x^k - x^*\|_2^2 - 2\gamma \langle \mathbb{E}\left[d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k\right], x^k - x^* \rangle + \gamma^2 \mathbb{E}\left[\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right].$$

Координатный метод: доказательство

- Работаем с $\mathbb{E}\left[d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k\right]$:

$$\mathbb{E}\left[d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k\right] = \frac{1}{d}\sum_{j=1}^{d} d[\nabla f(x^k)]_i e_i$$
$$= \nabla f(x^k)$$

Координатный метод: доказательство

- Теперь работаем с $\mathbb{E}\left[\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right]$:

$$\begin{aligned}\mathbb{E}\left[\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right] =& \mathbb{E}\left[\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right] \\ =& d^2\mathbb{E}\left[\|[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right] \\ =& d^2 \cdot \frac{1}{d}\sum_{j=1}^{d} \|[\nabla f(x^k)]_j e_j\|_2^2 \\ =& d\|\nabla f(x^k)\|_2^2\end{aligned}$$

Координатный метод: доказательство

- Промежуточный итог:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] = \|x^k - x^*\|_2^2 - 2\gamma\langle \mathbb{E}\left[d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k\right], x^k - x^*\rangle + \gamma^2 \mathbb{E}\left[\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right].$$

$$\mathbb{E}\left[d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k\right] = \nabla f(x^k)$$

$$\mathbb{E}\left[\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right] = d\|\nabla f(x^k)\|_2^2$$

Координатный метод: доказательство

- Промежуточный итог:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2 \mid x^k\right] = \|x^k-x^*\|_2^2 - 2\gamma\langle \mathbb{E}\left[d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k\right], x^k - x^*\rangle + \gamma^2 \mathbb{E}\left[\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right].$$

$$\mathbb{E}\left[d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k\right] = \nabla f(x^k)$$

$$\mathbb{E}\left[\|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k\right] = d\|\nabla f(x^k)\|_2^2$$

- Собираем вместе:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2 \mid x^k\right] \leq \|x^k-x^*\|_2^2 - 2\gamma\langle \nabla f(x^k), x^k - x^*\rangle + d\gamma^2\|\nabla f(x^k)\|_2^2.$$

Координатный метод: доказательство

- Сильная выпуклость и гладкость функции f:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] \leq (1 - \mu\gamma)\|x^k - x^*\|_2^2 - 2\gamma(1 - d\gamma L)(f(x^k) - f(x^*)).$$

Координатный метод: доказательство

- Сильная выпуклость и гладкость функции f:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] \leq (1 - \mu\gamma)\|x^k - x^*\|_2^2 - 2\gamma(1 - d\gamma L)(f(x^k) - f(x^*)).$$

- Пусть $\gamma \leq \frac{1}{dL}$:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] \leq (1 - \mu\gamma)\|x^k - x^*\|_2^2.$$

Координатный метод: сходимость

Теорема сходимость (координатный метод))

Пусть задача безусловной оптимизации с L-гладкой и μ-сильно выпуклой целевой функцией f решается с помощью координатного метода с $\gamma \leq \frac{1}{dL}$. Тогда справедлива следующая оценка сходимости

$$\mathbb{E}\left[\|x^k - x^*\|_2^2\right] \leq (1 - \mu\gamma)^k \mathbb{E}\left[\|x^0 - x^*\|_2^2\right]$$

Координатный метод: сходимость

- Подставив $\gamma = \frac{1}{dL}$, получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu}\log\frac{1}{\varepsilon}\right).$$

Координатный метод: сходимость

- Подставив $\gamma = \frac{1}{dL}$, получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu}\log\frac{1}{\varepsilon}\right).$$

Вопрос: есть ли улучшения по сравнению с обычным градиентным спуском?

Координатный метод: сходимость

- Подставив $\gamma = \frac{1}{dL}$, получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu}\log\frac{1}{\varepsilon}\right).$$

Вопрос: есть ли улучшения по сравнению с обычным градиентным спуском? В общем случае нет. Это доказуемо так.

Координатный метод: сходимость

- Подставив $\gamma = \frac{1}{dL}$, получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu}\log\frac{1}{\varepsilon}\right).$$

  Вопрос: есть ли улучшения по сравнению с обычным градиентным спуском? В общем случае нет. Это доказуемо так.

- Если есть дополнительная информация о задаче (например, свойства констант Липшица градиента по направлению), то улучшения можно получить.

Координатный метод: сходимость

- Подставив $\gamma = \frac{1}{dL}$, получаем следующую итерационную сложность
$$\mathcal{O}\left(\frac{dL}{\mu}\log\frac{1}{\varepsilon}\right).$$
Вопрос: есть ли улучшения по сравнению с обычным градиентным спуском? В общем случае нет. Это доказуемо так.
- Если есть дополнительная информация о задаче (например, свойства констант Липшица градиента по направлению), то улучшения можно получить.
- Еще координатный метод часто хорошо себя проявляет на практике.

Координатный метод: сходимость

- Подставив $\gamma = \frac{1}{dL}$, получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu}\log\frac{1}{\varepsilon}\right).$$

  Вопрос: есть ли улучшения по сравнению с обычным градиентным спуском? В общем случае нет. Это доказуемо так.
- Если есть дополнительная информация о задаче (например, свойства констант Липшица градиента по направлению), то улучшения можно получить.
- Еще координатный метод часто хорошо себя проявляет на практике.
- Результат обобщается и на случай выбора нескольких координат.
- Возможно ускорение с помощью двух моментумов.

Координатный метод: редукция дисперсии

- Идею SAGA можно использовать и здесь.